ECE 5325/6325: Wireless Communication Systems
Lecture Notes, Fall 2011

Prof. Neal Patwari
University of Utah
Department of Electrical and Computer Engineering

Contents

Lecture 1

Today: (1) Syllabus, (2) Cellular Systems Intro

1 Cellular Systems Intro

1.1 Generation Zero

The study of the history of cellular systems can help us understand the need for the system design concepts we have today.

One of the major developments in WWII was the miniaturization of FM radio components to a backpack or handheld device (the walkie-talkie), a half-duplex (either transmit or receive, not both) push-to-talk communication device. After returning from war, veterans had the expectation that wireless communications should be available in their civilian jobs [26]. But the phone system, the Public Switched Telephone Network (PSTN) was: wired, and manually switched at telephone exchanges. In 1952, the Mobile Telephone System (MTS) was designed to serve 25 cities in the US [11] (including one in Salt Lake City [10]). In each city, an additional telephone exchange office was created for purpose of connection with the mobile telephones [26]. The MTS and later the improved mobile telephone system (IMTS), introduced in 1964, were not particularly spectrally efficient.

- They were allocated a total bandwidth of about 2 MHz. Frequency modulation (FM) was used. For multiple user access, the system operated frequency division multiple access (FDMA), in which each channel was allocated a non-overlapping frequency band within the 2 MHz.
- The PTSN is full duplex (transmit and receive simultaneously) in IMTS, so it required two channels for each call, one uplink (to the base station) and one downlink (to the mobile receiver). Note MTS had been half duplex, *i.e.*, only one party could talk at once.
- The FCC required them to operate over an entire city (25 mile radius). Since the coverage was city wide, and coverage did not exist outside of the cities, there was no need for handoff.
- Initially channels were 120 kHz [7], due to poor out-of-band filtering. The channel bandwidth was cut to 60 kHz in 1950 and again to 30 kHz in 1965. Thus there were 2 MHz / 2 / 120 kHz or 8 full duplex channels at the start, and up to 32 in 1965, *for the entire city*.

Control was manual, and the control channel was open for anyone to hear. In fact, users were required to be listening to the control channel. When the switching operator wanted to connect to any mobile user, they would announce the call on the control channel. If the user responded, they would tell the user which voice channel to turn to. Any other curious user could listen as well. A mobile user could also use the control channel to request a call to be connected. The system was congested, so there was always activity.

The demand was very high, even at the high cost of about $400 per month (in 2009 dollars). There were a few hundred subscribers in a city [11] but up to 20,000 on the waiting list [26]. The only way to increase the capacity was to allocate more bandwidth, but satisfying the need would have required more bandwidth than was available.

The downsides to MTS took a significant amount of technological development to address, and the business case was not clear (AT&T developed the technologies over 35 years, but then largely ignored it during the 1980s when it was deployed [11]).

1.2 Cellular

The cellular concept is to partition a geographical area into "cells", each covering a small fraction of a city. Each cell is allocated a "channel group", *i.e.*, a subset of the total list of channels. A second cell, distant from a first cell using a particular channel group, can reuse the same channel group. This is called "frequency reuse". This is depicted in Figure 3.1 in Rappaport. This assumes that at a long distance, the signals transmitted in the first cell are too low by the time they reach the second cell to significantly interfere with the use of those channels in the second cell.

There are dramatic technical implications of the cellular concept. First, rather than one base station, you need dozens or hundreds, deployed across a city. You need automatic and robust mobility management (handoff) to allow users to cross cell lines and continue a phone call. Both of these are actually enabled by semiconductor technology advancement, which made the base stations and the automated wired PSTN cheaper [26].

Frequency reuse and handoff are topics for upcoming lectures.

1.3 Key Terms

Communication between two parties (a "link"), in general, can be one of the following:

- *Simplex*: Data/Voice is transferred in only one direction (*e.g.*, paging). Not even an acknowledgement of receipt is returned.
- *Half Duplex*: Data/Voice is transferred in one direction at a time. One can't talk and listen at the same time. One channel is required.
- *Full Duplex*: Data/Voice can be transferred in both directions between two parties at the same time. This requires two channels.

In a cellular system, there is full duplex communication, between a base station and a mobile. The two directions are called either uplink (from mobile to base station) or downlink (from BS to mobile). The downlink channel is synonymous with "forward channel"; the uplink channel is synonymous with the "reverse channel".

Simultaneous communication on the many channels needed for many users (radios) to communicate with a base station can be accomplished by one (or a combination) of the following *multiple access* methods.

- *Frequency division multiple access (FDMA)*: Each channel occupies a different band of the frequency spectrum. Each signal can be upconverted to a frequency band by multiplying it by a sinusoid at the center frequency of that band, and then filtering out any out-of-band content (see ECE 3500).
- *Time division multiple access (TDMA)*: Every period of time can be divided into short segments, and each channel can be carried only during its segment. This requires each device to be synchronized to have the same time clock.

- *Code division multiple access (CDMA)*: Many channels occupies the same frequency band, at the same time. However, each channel occupies a different "code channel". Like sinusoids at different frequencies are orthogonal (non-interfering), sets of code signals can also be made so that all code signals are orthogonal to each other. One user's channel is multiplied by one code in the set, and at the receiver, can be separated from the other signals by filtering (like frequency bands can be filtered to remove out-of-band content).

See Figures 9.2 and 9.3, pages 450-453, in the Rappaport book.

Physical "parts" of a cellular system:

1. Public switched telephone network (PSTN): Wired telephone network, connecting homes, businesses, switching centers.
2. Mobile switching center (MSC), a.k.a. mobile telephone switching office (MTSO): Controls connection of wireless phone calls through the base stations to the PSTN. Connected either by wire or by wireless (microwave relay) to the base stations.
3. Base station (BS): Maintains direct wireless connection to cell phones in its cell. Typically maintains many connections simultaneously. Has multiple antennas, some for downlink and some for uplink.

See Figure 1.5, page 14, in the Rappaport book.

In cellular systems, there are actually two types of channels: (1) Control, and (2) Communication. The control channel is needed to tell the mobile device what to do, or for the mobile to tell the BS or MSC what to do. The communication channel is the "voice channel" or data channel, the actual information that the user / system needs to convey in order to operate. Since we also have a forward and reverse channel (for full duplex comms), we have

1. FCC: Forward control channel
2. FVC: Forward voice channel(s)
3. RCC: Reverse control channel
4. RVC: Reverse voice channel(s)

Quite a bit of work goes into planning for frequency reuse. We have two goals. First, a radio should be in range of at least one BS; so BSes must have a certain density so to cover all of an area. Next, a radio must avoid *co-channel interference* from other BSes using the same channel, which means that base stations using the same channel should be widely separated. These are conflicting goals! The first section of this course will teach you how to engineer a cellular system that works to a desired specification.

Lecture 2

Today: (1) Frequency Reuse, (2) Handoff

2 Frequency Reuse

2.1 Transmit Power Limits

A cell is the area in which a mobile is served by a single BS. What is the power transmitted by the radios in a cell system? Limits differ by country.

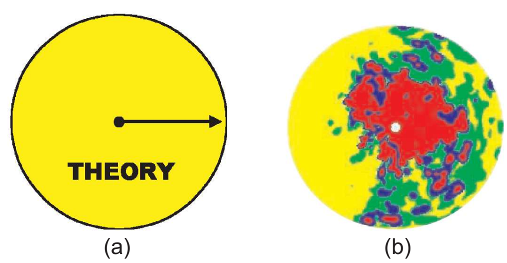
1. Base station maximum = 100 W maximum Effective Radiated Power (ERP), or up to 500 W in rural areas [27]
2. Cell phone maximum: typically 0.5 W; but limited by power absorbed by human tissue in test measurements (of specific absorption rate).

Cell phone exposure limits are typically set to meet strictest of US / European / other standards.

2.2 Cellular Geometry

When the signal from the base station becomes too weak, the mobile will not be able to be served by the BS. This defines the outer limit of a cell's coverage area. What shape is a cell? See Figure 1. These are in order from most to least accurate:

1. A shape dependent on the environment.
2. Circular (theoretical): If path loss was a strictly decreasing function of distance, say, $1/d^n$, where d is the distance from BS to mobile and n is the "path loss exponent", then the cell will be a perfect circle. This is never really true, but is often used to get a general idea.
3. An approximation to the theoretical shape: required for a tessellation (non-overlapping repetitive placement of a shape that achieves full coverage. Think floor tiles.) Possible "tile" shapes include triangles, squares, hexagons. Hexagons are closest to reality.



(a) (b)

Figure 1: Theoretical coverage area, and measured coverage area. In (b), from measurements, with red, blue, green, and yellow indicating signal strength, in decreasing order. From Newport et. al. [19].

As we mentioned in lecture 1, a cellular system assigns subsets, "channel groups", of the total set of channels to each cell. Call the total number of channels S, and the number of channel groups N. Then there are on average $k = S/N$ channels per cell, and N cells per cluster. (In reality, k may vary between groups.) Then with N channel groups, how do we assign them? We want cells that reuse group A, for example, to be as far apart as possible.

The total number of channels in a deployment are S times the number of clusters in our deployment area. If we're limited by spectrum (number of channels) and want to increase the capacity over a fixed area, we want to maximize the number of clusters, or minimize the area covered by any particular cluster. This is why we might use smaller and smaller cell diameters as we want to increase our system capacity. *What is the radius R of a cell?* (From C. Furse) *Macrocell*: $R > 2000$ feet, up to 25 miles; *Microcell*: $200 < R < 1000$ feet; *Picocell*: $R \approx 100$ feet.

2.2.1 Channel Assignment within Group

See Section 3.3. Which channels should be assigned to a cell? First, it is best to separate channels in the group in frequency as much as possible to reduce *adjacent channel interference* (studied later). But which channels are assigned? Two ways:

1. *Fixed assignment*: Each base station has a fixed set of channels to use. Simple, but a busy cell will run out of channels before a neighboring cell. System performance will be limited by the most crowded cell.

