Circuit-Aware System Design Techniques for Wireless Communication

by

Everest Wang Huang

S.B., Massachusetts Institute of Technology (1996)
S.B., Massachusetts Institute of Technology (1998)
M.Eng., Massachusetts Institute of Technology (1998)

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

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2006

Author .....
Department of Electrical Engineering and Computer Science
January 23, 2006

Certified by .....
Gregory W. Wornell
Professor
Thesis Supervisor

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

Circuit-Aware System Design Techniques for Wireless Communication

by
Everest Wang Huang

Submitted to the Department of Electrical Engineering and Computer Science
on January 23, 2006, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

When designing wireless communication systems, many hardware details are hidden from the algorithm designer, especially with analog hardware. While it is difficult for a designer to understand all aspects of a complex system, some knowledge of circuit constraints can improve system performance by relaxing design constraints. The specifications of a circuit design are generally not equally difficult to meet, allowing excess margin in one area to be used to relax more difficult design constraints.

We first propose an uplink/downlink architecture for a network with a multiple antenna central server. This design takes advantage of the central server to allow the nodes to achieve multiplexing gain by forming virtual arrays without coordination, or diversity gain to decrease SNR requirements. Computation and memory are offloaded from the nodes to the server, allowing less complex, inexpensive nodes to be used.

We can further use this SNR margin to reduce circuit area and power consumption, sacrificing system capacity for circuit optimization. Besides the more common transmit power reduction, large passive analog components can be removed to reduce chip area, and bias currents lowered to save power at the expense of noise figure. Given the inevitable crosstalk coupling of circuits, we determine the minimum required crosstalk isolation in terms of circuit gain and signal range. Viewing the crosstalk as a static fading channel, we derive a formula for the asymptotic SNR loss, and propose phase randomization to reduce the strong phase dependence of the crosstalk SNR loss.

Because the high peak to average power (PAPR) that results from multicarrier systems is difficult for analog circuits to handle, the result is low power efficiencies. We propose two algorithms, both of which can decrease the PAPR by 4 dB or more, resulting in an overall power reduction by over a factor of three in the high and low SNR regimes, when combined with an outphasing linear amplifier.

Thesis Supervisor: Gregory W. Wornell
Title: Professor

Acknowledgments

Since this thesis was a long time labor of love, the number of people who have helped me over these many years have grown to almost epic proportions. As a result, I may accidentally neglect to mention some people by name, but be assured, I am most grateful for your help and wish I had a better memory.

First, I would like to thank Greg Wornell, who I consider not just my advisor, but also my friend. As an advisor, his hands-off philosophy to research allowed me to explore different topics of interest until I found what I was really interested in, so that my research felt like my own. While I was free to shape my research topic, he was always there to nudge me towards areas where fruitful results may lie, and away from areas where I might fall over the edge of a cliff, figuratively speaking. In addition he was always friendly and available to talk about just about any topic (which we did), and oh, all those confounding puzzles!

I would also like to thank my committee members, Anantha Chandrakasan and Charles Sodini, who along with Greg gave me excellent feedback and asked many probing questions, which improved my thesis by at least and order of magnitude. Although I collaborated with many students over the course of my research, I am most grateful to Lunal Khuon and Anh Pham for teaching about the world of analog circuits, and their own circuits in particular. From their test measurements, and from discussions with each of them and with Farinaz Edalat, I was able to understand, at least a little, how the analog world works.

During my doctoral studies, I have been fortunate to be a member of both the Signals, Information, and Algorithms Laboratory and the Digital Signal Processing Group. Interacting with all of the many group members has made my time at MIT a real joy, and hard to give up. I enjoyed interacting with all of you, but I would especially like to thank Richard Barron, Lane Brooks, Albert Chan, Vijay Divi, Stark Draper, Uri Erez, Ashish Khisti, Nick Laneman, Mike Lopez, Emin Martinian, Charles Sestok, Charles Swannack, and Huan Yao for many fruitful research discussions. I would also like to thank Giovanni Aliberti for his many efforts to keep the computers up and running even given a resource hog such as myself, and Tricia Mulcahy for taking care of all the many behind-the-scene details to make sure things ran smoothly, especially near the end. Not to be forgotten as well is Cindy LeBlanc, who was always looking out for me, even though I wasn't even in her group.

I also learned a great deal from the many summer internships I had, both at Texas Instruments and at Lincoln Laboratory. My supervisors at these places, Don Boroson, Alan Gatherer, and Mike Polley, along with Mike Direnzo, Dale Hocevar, Tarik Muharemovic, and Eko Onggosanusi were all great sources of ideas and information which helped my learn about the world from a different perspective than I would get just staying at MIT.

Among the many friends I made at MIT, I'd like to thank James Geraci, Brian Heng, and Wade Wan, both for helping me push around heavy objects at the gym to keep my from getting too flabby, but also for hanging around MIT as long as they did to make sure there was a way for me to get away from graduate student life for a while and see the rest of the world.

Most importantly, I want to thank my wife, Regina, who has always been there for me, loving and supportive through all my (many) years at MIT and beyond. I couldn't have gotten this far without you in my life. I would also like to thank the rest of my family, both my and Regina's parents, and her sister Carol, for their ever-present love and support. Finally, I would like to thank our new baby daughter Megan, who has not only become with Regina my two favorite people in the whole world, but was also considerate enough to wait three days to be born so that I would have enough time to finish the final draft of this thesis.

This research was generously funded by MIT, the Semiconductor Research Corporation and MARCO C2S2, and Lincoln Laboratory.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In all but the simplest systems, it is often necessary to separate the design into many layers separated by abstraction barriers. The abstraction barriers isolate the details of a section of the system from all other system parts. These layers form a stacked protocol, in which a layer can only communicate with the layer immediately above and below it. In principle, all of the implementation details within a given layer are hidden from any layer above or below it. This allows a designer to focus on designing the algorithms in his/her own layer without having to know how the entire system is going to work. In addition, the layering allows the algorithms in a particular layer to be changed without affecting any of the other layers, enabling algorithm reuse for many different applications, as well as allowing improvements to be made without disrupting existing infrastructure.

In practice, however, this strict layering can cause design choices in one layer to make unreasonable demands on the neighboring layers. For this reason, the design in each layer often incorporates some knowledge of the details of nearby layers. The amount of information that is shared between the layers depends on their proximity, with neighboring layers more likely than distant layers to share information.

1.1 The Hardware Abstraction Layer

For a communications system, Figure 1-1 represents the bottom few layers of the protocol stack (adapted from the TCP/IP stack). The Physical layer is the bottom-most layer, and is concerned with how the data passed to it from the MAC (Media Access Control) layer is going to be transmitted and received. At the transmitter, the Physical layer receives a bit stream and processes it to an analog waveform to be transmitted. At the receiver, the Physical layer receives an analog waveform from the communications channel, and converts it back into a bit stream, which is passed to the layer above. In this thesis we will focus on aspects of the Physical layer.

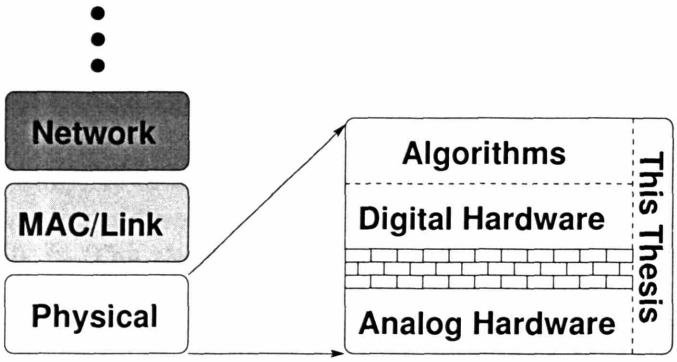


Figure 1-1: The bottom layers of the communications stack, with a brick wall indicating the relative lack of interaction between the algorithm and analog hardware layers. We exploit some knowledge of analog hardware impairments to influence algorithm design.

The Physical layer can be further subdivided into algorithm, digital, and analog hardware implementation layers. Although these layers do not officially exist, they are present in practice as algorithms, digital and analog hardware are all typically designed with only limited information about the other layers. The algorithm designer usually works in the theoretical realm, where precision can be exact and linearity is perfect, with infinite dynamic range and noise is absent or artificially introduced. These algorithms are then implemented on digital hardware, and then output through the analog hardware to interface with a communications channel. Each of these steps is separated by an abstraction barrier, but the strength of these barriers is very different.

Digital circuit designers must contend with issues of power, computational complexity and speed, dynamic range, and quantization errors. If the algorithm designer is aware of these limitations, the algorithms can be designed to be less computationally taxing or more tolerant to quantization errors. Similarly, the digital hardware designer can optimize the circuit architecture for running a particular type of algorithm so that it is more efficient for the given task. Because of these advantages, the algorithms and digital hardware are often designed with each other in mind.

Analog circuit designers must also contend with hardware that is far from ideal. For example, issues of nonlinearity, limited dynamic range, component noise and crosstalk, and power efficiency can all hinder analog circuit performance. The algorithm designer, unaware of the nature of this nonideal behavior, may place unrealistic demands on the analog hardware while optimizing algorithm performance, requiring large design margins for the hardware to function. Unlike with digital circuits, analog circuitry cannot be automatically synthesized, with difficult requirements placing greater demands on the circuit designer. In addition, analog circuits are tuned to

the specifications of a given system, often preventing them from being reused for a different system without sometimes extensive modification.

In Figure 1-1, a dotted line separates the algorithm and digital hardware layers, representing the weak abstraction barrier which allows significant details about each layer to be shared. On the other hand, a brick wall separates the analog hardware layer from the digital hardware and algorithm designer, representing the very strong abstraction barrier that separates them. The algorithm designer is already one layer removed from the analog hardware, and due to the large differences in technical knowledge required, fewer of the details of the analog hardware design passes up to the algorithm designer. For example, the algorithm designer will often have a model of the nonideal circuit behavior, but without detailed enough knowledge to know how changing the circuit specifications will affect other circuit parameters. This thesis attempts to bridge this gap by using some knowledge of analog hardware impairments and implementation to guide algorithm design.

Removing the abstraction barrier would seem to be a good solution, since a joint optimization should be better than optimizing each layer individually. Removing the layering allows more optimization, but at the cost of a greatly increased workload on the system designer. However, even some cross-layer knowledge can significantly increase the efficient use of resources. Many things are easier to do digitally than with analog components, but the digital versions may come at the expense of circuit size and power. Similarly, there are many tasks that are well-suited to analog circuit techniques and cannot easily be done with digital circuits.

As a general rule, the different specifications of a design are not all equally difficult to meet, thus the performance of a system is limited by its most difficult to meet constraint. By understanding some details of the circuit behavior, the algorithm designer can learn which specifications are particularly difficult for the analog circuit designer to meet, and which are easy to meet. The excess margin from the easier to meet constraints can then be traded off to relax the more difficult design constraints, leading to novel architectures and improved designs. Significant gains in power and area can be made with these cross-layer designs. The effects of nonidealities from the analog circuits can be reduced by these techniques as well. We explore several problems that can occur with analog circuits and algorithmic techniques that can either alleviate or are tolerant of these limitations.

1.2 High-level Physical Layer System View

A block diagram of a typical wireless system is shown in Figure 1-2. The data stream to be transmitted is first encoded to provide error protection, modulation, pulse shaping, etc. in the digital domain. The digital waveform is then converted to analog and

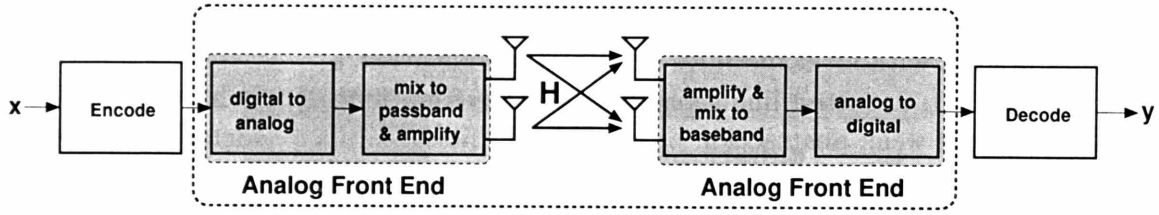


Figure 1-2: Wireless system block diagram. The circuits and other hardware inside the dashed rectangle are typically black boxed by the algorithm designer.

mixed up to the carrier frequency. A power amplifier (PA) drives the signal onto the antennas and through the wireless channel. At the receiver, the signals picked up by the receive antennas are then amplified by a low noise amplifier (LNA) and mixed back down to a baseband signal. This waveform is then sampled by the analog to digital converter. The shaded boxes are part of the analog front ends, which are squarely in the domain of the analog circuit designer. For the communication algorithm designer, the entire dashed box is usually abstracted away into a simplified wireless channel model. Hardware impediments are encapsulated by simplified models, which do not capture the interdependence of the different circuit characteristics. As we will show, some cross-layer knowledge allows the system to be improved in circuit area and power, and more resistant to circuit nonidealities.

1.3 The WiGLAN

For wired networks, gigabit per second (Gb/s) Ethernet networks are already commercially available. Although the wireless channel is much more hostile than the wired one, it would be desirable to have such a high bandwidth connection for wireless networks as well. Such a network might be the Wireless Gigabit LAN (WiGLAN) [14].

A central server with few constraints on transmit power, computational ability or memory controls and assists the peer-to-peer traffic among the multiple mobile nodes in the network. The nodes have limited battery and computational power, and are equipped with one or more antennas for transmission and reception. The central server is equipped with several antennas (for example, four), which can be used in any combination to transmit and receive information from the other nodes in the system. Within the WiGLAN, appliances throughout the home or office environment can communicate wirelessly among themselves and to the central network controller. Depending on the needs and type of each node, single and multiple antenna battery-powered portable adapters are attached to the various appliances. Devices that require high data rates and/or excellent link quality for large ranges (e.g. an HDTV or

DVD player) use multiple antennas while short range, low data rate appliances use single antennas.

The WiGLAN operates in the U-NII band at 5.25 GHz with a bandwidth of 150 MHz. The wide bandwidth means that the network only needs to be able to achieve an overall spectral efficiency of about 6-7 b/s/Hz. However, the wide bandwidth also brings frequency selective fading and complexity issues which must be dealt with. For these reasons, and to facilitate resource allocation, the total bandwidth is divided into many subcarriers, or frequency bins through Orthogonal Frequency Division Multiplexing (OFDM).

1.4 Thesis Outline and Contributions

In the following chapters, we explore in detail the implications and advantages of optimizing algorithms across the analog circuit abstraction barrier within the framework of the WiGLAN.

The wireless channel model is described in Chapter 2, both for single- and multiple-antenna links. The achievable performance is shown, as well as the trade offs involved in power, data rate, and error tolerance. Some basic circuit principles are also described to provide the necessary cross-layer knowledge of the analog circuitry.

Chapter 3 introduces the system architecture of the WiGLAN and the communications protocol it uses. Because of the asymmetrical capabilities of the nodes and the central server, the majority of the computation and memory requirements have been offloaded to the server. The uplink/downlink protocol allows the network nodes to achieve multiplexing gain without coordination, reduce transmit power, and increase transmission range by exploiting the capabilities of the central server.

Chapter 4 describes some of the limiting constraints for circuit area and power of receiver circuit designs. The SNR gain from using multiple antennas can be used to reduce the analog circuit area or power. Removing physically large passive components can reduce circuit area by a factor of three, while reducing bias currents to lower amplifier gain cuts amplifier power consumption by a factor of two or more. Gains in area and power can also be made for digital circuits by using the SNR gain to reduce the computational complexity.

Chapter 5 describes the effect of crosstalk between parallel circuits, with design guidelines as well as techniques to mitigate the severity of circuit crosstalk. This is especially important when there are parallel amplifiers present in a circuit which may create unstable feedback loops. A simple relationship between amplifier gain and crosstalk isolation prevents deep nulls or unstable feedback behavior for parallel circuit paths. A relation for the asymptotic SNR loss caused by the crosstalk is derived, and the beneficial effects of phase randomization is explored.

Chapter 6 outlines the problem of the high dynamic range requirements of many wideband systems including the WiGLAN, with the associated problem of poor amplifier efficiency for these waveforms. Peak to average power reducing algorithms, both with and without penalties in data rate are described which reduce the dynamic range requirement and significantly increases overall power efficiency. A precoding algorithm decreases the energy per bit requirement by more than a factor of three when coupled with an efficient linear amplifier at high SNR. At low SNR a synthesis algorithm achieves similar gains. It is also shown that the major distortion caused by hard signal clipping is not a loss in SNR, but instead bandwidth expansion.

Finally, Chapter 7 has conclusions and future research directions.

Chapter 2

Background

Since the WiGLAN system has 150 MHz of bandwidth, the spectral efficiency required to achieve data rates at or near a gigabit per second is less than seven bits per second per Hertz of bandwidth (b/s/Hz). Although it may be possible to transmit at or above this rate over some of the frequency band, the frequency- and time-dependent nature of the wireless channel requires adaptively changing the encoding to match the channel conditions. Dividing the available bandwidth into frequency bins via OFDM is described, as well as the achievable limits of the communication links. In addition, a brief overview is given for some of the nonidealities that must be accounted for in analog circuitry, along with some of the major concerns for both analog and digital circuit designs. This knowledge presents opportunities for algorithms to take advantage of circuit details that are often otherwise hidden from the algorithm designer.

2.1 Transmission via OFDM

Although the WiGLAN has a large amount of bandwidth to work with, it is not feasible, or even desirable for each transmitter to use all of the available bandwidth for each transmission. In a multi-user environment, it is inefficient for each user to be assigned time slots in which the entire frequency bandwidth is available for use. In addition to the requirement of much faster data converters and hardware to process data so quickly, the multipath fading over such a large bandwidth is not independent of frequency. It then becomes necessary to train and equalize the signals for the path between any two nodes which might be transmitting. This is especially taxing for small mobile nodes with very limited computational power and low data requirements. Additionally, even if the node has the capability to process the entire band, it might not need all of the available bandwidth, wasting the surplus for that time slot.

One method of allocating the bandwidth is to use Quadrature Amplitude Modu-

lation (QAM) on the subcarriers of an OFDM symbol. The advantage of OFDM is that the symbol is divided into a large number of frequency bins or subcarriers which can then be allocated to the nodes that wish to communicate at a given time in ratios according to their requested rates. The bandwidth of each frequency bin is small enough that the the multipath fading is essentially constant over the bin, reducing the equalization to only a single multiplicative constant to account for the amplitude and phase shift of that subcarrier.

2.1.1 Shape of OFDM Frequency Bins

A convenient subchannelization for an OFDM symbol is to use the frequency bins of a Discrete Fourier Transform (DFT). To convert the modulated subcarriers (frequency bins) into a time signal, one can simply apply the Inverse Fast Fourier Transform (IFFT), which has complexity that grows with $\mathcal{O}(N \log N)$ where N is the number of frequency bins. At the receiver, the Fast Fourier Transform (FFT) will convert the OFDM time symbol back to frequency bins, again with $\mathcal{O}(N \log N)$ complexity.

As shown in Figure 2-1, these frequency bins have the shape of a $\sin(x)/x$ pulse, which are orthogonal to each other, but have side lobes that fall off very slowly as $1/x$. As a result, if there is any kind of inter-symbol interference (ISI), then a frequency bin will interfere significantly with neighboring bins. To deal with the impulse response of the channel, a cyclic prefix can be prepended to the OFDM symbol, which has the effect of making the convolution of the OFDM symbol with the channel impulse response the same as the circular convolution. When the OFDM symbol is transformed to the frequency domain, each sample (frequency bin) is simply scaled by a complex number that is the channel frequency response at that frequency. As long as the prefix length is at least as long as the channel impulse response, this property holds, even if the channel characteristics change with every symbol. In order to be able to correct for the channel, however, there still needs to be some training data sent along with each OFDM symbol.

This is a very optimistic result, however, and not always true in practice because the length of the channel impulse response will invariably be longer than the chosen prefix length, or the required prefix length will be so long that it significantly impacts the data rate. The cyclic prefix then serves to reduce the ISI, making equalization easier, but not eliminating it altogether.

2.1.2 The Cyclic Prefix

If the channel impulse response is not finite length (or is very long), then the cyclic prefix will be so large that the code rate of an OFDM symbol becomes too small. If the frequency bins were more localized in frequency, then less work could be put

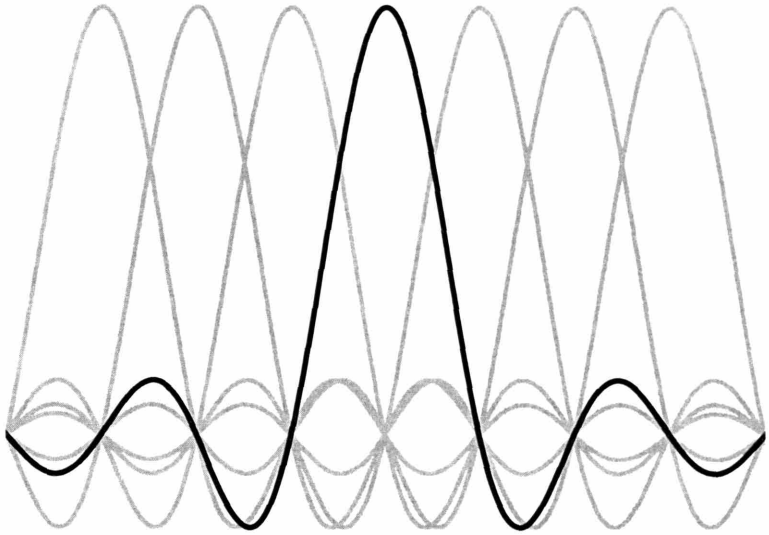


Figure 2-1: OFDM frequency bin shapes, with single bin highlighted.

into removing the ISI caused by the channel. In the extreme, the frequency bins could not overlap at all, but that makes the time series associated with each channel arbitrarily long. Although arbitrary subchannel shapes can be used as long as they are orthogonal, it is desirable to have an efficient algorithm, such as the FFT, to easily code and decode the OFDM symbol. The length of the cyclic prefix is dependent on the delay spread of the wireless channel, which is a measure of the average impulse response of the channel.

OFDM works well in the wireless domain because the delay spread in the channel is not considerable compared to the symbol duration as long as the signal bandwidth is relatively narrow. In a relatively wide bandwidth system like the WiGLAN, the cyclic prefix needed to combat ISI can be a significant fraction of the total symbol time [30].

2.1.3 Size of the OFDM Frequency Bins

The measure of how small a frequency bin needs to be to have essentially flat (frequency nonselective) fading is the coherence bandwidth, which in the 5 GHz band has been measured to be about 10 MHz for line of sight transmissions and about 4 MHz for transmissions without line of sight [39]. This suggests that the frequency bins should not be more than 4 MHz wide. There is a trade off in the number of frequency bins, since narrower bins facilitate resource allocation and are more likely

to experience flat fading, but the larger number of bins requires more computations. Because the number of frequency bins is the length of the IFFT input, increasing the number of frequency bins also increases the resulting symbol duration. More frequency bins will also increase the peak to average power of the resulting time signals, which places increased demands on the capabilities of the analog front ends.

The computational penalty for a larger number of frequency bins is offset by potential for increased parallelism. Since each bin can be decoded individually, a large number of slower, cheaper processors can be employed rather than a single expensive one. More frequency bins also allows finer granularity in the data rates that can be assigned to individual nodes, and make it easier to reject a narrowband interferer by only masking out a small chunk of frequency.

With the flat fading of the frequency bins, all that needs to be measured is the frequency response of the channel, which can be learned from training data included at the start of each packet. The receiver does not need a separate equalizer to equalize the channel, since each frequency bin can be equalized by multiplying with the appropriate constant. Even if the channel characteristics are different with every packet, since training data is always present, the channel parameters can be estimated. If an equalizer was used, then it might not be able to keep up with the changing channel parameters. This greatly reduces the amount of computation needed by a node to transmit, and the amount of training and computation needed increases with the requested data rate, so low-rate devices do not need to do as much calculation as a high-rate device, and can therefore be smaller and simpler. With all of the frequency bins, it is also possible for a node to slowly hop in frequency by being assigned different frequency bins over time by the central server.

2.1.4 Adaptive Modulation

The frequency bins also allow the system to adaptively deal with time variations in the channel. If a bin has a large amount of interference from outside sources (such as from a nearby, but separate, network), then the central server can neglect to allocate that bin to any node wishing to transmit. Additionally, if the data rate requested for a node varies with time, then additional frequency bins can be added or subtracted from the bandwidth allocated to that node to account for the change in the data rate. The node could instead be allocated differing numbers of time slots to vary its data rates as well. The modulation for each subcarrier can also adaptively change given the current signal to noise ratio (SNR) (such as 256-QAM for a bin with a very high SNR, 4-QAM for a bin with low SNR, or no transmission at all for a very noisy frequency bin) to maximize the amount of data that can be transmitted at a given time.

Figure 2-2 gives an example of the channel allocation and adaptive modulation for

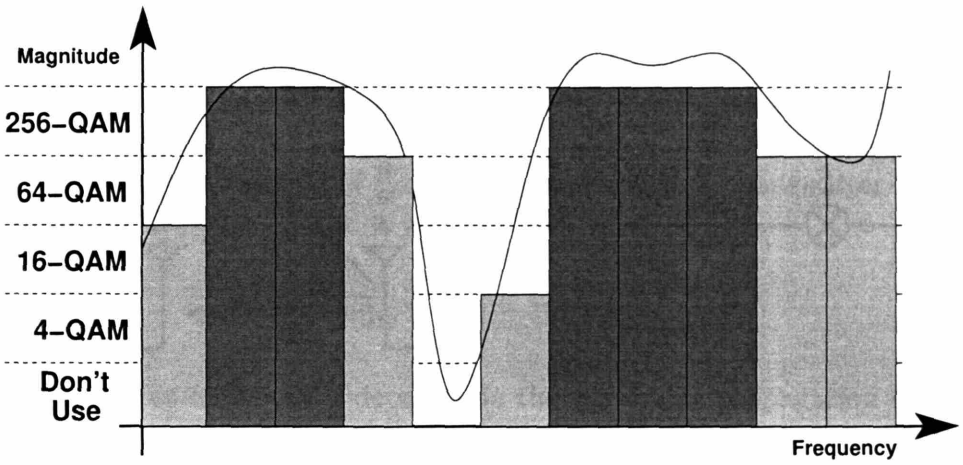


Figure 2-2: Adaptive modulation per frequency bin vs. measured channel response.

a hypothetical channel frequency response. The curved line is the frequency response of the channel, and the rectangular boxes represent idealized frequency bins. The shaded bins have been allocated according to the magnitude of the frequency response. If there is a deep fade (such as in the fifth bin), then that bin is not used at all.

This adaptive modulation approaches the optimal waterpouring scheme from a capacity standpoint [11]. However, most of the capacity can be achieved by simply deciding whether or not to use a use a frequency bin based on its relative SNR, and then using the largest supportable modulation across all of these bins. In Figure 2-2, the dark shaded bins show one way to do this. The total raw data rate for this particular channel is 64 b/s. If we use only the dark shaded bins, the data rate will be 40 b/s. If, on the other hand, we use all bins that can support 64-QAM, the data rate can be 48 b/s. This is a fairly simple optimization problem, but it saves the transmitter and receiver having to agree on the modulation for each individual bin. It is important to note that this threshold is in general different for each node as well as for each packet transmitted.

2.2 The Wireless Channel Model

The wireless channel is modeled as an equivalent discrete-time baseband system, as shown in Figure 2-3. An $M \times N$ system has M transmit and N receive antennas, resulting in MN distinct wireless channels. The system can be modeled as $\mathbf{y} = \mathbf{H}\mathbf{x}+\mathbf{n}$. Here, $\mathbf{x} = [x_1\, x_2\, x_3\, \cdots\, x_M]^{\mathrm{T}}$, where x_i is the symbol transmitted from the ith antenna. The received signal $\mathbf{y}$ and the receiver noise $\mathbf{n}$ are defined similarly for the N receive

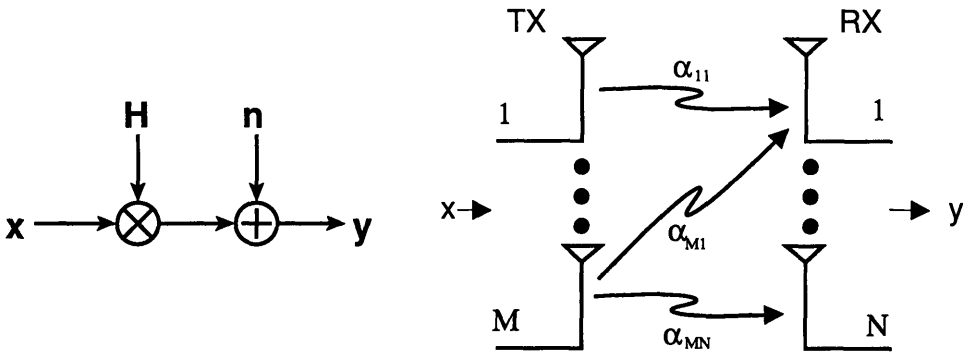


Figure 2-3: The wireless channel model.

antennas, and the channel is a matrix **H**, where $h_{ij} = \alpha_{ij}$ is the fading constant between the ith transmit and jth receive antenna.

Each receive channel is given by $y_j = \sum_i \alpha_{ij} x_i + n_j$, where x_i is the (usually complex) transmitted signal, α_{ij} is the channel fading parameter, and n_j is noise. The noise is modeled as complex Gaussian white noise with zero mean and variance $N_o/2$ in each dimension. This corresponds to the thermal noise of the amplifiers at each receive antenna. The channel fading parameter is also a complex zero mean Gaussian variable with variance 1/2 in each dimension. Equivalently, the amplitude of α_{ij} is Rayleigh-distributed with unit mean and uniformly distributed phase. This models a system in which there is no line of sight path between the transmitter and receiver, so all the energy from the transmitter arrives with random phases from all directions.

The value of n_j is randomly chosen according to its probability distribution at every channel use, but α_{ij} is constant over an entire packet of data (block fading). The fading constant is independent and identically distributed between packets. Thus for each packet, the system only needs to learn the statistics of the channel once, but it must relearn them for every packet. The value of α_{ij} is randomly chosen for each packet and each channel, so for multiple antenna systems, all channels are independent and identically distributed. We use $\mathcal{N}(m, \sigma^2)$ to denote a Gaussian random variable with mean m and variance σ^2, and similarly $\mathcal{N}^c(m, \sigma^2)$ to denote a corresponding complex Gaussian random variable with mean m and complex variance σ^2.

Not included in the channel matrix **H** is the path loss associated with the physical separation plus any power absorbing material between the transmitter and receiver. The average signal power at the receiver front end input is

$$P_R = P_T - P_L, \tag{2.1}$$

where P_R is received power, P_T is transmitted power, and P_L is the path loss (all in dB). Because the system noise is modeled at the receiver only, the receive SNR varies directly with the path loss, so an additional 10 dB in path loss translates to a 10 dB reduction in the SNR.

The path loss is a function of the carrier wavelength λ, transmitter-receiver distance d, and loss exponent n and is given by [55]

$$P_L = 20 \log_{10}\left(\frac{4\pi}{\lambda}\right) + 10n \log_{10}\left(\frac{d}{d_\circ}\right). \tag{2.2}$$

While the path loss exponent n depends on the particular propagation environment in effect, $n = 3$ is used as a typical choice for an indoor office environment. Typically, the loss at $d_\circ = 1$ m is measured and losses at a distance greater than 1 m are related through the second term of (2.2). Figure 2-4 plots the path loss as a function of the transmitter-receiver separation for various values of n assuming a carrier frequency of 5.25 GHz. The path loss for isotropic free space radiation is $n = 2$, while $n = 3$ corresponds to a Rayleigh fading environment with no line of sight signal. Higher values for n represent more severe path losses caused by obstructions such as walls and floors, although it is possible for the path loss to decrease even slower than in free space [38]. At a distance of 10 m, there is a a 10 dB difference in the path loss for each integer increment of n, so the choice of n can make a significant difference in the modeled performance of system. Measured indoor channel parameters for a house [40] and office [41] environment show values of n up to $n = 7$, with $n = 3$ being representative of most conditions. Some indoor channel measurements suggest that the constant term of (2.2) is an overly pessimistic model of actual conditions [40], but that the $n = 3$ exponent is still valid.

2.3 Capacity of a Wireless Link

After dividing the available bandwidth via OFDM, each frequency bin now has essentially frequency-independent fading. We can thus look at the capacity of a single bin by treating the fading as independent of frequency in each bin.

The capacity of a wireless link increases roughly linearly with the number of transmit and receive antennas [17]. However, this increase in capacity will only materialize with sufficient coding of the input data, which can greatly increase encoding and decoding complexity. For a single channel usage, the maximum achievable data rate is given by

$$C = \sum_{i=1}^{M} \log_2 (1 + \lambda_i \mathrm{SNR}), \tag{2.3}$$

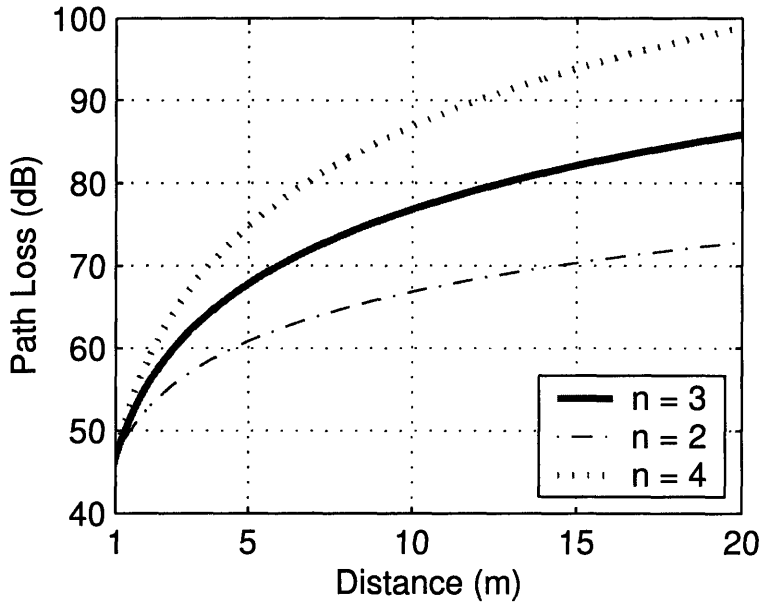


Figure 2-4: Path loss for $n = 2, 3, 4$ vs. distance.

where the λ_i's are the (possibly zero) eigenvalues of $\mathbf{H}^\dagger\mathbf{H}$ (with $\mathbf{H}^\dagger$ being the conjugate transpose of $\mathbf{H}$), and SNR is the received signal to noise ratio [66]. Alternately, the λ_i's are the squares of the singular values of $\mathbf{H}$. Since the channel is going to be used many times, we are interested in the maximum achievable throughput that we can expect for a given SNR. This is the ergodic capacity, and is only achievable with long codewords that span many different fades, or realizations of $\mathbf{H}$. This is counter to the block fading model that we are using, but serves as a useful metric to compare against. Alternately, for data blocks that only experience a single fade (as in the block fading model) we can look at the outage probability, or how likely the channel is not going to be able to support a given rate.

2.3.1 Ergodic Capacity

The ergodic capacity for an $M \times N$ system in b/s/Hz can be written as [66]

$$C = E\left[\sum_{i=1}^{M} \log_2\left(1 + \lambda_i \mathrm{SNR}\right)\right], \tag{2.4}$$

where the expectation is over all $\mathbf{H}$. Figure 2-5 shows the significant difference in the capacity curves for a 1×1 and a 4×4 system. When the SNR is sufficiently high,

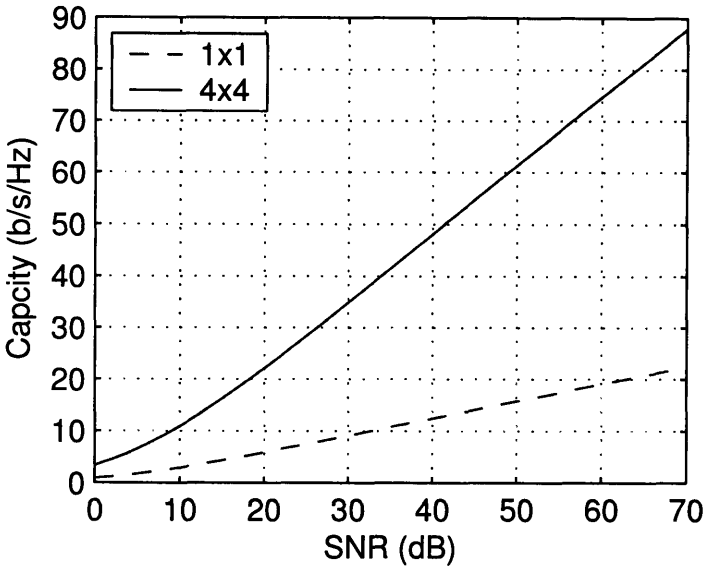


Figure 2-5: Capacity comparison between 1×1 and 4×4 wireless systems.

(2.4) can be simplified to

$$C \approx E\left[\sum_{i=1}^{k} \log_2\left(\lambda_i \mathrm{SNR}\right)\right], \tag{2.5}$$

where k is the number of nonzero eigenvalues. With high probability, $\mathbf{H}$ is full rank, so $k = \min(M, N)$ [17]. Since the SNR is a deterministic value

$$\begin{aligned} C &\approx \sum_{i=1}^{k} E\left[\log_2\left(\lambda_i\right)\right] + \sum_{i=1}^{k} E\left[\log_2\left(\mathrm{SNR}\right)\right] \\ &= \sum_{i=1}^{k} E\left[\log_2\left(\lambda_i\right)\right] + k\log_2\left(\mathrm{SNR}\right). \end{aligned} \tag{2.6}$$

If we define

$$k\log_2(\beta) = \sum_{i=1}^{k} E\left[\log_2\left(\lambda_i\right)\right], \tag{2.7}$$

then (2.6) can be simplified to

$$C \approx k\log_2(\beta \mathrm{SNR}). \tag{2.8}$$

The variable β is a constant that determines the offset of the capacity curve, and is dependent on the specific characteristics of the channel matrix $\mathbf{H}$. The parameter β can also be written as the kth root of the expected determinant of $\mathbf{H}^\dagger\mathbf{H}$. If $\mathbf{H}$ is a square matrix with Rayleigh-distributed entries, then the value of this expected determinant is $k!$ [15].

In the high SNR regime, the change in capacity ΔC is given by

$$\begin{aligned}
\Delta C &= C - C' \\
&\approx k \log_2\left(\frac{\beta \mathrm{SNR}}{\beta \mathrm{SNR}'}\right) \\
&= k[\log_2(\mathrm{SNR}) - \log_2(\mathrm{SNR}')] \\
&= \frac{k}{10\log_{10}(2)}(10\log_{10}(\mathrm{SNR}) - 10\log_{10}(\mathrm{SNR}')) \\
&= \frac{k}{10\log_{10}(2)}(\mathrm{SNR}_{\mathrm{dB}} - \mathrm{SNR}'_{\mathrm{dB}}) \\
&= \frac{k}{10\log_{10}(2)}\Delta\mathrm{SNR}_{\mathrm{dB}},
\end{aligned} \tag{2.9}$$

where $\Delta\mathrm{SNR}_{\mathrm{dB}}$ is the difference in SNR in dB. Simplifying results in

$$\Delta C \approx .33k\Delta\mathrm{SNR}_{\mathrm{dB}}. \tag{2.10}$$

This is the amount of capacity that must be sacrificed when reducing the signal to noise ratio from SNR to SNR′. As a result, every 3 dB reduction in SNR costs k b/s/Hz in capacity.

For example, referring to Figure 2-6, a 1×1 system at an SNR of 20 dB has a capacity of 6 b/s/Hz. For a 4×4 system at the same SNR, the capacity is 22 b/s/Hz. Using (2.10), this capacity difference of 16 b/s/Hz results in an SNR gain of 48.5 dB from the 1×1 system. Comparing this to the 1×1 curve in Figure 2-6, a capacity of 22 b/s/Hz requires approximately 69 dB SNR for a 1×1 system, a difference of 49 dB. However, physically achieving this SNR with an actual system would require a tremendous amount of transmit power for any reasonable transmitter-receiver separation. Thus the 4×4 system is able to reduce the required transmit power by more than a factor of sixty thousand compared to a 1×1 system for the same data rate.

The expression for capacity in (2.4) requires an expectation over all realizations of $\mathbf{H}$. In order to achieve the data throughput allowable according to the channel capacity, information bits need to be coded in blocks that are allowed to grow to very long lengths so that each block experiences many different fades. If a block only experiences a single fade, the instantaneous capacity may be below the ergodic capacity shown in Figure 2-5.

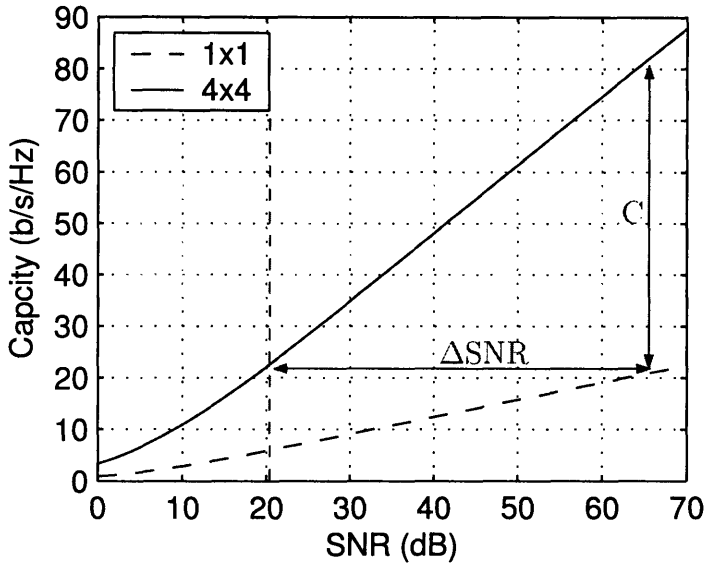


Figure 2-6: Capacity tradeoff with SNR.

2.3.2 Outage Probability

In a wireless network, there is a limit to how much delay can be tolerated, so the code blocks cannot reach arbitrarily long lengths, limiting how powerful the codes can be in terms of error correction capability. Since the transmitter might not know beforehand the channel conditions, a particular block may be transmitted during a fade that is deep enough that the channel is not able to support the data rate. This is an outage event, and the outage "capacity" in this sense is the highest rate that is achievable with a given probability for a given SNR. Figure 2-7 plots the probability of outage versus SNR for a 1×1 and a 4×4 system. The outage probability curves asymptotically become straight lines with a slope of $-MN$ [71]. As with the ergodic capacity plots, the 4×4 system shows a significant improvement in performance over the 1×1 system. At an outage probability of 1%, the 4×4 system is over 40 dB better than the 1×1 system for a total transmission rate of 6 b/s/Hz. Additionally, the curve for the 4×4 system transmitting at 24 b/s/Hz (6 b/s/Hz per transmit antenna) also shows an improvement of over 10 dB compared to the 1×1 system.

2.3.3 The Diversity-Multiplexing Tradeoff

The ergodic capacity is the maximum achievable data rate a system is able to support for a given SNR with arbitrarily low error rates. It assumes that the code words

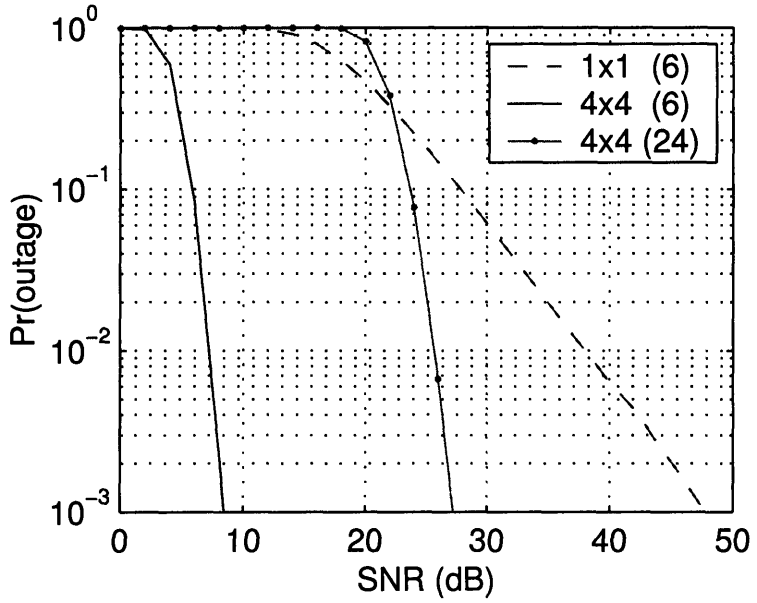


Figure 2-7: Outage probability for transmitting at rates of 6 b/s/Hz for 1×1 and 4×4 systems, and at 24 b/s/Hz (6 b/s/Hz per transmit antenna) for a 4×4 system.

are able to grow infinitely long to experience many fades, resulting in the channel's average behavior. On the other hand, the outage probability shows how quickly the error probability decays with SNR for a fixed data rate. There is a trade-off between these two fundamentally different ways to approach the limits of the wireless channel. This tradeoff between the multiplexing and diversity is described in [82].

The multiplexing gain describes how the capacity increases with SNR, and is defined as

$$r = \lim_{\text{SNR}\to\infty} \frac{R(\text{SNR})}{\log(\text{SNR})}, \tag{2.11}$$

where $R(\text{SNR})$ is the achievable data rate as a function of SNR. Comparing this to (2.8), we see that $r = \min(M, N)$. It should also be noted that a fixed data rate scheme (such as 64-QAM) results in $r = 0$. Intuitively, as the SNR increases in the case where $r = 0$, all the extra capacity gained is used to provide maximum redundancy for robustness to errors, so the diversity gain will be as large as possible.

The diversity gain describes the slope of the error probability curve (on a log scale), and is defined as

$$d = -\lim_{\text{SNR}\to\infty} \frac{\log(P_e(\text{SNR}))}{\log(\text{SNR})}, \tag{2.12}$$

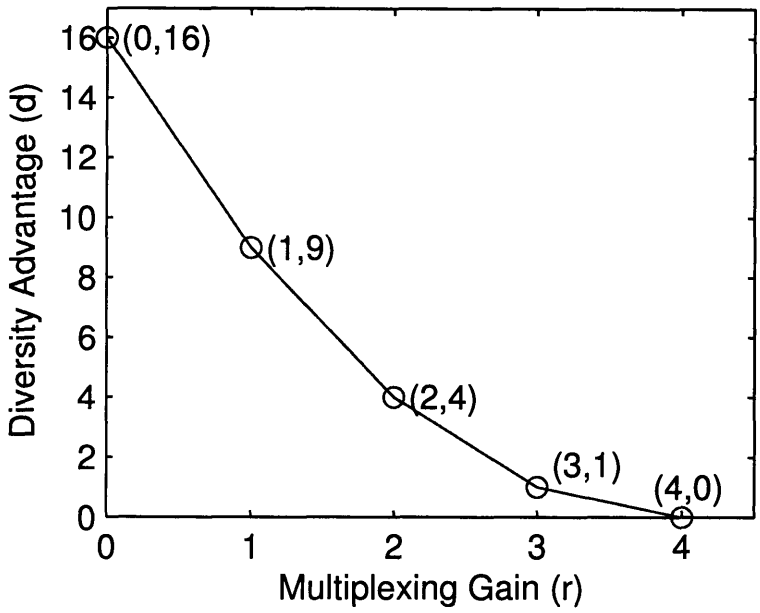


Figure 2-8: Diversity-multiplexing tradeoff for a 4×4 system.

where $P_e(\text{SNR})$ is the probability of error as a function of SNR. Since d is the slope of the error probability, we see that the probability of error asymptotically can fall as fast as SNR^{-d}. The case of $d = 0$ means that as the SNR increases, the probability of error does not grow smaller. Intuitively, when $d = 0$, all of the extra capacity from the increasing SNR is used to transmit at higher rates, so there is no rate left to use for redundancy to increase robustness to errors.

These two endpoints, when $r = 0$ and $d = 0$ comprise the extreme values of the achievable (r, d) pairs. There are a whole range of achievable pairs in between these two extremes, however. Figure 2-8 plots the diversity-multiplexing tradeoff for a 4×4 system, showing that the maximum diversity gain is 16 and the maximum multiplexing gain is 4. In general, (r, d) pairs of the form $(r, (M - r)(N - r))$ can be achieved [82]. The points on the lines connecting these discrete points can be achieved by timesharing between the endpoints.

For the purposes of this thesis, however, we will be focusing on the maximum diversity and maximum multiplexing endpoints. It should be noted that two schemes that achieve the same diversity gain d may still have a constant SNR offset due to the coding gain. Similarly, a coding scheme that achieves the maximum multiplexing gain $r = \min(M, N)$ may not necessarily achieve the ergodic capacity given by (2.4). Nevertheless, a simple scheme that achieves the maximum available diversity or multiplexing is desirable for its efficient use of resources.

2.3.4 Uncoded Bit Error Rate Performance

Any data rate smaller than the capacity given by (2.4) can be transmitted with arbitrarily low probability of error but requires large block size and computationally intensive encoding and decoding. If computation or memory resources are scarce, then encoding and decoding the data streams becomes difficult. An additional and more fundamental problem is that of delay. In order to encode with very long block lengths, the transmitter must first accumulate enough data. Similarly, the receiver must accumulate a full block's worth of symbols before any bits can be decoded. For a low-bandwidth node, a large block size can require unacceptable amounts of delay. For example, a cordless phone or an intercom requires low latency for two-way communications. Since these are low-bandwidth devices, they should only be assigned one frequency bin. With a 1 MHz channel, there can be several megabits of available data rate. However, since the information rate is only a few kilobits per second, it can take a significant fraction of a second to accumulate enough bits to fill up an information block for transmission and again for reception.

At the other end of the spectrum, uncoded transmissions have the minimum delay, since information can be transmitted as soon as data is ready and decoded immediately upon reception. The cost of uncoded transmissions is lack of any redundancy of error protection, though at high enough SNR this redundancy is unnecessary. Since the capacity of multiple antenna systems can be so high, it is reasonable to trade off some of the excess to reduce the amount of coding required. As lower bounds to performance, we can also examine the performance of uncoded (or minimally coded) data streams and compare to that of optimally coded systems.

In an uncoded system, the focus is not on capacity but diversity, or how quickly the receiver BER falls with respect to SNR. In general, the number of independent wireless channels that a bit experiences is the order of diversity it can see. Because the bit experiences multiple channels, it effectively sees an averaged channel, which has a lower variance than the individual channels. An alternate viewpoint is that it is less likely for all of the channels to be bad (deeply faded) than for a single channel. Although there are many ways to achieve high diversity, only spatial diversity will be considered in this thesis. Time diversity increases delay by requiring interleaving of data across many packets due to the block fading model, while frequency diversity uses up valuable bandwidth by coding across subcarriers. Spatial diversity does require additional hardware, but Chapter 4 will show how this cost can be reduced.

For a single transmit antenna and a single receive antenna (a 1×1 system), the uncoded bit error performance $P_e \propto 1/\text{SNR}$, where SNR is the ratio of total transmitted power P to the noise variance N_o [71]. By increasing the number of receive antennas, not only is there a gain in SNR from the increased received power, but more importantly, the slope of the bit error rate (BER) curve increases due to the

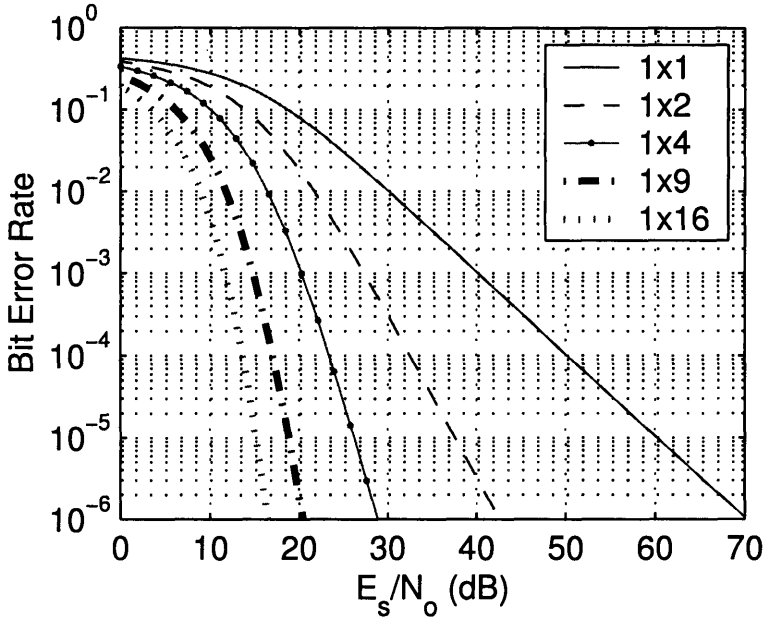


Figure 2-9: Uncoded 64-QAM BER for $1 \times N$ systems, $N = 1, 2, 4, 9, 16$ (right to left).

increased diversity. Figure 2-9 plots the BER curves (adapted from a closed-form formula in [51]) for uncoded 64-QAM constellation points transmitted with average power E_s and received by N antennas. For each of these curves, the diversity gain is $d = N$. Increasing the number of transmit antennas similarly increases the slope of the BER curve, but the transmit power per antenna is lowered to keep the total transmit power constant for a fair comparison.

For an $M \times N$ system in a Rayleigh-fading environment, the maximum diversity achievable is MN, which means the BER is proportional to SNR^{-MN} [71]. Maximum diversity reduces the achievable data rate to that of a 1×1 system, but it requires very little added complexity at the transmitter and receiver. The advantage of the maximum diversity case is that even without coding it is possible to significantly reduce the required SNR. However, the tradeoff between diversity and multiplexing gain allows most of the gains from diversity without sacrificing as much capacity [82]. As we can see from Figure 2-9, the difference in SNR is greatest between $d = 1$ and $d = 2$. Further increasing the diversity order does result in larger gains in SNR, but it is a case of diminishing returns. For example, the SNR gain between $d = 4$ and $d = 16$ is about the same as that between $d = 2$ and $d = 4$. However, going from $d = 4$ to $d = 16$ requires adding at least five more antennas (1×2 to 4×4), while going from $d = 2$ to $d = 4$ only requires one more antenna (1×2 to 2×2).

To achieve maximum diversity, the data x to be transmitted must be encoded as the vector $\mathbf{x}$ to be output by the transmit antennas. Using the singular value decomposition (SVD), the channel matrix can be decomposed as $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\dagger$, resulting in

$$\mathbf{y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\dagger\mathbf{x} + \mathbf{n}, \tag{2.13}$$

where $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices and $\mathbf{\Sigma}$ is a diagonal matrix of the singular values. Multiplying both sides by $\mathbf{U}^{-1} = \mathbf{U}^\dagger$ then gives

$$\begin{aligned} \mathbf{U}^\dagger\mathbf{y} &= \mathbf{U}^\dagger\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\dagger\mathbf{x} + \mathbf{U}^\dagger\mathbf{n} \\ &\downarrow \\ \mathbf{y}' &= \mathbf{\Sigma}\mathbf{x}' + \mathbf{n}', \end{aligned} \tag{2.14}$$

with $\mathbf{x}' = \mathbf{V}^\dagger\mathbf{x}$, $\mathbf{y}' = \mathbf{U}^\dagger\mathbf{y}$, and $\mathbf{n}' = \mathbf{U}^\dagger\mathbf{n}$. Note that $\mathbf{n}'$ has the same statistics as $\mathbf{n}$ since $\mathbf{U}$ is unitary. Since $\mathbf{\Sigma}$ is diagonal, the SVD has decomposed the $M \times N$ system into $\min(M, N)$ parallel channels with gains corresponding to the nonzero singular values. To achieve maximum diversity, each symbol $\mathbf{x}' = c[11\cdots1]^{\mathrm{T}}x$. In order to make fair comparisons with the single antenna case, the constant c normalizes the total power of $\mathbf{x}'$ to be the same as x. This is equivalent to repetition coding, which has no coding advantage over the uncoded data and requires minimal computation. The scalar input x is chosen from square M-QAM constellations that have been normalized to have the same average power. For this example, the input constellation is restricted to 64-QAM. This represents data rates near a gigabit per second for the WiGLAN, although typically the transmitter would adjust the constellations based on channel conditions. At the receiver, decoding is performed via maximal ratio combining (MRC), with $r = [\sigma_1 \cdots \sigma_N]\mathbf{y}'$, where the σ_i's are the singular values of $\mathbf{H}$.

Figure 2-10 shows the simulated BER vs. SNR for a 1×1 and a 4×4 system using 64-QAM. The SNR is plotted as E_s/N_o, where E_s is the average energy of the transmitted constellation points, and N_o is the variance of the complex Gaussian noise. Note that the slope of the BER curve for the 4×4 system is much steeper than for the 1×1 system, which is a result of the higher diversity gain. Thus, the SNR gain for a 4×4 system over a 1×1 system depends on the desired error rate, and increases without bound for very low error rates. Comparing the BER curve for the 4×4 system in Figure 2-10 to the $N = 16$ curve in Figure 2-9, we see a difference of 6 dB. This is because the 1×16 system has 16 receive antennas, which gathers four times, or 6 dB, as much receive power as a 4×4 system.

Assuming a packet size of around 1000 bits, an error rate of 10^{-5} corresponds to approximately one percent packet errors. For the target BER of 10^{-5}, the difference in SNR between the 1×1 and 4×4 systems is 40 dB. Uncoded systems are not typically operated at such low error rates due to the excessive SNR requirements.

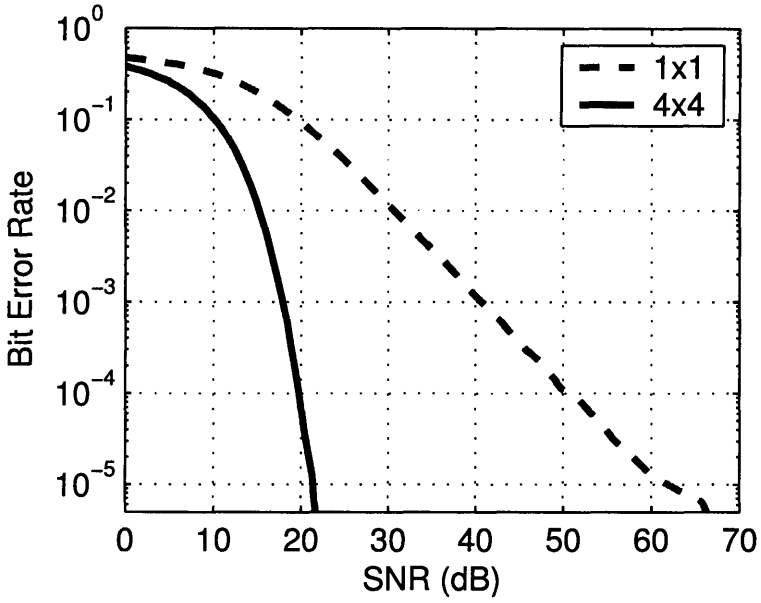


Figure 2-10: BER for a 1×1 vs. 4×4 system for a 64-QAM input constellation.

However, the high diversity results in SNR gains large enough to allow the possibility of using an uncoded system, with significant savings in computational complexity and memory requirements. The diversity gain possible for a 4×4 system is achievable if the wireless environment is rich enough to support so many degrees of freedom. Even if the available degrees of freedom is minimal, however, much of the SNR gain can still be achieved by lower diversity system. For example, a 2×2 system would still have a gain of 30 dB over a 1×1 system under the same conditions.

The above procedure assumes that the transmitter knows the channel characteristics, which is true for the WiGLAN and is reasonable for certain kinds of wireless systems. However, it is not always pragmatic to expend the resources to communicate the channel knowledge back to the transmitter, or that information may not even be useful. For example, if a transmitter is going to broadcast to more than one user, even if it knows the channel matrix for each receiver, it will not be able to use the SVD to diagonalize both links simultaneously. In this case, it is still possible to achieve maximum diversity without any significant coding. The transmitter, instead of repetition coding across the parallel channels, can transmit the desired signal on one antenna at a time, cycling through all the transmit antennas in M time steps. At each receiver, the N received signals at each of the M time steps can be summed using the same MRC procedure as before. This will allow all of the receive nodes to achieve the same performance as in Figure 2-10, but the at a cost of reducing the

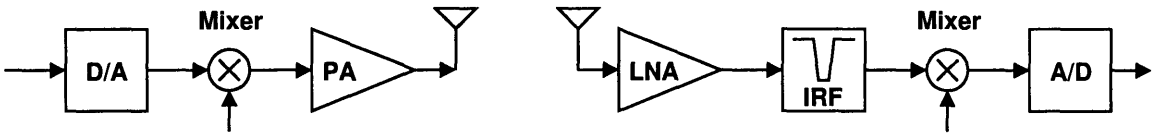


Figure 2-11: A typical analog transmit and receive front end.

data rate by $1/M$.

The diversity gain of multiple antenna systems can be viewed as a low complexity mechanism for SNR gain that would otherwise require computationally intensive coding. This allows different circuit architectures to be used for minimizing power or area, or both. The low complexity of the diversity system also reduces the number of digital circuits required to process the signals.

2.4 Analog Circuit Considerations

In order to decide how the gains from multiple antennas will help ease the circuit requirements of the analog hardware, we first need to understand some of the sources of nonideal behavior in analog circuitry. Referring back to the high-level system view of a wireless system in Figure 1-2, we focus on the circuits in the shaded areas while ignoring the details of the antennas.

As shown in Figure 2-11, the data stream is first converted to analog through the digital to analog converter (D/A), mixed up to the carrier frequency, and amplified by the power amplifier (PA). At the receiver, the received signal is first amplified by a low noise amplifier (LNA), then mixed down to baseband and sampled by an analog to digital converter (A/D). Depending on the particular system, the incoming signal may pass through an image reject filter (IRF) before entering the mixer. Of these components, we focus on the design of the amplifiers at both the transmitter and receiver.