2. *Dynamic allocation*: Each base station can change the channels it uses. Channels in neighboring cells must still be different. This requires more careful control, but increases the capacity.

For example, a typical city needs more channels in its business districts during the day, and in its residential areas at night and on weekends.

For general shapes, this can be seen as a graph coloring problem, and is typically covered in a graph theory course. For hexagons, we have simple channel group assignment. Consider $N = 3$, 4, 7, or 12 as seen in Figure 2. A tessellation of these channel groupings would be a "cut and paste" tiling of the figure. The tiling of the $N = 4$ example is shown in Figure 3.

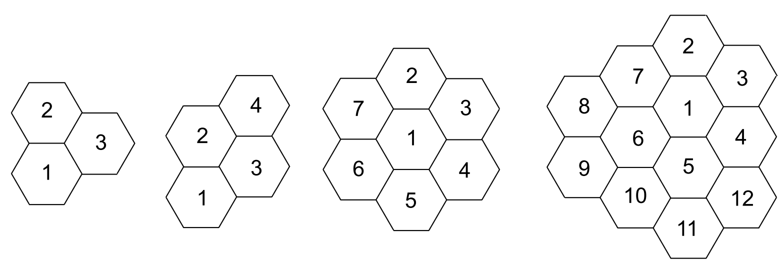


Figure 2: Hexagonal tessellation and channel groupings for $N = 3$, 4, 7, and 12.

Example: Call capacity of $N = 4$ system
Assume that 50 MHz is available for forward channels, and you will deploy GSM. Each channel is 200 kHz, but using TDMA, 8 simultaneous calls can be made on each channel. How large is k? How many forward calls can be made simultaneously for the cellular system depicted in Figure 3?
Solution: There are 50 MHz / 0.2 MHz or 250 total channels. With $N = 4$, then $k = 250/4 = 62.5$, and with (about) 62.5 channels, $8(62.5) = 500$ calls can be made simultaneously in each cell. There are 28 cells on the cell map in Figure 3, so the total forward calls is $28(500) = 14 \times 10^3$ calls can be made simultaneously.

Why wouldn't you choose N as low as possible? There are interference limits, which will be discussed in more detail in Section 2.4.

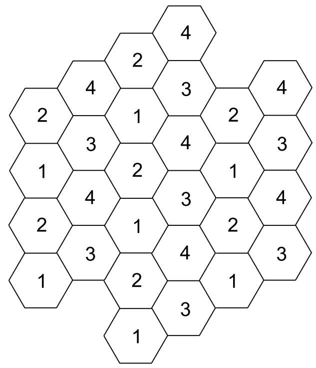


Figure 3: Frequency reuse for $N = 4$.

How do you generally "move" from one cell to the co-channel cell (a second cell assigned the same channel group)? All cellular tiling patterns can be represented using two non-negative integers, i and j. The integer i is the number of cells to move from one cell in one direction. Then, turn 60 degrees counter-clockwise and move j cells in the new direction. For Figure 3, this is $i = 2$, $j = 0$. In this notation, the number of cells can be shown to be:

$$N = i^2 + ij + j^2$$

What is the distance between two co-channel cell BSes? If the distance between the BS and a vertex in its cell is called R, its "radius", then you can show (good test question?) this co-channel reuse distance D is:

$$D = R\sqrt{3N}$$

The ratio of $D/R = \sqrt{3N}$ is called Q, the co-channel reuse ratio.

2.3 Large-scale Path Loss

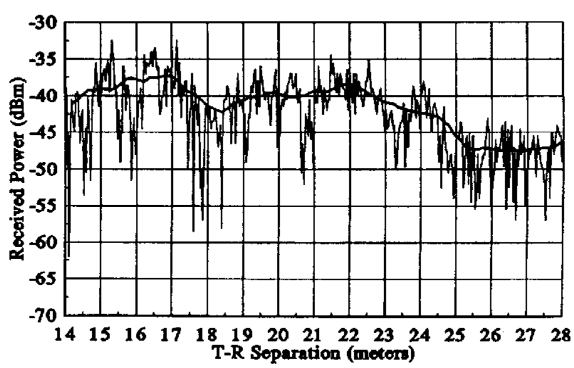


Figure 4: Figure 4.1 from Rappaport, measured received power along a linear path, along with its windowed average.

We will discuss fading in detail in subsequent lectures. For now, consider that received power varies quickly (see the measurement in Figure 4). Figure 4 also shows that a time-average of the received power is less variable but is not purely decreasing as distance between TX and RX (*path length*) increases. However, broadly speaking, if we could average out many measurements of received signal power over many links with path length d, we would see a received power P_r (in mW) that decayed proportional to $1/d^n$, where n is called the path loss exponent. Proportionality gives us $P_r = c/d^n$ for some constant c. If we choose $c = P_0 d_0^n$, then

$$P_r = (P_0 d_0^n)\frac{1}{d^n} = P_0\left(\frac{d_0}{d}\right)^n = P_0\left(\frac{d}{d_0}\right)^{-n}$$

for *reference power* P_0 and *reference distance* d_0. In this form, P_0 is the average power at a reference distance d_0. Typically d_0 is taken to be something small, like 1 meter or 10 meters. Converting to power in dBm,

$$\underbrace{10\log_{10} P_r}_{P_r(\text{dBm})} = \underbrace{10\log_{10} P_0}_{P_0(\text{dBm})} - \underbrace{10n\log_{10}\frac{d}{d_0}}_{\text{dB Pathloss}}$$

Question: How much does your average received power change when you double your path length?

Sidenote: How can you read your phone's received power in dBm? There are "field test modes" that differ by phone. For Nokia (I'm told) and iPhones, dial the number *3001#12345#* which puts you into field test mode. On my iPhone 3GS, this puts the dBm received power in the upper left corner of the screen.

2.4 Co-Channel Interference

What is the ratio of signal power to interference power? This is the critical question regarding the limits on how low we may set N. This ratio is abbreviated S/I. Signal power is the desired signal, from the base station which is serving the mobile. The interference is the sum of the signals sent by co-channel base stations, which is not intended to be heard by mobiles in this cell. The S/I ratio is defined as:

$$\frac{S}{I} = \frac{S}{\sum_{i=1}^{i_0} I_i}$$

where I_i is the power received by the mobile from a co-channel BS, of which there are i_0, and S is the power received by the mobile from the serving BS. NOTE: All powers in the S/I equation above are *LINEAR* power units (Watts or milliWatts), not dBm.

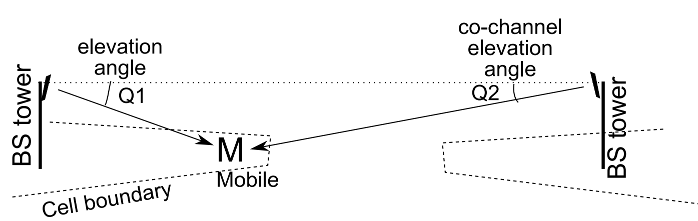


Figure 5: Desired, and interfering signal for a mobile (M) from a serving and co-channel base station.

As a first order, before we get more complicated, we model the received power as inversely proportional to distance to the n power, for some constant path loss exponent n:

$$S = cd^{-n}$$

for some real valued constant c.

We typically look at the worst case, when the S/I is the lowest. This happens when the mobile is at the vertex of the hexagonal cell, *i.e.*, at the radius R from the serving BS. So we know $S = cR^{-n}$. What are the distances to the neighboring cells from the mobile at the vertex? This requires some trigonometry work. The easiest approximation is (1) that only the first "tier" of co-channel BSes matter; (2) all mobile-to-co-channel-BS distances are approximately equal to D, the distance between the two co-channel BSes. In this case,

$$\frac{S}{I} = \frac{S}{\sum_{i=1}^{i_0} I_i} \approx \frac{cR^{-n}}{i_0(cD^{-n})} = \frac{(D/R)^n}{i_0} = \frac{(3N)^{n/2}}{i_0} \tag{1}$$

where i_0 is the number of co-channel cells in the first tier. For all N, we have $i_0 = 6$ (try it out!); this will change when using sector antennas, so it is useful to leave i_0 as a variable in the denominator. It is useful to report the S/I in dB, because S/I requirements are typically given in dB.

Example: AMPS design
Assume that 18 dB of S/I is required for acceptable system operation. What minimum N is required? Test for $n = 3$ and $n = 4$.
Solution: 18 dB is $10^{18/10} = 10^{1.8} = 63.1$. Using (1), we need $\frac{(3N)^{n/2}}{6} \geq 63.1$, so

$$N \geq \frac{1}{3}[6(63.1)]^{2/n}$$

For $n = 3$, $N = 17.4$; for $n = 4$, $N = 6.5$. Clearly, a high path loss exponent is important for frequency reuse.

2.4.1 Downtilt

The Rappaport does not cover antenna downtilt, but it is an important practical concept. Compare the elevation angles from the BS to mobile (Q1 in Figure 5) and co-channel BS to the mobile (Q2 in Figure 5). Note Q2 is lower (closer to the horizon) than from the serving BS. The great thing is, we can provide less gain at angle Q2 than at Q1, by pointing the antenna main lobe downwards. This is called downtilt. For example, if the gain at Q1 is 5 dB more than the gain at Q2, then the we have added 5 dB to the S/I ratio. Having a narrow beam in the vertical plane is also useful to reduce the delay spread and thus inter-symbol interference (ISI) [4], which we will introduce in the 2nd part of this course. This narrow vertical beam is pointed downwards, typically in the range of 5-10 degrees. The effect is to decrease received power more quickly as distance increases; effectively increasing n. This is shown in Figure 6. How do you calculate the elevation angle from a BS to a mobile? This angle is the inverse tangent of the ratio between BS height h_t and horizontal distance from the mobile to BS, d. But, at very low ratios, we can approximate $\tan^{-1}(x) \approx x$. So the angle is h_t/d.

Ever wonder why base station antennas are tall and narrow? *The length of an antenna in any dimension is inversely proportional to the beamwidth in that dimension.* The vertical beamwidth needs to be low (5-10 degrees), so the antenna height is tall. The horizontal pattern beamwidths are typically wide (120 degrees or more) so the antenna does not need to be very wide. For more information consult [12].

Discussion: What are some of the problems with coverage and frequency reuse vs. what the Rappaport book has presented?

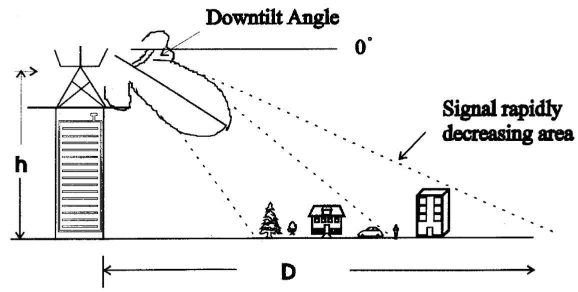


Figure 6: A diagram of a BS antenna employing downtilt to effectively increase the path loss at large distances. From [15].

2.5 Handoff

See Section 3.4. As a mobile travels beyond the coverage region of its serving BS, it must be transferred to better BS. If the average received power drops too low prior to handoff, the call is "dropped". Rappaport denotes this minimum average received power, below which a call cannot be received, as $P_{r,minimum\,useable}$. We want to initiate a handoff much prior to this point, so we set a higher threshold $P_{r,handoff}$ at which the MSC initiates the handoff procedure.

Instantaneous power may go down or up very quickly due to multipath fading. The time-averaged received power will be less variable, but will still vary due to changes in the path length (and thus the large-scale path loss) and due to shadowing. Regardless, at high mobile speeds, this handoff needs to happen quickly. In GSM, handoff is typically within 1-2 seconds. In AMPS, this was 10 seconds (higher potential for dropped calls!)

Define handoff margin as Δ

$$\Delta = P_{r,handoff} - P_{r,minimum\,useable}.$$

How much margin is needed to handle a mobile at driving speeds?

Example: Handoff Margin
Let the speed of a mobile be $v = 35$ meters/sec. For $n = 4$, a cell radius of 500 meters (the distance at which the power is at the threshold), and a 2 second handoff, what Δ is needed?
Solution: Assume the mobile is driving directly away from the BS, so distance d changes by 70 meters in two seconds. Consider the received power at the two times:

$$\begin{aligned} P_{r,minimum\,useable} &= \Pi_0 - 10n\log_{10} d \\ P_{r,handoff} &= \Pi_0 - 10n\log_{10}(d - 70) \end{aligned}$$

Taking the difference of the two equations (the 2nd minus the 1st),

$$\Delta = 10n\log_{10} d - 10n\log_{10}(d - 50) = 10n\log_{10}\frac{d}{d-70}$$

Plugging in that the call is dropped at $d = 500$ meters, we have $\Delta = 40\log_1 0\frac{500}{430} = 2.6$ dB.

Note that in this simple example, the propagation equation used is for "large scale path loss" only, which changes slowly. Typically, shadowing (caused by large geographical features and buildings blocking the signal) will play a more important role in quick changes in received power.

Mobile handoff (in GSM) is *mobile-assisted hand-off* (MAHO), in which the mobile measures the FCC from neighboring BSes, and reports them to the MSC.

Handoff assumes that there is a channel in the new BS to offer the entering mobile! But there may not be, and the call may be dropped for this reason. Users complain about dropped calls. So BSes may reserve "guard channels" purely for handoff purposes, which then are not offered to mobiles making new calls.

CDMA (Verizon 2G, and most 3G standards) phones do not require the same type of handoff as in GSM. In CDMA, a user does not need to switch "channel". Multiple base stations simultaneously receive a mobile's signal, and the MSC can combine the signals in some way (to obtain the best received signal). "Soft" handoff changes: which BSes are receiving the mobile's signal, and which BS (singular) is sending the replies.

Lecture 3

Today: (1) Interference (from Lecture 2); (2) Trunking

2.6 Review from Lecture 2

- RX power model: $P_r = c/d^n$ or $P_r(\text{dBm}) = P_0(\text{dBm}) - 10n\log_{10} d/d_0$.
- As a coarse approximation, cells are hexagons.
- A group of N cells is a cluster. Each cluster shares the total # of channels S by assigning each cell S/N of the channels.
- Co-channel base stations are approximately $R\sqrt{3N}$ apart.
- The signal to interference power ratio (SIR) $= S/\sum_i I_i$ where S is the received power from the desired signal and I_i is the received power from the ith interferer. The SIR in linear units is approximately SIR$= (3N)^{n/2}/i_0$ where $i_0 = 6$ (for now).

We did an example for SIR (dB) = 18. Rearranging,

$$N = \frac{1}{3}(i_0 \text{SIR})^{2/n}$$

Plugging in, for $n = 2, 3, 4$ we have N =126.2, 17.4, and 6.5, respectively.

2.7 Adjacent Channel Interference

Standard (non-ideal) radios do not perfectly filter out any out-of-band signals. Any signal that a mobile sends in another channel (besides its assigned channel) is interference at the BS w.r.t. the desired signal sent by another mobile in that channel. Each mobile's receiver also must filter out out-of-band signals from the BS, which does send signals on all channels. One standard way of making this problem less difficult is to assign non-adjacent channels within each cell's channel group.

We did an example last lecture in which we assigned $S = 70$ channels into groups for $N = 7$. There were $k = 70/7 \approx 10$ channels per group. For group 1, use channels $\{1, 8, \ldots, 57, 64\}$. For group i, use channels $\{i, i+7, \ldots, i+56, i+63\}$.

There is still the *near-far effect*. If a TX near the BS is producing just a little bit of out-of-band noise, it might swamp out the desired signal transmitted by a TX far away to the same BS.

One solution is *power control*, *i.e.*, reducing the TX power of mobiles close to the BS, since a high TX power is unnecessary. This reduces their out-of-band noise as well. Compared to a mobile transmitting full power all the time, power control extends battery life when close to a BS, and generally reduces even co-channel interference on the reverse channel. However, controlling a mobile means communication from the BS to the mobile to inform it whether to increase or decrease its power, which then requires data overhead. Tight power control is particularly required in all CDMA systems, which has a big "near-far problem".

3 Trunking

Trunking refers to sharing few channels among many users. Let U be the number of users, and C be the number of channels. Each user requires a channel infrequently, so a dedicated channel for each user is not required. But, the request for a channel happens at random times, and so for any $C < U$, it is possible that there will be more requests than channels.

- *Erlang*: A "unit" of measure of usage or traffic intensity. One Erlang is the traffic intensity carried by one channel that is occupied all of the time. 0.1 Erlang is the same channel occupied only 10% of the time.
- *Average holding time*: Average call duration, denoted H.
- *Call rate*: Average number of calls per unit time, denoted λ. Typically taken to be at the busiest time of day.
- *Total offered traffic intensity*: The total amount of traffic users request of the system, denoted A.
- *Grade of Service* (GOS): The probability an offered call will be blocked (and thus not served, or carried by the system).

Rappaport presents that an average user will request (offer) this much traffic, $A_u = \lambda H$. For example, if a user makes on average, two calls per hour, and that call lasts an average of 3 minutes, $A_u = \frac{2}{60 \text{ min}} 3 \text{ min} = 0.1$ Erlang. (Check your units!)

Then, to compute the total offered traffic intensity, and the total offered traffic intensity per channel (denoted A_c),

$$A = UA_u, \quad A_c = A/C$$

For the above example, assume that there are 1000 users and 200 channels. Then $A = 1000(0.1) = 100$, and $A_c = 100/200 = 0.5$.

Note that A_c is a measure of the *efficiency* of the utilization of the channels.

How should we design our system? Obviously, A_c should be less than one ($A < C$); or we'll never satisfy our call demand. But how should we set U, A_u, C to satisfy our customers?

First choice: what do we do when a call is offered (requested) but all channels are full?

- *Blocked calls cleared*: Ignore it.
- *Blocked calls delayed*: Postpone it!

3.1 Blocked calls cleared

1. Call requests are a Poisson process. That is, the times between calls are exponentially distributed, and memoryless.
2. Call durations are also exponentially distributed.
3. Rather than a finite number U of users each requesting A_u traffic, we set the total offered traffic as a constant A, and then let $U \to \infty$ and $A_u \to 0$ in a way that preserves $UA_u = A$. This is the "infinite number of users" assumption that simplifies things considerably.

These assumptions, along with the blocked calls cleared setup of the system, leads to the **Erlang B formula**:

$$GOS = P\,[\text{blocking}] = \frac{A^C/C!}{\sum_{k=0}^{C} A^k/k!} \tag{2}$$

Since C is very high, it's typically easier to use Figure 3.6 on page 81. By setting the desired GOS, we can derive what number of channels we need; or the maximum number of users we can support (remember $A = UA_u$); or the maximum A_u we can support (and set the number of minutes on our calling plans accordingly).

3.2 Blocked calls delayed

Instead of clearing a call; put it in a queue (a first-in, first-out line). Have it wait its turn for a channel. ("Calls will be processed in the order received"). There are now two things to determine

1. The probability a call will be delayed (enter the queue), and
2. The probability that the delay will be longer than t seconds.

The first is no longer the same as in (2); it goes up, because blocked calls aren't cleared, they "stick around" and wait for the first open channel.

Here, we clarify the meaning of GOS for a blocked calls delayed system. Here it means the probability that a call will be forced into the queue AND it will wait longer than t seconds before being served (for some given t).

We need a couple additional assumptions:

1. The queue is infinitely long. In a computer system, this translates to infinite memory.
2. No one who is queued gives up / hangs up (rather than wait).