2.4.1 The Transmitter Power Amplifier

For a power amplifier (PA), a major concern is efficiency, which can be defined as the ratio of the amplifier transmit power output to its total power consumption. For a typical wireless system, the transmitter power amplifier consumes a significant fraction of the total power used. The efficiency of the amplifier then becomes an important consideration, especially for battery-powered nodes. For example, if a wireless system transmit at a power of 10 mW, then a 10% efficient amplifier would draw 100 mW of power, while a 50% efficient amplifier would draw only 20 mW. If

the PA was the dominant power drain in the system, then the difference in efficiencies would lead to a factor of five difference in battery life. A major impediment to the efficiency of the PA comes from the fact that linearity and efficiency are opposing goals in amplifier design. In this thesis, we will restrict our attention to linear amplifiers in order to meet the high linearity demands of the WiGLAN.

A class A amplifier is highly linear, but it has a very low efficiency. Even though the theoretical peak efficiency is 50%, it is much less efficient when not at peak power [36]. Additionally, since current is always flowing regardless of the output power, the class A amplifier draws a constant amount of power regardless of the output. For comparison, a class B amplifier is a nearly linear amplifier that prevents this wasted current from flowing, increasing the peak efficiency to nearly 80%. The cost of this increased efficiency is a loss in linearity, however.

One method to increase amplifier efficiency is to adaptively bias a class A amplifier so that the maximum output power of the amplifier corresponds to the instantaneous power of the output signal [48]. With this method, the amplifier is always running close to its maximum power, while still keeping the linear characteristic of the class A amplifier. The left plot of Figure 2-12 shows a sinusoidal waveform output at three different bias levels for the same input level. When the maximum power of the desired output signal is less than the maximum output power of the amplifier (the two dashed curves), the bias currents can be reduced to lower the amplifier gain accordingly. In this way, the maximum of the output signal will always correspond to the maximum available output power of the amplifier, increasing its efficiency, although not as much as that of a class B amplifier. Its efficiency curve is similar to that of a class B, but with a maximum of 50%, the same as for a standard class A amplifier. We will refer to this type of amplifier as an adaptive class A, or adaptive A.

There are many classes of more highly efficient amplifiers with theoretical efficiencies approaching 100%, but they are highly nonlinear. The extreme case is a switching amplifier, which achieves a theoretical efficiency of 100% by acting like a switch, with either voltage but zero current (open circuit), or current with zero voltage (short circuit). The output of these amplifiers are square waves, thus severely distorting the output waveform. One method to linearize a highly efficient amplifier is called outphasing, which uses vector addition to combine the output of two constant amplitude amplifiers to create a single linear amplifier [8]. The right diagram in Figure 2-12 shows how a desired amplitude and phase output can be achieved by combining the outputs of two constant amplitude amplifiers. The two vectors have amplitude A, and their relative phase ϕ are varied such that the vector sum can be anywhere from 0 to $2A$. The outphasing amplifier has a maximum theoretical efficiency of 100%, but its efficiency curves drops off rapidly if the output level is not near the maximum. Outphasing amplifiers are described in more detail in Appendix C.

Figure 2-13 shows how the efficiency of the linear class A, adaptive class A, and

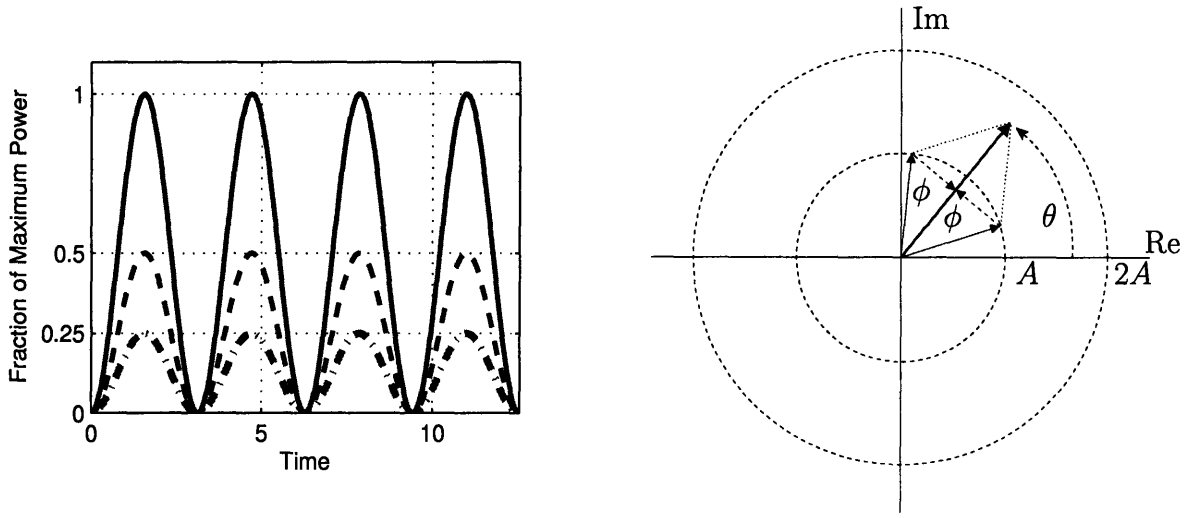


Figure 2-12: Two methods for increasing the efficiency of a linear amplifier. Varying the bias currents changes the gain of the adaptive A amplifier (left), while an outphasing amplifier sums the outputs of two highly efficient, constant amplitude amplifiers (right).

outphasing amplifiers vary with the output voltage (power is proportional to voltage squared). The efficiency curve for the nonlinear class B amplifier is also shown for comparison. Note that the class A has the worst efficiency, while the nonlinear class B amplifier is more efficient than the linear amplifiers over most of its range. The outphasing amplifier has the highest peak efficiency, but its efficiency falls off more rapidly with decreasing output level than the other amplifiers being considered.

In addition to linearity, dynamic range, or how much the envelope of the output waveform varies, is an important consideration. An ideal amplifier would have an output $y = Ax$ where A is the gain of the amplifier. However, since no amplifier is ideal, the output in general can be represented as a power series

$$y = \sum_{i=0}^{\infty} A_i x^i. \tag{2.15}$$

The closer an amplifier is to an ideal amplifier, the smaller the coefficients A_i are for $i \geq 2$. When the output voltage gets close to the power rails, the output will saturate. A model for the output of an amplifier (ignoring gain) due to Rapp is [54]

$$V_o = \frac{V_i}{\left[1 + \left(\frac{|V_i|}{V_{\text{sat}}}\right)^{2P}\right]^{1/2P}}, \tag{2.16}$$

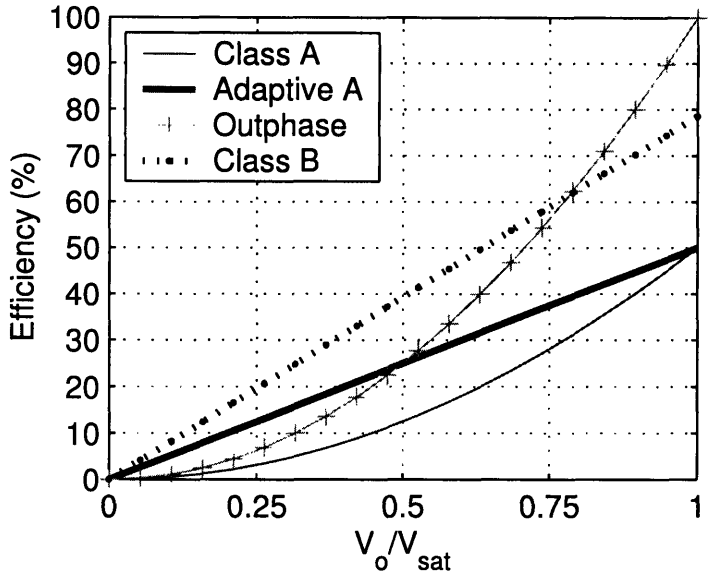


Figure 2-13: Efficiency of several linear (Class A, Adaptive A, and Outphase) amplifiers vs. output voltage swing. The nonlinear class B amplifier is shown for comparison.

where V_i and V_o are the input and output voltage, respectively. The output saturation voltage V_{sat} represents the highest value that the output voltage will reach. The parameter P can be varied to fit measured amplifier characteristics. The larger the value for P, the more linear the amplifier is. Figure 2-14 shows the Rapp model's output characteristic for several values of P. The curve for $P = 100$ is very close to an ideal amplifier, with an essentially linear characteristic and a hard saturation. The Rapp model only accounts for the amplitude distortion caused by the nonlinearities in the amplifier, but it does not model any phase distortion of the amplifier. Although we do not model the phase distortion of the amplifier, this distortion can adversely impact the error rate for higher order constellations such as 64-QAM.

2.4.2 The Receiver Low Noise Amplifier (LNA)

For the LNA, efficiency is not a great concern because it does not dissipate a significant fraction of the total receiver power. As the name implies, the most important feature of the LNA is low noise. The Noise Figure (NF) is a figure of merit for LNAs which encapsulates the total noise contribution of the LNA to the signal output. If an LNA has a NF of 2 dB, then the output SNR will be 2 dB less than the input SNR. Although each analog component has a noise figure associated with it, the first element in the receive chain is usually most important because it has the largest effect

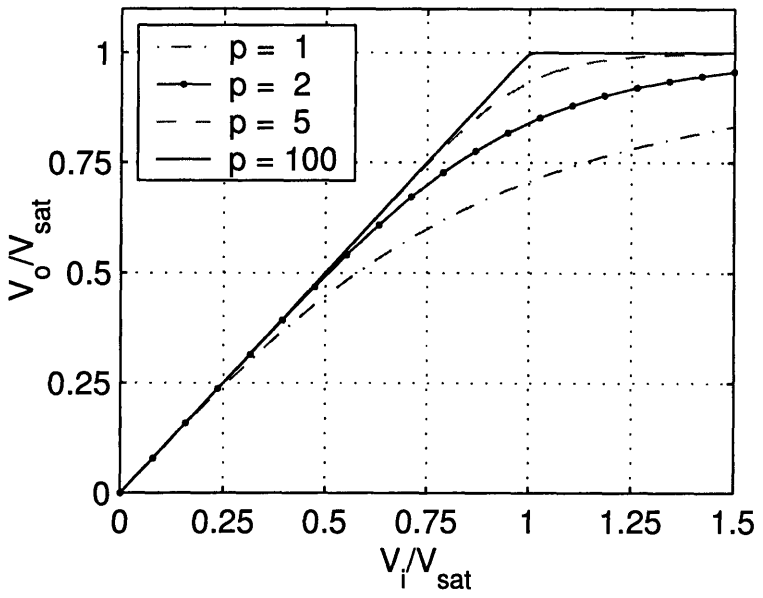


Figure 2-14: Rapp model for amplifier saturation (assuming unity gain).

on the overall noise figure. The noise can come from thermal variations and amplifier nonlinearities. Since efficiency is not much of a consideration for a LNA, there is no need to operate the amplifier close to saturation. This keeps the amplifier in the more linear region of its operating characteristic.

The LNA is also concerned about dynamic range, because the receiver is likely to experience a wide range of input powers depending on the transmitter-receiver separation. If the input is too small, then the receiver SNR will be too low for accurate decoding of the received signal. if the input is too large, due to very close range transmission for example, the receiver might saturate, which again would prevent accurate decoding. This has the effect of placing a bound on both how close and how far the transmitter-receiver separation can be.

Using a sinusoidal input, we can see what the effect of the LNA nonlinearities are in terms of the harmonics. Figure 2-15 shows the frequency spectrum of a 10 MHz sinusoid after passing through an amplifier with output characteristic given by (2.16). The maximum value of the input sinusoid is scaled so that the unmodified output would reach the full range of output voltages. The fundamental frequency is apparent, as well as the odd harmonics. The third order harmonic at 30 MHz is the dominant nonlinearity, with the fifth order also visible. In general, the third order harmonic is going to be the dominant nonlinearity. The relative power of the nonlinear harmonics depends on the shape of the amplifier transfer characteristic, but as Figure 2-15 shows,

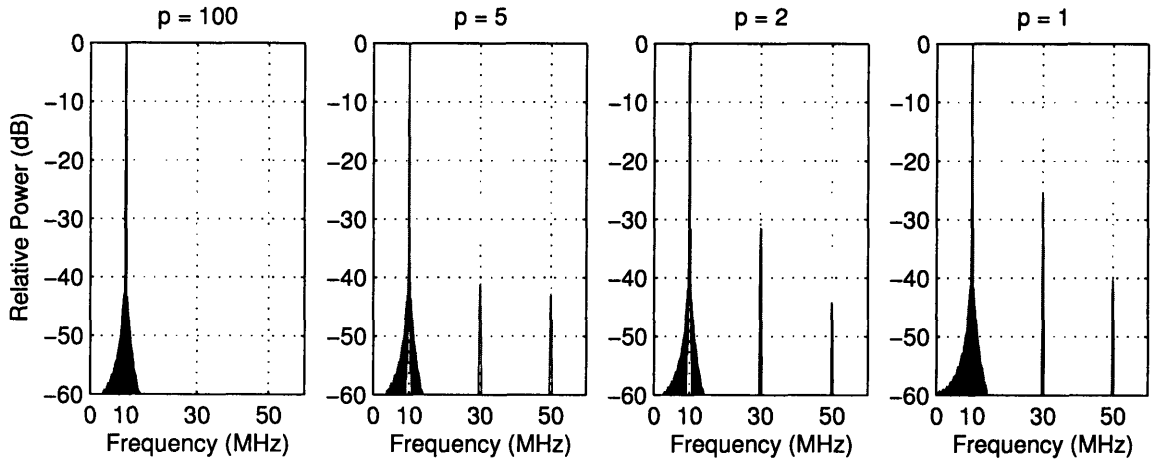


Figure 2-15: Sinusoidal harmonics from the Rapp model.

the strength of the harmonics decrease as the amplifier characteristic becomes more linear (as $p \rightarrow \infty$ in the Rapp model).

For the LNA, these signal harmonics or "spurs" caused by nonlinearities can have a higher power than the thermal noise floor, and thus will affect the noise figure. The maximum input value that keeps these output harmonics below the thermal noise floor is used to mark the Spurious Free Dynamic Range (SFDR), which is useful as a dynamic range measurement for the LNA. The SFDR is the difference between this maximum power and the noise floor.

2.4.3 Digital Circuit Considerations

For digital circuits, the complexity of the processing that is required for encoding and decoding the data stream will directly relate to the area and power consumption of the digital circuitry. Encoding techniques which approach the capacity given by (2.4) require very long block lengths, and high complexity at the decoder and possibly the encoder as well. The long block lengths require large blocks of memory circuits to hold and manipulate the data, and the complex calculations require a more capable and thus larger computational unit.

Although for analog circuits, the size and topology of circuits depend strongly on the actual specifications that must be met, digital circuits are able to follow some scaling laws. For example, the area required for memory grows approximately linearly with the amount of memory required, therefore doubling the block length of a coding scheme requires twice as much chip area for the extra memory cells. The power

consumption per transistor can be written as [7]

$$P = p_t \cdot C_L \cdot V_{dd}^2 \cdot f_{clk}, \tag{2.17}$$

where C_L is the loading capacitance (usually the gate capacitance of the following transistor(s) plus any interconnects), V_{dd} is the supply voltage, and f_{clk} is the system clock. Thus more transistors will increase the power consumption, as expected. The activity factor p_t denotes what fraction of the time the transistor is actually switching states, which will be lower if the transistor circuits are idle, for example. The amount of time required to switch a logic gate is

$$T \propto 1/V_{dd}. \tag{2.18}$$

As a result, increasing the supply voltage will lower the switching time, reducing the delay of a logic gate. In order to use more complex processing, the system must either allow more time for the extra computations, or increase the frequency of the system clock to allow more computations to be done in the same amount of time. However increasing the clock frequency requires the logic delays to decrease, which requires the supply voltage to be increased. However, from (2.17), it is clear that the power dissipated will increase significantly as the operating frequency as well as the supply voltage is increased. These scaling laws indicate that low-complexity algorithms can make a significant difference in both the area and power consumption of the digital circuits.

Chapter 3

The WiGLAN Architecture

In the WiGLAN, nodes communicate with each other using the central server to help schedule and allocate resources for node-to-node communications. The nodes can either communicate through the central server similar to a cellular network, or (more rarely) directly with each other as in an ad hoc network. In the latter case, the central server only directs the communication between nodes, which does not fully use the considerable resources that are available on the server.

In this chapter, we propose a communications protocol for a wireless network such as the WiGLAN, which takes advantage of the powerful central server to offload a significant amount of computation and memory resource requirements away from the mobile nodes and onto the central server, allowing the nodes to be inexpensive and less capable. All network communication is made through the central server, with node transmissions composed of an uplink from the transmitter to the central server, followed by a downlink from the server to the receiver. The nodes in the network are able to form virtual antenna arrays to increase the data throughput through multiplexing gain without requiring the nodes to communicate among themselves to coordinate their transmissions.

The uplink/downlink communication protocol is also able to provide high diversity gain to the nodes, decreasing the transmit power required for the mobile nodes, and increasing the effective range such that direct node-to-node transmissions are only rarely favorable from a power standpoint. The SNR gain from using multiple antennas can also allow little or no error correcting coding to be used, decreasing complexity and memory requirements to allow very simple nodes. Since each packet makes only a single relay hop by passing through the central server, packet delay times are bounded and small, facilitating time-sensitive communications.

Section 3.1 describes the typical network topology, with single and multiple antenna mobile nodes and a central server. The uplink/downlink communication protocol is described in Section 3.2, with emphasis on the multiplexing gain that can be

achieved with virtual node arrays. Section 3.3 explores the SNR gain possible from instead focusing on diversity rather than multiplexing gain. The SNR gain allows little or no coding to be used for reliable network transmissions. Section 3.4 describes how the SNR gain can be used to increase the transmission range, allowing nodes to transmit with less power through the server that directly with each other. The effect of the time-varying channel conditions on throughput is explored in Section 3.5. When applied to the WiGLAN in Section 3.6, we find that the indoor channel requires negligible overhead due to the coherence time, while the delay spread of the channel requires a similar amount of overhead as in an 802.11a system. This section also determines the minimum number of antennas required for a given throughput for a variety of network conditions.

3.1 Network Topology

A network like the WiGLAN is controlled by a central server with few constraints on transmit power or computational and memory resources. This server controls and assists the node-to-node traffic among the multiple mobile nodes in the network. The central server is equipped with several antennas (for example, four), which can be used in any combination to transmit and receive information from the other nodes in the system. The nodes have limited battery, memory, and computational power, and are equipped with one or more antennas for transmission and reception.

A typical setup for the proposed wireless network would be in a home or single-level office, and schematically might look like Figure 3-1. The central server controls the communications between all the other nodes in the network, as well as participates in the relaying of information between nodes. It is able to communicate directly with any node in the network without needing to relay information through any other node that is part of the network. The purpose of the central server is to allocate data bandwidth to nodes which request a certain data rate, as well as to serve as a conduit for data transfer between nodes. It may also communicate timing information and arbitrate communications across the network. Nodes may enter and leave the network randomly as well as move about within the network area. The server can also be connected to other servers in other networks via an independent (possibly wired) link, much like base stations in a cellular network, as well as to the Internet or other external networks.

The central server has several antennas that are independently capable of both transmission and reception, and is modeled as having unlimited transmit and computational power, as well as a large amount of memory to store packets for delayed routing or other uses. The mobile nodes are also capable of transmitting and receiving wireless data, and may or may not have multiple antennas.

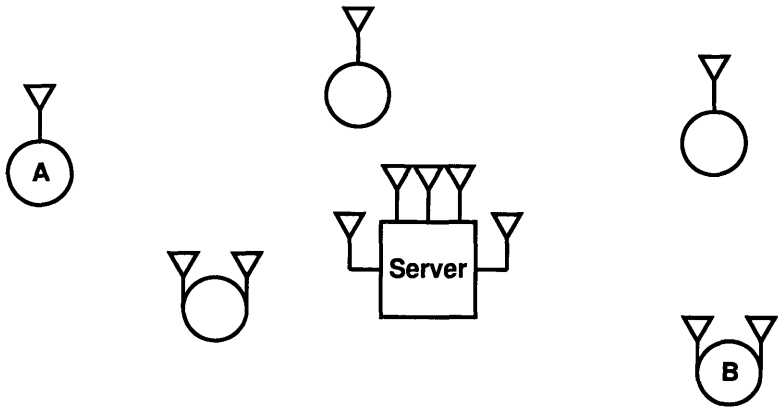


Figure 3-1: A schematic of a typical network, including both the high-bandwidth, multi-antenna nodes (B), as well as the less capable, power- and memory-limited nodes (A).

Some nodes in the network model things that only need to communicate wirelessly, but are otherwise able to be plugged into an electrical outlet. These nodes move only rarely, and are generally large in size. Examples of these nodes would be a TV and DVD player which are able to communicate wirelessly, a stereo and its speaker system, or a computer and the associated monitor. These nodes have modest power constraints, but have limited computational and memory resources. Since these are larger devices, they are more likely to have multiple transmit and receive antennas. Other nodes model devices such as cordless phones, walkie-talkies, or PDAs, which are battery powered, and thus have limited transmit power as well as few computational and memory resources. In addition, their small size dictate that only have a single antenna is available. These devices also may not be able to transmit and receiver across all frequencies of the system bandwidth due to less capable hardware.

3.2 The Communication Protocol

The communication protocol between the transmitter (TX) node and the receiver (RX) node is first described assuming a single frequency bin. If arbitrary node-to-node communications are allowed (ad hoc mode), the capacity of the network per node falls to zero as the size of the network increases [23]. As the number of nodes increase, each node spends a greater proportion of its time relaying messages to and from other nodes in the network. This ad hoc mode does not adequately utilize the special nature of the central server, causing the overall network throughput to suffer.

The central server, however, is a special node because it has a large number of antennas which are able to cooperate without using any wireless resources. In

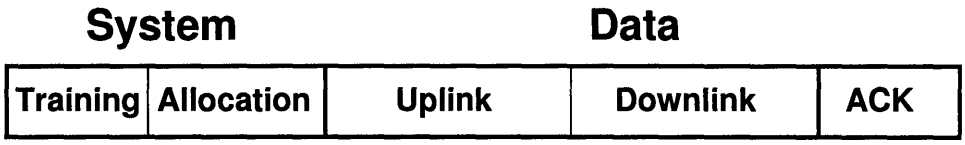


Figure 3-2: The five phases of a communications frame.

addition, it is able to communicate with all the other nodes in the network without assistance, and has large amounts of memory and computational power. This suggest that the server should be involved in the majority of node communications leading to hub and spoke model for communications with two phases. First, a transmitting node uplinks to the server, and then the server downlinks to the corresponding receiving node, much like the communications path for a cellular phone network. Additionally, there need to be two system phases, similar to the uplink and downlink, that are used for training and resource allocation. Finally, there is a set of two phases which allow the receiver to send a packet acknowledgment (ACK) to the transmitting node. These last two phases are exactly the same as the uplink and downlink data phases, so we will lump them together.

We can therefore define a frame as in Figure 3-2, with five phases. The first two phases are for learning the channel and resource allocation. Nodes that will transmit in this frame send training data to the server over many frequency bins. The server then allocates frequency bins and transmits the bin assignments plus channel information to the nodes. The next two phases are the uplink and downlink phases, and a packet node-to-node transmission requires both phases. A packet is uplinked to the central server from the TX node and then downlinked to the RX node. There can be multiple packets sent for each node in the data phases, depending on how often the channel needs to be relearned. Even more packets can be sent if the packets themselves can be used to update the channel estimates. Either the uplink or downlink phase can be skipped if the transmission is not peer-to-peer, but only between the node and the server, such as when communicating with an external network. Finally, there is the ACK phase, in which the receiver either acknowledges the receipt of the transmitted packets or requests retransmissions.

3.2.1 The Server as a Base Station

The antennas on the server are all able to coordinate and share information without using up any wireless bandwidth, whereas an equivalent number of individual nodes are only able to communicate with each other by using bandwidth that could otherwise be allocated for data transmissions. A collection of nodes is not able to achieve

diversity or multiplexing gain without some coordination between the transmitting or receiving nodes. Because the server antennas can coordinate, however, node transmissions that involve the central server are able to achieve these gains. As a result, the SNR gain from diversity can allow the transmit power to be greatly reduced, while the multiplexing gain effectively increases the total available bandwidth.

In order to take advantage of the server's multiple antennas we use the server as a base station, with a node-to-node transfer composed of an uplink from the TX node to the server, followed by a downlink from the server to the RX node. If we assume that the server has as many antennas as there are communicating nodes (and that each node has a single antenna), then it is possible to transmit and receive to all the nodes at once by using the maximum multiplexing gain. This is very similar to the virtual antenna arrays algorithm in [24], with the central server taking on the role of one of the virtual arrays, and the nodes forming the other array. While in [24] the nodes in each virtual array must use bandwidth to communicate with each other, the server antennas can freely communicate. The following uplink and downlink protocols also do not require the nodes to communicate with each other while still achieving the multiplexing gain. Thus the server enables the formation of virtual arrays of nodes for multiplexing gain, but without the bandwidth overhead for node coordination.

3.2.2 The Uplink Protocol

The N nodes in the network can be viewed as a single node with N transmit antennas, and the receiving node (the server) has an array of M receive antennas, forming a virtual $N \times M$ system. In this setup, the transmitting antennas are not able to coordinate their actions, so the uplink protocol needs to be able to accommodate N independent bit streams at the transmitter. The VBLAST architecture allows for exactly this setup, and can achieve a significant fraction of capacity (and the maximum multiplexing gain) even with uncoded bit streams, as long as the number of transmit antennas is not more than the number of receive antennas [78]. The encoding at each node is then quite simple: each node transmits its own uncoded bit stream without regard to how many other nodes are also transmitting. To maximize the throughput, as many nodes in the network can transmit as there are server antennas in this phase.

At the receiver, the nodes are detected sequentially, starting with the node that experiences the strongest channel in terms of its effective SNR. This is based on the amount of fading and the interference from other transmitting nodes which that node experiences. Once a node is decoded, its output is then subtracted from the received signal before the next node is decoded. This subtraction (assuming the decoding of that node was correct) removes the interference caused by that node on all the remaining nodes. When the last node is reached, it has no other interference except

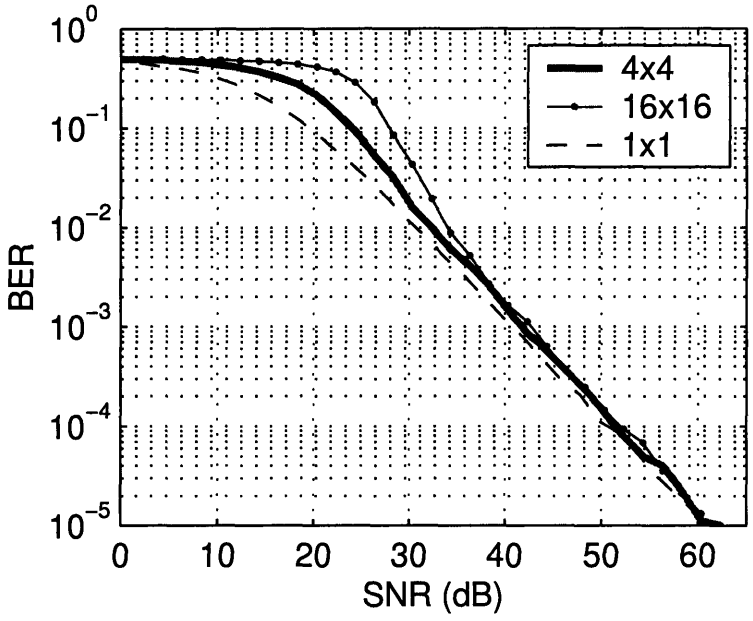


Figure 3-3: Uplink BER performance for 4 (dashed line) and 16 (solid line with dots) nodes for 64-QAM. The performance closely matches a 1×1 system at high SNR.

for the noise from the channel. The average bit error rate of this scheme is similar to the single transmit, single receive antenna case, with $P_e \propto 1/\text{SNR}$. Thus the uplink protocol has the effect of forming N parallel systems which do not interfere with each other.

This can be seen in Figure 3-3, which shows the average BER for 4 and 16 transmitting nodes (and an equal number of server receive antennas). The BER curves for the 4×4 and 16×16 system are the same as for a 1×1 system at high SNR, but with four and sixteen times as much throughput, respectively. Since the VBLAST algorithm assumes correct decisions are always made when subtracting interference, when the SNR is low, the probability of incorrect decisions rise, causing the BER curves to diverge from the 1×1 case. The decoding of the nodes is done on a symbol by symbol basis, allowing the nodes to transfer uncoded data, as simulated in Figure 3-3. The TX and RX nodes can agree to add error correcting codes to decrease the required SNR, but this is neither required for transmission nor does the server require any knowledge of this coding.

We can now describe the algorithm in more detail, which follows directly from the VBLAST algorithm [78]. The algorithm for encoding and decoding the uplink works on a symbol by symbol basis. First, each TX node transmits its symbol without consideration to what the other nodes might be sending. This forms the input vector

$\mathbf{x} = [x_1\, x_2\, x_3 \cdots x_N]^{\mathrm{T}}$ where x_i is the symbol transmitted from the ith node. Given the matrix of channel fades $\mathbf{H}$ and the noise $\mathbf{n}$, at the receiver the signal is

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}. \tag{3.1}$$

The first step at the decoder is to multiply by the matched filter, $\mathbf{G} = \mathbf{H}^{\dagger}$, where $\mathbf{H}^{\dagger}$ is the conjugate transpose of $\mathbf{H}$, resulting in

$$\mathbf{G}\mathbf{y} = \mathbf{G}\mathbf{H}\mathbf{x} + \mathbf{G}\mathbf{n}. \tag{3.2}$$

The matrix $\mathbf{GH}$ is square, and invertible with high probability, with $\mathbf{F} = (\mathbf{GH})^{-1}$. In general, $\mathbf{F}$ is the pseudoinverse of $\mathbf{GH}$ since it might not be invertible. The next step is to do zero-forcing by left-multiplying with the pseudoinverse,

$$\mathbf{F}\mathbf{G}\mathbf{y} = \mathbf{F}\mathbf{G}\mathbf{H}\mathbf{x} + \mathbf{F}\mathbf{G}\mathbf{n} = \mathbf{x} + \mathbf{F}\mathbf{G}\mathbf{n}. \tag{3.3}$$

What remains is the original symbols in the presence of colored noise. Instead of decoding all the symbols, only the symbol that experiences the least noise and interference is decoded. The noise power is assumed to be identical on all channels, but the interference is not because each symbol experienced N different wireless channels. The "best" symbol can be chosen by looking at the diagonal elements of $\mathbf{F}$, which are a measure of the strength of the channel experienced by each transmitted symbol. The row corresponding to the diagonal element of $\mathbf{F}$ which has the smallest magnitude is then decoded to get the estimated symbol $\hat{x}_i$. Assuming that this decoding step is correct, the interference caused by this symbol on all of the other symbols can be canceled by subtracting the decoded symbol. In the high SNR regime, this approximation is a good one, since the BER will be low. Defining $\mathbf{r} = \mathbf{F}\mathbf{G}\mathbf{y}$ as the symbols to be decoded, $\mathbf{r}$ is updated to remove the interference from $\hat{x}_i$, with

$$\mathbf{r} = \mathbf{r} - \mathbf{H}(:, i)\hat{x}_i, \tag{3.4}$$

where the interference of the detected symbol $\hat{x}_i$ on the other symbols is $\mathbf{H}(:, i)$ (the ith column of $\mathbf{H}$) times $\hat{x}_i$. Now that the interference caused by x_i has been canceled, the ith row and column of $\mathbf{H}$ and the ith row of $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ can be deleted. This reduces the problem from $M \times N$ to $(M-1)\times(N-1)$. The same procedure of calculating the pseudoinverse, then decoding and canceling the "best" symbol from the remaining ones is performed until all the symbols have been decoded. The decoding of the last symbol is equivalent to decoding a $1 \times (M - N)$ system. Note that there must always be at least as many receive antennas as transmit antennas for the VBLAST algorithm to work. In summary, the decoding procedure is

1. Multiply received signal by $\mathbf{G}$ (matched filter).

2. Multiply by pseudoinverse **F** (zero forcing).

3. Find and decode the strongest receive antenna i (ordering).

4. Subtract ith column of **H** times the decoded symbol (cancel interference).

5. Remove ith column of **H** and ith row of received signal (reduction).

6. Repeat procedure if there are still rows remaining (iterate).

By ordering the decoding of the symbols according to the channels they experienced, the symbols with the best channels (and therefore the most resistant to noise and interference) are decoded first. The symbol that is decoded last is the least resistant to noise and interference, but all of the interference from the other transmitted symbols have already been canceled, so it sees the least interference. If there are any decoding errors in the early iterations, then all subsequent decisions will also be wrong. However, since we are interested in high rate and reliable communications, the SNR is assumed to be high and errors are rare. The performance is then limited by the weakest symbol, where the effective SNR is at its lowest.

If there are more receive than transmit antennas, then each additional receive antenna adds diversity to every transmitted stream, so one extra receive antenna results in an error probability P_e proportional to $1/\text{SNR}^2$, for example. The performance of a system with four transmitting nodes and varying numbers of server receive antennas is given in Figure 3-4, showing the increasing diversity gain. In terms of typical operations, any given frequency bin will likely have more antennas available than nodes assigned to transmit/receive on it, allowing for significant performance gains. For example, at 10^{-3} bit error rate, doubling the number of antennas from four to eight gives a performance gain of about 14 dB, and at 10^{-5} the gain is almost 35 dB. Coding of each individual bit stream may also be done to improve error performance at the expense of rate, but again this is without requiring the nodes to share information.

If a node has more than one antenna with which to transmit, then its maximum throughput is multiplied by the number of antennas. For example, a node with four antennas is able to transmit data four times as fast as a single antenna node. The multi-antenna node does not have to do any more work in coding for the antenna, since each antenna can be coded without taking into account what is being sent by any of the other antennas. The simulations in Figures 3-3 and 3-4 assume equal average received signal power at the server (receive) antennas. To accommodate the different path losses caused by nodes being at varying distances, we employ power control at the transmitting nodes, which is described in the allocation phase, below.

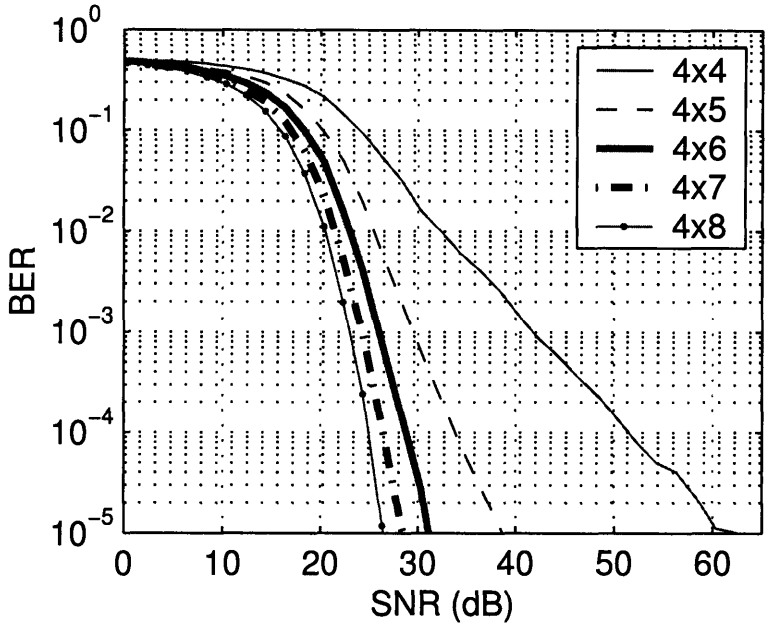


Figure 3-4: Uplink uncoded 64-QAM BER performance for 4 nodes with extra server antennas. From right to left, the curves represent 4, 5, 6, 7, and 8 server antennas.

3.2.3 The Downlink Protocol

In this phase, the transmit antennas of the server are able to share information, but the receive antennas of the nodes are not. The strategy used in the uplink phase cannot be used for the downlink because the receiving node antennas are not able to cooperatively share information without affecting their available bandwidth. However, given that the transmitter knows the channel characteristics as well as non-causal knowledge of what it is transmitting, precoding [20] of the data streams can produce the same effect of N non-interfering subcarriers as in the uplink phase.

This precoding strategy is very similar to Tomlinson-Harashima (TH) precoding [70, 25], in which the ISI caused by each transmitted symbol is canceled at the transmitter by subtracting out the ISI components in subsequent symbols. In this case, the interference is caused by the non-diagonal terms of the channel matrix. The downlink precoding has the effect of inverting the channel, so that at the receiver node antennas it appears that the each receive antenna experiences no interference from the signals meant for the other antennas. Thus the received signal at each node is as if only a single antenna had been transmitting. Because the precoding has the effect of inverting the channel, there is the potential of greatly increasing the average transmitted power. To prevent the power increase, the precoding employs modulo

arithmetic, which leads to only a small increase in transmit power. This adds a small amount of complexity at the receiving nodes, which need to employ modulo arithmetic to decode. This is not a significant complexity increase, however, and the decoding can again be done symbol by symbol.

Starting again from the uncoded system $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}$, we focus on the channel matrix $\mathbf{H}$. For simplicity, it is assumed that we have an equal number of server transmit antennas as receive nodes (each node with one antenna). First, a QR factorization on $\mathbf{H}^\dagger$ results in

$$\mathbf{H}^\dagger = \mathbf{Q}\mathbf{R} \Longrightarrow \mathbf{H} = \mathbf{R}^\dagger\mathbf{Q}^\dagger. \tag{3.5}$$

The matrix $\mathbf{Q}^\dagger$ is unitary, which does not increase the energy of the signal as it is only a rotation of the signal space. Ultimately, each receive node should have no interference from the symbols sent to all the other receive nodes, therefore the product of the transmit precoding with $\mathbf{H}$ should be a diagonal matrix. Because $\mathbf{R}^\dagger$ is lower triangular ($\mathbf{R}$ is upper triangular in a QR decomposition), we can precode $\mathbf{x}$ iteratively by doing back substitution. For example, for four receive nodes,

$$\mathbf{R}^\dagger = \begin{bmatrix} r_{11} & 0 & 0 & 0 \\ r_{21} & r_{22} & 0 & 0 \\ r_{31} & r_{32} & r_{33} & 0 \\ r_{41} & r_{42} & r_{43} & r_{44} \end{bmatrix}. \tag{3.6}$$

Given the intended input vector $\mathbf{x} = [x_1\, x_2 \cdots x_N]^{\mathrm{T}}$, we can define the vector $\tilde{\mathbf{x}}$ with elements given by

$$\begin{aligned} \tilde{x}_1 &= \mathrm{mod}\,(x_1) \\ \tilde{x}_2 &= \mathrm{mod}\left(x_2 - \frac{r_{21}}{r_{22}}\tilde{x}_1\right) \\ \tilde{x}_3 &= \mathrm{mod}\left(x_3 - \frac{r_{31}}{r_{33}}\tilde{x}_1 - \frac{r_{32}}{r_{33}}\tilde{x}_2\right) \\ \tilde{x}_4 &= \mathrm{mod}\left(x_4 - \frac{r_{41}}{r_{44}}\tilde{x}_1 - \frac{r_{42}}{r_{44}}\tilde{x}_2 - \frac{r_{43}}{r_{44}}\tilde{x}_3\right), \end{aligned} \tag{3.7}$$

where $\mathrm{mod}(\cdot)$ is the modulo operation, which wraps the value of its argument so that it doesn't get too large. Specifically, for a square QAM constellation with distance d between constellation points, the $\mathrm{mod}(\cdot)$ function wraps all values inside the square region that is $d/2$ larger than the maximum amplitude of the constellation. For a QAM constellation, the increase in transmit power is $\frac{M}{M-1}$, where M^2 is the number of constellation points [20]. In Figure 3-5, for example, the modulo operation is diagrammed for a 16-QAM constellation, which maps all of the points in the plane into

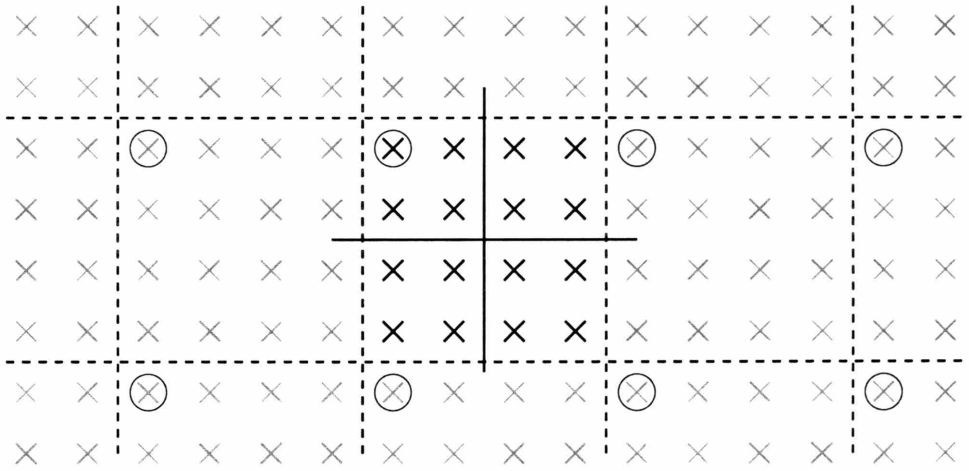


Figure 3-5: Downlink precoding: all of the shaded constellation points are mapped to the center constellation, so all the circled points are equivalent, for example.

the dashed square, and all the gray constellation points are mapped into the original (black) constellation points inside the dashed square. The modulo operation in effect turns the 16-QAM constellation into an infinite lattice, where each constellation point of the original constellation represents a coset of the infinite lattice. Then to complete the precoding procedure, we apply $\mathbf{Q}$, resulting in

$$\mathbf{x}' = \mathbf{Q}\tilde{\mathbf{x}}. \tag{3.8}$$

Sending $\mathbf{x}'$ across the channel results in [20]

$$\begin{aligned} \mathbf{y} &= \mathbf{H}\mathbf{x}' + \mathbf{n} \\ &= \mathbf{R}^\dagger\mathbf{Q}^\dagger\mathbf{Q}\tilde{\mathbf{x}} + \mathbf{n} \\ &= \mathbf{R}^\dagger\tilde{\mathbf{x}} + \mathbf{n} \\ &= \mathbf{D}\mathbf{x} + \mathbf{n}, \end{aligned} \tag{3.9}$$

where $\mathbf{D}$ is a diagonal matrix with $d_{ii} = r_{ii}$. To decode, the received signal is multiplied by $\mathbf{D}^{-1}$, and then lattice decoding is performed. The resulting signal y_i' that each receive node sees is

$$y_i' = \mathrm{mod}\left(x_i + \frac{n_i}{r_{ii}}\right). \tag{3.10}$$

The performance of this algorithm is shown in Figure 3-6 for both four and sixteen receive nodes. Note that there is hardly any difference in the performance between the two cases, which closely match the performance of a single antenna system.

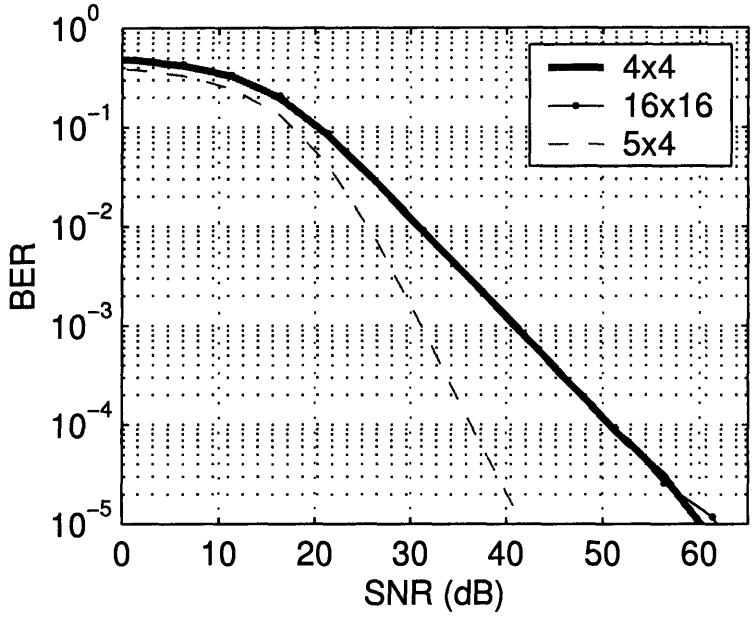


Figure 3-6: Downlink BER performance for 4 (dotted line) and 16 (dashed line) downlink nodes, with an equal number of server antennas. The diversity gain for an extra transmit antenna is also shown.

The amount of power that is allocated to each transmit antenna is such that the receive antennas all see the same signal strength. Since the receive antennas are going to be differing distances from the server, the path loss to each one will in general be different. The magnitude of these path losses are estimated in the training and allocation phases below, and scale the transmit powers accordingly.

If there are more antennas at the server than there are nodes receiving, they are automatically used by the procedure outlined above. In this case, the triangular matrix **R** will have rows that are all zero, but that signifies that the rank of the transmitting matrix is the same as the number of receive nodes, which is less than the number of transmitting antennas. Each extra transmit antenna gives an extra degree of diversity to each node, mirroring the way extra antennas are used in the uplink phase. This diversity gain from an extra transmit antenna is also shown in Figure 3-6.

3.2.4 The ACK Phase

Given the fluctuating nature of the wireless channel, there will be times when the channel is not able to support the data rate of the transmitted packet. There might

also be interfering networks or other sources of noise which severely degrade the channel. In these cases, no amount of coding will allow the packet to be successfully transmitted, and hence the packet will be lost. A standard automatic repeat request (ARQ) protocol allows the receiver node to send a NACK (not acknowledge) packet back to the transmitter if it is not able to successfully decode the packet, otherwise it will send an ACK (acknowledgment). The ACK/NACK will go through the same uplink and downlink phases as the original packet, but in the reverse direction. If the transmitter receives a NACK (or no reply at all), it will retransmit the packet, whereas an ACK will signal that the next packet can be sent.

3.2.5 The Control Channels

One of the assumptions that is made in the uplink phase is that there are as many receive antennas at the server as there are nodes that wish to transmit. Similarly, during the downlink, the number of receive nodes is assumed to not be more than the number of server antennas. The uplink and downlink algorithms are able to utilize any extra antennas to increase the diversity order of the transmissions, but these algorithms break down when there are more nodes that wish to transmit than antennas at the server. If the server has N antennas, then the network is able to support N nodes transmitting simultaneously in each frequency bin. In a given network, however, there may be many times as many nodes as there are server antennas and frequency bins. For example, in a lecture hall, there may be several hundred people, each with one or more cellular phones, pagers, PDAs, or laptop computers. Equipping a server with hundreds of antennas is clearly not viable, nor is the room likely to be able to support so many degrees of freedom.

Even if only a small subset of the nodes in the network want to transmit during any given frame, that number may still exceed the network resources, especially since a high bandwidth node with multiple antennas can completely occupy several frequency bins. One possible solution is to employ a pair of control channels to allow nodes to reserve transmission slots for a frame. The control channels are two reserved frequency bins of the OFDM symbol (one each for the uplink and downlink), and may vary in location from frame to frame in a predetermined pattern to minimize losses due to slow fades. The control channels allow the nodes to request bandwidth and the central server to allocate resources for the next frame.

When a node wishes to transmit packets in the next frame, it will try to reserve a space during the current frame using the uplink control channel. The protocol for transmission is similar to slotted ALOHA [1], where the slots are aligned to the samples of the OFDM symbol. A node transmits a very small packet which contains its unique identifier string, as well as the rate it wishes to transmit at. Because the server has N antennas, the node's packet can be decoded even if $N-1$ other nodes

are simultaneously trying to reserve transmission time. When the server receives the packets, it decides whether each node will be allowed to transmit during the next frame (i.e., if there is enough bandwidth available). The server then transmits a reply packet on the downlink channel to each node. If a node is allowed to transmit, it will receive a set of frequency bin assignments which it will use during the training phase of the proceeding frame. If a node is not allowed to transmit, it will receive a deny packet, and must wait until the next frame before trying again to request transmission access. Rather than forcing a node to keep re-requesting bandwidth for every single frame, a request for bandwidth is honored for successive frames until a a period of unuse is detected by the server, or the node sends a cancel request. While the connection is active, the server will send the node frequency assignments at every frame interval, so the node is still only allowed to transmit if it has received an assignment from the server, but it does not have to continuously request a connection.

If more than N nodes try to reserve at the same time, then their transmissions will collide and the server will not be able to decode the packets. The nodes will not receive a reply packet from the server, and so each will assume that the request packet was not received. If there is enough time left in the current frame, the nodes can try to retransmit their request. Each node, before retransmitting, will wait a random amount of time to prevent all the nodes from transmitting at once and colliding once again. If the node is not able to retransmit before the start of the next frame, then it is not allowed to transmit its data packets in the following frame, and must try again to request transmission bandwidth. Additionally, if the server determines that all the bandwidth has been allocated for the next frame, it will send a general deny packet on the downlink control channel to keep nodes from futilely requesting bandwidth.

This scheme works best when the nodes that are requesting transmission bandwidth are trying to send large amounts of data per frame, or long streams of data and so will tend to be assigned multiple frequency bins or multiple time frames. If the nodes that wish to transmit only want to send very small amounts of data, the network should be able to support $\mathcal{O}(N \times \#$ frequency bins) nodes. However, given the uplink/downlink request structure, there is a considerable amount of overhead for the nodes to be allocated channels because the probability that more than N nodes will try to request access at the same time is high, resulting in collisions and retransmissions.

3.2.6 The Training Phase

In the training phase, each node allowed to transmit in the current frame sends a packet with a unique identifier and a predetermined training sequence on its assigned frequency bins. The power that the node transmits at is also predetermined, so that the server knows what the path loss to the node is. The number of assigned frequency

bins is larger than the number of bins needed, on average, to support the rate requested by the node, allowing the server some latitude in the final bin assignments in the allocation phase. Ideally, each node would send its training sequence in all of the frequency bins so that the server could optimally choose frequency bin assignments for each node, but this limits the number of nodes allowed to transmit in a given frame to the number of server antennas to ensure that the server can untangle all of the training sequences. Alternately, the training sequences could be lengthened to increase the number of sequences that could be decoded in a frequency bin, but that reduces the amount of data that can also be transmitted in the same frame.

The length of the training sequences also determine how accurately the channel can be estimated. The minimum length of the training sequence is determined by how long the impulse response of the channel is. Since the cyclic prefix for the OFDM symbol is chosen to be as long as the channel response, the training sequence should also be at least as long as the cyclic prefix. For the uplink phase, perfect knowledge of the channel is assumed. However, with channel estimation errors, there is additional noise which ultimately limits performance. Given that the channel is known exactly, the interference from other antennas can be exactly canceled at each receiver. Errors in the channel estimation lead to additional interference from the other antennas. These errors raise the noise floor of the signal, so the error performance will not improve with received SNR after a certain point. It is therefore important to estimate the channel accurately, but perfect knowledge is not required.

At the server, decoding is done in much the same way as the uplink phase. The unique identifier tells the server which node it is currently detecting, and allows it to know what training sequence has been sent as well as the power it was transmitted with. After the server has estimated the path loss and frequency response of the channel between itself and each node, it stores that information along with the requested rates and quality of service for the allocation phase.

3.2.7 Bandwidth Allocation Phase

Once all the fading coefficients have been learned, the server decides how to allocate the frequency bins to each node. Multiple nodes can use the same frequency bin as long as there are at least as many antennas as nodes. A node can be assigned different numbers of frequency bins, sizes of constellations for each assigned bin, as well as transmit power per bin.

One possible algorithm for allocating the available bandwidth is to employ spatio-temporal waterpouring to maximize the total throughput of the network. For a single node, the optimal power allocation strategy is to allocate power across the frequency band by "waterpouring" into the inverse of the SNR function of the wireless channel for that node [11]. Thus the frequency bins with the higher SNR will be allocated

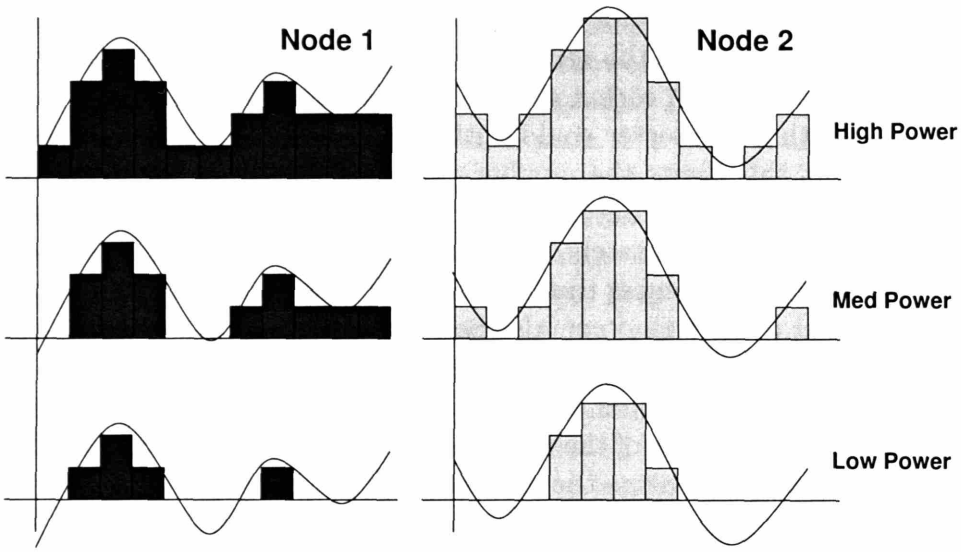


Figure 3-7: Bandwidth allocation for two nodes and several total power levels.

power first, and the total power will be spread across the frequency bins according to the relative SNR in those bins. This can be extended to the case of multiple nodes by waterpouring across a two-dimensional surface formed by stacking the frequency response curves of all the nodes which are eligible to transmit. The power is allocated across both frequency and space (since the nodes are spatially separated), and this allocation varies with time as the SNR functions for the nodes fluctuate over time.

An example of this algorithm for two nodes is shown in Figure 3-7. The SNR functions for both nodes and the power allocations for several total power levels are represented, showing the effect of the waterpouring on the power allocation over the frequency bins for the two nodes. When the total amount of power to be allocated is high, nearly all of the frequency bins are used for both nodes. As the total transmit power falls, the threshold SNR for a frequency bin to be allocated some power increases, until only the few bins with the highest SNRs are employed. This algorithm maximizes the total throughput of the network for a given total power. The algorithm also scales easily with the types of nodes trying to transmit simultaneously. The network can support a small number of high-power nodes which transmit at high data rates, or a larger number of low-power nodes which transmit at lower data rates, or any combination of these. To deal with the differences in the power requirements and requested data rates, the SNR functions of the nodes can be scaled accordingly before the waterpouring to change the allocation of power between nodes without affecting the ratios of power allocated to each frequency for each node.

There are several other issues that the power allocation algorithm must take into

account, however. If the total power to be allocated is very high, then there might be too many nodes using a certain frequency bin according to the strict waterpouring scheme. Since increasing the bandwidth used by a node increase the capacity linearly, while increasing the power only increases the capacity logarithmically, the most efficient way to increase the total throughput would be to allow all of the nodes to use as much bandwidth as possible. In either case, the situation could easily arise where certain frequency bins would need to be assigned to more nodes than there are server antennas, preventing the signals from being able to be decoded at the receiver.

Additionally, since the training sequences sent out by each node only cover a subset of the frequency bins, the unknown frequency bins will not be used (assumed to have an SNR of $-\infty$). A node may also request a high data rate, but may be stuck in an area where the SNR is too low to support that rate, while another node requests a lower rate but has access to a channel with a much better SNR. The server needs to distribute the power in a fair manner, so that the node with the poor SNR is not "starved" of bandwidth. A proportional fair allocation [73] is one method to insure that all nodes are allocated sufficient resources on average. Finally, some nodes may have requested a quality of service which requires a minimum allowable data rate, or a maximum response time, which the server must take into account when allocating power and bandwidth to the nodes.

Once the power and frequency allocations have been determined, the server can relay this information back to the nodes in the same manner as a downlink transmission. The server transmits back to each node the bandwidth it is allocated, along with the power it should transmit at for each of those frequency bins, and the constellation to use in each one.

3.2.8 Concatenated Code Viewpoint of Uplink/Downlink

The uplink and the downlink coding schemes are duals of each other, with the uplink employing one-dimensional encoding (the transmit antennas are not required to coordinate transmissions) with iterative decoding, and the downlink employing iterative encoding with one-dimensional decoding. These two coding schemes can also, however, be described in the guise of concatenated coding schemes which treat the transmitting and receiving antennas as virtual arrays.

For the uplink, the outer encoder is the individual uncoded bit stream for each node antenna. The nodes may optionally use error correction coding, but these would occur before the outer encoder. With a general multiple-input system, the encoding is more efficient if the coding is performed jointly across all of the data streams, but this requires communication between the transmitting nodes. The inner encoding is performed by the wireless channel, which is known to the transmitter and receiver but cannot be changed. At the receiver, the decoding for the VBLAST algorithm can

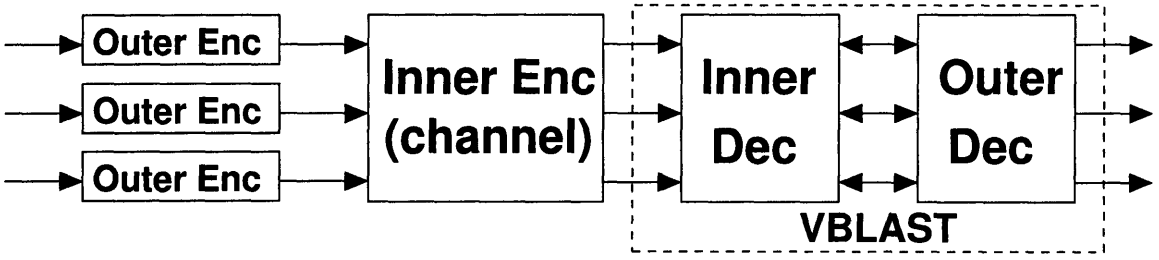


Figure 3-8: Concatenated code viewpoint for the uplink. One encoder outputs the bit streams for each node antenna, with the inner and outer decoders corresponding to the pseudoinverse and cancellation operations, respectively, of the VBLAST decoder.

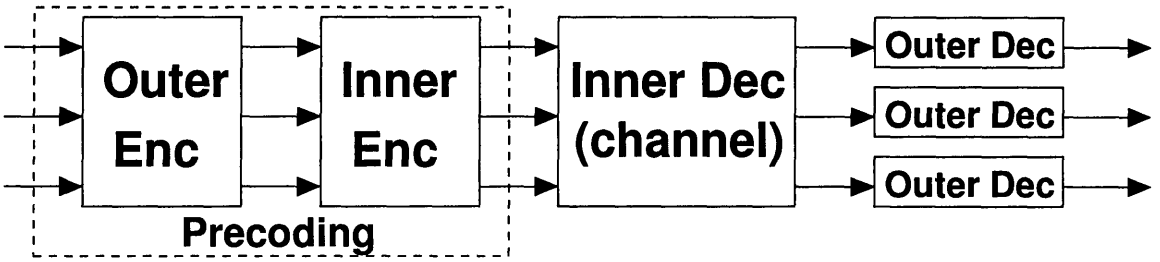


Figure 3-9: Concatenated code viewpoint for the downlink. The outer and inner encoders are the back-substitution and rotation operations, respectively, of the transmit precoder.

be broken up into two parts. The first, or inner, decoder removes the interference effect of the channel from the undecoded nodes, and corresponds to the pseudoinverse operation of the VBLAST decoder. The second, or outer decoder, removes the effects of the decoded node on all of the remaining nodes, and corresponds to the cancellation step of the VBLAST decoder. This is diagrammed in Figure 3-8, where the dotted box represents the VBLAST decoder.

The downlink phase can be similarly defined, with the roles of the encoder and decoder reversed. Here, the outer decoder treats the data stream from the N nodes as N separate data streams to be decoded individually. For the encoder, the precoding operation can be separated into two parts. The outer code is the back-substitution matrix $\mathbf{R}$, which precancels the effects of previously encoded nodes. The inner code is the unitary matrix $\mathbf{Q}$, which then precancels the rotation from the channel. The decoding of the inner and outer codes is performed by the wireless channel, resulting in individual data streams at the nodes, as shown in Figure 3-9. Here, the dotted box encloses the operations for the precoding algorithm which encodes the data streams from the server.

The separation of the encoder streams during uplink and the similar separation of the decoder streams during downlink allow the nodes to communicate with each other as part of a virtual array without requiring extra communication to coordinate. This is an inherently limiting constraint from the standpoint of the diversity-multiplexing tradeoff, but the VBLAST uplink architecture is optimal in terms of achievable rate for single antenna nodes [71]. However, for nodes with multiple antennas, using each antenna to transmit a separate data stream is not in general as good as if both antennas were used jointly. The multiple antenna node then has the option of either using its antennas separately or jointly. This can be done transparently to the receiving node, because the central server receives each packet during the uplink, and can reencode the data for the downlink. Thus it is possible for a multiple antenna node to achieve multiplexing gain when transmitting even if the receiver node has only a single antenna.

Similarly, the downlink precoding achieves the maximum multiplexing gain, but multiple antenna nodes are able to achieve better SNR performance if they decode their antennas jointly. As with the uplink, a multiple antenna node can achieve a high multiplexing or diversity gain even if the transmitting node has only a single antenna due to the relaying of the packets through the central server.

3.3 SNR Gain From Multiple Antennas

As we have seen, the uplink/downlink mode of communication allows the network nodes assemble into virtual arrays for transmission and reception. The use of the central server allow these virtual arrays achieve multiplexing gain without needing to coordinate with each other. The server multiple antennas can also be used for diversity, providing SNR gain when compared to a single antenna system. Figure 3-10 shows the simulated SNR gains relative to a 1×1 system assuming fixed channel conditions.

As the left plot shows, even with capacity-achieving codes, the SNR gain can be very large as the data rate grows. For example, a 2×2 system can achieve 6 b/s/Hz with 10 dB lower SNR than a 1×1 system could. A multiple antenna node could in principle achieve this SNR gain if there were no constraints on computation or delay. Since the capacity of a $1 \times N$ system is the same as that of a 1×1 system (to first order), single antenna nodes cannot achieve this SNR gain.

For uncoded systems, however, even single antenna nodes can experience large SNR gains from the diversity achievable via the server antennas. The middle and right plots of Figure 3-10 show the SNR gain for uncoded communications at a fixed data rate (6 b/s/Hz) using 64-QAM for $1 \times N$ and $N \times N$ systems. The $N \times 1$ systems (such as for the downlink) have the same performance as the $1 \times N$ curves since the

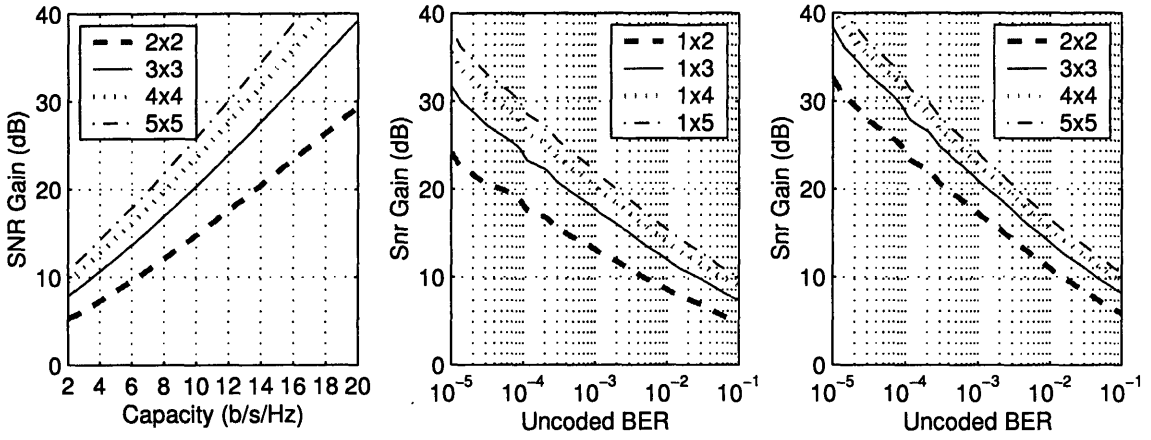


Figure 3-10: Achievable (left) and uncoded (middle, right, for 64-QAM) SNR gain for $1 \times N$ and $N \times N$ systems for $N = 2$ to 5, relative to a 1×1 system.

server knows the channel. The SNR gain for a single extra antenna at the server is over 20 dB for an uncoded BER of 10^{-5}, increasing to almost 40 dB for a 1×5 configuration. The added SNR gain can be almost 10 dB for an $N \times N$ compared to a $1 \times N$, but each additional antenna results in a smaller incremental gain in SNR than from the previous one.

3.3.1 Average Throughput With Adaptive Modulation

While Figures 2-10 and 3-10 shows the large gain in SNR of multiple antenna systems over the 1×1 system for 64-QAM constellations, the actual constellations used at any given time varies depending on the channel conditions. To compare the performance of the different systems, the average throughput must be calculated. The left plot of Figure 3-11 shows the BER curves for a 1×1 and a 1×4 system for M-QAM constellations of sizes 4, 16, 64, and 256. The dotted lines show the two families of BER curves, with the solid lines showing the parts of the curves used for a target maximum BER of 10^{-3}. For the 1×1 system, for example, when the SNR is below 30 dB, nothing is transmitted in that frequency bin. Between 30 dB and 36 dB, a 4-QAM constellation is used, 16-QAM from 36 dB to 41 dB, and so on up to 256-QAM. Since 256-QAM is the highest order constellation used, the maximum spectral efficiency achievable will be 8 b/s/Hz. The horizontal dashed line marks an average throughput of 6.67 b/s/Hz, which represents 1 Gb/s if the total available bandwidth is 150 MHz, as with the WiGLAN. This throughput will be used as an example target throughput for SNR comparisons.

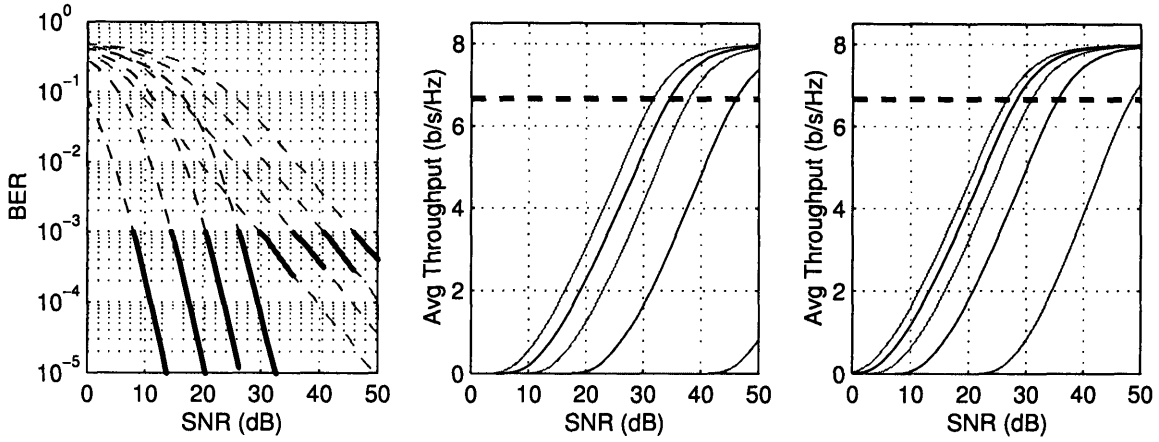


Figure 3-11: BER curves (left plot) for a 1×1 and 1×4 system for uncoded 4-, 16-, 64-, and 256-QAM constellation sizes (dashed, left to right), with BER curves for adaptive modulation at 10^{-3} BER (solid). Also plotted is expected throughput vs. SNR for 10^{-5} (middle) and 10^{-3} (right) BER, respectively for 1×1 to 1×5 systems (bottom to top). The horizontal dashed line represents 6.67 b/s/Hz, or 1 Gb/s for 150 MHz bandwidth.

The center and right plots of Figure 3-11 show the expected throughput for the WiGLAN given an average SNR for Rayleigh fading. The curves are for $1 \times N$ systems with N ranging from five (top curve) to one (bottom curve). The middle plot has a maximum BER of 10^{-5}, with the 1×1 curve barely visible at the bottom. The right plot has a maximum BER of 10^{-3}, and shows a higher throughput curve vs. SNR for all of the antenna configurations, including the 1×1 curve. For a target average throughput 6.67 b/s/Hz at an uncoded BER of 10^{-5}, the 1×4 system requires an SNR of 39 dB, while a 1×1 system requires 67 dB.

3.3.2 Using Error Correcting Codes

Although a BER of 10^{-5} is low enough that uncoded transmissions are feasible, the SNR required may still be too high even with the SNR gain from using multiple antennas. As Figure 3-11 shows, operating at a higher BER, such as 10^{-3}, can result in a reduction in the required SNR. An uncoded BER of 10^{-3} decreases the required SNR for a 4×4 system by 3.5 dB, and is a low enough error rate that a relatively uncomplicated algebraic block code such as a Reed-Solomon (RS) code can be used to protect against errors. Reed-Solomon codes are chosen as the error correcting code here because it is easy to vary the strength and rate of the code to change the error correcting abilities. At an even lower uncoded BER of 10^{-2}, even the 2×2

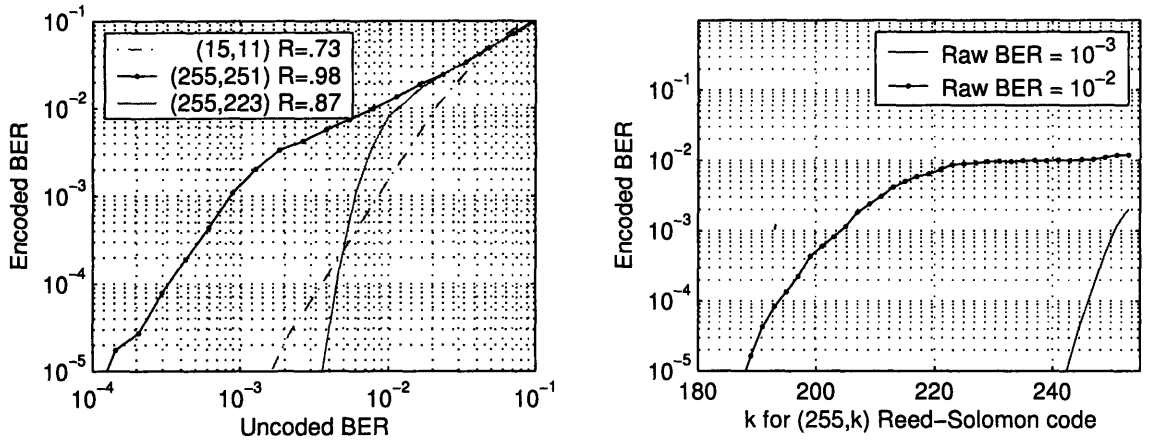


Figure 3-12: Encoded BER for Reed-Solomon codes of varying lengths (left), and optimizing code parameters for a given target BER using the (255,k) code family (right).

system can surpass 6.67 b/s/Hz with less than 30 dB SNR. Even at these lower BER requirements, however, the 1×1 system is still unable to achieve significant data rates without an unrealistic SNR or very powerful coding with long block lengths, so multiple antennas are still likely to be needed even for average rates of 2 b/s/Hz.

Figure 3-12 shows how the BER of uncoded systems can be improved by using a Reed-Solomon block code of varying lengths. An (n, k) RS has a code rate of k/n and requires a block length of n symbols. Each symbol is $\log_2(n-1)$ bits long, for a total block size of $n \log_2(n-1)$ bits. Thus a (15,11) RS code requires 60 bit block lengths and has a code rate of almost 3/4, while a (255,223) RS code requires 2040 bit blocks and has a code rate of almost 7/8. As the left plot shows, the (15,11) RS code can reduce the raw BER from 10^{-3} to less than 10^{-5}. By increasing the symbol size, such as with the (255,223) RS code, even better BER gains can be made with less rate loss. The cost is longer block lengths and increased complexity, however.

The right plot of Figure 3-12 shows how the BER improvement of a (255,k) RS code varies as k is varied. The bit errors are assumed to occur uniformly in a data block. Since the (255,k) RS codes use eight bit symbols, a single encoded block is 2040 bits long. For an uncoded BER of 10^{-3}, a (255,241) code is sufficient to reduce the BER to 10^{-5} or less, for a code rate near 19/20. As we can see from Figure 3-11, that allows a 3×3 system to achieve a BER of 10^{-5} at an SNR of 28 dB. Similarly a (255,187) RS code can reduce a raw BER of 10^{-2} to a coded BER of 10^{-5} with a code rate near 3/4, allowing a 2×2 system to achieve the target 6.67 b/s/Hz with an SNR of 28 dB. Thus using a mild amount of coding can significantly reduce the SNR required, which can reduce the number of antennas required by the system.

3.4 Managed vs. Ad Hoc Network

Although the uplink/downlink architecture is able to increase the throughput of the network as a whole, it is not always the best choice for all possible network conditions. This architecture is designed to achieve the maximum overall network throughput, but because of transmit power limitations, it is not always the best choice for communication between nodes.

3.4.1 Diversity vs. Multiplexing

If the nodes are very far away from the server, a great deal of transmit power may be required for the nodes to communicate with the server. The uplink/downlink architecture is designed to achieve the maximum multiplexing gain for throughput, but it may require too much transmit power by the transmitting nodes if they are very far or shielded from the server. In lieu of increasing the transmit power, the server can instead switch to a high diversity mode of communications to save transmit power. For a single antenna node in the absence of other nodes, the SNR required for successful decoded can be greatly reduced if the central server used all of its antennas to receive the transmitted signal. As Figure 3-10 shows, increasing the diversity order from one to two can decrease the required SNR by 20 dB or more.

In the case of low SNR from a transmitting node, the server should devote at least one extra antenna to receiving the virtual node array signals to reduce the required SNR as shown in Figure 3-4. Thus, for example, if there are four server antennas, then a frequency bin should not have more than three nodes transmitting simultaneously. This gives up some of the multiplexing gain, since a 3×4 system can only support three independent data streams, but the diversity gain allows nodes to transmit at a greater range that they could otherwise. If the path loss is severe, the server may allocate all of its antennas to the node for maximum diversity, although the incremental SNR gain for each additional antenna decreases significantly. The SNR gain from diversity for the downlink phase is not as important because of the greater power capabilities of the server, although 20 dB or more is a very significant increase in power. In this case, the receive nodes can experience extra diversity gain if the central server similarly devotes extra antennas for each frequency bin.

3.4.2 Comparison to Ad Hoc Communication

If two nodes which wish to communicate are both close to each other and far away from the server (in terms of path loss), it could be advantageous for them to communicate directly rather than through the server. The server could allow direct node to node communications, but it still needs to know the path loss between the two nodes.

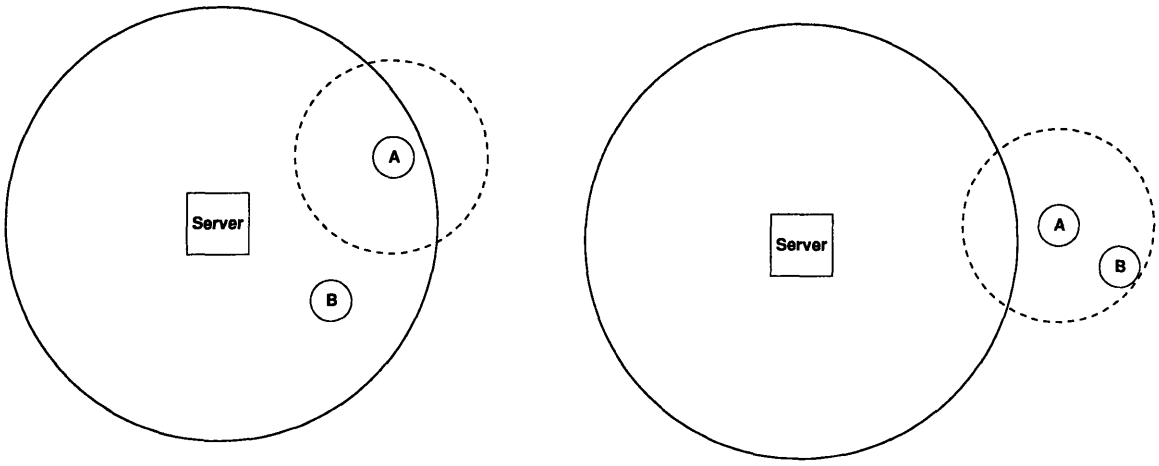
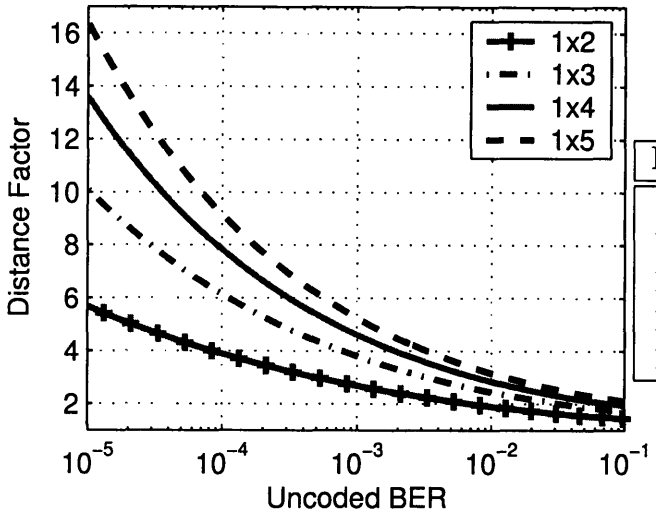


Figure 3-13: Circles represent the distance that the BER is not greater than some threshold for direct (dashed) and server-assisted (solid) communication between nodes A and B for a fixed transmit power.

Without this knowledge, two nodes communicating directly may use more transmit power than if they used the uplink/downlink with the server due to the diversity gain that the server can provide. This higher power then creates unnecessary interference to any other transmitting node in the same frequency bins, as well as shortens the battery life of the transmitting nodes.

From the uncoded SNR gain curves in Figure 3-10, we can see that for low bit error rates, the difference in SNR when the server uses extra antennas for diversity during the uplink phase can be very large, which translates into a distance gain. We can define a distance factor d_f, which is the ratio of the distance between the nodes and the central server to the distance between the two nodes. As Figure 3-13 shows, if node to node communication uses a transmit power P to achieve a certain uncoded BER at a distance x, then the node to server communication at the same transmit power achieves the BER at a distance xd_f. The dashed circle has radius x, while the solid circle has radius xd_f. These circles bound the areas in which the node to node and node to server communications achieve a BER threshold for a fixed transmit power. The left case is more common, favoring the server, while the right configuration shows an instance of when direct node communication requires less power. Since the area of the two circles are related by d_f^2, to first order the probability that this case occurs is $1/d_f^2$.

Figure 3-14 shows the distance factor for a given BER compared to the minimum diversity case assuming the signal strength degrades according to (2.2) with the path loss exponent $n = 3$ corresponding to Rayleigh fading. For example, if there is a



BER	1×2	1×3	1×4	1×5
10^{-1}	1.45	1.73	1.94	2.11
10^{-2}	1.90	2.45	2.85	3.16
10^{-3}	2.69	3.81	4.62	5.26
10^{-4}	3.89	6.16	7.85	9.18
10^{-5}	5.69	10.13	13.63	16.43

Figure 3-14: Multiplicative distance gain factor compared to a 1×1 system for diversity orders 2, 3, 4, and 5.

30 dB difference in the SNR, then for the same transmit power, the node can be ten times as far away and have the same error rate. For a BER of 10^{-5}, a single extra receive antenna allows the node to be over five times as far away. In order for it to be advantageous for two nodes to communicate directly, they have to be 5.69 times closer to each other than either is to the central server, which is expected to happen less than 5% of the time. With four antennas, the difference is much greater, with the two nodes needing to be 13.63 times closer together than either is to the server, which occurs less than 1% of the time. In the majority of cases, two nodes that wish to communicate will be "closer" to the central server than to each other. Using Reed-Solomon codes at different operating points of 10^{-2} and 10^{-3}, the diversity gain still decreases the apparent distance to the server by factors of 2–5.

In order for the server to learn the path loss between nodes, each candidate node in turn could send out a signal at a known power for all the other nodes to listen to. Since the transmit power is known, the received SNR for each node is a measure of the path loss between the two nodes. If there are N nodes, then each node builds a list of the path loss to each of the other $N-1$ nodes, and transmits the list back to the server. The server then has a total of $N(N-1)$ path losses to sort out into uplink/downlink nodes and ad hoc nodes. Due to node mobility or changing channel conditions, this needs to be recalculated on a regular basis — at least at intervals comparable to the coherence time. This extra communication can add an appreciable amount of overhead to the training and allocation phases, especially since it should be done over all frequency bins.

However, a much simpler way to enable direct node communications is for each

node not transmitting to listen during the uplink transmissions of the other nodes. If a node is able to decode a packet it is meant to receive, then it will also receive the same packet on the downlink from the server. When returning the ACK to the transmitting node, the received node can notify the server that it was able to decode directly from the transmitting node without aid from the server. It is up to the server to decide whether it should allow the nodes to communicate directly with each other or to continue communication through the server. Changing channel conditions plus the added diversity that the server can provide for transmission skews the preference to the uplink/downlink communication mode, unless the difference between the cellular and ad hoc models is large and resources are scarce at the server.

In addition, the ad hoc communications is only favorable if the nodes are able to communicate directly with each other without using other nodes as relays as would be suggested by [22, 23, 24] because the central server is essentially a relay node that is available to every node in the network. The resources of the server allow it to require a much lower transmit power for communication, and the single relay of the uplink and downlink allows delay sensitive transmissions to proceed without difficulty. By contrast, the ad hoc methods in [22, 23, 24] can require arbitrarily long delays, or still require some sort of centralized network control, which require the nodes to communicate with the central server anyway.

3.5 Effect of Coherence Time on Throughput

As described in Section 2.1.3, the size of the OFDM frequency bins needs to be less than the coherence bandwidth of the wireless channel so that the flat-fading approximation is valid. The coherence bandwidth does not say anything about how quickly the channel conditions change with time, however. The coherence time of the channel T_o is the amount of time that the channel statistics are essentially constant, and is dependent on the relative motion of the transmitter, receiver, and any objects in the wireless channel environment [63]. For the block fading model to accurate, a packet must not be longer than the coherence time. If the packet is longer than the coherence time, then the later time samples may experience a different fade from the earlier time samples, including the ones used to learn the channel statistics. Each packet has a duration T_p that is divided between the data time T_d and the system time T_s such that

$$T_p = T_s + T_d. \tag{3.11}$$

The system time T_s includes any header information, plus training data, while the data time T_d consists of several OFDM symbols of length T_{OFDM}, each with a cyclic prefix of time T_{CP}, so

$$T_p = T_s + N(T_{OFDM} + T_{CP}), \tag{3.12}$$

where N is the number of OFDM symbols in the packet. The total packet time must be no longer than the coherence time, so the maximum number of OFDM symbols that can be sent in a given packet is

$$N = \frac{T_o - T_s}{T_{OFDM} + T_{CP}}. \tag{3.13}$$

The fraction of time ζ in which data can be transmitted accounting for overhead in each packet can be calculated as

$$\zeta = \frac{NT_{OFDM}}{T_o} = \left(\frac{T_o - T_s}{T_o}\right)\left(\frac{T_{OFDM}}{T_{OFDM} + T_{CP}}\right), \tag{3.14}$$

where the first term in the multiplication is the fraction of time that is available for data, and the second term accounts for the overhead of the cyclic prefix for each OFDM symbol.

3.6 Implications for the WiGLAN

The WiGLAN operates in an indoor environment, and is capable of communication at or near gigabit data rates within a range of about ten meters with a BER of 10^{-5} or less. A target received SNR of 30 dB is chosen as the maximum that the network can reasonably expect to experience. For the WiGLAN, we can determine the minimum number of antennas required transmit at the target data rate of 1 Gb/s, both for uncoded and coded transmissions. We also consider the overhead required for the coherence time and bandwidth of the wireless channel in the WiGLAN system bandwidth.

3.6.1 Number of Server Antennas

In the WiGLAN bandwidth of 150 MHz, the OFDM frequency bins need to be less than the coherence bandwidth for flat fading across the bin. Since the coherence bandwidth of the 5 GHz band is about 4 MHz [39], the frequency ins can be assumed to experience flat fading as long as their size is less than 4 MHz. The size of the frequency bins has been chosen to be around 1 MHz wide, ensuring that the flat fading assumption is valid, and that there is sufficient flexibility for resource allocation.

Figure 3-15 summarizes the SNR required for the different antenna configurations to achieve an average throughput of 6.67 b/s/Hz, which corresponds to the 1 Gb/s target for the WiGLAN. As expected, adding antennas will lower the required SNR, though each additional antenna provides a smaller gain than the previous one.

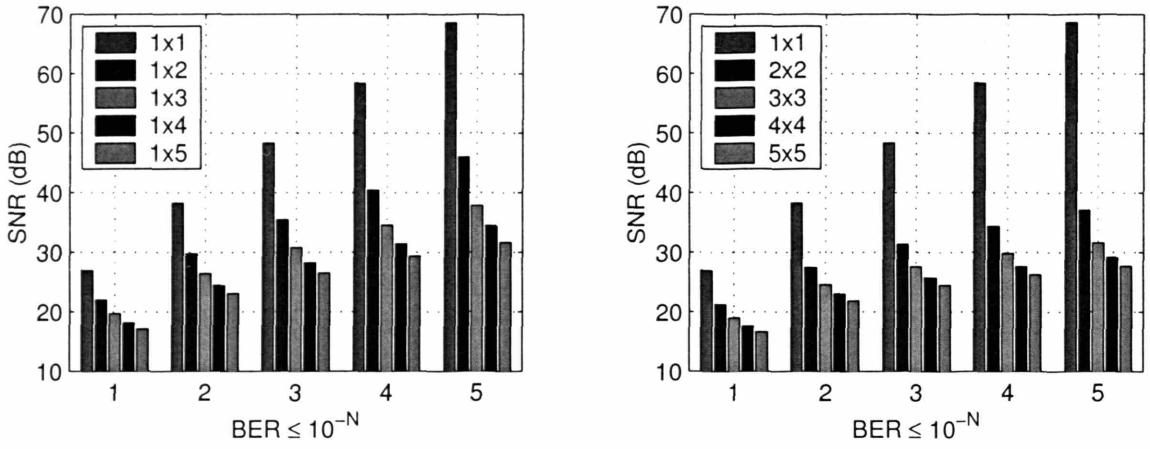


Figure 3-15: Average receive SNR required to achieve 1 Gb/s at different bit error rates using uncoded communications with adaptive modulation.

The 1×1 case requires almost 70 dB to reach 1 Gb/s at a BER of 10^{-5}, due to the large region (SNR below 30 dB) in which no data can be reliably sent. A BER of 10^{-5} is low enough that it is reasonable to consider sending uncoded data, with the associated savings in memory, computational complexity, and delay. Adding a transmit and receive antenna to create a 2×2 system results in a large reduction in the required SNR, achieving 1 Gb/s at 37 dB. The reduction in required SNR for each additional antenna shrinks rapidly, with diminishing returns making the case for more than a small number of antennas unfavorable. A 4×4 system is able to communicate at 10^{-5} BER with less than 30 dB SNR. Assuming that 30 dB is the limit to the required SNR, this is the minimum number of antennas that must be employed for reliable uncoded communications.

While a 4×4 system is the minimum required for uncoded transmission, a 3×3 system is feasible for a lightly coded system. For example, simulations have shown that a (255,241) RS code can correct an uncoded BER of 10^{-3} to less than 10^{-5}. The rate of this code is 19/20, saving 3.5 dB SNR for a 4×4 system, or 4 dB for a 3×3 system. Some of this savings in SNR is offset by the loss in rate, but the 3×3 system is still able to transmit at 1 Gb/s even after compensating for the rate loss.

For single antenna nodes, the uplink and downlink communications with the server can also benefit from the diversity gain that extra antennas at the server can provide. For the 30 dB SNR limit, all of the $1 \times N$ systems in Figure 3-15 require some coding as none are able to achieve a BER of 10^{-5} with the SNR less than 31 dB. Using the (255,241) RS code to operate at a BER of 10^{-3} results in larger reductions in SNR (6 dB for a 1×4 system), but the stronger (255,187) code is required to allow a 1×3

and higher antenna systems to operate below 30 dB. The 1×4 system is only very slightly above 30 dB when using the (255,241) code, however, so it almost qualifies.

3.6.2 Coherence Time

For the WiGLAN, the OFDM symbols are 128 time samples at 150 MHz, resulting in $T_{OFDM} = 853$ ns. The length of the cyclic prefix should be long enough to contain the maximum expected channel impulse response, which can be approximated by five times the delay spread [63]. Measured delay spreads for the indoor 5 GHz channel show delay spreads that range from 10–60 ns, with values over 100 ns for large open areas [4, 42]. For the 802.11a standard, the cyclic prefix time is 800 ns, allowing for even longer delay spreads [30]. Since the WiGLAN is primarily meant for indoor use, a 60 ns delay spread is a reasonable choice, resulting in $T_{CP} = 300$ ns. The system time T_s can also be taken from the 802.11a standard, which uses a value of 16 μs for the header. The coherence time can be written as a function of the relative velocity between the transmit, receiver, and/or an object that alters the wireless channel as [63]

$$T_o = \frac{\lambda/2}{V}, \tag{3.15}$$

where λ is the wavelength of the carrier frequency ($\lambda \approx 5.7$ cm for 5.25 GHz) and V is the relative velocity. The left plot of Figure 3-16 shows what fraction of the maximum packet length is available for data use after the header and cyclic prefixes are accounted for. Even if the coherence time is 1 ms or greater, the fraction of packet time that must be devoted to header and training information is 25%, which is close to the 20% required for standard 802.11a packets. The plot also shows the severe penalty for using the larger cyclic prefix length of 802.11a with the 128 bin OFDM symbols. In order to use the WiGLAN outdoors it is necessary to increase the number of frequency bins to lengthen the data time of the OFDM symbol. For the WiGLAN, 480 (or 512 for a power of 2) frequency bins are needed to bring the overhead penalty to under 20%. The computation penalty for increasing the number of bins by a factor of four is fairly modest due to the use of the FFT, however. If the peak to average power reduction algorithms in Chapter 6 are also used, however, the complexity grows linearly with the number of frequency bins, which could be prohibitive.

As Figure 3-16 shows, the fraction of the packet time available for data is not affected by the coherence time unless it is less than 100 μs. This corresponds to a relative velocity of over 25 m/s, or 90 Km/h (55 miles per hour), which is much faster that what would be normally expected in an indoor environment. Using the WiGLAN in a vehicular environment, however, could result in non-negligible overhead.

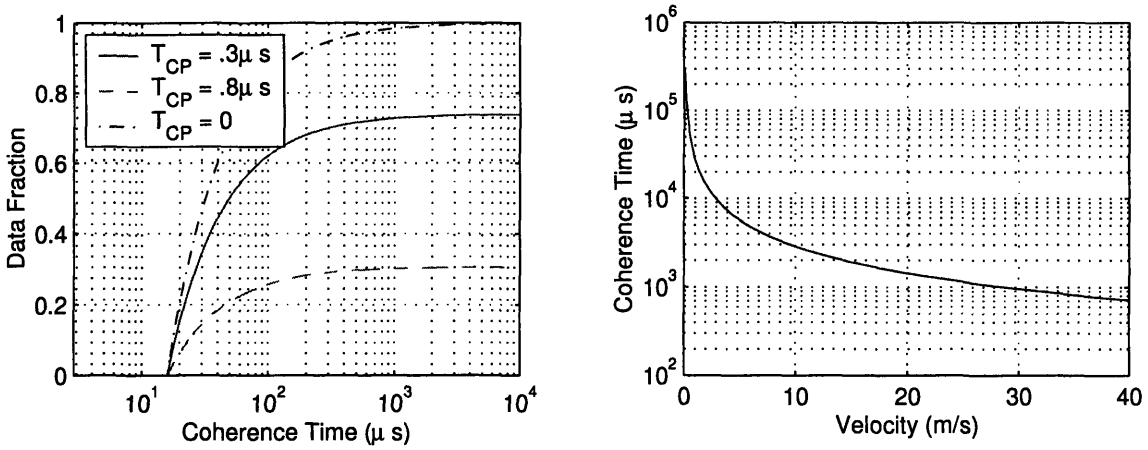


Figure 3-16: Coherence time vs. the fraction of data available data capacity (left), and how the coherence time varies as the relative velocity changes (right).

3.7 Design Guidelines

The cellular base station model for node communications is employed by the WiGLAN to take advantage of the resources of the central server. The uplink/downlink transmission protocol is designed to maximize throughput while offloading computational complexity onto the central server. Since each packet goes through only one uplink and one downlink, the packet delay is bounded and short. The server is able to provide a large amount of diversity gain for distant nodes, making uplink/downlink communications through the central server advantageous over ad hoc communications for the majority of node transmissions. Direct node to node communications is not ruled out completely, but its use is limited to those special cases (such as no delay requirements or distant, but closely spaced nodes) where the cellular model does not make sense.

In general a larger number of antennas at the server provides more flexibility in terms of allocation and allows the server to provide more diversity to aid node communications. Using at least one extra server antenna can greatly reduce the required SNR for both single- and multiple-antenna nodes during both the uplink and downlink.

In the context of the WiGLAN, single antenna nodes require using three antennas at the server to allows the node to transmit at 1 Gb/s using a moderate amount of coding with a (255,187) RS code.. For multiple-antenna nodes, it is possible to transmit uncoded at gigabit rates by employing four antennas at the node, plus four at the server. Using a (255,241) Reed-Solomon code allows requires three antennas

at the node and the server. Although four antennas at the server allow uncoded transmissions to reach gigabit rates, three antennas requires somewhat less hardware resources and is able to operate with only modest amounts of coding. For indoor use, the overhead for the WiGLAN packets for channel variations is around 25%. If outdoor or vehicular use is considered, however, this overhead increases significantly. The overhead caused by the length of the channel impulse response is the main source of overhead even at automotive speeds, however.

Chapter 4

Circuit Optimizations

As we have seen, multiple antennas at the transmitter and receiver promise greatly increased performance, both in terms of achievable capacity and the required SNR for uncoded BER performance. However, these arrays of antennas require extra digital hardware to handle the increased computation, and parallel analog RF chains to connect to the additional antennas. The larger required chip area increases cost and the additional power consumption reduces battery life. Using a separate chip for each analog front end would be prohibitively costly, especially as the number of antennas increases. In addition, approaching the data capacity of multiple antenna systems places similarly large demands on the digital processing circuitry, with powerful coding techniques requiring long data blocks and correspondingly large memory and computational resources.

Integrating the parallel RF chains onto a single chip saves chip area by allowing some sharing of common circuitry, but this alone does not sufficiently reduce the amount of area required. Because of the high capacity of multiple antenna systems, some excess system capacity can be traded off in favor of diversity to obtain large SNR gains. These gains in turn can be used for area as well as power savings for the analog and digital circuits.

In this chapter, we describe how diversity SNR gain available to multiple antenna systems can be used to relax circuit design constraints. Instead of only using this SNR gain to reduce transmit power or increase range as is the norm, we can use the SNR gain to decrease circuit area or power dissipation by using some knowledge of the analog hardware. Section 4.1 describes how some of the system capacity of a multiple antenna system can be sacrificed for diversity gain. The reduced SNR requirements can be viewed as an SNR gain which can enable circuits with poorer characteristics that are still within specifications, but reduce area and/or power consumption. A link budget analysis is given in Section 4.2 which details the factors that contribute to the SNR at the receiver input and how to calculate the available SNR margin.

Section 4.3 describes how circuit area and power consumption can be reduced for analog circuits. Because passive analog components (inductors, capacitors, and resistors) do not scale with improvements in process technology, significant area can be saved by removing the largest passive components. The SNR gain from multiple antennas offset the degradation in circuit performance from these alterations. The SNR gain can also allow the bias currents in the analog circuits to be reduced to lower power consumption, or allow the transmit power to be reduced.

In Section 4.4, we describe similar area and power reduction strategies for digital circuits. While digital circuits do not have SNR requirements, the reduction in computation and memory requirements for the high diversity system can reduce the amount of circuits required, reducing both area and power.

Section 4.5 applies these area and power savings ideas to the WiGLAN, focusing on the receiver RF front end. At the receiver, area savings allow four analog front ends to occupy only 50% more space than a single unmodified front end. Similarly, the power consumption can be reduced by more than a factor of two by reducing the bias currents at the receiver, or by reducing the transmitter power.

Some of the material presented in this chapter has been previously published in [28, 34].

4.1 Sacrificing Capacity for SNR Gain

Figure 4-1 (replicated from Figure 2-10) shows the bit error rate vs. SNR for an uncoded 1×1 and 4×4 system using 64-QAM and maximum diversity. The SNR is plotted as E_s/N_o, where E_s is the average energy of the transmitted constellation points, and N_o is the variance of the complex Gaussian noise. Note that the slope of the BER curves for the 4×4 system are much steeper than for the 1×1 system, which is a result of the greater diversity gain. Thus, the SNR gain for a 4×4 system over a 1×1 system depends on what error rate is desired, and increases without bound for low enough error rates. Assuming a packet size of 1000 bits, an error rate of 10^{-5} corresponds to approximately one percent of packets having an error. For a target bit error rate of 10^{-5}, the difference in SNR between the 1×1 and 4×4 systems is 40 dB, or a factor of 10,000. This 40 dB of SNR gain can be used to reduce the chip area and/or power consumption of both the digital and analog circuitry.

4.1.1 Using SNR Gain to Ease Circuit Requirements

Referring to Figure 2-5 and (2.4), the capacity of a 1×1 system at an SNR of 20 dB is 6 b/s/Hz. We can see from Figure 4-1 that the 4×4 system is able to achieve a BER of 10^{-5} for uncoded 64-QAM at almost the same SNR. Since 64-QAM encodes 6 bits

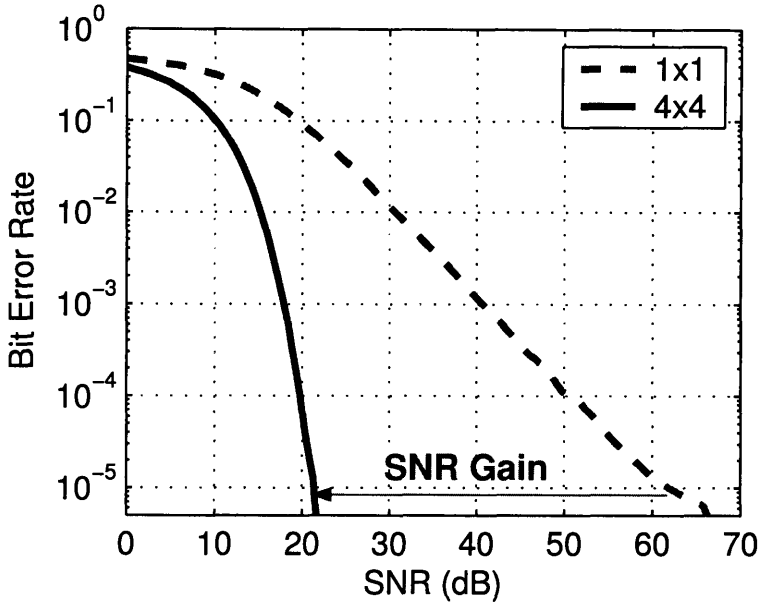


Figure 4-1: BER for a 1×1 vs. a 4×4 system for a 64-QAM input constellation.

per symbol, this will lead to the same overall data rate. The diversity gain allows an uncoded 4×4 system is able to achieve the same data rate as the capacity of a 1×1 system. The capacity of a 4×4 system at 20 dB is 22 b/s/Hz, however, so this SNR gain comes at the cost of 16 b/s/Hz in capacity. At first glance it appears the only benefit is complexity reduction in the encoding and decoding, which comes at the cost of additional hardware for the extra antennas. For the target bit error rate, however, an uncoded 1×1 system is not feasible due to the excessive SNR required.

For a constant transmit power, the receiver SNR is dependent on distance, so at closer ranges the 4×4 system no longer has to sacrifice as much capacity for diversity gain. Figure 4-2 plots the capacity as a function of distance assuming a transmit power of 100 mW and 150 MHz of system bandwidth. For the 1×1 system, every 3 dB SNR increase adds 1 b/s/Hz to the capacity. Similarly, for the 4×4 system, an increase of 6 dB allows the next higher constellation size to be used, which adds 2 b/s/Hz to the data rate. Thus it is possible to use the 4×4 system at maximum diversity with uncoded transmissions and keep up with the maximum achievable capacity of the 1×1 system. At closer ranges, the 4×4 system would always be able to sustain higher data rates at the same transmit power, as Figure 4-2 shows. Due to the larger capacity of the 4×4 system at all distances, trading off some capacity for diversity gain will not adversely affect the data rate compared to a 1×1 system.

The diversity gain of multiple antenna systems can be viewed as a low complexity

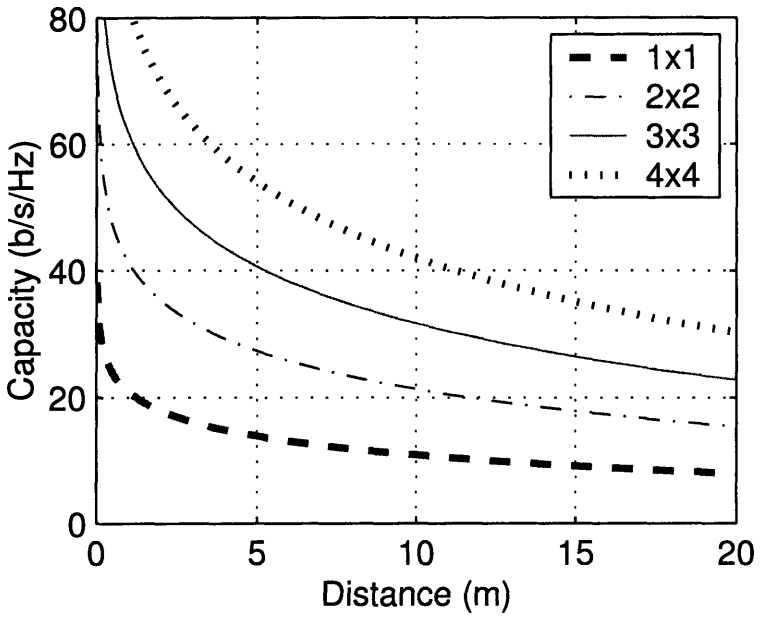


Figure 4-2: Achievable ranges of 1×1 to 4×4 systems.

mechanism for SNR gain. This allows different circuit architectures to be used for minimizing power or area, or both. By choosing a target range and bit error rate, the link budget is used to see what noise figure or other parameters are required for the circuits, and also how much SNR margin the diversity gain gives.

4.1.2 SNR Gain vs. Antennas

The amount of the SNR gain that can be realized is dependent on both the number of antennas as well as the amount of coding used. Figure 4-3 (replicated from Figure 3-10) shows the SNR gain in terms of both heavily and lightly coded or uncoded systems. The left plot shows the SNR gain compared to a 1×1 system for a heavily coded system that approaches the system capacity. For example, a 2×2 system at 6 b/s/Hz requires 10 dB less SNR than a 1×1 system, while a 4×4 system saves 16 dB. The SNR gains increase as more antennas are added, but the incremental gain for each additional antenna is small at the lower capacities in the regime which the WiGLAN operates. Because of the slope change of the capacity curves, each additional antenna will provide an arbitrarily large SNR gain for high enough capacities.

The right plots show the SNR gain for uncoded 64-QAM transmissions for both $1 \times N$ and $N \times N$ systems. For a BER of 10^{-5}, the SNR gain of a 2×2 system

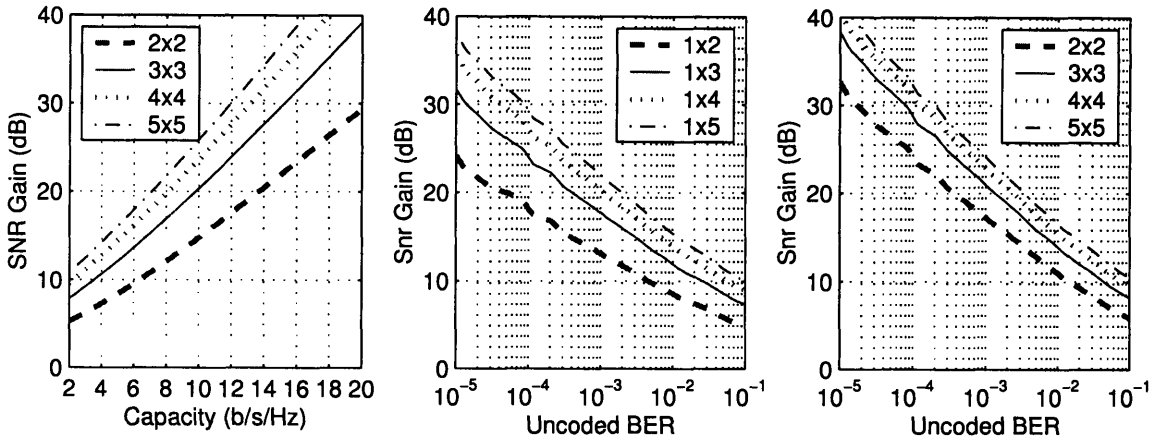


Figure 4-3: Achievable and uncoded SNR gain for 2×2 to 5×5 systems relative to a 1×1 system. Gains for single antenna nodes (middle) are also shown for comparison.

is over 32 dB, while a 4×4 system provides over 40 dB of SNR gain. As in the case for capacity, in the BER region of interest, the incremental gain shrinks for each additional antenna. Similar large gains can be had with the 1×2 and 1×4 systems, which provide 24 and 35 dB gain, respectively. As expected, these gains are not as large as with an $N \times N$ system, but are nevertheless appreciable.

The SNR gain is not as large for a lower raw BER with Reed-Solomon encoded transmissions, where the SNR gain at a raw BER of 10^{-3} is 17 dB for a 2×2 system, and 11 dB at 10^{-2}. A 4×4 system results in 23 dB and 16 dB gains for the same error rates. Doubling the number of antennas (resulting in an apparent doubling of area and power consumption) results in a gain of 6 dB in required SNR. The situation is very similar for $1 \times N$ systems. The biggest SNR gain comes from the first increase in antennas from a 1×1 to a 1×2 or 2×2 system, with diminishing returns making each additional antenna less beneficial than the previous one. In terms of absolute SNR gains, however, the extra antennas can push the SNR requirements low enough for the proceeding circuit area and power gains.

4.2 Link Budget Analysis

With a lower SNR requirement, more noise power is tolerated for a fixed signal strength. Alternately, less signal power is needed for a fixed noise power. For wireless circuit designers, this translates to an LNA with a more relaxed noise design at the input of the receiver, or reduced PA output power at the transmitter. A relaxed noise requirement for the receiver allows alternate circuit designs that can operate

at reduced power or occupy a smaller chip area. The effect of lowering the output power of the PA is not only reduced power consumption, but also increased linearity due to being able to operate farther from saturation. The link budget allows the system designer to summarize the impact of the overall SNR caused by the different components which the incoming signal passes through during processing. It is also a tool which allows the designer to trade off the performance of one component versus another given an SNR requirement of the overall system. We can use the link budget to determine the margin between the received SNR and the SNR required for the target system performance.

To determine the SNR margin for the receiver LNA, we look at the effective SNR at the input to the receiver decoder after processing by the analog front end. For the receiver, the input signal to noise ratio (SNR_i) is given by

$$\mathrm{SNR}_i = \frac{P_R}{N_i W}, \tag{4.1}$$

where P_R is the received power, N_i is the input noise power density and W is the noise equivalent bandwidth. The analog circuits add noise to the signal as it is processed, which degrades the SNR. The resulting output SNR_o in dB is then given as

$$\mathrm{SNR}_o = \mathrm{SNR}_i - NF, \tag{4.2}$$

where NF is the noise figure of the electronics. The noise figure captures the circuit noise contribution into a single number that represents the effective loss in SNR. While each individual component also has its own noise figure, the overall receiver noise figure is dominated by the noise figure of its first component. Typically, this component is the LNA. Although the horizontal axis of Figure 4-1 plots E_s/N_o versus BER, SNR_o can be related to E_s/N_o through

$$\mathrm{SNR}_o = \left(\frac{E_s}{N_o}\right)\left(\frac{R_s}{W}\right), \tag{4.3}$$

where R_s is the symbol rate (or the frequency of the time samples), and W is the system bandwidth. These two values are typically identical, making $\mathrm{SNR}_o = E_s/N_o$. The difference in the required SNR for the desired system performance and SNR_o limits the noise figure of the analog circuitry. Since the noise figure is one of the constraints of the analog front end, the larger this margin, the less constrained the design of the analog circuitry can be.

Table 4.1 shows a sample link budget comparing uncoded 1×1 and 4×4 systems. The center frequency is 5.25 GHz, with a system bandwidth of 150 MHz. The required noise figures are calculated assuming that a 64-QAM constellation should be

Determine Receive Power	
λ (m)	0.0571
Total TX power P_T (dBm)	20
d (m)	10
path exponent n	3
Path loss @ 10 m (dB)	76.8
Average RX power/ant P_R (dBm)	-56.8
Determine Max RX Noise Figure	
W (Hz)	1.50E+08
$N_i * W$ (dBm)	-92.2
SNR_i (dB)=$P_R/(N_iW)$	35.4
E_s/N_o 1×1 (dB)	61.0
E_s/N_o 4×4 (dB)	21.3
RX NF (dB) 1×1	-25.6
RX NF (dB) 4×4	14.1

Table 4.1: Sample link budget for a 1×1 and 4×4 system for uncoded 64-QAM at 10^{-5} BER.

decodable at a distance of ten meters with an uncoded BER of 10^{-5}. This distance is appropriate for a home or office environment. For this application, a path loss exponent of $n = 3$ corresponds to a Rayleigh fading channel and is a reasonable choice for an environment such as an open office space with no direct line of sight between the transmitter and receiver. The transmit power P_T is fixed at 100 mW, or 20 dBm, and the E_s/N_o for each system is taken from Figure 4-1. From Table 4.1, we see that the required noise figure at 10 m is 14.09 dB for a 4×4 system, and -25.91 dB for a 1×1 system. Note that the negative NF of the 1×1 system implies $\text{SNR}_o > \text{SNR}_i$, which is not achievable in a realizable system. For the 4×4 system, however, the rather large allowable noise figure allows unconventional circuit choices to be made which can reduce the circuit area or power consumption.

4.3 Analog Circuit Optimizations

The sample link budget analysis shows that there is a margin of 14 dB between the received SNR and the SNR required for good uncoded error performance. Although the actual SNR margin is dependent on the particular application, this sample SNR margin can give an approximate guideline for how large a circuit noise figure can be. The excess SNR can be utilized to optimize in several different directions, as shown in Figure 4-4. At the transmitter, since PA efficiency is a major concern, the SNR

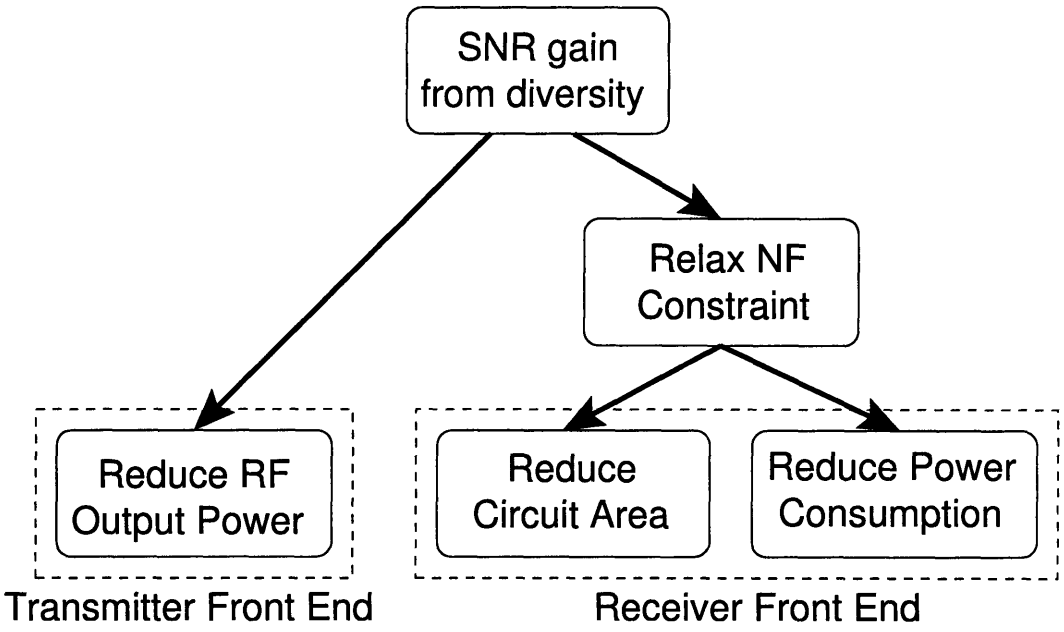


Figure 4-4: Roadmap for analog circuit optimizations.

margin allows a reduction of the RF transmit power. This leads to an even larger reduction in the total power consumption at the transmitter. At the receiver, the SNR margin can translate into a relaxation of the noise figure specifications, allowing the LNA to reduce its power consumption or circuit area, or both.

Another method to use the SNR gain is to increase the range of the network. As Figure 3-14 shows, the increased diversity of a $1 \times N$ system increases the distance that can achieve a given BER for a fixed transmit power. A large increase in range could be useful in an outdoor setting which has a large network coverage area, but the indoor environment typically has a smaller required range, so the SNR margin is better spent on area and power reductions.

4.3.1 Reducing RF Output Power

The power amplifier consumes the majority of the DC power consumed by the transmitter circuitry. Since the receiver input SNR is directly related to the RF transmit power, the transmitter RF output can be lowered by the SNR margin minus the noise figure of the receiver circuits. For the sample link budget in Table 4.1, the SNR margin for a 4×4 system is 14 dB, and we will temporarily assume the receiver noise figure is 0 dB. From the link budget analysis the original amplifier output is 20 dBm or 100 mW. If the amplifier is 25% efficient (optimistic given a theoretical maximum

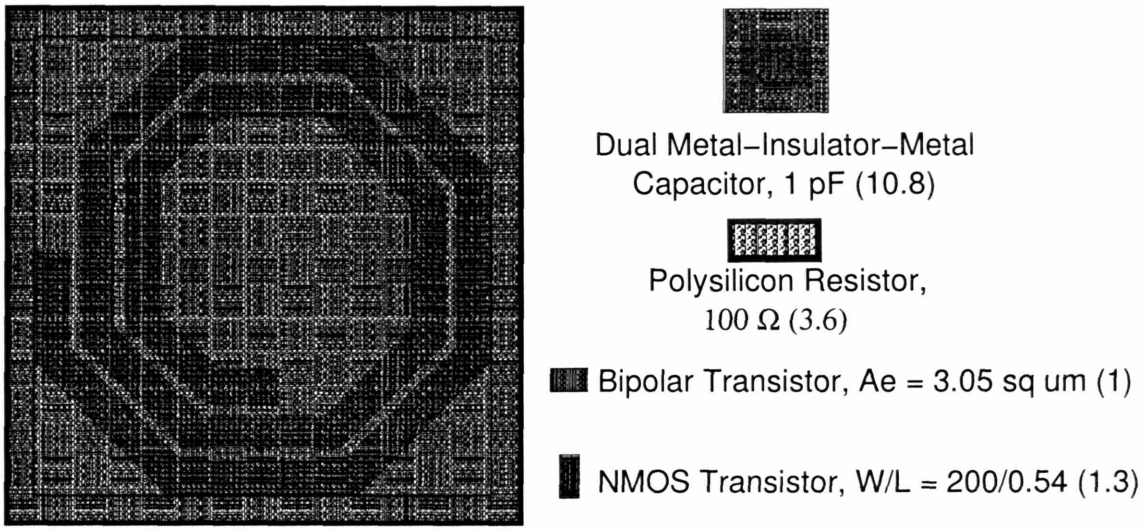


Figure 4-5: Sizes of typical on-chip components (in parentheses) relative to the bipolar transistor.

efficiency of 50% for a class A amplifier), then 400 mW of power is drawn to transmit 100 mW. Using the SNR margin to lower the transmit power reduces the output power to 6 dBm (4 mW). This translates to a power consumption of 16 mW, which is a significant improvement (14 dB, as expected, though less in practical terms due to the receiver noise figure).

4.3.2 Reducing RX Circuit Area

Multiple antenna systems require an individual analog RF receiver per antenna, suggesting that the area increases proportionally with the number of antennas. Thus, for example, a receiver with four antennas consumes four times the area of a single antenna receiver. As area translates directly to chip fabrication cost, larger die area results in fewer chips per wafer. Although decreasing feature sizes of each generation of process technology allows digital circuits to shrink, on-chip passive components (inductors, capacitors, and resistors) do not scale with technology. As a result, these components become major stumbling blocks to shrinking the physical size of circuits for wireless systems. Figure 4-5 compares the chip area required for the different types of components, sized to the values that would be typical for a 5 GHz LNA fabricated in a standard 0.18 μm SiGe BiCMOS.

Clearly, inductors will dominate the size of any analog circuit that uses it. Com-

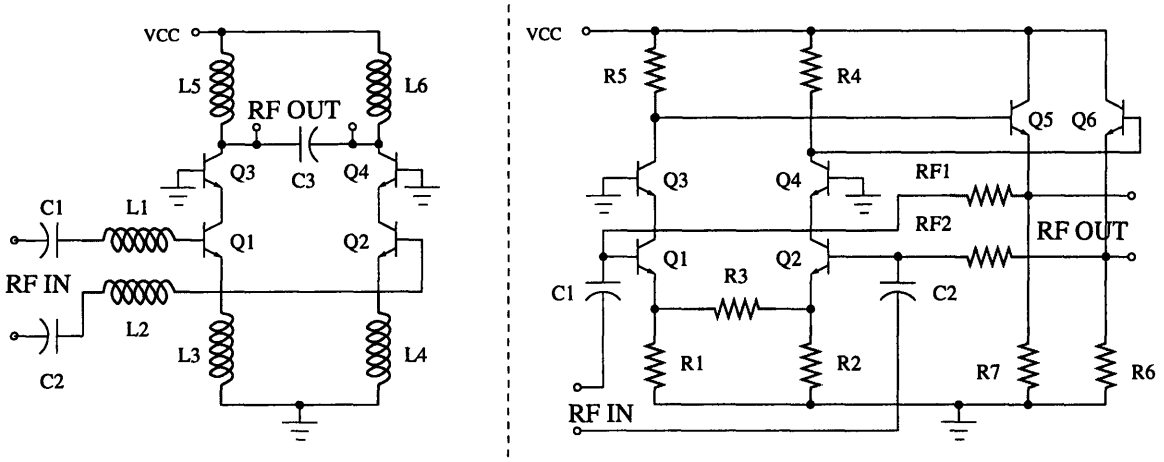


Figure 4-6: Simplified circuit comparison for narrowband (left) and broadband (right) LNA (biasing not shown). The inductors in the narrowband LNA are replaced with physically smaller resistors and transistors. (Diagram by L. Khuon)

paring their relative sizes, a single inductor is over 300 times larger than a bipolar transistor, and 250 times larger than an MOS transistor. Although inductors are often a necessary component in circuit designs, reducing their number could allow a significant decrease in area. Depending on their values, capacitors can also take up a significant fraction of the area. In comparison, transistors and resistors are relatively small and take up much less area. Therefore, an area-efficient design approach reduces the use of inductors (and to a lesser extent capacitors) in favor of resistors and transistors.

Figure 4-6 shows simplified circuit diagrams for two different LNA topologies. The narrowband LNA on the left uses inductors, capacitors, and transistors. As six inductors are required per LNA, that totals to 24 inductors for the four antenna receiver of a 4×4 system. The broadband LNA on the right, however, does not use any inductors, replacing them with resistors and transistors.

As stated in Section 4.2, the LNA dominates the noise figure of the receiver chain. Noise figure is therefore one of the biggest constraints on LNA design. A well-designed bipolar narrowband LNA will have a noise figure under 2 dB, but requires significant chip area due to the usage of on-chip inductors. For example, a narrowband cascode LNA such as the left circuit shown in Figure 4-6 requires six inductors (and several capacitors) per front end, for a total of 24 inductors for each end of a 4×4 system. The capacitors and inductors that connect the input to Q1 and Q2, form a resonant circuit, which is tuned to the desired frequency band. Since the bandwidth of the resonant circuit is relatively narrow, the equivalent bandwidth for the noise is not

much bigger than the signal bandwidth.

The broadband LNA, on the other hand, uses no inductors, replacing them with resistors. As the right circuit of Figure 4-6 shows, the topology of the amplifier has also changed, with the addition of several extra resistors plus two additional transistors for an output buffer. Although the broadband LNA requires several more components than the narrowband LNA, the total circuit area is reduced because a single inductor occupies the same area as hundreds of resistors and transistors. The broadband LNA is therefore significantly smaller in size than the narrowband LNA. Instead of a narrowband resonant circuit, the input of the broadband LNA only has a DC blocking capacitor, resulting in an equivalent noise bandwidth at the amplifier input that is much larger than the signal bandwidth. This leads to a degradation in received SNR figure for the broadband LNA. The resistors also add some thermal noise, which can further degrade the noise figure. However, given the large SNR margin that can be achieved with a high diversity system, the added noise can be a good trade off for area savings. The output buffer (and to a lesser extent, the resistors) additionally dissipate power while the inductors they replaced do not, raising the total power consumption of the broadband LNA compared to the narrowband LNA.

Considering the 4×4 case from the sample link budget in Table 4.1, and assuming that the receiver noise figure is dominated by the LNA, a margin of about 14 dB exists for the LNA. Some of this margin can be traded off if the narrowband LNA is replaced with a broadband amplifier. As the next section shows, however, if power reduction is the goal, then this choice is reversed.

4.3.3 Reducing RX Circuit Power Dissipation

Rather than minimizing area, the relaxed noise figure requirement can be used to lower the DC power consumption of the analog front ends. The resistors and output buffer of the broadband LNA dissipate power while the inductors they replaced in the narrowband LNA comparatively do not. The output buffer consumes a significant amount of power, but the resistors shown consume a relatively small fraction of the total power. With the narrowband LNA, the only components that will dissipate power are the transistors and biasing resistors, which cannot be removed as they are responsible for the amplifier gain. Thus, to reduce power consumption we should first start with a narrowband rather than a broadband LNA.

Reducing the bias currents in an LNA lowers the DC power dissipation, but also reduces its gain. A lower gain increases the noise figure as the contribution from the noise sources at the LNA output increases. As shown in Figure 4-7 for an example narrowband LNA, as the bias current is lowered from its original value I_o, the gain falls while the noise figure rises. At $0.2I_o$, the noise figure degrades by approximately 1 dB while the gain drops to 4 dB. Further reducing I_o eventually drops gain to 0 dB

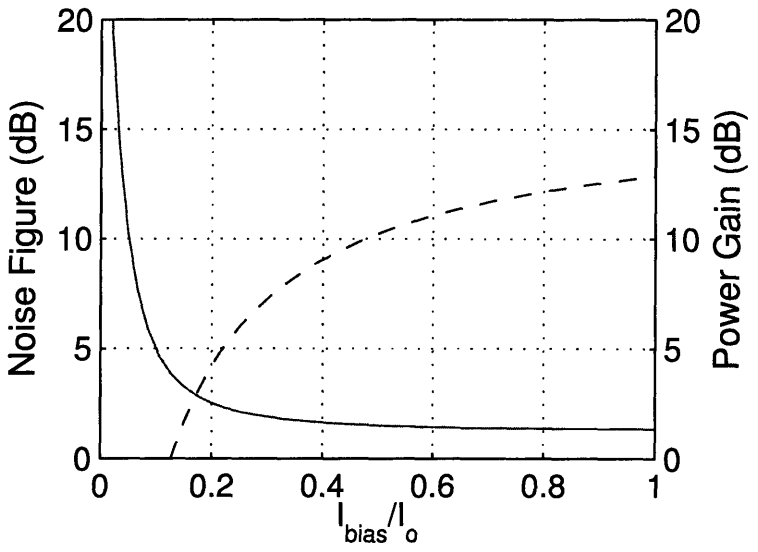


Figure 4-7: Example noise figure (solid) and LNA power gain (dashed) vs. bias current.

and raises the noise figure to 3.8 dB. The noise figure rises precipitously as the bias current is dropped further, but once the gain has reached 0 dB or unity, there is no amplification of the signal. At this point, then, the LNA is no longer useful and can potentially be removed.