With these assumptions, we can derive the Erlang C formula, for the probability that a call will be delayed:

$$P\,[\text{delay} > 0] = \frac{A^C}{A^C + C!\,(1 - A/C)\sum_{k=0}^{C-1} A^k/k!} \tag{3}$$

It is typically easiest to find a result from Figure 3.7, on page 82. Once it enters the queue, the probability that the delay is greater than t (for $t > 0$) is given as

$$P\,[\text{delay} > t|\text{delay} > 0] = \exp\left(-\frac{C - A}{H}t\right) \tag{4}$$

The two combined are needed to find the marginal (overall) probability that a call will be delayed AND experience a delay greater than t, the event that we are quantifying in GOS.

$$\begin{aligned} GOS = P\left[\text{delay} > t\right] &= P\left[\text{delay} > t|\text{delay} > 0\right] P\left[\text{delay} > 0\right] \\ &= P\left[\text{delay} > 0\right] \exp\left(-\frac{C-A}{H}t\right) \end{aligned} \tag{5}$$

Example: $N = 7$ cell cluster
A 7 cell cluster (with $N = 7$) has 30 MHz allocated to it for forward channels and each channel is 200 kHz. Assume blocked-called-delayed and a probability of delay of 1%, and each user makes one 10 minute call every 3 hours. (a) What is the number of users that can be supported? (b) What is $P\left[\text{delay} > 10\right]$ seconds? (c) What if it was a blocked-calls-cleared system with QOS of 1%?
Solution: 30 MHz / 200 kHz = 150 channels, divided among 7 cells, so about 20 channels per cell (after 1 control channel per cell). (a) With 20 channels, and probability of delay of 1%, looking at figure 3.7, we see A is about 12. With $11 = A_u U$ and $A_u = 10/(3 \times 60) = 1/18$, we have that $U = 198 \approx 200$. But this is per cell. So there can be $7(200) = 1400$ users in the 7 cells. (b) in each cell, $C = 20$, $A = 11$, $H = 10$ min or $H = 600$ sec. So $P\left[\text{delay} > t\right] = (0.01)\exp[-(20-11)(10)/600] = 0.01\exp(-0.15) = 0.0086$. (c) From Figure 3.6, $A \approx 13$, so $U = 234$, for a total of 1634 total users.

3.3 Discussion

What are the problems or benefits we see from the assumptions we've made? Are call requests "memoryless"? Is the exponential interarrival time assumption accurate? When catastrophic events occur, or major news breaks, what happens? How should a communications system be designed to handle these cases?

Lecture 4

Today: (1) Sectoring (2) Cell Splitting

4 Increasing Capacity and Coverage

4.1 Sectoring

In sectoring, we divide each cell into three or six "sectors" which are then served by three or six separate directional antennas, each with beamwidth of about 120 or 60 degrees.

We showed in Lecture 2 that the S/I ratio is given by $\frac{(3N)^{n/2}}{i_0}$, where N is the reuse ratio, and i_0 is the number of first-tier co-channel base stations. When we used omnidirectional antennas at each BS, we saw that $i_0 = 6$ regardless of N. By using sector antennas at the BSes, we will show that i_0 reduces. By reducing the S/I ratio for a given N, we allow a system to be deployed for a lower N, and therefore a higher capacity system.

However, each cell's channel group must be divided into three sub-groups. These new groups have 1/3 or 1/6 the number of channels, and thus the trunking efficiency will be lower.

Example: Decrease in trunking efficiency for constant N
Let $N = 7$, each cell has $C = 100$ channels, and users who make calls with $\lambda = 0.01$ per minute with average holding time 3 minutes. For blocked-calls-cleared and a GOS of 2%, what is the number of users which can be supported in this cell? Next, using 120 degree sectoring, and otherwise identical system parameters, what is the number of users which can be supported in this cell? What percentage reduction in capacity does sectoring with constant N cause?
Solution: For $C = 100$ and GOS= 0.02, from Figure 3.6, I read $A \approx 99$. Thus with $A_u = 0.01(3) = 0.03$, we could support $U = 99/0.03 = 3300$. For the sectoring case, $C = 33.3$ in each sector, and from Figure 3.6, $A = 24$. So we could support $U = 24/0.03 \approx 800$ per sector, or 2400 total in the cell. The number of users has reduced by 28%.

Example: Reducing N with sector antennas
For the same system, now assume that with 120 degree sectoring, that N can be reduced from 7 to 4. What number of users can be supported?
Solution: Now, the number of channels in each cell goes up to $100(7/4) = 175$. So each sector has $C = 58$ channels. With GOS = 2%, from Figure 3.6, $A \approx 48$, so $U \approx 1600$, for a total of 4800 users per cell. This is a 45% increase upon the $N = 7$ non-sectored cell.

Why does i_0 reduce? Consider again a mobile at the edge of a cell. We need to determine which of the first tier BSes contribute significantly to the interference signal. Refer to Figures 3.10, 3.11, for $N = 7$, P3.28(b) for $N = 3$, and to Figure 7 for $N = 4$.

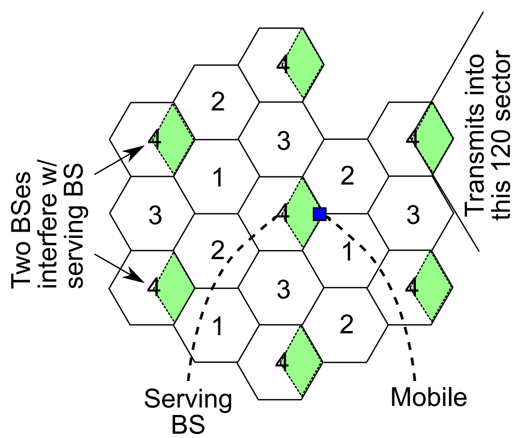


Figure 7: 120 degree sectoring for cellular system with $N = 4$. Only two first tier BSes significantly interfere with the middle BS.

Compared to when $i_0 = 6$, how much does S/I improve with sectoring?
Recall that $S/I = \frac{(3N)^{n/2}}{i_0}$. In dB terms,

$$\frac{S}{I}(dB) = 5n \log_{10}(3N) - 10 \log_{10} i_0$$

So with $i_0 = 6$, the latter term is 7.8 dB. If $i_0 = 1, 2$, and 3, the same term is 0, 3.0, or 4.8 dB. So, the improvement is 3, 4.8, or 7.8 dB. The particular value of i_0 that can be obtained is a function of N and whether 60 or 120 degree sectoring is used.

For a particular SIR and path loss exponent, how does i_0 affect the necessary N? From lecture

3,

$$N = \frac{1}{3}(i_0 \text{SIR})^{2/n}$$

So N is proportional to $i_0^{2/n}$.

4.1.1 Determining i_0

What is i_0 for 120 or 60 degree sector antennas? In short: it depends on N. You need to check on the hex plot to see how many sectors' base stations will "cover" the serving sector. My argument (not proven) is that when $i \neq j$, we have $i_0 = 2$ for 120^o antennas and $i_0 = 1$ for 60^o antennas. But for $i = j$, you need $i_0 = 3$ for 120^o antennas and $i_0 = 2$ for 60^o antennas. The case of $i = j$ happens at $N = 3$, and $N = 12$ (and $3i^2$ in general).

4.1.2 Example

Example: Assume we have $S = 533$ full-duplex channels. Assume blocked-calls cleared with a GOS of 2%, and per user offered traffic of 0.015 Erlang. Further assume we're using modulation with minimum required SIR(dB) of 19.5 dB and we've measured for our deployment area that $n = 3.3$. Find the total number of users possible per channel assuming (a) omni-directional antennas and (b) 120^o sector antennas.

Solution: Note linear SIR $= 10^{19.5/10} = 89.1$. (a) For omni antennas, $i_0 = 6$ so

$$N \geq \frac{1}{3}(6 \cdot 89.1)^{2/3.3} = 15.0$$

Since the given SIR is a minimum, we need $N \geq 15.0$. Since there is no 15-cell reuse, we need to increase to $N = 16$, which is possible with $i = 4$ and $j = 0$. Thus there are $533/16 = 33$ channels per cell available. With a GOS of 2%, from the Erlang B chart, $A \approx 25$. With $A_u = 0.015$, this means $U = A/A_u = 25/0.015 = 1667$ users per cell. (b) For 120^o antennas, we need to guess at N since i_0 is a function of N. For larger N, $i_0 = 2$ when using 120^o antennas. So let's plug in $i_0 = 2$ and see what N we get:

$$N \geq \frac{1}{3}(2 \cdot 89.1)^{2/3.3} = 7.7$$

So $N = 9$ would work. (Checking, sure enough, $i_0 = 2$ for $N = 9$.) Thus there are $533/9 = 59.22$ channels per cell or $533/(9 \cdot 3) = 19.7$ channels per sector available. With a GOS of 2%, from the Erlang B chart, $A \approx 14$ per sector. With $A_u = 0.015$, this means $U = A/A_u = 14/0.015 = 933$ users per sector, or 2800 per cell. This is a $(2800 - 1667)/1667 = 68\%$ improvement over the omni case.

4.2 Microcells

When we introduced "cells" we said the radius was a variable R. The idea of using microcells is that for a densely populated area, we cut the size of the cell by half. In this microcell-covered area, the concept of frequency reuse occurs described earlier, only with smaller R. The smaller R also has the benefit that transmit powers would be cut by a factor of 2^n (see Rappaport 3.7.1 for details). The other main benefit is that by reducing the area of a cell by a factor of four (forced by cutting R by two) the capacity in the microcell area is increased by four. For example, consider Figure 8, which shows an original macrocell grid, next to an "inserted" microcell area.

However, at the edges of the microcell area, there is a conflict. Cells that were separated by distance $R\sqrt{3N}$ for the initial R are no longer separated by that much. Conflicts in channel assignments at the edges are solved by splitting the channel group into two sub-groups. These subgroups can have different sizes, *e.g.*, the sub-group used for the microcell might have fewer channels assigned to it compared to the macrocell.

Another problem in GSM is that the number of handoffs is increased, since users travel through microcells more quickly. This can be addressed using umbrella cells (page 66) or microcell zones (Section 3.7.4).