Not shown in Figure 4-6 are the biasing circuits, which are present in both the broadband and narrowband cases, and can include both transistors and resistors. Although resistors replaced inductors in the broadband LNA, their power dissipation is small compared to the biasing networks and the output buffer. The output buffer in the broadband LNA is not critical to the operation of the LNA, but rather used to facilitate its connection to other circuits. The narrowband LNA might also need an output buffer, increasing its power usage significantly, and narrowing the difference between the two amplifier types. Just as with the narrowband LNA, the broadband LNA can also be used to reduce power by lowering bias currents, but its higher noise figure puts it at a disadvantage when compared to the narrowband LNA.

4.4 Digital Circuit Implications

Unlike analog circuits, digital circuits do not really have an SNR requirement, and noise immunity is one of the main features of digital circuits. Thus the SNR margin created by the antenna diversity does not directly affect the circuit area or power consumption for the digital processor. In addition, since digital circuits typically

consist only of transistors, their size shrinks as advancing process technology reduces the minimum feature size. Because digital circuit power and area consumption both increase with processing complexity, low-complexity algorithms will be able to reduce both simultaneously.

4.4.1 Reducing Digital Circuit Area

Comparing the sizes of an analog front end to a digital processor shows that the digital processor often takes up at least as much space as the analog components. For example, an integrated 802.11g system on a chip (SoC) was implemented in [43]. The die photo shows that the digital processor for the Physical and MAC layers occupy over two-thirds of the chip area, with the remainder being the analog front ends for both the transmit and receive RF chains. Thus an area savings for the digital circuits could have a larger impact than similar percentage savings in the analog circuitry.

Although digital circuits do not have SNR constraints, they also benefit from the SNR gain of the high diversity system. A heavily coded system, like one that would be used for a 1×1 system, would require long block lengths and complex encoding and/or decoding of the data streams. The long block lengths require the transmitter and receiver digital processors to queue up enough data to fill up a block before processing. This can require a significant amount of memory to hold the data blocks, plus several times more memory to allow enough processing space to encode and decode the blocks. For example, turbo codes and low-density parity check (LDPC) codes, which are both very powerful techniques that can approach the system capacity, often require block lengths in the thousands to millions of bits to be effective. An example of a turbo decoder chip is presented in [62], with a labeled die photo showing that about half of the chip area is reserved just for memory. Reducing the block sizes reduces the SNR gain of the error correction code, but this can be offset by the diversity SNR gains. In addition, the decoding algorithms for turbo codes and LDPCs require many iterations to converge to an answer, which requires a capable processing unit to allow decoding in real time applications.

As we have seen, the 4×4 system with maximum diversity and uncoded data streams has similar error performance as a heavily coded 1×1 system. The advantage of the high diversity system is that it requires significantly less processing power and memory at both the transmitter and the receiver than a powerful encoding scheme. For example, a relative processing requirement shows that 70–90% of the total computational complexity can be occupied by the coding for a heavily coded system [18].

The uncoded high diversity system only requires learning the channel parameters, plus one SVD per packet at the transmitter and at the receiver, which in this case is equivalent to diagonalizing a 4×4 matrix. Very little memory or computation is

required for the decoding because the MRC is symbol by symbol. Instead of queuing up a large number of information bits to encode or received signal samples to decode, only a single symbol needs to be manipulated or stored in memory at any given time. Less memory and simpler processing translates to a smaller digital processor, which can lead to significant area savings.

4.4.2 Reducing Digital Circuit Power

The low-complexity diversity system similarly results in lower power consumption for the digital processor. For the digital processor, the power dissipated P and switching time T per CMOS transistor is (repeated from (2.17) and (2.18)) [7]

$$P = p_t \cdot C_L \cdot V_{dd}^2 \cdot f_{clk} \tag{4.4}$$

$$T \propto 1/V_{dd}, \tag{4.5}$$

where p_t is the activity factor, C_L is the load capacitance (usually the input gates of other digital logic circuits), V_{dd} is the power rail, and f_{clk} is the switching frequency. In order to reduce the power consumption, either the activity factor, load capacitance, power rail, or clock frequency can be reduced. The load capacitance is a function of the circuit layout and the process technology, and is not easily reduced. Lowering the clock rate will reduce the power consumption, but at the same time reduces the amount of computation that can be done per unit time. This loss can be offset by the reduced complexity of the high diversity system with either uncoded communications or the Reed-Solomon coding. Power reduction can be easily made by reducing the activity factor. Since coding can account for up to 90% of the total computational complexity [18], using uncoded transmissions can reduce the activity factor by the same amount, lowering digital power consumption by a factor of ten due to the reduced computational requirement.

Because of the squared term, lowering the voltage results in a big gain in power consumption, but it also slows down the transistor switching time. [7]. Slower switching time requires reducing the clock frequency, which will additionally reduce the power consumption. Thus, the reduced complexity of the diversity system can greatly reduce the power consumption of the digital processor by reducing the number of transistors, their operating frequency, and the system voltage. The only variable that is not easily changed is the gate capacitance, but those will shrink as feature size shrinks with new processing technology generations.

	Uncoded			RS (255,241)			RS (255,187)	
	1000	900	540	1000	900	540	733	540
1×1	(-33.2)	(-31.0)	(-24.6)	(-14.7)	(-11.7)	(-4.5)	(-2.9)	2.6
2×2	(-1.7)	0.7	8.0	2.3	5.3	13.2	7.9	13.6
3×3	3.8	6.2	13.8	6.0	9.0	16.9	10.8	16.5
4×4	6.2	8.7	16.1	7.8	10.9	18.8	12.4	18.1
5×5	7.7	10.1	17.7	9.1	12.1	20.1	13.5	19.2
1×2	(-10.6)	(-8.2)	(-1.0)	(-1.9)	1.2	8.8	5.6	11.2
1×3	(-2.5)	(-0.1)	7.1	2.8	5.8	13.6	9.0	14.6
1×4	0.9	3.4	10.9	5.3	8.3	16.2	11.0	16.6
1×5	3.7	6.1	13.6	7.0	10.1	17.9	12.3	18.0

Table 4.2: SNR margin in dB for adaptive modulation in the WiGLAN assuming $SNR_i = 35.4$ dB. Uncoded numbers are for rates in Mb/s for 10^{-5} BER. All indicated rates are normalized to show actual information rates.

4.5 WiGLAN Analog Circuit Optimization

We can apply these area and power optimization ideas for the case of the WiGLAN. Because the central server is involved in all aspects of the network communication, it has significant resources and capabilities. The added power and circuit area cost for multiple antennas at the server is easily justified to allow high SNR gains from diversity. For the nodes, however, the area and power cost can be significant due to the desire to use inexpensive radios and battery power. The SNR gains and potential increase in data rate from multiplexing motivate the desire for nodes to have multiple antennas. This SNR gain can then be used to reduce the circuit area or power consumption. Because the WiGLAN is intended for an indoor environment, it is more important to reduce circuit area and power than increase communication range.

4.5.1 SNR Margin from Link Budget

Using the example link budget in Table 4.1, we can look at the SNR margin for different antenna configurations and data rates. The example link budget uses the same parameters as the WiGLAN, so we can see that the received SNR is 35.4 dB. Table 4.2 compares the SNR margins for several different average data rates in Mb/s for the WiGLAN using adaptive modulation. The data rates are normalized so that, for example, 1000 Mb/s represents the same overall throughput regardless of the encoding scheme used. The antenna combinations that have an SNR margin near or less than zero are unusable as the RF front end will have a nonzero noise figure.

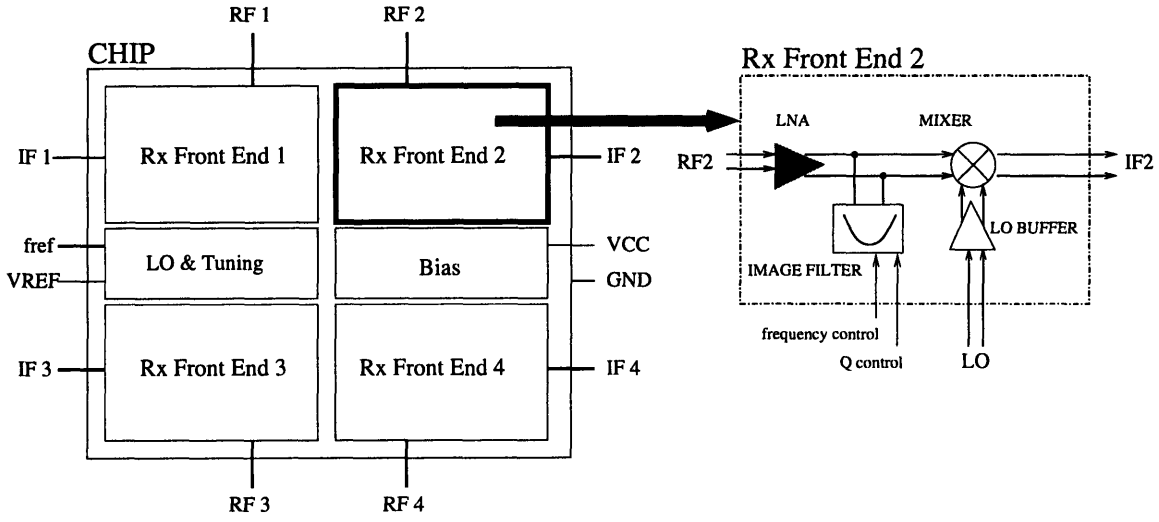


Figure 4-8: Schematic for WiGLAN integrated receiver test chip with four RF front ends. The LNA (highlighted) is the target for area and power optimizations. (Diagram by L. Khuon)

4.5.2 Reducing Circuit Area

For multiple antenna nodes, integrating the parallel RF chains onto a single chip allow many common circuits to be shared, since only the front end of each RF chain needs to be independent. As shown in Figure 4-8 for the WiGLAN receiver, each front end per antenna includes an LNA, image reject filter and mixer. Each front end, shares the local oscillator, bias circuits, and filter tuning circuits. To further reduce the circuit area, the LNAs can be switched from the more traditional narrowband to the broadband circuit topology.

To quantify the comparison, LNAs were designed for the WiGLAN with both broadband and narrowband approaches [34]. Table 4.3 compares the simulated size and noise figure for the two LNA topologies sized for the WiGLAN. While the broadband approach has a simulated noise figure 3 dB higher than the narrowband approach, it occupies roughly one-third the area. By incorporating the broadband-based LNA into the design of four receiver front ends on a single chip, a simulated noise figure of 8.8 dB and an estimated area of 4 mm^2 results.

Table 4.4 compares the estimated area and noise figure of an integrated parallel quad receiver front end for the WiGLAN using both the narrowband and broadband LNA designs. Also included are the areas and noise figures of several 5 GHz single receiver front ends found in the literature. As expected, the WiGLAN receiver front ends exhibit higher noise figure but occupy an area less than some published single

	Area (sq. mm)	NF (dB)
Narrowband	0.6	1.4
Broadband	0.23	4.4

Table 4.3: Comparison of a single WiGLAN narrowband and broadband LNA.

Reference	Area (sq. mm)	NF (dB)
[56]	1	5.2
[79]	3	8.5
WiGLAN (bb)	4	8.8
WiGLAN (nb)	5.5*	5–6 (est)
[74]	6*	5.9
[72]	12.5*	5.2
[81]	14*	8.0

Table 4.4: Area and noise comparison for 5 GHz single antenna receiver front ends with both broadband and narrowband four antenna WiGLAN receiver front ends. (* approximate receiver area for a combined transceiver chip).

front ends. It should be noted that although the circuit in [56] appears to just as small and has a lower noise figure, that design relies on custom inductors that were hand optimized to take up as little space as possible in the circuit. In contrast, the WiGLAN amplifier design uses the standard inductors shown in Figure 4-5, which occupy considerably more space. Note that the narrowband receiver circuitry was not laid out for fabrication, so the estimated receiver area and noise figure is only approximate.

After simulation, the test chip was fabricated to verify the design parameters [35]. Figure 4-9 is a die photo of the test chip, which was manufactured using a standard 0.18 μm SiGe process, with an active area of about 4 mm^2. The active area does not include the bond pads around the edges of the chip, which are used to connect the chip to other chips or to the packaging pins. The inductors visible at the top and bottom of the chip are for the image reject filters, and the boxed area is one of the broadband LNAs. Note that a single LNA does not occupy significantly more area than a single inductor. The other circuitry in the chip correspond to the ones shown in the schematic in Figure 4-8. The broadband approach has about a 3 dB higher noise figure than the narrowband design, but four receiver front ends using broadband LNAs are only 1.5 times larger in area than a single receiver front end with narrowband circuits.

Now that the noise figure of the broadband LNA is known, we can return to Table 4.2 to determine how many antennas are needed for gigabit data rates at a

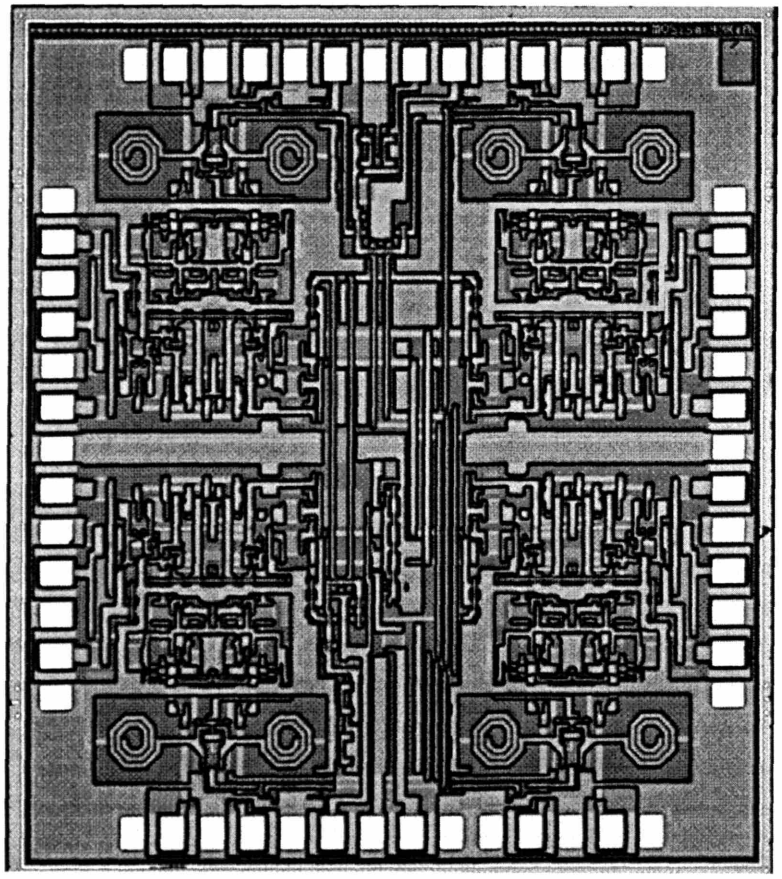


Figure 4-9: Die photo of WiGLAN receiver chip with four RF front ends. The boxed region is the area occupied by a single broadband LNA. (Chip design by L. Khuon)

range of 10 m. For a 4×4 system, the allowed RX noise figure is 6.2 dB for uncoded transmissions, which is enough for the narrowband but not the broadband design. Using the weaker (255,241) RS code, the broadband approach requires five antennas, while the narrowband approach only requires three antennas. It is not possible to use the (255,187) RS code to reach gigabit data rates because the maximum throughput is limited to 1.2 Gb/s due to the maximum constellation size of 256-QAM. However, a raw data rate of 1 Gb/s translates to 733 Mb/s and allows the broadband approach to use only three antennas, but the narrowband approach requires only two. If we instead targeted a near-gigabit rate of 900 Mb/s, the 4×4 system is nearly able to support the broadband approach for an uncoded system, whereas a lightly coded system with the (255,241) RS code is able to use only a 3×3 system.

The reduced size of the broadband LNAs can allow extra antennas to be added without necessarily increasing circuit area, however. As Table 4.3 shows, two broadband LNAs require less area than a single narrowband LNA. For the full receiver front end, this area gain allows the broadband approach to occupy only 50% more are than the narrowband design, but with four times as many antennas. Having the extra antennas is advantageous from a capacity standpoint, so the reduced size of the broadband approach can achieve a higher capacity than the narrowband for a fixed circuit area. Because the parallel front ends have some shared circuits, the difference in area for other numbers of antennas is not completely clear.

If it is not possible to have multiple antennas at the transmitter and the receiver, then the SNR gain decreases somewhat, but it is still possible to achieve large SNR margins due to the multiple antennas at the central server. For example, a single antenna node can use seven server antennas with the (255,241) Reed-Solomon code to achieve gigabit data rates with the narrowband design, or 900 Mb/s with the broadband design. Even though the LNAs are only one component of the front ends, altering their circuit topology results in significant changes in circuit area.

It is not straightforward to design an LNA that has a desired noise figure, as the changes in NF come mainly from topology changes rather than a smooth variation of circuit parameters. This leads to threshold behavior in the allowed antenna configurations. Any surplus SNR gain can be used to reduce the circuit power, as in the next section, or to reduce the transmit power, both decreasing the overall circuit power consumption.

4.5.3 Reducing Power Consumption

For overall power consumption, the most effective reduction occurs at the transmitter. The SNR margins from Table 4.2 translate directly into power reductions at the transmitter power amplifier, after the receiver noise figure has been accounted for. For example, if we assume the receiver noise figure is 0 dB, the 3×3 system has an

SNR margin of 3.8 dB for uncoded communication at an average data rate of 1 Gb/s. This margin means the transmitter output power can be reduced by 3.8 dB. For 20 dBm output power, that translates into a reduction from 100 mW to 42 mW of transmit power. If the transmitter PA was 50% efficient, then the power dissipation of the PA is reduced from 200 mW to 83 mW. Although the multiplicative factor is the same, the actual power reduction level is significant. If, for example, the efficiency was only 25%, then the PA power dissipation is reduced from 400 mW to 167 mW, which is an much larger reduction in terms of actual power. Since the output power can in principle be finely controlled, all of the SNR margin can be used to reduce the power dissipation at the transmitter. This is most effective during the uplink, as the node transmissions are more power constrained than the server transmissions.

At the receiver, the power consumption can be lowered by reducing the bias currents of the narrowband LNA as in Figure 4-7, which shows that even at 20% of the original bias current the noise figure is less than 3 dB. However, it does become necessary to amplify the input signal at some point before decoding, so there must be amplification somewhere in the receive chain. Usually this happens right away in the LNA, as the first component typically dominates the noise figure of the entire system. If the LNA has low gain, then the noise figures from the following components can significantly increase the overall noise figure of the front end. Circuit designers therefore normally keep the gain above 10 dB to ensure that there is sufficient LNA gain to dominate the noise performance of the receiver. We can see from Figure 4-7 that the narrowband LNA at 10 dB gain could be operated at half the power (due to reducing the bias current by half) with a slight increase of 1–2 dB in noise figure. As a comparison to the broadband LNA, its output buffer consumes almost as much power as the narrowband LNA, for a total power consumption four times higher than the narrowband LNA with the reduced bias current. Just as with the transmitter power reduction, the bias current can in principle be reduced to use up all of the SNR margin, but the gain of the LNA limits how far the current can reasonably be decreased.

4.6 Design Guidelines

The SNR gain from a high diversity system can significantly reduce the SNR required at the receiver of two communicating devices. The SNR gain for a 4×4 system can reach 40 dB or more when compared to a 1×1 system. This SNR gain can be used to reduce the circuit area of the receiver, or the power consumption at either the transmitter or receiver. The largest benefit in power consumption comes from reducing the transmit power due to the high power levels and low amplifier efficiency. A 10 dB extra margin in noise figure can be directly used to reduce the transmit

power by 10 dB.

For the WiGLAN, the SNR margin is sufficiently large for an uncoded 4×4 system, or a 3×3 system with the (255,241) RS code, to achieve gigabit data rates. For single antenna nodes, gigabit data rates are still possible with the (255,241) RS code if the server has four or more antennas. Reducing the bias currents to reduce power at the receiver is less effective because the LNA needs a minimum amount of gain to keep the overall noise figure low, only saving about 3 dB in power compared to a single antenna system.

While power consumption is the major concern for the transmitter, the lack of a power amplifier at the receiver makes the circuit area more important. Removing the large passive inductors will significantly reduce the area of the LNA, which can reduce the total front end chip area by more than a factor of two. A four antenna receiver broadband front end is only 50% larger than a single antenna narrowband front end, compared to an expected quadrupling of circuit area. The small additional area for the multiple antenna design is offset by the added flexibility and data capacity achievable with multiple antenna systems.

In the WiGLAN, the number of antennas required depends on the desired operating point. To allow uncoded transmissions at 1 Gb/s, more than a 5×5 system is needed. Lowering the throughput slightly to 900 Mb/s allows a 4×4 system to be used. With the (255,241) RS code, gigabit data rates can be achieved with a 5×5 system, however, or 900 Mb/s with only a 3×3 system. Although four and five antenna front ends are larger, the three antenna broadband front end should compare very favorably in size with the single narrowband front end. A single antenna node is capable of 900 Mb/s communication with coding if the server has four or more antennas.

Since digital circuits scale very well with improvements in technology, these area and power savings are not as important as in the analog circuitry. However, the use of uncoded communications can greatly reduce the memory and computational requirements of the digital circuits, allowing both their area and power to be reduced. Given the SNR margins available to the WiGLAN, four server antennas is the minimum required to allow even single antenna nodes to communicate at or near gigabit data rates at a distance of ten meters.

Chapter 5

Circuit Crosstalk

For multiple antenna systems, the MIMO processing algorithms of the parallel data streams at both the transmitter and receiver typically assume there is no crosstalk aside from the interference caused by the multipath channel. Correlation between the antenna elements has been studied in some detail, but the circuitry is still assumed to have no crosstalk.

When multiple RF chains for antenna arrays are located on a single chip, the potential exists for significant amounts of crosstalk between the otherwise separate signal chains. Careful circuit design and layout can reduce the signal leakage, but may come at the expense of chip area or effort on the part of the designer. Since larger chip area translates directly into higher cost, it is often an important consideration for a circuit design. It is advantageous for the circuit components and interconnect wires to be as closely spaced as possible, but this will increase the crosstalk coupling between independent circuit paths. The crosstalk between two devices can be decreased in magnitude by increasing physical separation, which increases circuit area. Isolation structures, such as insulating wells, guard rings, trenches, or even Faraday cages can also be placed in the space between the elements, increasing circuit isolation by 30 dB or greater [80]. These additional structures can lead to significant increases in chip area if they are used repeatedly in a circuit, and also may not be compatible with the particular processing technology that is being used for the circuit.

In this chapter, we define two crosstalk models to study the effect of crosstalk between parallel circuits in terms of the effective loss in SNR. We derive a relation for the asymptotic loss in SNR due to crosstalk in terms of the singular values of the crosstalk matrix, and propose using phase randomization to increase the "diversity" of the crosstalk and improve the SNR loss in the worst case. Section 5.1 analyzes the effect of crosstalk on two parallel amplifier chains. We determine that the crosstalk between the input and output ports form a feedback loop, which can become unstable due to positive feedback or form deep nulls in the frequency response. The crosstalk

isolation should be at least as large as the power gain of the amplifiers to prevent one signal from seriously interfering with another signal or itself.

Section 5.2 describes a more detailed linear model for the crosstalk. A general expression for the asymptotic SNR loss caused by the crosstalk is derived in Section 5.3 in terms of the singular values. Randomization of the crosstalk phase reduces the strong phase dependence of the crosstalk SNR loss when the isolation poor, allowing all circuit realizations to experience the same circuit performance. For isolations greater than 20 dB, the SNR loss is small and phase randomization will not help. By improving the worst case performance, phase randomization potentially allows circuits to be designed with less attention to crosstalk isolation and the associated isolation structures.

While Section 5.3 focuses on a 2×2 crosstalk matrix, Section 5.4 extends this to look at the behavior of larger systems. Section 5.5 presents several ideas on how to physically achieve phase randomization, with frequency hopping being the most viable. Finally, crosstalk measurements from a test chip are presented in Section 5.6, and simulations of the SNR loss from crosstalk are presented in Section 5.7.

5.1 Feedback Crosstalk Model

For the central server or a node with multiple antennas, each antenna requires a separate analog front end. As was discussed in Chapter 4, integrating these circuits onto a single chip can have advantages in area and power savings. Since all the circuits are on the same chip, however, crosstalk between circuit paths can create unwanted interference. At the power amplifier (PA) of the transmitter, the signal power is by far the largest compared to any other signal in the circuit. This large signal can couple into any other circuit path on the chip, adding unwanted noise to that signal. If there are multiple amplifiers, the output signal of one amplifier may couple into the input of another amplifier. This very large signal from an amplifier output can overwhelm the desired signal at the input of another amplifier.

5.1.1 Crosstalk Feedback Loop between Multiple Front Ends

Figure 5-1 shows a pair of amplifiers that have coupling between their inputs and outputs. Coupling between the inputs only or outputs only are ignored, although they can exist as well. The coupling between an amplifier output and the input of the other amplifier has the most significant effect due to the large difference in signal powers between the input and output ports of an amplifier. It is clear that under certain conditions it is possible to have an unstable system due to positive feedback.

The gain of each amplifier and the coupling magnitude between the amplifiers is

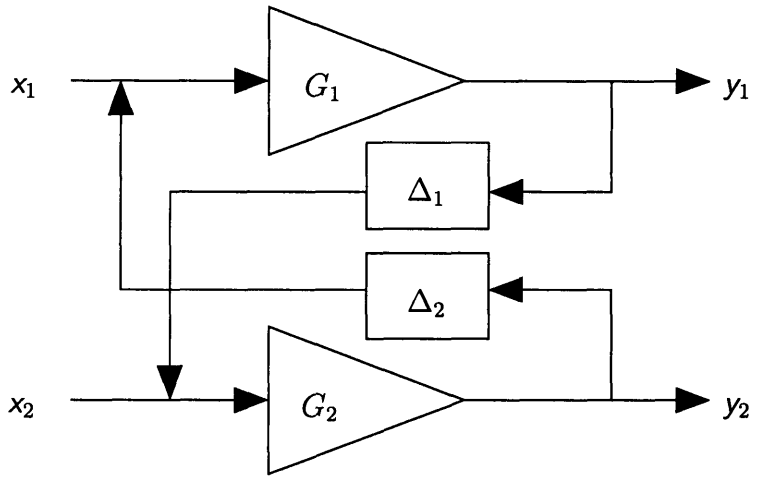


Figure 5-1: Crosstalk coupling model for two parallel amplifiers.

represented by g_i and δ_i, respectively. Both of these terms have an associated phase and magnitude, with

$$g_i = G_i e^{j\phi_i} \tag{5.1}$$
$$\delta_i = \Delta_i e^{j\theta_i}, \tag{5.2}$$

where G_i and Δ_i are positive real constants. It is assumed that the amplifier designer is able to control the gain magnitude G_i, but does not have significant control over the gain phase ϕ_i. We also assume that the crosstalk magnitude Δ_i is fixed by the circuit design, and the crosstalk phase θ_i is a fixed random constant that is dependent on the particular instance of that circuit. When a chip is fabricated, variations in processing conditions will lead to slight changes in the behavior of the circuit. These variations have little effect on the magnitude of the signal coupling between circuits, but may show a noticeable difference in the phase of the coupled signal. The crosstalk parameters could slowly vary as the temperature changes or the circuit ages, but these timescales are so long compared to the data transmission times that the crosstalk can be assumed to be constant. There are no noise sources modeled because the signal levels are typically far above the thermal noise levels, making crosstalk coupling the major source of interference. As Table 4.1 shows, the SNR at the transmitter for 20 dBm transmit power is over 110 dB.

The output of the two amplifiers can be written as

$$y_1 = g_1 x_1 + \delta_2 g_1 y_2 \tag{5.3}$$
$$y_2 = g_2 x_2 + \delta_1 g_2 y_1, \tag{5.4}$$

which after simplifying yields

$$\begin{aligned} y_1 &= g_1 x_1 + \delta_2 g_1 (g_2 x_2 + \delta_1 g_2 y_1) \\ &= g_1 x_1 + \delta_2 g_1 g_2 x_2 + \delta_1 \delta_2 g_1 g_2 y_1 \\ &\downarrow \\ y_1 &= \frac{g_1 x_1 + \delta_2 g_1 g_2 x_2}{1 - \delta_1 \delta_2 g_1 g_2}, \end{aligned} \tag{5.5}$$

with a similar result for y_2. Assuming that the two amplifiers have equal gain and crosstalk magnitudes (though this is not guaranteed by reciprocity since the two crosstalk paths are not the same), we can define $G = G_1 = G_2$ and $\Delta = \Delta_1 = \Delta_2$. Substituting into (5.5) results in

$$y_1 = G e^{j\phi_1} \frac{x_1 + \Delta G e^{j(\theta_2+\phi_2)} x_2}{1 - \Delta^2 G^2 e^{j(\theta_1+\theta_2+\phi_1+\phi_2)}}. \tag{5.6}$$

We can define a few auxiliary variables

$$\begin{aligned} \rho &= \Delta G \\ \psi &= \theta_2 + \phi_2 \\ \beta &= \theta_1 + \theta_2 + \phi_1 + \phi_2, \end{aligned} \tag{5.7}$$

which simplifies (5.6) to

$$y_1 = G e^{j\phi_1} \frac{x_1 + \rho e^{j\psi} x_2}{1 - \rho^2 e^{j\beta}}. \tag{5.8}$$

The angle β is the total accumulated phase for a signal round trip through both amplifiers, and ψ is the relative phase shift of second input x_2 with respect to x_1. The product ρ is the ratio of the amplifier gain and the crosstalk isolation magnitude. For different values of ρ, there are several distinct behaviors which are sketched in the left diagram of Figure 5-2. It shows the output of the first amplifier y_1 given the intended input x_1 and the coupled signal from the other amplifier ρx_2. The relative phases are chosen such that the interfering signals are out of phase ($\psi = \pi$ radians). We examine these behaviors in detail for two regimes, which correspond to the cases when $\rho < 1$ and $\rho \geq 1$.

The right plot of Figure 5-2 shows how bad the amplitude loss can be for various ratios of the coupled and the desired signal amplitude. Here, the interferer/input magnitude ratio corresponds directly to ρ, but is determined by the time-varying inputs to the multiple amplifiers. As expected, the SNR loss is most pronounced when the ratio is one. However, there is a wide range over which there is a significant drop in the SNR. In addition, although there appears to be little difference when

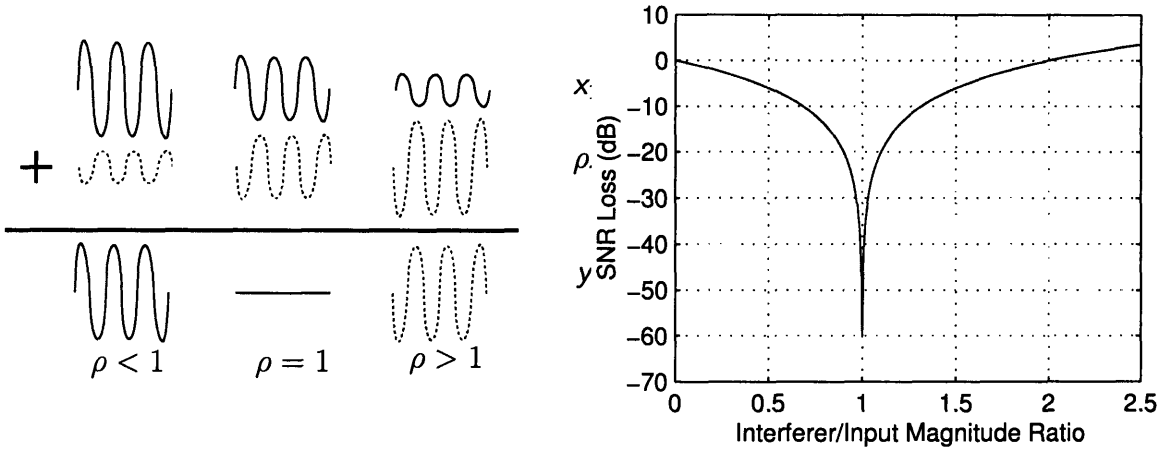


Figure 5-2: Three crosstalk cases for different values of the interferer/input ratio ρ. Unfavorable crosstalk conditions can degrade the SNR for a large range of values, with $\rho > 1$ resulting in the crosstalk signal overwhelming the intended signal.

$\rho > 1$ or $\rho < 1$ according to Figure 5-2, when $\rho > 1$ the crosstalk signal actually overwhelms the desired signal, severely degrades the effective SNR.

5.1.2 Crosstalk behavior when $\rho < 1$

When $\rho < 1$, the magnitude of the crosstalk isolation is greater than the amplifier gain. As shown in the left plot of Figure 5-2, the output of the amplifier is close to the desired signal. This is similar to the case without amplifiers, in which the coupled signal is smaller than the desired signal to be amplified. The result in this case is a loss in SNR that is dependent on the effective crosstalk isolation. The effective crosstalk isolation ρ is the original isolation Δ reduced by the amplifier gain G. Thus, for example, if the crosstalk isolation is 20 dB, and the amplifier gain is 15 dB, then the effective crosstalk isolation is 5 dB. There is also an additional SNR loss due to the noise coupled in from the second amplifier, but this is assumed negligible compared to the interference from the crosstalk signal.

Measurements for unshielded structures at 5 GHz have shown crosstalk isolation of around 15 dB [65]. With amplifier gains near this value, the effective level of crosstalk isolation is very low, and the SNR loss can be significant. The effective isolation gain from phase randomization can be used to offset the isolation losses caused by the amplifier gains. Adding isolation structures can increase the crosstalk isolation, but at the cost of added circuit area and may require unavailable process technology or significant design effort. Phase randomization, on the other hand, does not require additional circuitry and can be implemented by node bandwidth allocation.

5.1.3 Crosstalk behavior when $\rho \geq 1$

When $\rho = 1$, the magnitude of the desired input signal and the coupled crosstalk signal are the same. Depending on the phase in this case, there is the possibility of positive feedback. Focusing on the denominator, we see that if $\rho = 1$ and β is an integer multiple of 2π, then the amplifier output will grow without bound. For an actual system, this positive feedback will cause the amplifier output to quickly saturate. If the crosstalk phases are nearly a multiple of 2π apart, the gain from the amplifier feedback loop can still cause the amplifier output to saturate. Since the crosstalk is not easily measured and may slowly fluctuate (on the order of seconds or longer due to temperature changes or aging, for example), having too large an amplifier gain may lead to an unstable system, unless the crosstalk isolation is large enough to accommodate the larger gain. Consequently, larger gain stages require more attention to be paid to the isolation between the parallel amplifier chains.

Even when β is not a multiple of 2π, the numerator of (5.8) shows there can be destructive interference as well. If (assuming $x_1 = x_2$) $\rho = 1$ and $\phi = \pi$, then the two input signals will destructively interfere, resulting in no output. More generally, if the phase difference between x_1 and x_2 and the phase shift ϕ total an odd multiple of π, then destructive interference can occur, and when the the desired signal and the coupled crosstalk signal are nearly out of phase, the resulting output amplitude will be significantly attenuated. Any frequency inside the system bandwidth that experienced these phase alignments would have a severely degraded SNR.

These particular phase alignments cause the visible effects of crosstalk, namely a deep null or amplifier saturation from positive feedback. At other phases the situation is similarly bad, with the coupled signal magnitude very close to that of the desired signal. The resulting amplified signal can vary greatly in both amplitude and phase depending on the inputs.

In the case when $\rho > 1$ the coupled signal magnitude is actually larger than the desired input signal, thus the original signal is drowned out by the coupled crosstalk signal. If the phase difference between the desired and coupled signal is at or near a multiple of π, the amplifier output will either saturate or have a very small amplitude. At other phases, the desired signal is still drowned out by the coupling signal, resulting in a low SNR.

5.2 Matrix Crosstalk Model

As we have seen with parallel amplifiers, the coupled crosstalk signal can cause a significant loss in the SNR, though this is dependent on the relative amplitudes of the desired and coupled signals, as well as their phase difference. Even when they do not destructively interfere, the crosstalk signal will reduce the SNR of the desired

signal. However, since the magnitude of this effect will be dependent of the phase of the crosstalk, a more detailed model is needed in order to study these degradation caused by the crosstalk.

5.2.1 The Crosstalk Matrix

We use a linear model for the crosstalk, with the effect of crosstalk being the original intended signal, plus scaled and phase-shifted versions of neighboring signals. Thus, if we consider $\mathbf{x}$ as the input to the receiver and $\mathbf{y}$ as the output, then

$$\mathbf{y} = \mathbf{C}\mathbf{x} \tag{5.9}$$

in the absence of noise, where $\mathbf{C}$ is the crosstalk matrix. Looking more closely at the structure of $\mathbf{C}$,

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\mathbf{y}} = \alpha \underbrace{\begin{bmatrix} 1 & \delta_{12} & \cdots & \delta_{1n} \\ \delta_1 & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \delta_{n1} & \cdots & \cdots & 1 \end{bmatrix}}_{\mathbf{C}} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}}_{\mathbf{x}}, \tag{5.10}$$

where the δ_{ij}'s are complex numbers representing the crosstalk between signal paths, and α is a constant to ensure that the crosstalk matrix is passive (*i.e.*, there is no gain in energy). When the signal energy from one path leaks into an independent path, the amplitude is attenuated and the phase is shifted due to the propagation delay. In general, this coupling will be frequency dependent, but for simplicity we will consider a small enough frequency band that the crosstalk coupling is essentially constant. Assuming symmetric coupling, if the outputs

$$y_1 = x_1 + \delta_{21}x_2 \tag{5.11}$$

$$y_2 = \delta_{12}x_1 + x_2\,, \tag{5.12}$$

where δ_{ij} is the coupling coefficient between the jth input and ith output, then $\delta_{ij} = \delta_{ji}$ and the crosstalk matrix $\mathbf{C}$ is symmetric. The crosstalk matrix is not symmetric in general, however, and this is not crucial to the following results. For a given data path, there may be many sources of signal leakage, several of which are shown in Figure 5-3. Along with the inputs and outputs of other amplifiers, the signals can also couple through the power and ground planes as well.

Each crosstalk term can be separated into a magnitude and a phase, with

$$\delta_i = \Delta_i e^{j\theta_i}. \tag{5.13}$$

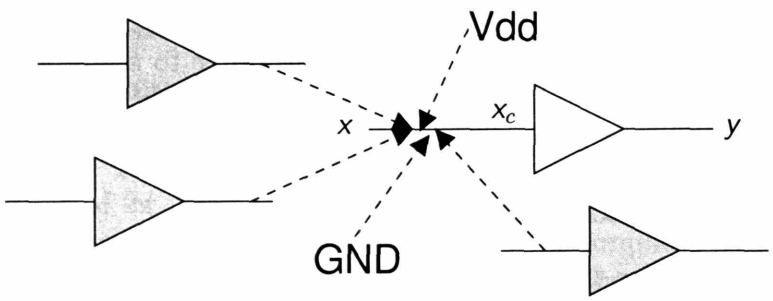


Figure 5-3: Some sources that can couple into an amplifier input.

The crosstalk magnitude is modeled by a constant that is fixed by the particular processing and environmental conditions present at the time of the particular circuit's manufacturing. The crosstalk magnitude is assumed to be identical between all pairs of signals which may experience coupling. This is a worst case scenario, since capacitative coupling usually decreases with circuit separation. There is no reason, however, to believe that distance is the only indicator of crosstalk coupling. For example, with coupling through the ground plane, all circuits might appear as nearly the same distance from each other. The crosstalk phase is assumed to be uniformly distributed, on the other hand, as it is an unknown function of the propagation delay as well as the frequency band of interest. The crosstalk magnitude and phase are in general frequency dependent, but the values vary slowly enough over time (if at all) that they can be assumed to be fixed for a given circuit. We define crosstalk isolation as the ratio of the original to the coupled signal power, or $1/\Delta^2$.

The aggregate effect of all the crosstalk signals is a scaled and phase shifted version of the desired input signal at each frequency. For example, every signal that is present at any point in the circuitry can be written in terms of its Fourier transform. Looking only at the signal components at a given frequency f_o, the amplifier input will see

$$\begin{aligned} x_c(f_o,t) &= x + \sum_i A_i(f_o) e^{j(2\pi f_o t + \theta_i(f_o))} \\ &= \sum_i A_i(f_o) e^{j(2\pi f_o t + \theta_i(f_o))}, \end{aligned} \tag{5.14}$$

where the $A_i(f_o)$'s and $\theta_i(f_o)$'s are the frequency-dependent magnitudes and phases of the ith coupled signal, with $i = 0$ being the desired input signal. The summation can be rewritten as

$$x_c(f_o,t) = e^{j2\pi f_o t} \sum_i A_i(f_o) e^{j\theta_i(f_o)}$$

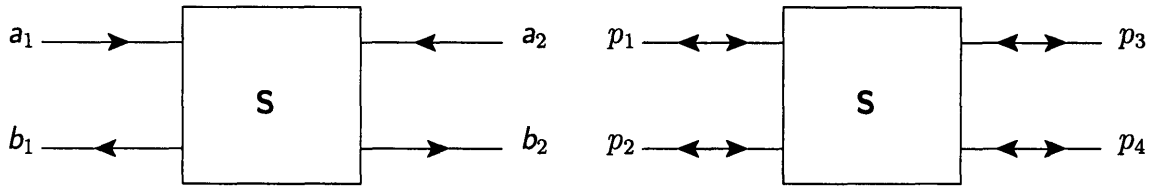


Figure 5-4: Two- (left) and four-port (right) model for a circuit with scattering matrix **S**.

$$\begin{aligned} &= A(f_o)e^{j2\pi f_o t} \\ &= |A(f_o)|e^{j(2\pi f_o t + \angle A(f_o))}, \end{aligned} \tag{5.15}$$

where $A(f_o)$ is given by [75]

$$\begin{aligned} |A(f_o)|^2 &= \sum_i A_i^2 + 2\sum_i \sum_{j>i} A_i A_j \cos[\theta_i(f_o) - \theta_j(f_o)] \\ \angle A(f_o) &= \tan^{-1}\left(\frac{\sum_i A_i \sin(\theta_i)}{\sum_i A_i \cos(\theta_i)}\right). \end{aligned} \tag{5.16}$$

Thus the total contribution of all the coupled sources at a certain frequency is a sinusoid with amplitude and phase given by (5.16).

5.2.2 Ensuring Passivity with the Scattering Matrix

In order to ensure that the modeling of the crosstalk matrix is physically realizable, the matrix must be passive, which can be done by proper choice of α. The crosstalk between two signal paths can be modeled as a four-port network as shown in the right side of Figure 5-4, with ports p_1 and p_3 being the two ends of a signal path, and p_2 and p_4 being the two ends of the other signal path. For simplicity of analysis and clarity, we first use the two-port model on the left side of Figure 5-4.

Each of the two ports (left and right) consists of an incident wave a_i and a reflected wave b_i. The scattering matrix **S** is defined in terms of these incident and reflected waves as

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}, \tag{5.17}$$

with s_{ij} being the ratio of the reflected wave at the ith port resulting from the incident wave from the jth port, or

$$s_{ij} = \frac{b_i}{a_j}. \tag{5.18}$$

For example, for an amplifier that is well matched at both ports, the scattering matrix

might look like

$$\mathbf{S} = \begin{bmatrix} 0 & \epsilon \\ A & 0 \end{bmatrix}, \tag{5.19}$$

where $s_{11} = 0$ implies that port 1 (input) has no reflected power, and $s_{22} = 0$ says the same thing about port 2 (output). Thus the amplifier is matched at the input and output. The typically large gain of the amplifier $s_{21} = A$ is the ratio of the incoming wave on the left port and the outgoing wave on the right port. Finally, $s_{12} = \epsilon \ll 1$ indicates that the input is well isolated from the output, as the ratio of the output wave that couples to the input is nearly zero.

For our crosstalk model, the ports 1 and 2 can be viewed as the two separate signal paths, which in the perfect isolation case, will have a scattering matrix

$$\mathbf{S} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{5.20}$$

so that the incoming wave into each port is reflected back and there is no crosstalk transmission between ports 1 and 2. With the introduction of crosstalk, however, the scattering matrix would look like

$$\mathbf{S} = \alpha \begin{bmatrix} 1 & \delta \\ \delta & 1 \end{bmatrix}, \tag{5.21}$$

where δ is the leakage between the two ports, and α is a constant to ensure that the crosstalk mechanism is passive.

In order to ensure a passive matrix representation, the scattering matrix for the two-port needs to be bounded real [3], satisfying

$$\mathbf{I} - \mathbf{S}^\dagger \mathbf{S} \geq 0, \tag{5.22}$$

where $\mathbf{A} \geq 0$ means that all eigenvalues of $\mathbf{A}$ are nonnegative and $\mathbf{I}$ is the identity matrix. Using the SVD,

$$\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\dagger, \tag{5.23}$$

(5.22) can be simplified to

$$\begin{aligned} \mathbf{I} - \mathbf{S}^\dagger \mathbf{S} &= \mathbf{I} - \left(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\dagger\right)^\dagger \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\dagger \\ &= \mathbf{I} - \mathbf{V}^\dagger \mathbf{\Sigma} \mathbf{U}^\dagger \mathbf{U} \mathbf{\Sigma} \mathbf{V}^\dagger \\ &= \mathbf{V}\mathbf{V}^\dagger - \mathbf{V}^\dagger \mathbf{\Sigma}^2 \mathbf{V}^\dagger \\ &= \mathbf{V}\left(\mathbf{I} - \mathbf{\Sigma}^2\right)\mathbf{V}^\dagger. \end{aligned} \tag{5.24}$$

We can define

$$\begin{aligned} \mathbf{\Lambda} &= \mathbf{I} - \mathbf{\Sigma}^2 \\ \mathbf{A} &= \mathbf{V}\left(\mathbf{I} - \mathbf{\Sigma}^2\right)\mathbf{V}^\dagger = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}, \end{aligned} \tag{5.25}$$

where $\mathbf{V}^\dagger = \mathbf{V}^{-1}$ since $\mathbf{V}$ is unitary. Since $\mathbf{\Sigma}$ is diagonal, $\mathbf{\Lambda}$ is a diagonal matrix as well, thus (5.25) is an eigenvalue decomposition of $\mathbf{A}$. Finally, since the eigenvalues of $\mathbf{A}$ are the diagonal elements of $\mathbf{\Lambda}$,

$$\begin{aligned} [\mathbf{\Lambda}]_{ii} &= 1 - s_i^2 \geq 0 \\ &\downarrow \\ s_i^2 &\leq 1, \end{aligned} \tag{5.26}$$

or $s_i \leq 1$ where s_i are the singular values of $\mathbf{S}$ (note that the singular values are always nonnegative). Thus the passivity constant α should be chosen such that the largest singular value is bounded by one, or

$$\alpha = \begin{cases} 1/s_{\max} & s_{\max} > 1 \\ 1 & s_{\max} \leq 1 \end{cases} \tag{5.27}$$

where $s_{\max}$ is the largest possible singular value. In order to ensure the crosstalk matrix is passive for all realizations, the value for α is determined by the largest possible singular value for all values of the phases of the off-diagonal elements.

The largest singular value of a matrix can be related to its eigenvalues through its spectral norm and radius. The spectral norm of a matrix is given by [27]

$$||\mathbf{S}||_2 = \max_{\mathbf{x}} \frac{||\mathbf{S}\mathbf{x}||}{||\mathbf{x}||} = s_{\max}, \tag{5.28}$$

where $||\cdot||$ represents the Euclidean norm of the vector and $s_{\max}$ is the largest singular value of $\mathbf{S}$. The spectral norm can be upper bounded by

$$||\mathbf{S}||_2^2 \leq ||\mathbf{S}||_1 ||\mathbf{S}||_\infty, \tag{5.29}$$

where

$$||\mathbf{S}||_1 = \max_j \sum_i |s_{ij}| \tag{5.30}$$

$$||\mathbf{S}||_\infty = \max_i \sum_j |s_{ij}| \tag{5.31}$$

are the maximum absolute column and row sums, respectively [27]. The spectral radius of a matrix can similarly be defined as the magnitude of the largest eigenvalue, which then lower bounds the spectral norm of the matrix [27].

The crosstalk matrix can be written as

$$\mathbf{C} = \mathbf{I} + \mathbf{\Delta}, \tag{5.32}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{\Delta}$ has zeros along its main diagonal. The off-diagonal elements of $\mathbf{\Delta}$ all have magnitude Δ and random phases. The eigenvalues of the matrix $\mathbf{C}$ all lie within Geršgorin discs which have radius

$$R_i = \sum_{i \neq j} |[\mathbf{\Delta}]_{ij}|, \tag{5.33}$$

with the centers given by the diagonal elements [27]. The largest eigenvalue of $\mathbf{C}$ is then upper bounded by $1+(N-1)\Delta$ for a $N \times N$ crosstalk matrix. It is easy to check that this bound is achievable if the phases of all the crosstalk terms is zero. Thus the maximum possible eigenvalue for the crosstalk matrix is $1 + (N-1)\Delta$, which lower bounds the maximum possible singular value.

Because the magnitudes of the off-diagonal elements of the crosstalk matrix are all Δ, the row and column sums are identical and equal to $1+(N-1)\Delta$, regardless of the phases of the crosstalk entries. From (5.29), we see that the largest singular value cannot be larger than $1 + (N-1)\Delta$. Because the upper and lower bounds coincide, we see that the largest possible singular value is the same as the magnitude of the largest possible eigenvalue, or $s_{\max} = |\lambda_{\max}|$. Thus because the crosstalk scattering matrix in (5.21) has eigenvalues $1 \pm \delta$, the passivity constant

$$\alpha = \frac{1}{1+\Delta}, \tag{5.34}$$

since $\Delta = |\delta|$ from (5.13).

We could also define the crosstalk in terms of the four-port model in Figure 5-4, in which ports d_1 and d_3 are two ends of the same signal path, and d_2 and d_4 are the two ends of the other signal path. In this case, the scattering matrix would be

$$\mathbf{S} = \alpha \begin{bmatrix} 0 & 0 & 1 & \delta \\ 0 & 0 & \delta & 1 \\ 1 & \delta & 0 & 0 \\ \delta & 1 & 0 & 0 \end{bmatrix}, \tag{5.35}$$

which has eigenvalues $\lambda = \{\pm(1+\delta), \pm j(1-\delta)\}$. Once again, the passivity constant is the same as for the two-port model above, and given in (5.34).

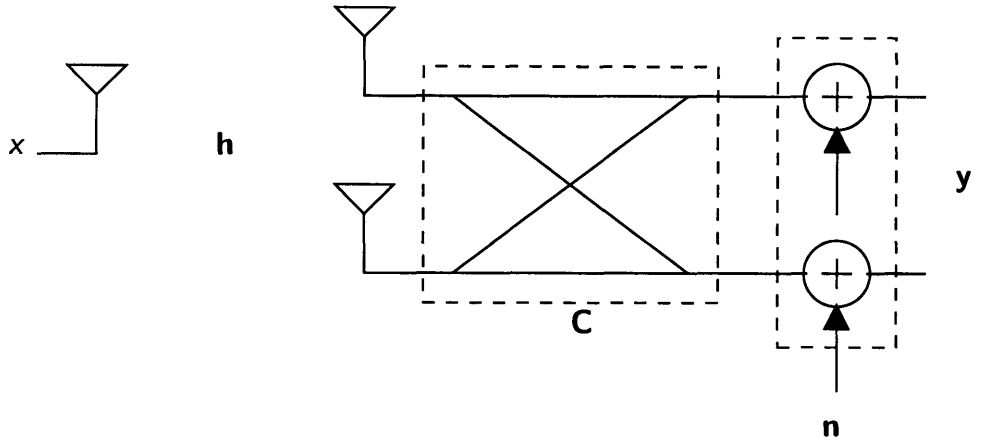


Figure 5-5: Crosstalk model for a 1 × 2 system.

For a 2 × 2 crosstalk matrix, then, the passivity constant ensures that the maximum possible eigenvalue does not have magnitude greater than unity, which corresponds to $\alpha = (1 + \Delta)^{-1}$. Although the form of α is simple for this case, it becomes significantly more complicated for larger numbers of antennas. In general, the bounded real property for the scattering matrix corresponds to assuring that the magnitudes of all singular values are equal to or less than unity. The value for α may seem very pessimistic, especially when the crosstalk matrix is large, but it does ensure that the crosstalk is physically realizable, and facilitates comparison between different phase configurations.

5.3 A 1 × 2 System with Crosstalk

As a the simplest case of circuit crosstalk, we can look at a 1 × 2 system, where crosstalk can be present between the two transmit RF chains, as well as between the two receiver RF chains and there is no crosstalk at the transmitter, as diagrammed in Figure 5-5.

The input x and output $\mathbf{y}$ of this system can be written as

$$\mathbf{y} = \mathbf{C}\mathbf{h}x + \mathbf{n}, \tag{5.36}$$

where $\mathbf{C}$ is the crosstalk matrix at the receiver and $\mathbf{n}$ is the receiver noise. Note that the behavior of the crosstalk matrix is similar to the channel matrix (in this case a vector) $\mathbf{h}$. In fact, when a receiver tries to learn the channel matrix from training data, it will not be able to learn $\mathbf{h}$ alone, but rather the combination of the channel and the

crosstalk. If the receiver tried to share the channel information with the transmitter it would not transmit the correct information, because the crosstalk matrices might be different for the transmit and receive circuits in a node. Alternately, in a two-way transmission between two nodes, the initial receiver might try (due to reciprocity) to use the channel information it learned to shape its transmission back to the original transmitter. Due to the fact that the transmit and receive circuitry of a node is necessarily different, the crosstalk matrix will not be the same, and thus the wireless channel will not be reciprocal. If the effects of crosstalk are large, this non-reciprocal nature could cause significant deviations from expected behavior.

Because of this, instead of trying to learn the crosstalk conditions, we use phase randomization to average out the effects of crosstalk across all circuit realizations. This prevents different instances of a circuit from having wildly different behaviors depending on the particular crosstalk phase it happens to experience. To gauge the difference in performance, we derive worst case and average performance as a function of the crosstalk matrix, as well as the SNR gain provided by the phase randomization.

5.3.1 Crosstalk as a Static Fading Channel

As previously noted, that the effect of the crosstalk matrix is similar to that of the wireless channel. Part of the hostile nature of the wireless channel is that it varies unpredictably with time, forcing the system to continuously relearn its characteristics. However, as we have seen, this random nature can be used as an advantage by averaging over many different channel realizations for diversity gain. For the crosstalk, however, even though the crosstalk matrices are also random, they are fixed for a given circuit. The particular crosstalk matrix for a given circuit is determined by the process variations during manufacturing, and even if it should change slightly over time due to heating or some other effect, this is many orders of magnitude slower than the timescales of communicating packets. Since there is no variation to the crosstalk, a circuit might be "stuck" forever with a particularly bad instance of crosstalk. In order to have acceptable yield rates for these circuits, a crosstalk margin must be added to the link budget. If chip yield needs to be very high, this margin may be unacceptably large. Since the effect of crosstalk is similar to that of the wireless channel, our strategy will be to try to use the averaging effect of diversity to make all circuit realizations experience average case behavior. Unlike the wireless channel, which may at times have a larger gain than the average, producing a channel that is actually better than if the fading were not there, the presence of the crosstalk will never improve the performance above what could be achieved if no crosstalk existed. The size of the margin that must be allocated for the circuit yield to be sufficiently high is dependent on the probability density of the crosstalk magnitudes, for which we do not have a good model.

5.3.2 Role of the Singular Values

We can expand (5.36) to include the structure of the crosstalk matrix,

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \alpha \begin{bmatrix} 1 & \delta \\ \delta & 1 \end{bmatrix} \mathbf{x} + \mathbf{n}, \tag{5.37}$$

where δ is a complex number representing the crosstalk and α is a constant to keep the matrix $\mathbf{C}$ passive. The vector $\mathbf{x}$ is defined as

$$\mathbf{x} = \begin{bmatrix} h_1 x \\ h_2 x \end{bmatrix}. \tag{5.38}$$

We can view the signal received by the antennas as the input signal, which is distorted by the crosstalk matrix. Ignoring noise, we are back to the simple crosstalk model of (5.9). The crosstalk parameter δ is further modeled as $\delta = \Delta e^{j\theta}$, where Δ and θ are the magnitude and phase of the crosstalk. We assume that the magnitude of the crosstalk is a random, fixed parameter that is set by the particular instantiation of the circuit. The system is also assumed to be narrowband (or uses narrow frequency channels), so that the delay of the crosstalk signal can be approximated by the phase θ. Although we are focusing on the effects of the receiver crosstalk $\mathbf{C}$, these results apply similarly for the case of transmitter crosstalk $\mathbf{D}$. We also assume that the receiver is able to learn the cascade of the crosstalk and channel matrices $\mathbf{Ch}$, but not either one individually.

For a given crosstalk realization, the effect on the input will depend on the singular values of $\mathbf{C}$. From the SVD, we know that the matrix $\mathbf{C}$ can be decomposed into $\mathbf{U\Sigma V}^\dagger$, with $\mathbf{U}$ and $\mathbf{V}$ being unitary and $\mathbf{\Sigma}$ a diagonal matrix with positive elements. If the input x is represented as a circle, then when it passes through the crosstalk matrix, the result is the circle scaled by the singular values into an oval where the axes lengths are the singular values of $\mathbf{C}$. The unitary matrices rotate the original signal before scaling and the resulting oval after scaling.

For example, Figure 5-6 looks at the effect of the crosstalk matrix on a unit input circle, with points highlighted for a 4-QAM constellation. The length of the arrows show the effect of the crosstalk singular values, with the right two diagrams showing the scaling effects of two different crosstalk orientations. In both these cases, the two singular values are assumed to be different, with the shortened length of the arrows in the right two diagrams representing the SNR loss by that particular point on the input unit circle. Because the performance of the input constellation is dominated by the minimum distance between constellation points, the net SNR loss is determined by the shortest of these arrows.

The middle plot shows the case where the axes if the input constellation are

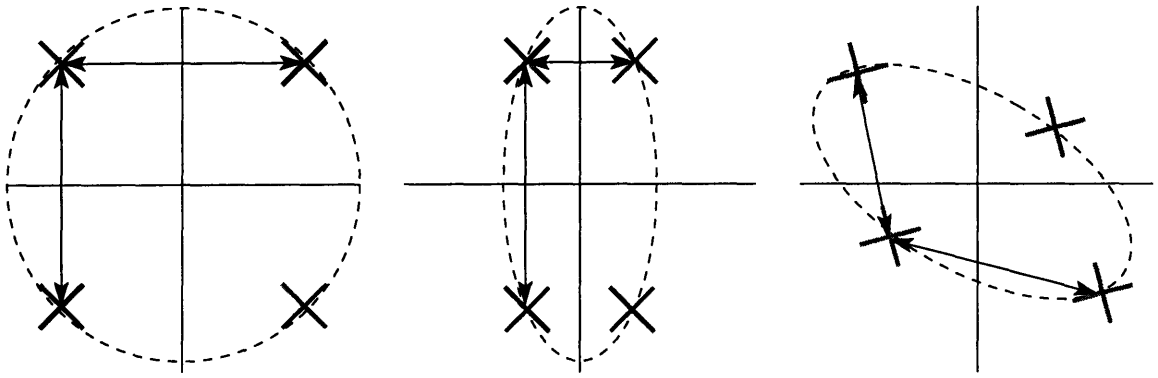


Figure 5-6: Effect of singular value ellipse on the input vector.

aligned with the singular value axes of the crosstalk matrix. The shorter arrow represent the worst case SNR loss, which occur when the input vector is collinear with the axis of the smallest singular value. Because of the passivity constant α, the longest possible arrow is the same length as the input circle, so there is no SNR gain possible from crosstalk. The arrows in the right diagram show the result of a similar crosstalk matrix, except that the axes of the ellipse are not aligned with the axes of the constellation. The length of the arrows is now given by a combination of the scaling of the two singular values, resulting in less SNR loss than the worst case.