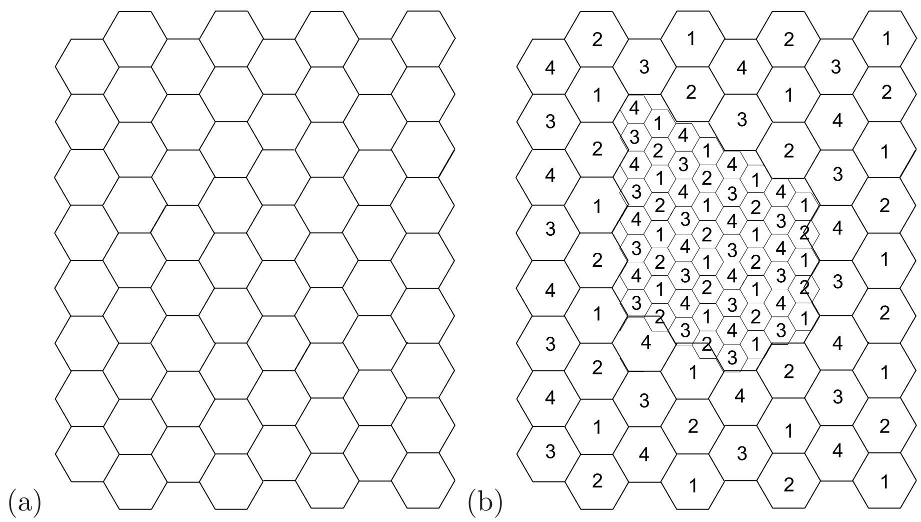


Figure 8: (a) 68 macrocells vs. (b) 53 macrocells plus 57 microcells.

4.3 Repeaters

This is Section 3.7.3 in Rappaport. Repeaters can be used to increase the coverage area, particularly into buildings, tunnels, and canyons. They are bidirectional (they amplify forward and reverse channels). However, repeaters don't add any capacity to the system, they just increase the reach of a BS or MS into "shadowed" areas.

4.4 Discussion

What are some of the problems with the assumptions made in this analysis?

Lecture 5

Today: (1) Free Space (2) Large Scale Path Loss (Intro)

Path loss models are either (1) empirical or (2) theoretical. We'll start to discuss both. As you'll see, empirical models were developed as modifications to theoretical models.

5 Free Space Propagation

Free space is nothing – nowhere in the world do we have nothing. So why discuss it?

Rappaport Section 4.3 describes the electric and magnetic fields produced by a small dipole antenna. This equation is only valid for a small dipole, and is only useful very close to (the near field of) the antenna. In the "far field" (distances many wavelengths from the antenna), the received power P_r in free space at a path length d is given in Section 4.2 as

$$P_r = P_t G_t G_r \left(\frac{\lambda}{4\pi d}\right)^2 \tag{6}$$

where G_t and G_r are the transmitter and receiver antenna gains, respectively; P_t is the transmit power; and λ is the wavelength. Notes:

- Wavelength $\lambda = c/f$, where $c = 3 \times 10^8$ meters/sec is the speed of light, and f is the frequency. We tend to use the center frequency for f, except for UWB signals, it won't really matter.
- All terms in (6) are in linear units, not dB.
- The effective isotropic radiated power (EIRP) is $P_t G_t$.
- The path loss is $L_p = \left(\frac{4\pi d}{\lambda}\right)^2$. This term is called the "free space path loss".
- The received power equation (6) is called the Friis transmission equation, named after Harald T. Friis [6].
- Free space is used for space communications systems, or radio astronomy. Not for cellular telephony.

In dB, the expression from (6) becomes

$$P_r(\text{dBm}) = P_t(\text{dBm}) + G_t(\text{dB}) + G_r(\text{dB}) - L_p(\text{dB}), \quad \text{where } L_p(\text{dB}) = 20\log_{10}\left(\frac{4\pi d}{\lambda}\right) \tag{7}$$

I like to leave $L_p(\text{dB})$ in terms of d/λ, which is a unitless ratio of how many wavelengths the signal has traveled. The terms $G_t(\text{dB})$ and $G_r(\text{dB})$ are clearly gains, when they are positive, the received power increases. And as distance increases, $L_p(\text{dB})$ increases, which because of the negative sign, reduces the received power. *We use "G" to denote gains and "L" to denote losses.* But a negative gain is a loss, and a negative loss is a gain.

5.1 Received Power Reference

Note either (6) or (7) can be converted to "refer to" a reference distance. For example, multiply the top and bottom of (6) by $(d_0/d_0)^2$ for some reference distance d_0:

$$\begin{aligned} P_r &= P_t G_t G_r \left(\frac{\lambda}{4\pi d_0}\right)^2 \left(\frac{d_0}{d}\right)^2 \\ &= P_r(d_0)\left(\frac{d_0}{d}\right)^2 \end{aligned} \tag{8}$$

where $P_r(d_0)$ is the received power at the reference distance d_0, according to (6). Now, we see that whatever the received power in free space is at distance d_0, the power at d decays as $(d_0/d)^2$ beyond that distance. In dB terms,

$$P_r(\text{dBm}) = \Pi_0(\text{dBm}) - 20\log_{10}\frac{d}{d_0} \tag{9}$$

where $\Pi_0(\text{dBm}) = 10\log_{10} P_r(d_0)$. This is actually an easier equation to deal with in practice, because we don't necessarily know the antenna gains and mismatches, and transmit power; but we can measure $\Pi_0(\text{dBm})$. Of course, not in free space – we don't exist there!

5.2 Antennas

Antenna gain is a function of angle. The only exception is the (mythical) isotropic radiator.

Def'n: *Isotropic Radiator*
An antenna that radiates equally in all directions. In other words, the antenna gain G is 1 (linear terms) or 0 dB in all directions.

(From Prof. Furse) An isotropic radiator must be infinitesimally small. Does not exist in practice, but is a good starting point.

Antenna gains can be referred to other ideal antenna types:

- dBi: Gain compared to isotropic radiator. Same as the dB gain we mentioned above because the isotropic radiator has a gain of 1 (or 0 dB).
- dBd: Gain compared to a half-wave dipole antenna. The 1/2 wave dipole has gain 1.64 (linear) or 2.15 dB, so dBi is 2.15 dB greater than dBd.

Technically, any antenna that is not isotropic is *directive*. Directivity is measured in the far field from an antenna as:

$$D = \frac{P_r(\text{maximum})}{P_r(\text{isotropic})}$$

where $P_r(\text{maximum})$ is the maximum received power (at the same distance but max across angle), and $P_r(\text{isotropic})$ is the power that would have been received at that point if the antenna was an isotropic radiator.

Antennas also have an efficiency. They lose some power without radiating it as EM waves. Thus the maximum gain is the directivity times the efficiency.

Commonly, we call an antenna *directional* if it is has a non-uniform horizontal pattern. A dipole has a "donut-shaped" pattern, which is a circle in its horizontal pattern (slice).

There are also antenna mismatches. We denote these as Γ_t and Γ_r. Both are ≤ 1, and only one if there is a perfect impedance match and no loss.

5.3 Power Flux Density

There is a concept in propagation of power flux density, the amount of power that travels through a given area. This is a far-field concept only. Power flux is denoted P_d in Rappaport, and has units of Watts per square meter, W/m^2. In free space,

$$P_d = \frac{|E|^2}{\eta}\text{W/m}^2$$

where η is the intrinsic impedance of free space, $120\pi\Omega = 377\Omega$, and $|E|^2$ is the magnitude squared of the electric field. The idea is that an antenna "captures" some of this power, according to, effectively, how large the antenna is. We call this the *effective antenna aperture*, and denote it A_e, with units m^2. In short, physically larger antennas are capable of larger A_e, although there is no exact proportionality. In this case the definition of the received power is

$$P_r(d) = P_d A_e$$

6 Large Scale Path Loss Models

Let's transition to the real world, where we exist. There are other effects besides radiation, including attenuation (transmission), reflection, diffraction, scattering, etc. We will discuss each of these in upcoming lectures. For now, suffice it to say that many signals arrive at the receiver, but with less power than would be indicated by the Friis equation. The received power varies strongly (5-25 dB) even for small changes in antenna position, center frequency, and time. But, there is a large effect caused when the distance (a.k.a. path length) increases by orders of magnitude. This large effect we call *large scale path loss*. It is analogous to L_p, but doesn't necessarily take the same form. We will re-write L_p as a function of distance in one of two ways:

1. Exponential decay: L_p will include a term proportional to $10^{\alpha d/10}$, where α is a loss factor, with units dB per meter. In this case, $L_p(dB) = \alpha d$, which makes it easier to see that α is dB loss per meter. Equation (6) is typically re-written as:

$$P_r = P_t G_t G_r \left(\frac{\lambda}{4\pi d}\right)^2 10^{-\alpha d/10} \tag{10}$$

   This works well in some conditions, for example, at 60 GHz, at which oxygen molecules absorb RF radiation, or due to rain at 30 GHz.

2. Power decay: L_p will be proportional to $1/d^n$, for some path loss exponent n. In free space, it was proportional to $1/d^2$, so this just lets n adjust to the particular environment. Typically, n ranges between 1.6 and 6, according to Rappaport. From my experience, I've seen n between 1.7 and 5.

6.1 Log Distance Path Loss

This is 4.9 and 4.11 in Rappaport. This is synonymous with what I call "power decay" above. Actually, it is the simplest of the models, and makes a big step towards better representation of actual large-scale path loss. In the log-distance path loss model, we can simply write the received power as a modification of (9) as

$$P_r(\text{dBm}) = \Pi_0(\text{dBm}) - 10n \log_{10} \frac{d}{d_0} \tag{11}$$

where $\Pi_0(\text{dBm})$ is still given by the Friis equation, but now the $L_p(\text{dB})$ term has changed to include a factor $10n$ instead of 20. Typically d_0 is taken to be on the edge of near-field and far-field, say 1 meter for indoor propagation, and 10-100m for outdoor propagation.

We mentioned that we can find the parameter $\Pi_0(\text{dBm})$ from measurements. We can also find the parameter n from measurements. For example, two measurement campaigns I did in office areas resulted in the estimates of $n = 2.30$ and 2.98 as shown in Figures 9 and 10.