For the case of two receive antennas, the crosstalk matrix $\mathbf{C}$ has the form

$$\mathbf{C} = \alpha \begin{bmatrix} 1 & \delta \\ \delta & 1 \end{bmatrix}. \tag{5.39}$$

Ignoring α, the singular values of $\mathbf{C}$ are $s_i = 1 \pm \delta$. Since the largest possible singular value is $s_{\max} = 1 + \Delta$, the squares of the singular values including α are given by

$$s_i^2 = \left\{ \frac{1 + 2\Delta\cos(\theta) + \Delta^2}{(1+\Delta)^2}, \frac{1 - 2\Delta\cos(\theta) + \Delta^2}{(1+\Delta)^2} \right\}. \tag{5.40}$$

If we plot the singular values with respect to each other as in Figure 5-7, we can see that increasing the value of one singular value decreases the other. The magnitudes of the singular values lie on a circle, with the radius determined by α, and hence the crosstalk magnitude. The reason the singular values lie on a circle comes from the fact that the Frobenius norm of a matrix is defined as [27]

$$||\mathbf{A}||^2 = \sum_i \sum_j |\mathbf{A}_{ij}|^2, \tag{5.41}$$

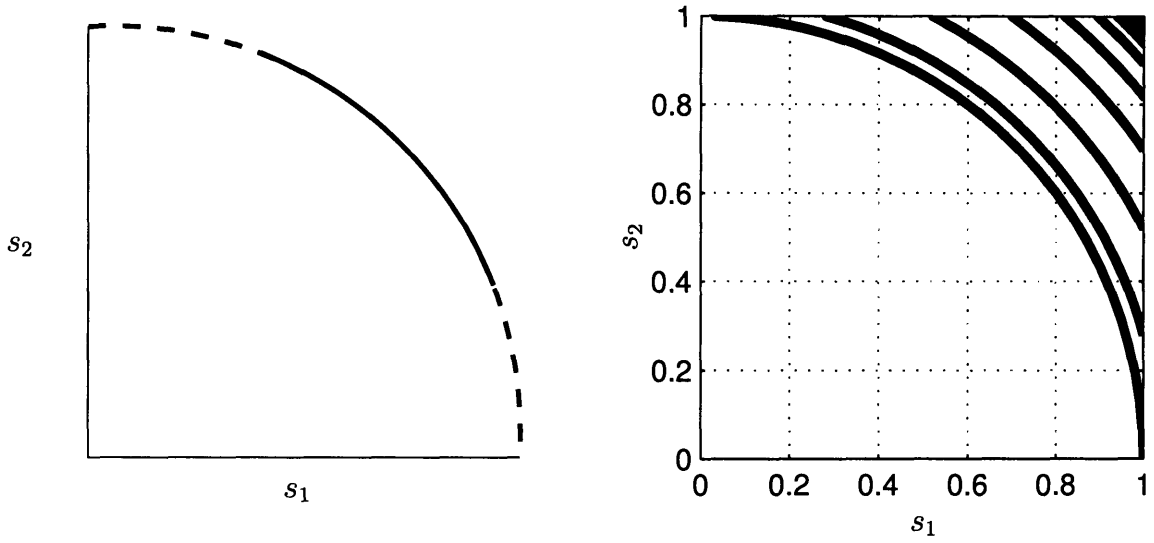


Figure 5-7: Singular values as a function of crosstalk magnitude (right) for 0 to 30 dB in 5 dB increments. Left side shows the range of singular values for a small (solid) and large (dashed) crosstalk magnitude. The crosstalk phase is uniformly distributed.

and is also the sum of the squares of the singular values,

$$||\mathbf{A}||^2 = \mathrm{Tr}(\mathbf{A}\mathbf{A}^\dagger) = \sum_i \lambda_i = \sum_i s_i^2, \tag{5.42}$$

where λ_i are the eigenvalues of $\mathbf{A}\mathbf{A}^\dagger$ and $\mathrm{Tr}(\cdot)$ is the trace of a matrix. Because only the phase of the crosstalk terms change but their magnitude remains constant, the Frobenius norm of the crosstalk matrix is constant as the phase varies. Thus the squared sum of the singular values is constant as well.

By superimposing the singular value curves as in the left plot of Figure 5-7, we can see that the value of Δ determines what fraction of the quarter circle the singular values can occupy. As the smallest singular value determines the worst case degradation, ideally the singular values would be confined to as small a range of the circle as possible. As expected, the larger the crosstalk magnitude, the wider range the singular value magnitudes can have.

5.3.3 Worst Case SNR Loss From Crosstalk

Since the performance of any encoding scheme decreases monotonically as the SNR falls, we can bound the worst case performance for a given magnitude of crosstalk as

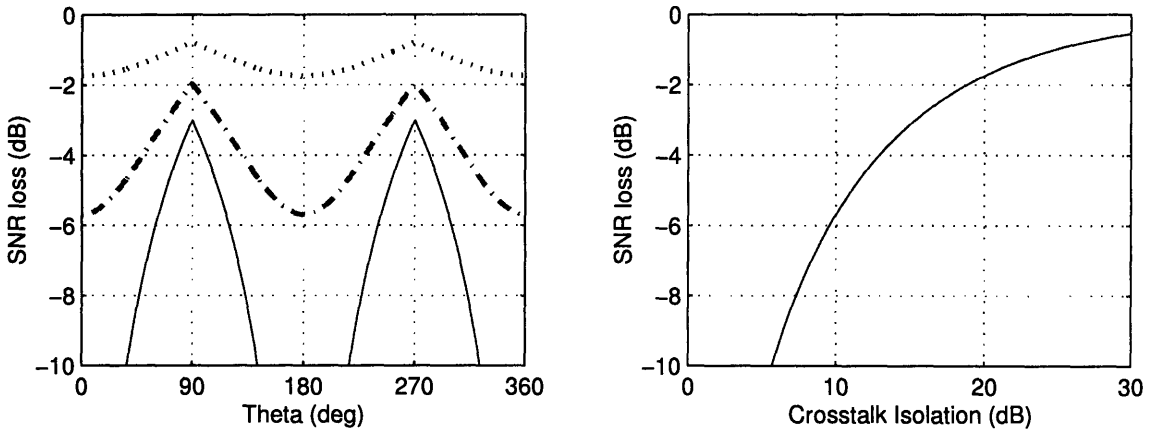


Figure 5-8: For a 2×2 crosstalk matrix, lower bounds on performance degradation due to crosstalk (left plot) as a function of phase for 0 dB (solid), 10 dB (dash-dot) and 20 dB (dotted) isolation, and worst case performance as a function of crosstalk isolation (right).

a function of θ, as shown in Figure 5-8. The left plot shows the worst case bound on the SNR loss due to the crosstalk as a function of the phase for several crosstalk magnitudes. When $\theta = 0°$ or $\theta = 180°$, one of the singular values has the minimum possible magnitude of $(1-\Delta)/(1+\Delta)$. The worst case bound shown in the right plot of Figure 5-8 is equaled when the received vector $\mathbf{h}x$ is such that the constellation axis is collinear with the axis associated with the minimum singular value. This corresponds to the middle case of Figure 5-6, where the minimum distance between constellation points is most significantly reduced.

The actual performance of the system at the worst case phase will be better than the bound because the received vector will in general be scaled by both singular values rather than only the smallest one. Note that as the isolation approaches 0 dB, the SNR loss becomes infinite. We also see that if the isolation is sufficiently high, even the worst case performance loss is not significant. At 10 dB, worst case loss is about 5.7 dB, and at 20 dB, it is about 1.7 dB.

A similar set of curves could be constructed for a best case bound, which occurs when the constellation lines up with the largest singular value instead of the smallest. However, since the direction of the received vector will change with the input as well as with changing channel conditions, keeping the input correctly lined up with the best singular value axis is overly constraining. To maximize performance, then, the goal is to maximize the minimum singular value. Since the Frobenius norm of the matrix is constant, the sum of the squares of the singular values is also a constant. Thus, the best case is when the singular values all have the same magnitude, resulting

in a crosstalk matrix that is a unitary transformation scaled by a constant.

5.3.4 Asymptotic Averaged SNR Loss From Crosstalk

Because the singular value axes of the crosstalk matrix are not going to line up in general with the axes of the received constellation, we can derive an expression for the SNR loss caused by the crosstalk matrix in the high SNR regime. This analysis is for a general $N \times N$ crosstalk matrix, which we can specialize to the 2×2 case.

Again using the SVD, we can diagonalize the crosstalk matrix $\mathbf{C}$ into

$$\mathbf{C} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\dagger}, \tag{5.43}$$

where $\mathbf{U}$ and $\mathbf{V}$ are unitary and $\mathbf{\Sigma}$ is a real diagonal matrix of the singular values of $\mathbf{C}$. The crosstalk matrix can then be rewritten as

$$\mathbf{C} = \sum_i s_i \mathbf{u}_i \mathbf{v}_i^{\dagger}, \tag{5.44}$$

where s_i is the ith singular value, and $\mathbf{u}_i$ and $\mathbf{v}_i$ are the ith columns of $\mathbf{U}$ and $\mathbf{V}$, respectively. If we define the projection

$$\tilde{x}_i = \langle \mathbf{x}, \mathbf{v}_i \rangle = \mathbf{v}_i^{\dagger}\mathbf{x}, \tag{5.45}$$

then the effect of the crosstalk matrix is

$$\mathbf{C}\mathbf{x} = \sum_i s_i \mathbf{u}_i \mathbf{v}_i^{\dagger}\mathbf{x} = \sum_i s_i \tilde{x}_i \mathbf{u}_i. \tag{5.46}$$

The expected magnitude of the vector $\mathbf{C}\mathbf{x}$ is then

$$E\left[||\mathbf{C}\mathbf{x}||^2\right] = E\left[\left(\sum_i s_i \tilde{x}_i \mathbf{u}_i\right)^{\dagger}\left(\sum_j s_j \tilde{x}_j \mathbf{u}_j\right)\right] \tag{5.47}$$

$$= E\left[\sum_i \sum_j s_i s_j \tilde{x}_i \tilde{x}_j \mathbf{u}_i^{\dagger}\mathbf{u}_j\right] \tag{5.48}$$

$$= E\left[\sum_i s_i^2 \tilde{x}_i^2\right] = \sum_i s_i^2 E[\tilde{x}_i^2], \tag{5.49}$$

where the terms with $i \neq j$ disappear because the $\mathbf{u}_i$'s are orthonormal. Assuming that the direction of $\mathbf{x}$ is uniformly distributed, all of the $\tilde{x}_i$'s have the same expected

power, which is $1/N$th the total power $||\mathbf{x}||^2$, thus

$$E\left[||\mathbf{Cx}||^2\right] = E[\tilde{x}_1^2]\sum_i s_i^2 = \frac{||\mathbf{x}||^2}{N}\sum_i s_i^2. \tag{5.50}$$

However, we know from (5.41) and (5.42) that the sum of the squares of the singular values does not vary with the crosstalk phase, thus thus the expected effect of the crosstalk matrix $\mathbf{C}$ on the magnitude of a vector $\mathbf{x}$ is a constant. This does not agree with the worst case bounds of Figure 5-8, in which the SNR loss should certainly be a function of the crosstalk phase. The degradation in performance due to a crosstalk matrix thus must be a nonlinear function of the matrix.

To determine the SNR loss, we must examine the crosstalk behavior in more detail. Starting from the general receiver crosstalk relation

$$\mathbf{y} = \mathbf{CHx} + \mathbf{n} = \mathbf{Cx}' + \mathbf{n}, \tag{5.51}$$

which we can decode using the matched filter

$$\tilde{\mathbf{y}} = \mathbf{H}^\dagger\mathbf{C}^\dagger\mathbf{y}. \tag{5.52}$$

Substituting into (5.51),

$$\begin{aligned}
\tilde{\mathbf{y}} &= \mathbf{H}^\dagger\mathbf{C}^\dagger\mathbf{CHx} + \mathbf{H}^\dagger\mathbf{C}^\dagger\mathbf{n} && (5.53)\\
&= \mathbf{H}^\dagger\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\dagger\mathbf{\Sigma}\mathbf{U}^\dagger\mathbf{Hx} + \mathbf{H}^\dagger\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\dagger\mathbf{n} && (5.54)\\
&= \tilde{\mathbf{H}}^\dagger\mathbf{\Sigma}^2\tilde{\mathbf{H}}\mathbf{x} + \tilde{\mathbf{H}}^\dagger\mathbf{\Sigma}\tilde{\mathbf{n}}, && (5.55)
\end{aligned}$$

where $\tilde{\mathbf{H}} = \mathbf{U}^\dagger\mathbf{H}$ and $\tilde{\mathbf{n}} = \mathbf{V}^\dagger\mathbf{n}$. Because $\mathbf{n}$ is i.i.d. Gaussian and $\mathbf{V}$ is unitary, $\mathbf{n}$ and $\tilde{\mathbf{n}}$ share the same probability distribution. Looking at the ith element of $\mathbf{y}$,

$$\begin{aligned}
y_i &= \tilde{\mathbf{h}}_i^\dagger\mathbf{\Sigma}^2\tilde{\mathbf{h}}_i x_i + \tilde{\mathbf{h}}_i^\dagger\mathbf{\Sigma}\tilde{\mathbf{n}}_i && (5.56)\\
&= \sum_i |\tilde{h}_i|^2 s_i^2 x_i + \sum_i \tilde{h}_i s_i n_i. && (5.57)
\end{aligned}$$

For a given realization of the channel $\mathbf{H}$ and the crosstalk $\mathbf{C}$, the left summation is a constant and the right summation is a Gaussian random variable with distribution

$$y_i \sim \mathcal{N}\left(\eta x_i, \eta\sigma^2\right). \tag{5.58}$$

Here η is defined as

$$\eta = \sum_{i=1}^{N}\eta_i = \sum_{i=1}^{N}|\tilde{h}_i|^2 s_i^2, \tag{5.59}$$

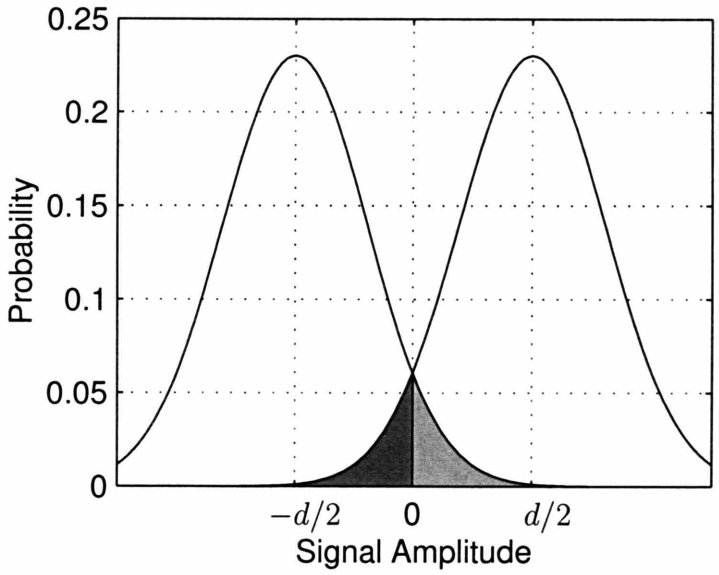


Figure 5-9: PDFs of constellation points at $\pm d/2$ in AWGN. The probability of error $Pr(\epsilon)$ is given by the integral of the shaded areas.

with $n_i \sim \mathcal{N}(0, \sigma^2)$. To find the probability of error for each y_i, we define d as the minimum distance between any two constellation points from which x_i is chosen. The most likely probability event is mistaking two points that are separated by d in amplitude, which is given by the integral of the shaded area in Figure 5-9. The error event ϵ is defined as the probability that the noise pushes the value of each constellation point into the shaded region (the darker shaded region on the left is for the constellation point at $d/2$, and similarly for the lighter shaded region and $-d/2$. Although there are other ways to make an error besides a nearest neighbor misidentification, those events are very unlikely at high SNR. Assuming x_i was chosen from a signal constellation with points at $\pm d/2$, after passing through the wireless channel and the crosstalk matrix, the PDFs are centered at $\pm d\sqrt{\eta}/2$.

The probability of an error event ϵ is given by

$$Pr(\epsilon) = 2Q\left(\frac{d}{2\sigma}\right) = 2Q\left(\sqrt{\frac{\eta}{\sigma^2}}\right), \tag{5.60}$$

where $Q(x)$ is the integral of the tail of a Gaussian random variable, and defined as

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{t^2/2}\, dt. \tag{5.61}$$

There is no closed form solution for $Q(x)$, but its value can be upper bounded by the Chernoff bound as

$$Q(x) \leq e^{-x^2/2}. \tag{5.62}$$

Thus, for the two constellation points located at $\pm d/2$, the probability of error is

$$Pr(\epsilon) \leq \exp\left(-\frac{d^2}{8\sigma^2}\right) = \exp\left(-\frac{\eta}{2\sigma^2}\right). \tag{5.63}$$

Since η depends on the particular crosstalk and channel matrix, in order to determine the average probability of error for a given crosstalk matrix, we must take the expectation of (5.63) over all values of $\mathbf{H}$,

$$E[Pr(\epsilon)] = 2E\left[\exp\left(-\frac{\eta}{2\sigma^2}\right)\right]. \tag{5.64}$$

This expectation can be evaluated via the moment generating function [53]

$$\begin{aligned} E[Pr(\epsilon)] &= 2\prod_{i=1}^{N} \frac{1}{1+\mathrm{E}\left[\eta_i\right]/\sigma^2} && (5.65)\\ &= 2\prod_{i=1}^{N} \left(1+\frac{s_i^2}{\sigma^2}\mathrm{E}\left[|\tilde{h}_i|^2\right]\right)^{-1} && (5.66)\\ &= 2\prod_{i=1}^{N} \frac{1}{1+s_i^2\mathrm{SNR}_i}, && (5.67) \end{aligned}$$

where SNR_i is the average received SNR for the ith receive antenna. Because $\mathbf{U}$ is unitary,

$$E[|h|^2] = E[|\tilde{h}|^2] = 1. \tag{5.68}$$

In the high SNR regime, the one in the denominator is negligible, thus

$$E[Pr(\epsilon)] = 2\left(\frac{1}{\mathrm{SNR}}\right)^N \prod_{i=1}^{N} \frac{1}{s_i^2}. \tag{5.69}$$

The singular values of $\mathbf{C}$ are defined as the square root of the eigenvalues of $\mathbf{CC}^\dagger$, and the determinant of a matrix is the product of its eigenvalues. Given that

$$|\mathbf{CC}^\dagger| = \prod_i \lambda_i = \prod_i s_i^2, \tag{5.70}$$

where λ_i are the eigenvalues of $\mathbf{CC}^\dagger$, the expected probability of error can be written

as

$$E[Pr(\epsilon)] = 2\left(\frac{1}{\mathrm{SNR}}\right)^N \frac{1}{|\mathbf{C}\mathbf{C}^\dagger|}. \tag{5.71}$$

The SNR degradation in error performance caused by the crosstalk matrix can then be written as

$$\mathrm{SNR}_{\mathrm{loss}} = |\mathbf{C}\mathbf{C}^\dagger|^{\frac{1}{N}} = \left(\prod_{i=1}^{N} s_i^2\right)^{\frac{1}{N}}, \tag{5.72}$$

since the effective SNR is the original SNR minus $\mathrm{SNR}_{\mathrm{loss}}$. Because of the high SNR approximation, this result is only valid in the asymptotic sense, but it is quite accurate for values of SNR larger than 10 dB. Thus the asymptotic SNR loss as a function of the crosstalk matrix is the geometric mean of the squared singular values of the crosstalk matrix, or alternately the Nth root of the squared crosstalk matrix determinant. Recalling that the sum of the squares of the singular values is a constant for all crosstalk matrices with a fixed crosstalk isolation magnitude, the smallest value for the SNR loss occurs when all the singular values are equal, and the loss grows without bound as one or more of the singular values goes to zero.

For the case of a 2×2 matrix, the singular values are given by (5.40), which results in an SNR loss of

$$\mathrm{SNR}_{\mathrm{loss}} = \frac{\sqrt{1 + 2\Delta^2\left(1 - 2\cos^2\theta\right) + \Delta^4}}{(1+\Delta)^2}. \tag{5.73}$$

Figure 5-10 shows the averaged SNR loss given by (5.73) for several crosstalk isolations. Compared to Figure 5-8, the average SNR loss is several dB less than the worst case bound at the worst case phases of 0° and 180°, although it still grows arbitrarily large when the crosstalk isolation approaches 0 dB.

5.3.5 Effect of Crosstalk Phase Randomization

Although the phase of the crosstalk is a random quantity, it is fixed for a particular circuit due to the conditions under which it was manufactured. A particular instance of a circuit may have a favorable crosstalk phase, but since circuits interact with each other, then extra SNR margin to account must be allocated for the worst case behavior.

If the phase of the crosstalk could be measured and controlled precisely, the optimal solution is to set the phase to 90° via a compensation circuit or other means. However, even if this were possible, it could require significant effort in the design of the circuit or control systems to allow this to be done. This is a similar situation to the wireless channel, which can produce deep fades, but the system cannot generally

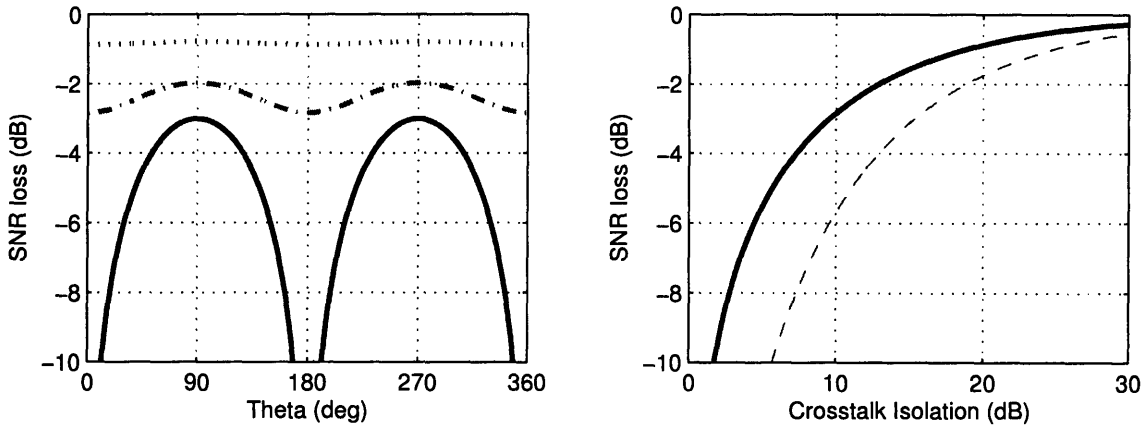


Figure 5-10: Asymptotic SNR loss for a 2×2 crosstalk matrix averaged over all inputs for a fixed crosstalk phase (left) for 0 (solid), 10 (dot-dash), and 20 dB (dotted) isolation. Right plot compares average (solid) to worst case performance (dashed) for $\theta = 180°$.

change its parameters and must deal with what it is given.

Looking again at (5.51), we see that the crosstalk matrix $\mathbf{C}$ has the same effect as the channel matrix $\mathbf{H}$. We also know that the wireless channel may temporarily produce a deep fade making transmission difficult. Using diversity allows the transmitter to transmit at a later time, or at a different frequency, or use different antennas all in an effort to reduce the likelihood that a deep fade will adversely affect the ability to transmit data. At worst, the transmitter can wait for the channel parameters to change before attempting to transmit data. The fixed nature of the crosstalk, however, prevents a similar strategy, since a circuit that is "stuck" in a bad crosstalk "fade" will be stuck forever.

If the phase of the crosstalk is allowed to vary randomly, however, each circuit then experiences the average crosstalk behavior. In other words, the crosstalk phase varies randomly between each transmission, allowing transmissions to experience the average singular value. This equalizes performance between different physical realizations of the circuit, which prevents a circuit from being "stuck" with a bad crosstalk phase. The averaging effect of the phase randomization works in the same way as diversity for the wireless channel. The left plot in Figure 5-11 shows the averaged SNR loss due to the crosstalk as a function of the crosstalk isolation. Since the phase randomization averages out the phase dependence in Figure 5-10, the resulting SNR loss is less than the worst case, but not as good as the performance when the phase is most favorable. What the phase randomization does do, however, is average out the variability in performance for different circuit realizations. This allows a smaller SNR design margin

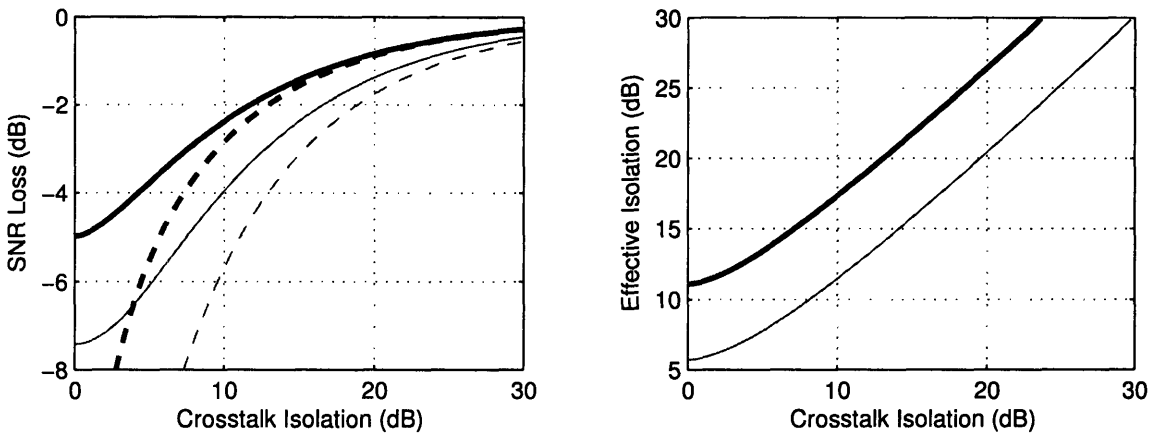


Figure 5-11: Left shows performance of randomized phase algorithm (solid lines) for all inputs (thick) and worst case inputs (thin). Dotted lines show the unrandomized performance for comparison, and right plot shows equivalent crosstalk isolation.

to be used to ensure proper circuit operation.

Even at 0 dB crosstalk isolation, the average SNR loss is not more than 5 dB. Also shown is the phase randomization result for the worst case bound of Figure 5-8. Even in this case, the average SNR loss is less than 8 dB. With the phase randomization, there are still times when the SNR loss can be arbitrarily large, but a circuit will only spend a small fraction of the time in the bad phase configurations.

Because of the reduced SNR loss from phase randomization, the required SNR design margin can be reduced. This can be viewed as an equivalent crosstalk isolation. At 10 dB, for example, the averaged worst case SNR loss is about 4 dB, which is almost 2 dB better than for the worst case phase. On the other hand, the average case SNR gain for 10 dB isolation is only a fraction of a dB. The largest gains in SNR occur when the crosstalk isolation is close to 0 dB, and is nearly zero when the crosstalk isolation is 20 dB or more. The right plot of Figure 5-11 shows the equivalent crosstalk isolation with phase randomization when compared to the worst case unrandomized case. Most notable is that the randomization provides a minimum effective isolation of 5–10 dB even if there is no actual crosstalk isolation. Although the effective isolation is nearly identical to the actual isolation at 20 dB or more for the average case, there an almost constant 6 dB isolation gain over the worst case behavior. The added crosstalk isolation can make it easier to accommodate higher gain stages without causing undesirable feedback effects. If we consider the crosstalk to be akin to the wireless channel, then phase randomization is analogous to turning a slow or static fading channel into a fast fading channel.

5.4 Crosstalk Performance of Larger Systems

Although the case of a 2×2 crosstalk matrix is easy to solve and the bounds can be calculated in a closed form, the general case of a $N \times N$ crosstalk matrix is more difficult. The best case goal is still to make all the singular values equal, and the worst case performance is given by the matrix that minimizes the determinant of $\mathbf{C}\mathbf{C}^\dagger$, which maximizes the SNR loss given by (5.72). Given that the Frobenius norm of the matrix will be constant as the crosstalk phase changes, we can bound the size of the lowest singular value as was done with the 2×2 case. Since the determinant of $\mathbf{C}\mathbf{C}^\dagger$ is the product of its eigenvalues, the worst case should occur when one eigenvalue is as small as possible. The magnitudes of the eigenvalues of $\mathbf{C}\mathbf{C}^\dagger$ are the squares of the singular values of $\mathbf{C}$, so this condition also corresponds to the smallest possible singular value of $\mathbf{C}$.

From (5.33), and given that all off-diagonal elements have the same magnitude Δ, the maximum or minimum possible singular value is at most $(N-1)\Delta$ away from one of the diagonal elements (all of which are unity). The smallest possible singular value is then $1-(N-1)\Delta$, with all the other singular values being $1+\Delta$ [27]. For the 2×2 system, finding the phase which achieves the best and worst case singular values is easy to do analytically.

For larger matrices, however, it is not as clear what the optimal crosstalk phases are. For a $N \times N$ matrix, there are $N(N-1)$ off-diagonal crosstalk elements which can have different phases, making a brute force search difficult. The eigenvalues (and thus the singular values) of a matrix are continuous functions of the matrix elements, so small perturbations in the element values will lead to small variations in the eigenvalues [27]. This reduces the complexity of a brute force search for the best and worst case crosstalk matrices for small values of N. For example, for a 4×4 crosstalk matrix, the phases which achieve the best and worst case crosstalk are found from a brute force search to be

$$\mathbf{C}_{\text{best}} = \begin{bmatrix} 1 & \Delta & \Delta & \Delta \\ \Delta & 1 & -\Delta & -\Delta \\ \Delta & -\Delta & 1 & -\Delta \\ \Delta & -\Delta & -\Delta & 1 \end{bmatrix} \tag{5.74}$$

and

$$\mathbf{C}_{\text{worst}} = \begin{bmatrix} 1 & \Delta & \Delta & -\Delta \\ \Delta & 1 & -\Delta & \Delta \\ \Delta & -\Delta & 1 & \Delta \\ -\Delta & \Delta & \Delta & 1 \end{bmatrix}. \tag{5.75}$$

The passivity constant α is given by the magnitude of the largest possible singular value, which is $1+(N-1)\Delta$. Unlike the best and worst case matrices, this can be

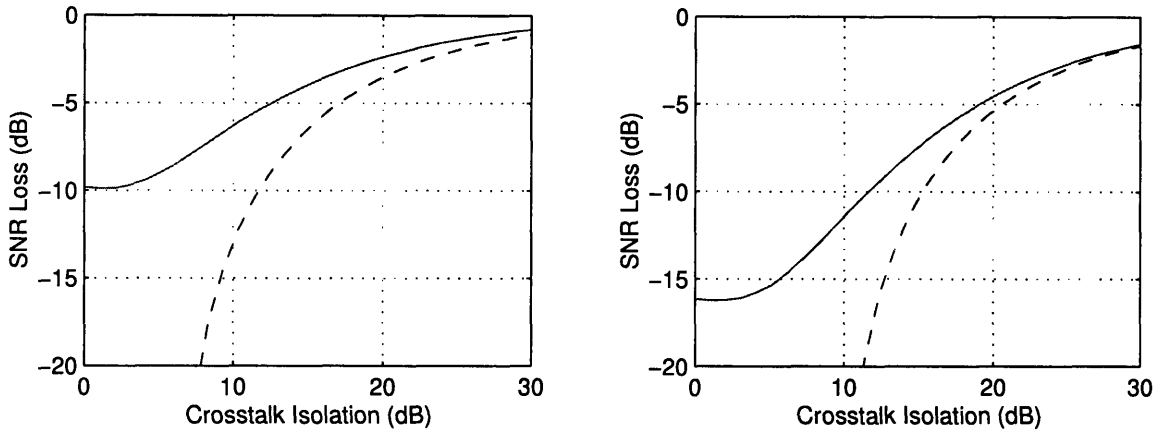


Figure 5-12: Worst case (dashed) and randomized (solid) SNR loss for 3×3 (left) and 4×4 (right) crosstalk matrices.

easily found for any size matrix. If the phases of the crosstalk matrix are all zero, then the crosstalk matrix

$$\mathbf{C}_{\max} = \begin{bmatrix} 1 & \Delta & \cdots & \Delta \\ \Delta & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \Delta & \cdots & \cdots & 1 \end{bmatrix} \tag{5.76}$$

has as one of its eigenvectors $\mathbf{x} = [1\,1 \cdots 1]^{\mathrm{T}}$ with an associated eigenvalue (and singular value) of maximum magnitude. Although a plot of the SNR loss as a function of crosstalk phase is difficult because of six phase variables, we can plot the difference in SNR loss between the worst case and after phase randomization. Figure 5-12 shows the average SNR loss for a 3×3 and 4×4 system, both with the worst case crosstalk matrix and randomized phase. Just as with the 1×2 system, phase randomization is not helpful when the crosstalk isolation is high, and the difference in SNR loss grows without bound as the crosstalk isolation falls to zero. Even though the maximum SNR loss is greater for the larger number of antennas, the phase randomization is also effective over a larger range of crosstalk isolations.

5.5 How to Achieve Phase Randomization

The task of achieving the phase randomization in an actual circuit can be a difficult one, because the parameters of the crosstalk matrix are determined by the circuit

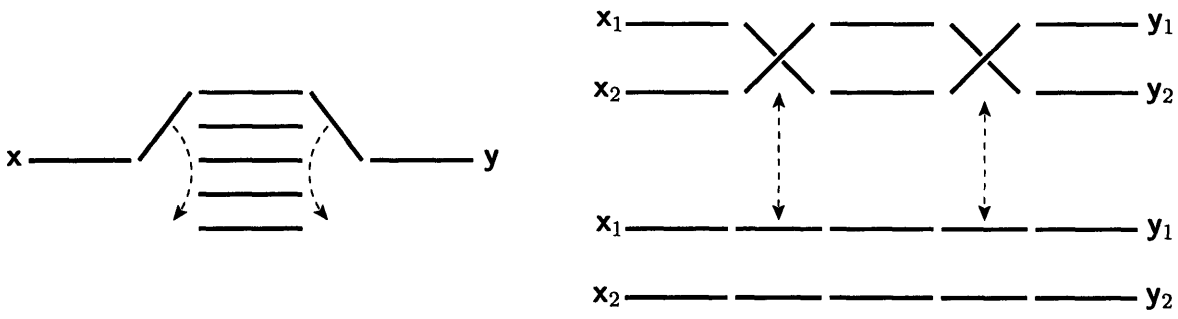


Figure 5-13: Two possible hardware methods of adjusting the phase.

layout as well as the variations that can occur during chip production. Altering these parameters requires physical changes to the circuit paths, or the use of alternate circuit paths. For example, Figure 5-13 shows two possible ways to alter the crosstalk on a given data path. The first way is to switch between several equivalent data paths. Since each data path will have a slightly different electrical environment, the crosstalk coupling it will experience will be different from other paths. The redundant circuitry will have an associated area cost, and if not carefully designed might actually decrease the crosstalk isolation. A second option is to switch between pairs of signal paths, so there is a minimum of extra circuitry. Both of these methods require analog switches, which can be difficult to make and costly in terms of circuit area. Ideally, the phase could be changed without using any additional hardware.

Adjusting the relative phases of the input constellations is not helpful for phase randomization because although it might prevent certain constellation point pairs from destructively interfering, it will allow others to do so. In order to adjust the orientation of the constellation depending on which points are chosen (so that they are always at the right phase), the phase rotations must be communicated to the receiver would be required, otherwise it cannot properly decode. Alternately, certain input pairs could be forbidden from being transmitted and known to both the receiver and transmitter. In addition to reducing the transmission rate, however, this also requires knowledge of the actual value of the crosstalk, which is again difficult and often unavailable.

While crosstalk "diversity" is not easily available through different hardware paths or by altering the input signals, the frequency dependent nature of the crosstalk can be used to achieve the save effect. Due to symmetry, only 180° of phase shift is necessary for the crosstalk to experience all the possible singular values. The amount of bandwidth needed to span the full range of phases may be more than is available within the system bandwidth, but an ultrawideband system, for example, should have enough bandwidth to achieve phase randomization.

Even partial phase randomization will have a beneficial effect, just as having two looks at the wireless channel is significantly better (in terms of diversity gain) than one. For example, a multicarrier system can utilize its entire band through frequency hopping for the purpose of randomizing the crosstalk phase and avoiding deep nulls that may appear due to destructive interference. Although there might not be enough bandwidth for the phase to be fully randomized, even partial phase randomization can reduce the worst case effects to a more manageable level.

5.6 Circuit Crosstalk Measurements

In order to test real circuit crosstalk conditions, the crosstalk isolation was measured on a test chip with four power amplifiers [47]. Figure 5-14 shows a die photo of the PA test chip which has four power amplifiers, one at each corner. No special attention was paid to isolating the amplifiers from one another, aside from placing them at the four corners of the chip to physically space them apart. The majority of the chip area is empty space, with filler structures placed to meet processing conditions. The crosstalk between amplifier inputs, amplifier outputs, and between one amplifier output and another amplifier's input were measured for both adjacent amplifiers (such as #1 and #2) as well as opposing amplifiers (such as #2 and #3). The main function of the PA test chip is to test the feedback crosstalk model, as the circuit layout is not particularly realistic. Similar crosstalk measurements were made for a LNA test chip over a wide bandwidth, with the input to output coupling including some matching networks and filters to simulate more realistic circuit conditions [33]. The total amount of isolation between the two ports is measured instead of only the amplifier due to the difficulty in probing traces on the chip die.

5.6.1 Frequency Nulls from Crosstalk Feedback

For the PA test chip, test signals were injected into the amplifiers to determine whether the crosstalk isolation at any point is low enough to create a significant loss in signal strength. For the test apparatus, a sine wave is passed through each of two amplifier inputs. A variable phase shifter and attenuator alters the input to one of the amplifiers to simulate the second signal. As an example, Figure 5-15 plots the effect of the two signals of the same frequency passing through two amplifiers on the chip for a range of frequencies. The attenuator and phase shifter values were held constant as the frequency was swept. The left plot of Figure 5-15 shows how the crosstalk isolation varies with frequency for three different settings of the variable phase shifter. A crosstalk power of -10 dB means that the power of the coupled crosstalk signal is 10 dB less than the power of the original signal. The crosstalk

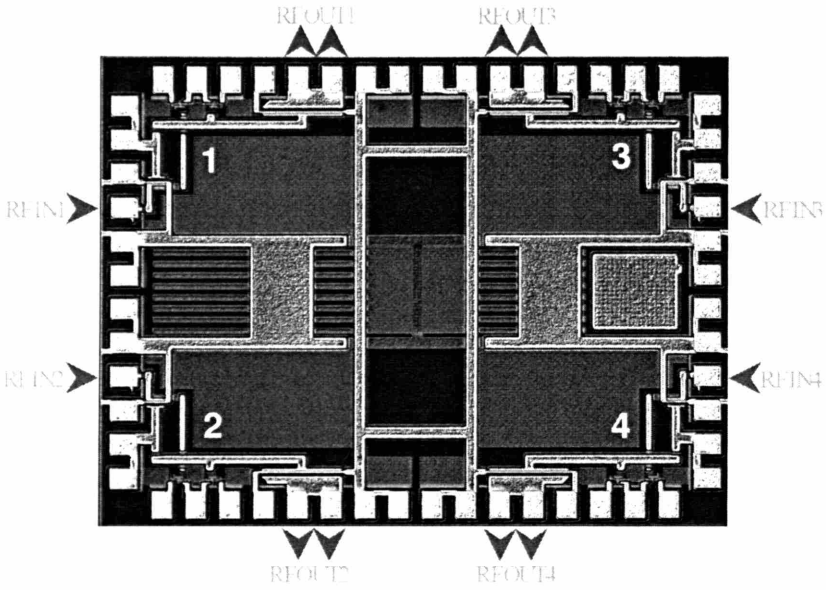


Figure 5-14: Die photo of PA test chip. (Chip design by A. Pham)

isolation is then defined as the inverse of the crosstalk power.

The thick solid line is the inverse gain of the PA, plotted on the same scale for comparison. Where the PA gain line and the crosstalk line cross, the coupled crosstalk signal and the desired signal will have the same amplitude. The variable phase shifter is essentially a tapped delay line, so the resulting phase shift is frequency dependent. As the right plot of Figure 5-15 shows, the particular value of the phase shift as well as the magnitudes of the crosstalk and the amplifier gain will affect where destructive interference will occur. The destructive interference can reduce the SNR of the region near the notch by more than 20 dB in some cases.

Looking more closely at the waveforms, the measured frequency response can be compared what the theoretical response should be using the measured crosstalk data. In Figure 5-16 two amplifier output waveforms are tested. In the left plot, the notch caused by the destructive interference is very deep, indicating that the amplitude match between the desired input signal and the coupled crosstalk signal is very close. Using the measured crosstalk amplitude and phase, we do not see the deep notch appear in the frequency response. Using only the phase of the crosstalk (and replacing the magnitude with a constant value), however, produces a very similar frequency response to the one that was measured. On the right plot, the fit is not as good, with the phase only information still producing a notch, but not in the same location or with the same shape as the measured notch. It may be the case that the measured amplitude information is not as reliable as the phase information.

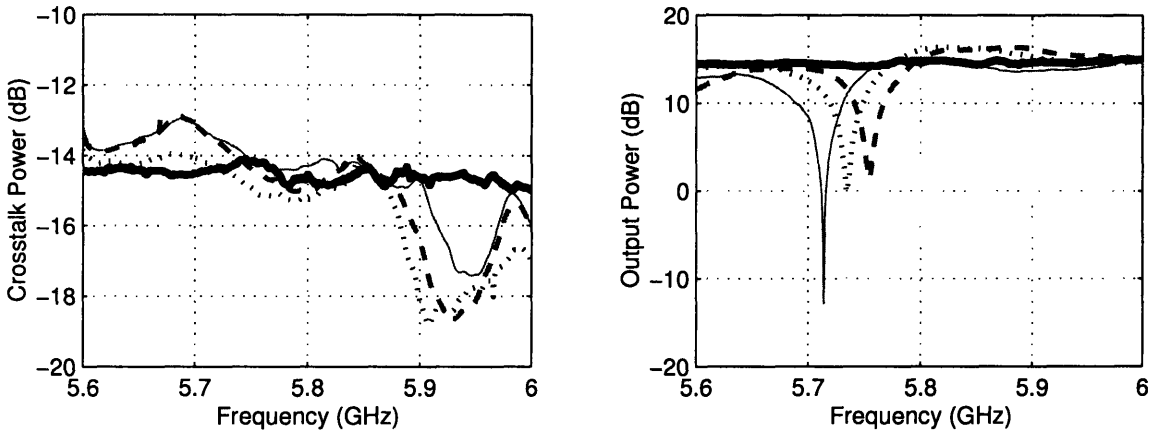


Figure 5-15: PA test chip crosstalk measured for different phase shifts (left). The thick solid line is inverse of amplifier gain for comparison. Depending on the phase, deep notches can appear at random places where the crosstalk is unfavorable (right).

The measured and simulated frequency response do both generally show the same behavior, however.

It is clear that the crosstalk between multiple antennas can severely degrade the SNR if the gain of the amplifiers is on the order of the amount of crosstalk isolation between amplifiers. Since the inputs to the two amplifiers can be different amplitudes, even more crosstalk isolation is required to prevent the larger signal from interfering with the smaller one. For example, if 64-QAM constellations are being used, the difference in power between the smallest and largest power constellation point is 16.9 dB. If the amplifier gain is 10 dB, then the crosstalk isolation must be greater than 26.9 dB to avoid destructive interference.

5.6.2 Wideband RX Circuit Crosstalk Measurements

While the PA test chip was able to demonstrate the potentially destructive effects of crosstalk, the circuits on the test chip are not representative of real working circuitry. In addition, crosstalk measurements for the PA test chip where made only over a narrow frequency band corresponding to the WiGLAN. A similar test chip had previously been constructed to test part of the receiver front ends as a predecessor to the RX test chip in Chapter 4, allowing crosstalk isolation measurements for more realistic circuit conditions [33]. The measurements were made between input of the LNA, and the output of the image reject filter. Measurements were also made over a much wider bandwidth, which reveal trends in the crosstalk phase.

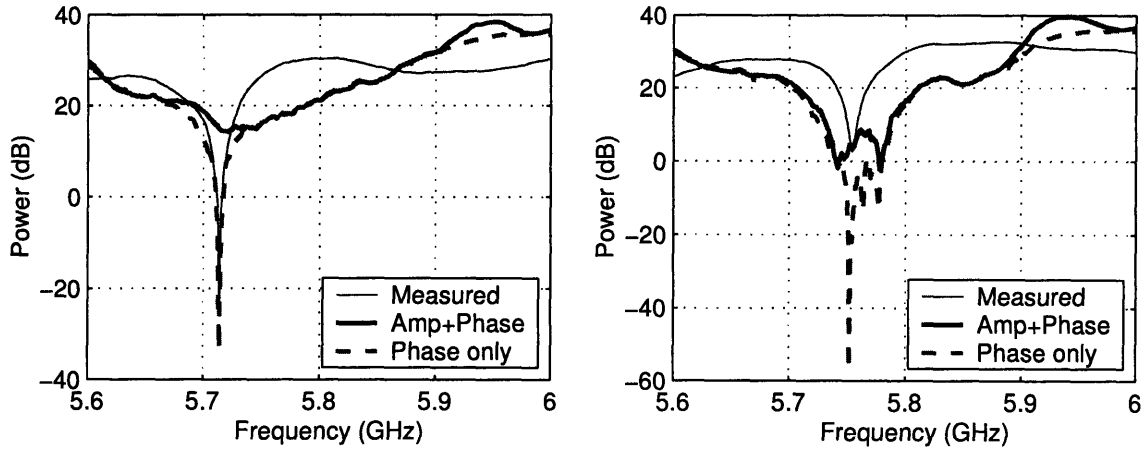


Figure 5-16: Simulated data fits using measured crosstalk amplitude and phase information compared to the measured frequency response.

Some of these crosstalk amplitudes are shown in Figure 5-17. The two dashed lines are the coupling between the output of one amplifier and the input of another for two amplifiers that are located physically closer to each other, and the solid lines are for two amplifiers that have a greater physical separation. The crosstalk power for the "close" amplifiers is about 23 dB in the 5 GHz frequency band, and up to 10 dB less for the "far" amplifiers. The crosstalk power for the "far" amplifiers appear to vary more widely, but in general are significantly less than the for the "close" circuitry. Note that the crosstalk is not symmetric due to the presence of circuitry between the measurement points.

The phase of the crosstalk is also plotted, with a generally linear phase which is characteristic of a time delay. It takes about 2 GHz of bandwidth for the crosstalk phase to vary through a full 360° (2π radians) range. In the WiGLAN frequency band, the crosstalk phase varies by about 50° ($\pi/4$ radians) within 150 MHz. The rapid changes in the phase visible between 4 GHz and 5 GHz for one of the "far" measurements are mostly likely measurement error caused by the low crosstalk power in that frequency range.

5.6.3 Effect of Partial Phase Randomization

For the test chips, the crosstalk phase was measured to have about a 50° phase shift across the 150 MHz frequency band, while a full 360° phase shift requires 2 GHz of bandwidth. Since 180° are needed for the phase randomization, this translates to a bandwidth of 1 GHz, which is over six times larger than the WiGLAN bandwidth.

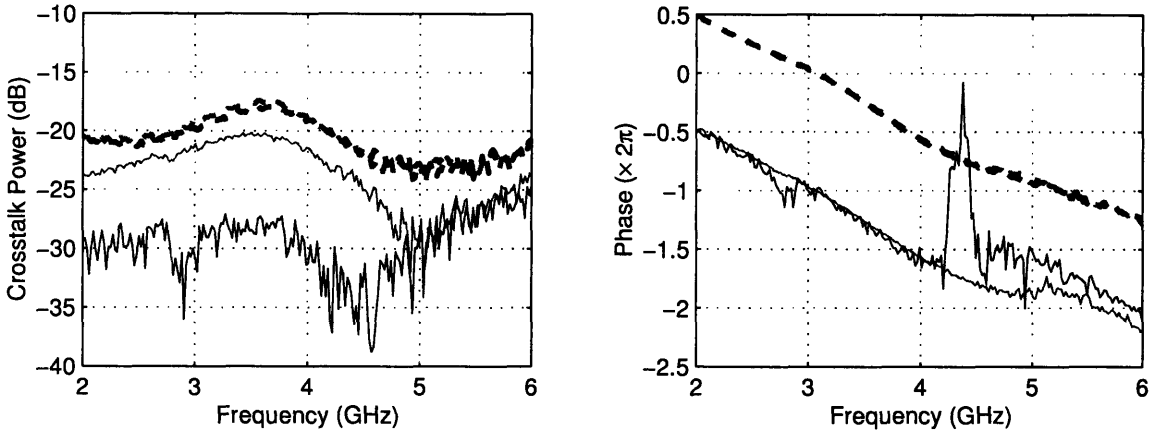


Figure 5-17: Measured crosstalk magnitude (left) and unwrapped phase (right) for "close" (dashed) and "far" (solid) circuits. The crosstalk magnitude is correlated with but not a function of distance.

Thus, for the WiGLAN, it is not possible to completely randomize the phase. For an ultrawideband system, however, the bandwidth is wide enough that the crosstalk phase can be fully randomized. There is still some benefit for partial phase randomization, although there is a penalty if not all phases are available.

Figure 5-18 shows the beneficial effects of partial phase randomization. While the full range phase randomization shown in Figure 5-11 has a worst case SNR loss of about 5 dB when the crosstalk isolation is 0 dB, the partial randomization range of 50° allowed by the WiGLAN bandwidth results in a worst case SNR loss of about 9 dB. At 5 dB isolation, the difference between the partial phase randomization and no randomization is very small. The phase randomization available for the WiGLAN bandwidth can still provide benefits when the crosstalk isolation is poor. From a different viewpoint, the difference in performance between partial and full phase randomization is not very large.

5.7 Simulation Results

The performance for a 1×2 system over a Rayleigh fading channel was simulated with MATLAB for uncoded 4-QAM transmit constellations and several values of crosstalk isolation. The resulting BER curves and are plotted in Figure 5-19. Although the input constellation only has four distinct points, the varying channel matrix $\mathbf{H}$ changes the amplitude and phase so that the phase is uniformly distributed. Each plot shows the worst case (dashed) and average (thick line) case performance for a 1×2 system

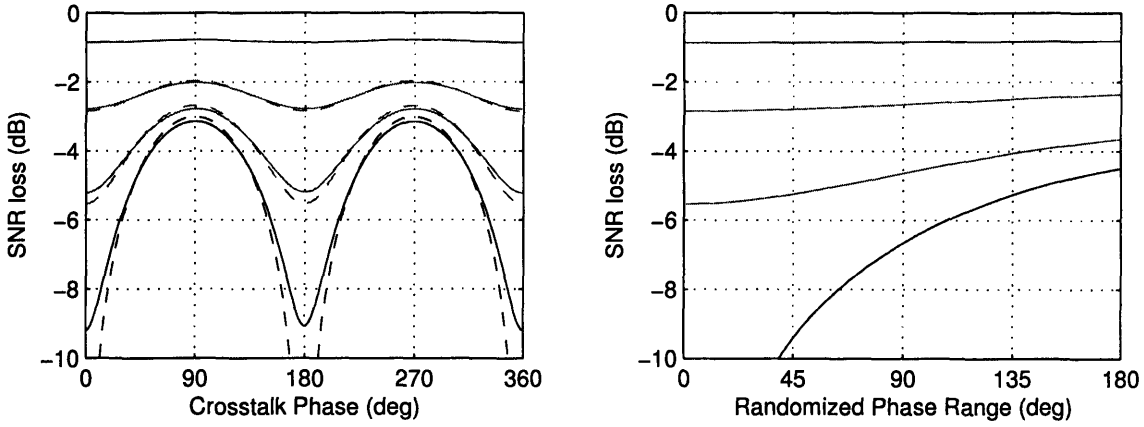


Figure 5-18: Average SNR loss for a 2×2 system for a fixed phase (dashed, left) and for randomization over the WiGLAN bandwidth (solid, left), or 50° centered on the fixed phase. Right plot shows the average SNR loss for the worst case phase (0° or 180°) with varying amounts of partial phase randomization. Crosstalk isolation is 0, 5, 10, and 20 dB from bottom to top in both plots.

with crosstalk at the receiver. The lowermost line in each plot is the BER performance for no crosstalk, showing SNR gap caused by the crosstalk passivity constant.

We see that the worst case performance is not significantly different from the averaged (phase randomized) case for crosstalk isolations above 10 dB. The larger SNR gap in performance for the 10 dB over the 20 dB case is a compelling reason to try and increase the crosstalk isolation, but this gain must be made from circuit or other optimizations, since phase randomization will only result an improvement of less than 1 dB. Note that at 0 dB, or no isolation, the worst case has a slope change due to the lost diversity, hence the asymptotically unbounded SNR loss. The random phase case also appears to have a slope change due to the averaging between the two different diversity orders of the worst case and other phases. At high enough SNR the slope of the random phase case would be -2 like the best and no crosstalk cases, however, because the asymptotic SNR loss remains bounded for the random case.

Figure 5-20 plots the best, worst, and average case performance for a 1×4 system with crosstalk at the receiver in the same way as Figure 5-19 did for the 1×2 system. There is a much larger SNR penalty for the crosstalk here, as well as a larger separation between the curves. Again, the effect of phase randomization is small if the crosstalk isolation is above 10 dB, and negligible for isolation above 20 dB.

These results also apply for crosstalk at the transmitter. If there are multiple transmit antennas and only a single receive antenna, then these results will translate

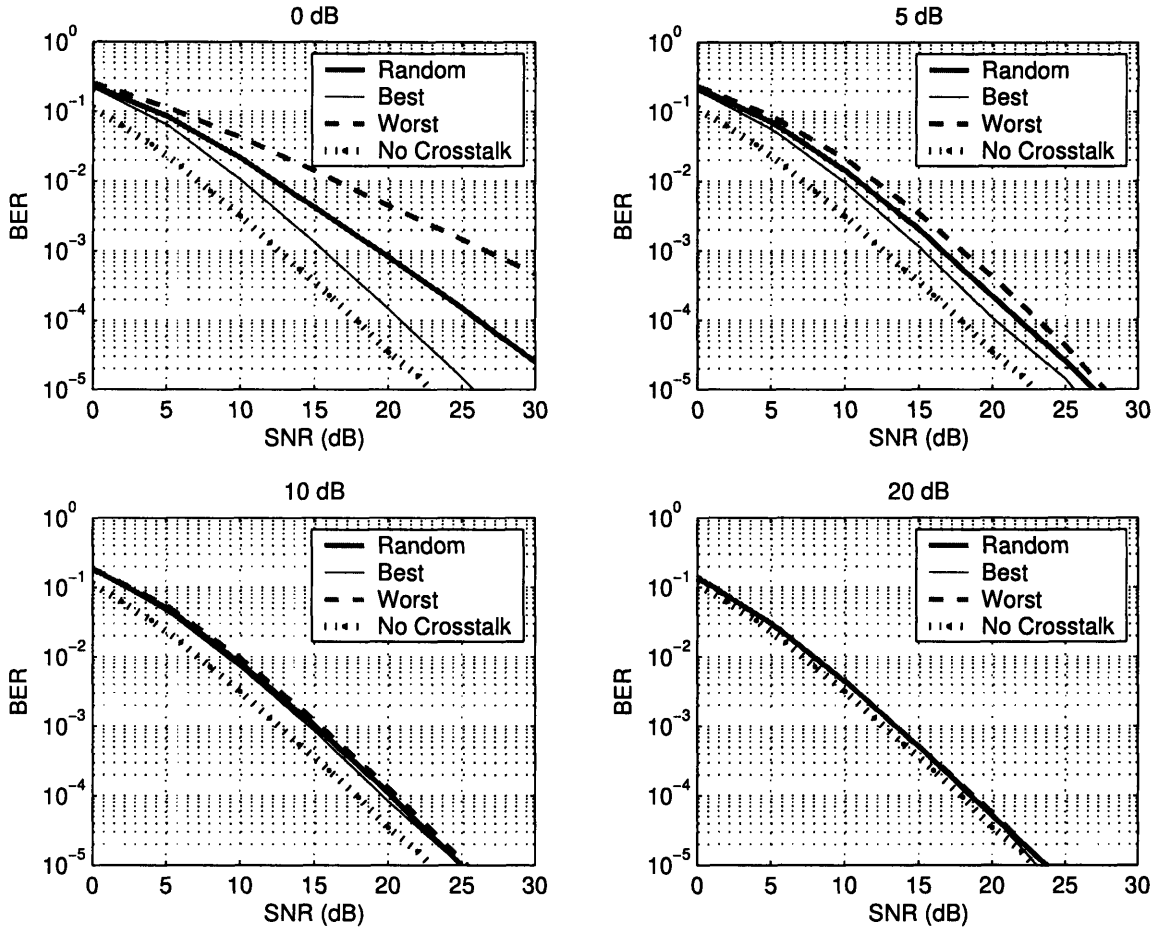


Figure 5-19: Worst (dashed), average (thick), and best (thin line) case crosstalk for several isolations for a 1×2 system. The dotted line represents no crosstalk.

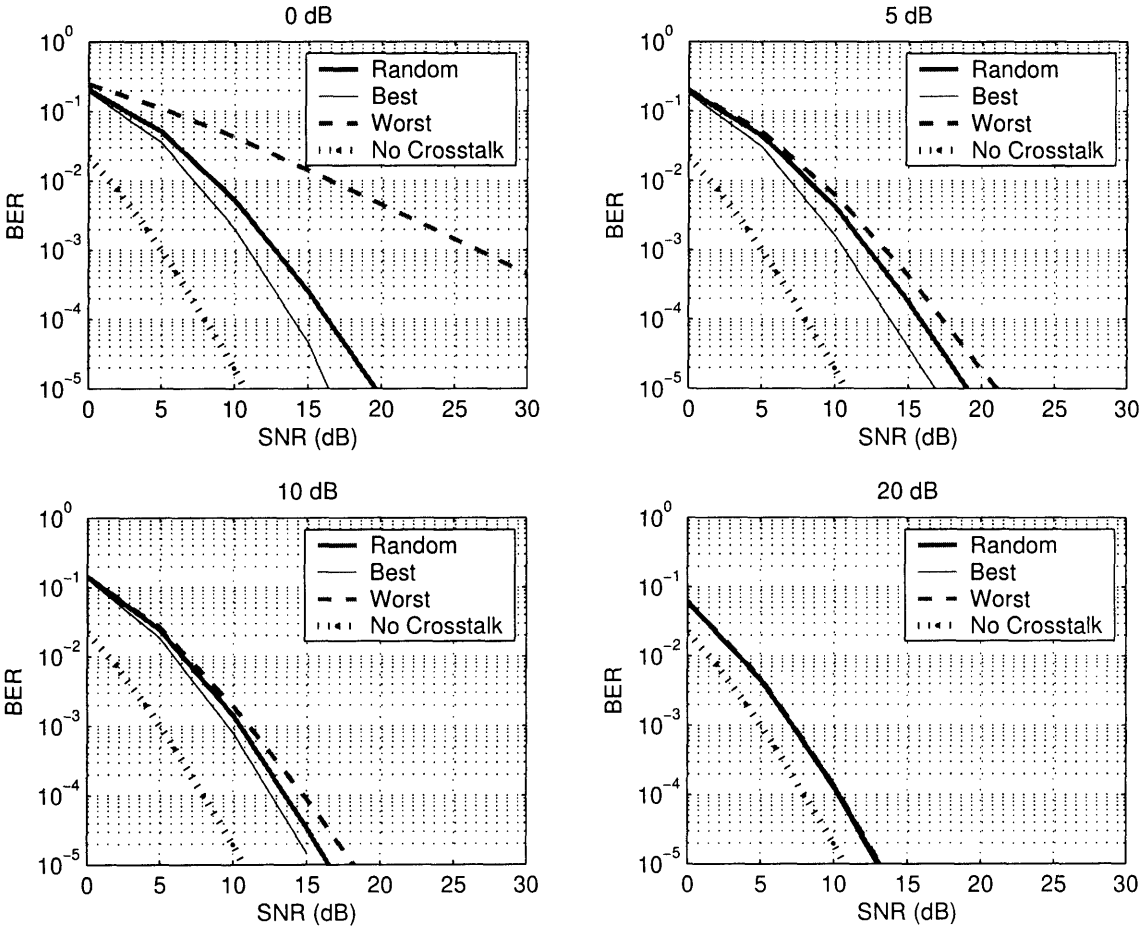


Figure 5-20: Worst (dashed), average (thick), and best case crosstalk for several isolations for a 1×4 system. The dotted line represents no crosstalk.

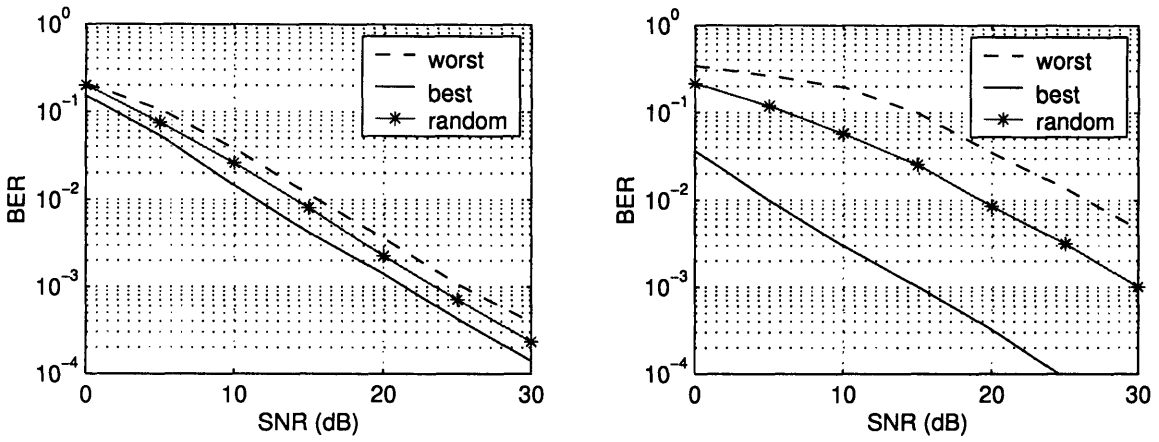


Figure 5-21: Crosstalk at transmitter for a 2×2 system using V-BLAST with 20 dB (left) and 3 dB (right) crosstalk isolation.

directly as the $N \times 1$ system with transmitter crosstalk is the dual of the $1 \times N$ system with receiver crosstalk. We can also use (5.72) for other encoding techniques. For example, Figure 5-21 shows the best, worst, and randomized behavior for a 2×2 system transmitting 4-QAM signal constellations using the VBLAST algorithm [78]. The VBLAST algorithm breaks down when the crosstalk matrix is singular, which occurs at 0 dB isolation. Thus, the poor isolation case is 3 dB rather than 0 dB in this case. The difference in performance for the 20 dB case is more noticeable due to the higher sensitivity of the VBLAST algorithm to low SNR.

5.8 Design Guidelines

The presence of multiple analog front ends on a chip present many parallel data paths whose signals may couple with each other. While digital circuits are resistant to this crosstalk, any signal that gets coupled into a signal path will decrease the SNR of that signal. The amplifiers, especially high gain power amplifiers at the transmitter, guarantee that there will be signals with very different amplitudes present at the same time on the chip. In order to prevent the destructive interference or a positive feedback loop from forming, the crosstalk isolation must be at least as large as the amplifier gain. If the input signals to the parallel amplifiers are of different strengths, then the difference in powers of these two powers must also be added to the crosstalk isolation. For example, if 64-QAM signals are sent over two independent data paths which each have a 20 dB gain amplifier, then the crosstalk isolation must be at least 37 dB to prevent the possibility of severe SNR loss.

Although this isolation rule avoids unstable feedback loops and destructive interference, it does not account for the effect that the smaller but non-negligible interference causes to the SNR of the desired signal. Using a linear crosstalk model, we can determine the worst case and average SNR loss as a function of the singular values of the crosstalk matrix. For cases of insufficient crosstalk isolation, these SNR losses have a strong dependence on the crosstalk phase. Using phase randomization allows the worst case behavior to be brought closer to the average behavior, even if there is insufficient bandwidth for full randomization. Since the crosstalk is fixed for a circuit, the improvement in the worst case performance decreases the SNR margin needed to ensure that few if any manufactured chips would not function due to a particularly bad crosstalk environment.

The crosstalk model shows that if the crosstalk isolation is larger than 20 dB, the phase randomization techniques have no effect and the SNR loss is less than 1 dB. However, adding an extra 20 dB to the crosstalk isolation now requires the hypothetical system above to have crosstalk isolation of greater than 57 dB. The phase randomization, by raising the worst case performance loss caused by crosstalk, can add more than 5 dB to the isolation. Test measurements have shown that the crosstalk magnitude varies little within the WiGLAN signal bandwidth, but the phase does vary significantly. It is therefore beneficial for all communications to experience all available frequencies in the signal band, either by using all of the bandwidth or by hopping within the signal band.

Some of this frequency hopping will occur automatically if a frequency bin is measured to have a low SNRduring node bandwidth allocation, but this will only happen if the training sequences happen to have the right phases across the parallel RF chains. Since the SNR loss depends not only on the crosstalk phase and magnitude, but also he relative phases of the input signals, the SNR loss seen by the training sequences are not good indicators of whether or not the actual input signals will experience large SNR losses.

Chapter 6

Peak to Average Power Ratio

From a communications standpoint, the goal of the transmitter is to maximize the reliable data rate for a given transmit power, or equivalently, the energy required per transmitted bit. An amplifier will saturate if the desired instantaneous output value is too high, however, so the peak output power is also a important constraint to consider. Using a multicarrier modulation scheme like OFDM tends to produce output signals with a high Peak to Average Power Ratio (PAPR). A high PAPR negatively affects amplifier efficiency, however, because the average signal power must be scaled so that the signal peaks do not cause the amplifier to saturate. This directly affects the SNR since the transmitter is forced to lower the signal power to accommodate the peaks. A high PAPR also requires high linearity over a large dynamic range for the amplifier. As we have seen, a highly linear amplifier tends to be less efficient than a nonlinear amplifier, which can significantly impact the total power consumption.

An output signal that has a relatively low PAPR is desirable not only because of the lowered demands on linearity and dynamic range, but also because of the increase in efficiency afforded by allowing nonlinear amplifiers to be used. This could allow a higher average transmit power to be used while still reducing the total power consumption. In the extreme case, a constant amplitude output signal has a PAPR of 0 dB, which would allow an extremely efficient nonlinear amplifier to be used without distortion. The increase in efficiency translates into less total energy used for data transmission, thus we can use PAPR reduction as a proxy for decreasing the energy per bit for transmission.

In this chapter, we describe two algorithms to reduce the PAPR for OFDM systems, which in turn increase the efficiency of the transmitter power amplifier. Neither of these algorithms require side information to be sent to the receiver, and can be easily added to an existing system without seriously disturbing any coding already present. The precoding algorithm can in fact be added transparently to an existing system, without any other parts of the system needing any knowledge of its presence.

Section 6.1 presents some existing algorithms for PAPR reduction, and details some of their shortcomings. Sections 6.2 and 6.3 present the contradictory problems of lowering the clipping probability of OFDM symbols and raising the low efficiency of linear amplifiers, respectively. As Section 6.4 shows, clipping an OFDM symbol is problematic due more so to bandwidth expansion rather than a loss in SNR.

Section 6.5 presents the PAPR precoding algorithm, which is suited for use in the high SNR regime. Using precoding similar to that used for the downlink transmissions of the WiGLAN, average power is traded for peak power. The precoding algorithm reduces the PAPR of the system by over 5 dB with only a slight increase in average power. When coupled with the outphasing amplifier, the overall energy per transmitted bit is reduced by a factor of 3.3. The rate loss for this algorithm is 1 b/s/Hz more than the minimum required for the PAPR reduction. This fixed rate loss makes this algorithm useful at high SNR, where the rate loss is tolerable.

Section 6.6 presents the phase synthesis algorithm, which achieves similar reductions in PAPR of around 5 dB with no rate penalty at low SNR, making it a good choice for low power systems. At high SNR, however, half of the rate is lost, making the synthesis algorithm unsuitable for high SNR applications.

Section 6.7 presents simulation results for these algorithms, showing the efficiency gains that can result using the two algorithms with several different types of linear amplifiers. A comparison of the two algorithms in Section 6.8 shows the suitability of the precoding algorithm for the WiGLAN, while the synthesis algorithm could be useful for very low power networks.

6.1 Existing Solutions for Reducing PAPR

There are many strategies that have been proposed to reduce the PAPR of an OFDM or other multicarrier signal. These techniques can be loosely divided into two categories, which either try to shape the signal while encoding the bits, or after the encoding has been performed. Performing the PAPR reduction while encoding has the potential of greatly decreasing the PAPR of the resulting OFDM time signal, but designing good codes that have low PAPRs is difficult without an excessive reduction in the achievable data rate. In addition, these low PAPR codes may have undesirable properties, such as poor resistance to errors or high complexity. Altering the OFDM signal after encoding typically leads to more modest reduction in the PAPR. These algorithms can have relatively small losses in data rate and low complexity while retaining good error correcting performance.

Since codes have been designed to have very good error correcting properties, it seems reasonable to believe that similar codes can be designed that have both good error correcting properties as well as a low PAPR. In [46], a bound is given on the

achievable PAPR given the rate of the code and how powerful it is. Unfortunately, due to the nonlinear nature of how the choice of a constellation point in the frequency domain affects the PAPR in the time domain, there are few known ways to approach the code design. It should be possible to limit the PAPR to $\ln N$ for a length N OFDM symbol without losing any appreciable code rate [61], however deterministic codes have only been found that limit to $c \ln N$ for some constant c [46, 59]. In [59], $c = 8$, which is a 9 dB increase from the theoretical limit. These codes still require a small amount of code loss, however. Complementary Golay codes have been found to limit the PAPR to only 3 dB [52, 12], but the code rate falls to zero rapidly as the length of the OFDM symbol increases, and has already fallen to one-fourth or less when $N = 32$ [12].

Because of the difficulty of designing low PAPR codes, most proposed algorithms take the already encoded OFDM symbol and then use various techniques to reduce the PAPR on a per-symbol basis. A simple method of reducing the PAPR is to clip the amplitude of the peaks and then bandpass filter to prevent the signal from leaking outside of its bandwidth, effectively introducing noise into the original waveform [37]. Clipping the signal is a nonlinear operation equivalent to a hard saturation of the amplifier. Other solutions use a nonlinear transformation such as companding [29] to try and reduce the dynamic range of the time signal.

Clipping or distorting the OFDM time waveform can be equivalently viewed as adding a carefully chosen bandlimited signal which has peaks in the same location as the original signal, but with opposite phase such that the sum is small. For example, a specially designed template signal can be used which is centered around the largest peak of the OFDM symbol[16]. This sort of solution not only requires determining the location of the largest peak, but also must somehow transmit this information to the receiver. A similar strategy inserts offsets into selected frequency bins to try and reduce the PAPR [32], but again, this requires sharing the information with the receiver. One of these methods, called Tone Insertion [68], allows the input constellation to be offset different amounts on a per-frequency bin-basis, effectively adding complex sinusoidal time signals to reduce the PAPR. This is in a sense a generalization of the transmit precoding approach that will be described in Section 6.5.

Rather than clipping or distorting the signal, dummy bits can be added in unused frequency bins with values chosen to reduce the peak power of the symbol [19, 67]. Alternately, the coding scheme can be specifically chosen for the reduction of the PAPR by, for example, "spreading" the bits across all of the frequency bins with specifically chosen phases [76]. Since these coding schemes are specifically designed to reduce the PAPR and do not add error correcting redundancy for the input bits, it is usually necessary to encode the data stream with an error correcting code before using these PAPR reducing coding schemes.

All of these methods are able to reduce the PAPR by several decibels and require

some (possibly negligible) loss in the code rate. The problem with these existing solutions is either they greatly reduce PAPR at the expense of losing almost all of the data rate, or they can be computationally expensive and produce more modest PAPR reductions. Many require side information to be transmitted to the receiver for decoding, which impacts the code rate and also increases the work required by the receiver. Given the uplink/downlink transmission protocol for WiGLAN communication, the goal is to offload as much of the computational power onto the central server and away from the mobile nodes. We explore the restrictions on code rate for PAPR reduction, as well as some deterministic ways to reduce the PAPR which provide good performance without requiring any extra information to be transmitted to the receiver. Both the precoding and the synthesis algorithm we propose do not require any extra information to be sent to the receiver, and the precoding algorithm is fully compatible with an existing encoding scheme without affecting the data rate.

6.2 PAPR Distribution and Clipping Probability for OFDM Symbols

Before trying to reduce the PAPR via algorithmic techniques, we first look at the distribution of PAPR values for OFDM signals for both the discrete- and continuous-time cases. This allows us to determine how often clipping occurs for a given PAPR threshold and how much rate must be used to reduce the PAPR below that threshold.

To determine the distribution of PAPR values for an OFDM symbol, we begin with the discrete-time frequency samples which correspond to the values in each frequency bin. The amplitudes and phases of the frequency bins of an OFDM constellation are chosen from an M-QAM constellation. Each time sample of the OFDM symbol is

$$x[n] = \sum_{k=0}^{N-1} X(f_k) e^{j2\pi kn/N}, \tag{6.1}$$

where $x[n]$ is the value of the nth time sample, and $X(f_k)$ is the value of the kth frequency bin of the OFDM symbol. The values of $X(\cdot)$ are the amplitude and phase values from the input constellation for that bin. We can model these values as independent random variables, each of which has a finite variance due to the finite constellation size. The values of $x[n]$ will then approach a zero mean, circularly-symmetric complex Gaussian distribution by the Central Limit Theorem.

The magnitude r of a zero-mean complex Gaussian random variable x is Rayleigh distributed with

$$p_r(r) = \frac{r}{\sigma^2} e^{-\frac{r^2}{2\sigma^2}}, \tag{6.2}$$

where $2\sigma^2$ is the variance of x. To determine the clipping probability, we note that an OFDM symbol will not be clipped if all of its time samples are less than the clipping threshold ν. The probability that the magnitude of a single sample is greater than the clipping threshold ν is

$$\Pr(r > \nu) = \int_{\nu}^{\infty} p_r(r)\, dr = e^{-\frac{\nu^2}{2\sigma^2}}. \tag{6.3}$$

The probability of a clip, or that at least one of the time samples of a OFDM symbol of length N is greater than the clipping threshold is therefore

$$\begin{aligned} \Pr(\text{clip}) &= \Pr(\text{PAPR} > \text{PAPR}_o) \\ &= 1 - \prod_{i=1}^{N} \Pr(||x[i]||^2 < \nu) \\ &= 1 - (1 - \Pr(r > \nu))^N \\ &= 1 - \left(1 - e^{-\frac{\nu^2}{2\sigma^2}}\right)^N. \end{aligned} \tag{6.4}$$

The largest amplitude is ν, making the peak power ν^2, and the average power is the variance of $x[n]$. Thus the peak to average power is the exponent of the exponential of (6.4), therefore

$$\Pr(\text{PAPR} > \text{PAPR}_o) = 1 - \left(1 - e^{-\text{PAPR}}\right)^N \tag{6.5}$$

is the Complimentary Cumulative Distribution Function (CCDF) of the PAPR of a discrete-time OFDM symbol.

Figure 6-1 plots the CCDF for the PAPR for an OFDM symbol. For length 128 symbols as would be used in the WiGLAN, we see that the PAPR will be greater than 6 dB with very high probability, but an OFDM symbol with a PAPR greater than 10 dB will occur less than 1% of the time. The type of modulation used does not affect the peak distribution, as the plots for 4-QAM and 64-QAM follow the theoretical plot closely, except at very low probabilities, which is mostly a result of insufficient data points. Note that the PAPR of the actual OFDM symbol is limited to N, the number of frequency bins, but the Rayleigh distribution has a nonzero probability for all positive values.

These are discrete-time results, however, and the analog amplifiers have as inputs the bandlimited interpolation of these signals. The left plot also shows the PAPR of the analog signals (approximated here by $10\times$ oversampling of the discrete-time signals), which have a slightly higher PAPR than the theoretical values suggest. However, the PAPR of the upsampled time signals have a distribution very close to the

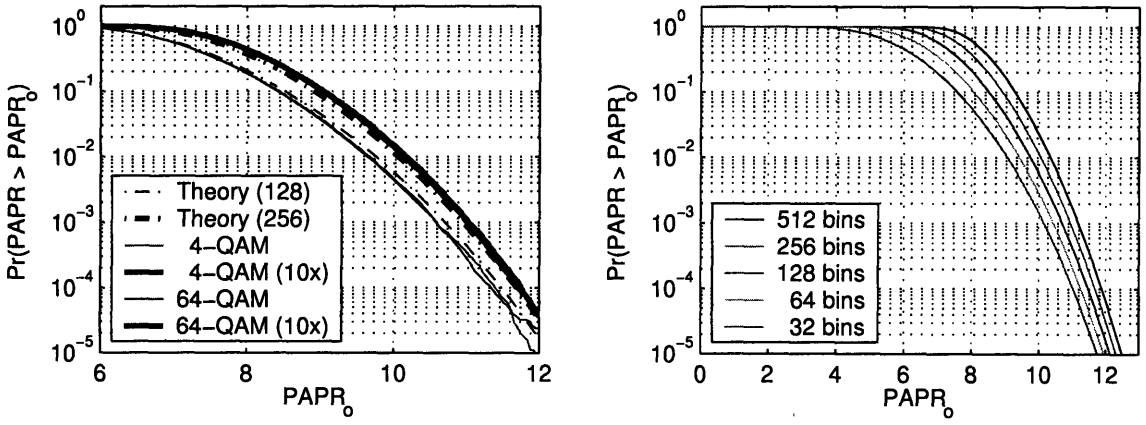


Figure 6-1: PAPR distribution for a 128 channel discrete-time OFDM symbol (thin lines) and the bandlimited interpolation (thick lines) of the digital signal (left). Right plot has theoretical CCDF's for several OFDM lengths (left to right, 32, 64, 128, 256, and 512).

theoretical plot for the PAPR of an OFDM symbol with twice as many frequency bins. The difference between the distributions of the analog signal and the sampled signal is only about 0.5 dB, however, and using twice as many frequency bins is a very good approximation to the analog signal distribution. The right plot shows how the average PAPR increases as the length of the OFDM symbol increases. For larger number of frequency bins, the curve drops off more sharply, implying that the probability density of PAPR values becomes more clustered around the mean value. In fact, as the number of frequency bins increases, the PAPR of a randomly selected OFDM symbol asymptotically approaches $\ln N$ with probability one [61].

One implication of this is that to reduce the PAPR to be less than $\ln N$ as $N \to \infty$ would require a significant drop in the code rate, while allowing a maximum PAPR of $\ln N$ requires a negligible loss in the code rate. However, the specific combinations of constellation points that will create the OFDM symbols with high peaks are not easily avoided or predicted due to the nonlinear relationship between the frequency- and time-domain signals. Coding strategies do exist which will keep the PAPR no larger than $c \ln N$ for a constant c [59], but for small values such as $N = 128$ this corresponds to a PAPR limit of almost 16 dB. It is also possible to have a code that produces OFDM symbols with less than some constant PAPR with a rate loss of $\log_q 2$ for a q-ary constellation size [60], though how to construct such a code without a brute force search is not known. It should also be noted that although the time samples of the OFDM symbol approach Gaussian random variables in distribution, the actual analog time signal will not be a Gaussian random process [58], but the difference between the Gaussian approximation and the actual waveform statistics do

not significantly affect our results.

From the shape of the CCDF for PAPR values, we can also make inferences on the amount of rate that must be used to limit the maximum PAPR. For example, from the left plot of Figure 6-1, we see that for the analog signal produced by a 128 bin OFDM symbol, the probability of clipping is 1% at about 10 dB, and 50% at about 8 dB. A code that avoids all of the OFDM symbols which have a PAPR greater than 10 dB would avoid 1% of the possible OFDM symbols, for a code rate slightly smaller than one. A code that avoids all OFDM symbols with a PAPR greater than 8 dB then does not allow half of all possible OFDM symbols, for a rate loss of 1 b/s/Hz.

6.3 Effect of PAPR on Amplifier Efficiency

Since the amplitude and phase of each frequency bin of an OFDM symbol are independently chosen, the corresponding time waveform can be very peaky. As an example of the worst case, the same constellation point is chosen in all the frequency bins. The time waveform will have a single peak at the first sample, and all other samples will be zero. Assuming N subcarriers, that gives a peak power proportional to N^2, and an average power proportional to N, resulting in a PAPR that is proportional to N. For a typical number of subcarriers, such a large peak to average power ratio can greatly reduce the average efficiency of the transmitter power amplifiers.

6.3.1 Some Example OFDM Symbols

For example, Figure 6-2 shows three sample OFDM symbols each comprised of 128 frequency bins using points from a 64-QAM constellation. The amplitudes are plotted relative to the RMS amplitude, so that the average power level corresponds to a relative amplitude of one. The middle plot shows a "typical" OFDM symbol, with a PAPR near the average value of about 7.4 dB. The entire signal is very peaky, with several amplitude peaks that stand out. The highest peak has a relative amplitude of about 2.34, which corresponds to a peak power that is about 5.5 times the average power, resulting in a PAPR of about 7.4 dB.

The left plot represents a "good" case, in which the PAPR is almost 3 dB less than the average value. Here we see that the peaks are almost uniform in height. The right plot shows a "bad" OFDM symbol, with a single very large peak. Clearly in this case, the particular values of the data stream have caused a very unfavorable OFDM symbol in terms of PAPR. Although these symbols are rare, they still occur often enough that a significant SNR margin may be needed to prevent amplifier saturation. The high value of the peak power requires the input of the power amplifier to be scaled so that the peak power does not cause the amplifier to saturate. With the

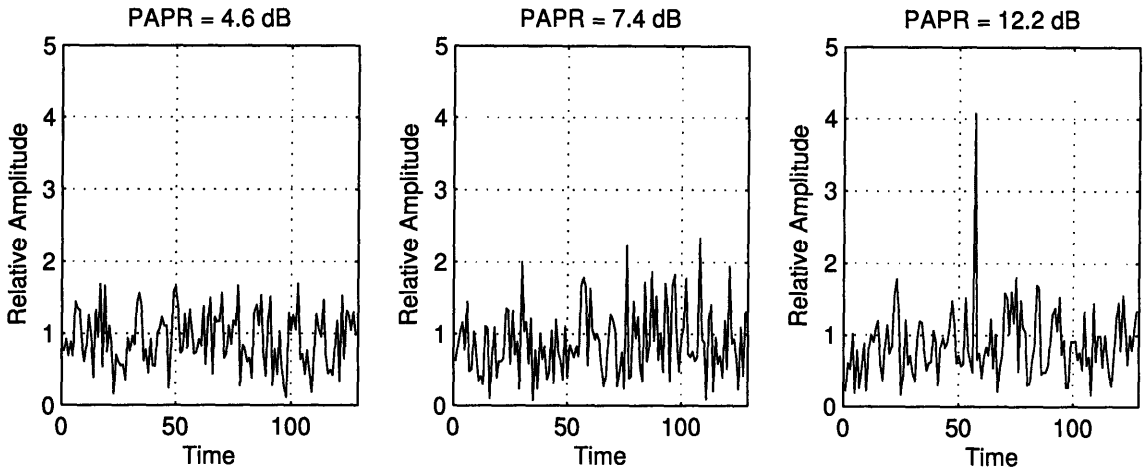


Figure 6-2: Three examples of OFDM symbols with 128 frequency bins and 64-QAM constellations scaled relative to the RMS amplitude. Depicted are a "typical" OFDM symbol (middle), as well as unusually "good" (left) and "bad" (right) symbols.

amplitude scaled down, most of the time the power amplifier is operating far from saturation. As Figure 2-13 shows, the efficiency of an amplifier is much lower when it is not near its maximum output value. We now explore how the averaged efficiency of an amplifier is affected by the maximum PAPR it is able to amplify without clipping of the output waveform.

6.3.2 Instantaneous Amplifier Efficiency Curves

The instantaneous efficiency of a power amplifier depends strongly on both the input level as well as the type of amplifier used. Figure 2-13 showed the theoretical efficiency curves of both linear amplifiers and a nonlinear class B amplifier. There are many other amplifier types, but because of the high linearity demands of OFDM symbols, we restrict the analysis to linear amplifiers only. As described in [48], the efficiency of a class A amplifier can be increased by adaptively changing the saturation level to follow the input level. We will denote this configuration as an adaptive class A (or adaptive A) amplifier. Another way to increase the efficiency is to combine the outputs of two high-efficiency constant amplitude amplifiers with a summing device to allow vector addition. Altering the relative phases of the two amplifiers allows the phase and amplitude of summed signal to be match the desired output. This mechanism is called outphasing and is described in more detail in Appendix C.

Figure 6-3 plots the efficiency curves of these three linear amplifier types with

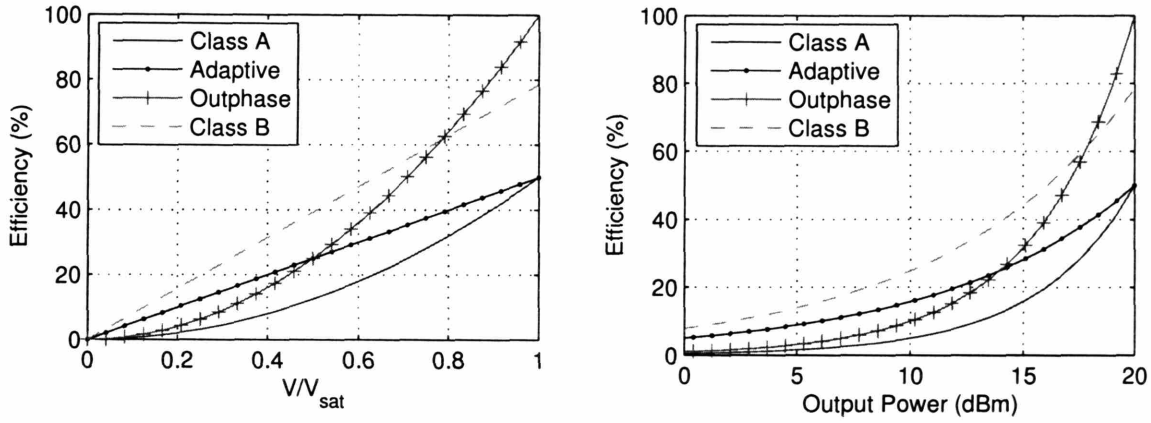


Figure 6-3: Comparison of the efficiency of various amplifier topologies as functions of the output magnitude (left) and power (right).

respect to both output magnitude normalized to the saturation voltage, and output power assuming a maximum output of 20 dBm (100 mW). The efficiency curve of the nonlinear class B amplifier is included for comparison. The class A amplifier has the worst efficiency, as expected, with the adaptive class A amplifier improving the efficiency somewhat. Because it is still a class A amplifier, the maximum efficiency is the same as a normal class A amplifier. The outphasing amplifier always has a higher efficiency than the class A amplifier, however it only has a higher efficiency than the adaptive class A amplifier for high output powers even though it uses highly efficient amplifiers as components. As the right graph of Figure 6-3 shows, the outphasing amplifier only outperforms the other linear amplifiers at output powers within about 5–6 dB of maximum. Since the class B amplifier is not considered a linear amplifier, it is not considered. As is generally the case, the nonlinear class B amplifier is more efficient than the linear amplifiers, although the outphasing amplifier is able to exceed its efficiency when very close to its maximum output power. Although the theoretical efficiencies of these amplifiers can approach 100% in theory, typical maximum amplifier efficiencies are closer to 30% for class A amplifiers, 40% for class B, and 47% for the outphasing amplifier [49].

6.3.3 Average Efficiency with Rayleigh Inputs

Since we model the OFDM time signal as a Gaussian random function (a better, but more difficult model would be a ideal bandlimited interpolation of samples of a Gaussian function), the probability density of the input magnitude to the power amplifier is Rayleigh distributed, with a probability density given in (6.2). The average power

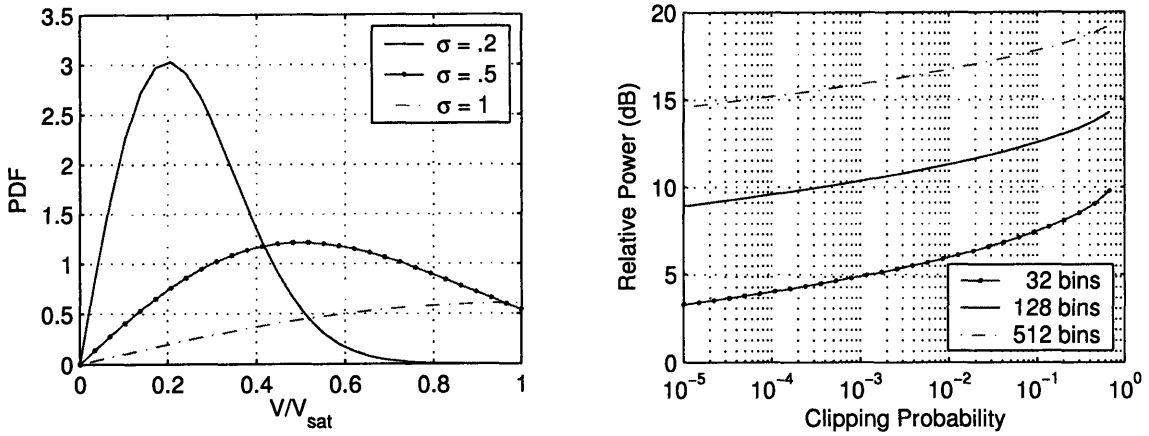


Figure 6-4: PDF of a Rayleigh distribution (left) with different average powers ($2\sigma^2$). Right shows the power savings relative to the average power corresponding to no clipping.

of the Rayleigh distribution determines how probable a magnitude above the clipping threshold will occur.

Figure 6-4 shows how raising the average power of the OFDM symbol will affect the distribution of the signal magnitudes, which are scaled to the clipping threshold, V_{sat}. We see that as the average power $2\sigma^2$ increases, the more likely the OFDM signal will be clipped. The right plot shows how much the average power can be increased given the clipping probability and the OFDM symbol length. The average power is relative to $\sigma = 0.0625$, which is equivalent to the having the clipping threshold 128 times, or 21 dB above the average power. Although given the Rayleigh distribution this would result in some clipping probability, the actual signal will never clip due to the allowable maximum PAPR of N.

The clipping probability is a strong function of the average power of the OFDM symbol. For a length of 128, doubling the average power from 9 dB to 12 dB increases the clipping probability by over three orders of magnitude. This strong dependence on the average power greatly limits the transmit power an amplifier is able to use to be much lower than the maximum power to keep from saturating too frequently. If we assume a clipping probability of 1% or less, we see that 11 dB can be saved from the worst-case SNR margin needed to prevent clipping completely. This still leaves a clipping margin of 10 dB between the average transmit power and the maximum output of the transmitter PA. Such a large clipping margin will reduce the amplifier efficiency significantly.

Given the instantaneous efficiency curves for different amplifiers (such as for the amplifiers in Figure 6-3), we can calculate the expected efficiency for a given clipping

probability for each amplifier. If the efficiency function of the amplifier is $f(r)$, where r is the normalized magnitude of the OFDM symbol such that $r \in [0, 1]$, then the expected efficiency of that amplifier is

$$\overline{\eta} = \int_0^1 f(r) p_r(r)\, dr. \tag{6.6}$$

For a generic amplifier, we can model the efficiency curve as

$$f(r) = a_0 + a_1 r + a_2 r^2. \tag{6.7}$$

This model is general enough to encompass all the amplifier types we are considering, however it can be easily extended should a more complicated efficiency curve be required. The average efficiency of an amplifier for a Rayleigh input is then

$$\overline{\eta}_R = \int_0^1 \left[a_0 + a_1 r + a_2 r^2\right] p_r(r)\, dr = a_0 r_0 + a_1 r_1 + a_2 r_2 + \eta_{\max} r_{\text{clip}}, \tag{6.8}$$

where we define the integrals

$$\begin{aligned}
r_0(\sigma) &= \int_0^1 p_r(r)\, dr = -\, e^{-\frac{r^2}{2\sigma^2}}\Big|_0^1 = 1 - e^{-\frac{1}{2\sigma^2}} && (6.9)\\
r_1(\sigma) &= \int_0^1 r p_r(r)\, dr = \left(-re^{-\frac{r^2}{2\sigma^2}} + \sqrt{\frac{\pi\sigma^2}{2}}\,\text{Erf}\left(\frac{r}{\sigma\sqrt{2}}\right)\right)\Bigg|_0^1 \\
&= \sqrt{\frac{\pi\sigma^2}{2}}\,\text{Erf}\left(\frac{1}{\sigma\sqrt{2}}\right) - e^{-\frac{1}{2\sigma^2}} && (6.10)\\
r_2(\sigma) &= \int_0^1 r^2 p_r(r)\, dr = \left(-(2\sigma^2 + r^2)e^{-\frac{r^2}{2\sigma^2}}\right)\Big|_0^1 \\
&= 2\sigma^2 - (2\sigma^2 + 1)e^{-\frac{1}{2\sigma^2}} && (6.11)\\
r_{\text{clip}}(\sigma) &= \int_1^\infty p_r(r)\, dr = 1 - r_0(\sigma) = e^{-\frac{1}{2\sigma^2}}, && (6.12)
\end{aligned}$$

with

$$\text{Erf}(x) = 1 - 2Q(x\sqrt{2}), \tag{6.13}$$

and $\eta_{\max}$ being the efficiency of the amplifier at saturation. Here, $Q(\cdot)$ is the integral of the tail of a Gaussian PDF which was defined in (5.61). The final term of (6.8) represents the clipped parts of the output signal. Although clearly the average efficiency increases as the amount of clipping represented by r_{clip} increases, the distortion caused by clipping can unacceptably decrease the signal SNR if the clipping

probability is too high.

6.3.4 Average Efficiency with Uniform Input

Since the clipping probability of the Rayleigh distribution rises strongly as the average power rises, the useful PAPR values are still quite high. As an alternative viewpoint, we can look at the average efficiency for a uniform amplitude distribution, which is given by

$$p_r(r) = \begin{cases} \dfrac{1}{1-\alpha} & \alpha \le r \le 1 \\ 0 & \text{otherwise} \end{cases}. \tag{6.14}$$

As (6.14) shows, the probability distribution of the OFDM amplitude is uniform between a lower limit α and the amplifier saturation level, which is again normalized to unity. With the uniform amplitude distribution, there is no clipping, so all PAPR values are useful. Similar to (6.8), we can define the average efficiency of an amplifier for a uniformly distributed input as

$$\overline{\eta}_U = \int_0^1 \left[a_0 + a_1 r + a_2 r^2\right] p_r(r)\, dr = a_0 u_0 + a_1 u_1 + a_2 u_2, \tag{6.15}$$

where we define the integrals

$$u_0(\alpha) = \int_\alpha^1 \frac{1}{1-\alpha}\, dr = 1 \tag{6.16}$$

$$u_1(\alpha) = \int_\alpha^1 \frac{r}{1-\alpha}\, dr = \frac{1}{1-\alpha}\frac{1-\alpha^2}{2} = \frac{1+\alpha}{2} \tag{6.17}$$

$$u_2(\alpha) = \int_\alpha^1 \frac{r^2}{1-\alpha}\, dr = \frac{1}{1-\alpha}\frac{1-\alpha^3}{2} = \frac{1+\alpha+\alpha^2}{3}. \tag{6.18}$$

The average power of the uniform distribution is $E[r^2] = u_2(\alpha)$ and the PAPR is $1/u_2(\alpha)$. Since the distribution is uniform, the peak power is never more than three times the average power (when $\alpha = 0$), so the maximum PAPR is never greater than 4.77 dB.

Table 6.1 summarizes the theoretical instantaneous and average efficiency curves for the different classes of linear and nonlinear amplifiers, with $\eta = 1$ indication 100% efficiency. Note that since the class D amplifier is a saturating amplifier, it has no average efficiency because it can only output a constant amplitude waveform and is therefore far too nonlinear to be used for an OFDM symbol. However, the outphasing amplifier combines the output of two saturating amplifiers to give a maximum efficiency of 100% but still be linear.

Type	η (instant)	$\overline{\eta_R}$ (Rayleigh)	$\overline{\eta_U}$ (Uniform)
Class A	$\frac{1}{2}r^2$	$\frac{1}{2}r_2(\sigma)$	$\frac{1+\alpha+\alpha^2}{6}$
Class B	$\frac{\pi}{4}r$	$\frac{\pi}{4}r_1(\sigma)$	$\frac{\pi(1+\alpha)}{8}$
Class D	1	—	—
Adaptive A	$\frac{1}{2}r$	$\frac{1}{2}r_1(\sigma)$	$\frac{1+\alpha}{4}$
Outphase	r^2	$r_2(\sigma)$	$\frac{1+\alpha+\alpha^2}{3}$

Table 6.1: Table of instantaneous and average efficiency for several types of amplifiers.

6.3.5 Average Efficiency Curves

Figure 6-5 plots the efficiency curves for these amplifiers for both a Rayleigh and a uniform magnitude distribution. Since the OFDM time signals have high linearity requirements, the performance of the linear amplifiers with a Rayleigh input distribution is most important. As expected, the linear amplifiers are less efficient than the class B amplifier for all ranges of PAPR for the Rayleigh distribution. Although the maximum efficiency of the outphasing amplifier is higher than the maximum for the class B, the class B has a higher efficiency over a much larger range of inputs.

Although the class A amplifier claims a theoretical maximum efficiency of 50%, the actual efficiency is often less than 10%, and with a maximum PAPR of 10 dB corresponding to a 1% clipping level, the efficiency is only 5%. The adaptive class A amplifier fares much better, but its efficiency is still less than 15% for a maximum PAPR of 10 dB or greater. If the PAPR could be limited to less than about 6 dB, then the outphasing amplifier becomes the most efficient linear amplifier, approaching an efficiency over 60%. Even though the outphasing amplifier has the highest maximum efficiency, the Rayleigh distribution emphasizes the low magnitudes, which is where the amplifier efficiency is at its lowest. If nothing is done to reduce the large amplitude "tail" of the magnitude distribution, then the expected PAPR for a 64-QAM system with 128 frequency/time samples is about 16 dB. None of the linear amplifiers are able to achieve an efficiency above 10% at this level of input backoff, with the outphasing amplifier having an efficiency of only a few percent. Significantly reducing the PAPR of the OFDM signal can greatly increase the efficiency of the amplifier by allowing the average power of the input to the amplifier to be increased. Without a method of PAPR control, the maximum PAPR must be high to prevent excessive clipping.

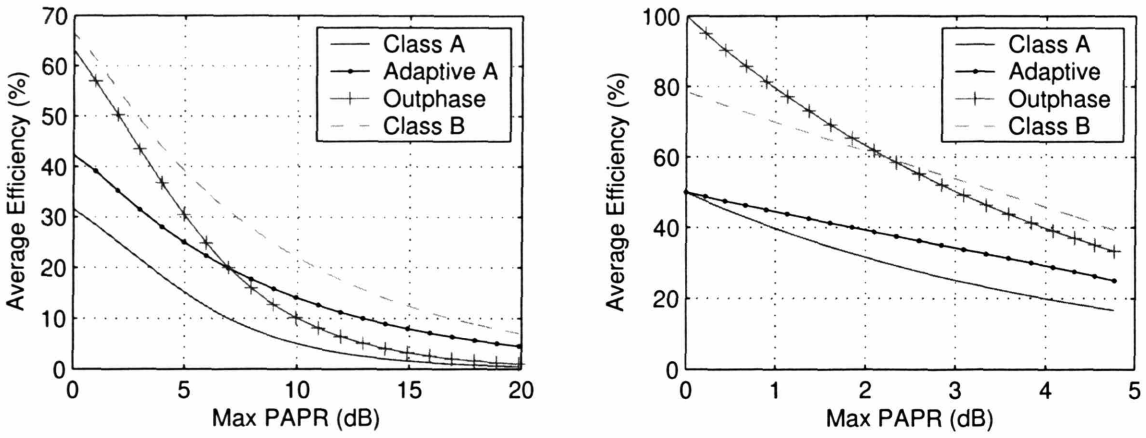


Figure 6-5: Average Efficiency for a Rayleigh distribution (left) and a uniform distribution (right) vs. maximum PAPR.

For the uniform distribution, the range of possible PAPR values is much smaller, but there is still a significant difference in the performance of the amplifiers. As the PAPR approaches zero (representing a constant magnitude output), the amplifiers reach their peak efficiencies. The outphasing amplifier has theoretical efficiencies approaching 100% when the output is constant amplitude, which even outperforms the less linear class B amplifier. In this case the outphasing amplifier has the highest efficiency of all the linear amplifiers regardless of the PAPR. Although this amplitude distribution is quite different from what an uncoded OFDM time signal would look like, it does reflect the fact that as the probability distribution of input magnitudes narrows to fewer values, the PAPR decreases and the efficiency can be quite high. Thus reducing the PAPR of the transmitted signal by changing the shape of the input probability distribution can significantly improve the overall efficiency of the amplifier.

6.4 Effects of Clipping an OFDM Symbol

As discussed in Section 6.2, the probability of clipping an OFDM symbol increases rapidly as the maximum allowed PAPR decreases. Without any algorithms for PAPR control, it is impractical to scale the input so that clipping will not occur (16 dB for $N = 128$, for example) due to the very low efficiency that results. Before trying to reduce the PAPR by algorithmic means, we first explore the effects of clipping the OFDM symbol to a lower maximum PAPR. While this increases the efficiency, it distorts the signal which can cause bandwidth expansion and reduce the effective

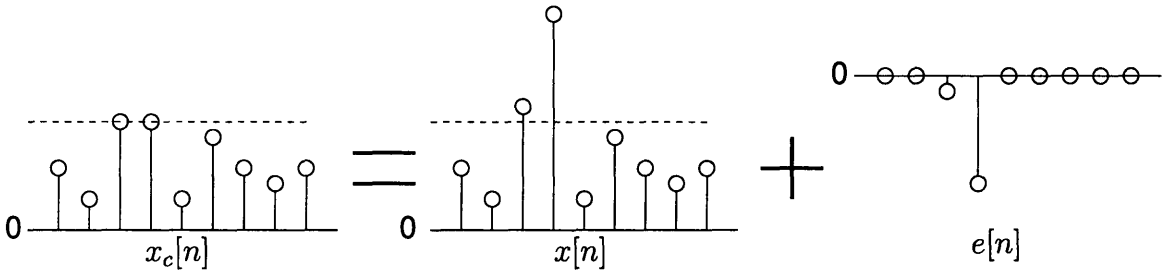


Figure 6-6: Clipping of a digital signal $x[n]$ is equivalent to adding a noise signal $e[n]$ given by the Fourier transform of the inverse signal.

SNR of the OFDM symbol. As we will see, the SNR loss from clipping is small unless the clipping probability is very high, but the bandwidth expansion caused by the nonlinear operation can exceed the allowed spectral mask even for small clipping probabilities.

6.4.1 Discrete-Time Clipping Model

When an OFDM symbol happens to have a peak that is higher than the dynamic range of the transmitter power amplifier, it will saturate and distort the output. Using a hard saturation model (equivalent to the Rapp model in (2.16) with $P \to \infty$) the clipped peak is equivalent to the inverse of the clipped peak being added to the original signal, or

$$x_c[n] = x[n] + e[n], \tag{6.19}$$

where $x[n]$ is the original signal, $e[n]$ is the error vector, and $x_c[n]$ is the resulting clipped signal. We can view this inverse peak as a noise signal, which affects every frequency bin. The clipping in the analog domain is a nonlinear function, so the bandwidth of the noise signal will in general be larger than that of the original signal, hence the clipped signal will also have increased bandwidth. As shown in Figure 6-6 for a discrete-time signal, the error signal is the inverse of the peaks that have been clipped. Since this clipping is done digitally, the bandwidth of the clipped signal is the same as that of the original signal.

Figure 6-7 shows an example of clipping a discrete-time OFDM symbol. The OFDM symbol is composed of 128 frequency bins with points chosen from a 16-QAM constellation. The input bits have been skewed slightly so that the probability of ones is greater than the probability of zeros, which favors some constellation points over others. This results in a higher PAPR by producing a larger peak at the first time sample. The magnitude of the time signal has been normalized so that the average power is unity, thus the PAPR can be read directly off the figure. In this

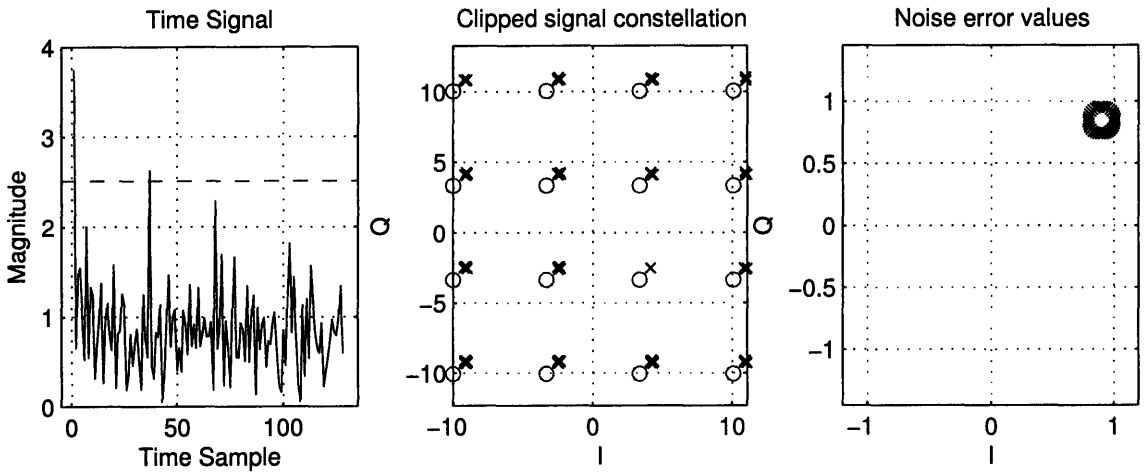


Figure 6-7: Two peaks of the discrete-time, 128 bin OFDM symbol (left,solid) are clipped to the threshold (left, dashed) level. The center plot compares the original frequency bin constellation points (o) with the clipped symbol constellation points (×). Right plot shows the constellation point offsets caused by the clipping.

case, the PAPR is around 11.5 dB. The clipping level is set at a PAPR of 8 dB, which corresponds to a discrete-time clipping probability of about 20% according to Figure 6-1.

We can see that two peaks have been clipped, with the largest peak being at the first time sample. The noise signal associated with this peak is $c\delta[n - n_o]$, where c is the amplitude of the clip and $\delta[n - n_o]$ is one at $n = n_o$ and zero elsewhere. The Fourier transform of a single clipped signal has a constant magnitude, with a linearly varying phase (except when $n_o = 0$, when it will have a constant phase). We can see the noise introduced by the clipping in the middle plot, which superimposes the values of each frequency bin as circles, with the ×'s representing the values in the frequency bins of the clipped signal. The effect of the larger clipped peak is apparent as a shift of all the constellation points, while the smaller clipped peak causes a smaller offset. Since the largest clip comes from the first time sample, the offset on each frequency bin is a constant with no phase shift. The smaller second clip does have a linear phase component, thus the offset is the smaller magnitude with a phase that is dependent on the frequency bin. This is apparent from the right plot of Figure 6-7, which shows the large constant offset caused by the large clip and the small circular displacement of the constellation points by the smaller clip.

Since the location of the peaks to be clipped are not known to the receiver, the noise vector degrades the SNR of the received signal and will be discussed in the

next section. If the transmitter were willing to spend some bits to communicate the clipped peak locations to the receiver, then their effects can be subtracted out without affecting the received SNR. This is similar to the strategy used by [16], which is not desirable both because of the need to transmit extra information to the receiver as well as the large number of bits that might be needed to encode the location and amplitude of each clip. To significantly reduce the PAPR by clipping requires a large number of the signal peaks, which could reduce the code rate unacceptably.

When the signal is converted to analog, there may be peaks that reappear because of the bandlimited interpolation of the digital signal. These peaks can reduce any PAPR gains from the clipping by 5 dB [37]. However, it is advantageous to detect and clip peaks in the digital domain to prevent bandwidth expansion. To reduce the peak regrowth, the time signal should be oversampled by at least a factor of four before clipping [57]. The digital clipping will no longer be bandlimited, but digital filtering is more accurate than analog filtering.

6.4.2 SNR Degradation from Clipping

As Figure 6-7 shows, a clipped time sample will lead to an offset of the constellation points in each frequency bin. Assuming the probability of clipping is small, then there will either be zero or one peak clips per OFDM symbol. Figure 6-8 looks at this situation in more detail. The gray constellation is the original signal constellation for each frequency bin. The black constellation shows the resulting constellation offset by a constant which is determined by the position and amplitude of the signal clip as well as the frequency bin in question. This offset r has the same magnitude but different phases for each of the frequency bins of the OFDM signal. The dotted lines denote the decision regions for each constellation point. While the original constellation points are centered in their decision regions, the clipped constellation points are shifted closer to one side of their respective decision regions, which has the same effect as additive noise on the constellation points. The clipping distortion can therefore be viewed as a deterministic noise source, which degrades the effective SNR. We can determine this expected SNR loss for low clipping probability levels.

With additive Gaussian noise, the uncoded error probability for the unclipped signal constellation is

$$\Pr(\epsilon) = Q\left(\frac{d}{2\sigma}\right) = Q\left(\sqrt{\text{SNR}}\right), \tag{6.20}$$

where $d/2$ is the distance from the constellation point to the nearest boundary of the decision region, σ^2 is the variance per dimension of the Gaussian noise, and $Q(\cdot)$ is the Q-function as defined in (5.61). In actuality, the probability of error is the sum of

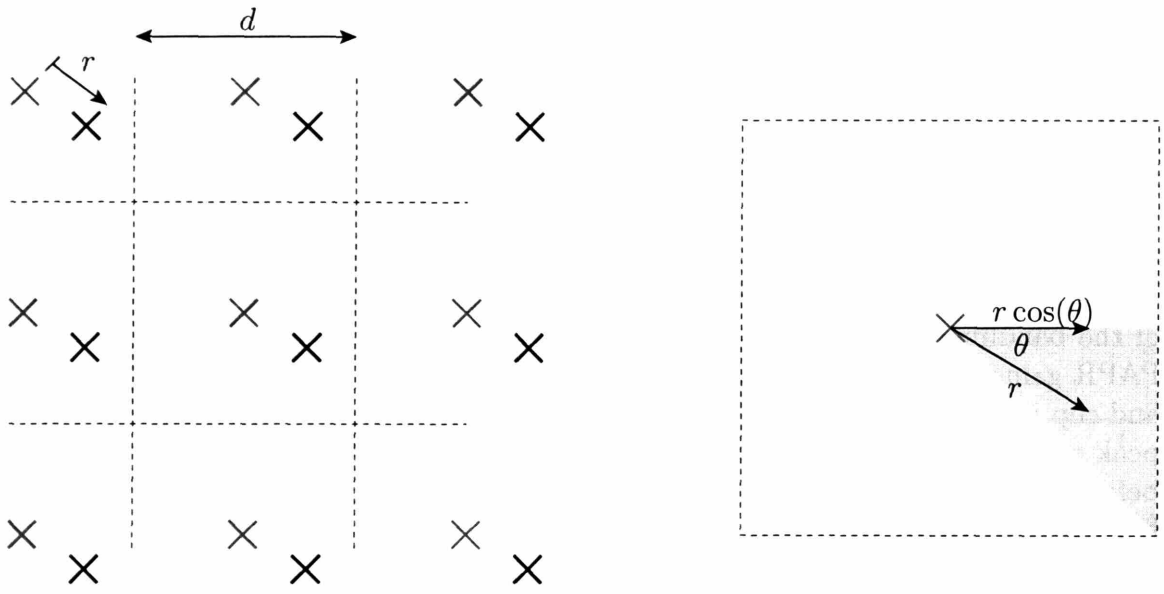


Figure 6-8: Constant offset caused by a single clip (left) shifts the resulting constellation closer to a decision region boundary, increasing the error probability. A zoomed picture of a single constellation point and its decision region is shown on the right.

several Q-functions, one for each border of the decision region. Since the Q-function is a strong function of its argument, however, the term that comes from the shortest distance dominates the error probability.

After the constellation has shifted, the distance to the nearest decision boundary has been reduced and is dependent on the relative orientation of the error vector to the decision boundaries. This distance can be determined exactly for each frequency bin if the details of the signal clip are known. The receiver, however, does not in general know this information and will only see an apparent drop in SNR due to the reduced distance from the decision region boundary. Both an average and a worst-case SNR loss can be calculated for a given clipping probability.

Given the situation shown in Figure 6-8, an offset r will cause the error probability to be

$$\Pr(\epsilon) = Q\left(\frac{d - 2r\cos(\theta)}{2\sigma}\right) = Q\left(\sqrt{\mathrm{SNR} - \frac{r^2\cos^2(\theta)}{\sigma^2}}\right), \tag{6.21}$$

where r is the magnitude of the constellation error vector caused by the clip, and θ is the angle between the error vector and the shortest distance to a decision region boundary. The error term in the Q-function can be viewed as an SNR loss caused by the clipping operation. In order to find the value of this SNR loss, the density

functions for r and θ need to be determined.

For a particular clip, the samples of the error vector $e[n] = r\delta[n - n_o]$ are all zero except for sample n_o where the clip occurred. The Fourier transform of this time signal results in a constant magnitude frequency signal with frequency-dependent phase given by

$$E(f) = \sum_{n=0}^{\infty} e[n]e^{j2\pi fn} = re^{j2\pi fn_o}. \tag{6.22}$$

Thus the magnitude of the clip of the time signal is also the magnitude of the error vector in the frequency constellations. The exact phase of the error vector depends on the phase and location of the clipped time signal, as well as the frequency bin, so it may be modeled as uniformly distributed over all values. Due to the symmetry of the decision region for an M-QAM constellation, it is only necessary to consider the shaded range of angles of shown Figure 6-8.

The probability density for r can be found from the probability density of the magnitudes of the time samples, which is Rayleigh distributed. If the time signal is modeled as Gaussian with complex variance $2\tau^2$, the probability density for the instantaneous magnitude is

$$p_\alpha(\alpha) = \frac{\alpha}{\tau^2} e^{-\frac{\alpha^2}{2\tau^2}} \qquad 0 \le \alpha \le \infty. \tag{6.23}$$

If the clipping level is A, then the error signal when a clip occurs is

$$e[n] = (\alpha - A)\delta[n - \beta], \tag{6.24}$$

where β is the location of the clip. The value of β is not considered, as it only affects the phase information of the error vector, which we do not analyze. The probability density of r, the magnitude of the nonzero sample, can be derived from the density for α. Thus the required probability densities are

$$\begin{aligned} p_r(r) &= \frac{r + A}{\tau^2} e^{-\frac{r(r+2A)}{2\tau^2}} \qquad 0 \le r \le \infty \\ p_\theta(\theta) &= \frac{4}{\pi} \qquad 0 \le \theta \le \frac{\pi}{4}. \end{aligned} \tag{6.25}$$

The clipping threshold A can be derived from the desired clipping probability. The probability of a clip given by (6.5) can be rearranged into

$$\text{PAPR}_{\max} = -\ln\left(1 - \sqrt[N]{1 - \Pr(\text{clip})}\right), \tag{6.26}$$

where N is the number of OFDM frequency bins, and $\text{PAPR}_{\max}$ is the maximum

allowed PAPR. As Figure 6-1 shows, it is sufficient to double the value of N to get the right probability distribution for the bandlimited interpolation of the sampled signal. Since

$$\text{PAPR}_{\max} = \frac{A^2}{2\tau^2}, \tag{6.27}$$

the clipping level can be written as

$$A = \left[-2\tau^2 \ln\left(1 - \sqrt[M]{1 - \Pr(\text{clip})}\right)\right]^{\frac{1}{2}}, \tag{6.28}$$

where $M = N$ for the sampled signal, and $M = 2N$ for the analog signal.

The distance to the decision boundary after clipping is $d/2 - r\cos\theta$, compared to the original distance of $d/2$. The SNR loss caused by the clipping error vector is then

$$\text{SNR}_{\text{loss}} = \frac{\left(\frac{d}{2} - r\cos(\theta)\right)^2}{\left(\frac{d}{2}\right)^2} = \frac{d^2 - 4dr\cos(\theta) + 4r^2\cos^2(\theta)}{d^2}. \tag{6.29}$$

The expected SNR loss is then

$$\text{E}\left[\text{SNR}_{\text{loss}}\right] = 1 - \frac{4\text{E}\left[r\right]\text{E}\left[\cos(\theta)\right]}{d} + \frac{4\text{E}\left[r^2\right]\text{E}\left[\cos^2(\theta)\right]}{d^2}, \tag{6.30}$$

where the expectations factor because r and θ are independent. The expectations for θ can be easily calculated to be

$$f_\theta = \text{E}\left[\cos(\theta)\right] = \int_0^{\pi/4} \cos(\theta) p_\theta(\theta)\, d\theta = \frac{2\sqrt{2}}{\pi} \tag{6.31}$$

$$g_\theta = \text{E}\left[\cos^2(\theta)\right] = \int_0^{\pi/4} \cos^2(\theta) p_\theta(\theta)\, d\theta = \frac{\pi + 2}{8}. \tag{6.32}$$

The expectations for r are more complicated, however, so will be evaluated numerically. If we define

$$f_r(\tau, A) = \text{E}\left[r\right] = \int_0^\infty r \frac{(r + A)}{\tau^2} e^{-\frac{r^2 + 2rA}{2\tau^2}}\, dr \tag{6.33}$$

$$g_r(\tau, A) = \text{E}\left[r^2\right] = \int_0^\infty r^2 \frac{(r + A)}{\tau^2} e^{-\frac{r^2 + 2rA}{2\tau^2}}\, dr, \tag{6.34}$$

then the expected SNR loss from the clip is

$$\text{E}\left[\text{SNR}_{\text{loss}}\right] = 1 - \frac{4}{d} f_\theta f_r(\tau, A) + \frac{4}{d^2} g_\theta g_r(\tau, A) \tag{6.35}$$

$$= 1 - \frac{8\sqrt{2}}{\pi d} f_r(\tau, A) + \frac{\pi + 2}{2d^2} g_r(\tau, A). \tag{6.36}$$

The moments of r are functions of both the signal power $2\tau^2$ and the target clipping probability, as (6.28) relates the clipping threshold A to the clipping probability. The functions f_θ and g_θ encapsulates the effect of the random phase on the SNR degradation for the average case. To find the worst-case SNR loss, we can let $f_\theta = g_\theta = 1$, which assumes that the error vector is always perpendicular to the nearest constellation decision boundary.

To calculate τ and relate it to d, we note that the power of the time series and the frequency bins must be related by Parseval's theorem to be [45]

$$\frac{1}{N} \sum_{k=0}^{N-1} |X[f_k]|^2 = \sum_{i=0}^{N-1} |x[t_i]|^2, \tag{6.37}$$

where $x[t_i]$ is the ith time sample, $X[f_k]$ is the kth frequency bin, and N is the total number of samples. The average power of the Rayleigh-distributed time samples is $2\tau^2$, so the total power is $2N\tau^2$. For a 4-QAM constellation, the average power is $d^2/2$. In general, for square M-QAM constellations with the same average power,

$$d_M = d\sqrt{\frac{3}{2^{2n} - 1}}, \tag{6.38}$$

where $n = .5 \log_2 M$ is the number of bits in each dimension ($n = 3$ for 64-QAM, for example), d is the spacing for a 4-QAM constellation, and d_M is the spacing for a M-QAM constellation. As Figure 6-1 showed, it is not the size of the constellation but the average power that affects the average power of the time signal.

Assuming a 4-QAM signal constellation, the signal power in each frequency bin is $d^2/2$, for a total of $Nd^2/2$. Combining (6.37) and (6.38) results in

$$\tau = \frac{d}{2\sqrt{N}} \sqrt{\frac{3}{M - 1}}. \tag{6.39}$$

Figure 6-9 shows the average SNR loss caused by a single clip at a given clipping probability for several different constellation sizes. As expected, the losses are greater for the higher order constellations due to the smaller spacing of the constellation points. The solid lines show the average loss in SNR for an OFDM that was actually clipped, and the dashed lines show the corresponding worst-case behavior (i.e., θ is always assumed to be zero). The SNR losses are quite modest, however, with less than a 1 dB loss for a 64-QAM constellation at 10^{-2} clipping probability per symbol. As Figure 6-7 shows, the magnitude of the error vector of even a fairly large clip

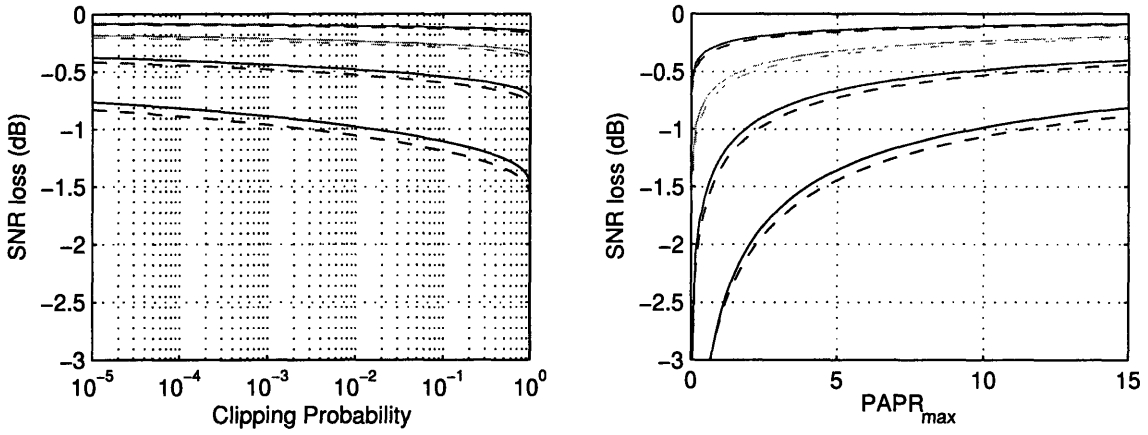


Figure 6-9: Average SNR loss due to clipping (left) and as a function of the maximum allowed PAPR (right) for length 128 OFDM symbols using (top to bottom) 4-, 16-, 64-, and 256-QAM constellations. Dashed lines indicate the worst-case behavior.

is not very large. It should be noted that these SNR curves assume only a single clip occurs, so they break down at high clipping probabilities. At these high clipping probabilities the SNR loss will actually be worse due to the larger number of clips.

Since the SNR loss is appreciable only when the clipping probability is high, however, the right plot instead shows how the SNR loss is affected by the maximum allowed PAPR. Again, the SNR loss is only significant when the clipping threshold is very low to raise the clipping probability. For example the SNR loss for 64-QAM constellations doesn't exceed 1 dB unless the signals are clipped at only 3 dB PAPR. In addition, these SNR losses are only considering those OFDM symbols which have been clipped. Using 64-QAM at a clipping probability of 10^{-2}, the SNR loss is about 0.5 dB. Since this SNR loss only happens 1% of the time, the other 99% of OFDM symbols do not suffer any SNR loss. Thus if the SNR loss is the only consideration, it is possible to operate at a fairly high clipping probability without a significant loss in the effective SNR.

6.4.3 Bandwidth Expansion from Clipping

Although the SNR loss from clipping is not severe, clipping of the analog signal is a nonlinear operation which causes bandwidth expansion. Although it is difficult to characterize the bandwidth expansion caused by the clipping operation, we can look at the effects of clipping at various levels on the waveform bandwidth. Figure 6-10 shows the cumulative maximum spectral height for several clipping levels including

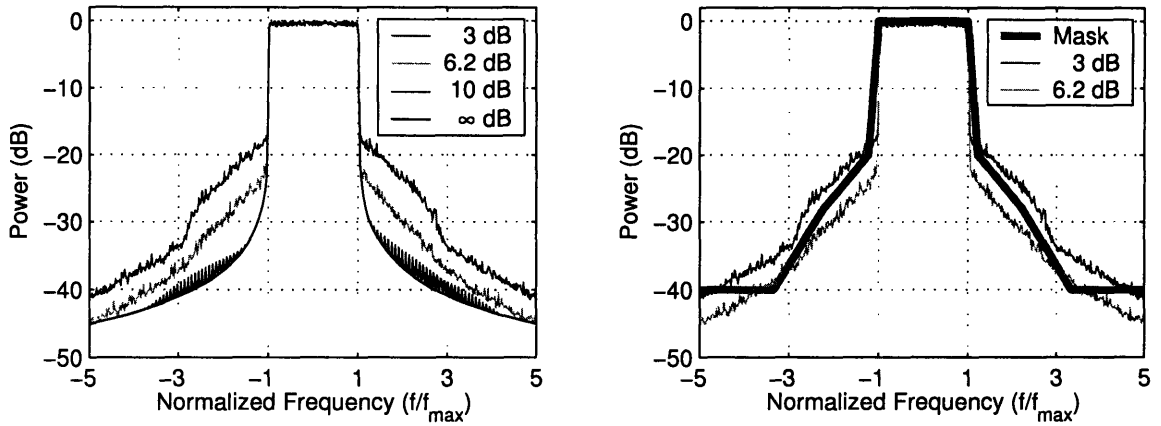


Figure 6-10: Cumulative maximum spectral heights for several clipping levels. The 6.2 dB and 10 dB levels correspond to the maximum PAPRs for a 10^{-2} clipping probability for the precoding algorithm (Section 6.5) and the original distribution. Right plot compares spectral heights to 802.11a transmitter spectral mask.

the original unclipped signals. The resulting traces show the highest spectral power encountered for 460,000 sample OFDM signals. The original signal bandwidth is normalized to ± 1, and shows sidelobes greater than 25 dB down. Limiting the PAPR to 10 dB or less results in a clipping probability of 10^{-2}, and raises the sidelobes slightly.