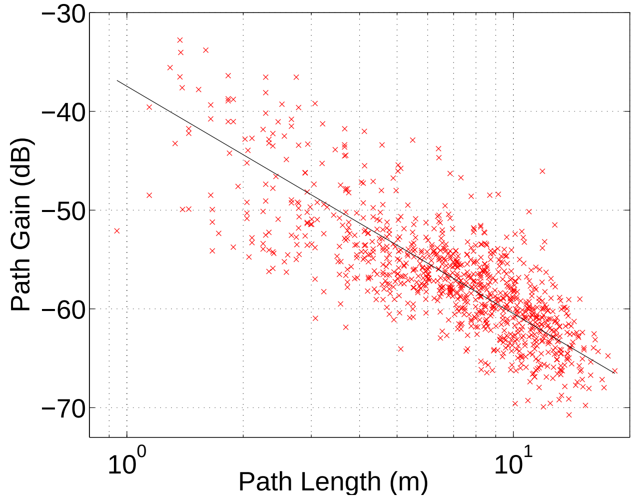


Figure 9: Wideband path gain measurements (x) at 2.4 GHz as a function of path length d. Linear fit (—) is with $d_0 = 1$m, $n = 2.30$, and $\sigma_{dB} = 3.92$. From [21].

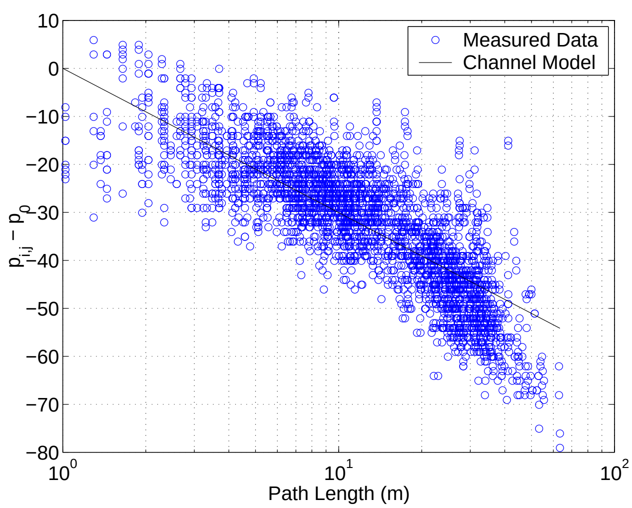


Figure 10: Narrowband measurements of path gain minus Π_0(dBm) (o) at 925 MHz as a function of path length d. Linear fit (—) is with $d_0 = 1$m, $n = 2.98$, with standard deviation $\sigma_{dB} = 7.38$. From [21].

6.2 Multiple Breakpoint Model

This is Rappaport Section 4.11.4. Empirically measurement studies have shown that the slope of the L_p vs. distance curve changes after a certain distance [9]. You can see this effect in Figure 10 for $d > 20$ meters; the path gains at $d = 50$ meters are all lower than the model, and one can see the slope changing to an n higher than 2.98. We will discuss theoretical reasons why this might happen in later lectures. Regardless, we can model the path loss as experiencing more than one slope in different segments of the $\log d$ axis.

$$P_r(\text{dBm}) = \begin{cases} \Pi_0(\text{dBm}) - 10n_1 \log_{10} \frac{d}{d_0}, & d \leq d_1 \\ \Pi_1(\text{dBm}) - 10n_2 \log_{10} \frac{d}{d_1}, & d > d_1 \end{cases} \tag{12}$$

where $\Pi_0(\text{dBm})$ is still the Friis received power at a distance d_0, and $\Pi_1(\text{dBm})$ is the received power (given by the first line of the equation) at distance d_1, and $d_0 < d_1$. Typically, the slope of the path loss increases, *i.e.*, $n_2 > n_1$.

Lecture 6

Today: (1) Reflection (2) Two-ray model

7 Reflection and Transmission

There are electric and magnetic waves that serve to propagate radio energy. The electric waves can be represented as a sum of two orthogonal *polarization* components, for example, vertical and horizontal, or left-hand and right-hand circular. What happens when these two components of the electric field hit the boundary between two different dielectric media?

We talk about the plane of incidence, that is, the plane containing the direction of travel of the waves (incident, reflected, and transmitted), and perpendicular to the surface (plane where the two media meet). See Figure 4.4 on page 115, reproduced in Fig 11.

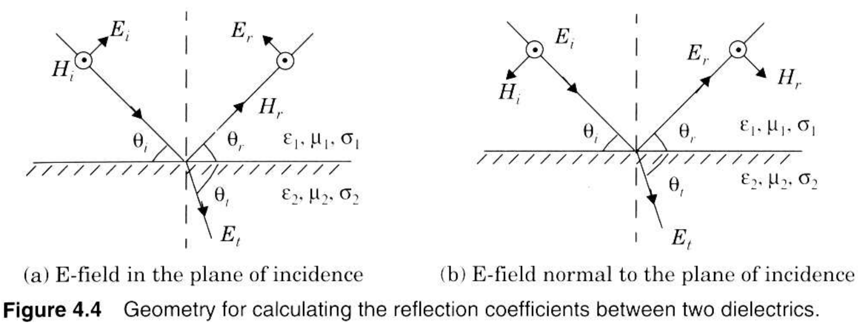


Figure 11: Figure 4.4 from Rappaport. The $\odot$ indicates an arrow pointing out of the paper.

Notes about the notation from Rappaport:

- "Parallel" refers to the E-field having direction parallel to the plane of incidence (as in Figure 4.4(a)); "perpendicular" means perpendicular (normal) to the plane of incidence (as in Figure 4.4(b)).

- Use subscripts i, r, and t to refer to the incident, reflected, and transmitted field.
- ϵ_1, ϵ_2, is the permittivity of medium 1 and 2. (units Farads/meter) (Note: F = sec / Ω)
- μ_1, μ_2 is the permeability of medium 1 and 2. (units Henries/meter) (Note: H = Ω sec)
- σ_1, σ_2 is the conductance of medium 1 and 2 (units Siemens/meter). (Note: S = 1 / Ω)
- The complex dielectric constant is $\epsilon = \epsilon_0\epsilon_r - j\epsilon'$ where $\epsilon_0 = 8.85 \times 10^{-12}$ F/m is free space permittivity, $j = \sqrt{-1}$, and $\epsilon' = \frac{\sigma}{2\pi f}$, and ϵ_r is the relative permittivity. Don't get confused by the subscripts in this equation.
- A material is a good conductor when $\sigma_k > f\epsilon_k$.
- The intrinsic impedance of medium k is $\mu_k\sqrt{\mu_k/\epsilon_k}$.

Then the reflection coefficients at the boundary between two non-conductive (dielectric) materials are given by

$$\begin{aligned}
\Gamma_{\|} \triangleq \frac{E_r}{E_i} &= \frac{\eta_2 \sin\theta_t - \eta_1 \sin\theta_i}{\eta_2 \sin\theta_t + \eta_1 \sin\theta_i} \\
\Gamma_{\perp} \triangleq \frac{E_r}{E_i} &= \frac{\eta_2 \sin\theta_i - \eta_1 \sin\theta_t}{\eta_2 \sin\theta_t + \eta_1 \sin\theta_i}
\end{aligned} \tag{13}$$

where θ_t is determined by *Snell's Law*:

$$\sqrt{\mu_1\epsilon_1}\sin(90^o - \theta_i) = \sqrt{\mu_2\epsilon_2}\sin(90^o - \theta_t) \tag{14}$$

Also, the angle of incidence is equal to the angle of reflection:

$$\theta_i = \theta_r$$

Finally, the reflected and transmitted field strengths are:

$$\begin{aligned}
E_r &= \Gamma E_i \\
E_t &= (1 + \Gamma)E_i
\end{aligned}$$

where you chose Γ based on the polarization of the incident E-field, *i.e.*, use either $\Gamma_{\|}$ or $\Gamma_{\perp}$.

There is a special case of (13) when the first medium is free space (or approximately, air) and $\mu_1 = \mu_2$. These two conditions are the case for most dielectric materials, in short, for the materials for we'd care to apply (13). In this case you can show (good test problem?) that

$$\begin{aligned}
\Gamma_{\|} &= \frac{-\epsilon_r \sin\theta_i + \sqrt{\epsilon_r - \cos^2\theta_i}}{\epsilon_r \sin\theta_i + \sqrt{\epsilon_r - \cos^2\theta_i}} \\
\Gamma_{\perp} &= \frac{\sin\theta_i - \sqrt{\epsilon_r - \cos^2\theta_i}}{\sin\theta_i + \sqrt{\epsilon_r - \cos^2\theta_i}}
\end{aligned} \tag{15}$$

See Figure 4.6 on page 118 of Rappaport. At some angle θ_i, there is no reflection of the parallel E-field from (15). This angle is called the "Brewster angle", which is given by

$$\sin\theta_B = \sqrt{\frac{\epsilon_1}{\epsilon_1 + \epsilon_2}}$$

When medium 1 is free space, and $\epsilon_2 = \epsilon_0 \epsilon_r$,

$$\sin\theta_B = \frac{1}{\sqrt{1+\epsilon_r}}$$

This is the same as Equation 4.28 in Rappaport.

Note that as $\theta_i \to 0$, $\Gamma_\parallel \to 1$ and $\Gamma_\perp \to -1$.

Also, for perfect conductors (as described in Section 4.5.3), we also have $\Gamma_\parallel = 1$ and $\Gamma_\perp = -1$.

Example: Reflection from ground
Find the reflection coefficients for typical ground at an incident angle of 15 degrees at 100 MHz.
Solution: Assume free space is medium 1 and that 'typical ground' has $\epsilon_r = 15$. Note $\sin 15^o = 0.259$, and $\cos^2 15^o = 0.933$, so from (15),

$$\begin{aligned}\Gamma_\parallel &= \frac{-15(0.259) + \sqrt{15-0.933}}{15(0.259)+\sqrt{15-0.933}} = -0.0176 \\ \Gamma_\perp &= \frac{0.259 - \sqrt{15-0.933}}{0.259+\sqrt{15-0.933}} = -0.871\end{aligned}$$

8 Two-Ray (Ground Reflection) Model

Section 4.6 in Rappaport develops a theoretical model for propagation slightly better than the free space assumption. This model includes not just one path, but also another path that reflects off of the ground, as shown in Figure 4.7 in Rappaport. The model isn't hard to develop, and provides an important theoretical underpinning to the multiple breakpoint model we covered in lecture 5.