The greedy precoding algorithm we describe in Section 6.5 is able to reduce the PAPR to 6.2 dB at the same clipping probability of 1%. For the original OFDM symbols, clipping the time signals to this level clips nearly all of the OFDM symbols, and raises the sidelobes by 10 dB or more. Clipping the OFDM symbols to a maximum PAPR of 3 dB like what would be available with the use of complementary Golay codes [12] leads to even greater sidelobe expansion. The raised sidelobes represent signal energy in frequencies outside the system bandwidth, which wastes power and may exceed the levels allowed by the spectral mask.

The right plot of Figure 6-10 compares the bandwidth of the clipped OFDM symbols with the allowable spectral mask for the transmitter in an 802.11a system [30]. The spectral mask specifies the maximum power that can be transmitted both within and outside the system bandwidth. When the PAPR is limited to 3 dB the sidelobes are certainly greater than the allowable spectral leakage, but even at 6.2 dB the sidelobes violate the allowable transmit frequency spectrum. Filtering the output signal to limit the sidelobes is one way to reduce the bandwidth expansion, but lossy filters at the output of the power amplifier will significantly reduce efficiency [13]. Filtering the signal before the amplifier can reduce this efficiency loss, but requires

construction of a sharp analog filter, increasing circuit complexity. In addition, the filtering can cause spectral regrowth [37], which counteracts the PAPR gains from clipping. Although it is not shown, the spectral sidelobes from the 10 dB maximum PAPR clipping level just barely meets this spectral mask. The spectral mask for a given system is application dependent, so these results may vary depending on the type of system, however, so the main message is that the limiting effect of clipping is on the bandwidth expansion rather than the loss in SNR. The PAPR algorithms presented in the following sections can be viewed as a digital filter designed to keep the sidelobes small at reduced PAPRs without causing spectral regrowth.

6.5 Precoding Algorithm for PAPR Reduction

As was shown in Section 6.3, the efficiency of the linear amplifiers are strongly dependent on the maximum PAPR that is allowed for the input signal. A common drawback of existing PAPR reduction algorithms is the rate loss that is associated with a large reduction in the PAPR. In an ideal situation, the rate loss would be negligible or at least small. However, as Figure 6-1 shows, the PAPR for a length 128 OFDM symbol is very likely to be above 6 dB, and has about a 50% chance of being above 7 dB. To limit the PAPR to 7 dB requires a code that avoids half of the possible code words, for a rate loss of 1 b/s/Hz.

The precoding algorithm presented in this section is able to reduced the PAPR to less than 7 dB at a rate loss of 2 b/s/Hz, which is only 1 b/s/Hz less than what is theoretically possible. From a different viewpoint, this algorithm uses less than 0.5 dB of extra power on average to reduce the peak power by 5 dB or more. The precoding algorithm is compatible with any upstream coding scheme which uses finite constellation sizes, and does not require any rate loss in relationship to that encoding scheme. Although the encoding at the transmitter is computationally complex, the receiver can decode without any side information from the transmitter with little additional complexity.

6.5.1 Precoding Connection to WiGLAN Downlink Protocol

Since it is not possible to construct a code with negligible rate loss while reducing the PAPR to 7 dB or less, we instead look for a PAPR-reducing code that can operate independently of any other coding that might have been used for error protection. Because it is the large peak power that limits the amplifier efficiency, trading off some average power to reduce the peak power can significantly reduce the PAPR.

Each frequency bins of an OFDM symbol corresponds to a complex sinusoid in time, with the amplitude and phase specified by the particular constellation point

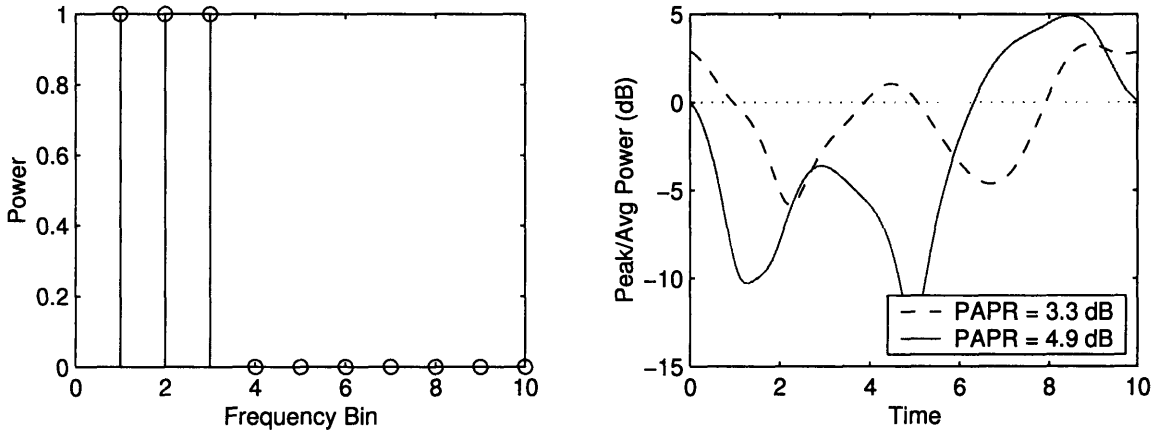


Figure 6-11: OFDM symbol with 10 frequency bins (left), and the corresponding PAPR (right) for two different phase configurations.

chosen for that frequency bin. Thus if only a single frequency bin is used, the OFDM time signal is constant amplitude, with a PAPR of 0 dB. Each additional nonzero frequency bin will add a complex sinusoid to the time signal, which adds constructively or destructively over the many time samples. The PAPR of the resulting time series is dependent on the phases of the values in each frequency bin.

Figure 6-11 illustrates this phase dependence using three frequency bins. While each of the frequency bins alone would produce a constant amplitude time signal, the sum of three different values causes the time signal to have a nonconstant magnitude. As the left plot shows, a length ten OFDM symbol has three nonzero frequency bins, all with unit magnitude but differing phases. The right plot shows two time signals, corresponding to two different phases for the third frequency bin. Rotating the third frequency bin by $\pi/2$ leads to a change in the PAPR of 1.6 dB. This idea is used in [44] to reduce the PAPR of OFDM symbols by grouping the frequency bins into blocks and then adding one of four phase offset to each block to minimize the PAPR. The phase offset used for each block must be transmitted to the receiver, which uses a varying amount of code rate depending on how many blocks are used.

The transmitter is able to determine how the frequency bins are going to interact to form the time signal (by taking the Fourier transform) before it has to transmit anything. These frequency bins can then be seen as creating known interference with each other, so the transmitter can do some precoding to reduce this interference, similar to the downlink precoding used in the WiGLAN architecture and described in Section 3.2.3. For the WiGLAN downlink precoding, the desired input constellation is replicated infinitely in space (Figure 3-5), with corresponding points in each copy

of the original constellation being equivalent. The downlink precoding algorithm then selects among the equivalent constellation point to counteract the effect of the interference caused by the channel. Because of the linear nature of the interference, the choice of which copy of the constellation points to choose can be made through a modulo operation, which requires little extra complexity at the receiver.

Similarly, the PAPR precoding algorithm can select for each frequency bin the copy of the desired constellation point that would minimize the PAPR of the time signal. However, unlike the previous precoding algorithms, the interference from the channel is caused by the same constellation points that are being selected from, therefore the choice of constellation points in each frequency bin will globally affect the interference experience by all the other frequency bins. To achieve a minimum PAPR via precoding must then require a search through the exponentially-large space of all possible equivalent points for each frequency bin. How this search is done as well as how to reduce the search complexity is described in the next section.

An existing idea, called Tone Insertion [68] uses a similar concept to reduce PAPR. The input constellation is also replicated, but instead of a regular tiling of the complex plane, the offset between constellation copies can be any value, with the regular infinite lattice being a special case. This still has an exponentially large search space, and iterative algorithms show a maximum PAPR of around 9 dB.

6.5.2 The PAPR Precoding Algorithm

As we have seen previously, the amount of rate required to reduce the PAPR to less than 7 dB for a 128 bin OFDM symbol is 1 b/s/Hz. One way to approach this rate is to use the ideas from TH precoding. In TH precoding, the input constellation is replicated to tile an infinite plane, with each copy being equivalent from the standpoint of the decoder. The encoder then chooses the offset to cancel out the interference from the channel such that the received signal is one of the constellation points equivalent the transmitted constellation points. The precoding algorithm could also choose from the entire infinite lattice of equivalent points for each frequency bin to minimize the PAPR of the OFDM symbol. Most of the equivalent points are undesirable choices, however, because they are far away from the origin, and hence require more power to transmit. Although we are willing to trade off some average power to reduce the peak power, increasing the average power significantly will waste transmit power. In addition, the fraction of power allocated between the frequency bins are chosen to maximize the data rate, but choosing the higher power alternate constellation points will redistribute the power away from the optimum. The precoding algorithm instead uses a slightly augmented constellation that is only twice as big as the original, but provides four equivalent choices for each constellation point.

Figure 6-12 shows the constellation that the precoding algorithm uses. There are

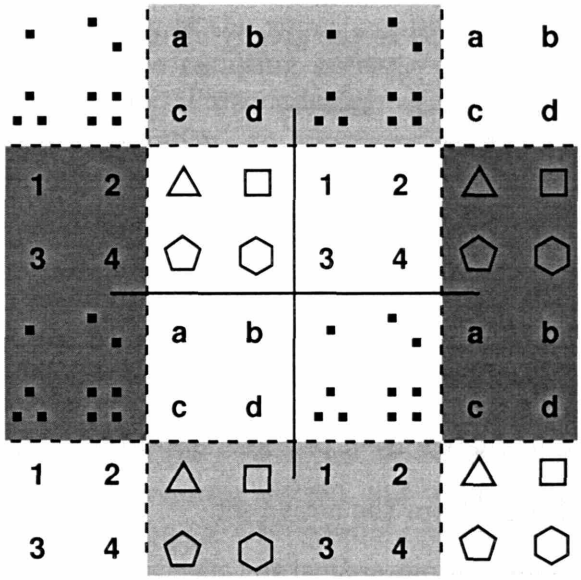


Figure 6-12: Constellation mapping for transmit precoding with 16-QAM showing four equivalent copies of the original constellation (center).

a total of four copies of the original constellation, which is 16-QAM in this case. The original constellation is in the center, with the three other copies divided into the corners and sides. Because of the periodic nature of the infinite constellation tiling, each of these copies is actually composed of sections of several different constellation copies, which are delineated by the dashed lines. The four lightly shaded corners of this precoding constellation form a single redundant copy of the original constellation.

These three constellation copies are the lowest power copies of the original constellation, so their use will minimize how much the average power is increased. Since there are four redundant choices for each constellation point, the rate loss is 2 b/s/Hz. However, this is offset by the increased size of the precoding constellation. In effect, this PAPR precoding strategy has no coding loss, with the penalty coming from the increase in average transmit power. In addition, the PAPR precoding works independently of any other coding that might be present in the system as long as the constellation used is finite in size.

While the expanded constellation allows the PAPR to be reduced, the nonlinear nature of the of the frequency to time conversion makes it very difficult to predict what the proper choice of constellation points should be. An exhaustive search requires four choices per frequency bin, for a total of 4^N possibilities for N bins. This exponential complexity makes an exhaustive search infeasible for all but the smallest values of N.

To reduce the computations needed to search through the space, a greedy iterative

algorithm can be used. The steps to the greedy algorithm are:

1. Encode OFDM signal using original constellation
2. Choose active set of M bins (default is all bins, $M = N$)
3. Enumerate the reduction in PAPR from changing the constellation point in a single bin to one of the alternate points ($4M$ computations)
4. Choose the bin b and equivalent constellation point that produces the greatest reduction of PAPR and recalculate the OFDM symbol
5. End algorithm if no change from previous iteration
6. Remove the used bin b from the active set
7. Go to Step 3 if maximum number of iterations has not been reached.

The greedy algorithm at each iteration always chooses the configuration that gives the biggest gain. There is no guarantee that this is the best choice, and in general, a greedy algorithm will not produce the global optimum. The computation required for each iteration of the greedy algorithm is linear in the number of bins searched over, with a worst-case complexity that is $\mathcal{O}(N^2)$. Note that rather than always recalculating the OFDM time symbol when a different constellation point is chosen, a single offset signal can be calculated and added to the previous time signal. The offset is a complex sinusoid with amplitude and phase governed by the difference between the old and new constellation point, and the frequency given by what frequency bin is being operated on. A random-greedy algorithm can further reduce computation by only considering a random fraction of all the bins in the active set, resulting in a small loss in performance.

6.6 Amplitude Synthesis for PAPR Reduction

An alternative method to reduce the PAPR is inspired by a related problem in phased array design. A phased array alleviates the problems of having a high powered linear amplifier by summing the outputs of many amplifiers in free space. A Wilkinson combiner can also be used to sum two amplifier outputs (and is used for the outphasing amplifier configuration), but as described in Section C.1.1, the combiner is very large for an integrated circuit. One of the features of a phased array is the reconfigurability of its far-field antenna pattern, allowing it to, for example, sweep a narrow beam of energy across a wide range of angles. The antenna pattern can be controlled by adjusting the phase and magnitude of the individual amplifiers to shape the overall

antenna pattern. If the outputs of the small amplifiers are treated as samples of a continuous-time signal, then the resulting antenna pattern is given by the Fourier transform of the time samples [2]. The problem of trying to specify the amplitude and phases for each of the amplifiers to achieve a desired antenna pattern is very similar to trying to specify the amplitude and phases of the frequency bins to make the time signal a desired shape (in this case, a constant amplitude signal).

As shown previously, it is not possible to create constant amplitude OFDM symbols without losing a significant amount of rate. Using 1 b/s/Hz of redundancy, however, can theoretically limit the PAPR to less than 7 dB. The PAPR precoding algorithm previously discussed only uses 2 b/s/Hz to achieve the same results. It can be computationally intensive, however, even with the greedy iterative algorithm. By potentially giving up data rate, the PAPR can be reduced to 7 dB similarly to the precoding algorithm from the previous section, but with less computation.

For each frequency bin in the OFDM symbol, M-QAM constellations are used, which encode the information in each bin with a magnitude and phase representing the appropriate constellation point. It is a known result that given only the frequency-domain magnitude of an unknown time signal which meets a mild set of conditions, it is possible to recover the frequency-domain phase of the unknown signal [69]. Given a different set of conditions (though still fairly easy to meet), it is possible to recover the frequency magnitudes given only the phases [26]. These procedures work by repeatedly transforming the known signal back and forth between time and frequency to regenerate the missing information. This procedure is very sensitive to noise, however, so it can not be used to send only the magnitude information of a signal without sending the phase information as well.

The PAPR synthesis algorithm only transmits information via the frequency magnitude of the OFDM symbol, and tries to "recover" the phase information that comes from a low PAPR time signal with the desired frequency magnitudes. At the transmitter, this operation comes at the cost of half the rate, since information could have also been sent via the phase channel. At the receiver, the decoding only uses the magnitudes of the frequency bins and ignores the phases. Like the precoding algorithm, the synthesis algorithm does not require information to be shared between the transmitter and receiver, and has a much higher complexity to encode rather than decode. The rate of 1/2 is very large when the SNR is high, but at low SNR, this rate loss actually falls to zero because the phase channel has zero capacity, making this algorithm more desirable than the precoding algorithm in this regime.

6.6.1 Capacity of Phase- and Magnitude-Only Channels

Before sacrificing either the frequency magnitude or phase to reduce the PAPR, it is useful to determine how much rate will be lost. Starting with a complex additive

Gaussian channel,

$$y = x + n, \tag{6.40}$$

where x is the input, y is the output, and n is additive white Gaussian noise, with complex variance $2\sigma^2$ and zero mean. The noise is uncorrelated with the input x. The complex channel can be viewed as two independent parallel channels, one each for the real and imaginary components. For the real or imaginary channel, the noise is a one-dimensional Gaussian with variance σ^2, so the capacity given an average input power P is [11]

$$C_{\text{real}} = C_{\text{imag}} = \frac{1}{2}\log_2\left(1 + \frac{P}{\sigma^2}\right) \tag{6.41}$$

in b/s/Hz where P/σ^2 is the SNR. The capacity of the complex channel C_{cplx} is then the sum of the capacities for the real and imaginary channels.

Alternately, the complex channel can be decomposed into a magnitude-only and a phase-only channel. This configuration may have a different total capacity than the real and complex channels, even though they both describe the same complex channel. Figure 6-13 shows what signal constellations might look like for the different channels. The left plot is a 16-QAM constellation, which is a typical constellation for the complex channel. This can be split up easily into two one-dimensional constellations, one each for the real and imaginary channels. The middle plot is what a constellation would look like for the magnitude-only channel, in which only the magnitude is specified, and the phase is arbitrary. The result is concentric rings, with each ring representing a constellation "point." The right plot shows the constellation "points" for the phase-only channel, which only specifies the phase of the input signal, with arbitrary magnitude. Each ray radiating from the origin represents a single constellation point, although given a average power constraint, the constellation points would be the intersection of these rays with a circle of radius $\sqrt{P}$.

The magnitude-only channel looks like

$$y = \left|\, |x|e^{j\theta} + n \,\right|, \tag{6.42}$$

where the magnitude $|x|$ encodes the information to be sent, and the angle θ is arbitrary. The noise n is complex Gaussian as before, and the output y is the magnitude of the noisy input signal. For this channel, the capacity in the high and low SNR regime is [31]

$$C_{\text{mag}} \approx \begin{cases} \frac{1}{2}C_{\text{cplx}} & \frac{P}{\sigma^2} \gg 1 \\ C_{\text{cplx}} & \frac{P}{\sigma^2} \ll 1\,. \end{cases} \tag{6.43}$$

At high SNR, the capacity of the magnitude-only channel is half that of the complex

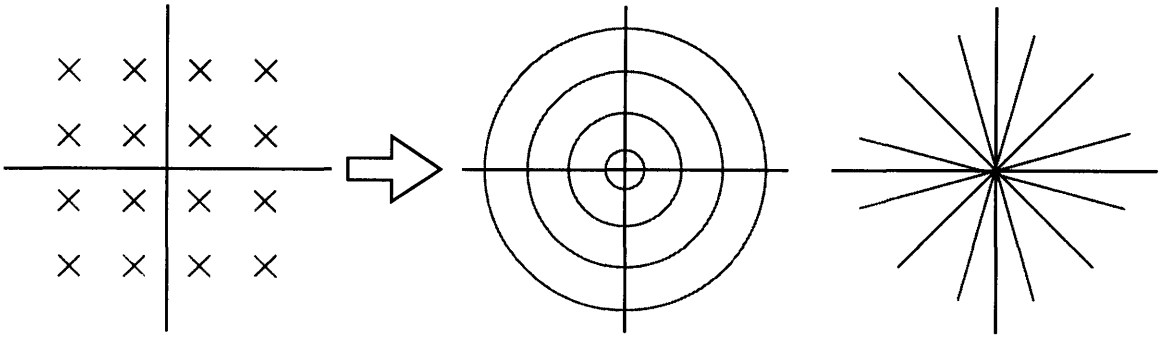


Figure 6-13: A typical constellation for the complex (left), magnitude-only (middle), and phase-only (right) channels.

channel, which is the same as if information was only transmitted on the real channel. Thus restricting the input to only transmit information via the signal magnitude imposes a rate loss of one half. The input distribution that achieves this is a Rayleigh distribution. This seems a natural choice since the optimal distribution for the complex channel is a Gaussian, and the magnitude of a Gaussian variable is Rayleigh distributed. At very low SNR, however, the optimum distribution is Pulse-Position Modulation (PPM) [21], and there is no loss in capacity by using the magnitude-only channel. This is unlike the case of the complex channel, for which the Gaussian distribution is optimal for all SNR, and the real channel always has half the capacity of the complex channel.

The phase-only channel looks like

$$y = \angle\left(\sqrt{P}e^{j\theta} + n\right), \tag{6.44}$$

where the input x is a constant magnitude signal with magnitude $\sqrt{P}$. The information is carried on the phase θ, and n is again complex Gaussian noise. The output y is the phase of the input signal with additive noise. At high SNR, the capacity of the phase-only channel is [5]

$$C_{\text{phase}} \approx log_2\sqrt{\frac{4\pi}{e}\frac{P}{\sigma^2}} = \frac{1}{2}log_2\left(\frac{P}{\sigma^2}\right) + \alpha, \tag{6.45}$$

where $\alpha \approx 1.10$. At high SNR, the capacity of the complex channel is

$$C_{\text{cplx}} \approx \log_2\left(\frac{P}{\sigma^2}\right), \tag{6.46}$$

so the capacity of the phase-only channel is nearly half that of the complex channel. At low SNR, since the capacities of the magnitude-only channel is the same as the capacity of the complex channel, the phase channel must have capacity zero.

Thus using only the magnitude or the phase of the channel to send information results in a rate loss of one half in the high SNR regime. The remaining dimension can be used with the following synthesis algorithm to reduce the PAPR of the OFDM symbols. The frequency magnitude of the OFDM is not well suited to reducing the PAPR because of the frequency dependent nature of the wireless channel. Since the frequency magnitude is usually scaled according to the channel conditions, allowing the magnitudes to be altered to reduce the PAPR would change the distribution of signal power in the frequency bins to an undesirable configuration. On the other hand, allowing the phase to be altered has no effect on the power distribution between the frequency bins. The PAPR synthesis algorithm will therefore use the phase of the frequency bins as adjustable parameters to try to reduce the PAPR of the OFDM symbol.

6.6.2 The Phase Synthesis Algorithm

Since only the magnitude of the frequency bins carries information, then the phase can be set to whatever values will minimize the PAPR of the signal. If the desired signal is a constant amplitude signal, with the frequency magnitudes specified by the information to be encoded, then the unknown frequency phase can be "recovered" to create the constant amplitude signal using the algorithm specified in [26]. The desired constant amplitude signal is

$$X[f] = M[f]e^{j\theta[f]}, \tag{6.47}$$

where $M[f]$ is the magnitude of the frequency bins and $\theta[f]$ is the corresponding phase.

The synthesis algorithm is then

1. Encode information bits in $M[f]$
2. Choose random phases for $\theta[f]$
3. Project $X[f]$ into time domain to get $x[t]$
   (a) Use IFFT to generate time domain signal
   (b) Clip signal to average power level to get $x[t]$
4. Project $x[t]$ into frequency domain and sample to get $X'[f]$ with $M[f]$ and $\theta'[f]$

(a) use FFT to generate frequency domain signal $X'[f] = M'[f]e^{j\theta'[f]}$

(b) Replace magnitude $M'[f]$ with $M[f]$, keep phase $\theta'[f]$ as is

5. Go to Step 3 if maximum number of iterations has not been reached.

For each iteration, the frequency magnitude is forced to remain at its original value, but the phases are allowed to change. If a solution exists, the phase function should converge to whatever is necessary for the resulting frequency signal $X[f]$ to have a constant amplitude in the time domain.

The algorithm is similar to a POCS (projection onto convex sets) algorithm, which is a fairly simple iterative procedure that converges to an element that is a member of two convex sets [10]. In this case, the two sets are amplitude-limited time signals, and frequency signals with a given frequency magnitude. Note that both the time and frequency signals of these sets are discrete, as the OFDM symbol only specifies the frequency magnitudes for a finite number of frequency bins. The phase synthesis algorithm is not a POCS algorithm because the sets are not both convex. For a convex set, if A and B are both elements of the set, then for any $0 \leq \alpha \leq 1$ the weighted sum of the two elements $\alpha A + (1-\alpha)B$ is also in the set. The set of amplitude-limited time signals is convex, but the set of all signals with a specified frequency magnitude is not. For example, the sum of two signals with the same frequency magnitude but are 180 degrees out of phase at every frequency is zero. A POCS algorithm is guaranteed to find a point in the intersection of the two sets, assuming it exists [10], but this is not guaranteed if the sets are not convex. It can be guaranteed, however, that the error signal never gets larger with each iteration, although it may never converge to zero even if a solution exists [6, 64].

As Figure 6-14 shows, in the POCS algorithm the initial guess alternately projected onto each of the target sets. The error signal of the ith iteration $e[i]$ is the part of the signal that is not in the projected set. For example, when $x[i]$ is projected onto the set of amplitude limited signals, the error signal is the difference between the amplitude limited signal $X[i]$ and the original. In Step 3, the time signal is clipped to the average signal power. Since adjusting the phase of the frequency bins will not affect the total power of the signal, clipping to the average power will ideally limit the time signal to a constant amplitude, reducing the PAPR of the OFDM signal. To project into the other set, the signal is converted to the frequency domain, and then the magnitude is reset to the original specified frequency magnitude. While the frequency magnitude is constantly forced to the original values, the phases are allowed to change to whatever values will lower the PAPR of the OFDM symbol.

Half of the data rate may seem like too much to sacrifice to reduce the PAPR. The resulting constant amplitude signal will make a significant improvement in the efficiency of the power amplifier, however, so this may be a good tradeoff. Addition-

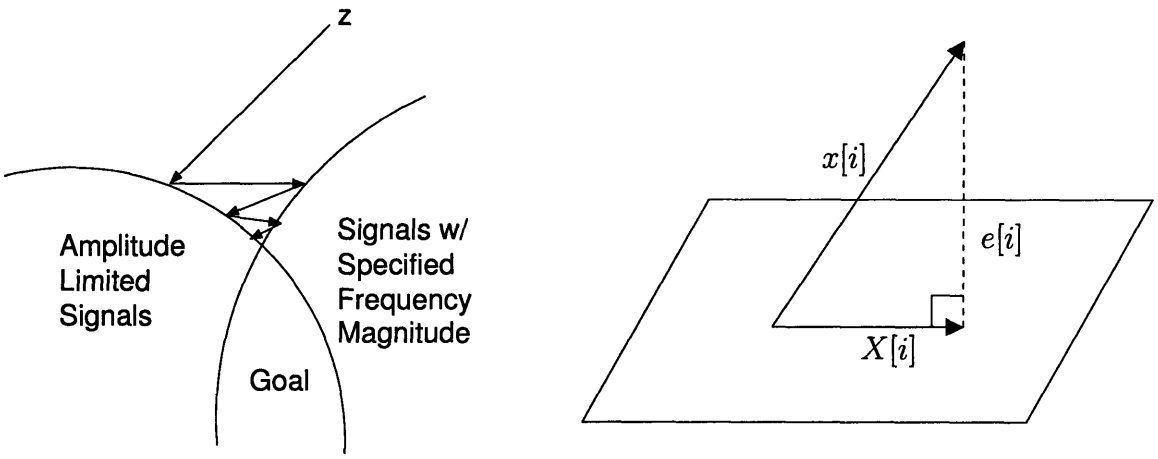


Figure 6-14: POCS algorithm dynamics: the initial guess converges to a point in the intersection of the two sets after repeated projections onto each set. The error signal $e[i]$ decreases with each iteration.

ally, in the low SNR regime there is no loss in capacity, so this synthesis algorithm is actually a better choice than the precoding algorithm.

We can try to analyze how many free parameters must be used to create a constant amplitude signal. In the simplest case, an OFDM symbol has length two, with the two frequency samples X_1 and X_2, which each have a magnitude and phase given by

$$\begin{aligned} X[1] &= M_1 e^{j\theta_1} \\ X[2] &= M_2 e^{j\theta_2} . \end{aligned} \tag{6.48}$$

Using the inverse discrete Fourier transform, the time signals are then

$$\begin{aligned} x[1] &= M_1 e^{j\theta_1} + M_2 e^{j\theta_2} \\ x[2] &= M_1 e^{j\theta_1} - M_2 e^{j\theta_2} , \end{aligned} \tag{6.49}$$

which have squared magnitudes

$$\begin{aligned} |x[1]|^2 &= M_1^2 + M_2^2 + 2M_1 M_2 \cos(\theta_1 - \theta_2) \\ |x[2]|^2 &= M_1^2 + M_2^2 - 2M_1 M_2 \cos(\theta_1 - \theta_2) . \end{aligned} \tag{6.50}$$

For the two time samples to have the same magnitude, the cosine term must disappear, or $\theta_1 - \theta_2 = \pi/2 + k\pi$ for some integer k. More simply put, the phases of the two frequency samples must differ by $\pi/2$ radians. Since only one of the phases needs to be changed, only one parameter must be sacrificed to make the time samples

have the same magnitude. The situation quickly becomes more complicated as the number of samples increases. In addition, for more than two samples, the relationship between the phases depends on all the magnitudes of the frequency samples. Empirical calculations show that for the cases of three and four samples, all but one of the phases needs to be used to equalize the magnitudes of all the time samples. In general, for N frequency samples, each time sample is a linear combination of the N frequency magnitudes. The coefficients are in turn functions of the frequency phases as well as the sample number. Since N equations require N parameters to solve, all the phases are expected to be needed to make all the time sample magnitudes equal. Multiplying all of the frequency samples by a constant phase shift will not affect the magnitude of the time samples, so one of the original phase values can be retained while shifting all of the others to retain the constant amplitudes in time. Thus in general, for N samples, $N-1$ of the phases must be used to make the time samples constant magnitude.

Although this phase synthesis algorithm has enough free variables to make the time samples constant amplitude, only time samples are specified rather than their bandlimited interpolation. The actual signal that is amplified is the bandlimited interpolation of these time samples, which is unlikely to be constant amplitude. In addition, although the error vector for the sampled signals is guaranteed to be non-increasing, the deviation of the associated continuous-time signal has no such guarantee. Simulations show that the although the time samples monotonically converge to constant magnitude, the bandlimited interpolation will sometimes decrease and then increase. For example, Figure 6-15 shows an example OFDM time signal, with a discrete-time PAPR of 7.3 dB in the left plot. After 100 iterations, the discrete-time PAPR is 0.25 dB, which is very nearly a constant magnitude signal. However, the right plot is the $10\times$ upsampled interpolation of the discrete-time signals, showing that the difference in PAPR is much smaller. The original PAPR is now 8.5 dB, while the algorithm results in a PAPR of 5.7 dB. This reduction of of 3 dB is still appreciable, but not as spectacular as one might hope for.

Thus it is necessary to check after each iteration what the continuous-time PAPR is, and save the current values if it is better than the last saved peak. After all iterations have been run, the best values are used rather than the values from the final iteration. The algorithm therefore always runs for the maximum number of iterations and chooses the iteration that resulted in the lowest PAPR. The algorithm is not able to use the information about the continuous-time PAPR unless the original signal is oversampled before the algorithm is used. The extra samples are equivalent to expanding the bandwidth of the signal, which may be undesirable given the additional loss in data rate that it will incur.

This synthesis algorithm can also be easily adapted to the antenna synthesis problem, in which a specified antenna pattern is required. Since the antenna pattern in

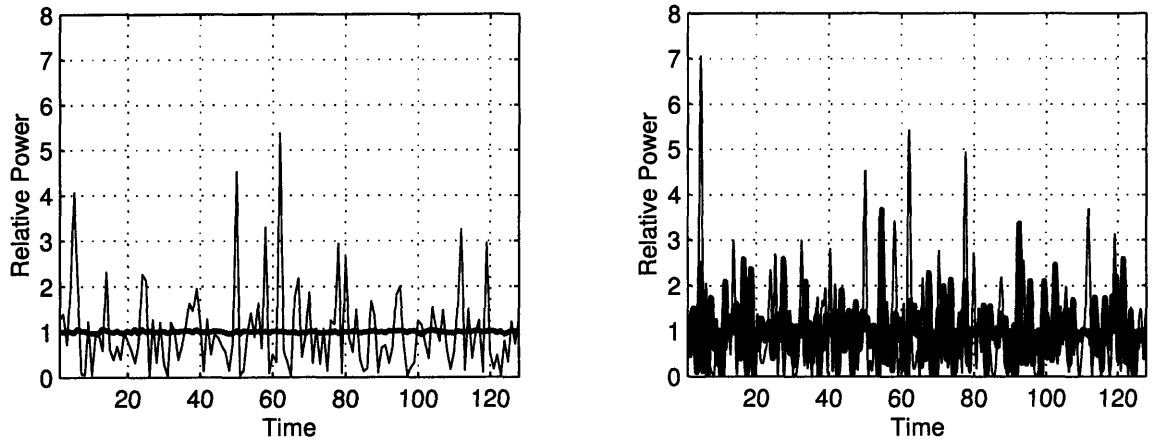


Figure 6-15: Phase synthesis example comparing discrete-time signal (left) and continuous-time signal (right). The thin line is the original signal, and thick line shows the result of the synthesis algorithm after 100 iterations.

the far-field is the Fourier transform of the array excitation, the phase synthesis algorithm can be used if the array excitation is viewed as the OFDM time series and the desired antenna pattern is the magnitude of the frequency signal. The phase of the frequency signal is not important for an antenna pattern, and the phase of the time signal is what is to be synthesized. A constant magnitude time signal is equivalent to having constant amplitude signals being transmitted by the array elements. This allows the array elements to use highly efficient amplifiers, and the antenna pattern is controlled by adjusting the phases of the antenna elements as specified by the synthesis algorithm.

6.7 Simulation Results

In order to determine the performance of the precoding and synthesis algorithms, simulations of both algorithms were run over 460,000 randomly-generated 128 bin OFDM samples using 64-QAM constellations. Both the precoding and the synthesis algorithms reduce the PAPR to around 7 dB, with the synthesis algorithm using significantly less computation.

6.7.1 Precoding Simulation Results

Figure 6-16 shows an example of a 128 bin OFDM symbol using 64-QAM with an initial PAPR of about 13.5 dB. The middle plot shows the original signal with its

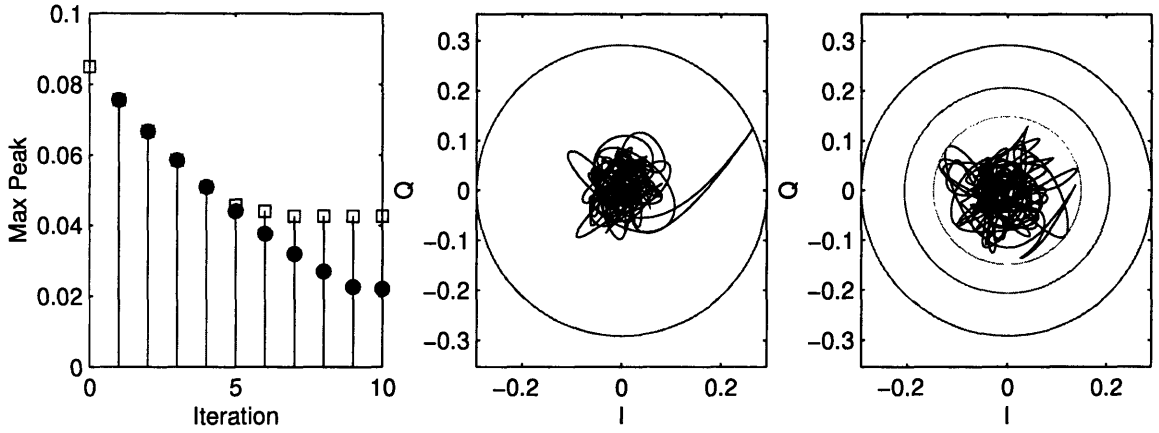


Figure 6-16: Example of PAPR precoding using 10 iterations for the greedy and random-greedy algorithm. Left plot shows the peak reduction with each iteration for the greedy (solid) and random-greedy (hollow) algorithm. Middle plot shows original signal, and right plot shows the result of the greedy algorithm. The middle circle represents the maximum amplitude after using the random-greedy algorithm.

prominent peak, with a circle representing the maximum amplitude. The maximum number of iterations is ten, with the random-greedy algorithm only modifying ten randomly selected bins. As the left plot shows, the random-greedy algorithm does not reduce the peak as much as the greedy algorithm (as expected), but the two algorithms remain close in performance for the first five iterations. For each iteration the algorithm is only able to reduce the peak amplitude by a certain amount, thus the random-greedy algorithm suffers in this case from having too few frequency bins to choose from. Because an OFDM symbol with such a high peak is rare, the greedy and random-greedy algorithms may be closer in performance for the more common OFDM symbols with a lower PAPR.

The right plot shows the result of the greedy algorithm, with a circle denoting the maximum amplitudes of the different algorithm results. The innermost concentric circle shows the maximum amplitude of the greedy algorithm, compared to the much larger circle for the original signal (outermost). The middle circle is the maximum amplitude for the random-greedy algorithm. Instead of a single peak dominating the size of the circumscribing circle, the greedy algorithm produces a time signal that is close to the bounding circle in several places. Also note that the greedy algorithm results in a signal with an evident increase in average power.

Figure 6-17 compares the results of these two algorithms on the example OFDM symbol from Figure 6-16. As the left two plots show, the greedy algorithm produces a

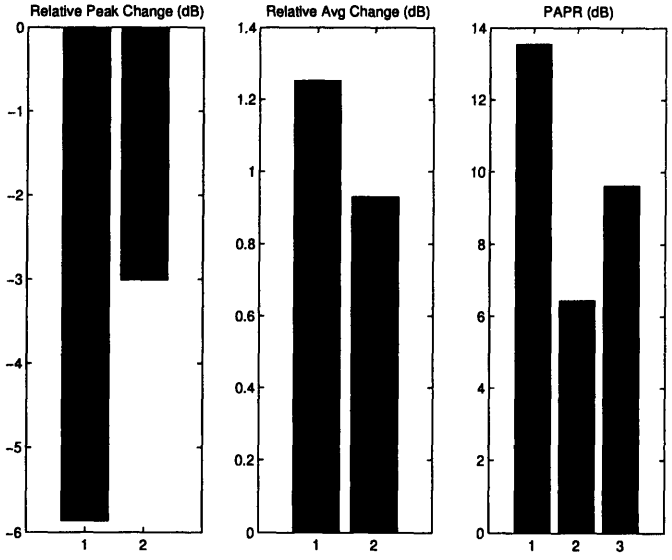


Figure 6-17: Results for example PAPR precoding in Figure 6-16, showing how the peak and average power changes result in PAPR improvements. Left two plots are for the greedy (1) and random-greedy (2) algorithm, and the right plot diagrams the original (1), greedy (2), and random-greedy (3) PAPR.

greater reduction in the peak power, gaining an extra 3 dB in peak reduction over the random-greedy algorithm. A consequence of this reduction is that the average power of the greedy algorithm is also larger than for the random-greedy one. However, the increase in average power is less than 1.3 dB, and only about 0.3 dB worse than that for the random-greedy algorithm. The greedy algorithm is able to reduce the PAPR by 7 dB, while the random-greedy algorithm is able to reduce the PAPR by almost 4 dB while requiring much less computation.

Although the greedy algorithm generally performs better than the random-greedy algorithm, this is not always the case, as the left histogram in Figure 6-18 shows. The histogram plots the difference between the PAPR for the greedy and the random-greedy algorithm for each OFDM symbol. A non-negligible fraction (about 20%) of the OFDM symbols actually performed better with the greedy-random algorithm. At first glance this appears counterintuitive, since the greedy algorithm searches over a much larger set of bins than the random-greedy algorithm. Due to the nonlinear nature of the interaction between the frequency bins, avoiding certain bins can sometimes actually improve performance of the algorithm. There is also a peak at 0 dB, which represents those OFDM symbols where the greedy and random-greedy algorithm produce the same results.

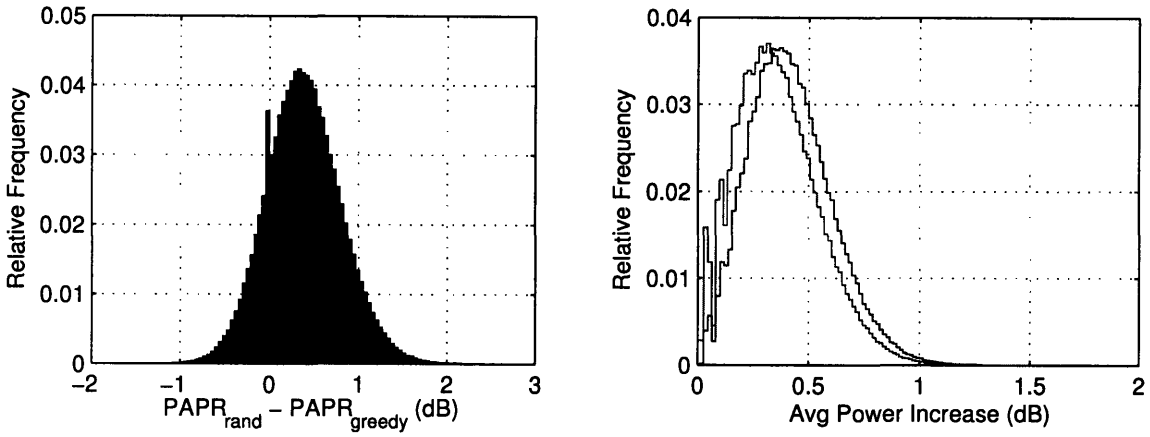


Figure 6-18: Histogram of PAPR reduction loss for random-greedy over greedy algorithm for 20% (25) active bins and the average power gain needed for the greedy (right curve) and random-greedy (left curve) algorithm.

The average power increases required by the greedy and random-greedy algorithms are only a fraction of a dB on average, and rarely above 1 dB. The right plot of Figure 6-18 shows how much the average power was increased by running the greedy and random-greedy algorithms. As expected, the greedy algorithm has a slightly higher increase in average power, but the mean increase in power is only 0.42 dB for the greedy algorithm and 0.36 dB for the random-greedy (25) algorithm.

As Figure 6-19 shows, both the greedy and random-greedy algorithms significantly improves the PAPR of the OFDM symbols. It plots the CCDF from 460,000 runs of the greedy and random-greedy algorithm for randomly generated 128 bin OFDM symbols using 64-QAM. The dashed line shows the probability that the PAPR of an OFDM is greater than $PAPR_o$ for the unaltered signal, and the solid line shows the CCDF resulting from the greedy algorithm. The dot-dashed lines are the CCDFs from the random-greedy algorithm with different fractions of selected bins. At a clipping probability of 1%, the maximum PAPR for the greedy algorithm is 6.2 dB, which is over 3 dB below the uncoded original symbols. At a clipping probability of 10^{-5}, this difference is over 5 dB. The random-greedy algorithm loses less than 1 dB compared to the greedy algorithm when 20% (25) of the bins are available to be precoded. The resulting PAPRs for the greedy algorithm are below the 7 dB required for the outphasing amplifier to be most efficient.

The number of iterations required for the greedy algorithm is not excessively large, as Figure 6-20 shows. A very small number of OFDM symbols cannot be improved upon by the greedy algorithm (less than 0.1%), so only require a single iteration.

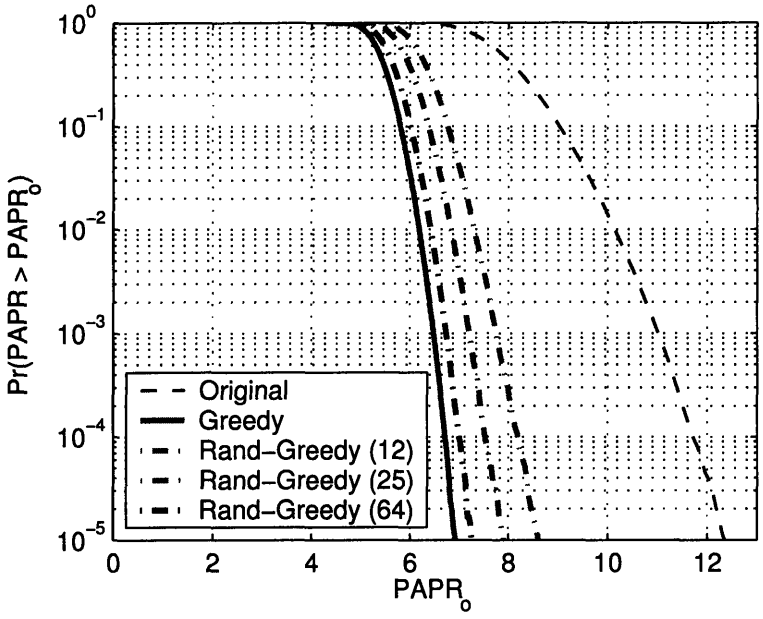


Figure 6-19: CCDFs for precoding algorithm using greedy and random-greedy algorithms for length 128 OFDM symbols and 10× oversampling. The greedy algorithm uses 25 iterations maximum, and the random-greedy algorithms use 10%, 20%, and 50% bins active (right to left).

Most OFDM symbols require between five and six iterations to converge, and less than one percent require more than ten iterations. Note that the last iteration always has the same result as the previous iteration, as that is the sign that the algorithm has converged. The right plot is a histogram of how often each bin is altered by the precoding algorithm. The relative frequency of the bins is normalized to the expected number of times a bin should be used given that all bins are equally good for reducing the PAPR. Although most bins are nearly equivalent, there is a prominent spike near the center bins, indicating that they are the most useful for reducing the PAPR of the OFDM symbol. These middle bins control the highest frequency components of the OFDM symbol, and are used almost twice as much as the other bins to reduce the PAPR. Biasing the random-greedy algorithm to favor the center bins in its active set may increase its performance, allowing fewer bins to be searched over while still maintaining good performance.

The advantage of the PAPR precoding algorithm is the significant reduction in the PAPR that can be achieved without disrupting any other encoding that may be used in the system. The loss in code rate is zero from the viewpoint of any encoder that is before the precoding algorithm, and the cost in increased power is quite small.

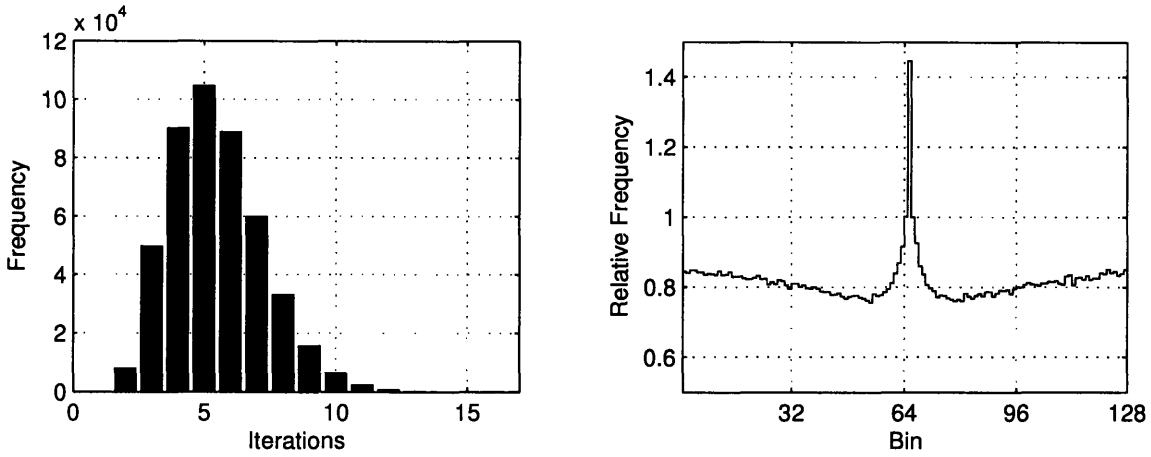


Figure 6-20: Iterations required for for 128 bins, 64-QAM, greedy (less than 1% requires more than 10 iterations) and histogram of which bins are most likely to improve PAPR.

To be fair, however, the constellation size has been increased by a factor of four, for a rate loss of 2 b/s/Hz compared to what could be transmitted if there were no restrictions on the choice of constellation points. Since Figure 6-1 shows that to limit the PAPR to be less than 7 dB requires 1 b/s/Hz, the greedy precoding algorithm loses only 1 b/s/Hz compared to what should theoretically be possible. This rate loss is tolerable in the high SNR regime, but if the SNR is very low, this fixed loss in rate can be quite severe. Thus the precoding algorithm is most useful for high rate, high SNR applications.

Although there is a moderate increase in computational complexity at the transmitter to run the greedy search algorithm, the receiver only has to incorporate modulo arithmetic to successfully decode the received signal. The precoding algorithm also does not have the disadvantage of not having perfect knowledge of the interference as TH precoding does, since the interference is caused by the IFFT operation.

6.7.2 Synthesis Simulation Results

Simulations run for the synthesis algorithm show results similar to those for the precoding algorithm. Figure 6-21 shows the results of the phase synthesis algorithm for length 128 OFDM symbols using 64-QAM constellations for the frequency bins. These OFDM symbols were the same ones used in Figure 6-19, so the results can be directly compared. Only the magnitudes of the frequency bins were retained from the 64-QAM encodings, so the resulting time series are no longer uniquely decodable. However, we are only interested here in the reduction in PAPR to compare with the

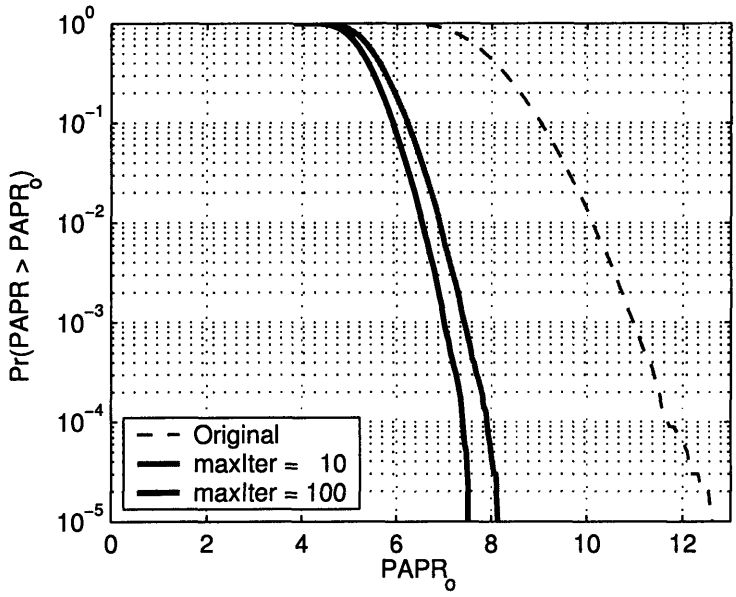


Figure 6-21: CCDF of the PAPR for the phase synthesis algorithm for length 128 OFDM symbols using 64-QAM and 10× oversampling for bandlimited interpolation. The difference between 10 (solid right) and 100 (solid left) iterations is less than 1 dB.

precoding algorithm. Using a magnitude-only constellation would not significantly change these distributions except at very low probabilities due to the finite number of sample signals used.

The PAPR gain of the phase synthesis algorithm is about 1 dB less than the gain from the precoding algorithm. The resulting PAPR is still near the 7 dB needed for the outphasing amplifier to be most efficient. The difference in performance between 100 and 200 iterations is negligible, and only about 0.5 dB is lost by restricting the number of iterations to 10.

It should be noted again that these numbers are for the continuous-time PAPR. If we only considered the discrete-time PAPR, the synthesis algorithm reduces the PAPR to almost 0 dB even though the actual signal is not as constant magnitude as the discrete-time PAPR would suggest. This is a pitfall that is not always considered in other PAPR reduction schemes that have been proposed.

6.8 Algorithm Comparisons

To evaluate the performance of the two algorithms, we consider how their characteristics compare in terms of use with the WiGLAN. The differences in computational

Algorithm	# iterations	MFLOPS
Precoding (FFT)	1	246
Precoding (fast)	1	89
Synthesis	10	7
Synthesis	100	65

Table 6.2: Computational complexity for precoding and synthesis algorithms for a sample OFDM symbol of length 128 using 64-QAM. (1 MFLOP is 10^6 FLOPS)

complexity, rate loss, amplifier efficiency gains, and transmitter energy per bit are compared and then evaluated for use in the WiGLAN.

6.8.1 Computational Complexity Comparisons

The phase synthesis algorithm is a significant reduction in computation over the precoding algorithm because it does not have an exponential search space to optimize over. Regardless of the specifics of how these operations are implemented, in a single iteration of the precoding algorithm the upsampled version of the time signal must be calculated four times for each of the M active bins. Since the average number of iterations required for the greedy precoding algorithm is 5.5, this requires approximately $20M$ upsample and compare operations per OFDM symbol. For the synthesis algorithm with M iterations, $2M$ FFT operations are required, along with M clipping operations. It is not immediately clear which algorithm is less computationally costly due to unknown details about how expensive the upsample is compared to the FFT. A MATLAB simulation of these algorithms with a FLOP (FLoating point OPeration) count is shown in Table 6.2 for several versions of the two algorithms for comparison. Since newer versions of MATLAB no longer support FLOP counts, these simulations were done on version 5.3.

For the precoding algorithm, an FFT version and a "fast" version are used. The FFT version upsamples the time signal via an FFT with zero padding, while the "fast" version calculates the appropriate complex sinusoid offset signal caused by the change in constellation point. The "fast" version of the precoding algorithm is almost four times faster than the FFT version. In comparison, the synthesis algorithm is able to run 100 iterations in fewer FLOPS than a single iteration of the fast precoding algorithm. Thus the synthesis algorithm requires 7.5 times less computation than the precoding algorithm, but loses 1 dB in performance. For less than an extra 0.5 dB loss in performance, the speedup is a factor of 70 over the precoding algorithm.

For the WiGLAN, the difference in computational complexity can be important depending on whether the node is in uplink or downlink. During the uplink, the nodes are transmitting to the server. Since the node has little computational power,

the precoding algorithm might be too complex for the node to support. Although the synthesis algorithm is less complex for nearly the same PAPR reductions, it may still be too complex for the node. During the downlink, the nodes are receiving from the server. Since the server does not have such computational constraints, the extra PAPR reduction of the precoding algorithm favor its use over the synthesis algorithm.

6.8.2 Rate Loss Comparisons

In the high SNR regime, the precoding algorithm uses a fraction of a dB extra transmit power to reduce the PAPR, which leads to a slight loss in the capacity. Alternately, since the transmit constellation is expanded, 2 b/s/Hz is used to lower the PAPR, which is a modest lost at the high capacities in the high SNR regime. The precoding algorithm, on the other hand, does not change the average transmit power, but it does reduce the rate by 1/2 to reduce the PAPR. This is a large reduction in rate at high SNR, making the precoding algorithm more desirable for high data rates.

In the low SNR regime, however, the situation is reversed. The precoding algorithm loses a fixed 2 b/s/Hz regardless of the constellation size, so when the SNR is sufficiently low such that 4-QAM is used, the rate loss for the precoding algorithm is the same as the synthesis algorithm. At this point, the lower complexity of the precoding algorithm prevails. If the SNR is extremely low, then the synthesis algorithm is actually superior, as it incurs no rate loss for very low transmit powers, while the precoding algorithm always has the overhead of the expanded constellation.

For the WiGLAN operational regime, the SNR is typically high to support the gigabit data rates. In this case, the precoding algorithm is most promising because it produces a greater reduction in the maximum PAPR without reducing the code rate as much as the synthesis algorithm.

6.8.3 Amplifier Efficiency Comparisons

Although the precoding algorithm does appear to perform slightly better than the synthesis algorithm in terms of the PAPR distribution, the ultimate goal is to improve the average efficiency of the amplifier. To compare how the precoding and synthesis algorithms affect the average efficiency of the power amplifier, the average efficiencies of the OFDM symbols resulting from the simulations in the previous section were calculated. The same set of OFDM symbols were run through both algorithms so their results can directly compare. To remove the effect of clipping, each OFDM signal was scaled such that the peak power of the signal is exactly at the clipping threshold. Since the transmitter can know exactly what the peak power for an OFDM symbol it is about to transmit, it can prevent any clipping by scaling the signal in this manner before it reaches the digital to analog converters or power amplifier.

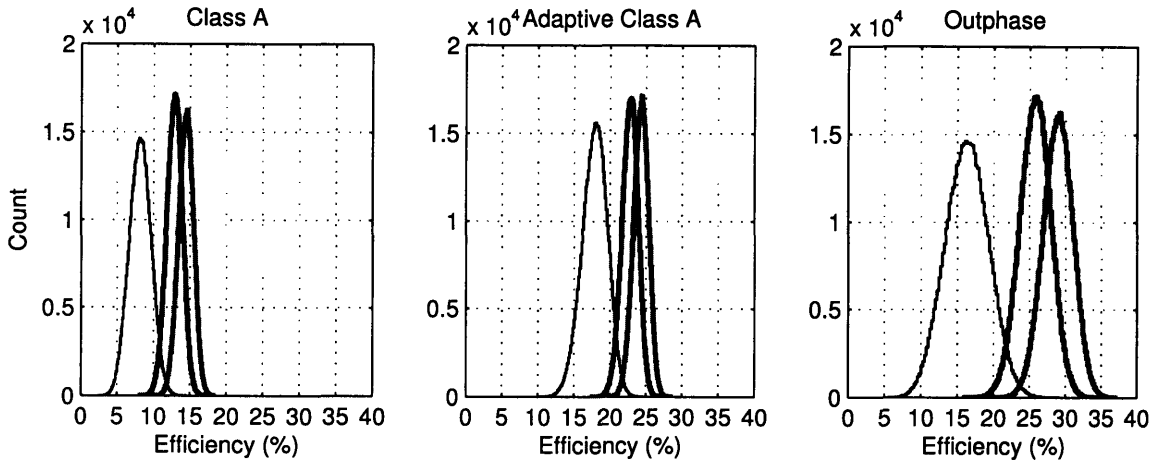


Figure 6-22: Precoding: Average Efficiency for Rayleigh distributed signals through ideal linear amplifiers using length 128 OFDM time signals with 10× oversampling and 64-QAM. Histograms show original PAPR, plus PAPR of random-greedy and greedy algorithms (left to right) with 20% active bins or 25 maximum iterations.

Figure 6-22 shows histograms of the PAPR for each OFDM symbol for each of the three classes of linear amplifiers that have been considered. The greedy algorithm was run with a maximum of 25 iterations (although, as shown earlier, the average number of iterations is 5.5), while the random-greedy algorithm was run with 12, 25, and 64 iterations (for 10%, 20%, and 50% active bins). For the baseline case of OFDM symbols through a class A amplifier, the average efficiency is only 8%. Thus, for example, it would require 1.25 watts to transmit 100 mW or 20 dBm of power. This efficiency is greater than the theoretical curve from Figure 6-5 because the input is scaled with each OFDM symbol, so an OFDM symbol with a low PAPR will be scaled larger than one with a high PAPR.

By only changing the amplifier hardware to the adaptive class A, the average efficiency rises to almost 18%. The outphase amplifier also increases the average efficiency, but only to 16%, which is less than the adaptive A performance. Looking at the histograms, it is apparent that the outphasing amplifier has a much larger spread of efficiencies than either of the class A amplifiers. This is a result of the larger variation in the efficiency curve of the outphasing amplifier.

After applying the greedy algorithm, the increase in efficiency ranges from 6% for the class A amplifier to over 12% for the outphasing amplifier. Note also that the outphasing amplifier outperforms the adaptive class A amplifier by 4% when using the greedy algorithm. As expected the computational savings from the random-greedy

Amplifier Type	Original Signal	R-G (bins) 10%	20%	50%	Greedy 25 max	Synthesis (iter) 10	20	100
Class A	8.1	11.9	12.8	13.8	14.4	14.0	14.3	14.7
Adaptive	17.8	21.8	22.7	23.7	24.2	25.0	25.3	25.8
Outphase	16.2	23.8	25.7	27.7	28.8	27.9	28.5	29.5

Table 6.3: Average efficiencies (%) for precoding and synthesis algorithms.

algorithm come at the cost of slightly reduced efficiency. However, even with only an active bin set of 10% there is a 4–7% improvement depending on the amplifier type. Also, note that the combination of the greedy algorithm and the outphasing amplifier leads to a greater efficiency gain that either is able to achieve alone. The outphasing amplifier alone is able to increase the efficiency by 8%, and the greedy algorithm improves class A performance by 6%, but the combination of the greedy algorithm with the outphasing amplifier increases the efficiency by almost 20%. This gain in efficiency means that the power required to transmit is almost one-fourth that required for the original signal with a class A amplifier. Similar gains can be made by pairing the precoding algorithms with the adaptive class A amplifier, which may require less hardware to implement.

Similarly, with the synthesis algorithm, large gains in the efficiency can be realized. Figure 6-23 shows the results of the synthesis algorithm when coupled with each of the linear amplifier types. With only 20 iterations, the synthesis algorithm performs nearly as well as the greedy algorithm for the class A and the outphasing amplifier. Even with 10 iterations, the performance is very close to what could be achieved with the precoding algorithms. The best performance of all comes with 100 iterations, which still requires fewer computations than just a single iteration of the precoding algorithm. The gains in performance with the increased iterations is only 1–2%, however, so using the smaller number of iterations would appear to be a better trade-off between performance and complexity.

Table 6.3 summarizes the average efficiencies of the precoding and synthesis algorithms with various parameters for the three types of linear amplifiers. In terms of performance increases, the best performance both in terms of total efficiency gain as well as efficiency gain per computation comes from using the synthesis algorithm with the outphasing amplifier. If the outphasing amplifier is too large for the target application, the adaptive class A amplifier is still able to achieve significant gains in efficiency when coupled with either of the PAPR reducing algorithms. Even if it is not possible to change the amplifier from a class A (in an existing system, for example), using either of the PAPR reducing algorithms can nearly double the amplifier efficiency.

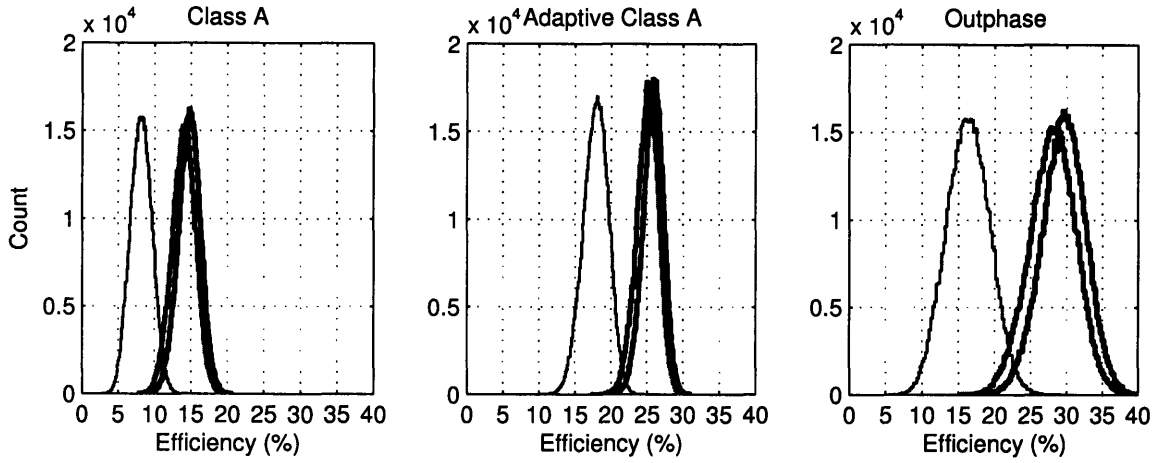


Figure 6-23: Synthesis: Average Efficiency for Rayleigh distributed signals through ideal linear amplifiers using length 128 OFDM time signals with 10× oversampling and 64-QAM. Histograms show original PAPR, plus PAPR of synthesis algorithm after 10 and 100 max iterations (left to right).