Remember, *powers of multipath DON'T add together*. Only voltages or field strength of multipath actually add together. The voltage on the antenna is proportional to the electric field at the antenna position. So let's talk about adding electric fields.

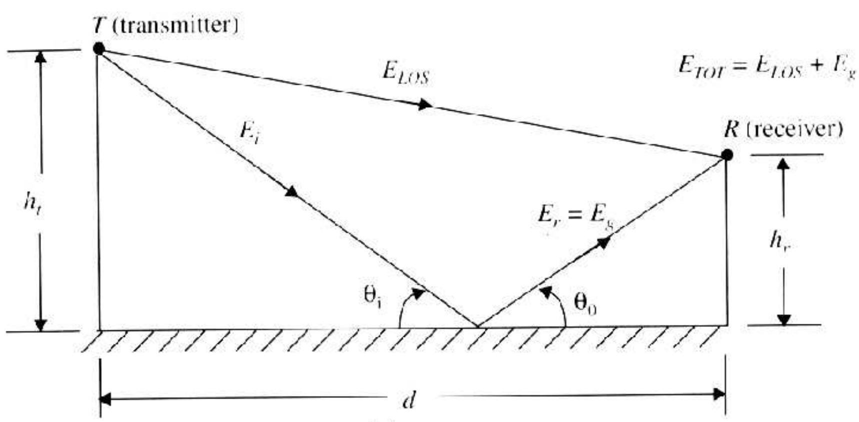


Figure 12: Figure 4.7 from Rappaport. TX and RX have are separated on the ground by d, but are at heights h_t and h_r respectively.

8.1 Direct Path

Recall that the electric field magnitude decays as $1/d$ in free space. So, similar to how we wrote the received power with a reference distance d_0, we write the E-field strength as the E-field strength at a reference distance, multiplied by d_0/d', for a path (distance of travel for waves) length d'. Also, assume the signal is a simple sinusoid at the carrier frequency, f_c. So

$$E(d,t) = E_0 \frac{d_0}{d'} \cos\left(2\pi f_c \left(t - \frac{d'}{c}\right)\right) \tag{16}$$

For the LOS path, given a distance along the ground of d, antenna heights h_t and h_r at the TX and RX, respectively, the $d' = \sqrt{d^2 + (h_t - h_r)^2}$. So

$$E_{LOS} = E_0 \frac{d_0}{\sqrt{d^2 + (h_t - h_r)^2}} \cos\left(2\pi f_c \left(t - \frac{\sqrt{d^2 + (h_t - h_r)^2}}{c}\right)\right) \tag{17}$$

8.2 Reflected Path

Two things change for the reflected path compared to the LOS path:

1. The path is longer than the LOS path for total length $\sqrt{d^2 + (h_t + h_r)^2}$ (use the "method of images" to show this).
2. The reflected field strength changes by a factor of Γ.

In general, we can write the E-field as

$$E_g = \Gamma E_0 \frac{d_0}{\sqrt{d^2 + (h_t + h_r)^2}} \cos\left(2\pi f_c \left(t - \frac{\sqrt{d^2 + (h_t + h_r)^2}}{c}\right)\right) \tag{18}$$

Let's assume that d is very long compared to the antenna heights. So, the angle of incidence is approximately 0. In this case the reflection coefficient (assume perpendicular polarization) is -1. Then (18) becomes,

$$E_g = -E_0 \frac{d_0}{\sqrt{d^2 + (h_t + h_r)^2}} \cos\left(2\pi f_c \left(t - \frac{\sqrt{d^2 + (h_t + h_r)^2}}{c}\right)\right) \tag{19}$$

8.3 Total Two-Ray E-Field

We are interested in the magnitude of the total E-field, $E_{TOT} = E_{LOS} + E_g$, that is, the quantity that multiplies the $\cos(2\pi f_c t)$ term. Using trig identities, and this approximation:

$$\Delta = \sqrt{d^2 + (h_t + h_r)^2} - \sqrt{d^2 + (h_t - h_r)^2} \approx \frac{2h_t h_r}{d}$$

we can show that

$$|E_{TOT}| \approx 2E_0 \frac{d_0}{d} \sin\left(\frac{2\pi f_c}{c} \frac{2h_t h_r}{d}\right)$$

But at large L, the argument of the sin is approximately 0, and the $\sin x \approx x$. This is when $x < 0.3$ radians, which in our equation, means that for

$$d > \frac{20 h_t h_r f_c}{c}$$

we can use this approximation :

$$|E_{TOT}| \approx 2E_0 \frac{d_0}{d} \left(\frac{2\pi f_c}{c} \frac{2h_t h_r}{d} \right) = \frac{\text{const}}{d^2}$$

This means that the power decays as $1/d^4$! See Figure 13.

In summary, when there are two paths, one direct and one ground reflection, the theoretical models show behavior that has two different path loss exponents, $1/d^2$ for d less than a threshold, and $1/d^4$ for d above the threshold. This matches what we've observed from measurements and presented as the empirical "multiple breakpoint" model.

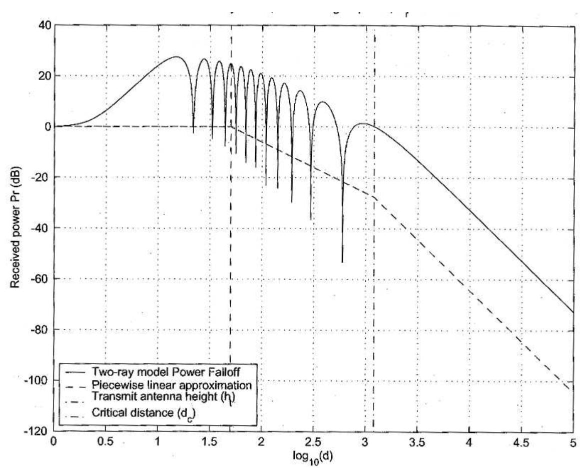


Figure 13: Received power as a function of log distance in two-ray model, Figure 2.5 from the Goldsmith book [11].

However, a note: this is just a theoretical model. Typical cellular or indoor channels do not have just two paths. One of the 6325 assignments is to build a "ray-tracing" simulation for a rectangular room. As the number of paths with significant amplitude increases, you tend not to see the $1/d^4$ behavior. This model tends to be accurate in outdoor areas with few obstructions, where the ground reflection is strong, and few other multipath components have significant power.

9 Indoor and Site-specific Large Scale Path Loss Models

9.1 Attenuation Factor Models

In previous lectures we explored models for large scale path loss as

$$P_r(\text{dBm}) = \Pi_0(\text{dBm}) - 10n \log_{10} \frac{d}{d_0}. \tag{20}$$

From the reflection equations, we can see that transmitted field has a multiplier Γ_t, and thus the transmitted signal power is multiplied by $|\Gamma_t|^2$ for each change in material. In dB terms, these are additive terms $20 \log_{10} |\Gamma_t|$. Thus in my large scale path loss equation, if I know what materials (obstructions) are in between the transmitter and receiver, I might add in these gains (equivalently subtract the losses) to my received power as follows:

$$P_r(\text{dBm}) = \Pi_0(\text{dBm}) - 10n \log_{10} \frac{d}{d_0} - \sum_i \text{PAF}_i \tag{21}$$

where PAF_i is the loss (partition attenuation factor) caused by transmission through obstruction i. For example, if the signal must go through one external wall and two internal walls, we'd find PAF for these three obstructions, and add them in. In fact, in this case, if the two internal walls are identical types of walls, I would find two different PAFs, and add the external wall PAF and twice the internal wall PAF in (21). Table 4.3 in Rappaport has a summary of reported PAF values for standard obstructions at some frequencies. If you know the signal goes from one floor to another, you can consider the floor as just another obstruction with its own PAF. Incidentally, you can estimate the PAFs if you measure many links and keep track of the path length, and how many of each type of obstruction that each link crosses through, using some linear algebra [8].

An example of the importance of this work is in determining the received power in buildings or in homes when the transmitter is located outside of the building. For example, in the late 1990s, wireless companies (and the FCC) proposed "local multipoint distribution services" (LMDS) as an alternative to cable TV, at either 6, 20, or 30 GHz. The service would put digital TV signal transmitters on a telephone pole in a neighborhood, and people in homes would just connect a special antenna to their TV and be able to download a wide bandwidth of video data. At 6 or 30 GHz, PAFs are higher, and thus the link budgets were strongly dependent on how many walls and trees the signal would need to pass through between tower and TV [8]. Incidentally, LMDS lost out in the competition with cable and DSL and satellite.

9.2 Ray-tracing models

An even more site-specific propagation model uses a GIS database of terrain and building locations and shapes in order to "ray-trace" contributing paths. Like we traced the ground-reflected path above, there will be other reflected paths (and diffracted and scattered paths) that can be computed from the geometry of the obstructions in the environment. Of course, these databases are incomplete and so they must be "corrected" using measurements of P_r at known positions.

One example of the importance of accurate RX power models is in the localization of cellular phones. The FCC has a requirement to reliably locate cell phones when they dial 911. There are specifics behind the word "reliably" and GPS won't cut it, because GPS doesn't accurately locate a phone that is indoors. One complementary technology to GPS (sold by Polaris Wireless) is to use the signal power recorded from neighboring base stations as a vector "signature" that is compared with those in a database. They do drive tests to come up with the signal strength vectors to put in the database. But they can't go in to every building, so they use a site-specific model to estimate what the signal strengths are inside of the building, based on the signals measured outside of the building. These estimated vectors are also put into the database. Finally, when a 911 call comes in, the phone records the RSS from neighboring BSes and the vector of them is compared to those in the database, and the location is guessed from the closest matches.