6.8.4 Rate Normalized Efficiency Gain (High SNR)

Although the overall efficiency has been increased by up to four times, the rate loss incurred by the by the precoding and synthesis algorithms requires extra transmissions to send the same amount of data. The efficiency gains from the PAPR algorithms must then be normalized to the rate. Although the amplifier efficiency improves, we are really interested in the total energy required to transmit a given amount of data, or the energy required to transmit per bit. The amount of energy used per unit time is proportional to the SNR, and the amount of possible bits is given by the capacity.

At high SNR, the capacity is (repeated from (2.8))

$$C \approx k \log_2(\beta \mathrm{SNR}), \tag{6.51}$$

where $k = \min(M, N)$ represents the number of degrees of freedom. However, by focusing on maximum diversity operation, only one degree of freedom is actually available, with $k = 1$ and $\beta = 1$. In this regime, doubling the capacity requires squaring the SNR, or doubling the SNR in dB.

At an operating point of 30 dB, for example, the capacity is about 35 b/s/Hz. The synthesis algorithm uses up half the capacity, resulting in a maximum throughput of 17.5 b/s/Hz. This could have been accomplished with an SNR of only 15 dB, which is a factor of almost 32 times less power. This is mitigated somewhat by the efficiency

Amplifier Type	Original Signal	R-G (bins) 10%	20%	50%	Greedy 25 max	Synthesis (iter) 10	20	100
Class A	100.0	73.2	68.8	64.4	62.0	115.7	113.3	110.2
Adaptive	45.5	40.0	38.8	37.5	36.9	64.8	64.0	62.8
Outphase	50.0	36.6	34.3	32.1	31.0	58.1	56.8	54.9

Table 6.4: Average energy units (original signal normalized to 100) required per bit after using the precoding and synthesis algorithms in the high SNR regime.

gain from the synthesis algorithm, but it is still a large difference in power. This difference only gets worse with higher SNR values, so the rate penalty is quite high in the high SNR regime.

However, from the standpoint of total energy use, the synthesis algorithm cuts the rate in half, so twice as much time is required to transmit the same amount of data. The factor of 3.64 increase in amplifier efficiency allows the amplifier to consume 5.6 dB less power, so the total energy required for transmission is about 55% that of the original system when using synthesis algorithm with the outphasing amplifier. Note, however, that not only is this not better than the performance of the precoding algorithm, but the overall energy per bit is actually greater than if no PAPR reduction algorithm is used. In the high SNR regime, then, the synthesis algorithm's rate loss of 1/2 is too large a penalty to pay for reducing the PAPR.

For the precoding algorithm, the expansion of the constellation appears to cost 2 b/s/Hz. However, all of the constellation points are not used equally, since only a fraction of the frequency bins are altered by the precoding algorithm. The average power increase for the greedy algorithm is 0.42 dB, and 0.36 dB for the random-greedy algorithm, both with 25 iterations. Using (2.10), this translates to a capacity loss of 0.14 bits for the greedy algorithm, and 0.12 bits for the random-greedy algorithm. Alternately, the 10% increase in power caused by the greedy algorithm only slightly decreases the efficiency increases shown in Table 6.3. Thus, the efficiency gain factor is reduced from 3.56 to 3.23, and from 3.17 to 2.88 for the random-greedy algorithm.

Table 6.4 details the overall savings in energy per bit dissipated by the power amplifier for the precoding and synthesis algorithms at high SNR. After normalizing, Table 6.4 reveals the best performance occurs with the greedy algorithm coupled with the outphasing amplifier, with an overall gain of 5.1 dB in the power dissipated by the PA at the transmitter, although 3.4 dB of this gain can be achieved just by using the adaptive class A amplifier. If the system uses a class A amplifier which cannot be altered, then using the precoding algorithm still results in a 2.1 dB reduction in power. In the high SNR regime, the precoding algorithm is clearly superior to the synthesis algorithm in terms of energy savings.

An interesting thing to note is that although the adaptive class A amplifier is the

Amplifier Type	Original Signal	Synthesis (iter) 10	20	100
Class A	100.0	57.9	56.6	55.1
Adaptive	45.5	32.4	32.0	31.4
Outphase	50.0	29.0	28.4	27.5

Table 6.5: Average energy units (original signal normalized to 100) required per bit after using the synthesis algorithm in the low SNR regime.

most efficient by itself, the combination of the outphasing amplifier with the precoding algorithm outperforms the same algorithm used with the adaptive class A. In addition, the synthesis algorithm actually requires more overall energy per bit to transmit unless the amplifier is upgraded as well. Even when combined with the outphasing amplifier, however, the performance of the synthesis algorithm does not outperform just the original signal passed through the upgraded amplifiers. Thus from an energy standpoint, the synthesis algorithm actually hurts performance because it requires twice as long to transmit a given amount of data. However, the synthesis algorithm may be useful to reduce the instantaneous power to meet a total power specification that the system might have (for example, to accommodate a smaller power supply), while using significantly less computation than the precoding algorithm.

6.8.5 Rate Normalized Efficiency Gain (Low SNR)

In the low SNR regime, the capacity is approximately equal to the SNR, thus halving the capacity halves the required SNR. In this realm, the loss of rate from using the magnitude-only channel falls to zero, thus the use of the synthesis algorithm incurs no rate penalty at all. The precoding algorithm, by contrast always loses 2 b/s/Hz due to the expanded constellation size. At very low SNR, the capacity is small, so the 2 b/s/Hz rate penalty overwhelms the capacity, making use of the precoding algorithm infeasible. Since there is no rate loss from the synthesis algorithm at low SNR, it is well suited for use in reducing the PAPR. The average energy per bit for the synthesis algorithm at low SNR is shown in Table 6.5, with slightly greater reductions in the energy per bit when compared to the precoding algorithm in the high SNR regime.

As was the case with the precoding algorithm at high SNR, the combination of the PAPR reducing algorithm (in this case the synthesis algorithm) and the outphasing amplifier produces the greatest gain in efficiency and the largest reduction in energy per transmitted bit. The two PAPR reducing algorithms are complementary to each other, with the precoding algorithm working well at high SNR, and the synthesis algorithm working well at low SNR.

6.9 Design Guidelines

For the WiGLAN operating regime, the SNR is typically high to support gigabit data rates. During the downlink phase, the central server is transmitting, thus it has significant computational resources that can be used. In addition, the server is typically transmitting at a high power, where efficiency can greatly reduce the total power consumption. The combination of the precoding algorithm with the outphasing amplifier produces the greatest increase in the overall efficiency of the power output for the transmitter, reducing the energy per bit by 5 dB, or more than a factor of three. If the amplifier cannot be changed from a class A amplifier, the precoding algorithm is still able to reduce energy per bit by 60%. On the other hand, the outphasing amplifier is not as effective as the adaptive class A amplifier if no PAPR algorithm is used.

The synthesis algorithm is much less computationally complex than the precoding algorithm, but the large reduction in the data rate can actually increase the power required per bit to transmit at high SNR. At low SNR, however, the synthesis algorithm has no rate loss, making it suitable for very low power, low data rate systems. Although the synthesis algorithm is about 100 times less complex than the precoding algorithm, it may still be too complex for a node with limited capabilities. These PAPR reducing algorithms are more suited to the central server and the more capable high data rate nodes.

For situations where little or no complexity can be tolerated, we have seen that the SNR loss caused by clipping the OFDM symbols is very small for moderate clipping levels, but the bandwidth expansion caused by the nonlinear clipping operation may exceed the spectral mask limits for even moderate clipping levels. The average efficiencies were calculated for three types of linear amplifiers and two input probability distributions, but similar calculations can be easily made for arbitrary input probabilities and amplifier efficiency curves using the same techniques.

Chapter 7

Conclusions and Future Work

As we have seen, some of the problems inherent in analog circuit design can be mitigated by the use of algorithms that have been designed to takes into account the details of analog circuit behavior. There is usually little information about the details of the analog hardware available to the algorithm designer, which can lead to algorithms that that significantly increase the burden on the circuit designer. By using some information about the details of the analog hardware, algorithm designs can be made which work with, not against the characteristics of the analog hardware. The contributions of this thesis and some future research directions are outline below.

7.1 Thesis Contributions

In this thesis, we have studied several aspects of designing algorithms for a wireless network. The major focus is on shaping algorithms to address some of the difficult problems of analog circuits, including area cost, power cost, circuit crosstalk, and peak to average power. Additionally, we have proposed a network architecture that uses a central server and antenna arrays to increase network performance.

7.1.1 Network Architecture

For a wireless network, and for the WiGLAN in particular, we proposed an architecture with a central server and multiple antennas. The uplink/downlink communications model takes advantage of the multiple antennas at the central server to either increase the network throughput or decrease the transmit power. Since all transmissions are made through the central server, the mobile nodes are able to achieve a high multiplexing gain without needing to coordinate their transmissions with each other. Alternately, the nodes can use the server antennas to increase the diversity gain, reducing the amount of transmit power needed. The network communications

are structured so that even single-antenna nodes can achieve multiplexing and diversity gain without needing to be aware of what other nodes are also transmitting. In addition, the computation and memory requirements for the nodes are minimized by offloading most of the processing to the server. These architecture choices allow the nodes to be less complex and therefore inexpensive.

7.1.2 Circuit Area and Power Optimizations

For a wireless node with multiple antennas, we proposed trading off the greatly increased capacity (as compared to a single antenna) off in favor of high diversity. The SNR gain that results for uncoded or moderately coded systems can be used to reduce the area and power consumption of analog circuits.

While digital circuits benefit greatly from the scaling laws allowed by new process technologies, analog circuits do not scale due to the presence of passive elements (inductors, capacitors, resistors). A single inductor, for example, occupies the same space as over 300 transistors. Removing some or all of the larger components significantly reduces the size of analog circuits at the expense of performance. For an LNA, removing the inductors allows four amplifiers to occupy only 50% more space than the original did. The performance loss in terms of noise figure is offset by the SNR gain that can be achieved with increased antennas.

For power considerations, reducing the power at the transmitter by the diversity SNR gain is a straightforward and effective way to reduce power, especially considering the high power and low efficiency that is more prevalent the transmitter than the receiver. Power savings can also be made by reducing the static power dissipation of amplifiers, which reduces their gain. These losses are again offset by the SNR diversity gain.

7.1.3 Circuit Crosstalk Mitigation

The parallel circuit pathways on an integrated circuit will invariably couple with each other, with the resulting crosstalk potentially causing severe amounts of interference. Using a feedback model for the crosstalk, we determined that the crosstalk isolation must be at least as high as the circuit gains to prevent unstable circuit behavior or deep nulls from forming within the system bandwidth. If the two signals which couple have very different powers, then this difference requires extra isolation margin. This can lead to a very high crosstalk isolation restriction if the input signals have a high dynamic range.

Since the crosstalk is essentially time-invariant, a circuit that has a bad crosstalk configuration could be forever stuck with some of its bandwidth unusable due to excessive crosstalk coupling. From a linear crosstalk model, a formula for the asymptotic

SNR loss due to the crosstalk was derived in terms of the matrix singular values. Due to the strong dependence on the phase of the crosstalk SNR loss, we proposed a phase randomization technique using frequency hopping which allows all instantiations of a circuit to experience the average behavior, reducing circuit variability even when there is insufficient bandwidth to span the full range of phases. By improving the worst case circuit performance, the effective crosstalk isolation is increased by 5 dB or more, making it easier to meet the feedback condition above.

7.1.4 Peak to Average Power Control

The high peak to average power that results from OFDM and other multicarrier systems require both high linearity and dynamic range. These demands are difficult for the analog designer to meet, and as a result the amplifier power efficiencies are low. We determined that clipping of the signal to reduce the PAPR causes the most significant degradation by bandwidth expansion rather than an SNR loss. To reduce the PAPR, we proposed two algorithms, both of which can decrease the PAPR by 4 dB or more, resulting in an overall power reduction by a factor of three in the high SNR regime, and almost a factor of four in the low SNR regime when combined with an outphasing linear amplifier.

The PAPR precoding algorithm increases the size of the input signal constellation, creating four equivalent choices for each constellation point. An iterative greedy search algorithm can then be used to find the best choice of points to reduce the PAPR. The rate loss due to the increased constellation size is 2 b/s/Hz, although in practice the transmit power has only been increased by a fraction of a dB. This algorithm works best in the high SNR regime with a threefold reduction in transmitter power consumption, with only a 1 b/s/Hz loss in excess of the minimum rate loss required.

The PAPR synthesis algorithm, on the other hand, works well in the low SNR regime, with no loss in rate while achieving similar gains in transmitter power as the precoding algorithm. With this algorithm, only the Fourier magnitude is used to transmit information, with the Fourier phase being used to reduced the PAPR. When coupled with the outphasing amplifier at low SNR, the transmit power consumption is reduced by almost a factor of four.

7.2 Future Research Directions

In addition to the areas that have been covered in this thesis, there are many other areas where knowledge of some of the details of the analog circuitry present in any wireless system can lead to better algorithm design. Following the general principle

of trading off some resource that may exist in abundance to relax a difficult requirement, algorithms can be designed to relax a difficult design constraint in other analog components.

For example, the circuit area and power optimizations focused on the receiver RF chain, specifically on the LNA. While removal of inductors reduced the LNA to one-third its original size, the LNA is only a fraction of the size of the receiver chain. Filters also use many passive components to sharply reject unwanted frequencies. The SNR gain may be able to be used here and in other circuits such as the mixer and data converters to enable novel circuit topologies which use less area or power at the expense of added noise.

Similarly, for the crosstalk isolation, the focus was on using diversity-like techniques to reduce the variability in circuit crosstalk degradation. However, since the crosstalk has the same effect as the wireless channel, coding techniques may be able to be developed which are resistant to interference from crosstalk. In addition, other methods besides frequency hopping to randomize the crosstalk phase may be found. The effective crosstalk gain from the phase randomization may also allow novel circuit designs to be made which have otherwise undesirably low crosstalk isolation.

The peak to average power reduction algorithms are able to reduce the PAPR significantly, but both have some shortcomings, either in complexity or rate loss. Although there has been significant amounts of research already done on PAPR reduction algorithms, a joint algorithm and circuit design, similar to what resulted with the outphasing amplifier, may result in even greater gains in power efficiency. In addition, the PAPR algorithms have typically been designed assuming high rate, high SNR operation. However, like the synthesis algorithm, other methods for reducing the PAPR for very low SNR regimes may exist which have better performance or less complexity.

Finally, for the network architecture, only the communications aspect of the physical layer was explored, but there are many other issues, such as the actual mechanics of resource allocation, packet design, etc, which have not been fully explored. The WiGLAN architecture was intended for a high bandwidth, high data rate indoor application, but the use of a central server and antenna arrays may also be advantageous for other operating regimes.

Appendix A

Notation

Below is a listing of the symbol notations used in this thesis:

c, x	scalar constant
x, y	scalar
$\mathbf{x}$, $\mathbf{y}$	vector
$\mathbf{H}$, $\mathbf{C}$, $\mathbf{\Sigma}$	matrix
x_i or $[\mathbf{x}]_i$	element of a vector
x_{ij}, $[\mathbf{H}]_{ij}$ or $\mathbf{H}(i,j)$	element of a matrix
$\mathbf{C}^{\mathrm{T}}$	non-conjugate transpose of $\mathbf{C}$
$\mathbf{C}^{\dagger}$	conjugate or Hermitian transpose
$M \times N$	matrix: m rows and n columns
	antenna system: M transmit antennas, N receive antennas
$\mathcal{N}(a,b)$	Gaussian distribution with mean a and variance b
$\mathcal{N}^{\mathrm{c}}(a,b)$	complex Gaussian distribution with mean a and complex variance b
$\mathcal{O}(a)$	complexity grows with order a
$\|x\|$	magnitude of a scalar
$\|\|\mathbf{x}\|\|$	Euclidean norm of a vector

Appendix B

Tables of SNR Values

The tables provided in this chapter are meant to aid decisions on the number of antennas to use for a system design based on the SNR margin that is calculated from the link budget as in Section 4.2. Tables B.1 through B.5 show the SNR required at the decoder input to decode the data stream at the specified BER. All of the throughputs are normalized for the rate loss from using the error correction codes. Tables B.6 and B.7 show numerical values for the SNR gain curves in Figures 3-10 and 4-3.

BER	1000 Mb/s		900 Mb/s		540 Mb/s		
10^{-1}	Uncoded	(255,241)	Uncoded	(255,241)	Uncoded	(255,241)	(255,187)
1×1	26.96	28.73	24.66	25.78	18.17	18.74	21.64
2×2	21.28	23.07	18.94	20.09	12.20	12.79	15.83
3×3	19.03	20.82	16.67	17.83	9.85	10.45	13.53
4×4	17.65	19.44	15.30	16.46	8.47	9.08	12.16
5×5	16.63	18.42	14.28	15.43	7.43	8.04	11.13
6×6	15.75	17.54	13.40	14.55	6.53	7.14	10.24
7×7	15.08	16.86	12.73	13.88	5.90	6.50	9.59
1×2	22.03	23.81	19.71	20.85	13.04	13.63	16.63
1×3	19.75	21.53	17.41	18.56	10.67	11.27	14.30
1×4	18.15	19.94	15.81	16.96	9.08	9.67	12.71
1×5	17.13	18.91	14.78	15.93	7.98	8.59	11.66

Table B.1: Table for adaptive modulation SNR values (dB) for 1000, 900, and 540 Mb/s assuming a system bandwidth of 150 MHz with a uncorrected BER = 10^{-1}. The throughput is normalized to account for the rate loss of the (255,k) RS codes.

BER	1000 Mb/s		900 Mb/s		540 Mb/s		
10^{-2}	Uncoded	(255,241)	Uncoded	(255,241)	Uncoded	(255,241)	(255,187)
1×1	38.26	40.05	35.92	37.07	29.15	29.74	32.79
2×2	27.49	29.30	25.09	26.27	17.88	18.52	21.81
3×3	24.61	26.43	22.20	23.39	14.91	15.56	18.89
4×4	23.02	24.83	20.61	21.80	13.30	13.95	17.30
5×5	21.89	23.70	19.47	20.66	12.15	12.81	16.16
6×6	20.97	22.79	18.55	19.74	11.21	11.86	15.22
7×7	20.19	22.01	17.78	18.97	10.43	11.08	14.44
2×2	29.78	31.58	27.40	28.57	20.32	20.95	24.17
3×3	26.44	28.24	24.05	25.22	16.89	17.53	20.79
4×4	24.44	26.25	22.04	23.22	14.85	15.49	18.77
5×5	23.09	24.90	20.68	21.87	13.42	14.06	17.38

Table B.2: Table for adaptive modulation SNR values (dB) for 1000, 900, and 540 Mb/s assuming a system bandwidth of 150 MHz with a uncorrected BER = 10^{-2}. The throughput is normalized to account for the rate loss of the (255,k) RS codes.

BER	1000 Mb/s		900 Mb/s		540 Mb/s		
10^{-3}	Uncoded	(255,241)	Uncoded	(255,241)	Uncoded	(255,241)	(255,187)
1×1	48.30	50.08	45.96	47.11	39.26	39.85	42.86
2×2	31.30	33.12	28.89	30.08	21.57	22.22	25.57
3×3	27.62	29.44	25.19	26.39	17.80	18.46	21.85
4×4	25.76	27.58	23.33	24.53	15.90	16.56	19.98
5×5	24.50	26.32	22.07	23.27	14.63	15.30	18.71
6×6	23.51	25.33	21.09	22.29	13.67	14.34	17.74
7×7	22.76	24.58	20.33	21.53	12.85	13.52	16.95
2×2	35.44	37.25	33.05	34.23	25.92	26.55	29.80
3×3	30.83	32.64	28.42	29.60	21.16	21.81	25.12
4×4	28.28	30.09	25.87	27.05	18.55	19.21	22.55
5×5	26.57	28.38	24.15	25.34	16.82	17.47	20.83

Table B.3: Table for adaptive modulation SNR values (dB) for 1000, 900, and 540 Mb/s assuming a system bandwidth of 150 MHz with a uncorrected BER = 10^{-3}. The throughput is normalized to account for the rate loss of the (255,k) RS codes.

BER	1000 Mb/s		900 Mb/s		540 Mb/s		
10^{-4}	Uncoded	(255,241)	Uncoded	(255,241)	Uncoded	(255,241)	(255,187)
1×1	58.42	60.22	56.07	57.22	49.25	49.85	52.92
2×2	34.40	36.21	31.98	33.17	24.62	25.28	28.65
3×3	29.82	31.64	27.39	28.59	19.94	20.60	24.02
4×4	27.63	29.45	25.20	26.40	17.75	18.42	21.84
5×5	26.28	28.10	23.85	25.05	16.41	17.08	20.50
6×6	25.26	27.09	22.82	24.03	15.30	15.97	19.43
7×7	24.44	26.26	22.00	23.20	14.48	15.15	18.61
2×2	40.47	42.28	38.07	39.25	30.93	31.55	34.79
3×3	34.59	36.41	32.17	33.37	24.84	25.49	28.85
4×4	31.45	33.27	29.03	30.22	21.63	22.29	25.68
5×5	29.31	31.13	26.89	28.09	19.51	20.16	23.55

Table B.4: Table for adaptive modulation SNR values (dB) for 1000, 900, and 540 Mb/s assuming a system bandwidth of 150 MHz with a uncorrected BER = 10^{-4}. The throughput is normalized to account for the rate loss of the (255,k) RS codes.

BER	1000 Mb/s		900 Mb/s		540 Mb/s		
10^{-5}	Uncoded	(255,241)	Uncoded	(255,241)	Uncoded	(255,241)	(255,187)
1×1	68.61	70.36	66.35	67.45	59.99	60.57	63.41
2×2	37.09	38.90	34.68	35.86	27.35	28.00	31.36
3×3	31.60	33.43	29.16	30.37	21.64	22.30	25.75
4×4	29.17	30.99	26.74	27.94	19.27	19.94	23.38
5×5	27.70	29.53	25.25	26.46	17.68	18.35	21.83
6×6	26.70	28.54	24.25	25.46	16.62	17.30	20.82
7×7	25.40	27.21	23.01	24.19	15.74	16.40	19.73
2×2	46.01	47.84	43.59	44.79	36.37	36.99	40.27
3×3	37.85	39.65	35.47	36.64	28.31	28.96	32.23
4×4	34.47	36.30	32.02	33.23	24.45	25.12	28.60
5×5	31.67	33.49	29.25	30.44	21.79	22.46	25.89

Table B.5: Table for adaptive modulation SNR values (dB) for 1000, 900, and 540 Mb/s assuming a system bandwidth of 150 MHz with a uncorrected BER = 10^{-5}. The throughput is normalized to account for the rate loss of the (255,k) RS codes.

	2 b/s/Hz	4 b/s/Hz	6 b/s/Hz	8 b/s/Hz	10 b/s/Hz
2×2	5.25	7.29	9.56	12.07	14.75
3×3	7.83	10.64	13.65	16.89	20.34
4×4	9.54	12.74	16.15	19.78	23.61
5×5	10.79	14.24	17.92	21.78	25.85

Table B.6: SNR gain (dB) for $N \times N$ systems compared to a 1×1 at the same capacity.

	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
2×2	5.88	10.95	17.28	24.12	32.65
3×3	8.21	13.90	20.98	28.74	38.41
4×4	9.51	15.46	22.85	30.90	40.58
5×5	10.52	16.59	24.12	32.23	42.26
1×2	4.98	8.58	13.09	18.16	24.13
1×3	7.36	11.94	17.72	23.95	31.66
1×4	8.98	13.97	20.28	27.10	35.50
1×5	9.95	15.39	22.01	29.25	38.09

Table B.7: SNR gain (dB) for uncoded systems compared to a 1×1 at the same BER.

Appendix C

Outphasing Amplifiers

C.1 Using Outphasing Amplifiers

The outphasing amplifier is a way to create a linear amplifier from two or more nonlinear ones [8]. The nonlinear amplifiers are each more efficient than the linear one they replace, but the overall efficiency of the outphasing amplifier is dependent on the characteristics of the signals it has to amplify. Figure C-1 shows how an outphasing system works. The input x is discrete time samples of the desired complex waveform. In the mapping block, it is decomposed into two complex signals s_1 and s_2 which are constant amplitude but with varying phases such that $x = s_1 + s_2$. The outputs of the mapping block are then converted to analog before feeding into the amplifiers. The amplifiers are both highly efficient and nonlinear. Since the input is constant amplitude at baseband, the actual input at passband is a single frequency tone, and the output of the amplifier (possibly after filtering), is an amplified version of the same frequency tone with some fixed phase offset. The two amplified signals y_1 and y_2 are then summed to produce the desired output y, which nominally is a scaled version of the input x.

As we have seen, the efficiency of the outphasing amplifier is higher than the other linear amplifiers only when the output is near the maximum. When the output power is low, however, the efficiency of the outphasing technique is less than other viable options, so use of the outphasing amplifier must be coupled with a input signal designed to have a low PAPR.

C.1.1 The Combining Circuit

The summing circuit is pictured on the right of Figure C-1. Because the two signals to be added are different and at RF, a waveguide-type combining circuit such as the Wilkinson combiner pictured is required [77]. As shown in the diagram, the length

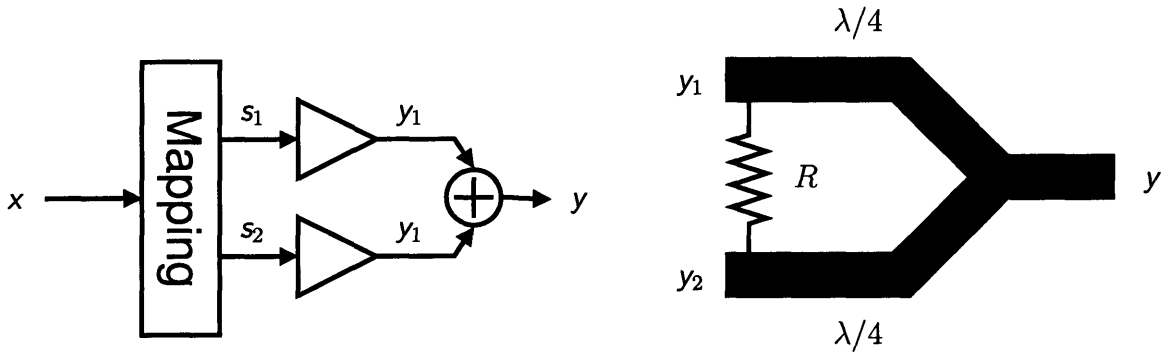


Figure C-1: Outphasing system (left) with Wilkinson combiner (right).

of each of the arms of the Wilkinson combiner is $\lambda/4$ where λ is the wavelength of the frequency to be summed. An implication of this is that the summing circuit only works ideally as a summer at one frequency (typically the carrier frequency), but as long as the signal bandwidth is sufficiently narrow the approximation is good. At 5.8 GHz, $\lambda/4$ is about 2.8 mm, which is very large for an integrated circuit [49]. Assuming a loop path for each $\lambda/4$ branch, that would require a square with perimeter of 5.6 mm, for a total chip area of almost 2 mm^2. By contrast, the RX test chip produced in Chapter 4 has four analog front ends but is only twice this size.

The behavior of the combining circuit can be seen in terms of the even and odd symmetric modes of the circuit [9]. It is assumed that the two waveguide branches are lossless, so the only places that power can go is through the output port, the input ports, or the resistor connecting the two inputs. If $y_2 = y_1$, then the voltage difference across the resistor R is zero, so no power is dissipated across the resistor and instead adds at the output y. If instead $y_2 = -y_1$, the output y is grounded by symmetry, so all of the power is dissipated across the resistor R.

Thus for an arbitrary input pair $\mathbf{y}$,

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}}_{\mathbf{y}} = \underbrace{\left(\frac{y_1+y_2}{2}\right)}_{y_e}\begin{bmatrix} 1 \\ 1 \end{bmatrix} + \underbrace{\left(\frac{y_1-y_2}{2}\right)}_{y_o}\begin{bmatrix} 1 \\ -1 \end{bmatrix}, \tag{C.1}$$

where y_e and y_o are the even and odd components of $\mathbf{y}$. Since the combiner adds the power of the even component, the output power is

$$||\mathbf{y}||^2 = 2|y_e|^2 = \frac{|y_1+y_2|^2}{2}. \tag{C.2}$$

The Wilkinson combiner can in theory losslessly add the two signals together if they

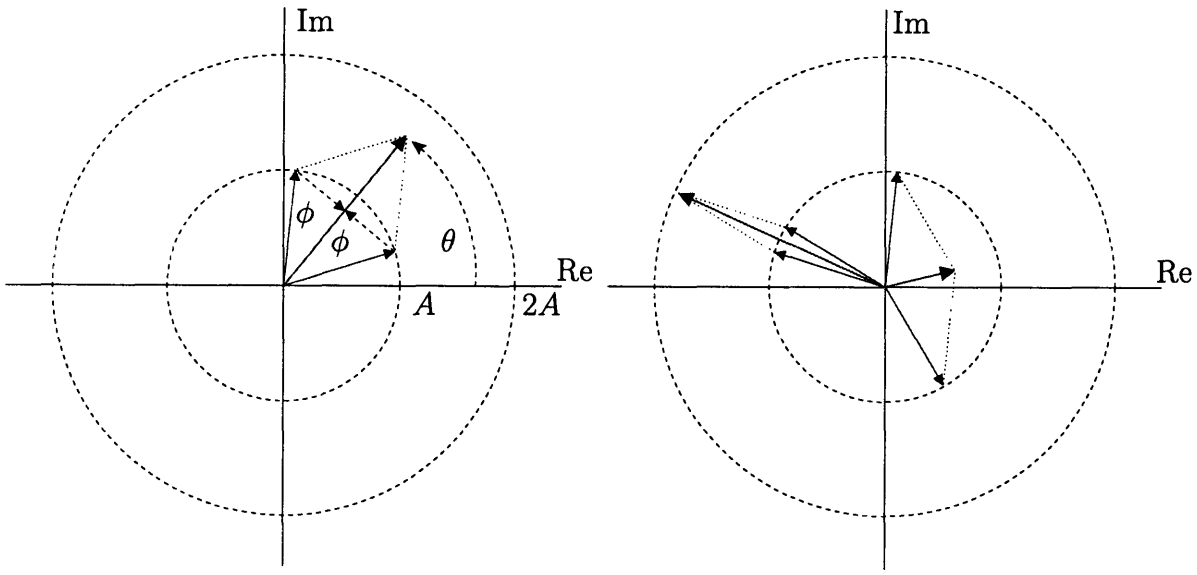


Figure C-2: Vector combining of two constant amplitude signals in the outphased amplifiers.

are equal, but if they are unequal or have different phases, then some of the input power is dissipated by the resistor.

C.1.2 Vector Representation of a Signal

The input signal x can be represented as a time-varying complex vector which is the sum of two constant amplitude signals s_1 and s_2, such that

$$x(t) = s_1(t) = s_2(t), \tag{C.3}$$

where the magnitude of s_i is a constant A. Since the s_i are complex as well, $s_i(t) = Ae^{j\theta_i(t)}$. These signals can be viewed as time-varying vectors on the complex plane, as shown in Figure C-2. The inner dotted circle has radius A and represents the possible values of the outphasing vectors $s_i(t)$. The outer dotted circle has radius $2A$ and represents the largest amplitude that can be formed by the outphasing vectors.

We can infer from the power combining in (C.2) that the vectors are added as

$$y = \frac{y_1 + y_2}{\sqrt{2}}, \tag{C.4}$$

where the combiner acts as an adder scaled by a fractional constant. If y_1 and y_2

are the same, the resulting amplitude is $\sqrt{2}|y_1|$, which has double the power of y_1, as expected. The vector addition picture therefore correctly models the phase while the amplitude is reduced by a constant factor.

If the angle between the two outphase vectors is 2θ, then the perpendicular component of each outphase vector has length $A\sin(\theta)$. These form the odd component of the overall signal (equal in magnitude and with opposite directions), and are dissipated by the combiner resistor. As diagrammed in Figure C-2, if the desired output amplitude is large, the angle θ between the two outphase vectors will be small, and very little power is dissipated by the combiner. If, on the other hand the desired output amplitude is very small, θ will be large and most of the power will be dissipated by the combiner. The efficiency curve of the outphasing amplifier in Figure 6-3 reflects this, as a constant amount of power is burned by the amplifier regardless of how much is actually output. Although this amplifier configuration has a very high peak efficiency, in order to have a high average efficiency, the PAPR of the output signal must be kept small.

C.2 Outphasing Bandwidth Expansion

Although the two outphasing signals are constant amplitude, their phases will vary significantly to form the desired signal. The desired output waveform is typically bandlimited to some set of frequencies, but the constant amplitude outphase signals do not have to be, and in general are not bandlimited. The individual outphase signals are infinite bandwidth in general, and rely on perfect additive sidelobe cancellation to keep the output signal bandlimited. Even though the effect of gain and phase mismatches in the two outphase amplifiers has been shown to be small [50], some mismatches will always exist which can lead to unwanted out of band leakage. One way to prevent bandwidth expansion is to form the outphase signals digitally rather than in the analog domain. The outphase signals still experience bandwidth expansion, but in the digital domain this out of band energy is aliased back into the original bandwidth.

If we represent the desired signal by

$$y(t) = M(t)e^{j\theta(t)}, \tag{C.5}$$

where $M(t)$ and $\theta(t)$ are the time varying magnitude and phase of the desired signal $y(t)$. The magnitude function $M(t)$ is assumed to be limited to the range $[0, 2]$ for convenience, and the signal is bandlimited to some maximum frequency W. From the outphasing equation (and ignoring the constant factor), we know that

$$y(t) = y_1(t) + y_2(t), \tag{C.6}$$

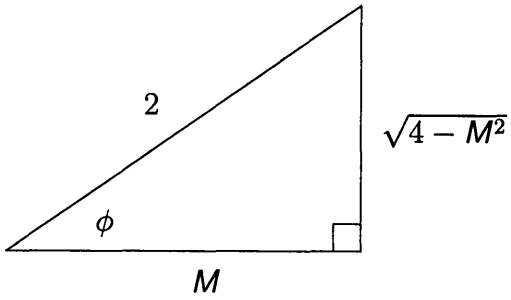


Figure C-3: Triangle for sin/cos property.

with

$$y_i(t) = e^{j\theta_i(t)}, \tag{C.7}$$

where $\theta_i(t)$ is the time varying phase of each outphase signal. Using the convention given by Figure C-2,

$$\begin{aligned} \theta_1(t) &= \theta(t) + \phi(t) \\ \theta_2(t) &= \theta(t) - \phi(t). \end{aligned} \tag{C.8}$$

Since the even parts of the two outphase signals will add to form the final amplitude, then

$$M(t) = 2\cos(\phi(t)) \Rightarrow \phi(t) = \cos^{-1}\left(\frac{M(t)}{2}\right). \tag{C.9}$$

Looking more closely at one of the outphasing vectors, $y_1(t)$ can be rewritten as

$$y_1(t) = e^{j\theta_1(t)} = e^{j\theta(t)}e^{j\phi(t)} \tag{C.10}$$

$$= e^{j\theta(t)}\left[\cos\phi(t) + j\sin\phi(t)\right] \tag{C.11}$$

$$= e^{j\theta(t)}\left[\cos\cos^{-1}\left(\frac{M(t)}{2}\right) + j\sin\cos^{-1}\left(\frac{M(t)}{2}\right)\right] \tag{C.12}$$

$$= \frac{M(t)}{2}e^{j\theta(t)} + je^{j\theta(t)}\sin\cos^{-1}\left(\frac{M(t)}{2}\right). \tag{C.13}$$

The trigonometric term can be simplified by using the triangle in Figure C-3, which shows a right triangle with appropriate lengths such that $\cos\phi = M/2$. Since the magnitude $M \in [0, 2]$, this triangle is valid for all values of $\phi \in [0, \pi/2]$.

Using Figure C-3, $y_1(t)$ can be simplified to

$$y_1(t) = \frac{1}{2}y(t) + je^{j\theta(t)}\frac{\sqrt{4 - M^2(t)}}{2} \tag{C.14}$$

$$= \frac{1}{2}y(t) + e^{j[\theta(t)+\pi/2]}\sqrt{1-\frac{M^2(t)}{4}}. \tag{C.15}$$

Although the first term of $y_1(t)$ is bandlimited to W, the bandwidth of the second term is unbounded. This is evident from the Taylor expansion

$$\sqrt{1-\frac{M^2(t)}{2}} = 1 - \frac{M^2(t)}{4} - \frac{M^4(t)}{32} + \cdots. \tag{C.16}$$

If $y(t)$ is bandlimited to W, then $M(t)$ is also bandlimited to W because $M(t)$ is equal to $y(t)$ convolved with an all pass (phase-only) function. If $M(f)$ is the Fourier transform of $M(t)$, then $M^2(t)$ corresponds to $M(f)$ convolved with itself. If a function $M(f)$ is nonzero for frequencies less than W, then $M(f) * M(f)$ (with $*$ denoting convolution) will in general be nonzero for frequencies less than $2W$. Similarly, $M^n(t)$ will have nonzero frequency components for frequencies less than nM. And since the Taylor series expansion is infinite, the resulting bandwidth of the outphase signal is infinite. However, since the coefficients of the higher order terms are also decreasing, only a finite amount of the excess bandwidth will be detrimental by rising above the existing noise floor. Using (C.6), the second outphase signal has the form

$$y_1(t) = \frac{1}{2}y(t) - e^{j[\theta(t)+\pi/2]}\sqrt{1-\frac{M^2(t)}{4}}, \tag{C.17}$$

where the second term is the odd part of the outphase signal which is dissipated by the combiner resistor.

Although the addition of the two outphase signals should result in the desired bandlimited output signal, if there is any mismatch in the gain or phase between the two branches of the combiner or amplifiers, then the odd part of the signal will not cancel exactly, and some of it will appear in the output. For example, Figure C-4 shows the frequency spectrum for an OFDM symbol of length 128 using 64-QAM. The frequency axis has been normalized so that the original bandwidth is two, and ten times the original bandwidth is shown. The right plot shows the frequency spectrum of one of the constant amplitude outphase signals, with its correspondingly larger bandwidth. The expanded bandwidth is as high as 10 dB below the original signal level, thus even a small amount of mismatch between the two outphase signals will result in significant power outside of the desired bandwidth.

Since it is undesirable to exceed the allocated system bandwidth, the signal output which is outside the allowed frequency band must be removed. The output of the outphase combiner can be filtered to remove the excess bandwidth, but this can result in the loss of a significant amount of output power. In the example outphase signal of Figure C-4, the power in the out-of-band frequencies is more than 10% of the total

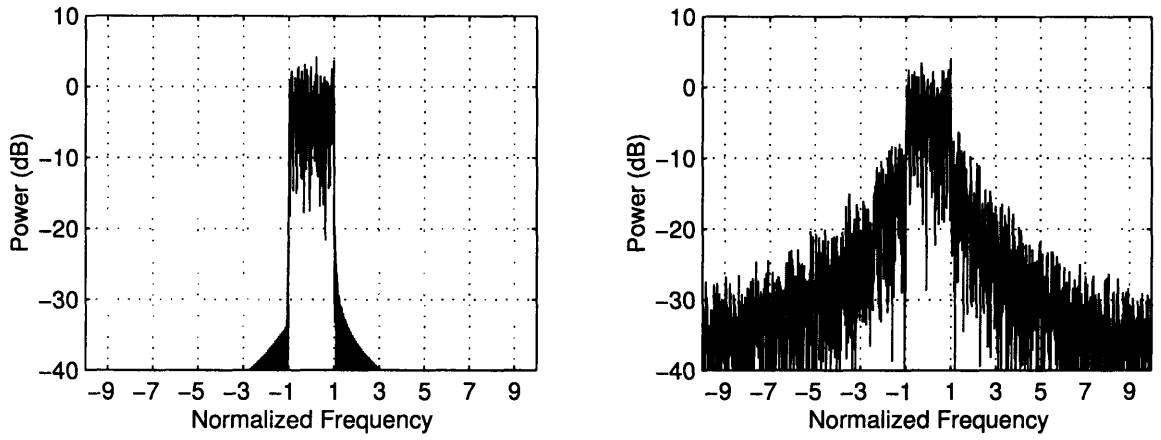


Figure C-4: Frequency spectrum of original signal (left) and outphase signal (right).

output power. Having a filter after the amplifier is also undesirable because of the significant loss in efficiency that occurs even through the passband of the filter. A typical insertion loss of a filter is 3 dB [13], which leads to a immediate efficiency loss of 50%, even ignoring the amplifier output energy that is out-of-band.

A way to avoid these efficiency losses is to remove the out-of-band frequency components before they are amplified, such as by placing the filters directly before the amplifier inputs. This will also reduce the efficiency losses because the filter inputs were not amplified. The process of filtering out some of the outphase signal destroys its constant amplitude property. For example, Figure C-5 shows the effects of limiting the bandwidth of one of the outphasing signals, as shown in Figure C-4.

Since the original signal has infinite bandwidth, limiting the bandwidth to any finite value will cause the resulting signal to no longer be constant amplitude. In the figure, the discrete time signal was oversampled by a factor of ten (representing ten times the system bandwidth) to approximate the continuous-time signal. For signal oversampling ratios larger than π, the actual peak of the continuous-time signal is at most

$$B(k) = \begin{cases} \dfrac{k^2}{k^2 - \pi^2/2} & k > \pi/\sqrt{2} \\ \dfrac{2}{\pi} \ln 2N + 2 & k = 1 \end{cases} \tag{C.18}$$

times the peak of the oversampled signal, where k is the oversampling ratio [57]. For 10× oversampling, the worst-case PAPR error is therefore 0.44 dB. The analog bandlimited interpolated signals are approximated by 10× oversampling in all the PAPR simulations. The power is scaled such that 0 dB represents the original constant

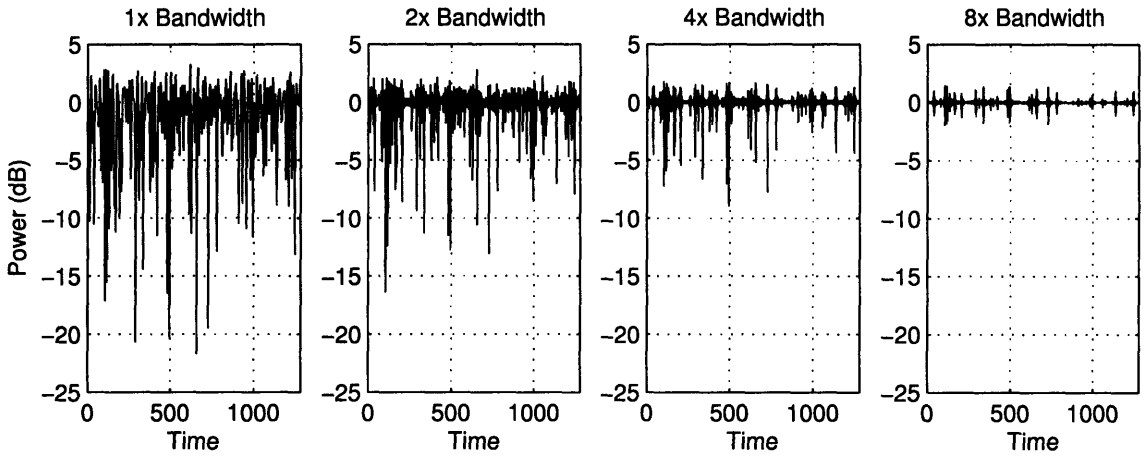


Figure C-5: Instantaneous power for bandlimited versions of the outphase signal.

amplitude signal. If the bandwidth of the outphase signal is limited to the system bandwidth as in the leftmost plot of Figure C-5, the resulting signal is very far from constant amplitude.

As expected, increasing the allowable bandwidth for the outphase signal brings it closer to a constant amplitude signal, but a great deal of excess bandwidth is required. For example, the rightmost plot shows that using eight times the system bandwidth results in low amplitude variations.

Bibliography

[1] N. Abramson, "The ALOHA system: another alternative for computer communications," in *AFIPS Conference Proceedings*, vol. 37, Montvale, NJ, 1970, pp. 281–285. 63

[2] N. Amitay, V. Galindo, and C. P. Wu, *Theory and Analysis of Phased Array Antennas.* New York: Wiley-Interscience, 1972. 171

[3] B. D. O. Anderson and S. Vongpanitlerd, *Network Analysis and Synthesis: A Modern Systems Theory Approach.* Prentic-Hall, Inc., 1973. 114

[4] C. Bergljung and P. Karlsson, "Propagation Characteristics for Indoor Broadband Radio Access Networks in the 5 GHz Band," in *The Ninth IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, vol. 2, Boston, MA, Sept. 1998, pp. 612–616. 79

[5] R. E. Blahut, *Principles and Practice of Information Theory.* Reading, MA: Addison Wesley, 1987. 173

[6] O. M. Bucci, G. Mazzaralla, and G. Panariello, "Reconfigurable Arrays by Phase-Only Control," *IEEE Transactions on Antennas and Propagation*, vol. 39, no. 7, pp. 919–925, July 1991. 175

[7] A. P. Chandrakasan, S. Sheng, and R. W. Brodersen, "Low-Power CMOS Digital Design," *IEEE Journal of Solid-State Circuits*, vol. 27, no. 4, pp. 473–484, Apr. 1992. 50, 96

[8] H. Chireix, "High power outphasing modulation," *Proc. IRE*, vol. 33, no. 6, pp. 1370–1392, Nov. 1935. 45, 203

[9] S. B. Cohn, "A Class of Broadband Three-Port TEM-Mode Hybrids," *IEEE Transactions on Microwave Theory and Techniques*, vol. MTT-16, no. 2, pp. 110–116, Feb. 1968. 204

[10] P. L. Combettes, "The Foundations of Set Theoretic Estimation," *Proceedings of the IEEE*, vol. 81, no. 2, pp. 182–208, Feb. 1993. 175

[11] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* New York: John Wiley & Sons, Inc., 1991. 31, 65, 172

[12] J. A. Davis and J. Jedwab, "Peak-to-Mean Power Control in OFDM, Golay Complementary Sequences, and Reed-Muller Codes," *IEEE Transactions on Information Theory*, vol. 45, no. 7, pp. 2397–2417, Nov. 1999. 145, 165

[13] X. Duo, L.-R. Zheng, H. Tenhunen, L. Chen, G. Zou, and J. Liu, "Design and implementation of a 5 GHz RF receiver front-end in LCP based system-on-package modules with embedded chip technology," in *Proc. IEEE Electrical Performance of Electronic Packaging*, New Jersey, Oct. 2003, pp. 51–54. 165, 209

[14] F. Edalat, "Effect of power amplifier nonlinearity on system performance metric, bit-error-rate (BER)," Master's thesis, MIT, 2003. 24

[15] A. Edelman, "Eigenvalues and Condition Numbers of Random Matrices," Ph.D. dissertation, MIT, 1989. 36

[16] D. Farnese, A. Leva, G. Paltenghi, and A. Spalvieri, "Pulse Superposition: A Technique for Peak-to-Average Power Ratio Reduction in OFDM Modulation," in *Proc. IEEE International Conference on Communications*, vol. 3, Milano, Italy, Apr. 2002, pp. 1682–1685. 145, 159

[17] G. J. Foschini and M. J. Gans, "On Limits of Wireless Communications in a Fading Environment when Using Multiple Antennas," *Wireless Personal Communications*, vol. 6, no. 3, pp. 311–335, Mar. 1998. 33, 35

[18] A. Gatherer and E. Auslander, Eds., *The Application of Programmable DSPs in Mobile Communications.* John Wiley & Sons, Ltd., 2002. 95, 96

[19] A. Gatherer and M. Polley, "Controlling Clipping Probability in DMT Transmissions," in *Conference Record of the Thirty-First Asilomar Conference on Signals, Systems, & Computers*, Nov. 1997, pp. 578–584. 145

[20] G. Ginis and J. M. Cioffi, "A multi-user precoding scheme achieving crosstalk cancellation with application to DSL systems," in *Conference Record of the Thirty-Fourth Asilomar Conference on Signals, Systems, & Computers*, 2000, pp. 1627–1631. 59, 60, 61

[21] M. J. E. Golay, "Note on the theoretical efficiency of information reception with PPM," *Proc. IRE*, vol. 37, p. 1031, Sept. 1949. 173

[22] M. Grossglauser and D. Tse, "Mobility increases the capacity of ad-hoc wireless networks," *Proceedings of INFOCOM '01*, vol. 3, pp. 1360–1369, 2001. 76

[23] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 388–404, Mar. 2000. 53, 76

[24] ——, "Towards an information theory of large networks: an achievable rate region," in *Proceedings of ISIT '01*, Washington, D.C., June 2001, p. 159. 55, 76

[25] H. Harashima and H. Miyakawa, "Matched-transmission technique for channels with intersymbol interference," *IEEE Transactions on Communications*, vol. 38, pp. 774–780, Aug. 1972. 59

[26] M. H. Hayes, J. S. Lim, and A. V. Oppenheim, "Signal Reconstruction from Phase or Magnitude," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, no. 6, pp. 672–680, Dec. 1980. 171, 174

[27] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. Cambridge: Cambridge University Press, 1999. 115, 116, 120, 130

[28] E. Huang, L. Khuon, C. Sodini, and G. Wornell, "An Approach for Area- and Power-Efficienct Low-Complexity Implementation of Multiple Antenna Transceivers," in *Proceedings of IEEE Radio and Wireless Symposium (RWS) 2006*, San Diego, CA, Jan. 2006. 84

[29] X. Huang, J. Lu, J. Chuang, and J. Zheng, "Companding transform for the reduction of peak-to-average power ratio of OFDM signals," in *Proceedings of the IEEE Vehicular Technology Conference (VTC) 2001, Spring*, 2001, pp. 835–839. 145

[30] *Supplement to IEEE standard for information technology telecommunications and information exchange between systems - local and metropolitan area networks - specific requirements. Part 11: wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: high-speed physical layer in the 5 GHz band*, IEEE Std 802.11a-1999. 29, 79, 165

[31] I. Jacobs, "Limits on the Power and Spectral Efficiency of Direct Detection Systems with Optical Amplifiers," in *Conference Record of the Thirty-Fifth Asilomar Conference on Signals, Systems, & Computers*, vol. 1, Pacific Grove, CA, Nov. 2001, pp. 8–12. 172

[32] D. L. Jones, "Peak Power Reduction in OFDM and DMT via Active Channel Modification," in *Conference Record of the Thirty-Third Asilomar Conference*

on Signals, Systems, and Computers, vol. 2, Pacific Grove, CA, Oct. 1999, pp. 1076–1079. 145

[33] L. Khuon, "LNA test chip and crosstalk measurement data," personal communication. 133, 135

[34] L. Khuon, E. Huang, C. Sodini, and G. Wornell, "Integrated Transceiver Arrays for Multiple Antenna Systems," in *Proceedings of the IEEE Vehicular Technology Conference (VTC) 2005, Spring*, Stockholm, Sweden, May 2005. 84, 98

[35] L. Khuon and C. G. Sodini, "Parallel Integrated Receiver Front-Ends for a 5.25 GHz Wireless Gigabit LAN," in *MTL Annual Research Report*, MIT, 2005. 99

[36] H. L. Krauss, C. W. Bostian, and F. H. Raab, *Solid State Radio Engineering.* John Wiley & Sons, 1980. 45

[37] X. Li and J. Leonard J. Cimini, "Effects of clipping and filtering on the performance of OFDM," in *Proceedings of the IEEE Vehicular Technology Conference (VTC) 1997, Spring*, Phoenix, AZ, May 1997, pp. 1635–1638. 145, 159, 166

[38] D. Lu and D. Rutledge, "Investigation of Indoor Radio Channels for 2.4 GHz to 24 GHz," in *IEEE Antennas and Propagation Society International Symposium*, vol. 2, June 2003, pp. 134–137. 33

[39] Iñigo Cuiñas, M. S. Varela, and M. G. Sánchez, "Wide band indoor radio channel measurements at 5.8 GHz," in *Proceedings of the IEEE Vehicular Technology Conference (VTC) 2000, Fall*, Boston, MA, Sept. 2000, pp. 695–702. 29, 77

[40] J. T. E. McDonnell, "5 GHz Indoor Channel Characterization: Measurements and Models," in *IEE Colloquium on Antennas and Propagation for Future Mobile Communications*, London, Feb. 1998, pp. 10/1–10/6. 33

[41] J. Medbo and J.-E. Berg, "Spatio-Temporal Channel Characteristics at 5 GHz in a Typical Office Environment," in *Proceedings of the IEEE Vehicular Technology Conference (VTC) 2001, Fall*, vol. 3, Atlantic City, NJ, Oct. 2001, pp. 1256–1260. 33

[42] J. Medbo, H. Hallenberg, and J.-E. Berg, "Propagation Characteristics at 5 GHz in Typical Radio-LAN Scenarios," in *Proceedings of the IEEE Vehicular Technology Conference (VTC) 1999, Spring*, vol. 1, Houston, TX, May 1999, pp. 185–189. 79

[43] S. Mehta, D. Weber, M. Terrovitis, K. Onodera, M. M. B. Kaczynski, H. Samavati, S. Jen, W. Si, M. Lee, K. Singh, S. Mendis, P. Husted, N. Zhang, B. McFarland, D. Su, T. Meng, and B. Wooley, "An 802.11g WLAN SoC," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, San Francisco, CA, Feb. 2005. 95

[44] S. H. Müller and J. B. Huber, "OFDM with reduced peak-to-average power ratio by optimum combination of partial transmit sequences," *Electronics Letters*, vol. 33, no. 5, pp. 368–369, Feb. 1997. 167

[45] A. V. Oppenheim and R. W. Shafer, *Discrete-Time Signal Processing.* Englewood Cliffs, NJ: Prentice-Hill, 1989. 163

[46] K. G. Paterson and V. Tarokh, "On the Existence and Construction of Good Codes with Low Peak-to-Average Power Ratios," *IEEE Transactions on Information Theory*, vol. 46, no. 6, pp. 1974–1987, Sept. 2000. 144, 145

[47] A. D. Pham, "PA test chip and crosstalk measurement data," personal communication. 133

[48] ——, "Biasing techniques for linear power amplifiers," Master's thesis, MIT, 2002. 45, 150

[49] ——, "Outphase Power Amplifiers in OFDM Systems," Ph.D. dissertation, MIT, 2005. 151, 204

[50] A. D. Pham, G. W. Wornell, and C. G. Sodini, "A Digital Amplitude-to-phase Conversion for High Efficiency Linear Outphse Power Amplifiers," in *Proc. ICASSP '06*, submitted. 206

[51] H. V. Poor and G. W. Wornell, Eds., *Wireless Commmunication: Signal Processing Perspectives.* Prentice-Hall, Inc., 1998. 41

[52] B. M. Popovič, "Synthesis of Power Efficient Multitone Signals with Flat Amplitude Spectrum," *IEEE Transactions on Communications*, vol. 39, no. 7, pp. 1031–1033, July 1991. 145

[53] J. G. Proakis, *Digital Communications*, 3rd ed. New York: Mcgraw-Hill, Inc., 1995. 126

[54] C. Rapp, "Effects of HPA-Nonlinearity on a 4-DPSK/OFDM-Signal for a Digital Sound Broadcasting System," in *Proceedings of the Second European Conference on Satellite Communications*, Liege, Belgium, Oct. 22–24, 1991, pp. 179–184. 46

[55] T. S. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. Upper Saddle River, NJ: Prentice Hall, 2002. 33

[56] H. Samavati, H. R. Rategh, and T. H. Lee, "A 5–GHz CMOS wireless LAN receiver front end," *IEEE Journal of Solid-State Circuits*, vol. 35, no. 5, pp. 765–772, May 2000. 99

[57] M. Sharif, M. Gharavi-Alkhansari, and B. H. Khalaj, "New Results on the Peak Power of OFDM Signals Based on Oversampling," in *Proc. IEEE International Conference on Communications*, vol. 2, Helsinki, Finland, Apr. 2002, pp. 866–871. 159, 209

[58] M. Sharif and B. Hassibi, "Asymptotic Probability Bounds on the Peak Distribution of Complex Multicarrier Signals Without Gaussian Assumption," in *Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems, & Computers*, vol. 1, Nov. 2002, pp. 171–175. 148

[59] ——, "A determinisitic algorithm that achieves the PMEPR of $c \log n$ for multicarrier signals," in *Proc. ICASSP '03*, 2003, pp. 540–543. 145, 148

[60] ——, "On the Existence of Codes with Constant Bounded PMEPR for Multicarrie Signals," in *Proc. ISIT '03*, Yokohama, Japan, June 29 – July 4 2003, p. 130. 148

[61] ——, "On Multicarrier Signals Where the PMEPR of a Random Codeword is Asymptotically $\log n$," *IEEE Transactions on Information Theory*, vol. 50, no. 5, pp. 895–903, May 2004. 145, 148

[62] M.-C. Shin and I.-C. Park, "A Programmable Turbo Decoder for Multiple 3G Wireless Standards," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, San Francisco, CA, Feb. 2003, p. 154. 95

[63] B. Sklar, "Rayleigh Fading Channels in Mobile Digital Communication Systems Part I: Characterization," *IEEE Communications Magazine*, vol. 35, no. 7, pp. 90–100, July 1997. 76, 79

[64] H. Stark, W. C. Catino, and J. L. LoCicero, "Design of phase gratings by generalized projections," *J. Opt. Soc. Am. A*, vol. 8, no. 3, pp. 566–571, Mar. 1991. 175

[65] S. Stefanou, J. S. Hamel, P. Baine, M. Bain, B. M. Armstrong, H. S. Gamble, M. Kraft, and H. A. Kemhadjian, "Ultralow Silicon Substrate Noise Crosstalk Using Metal Faraday Cages in an SOI Technology," *IEEE Transactions on Electron Devices*, vol. 51, no. 3, pp. 486–491, Mar. 2004. 109

[66] I. E. Teletar, "Capacity of multi-antenna gaussian channels," *AT&T-Bell Laboratories, Internal Tech. Memo*, June 1995. 34

[67] J. Tellado and J. Cioffi, "Efficient Algorithms for Reducing PAR in Multicarrier Systems," in *Intl. Symposium Info. Theory*, 1998, p. 191. 145

[68] J. Tellado and J. M. Cioffi, "Peak Power Reduction for Multicarrier Transmission," in *Proc. IEEE Globecom Comminication Theory Mini-Conference (CTMC)*, Sydney, Australia, Nov. 1998, pp. 219–224. 145, 168

[69] J. Thomas F. Quatieri and A. V. Oppenheim, "Iterative Techniques for Minimum Phase Signal Reconstruction from Phase or Magnitude," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-29, no. 6, pp. 1187–1193, Dec. 1981. 171

[70] M. Tomlinson, "New automatic equaliser employing modulo arithmetic," *Electronics Letters*, vol. 7, pp. 138–139, Mar. 1971. 59

[71] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication.* Cambridge University Press, 2005. 37, 40, 41, 69

[72] I. Vassiliou, K. Vavelidis, T. Georgantas, S. Plevridis, N. Haralabidis, G. Kamoulakos, C. Kapnistis, S. Kavadias, Y. Kokolakis, P. Merakos, J. C. Rudell, A. Yamanaka, S. Bouras, and I. Bouras, "A single–chip digitally calibrated 5.15–5.825-GHz 0.18 μm CMOS transceiver for 802.11a wireless LAN," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 12, pp. 2221–2231, Dec. 2003. 99

[73] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic Beamforming Using Dumb Antennas," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1277–1294, June 2002. 67

[74] S. P. Voinigescu, M. A. Copeland, D. Marchesan, P. Popescu, and M. C. Maliepaard, "5–GHz SiGe HBT monolithic radio transceiver with tunable filtering," in *Proc. IEEE RFIC '99*, Anaheim, CA USA, June 1999, pp. 131–134. 99

[75] E. W. Weisstein, "Harmonic addition theorem," From *MathWorld* – A Wolfram Web Resource., http://mathworld.wolfram.com/HarmonicAdditionTheorem.html. 113

[76] D. A. Wiegandt, C. R. Nassar, and Z. Wu, "Overcoming peak-to-average power ratio issues in OFDM via carrier-interferometry codes," in *VTC '01*, vol. 2, 2001, pp. 660–663. 145

[77] E. J. Wilkinson, "An N-Way Hybrid Power Divider," *IRE Transactions on Microwave Theory and Techniques*, vol. MTT-8, no. 1, pp. 116–118, Jan. 1960. 203

[78] P. W. Wolniansky, G. J. Foschini, G. D. Golden, and R. A. Valenzuela, "V-BLAST: an architecture for realizing very high data rates over the rich-scattering wireless channel," in *Proccedings of ISSSE'98*, Pisa, Italy, Oct. 1998, pp. 295–300. 55, 56, 141

[79] C. Wu and C. Chou, "A 5–GHz CMOS double quadrature receiver front–end with single-stage quadrature generator," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 3, pp. 519–521, Mar. 2004. 99

[80] J. H. Wu, J. Scholvin, J. A. del Alamo, and K. A. Jenkins, "A Faraday Cage Isolation Structure for Substrate Crosstalk Suppression," *IEEE Microwave and Wireless Components Letters*, vol. 11, no. 10, pp. 410–412, Oct. 2001. 105

[81] M. Zargari, D. K. Su, C. P. Yue, S. Rabii, D. Weber, B. J. Kaczynski, S. S. Mehta, K. Singh, S. Mendis, and B. A. Wooley, "A 5–GHz CMOS transceiver for IEEE 802.11a wireless LAN systems," *IEEE Journal of Solid-State Circuits*, vol. 37, no. 12, pp. 1688–1694, Dec. 2002. 99

[82] L. Zheng and D. Tse, "Diversity and Multiplexing: A Fundamental Tradeoff in Multiple-Antenna Channels," *IEEE Transactions on Information Theory*, vol. 49, no. 5, pp. 1073–1096, May 2003. 38, 39, 41