Lecture 7

Today: (1) Link Budgeting

Exam 1 (Tue, Sept 20, 3:40-4:40pm) covers: lectures 1-6; homeworks 1-3; Rappaport Chapters 3 and 4 (only the sections listed on the schedule). See `http://span.ece.utah.edu/5325-exam1prep`, for more practice. Note that I cannot test you on everything in one hour, the problems will be a random sample. So be knowledgeable about all possible problems that can be solved. DO NOT think that if you can handle the spring 2010 exam 1, that those are the only types of problems that will be covered, and thus you are ready.

10 Link Budgeting

Link budgets are, as the name implies, an accounting of the gains and losses that occur in a radio channel between a transmitter and receiver. We've talked about S/I – you need an acceptable signal to interference ratio. In addition, you need an acceptable signal to noise, or S/N, ratio. (a.k.a. SNR, C/N, or P_r/P_N ratio, where C stands for carrier power, the same thing we've been calling P_r, and N or P_N stands for noise power. Since we've already used N in our notation for the cellular reuse factor, we denote noise power as P_N instead.) Noise power is due to thermal noise.

In the second part of this course, we will provide more details on where the requirements for S/N ratio come from. For now, we assume a requirement is given. For a given required S/N ratio, some valid questions are: What is the required base station (or mobile) transmit power? What is the maximum cell radius (*i.e.*, path length)? What is the effect of changing the frequency of operation? Also, there is a concept of path balance, that is, having connectivity in only one direction doesn't help in a cellular system. So using too much power in either BS or mobile to make the maximum path length longer in one direction is wasteful.

As we've said, this is accounting. We need to keep track of each loss and each gain that is experienced. Also, to find the noise power P_N, we need to know the characteristics of the receiver.

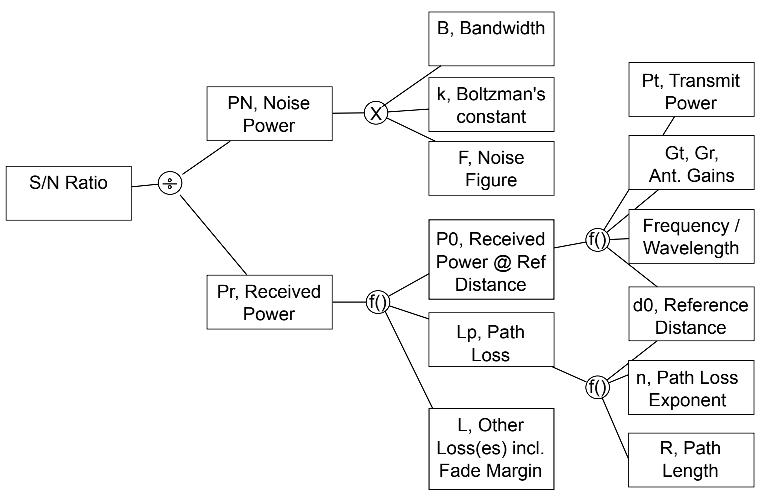


Figure 14: Relationship among link budget variables.

10.1 Link Budget Procedure

A universal link budget for received power is:

$$P_r(\text{dBW}) = P_t(\text{dBW}) + \sum \text{dB Gains} - \sum \text{dB Losses} \tag{22}$$

A universal link budget for S/N is:

$$S/N = P_r(\text{dBW}) - P_N(\text{dBW}) = P_t(\text{dBW}) + \sum \text{dB Gains} - \sum \text{dB Losses} - P_N(\text{dBW})$$

1. There's no particular reason I chose dBW instead of dBm for P_r and P_N. But they must be the same, otherwise you'll have a 30 dB error!
2. *If using EIRP transmit power*, it includes $P_t(\text{dBW}) + G_t(\text{dB})$, so don't double count G_t by also including it in the dB Gains sum.
3. The dB noise figure F (dB) is either included in $P_N(\text{dBW})$ or in the dB losses, *not both*!
4. **Gains** are typically only the antenna gains, compared to isotropic antennas.
5. There are also *coding, a.k.a. processing, gains*, achieved by using channel coding to reduce the errors caused by the channels. DS-SS (*e.g.*, CDMA) is a type of modulation which has a processing gain. These might be subtracted from the required S/N ratio, or added to the gains. Do one, but not both.
6. **Losses** include large scale path loss, or reflection losses (and diffraction, scattering, or shadowing losses, if you know these specifically), losses due to imperfect matching in the transmitter or receiver antenna, any known small scale fading loss or "margin" (what an engineer decides needs to be included in case the fading is especially bad), etc.
7. Sometimes the *receiver sensitivity* is given (for example on a RFIC spec sheet). This is the $P_N(\text{dB})$ plus the required $S/N(\text{dB})$.

10.2 Thermal noise

The thermal noise power in the receiver is P_N, and is given as $P_N = FkT_0B$, where

- k is Boltzmann's constant, $k = 1.38 \times 10^{-23} J/K$. The units are J/K (Joules/Kelvin) or W·s/K (1 Joule = 1 Watt × second).
- T_0 is the ambient temperature, typically taken to be 290-300 K. If not given, use 294 K, which is 70 degrees Fahrenheit.
- B is the bandwidth, in Hz (equivalently, 1/s).
- F is the (unitless) noise figure, which quantifies the gain to the noise produced in the receiver. The noise figure $F \geq 1$.

In dB terms,

$$P_N(\text{dBW}) = F(\text{dB}) + k(\text{dBWs/K}) + T_0(\text{dBK}) + B(\text{dBHz})$$

where $k(\text{dBWs/K}) = 10\log_{10} 1.38 \times 10^{-23}\text{J/K} = -228.6$ dBWs/K. We can also find F from what is called the equivalent temperature T_e. This is sometimes given instead of the noise figure directly.

$$F = 1 + \frac{T_e}{T_0}$$

10.3 Examples

Example: GSM uplink
Consider the uplink of a GSM system, given GSM requires an S/N of 11 dB [16]. Assume a maximum mobile transmit power of 1.0 W (30 dBm), 0 dBd antenna gain at the mobile, and 12 dBd gain at the BS. Assume path loss given by the urban area Hata model, with $f_c = 850$ MHz, BS antenna height of 30 meters, mobile height of 1 meter. Assume $F = 3$ dB and that the system is noise-limited. What is the maximum range of the link?
Solution:

- S/N required is 11 dB.
- $P_N = FkT_0B = 2(1.38 \times 10^{-23} J/K)(294K)(200 \times 10^3 Hz) = 1.62 \times 10^{-15} = -147.9$(dBW).
- $P_t = 0$ dBW.
- Gains: include 0 dBd and 12 dBd (or 2.15 dBi and 14.15 dBi) for a total of 16.3 dB gains.
- Losses: Path loss is via urban area Hata, for d in km,

$$\begin{aligned} L(\text{dB}) &= 69.55 + 26.16 \log_{10}(850) - 13.82 \log_{10}(30) + [44.9 - 6.55 \log_{10}(30)] \log_{10} d \\ &= 125.8 + 35.22 \log_{10} d \end{aligned}$$

So

$$\begin{aligned} 11(\text{dB}) &= 0(\text{dBW}) + 16.3(\text{dB}) - (125.8 + 35.22 \log_{10} d) + 147.9(\text{dBW}) \\ d &= 10^{27.4/35.22} = 6.0(\text{km}) \end{aligned} \quad (23)$$

Note 1 km = 0.62 miles, so this is 3.7 miles.

Example: Sensor Network

Assume two wireless sensors 1 foot above ground need to communicate over a range of 30 meters. They operate the 802.15.4 standard (DS-SS at 2.4 GHz). Assume the log-distance model with reference distance 1m, with path loss at 1 m is $\Pi_0 = 40$ dB, and path loss exponent 3 beyond 1m. Assume the antenna gains are both 3.0 dBi. The transmitter is the TI CC2520, which has max $P_t = 1$ mW, and its spec sheet gives a receiver sensitivity of -98 dBm. What is the fading margin at a 30 meter range? (Note: By the end of lecture 10 you will be able to specify fading margin given a desired probability that communication will be lost due to a severe fade. Here we're just finding what the margin is for these system parameters.)
Solution: The question asks us to find the difference between P_r at 30 meters and the receiver sensitivity (P_N(dB) plus the required S/N(dB). Rearranging (22),

$$-98dBm = S/N + P_N(\text{dBm}) = P_t(\text{dBm}) + \sum \text{dB Gains} - \sum \text{dB Losses} \quad (24)$$

1. $P_t(\text{dBm}) = 0$ dBm.
2. Gains: Two antennas at 3 dBi (the units are effectively dB), so the total gains are 6 dB.

3. Losses: There is the 40 dB loss to 1 m, then an additional $10(3.0)\log_{10}(30/1) = 44.3$ dB. Fading Margin is a Loss, so we have 84.3 dB + Fade Margin for the total losses.

So

$$-98dBm = 0(\text{dBm}) + 6(\text{dB}) - \text{Fade Margin} - 84.3(\text{dB})$$

Which, solving for Fade Margin, is 19.7 dB.

Example: IS-136

Compare IS-136 and GSM in terms of range. Compared the the GSM uplink example above, an IS-136 mobile has 0.6 W transmit power, and the required S/N is 15 dB [16], and IS-136 has a lower bandwidth of 30 kHz.

Solution:

- S/N required is 15 dB.
- $P_N = FkT_0B = 2(1.38 \times 10^{-23} J/K)(294K)(30 \times 10^3 Hz) = 2.43 \times 10^{-16} W = -156.1(\text{dBW})$.
- $P_t = -2.2$ dBW.

So

$$\begin{aligned} 15(\text{dB}) &= -2.2(\text{dBW}) + 16.3(\text{dB}) - (125.8 + 35.22\log_{10} d) + 156.1(\text{dBW}) \\ d &= 10^{29.4/35.22} = 6.8(\text{km}) \end{aligned} \quad (25)$$

Lecture 8

Today: (1) Diffraction (2) Scattering

11 Diffraction

In EM wave propagation *Huygens' principle* says that at each point, the wave field is effectively re-radiating in all directions. In free space, these *secondary reradiators* sum and "produce" the effect of a wave front advancing in the direction away from the source. When objects exist in free space that block or attenuate some of the wave field, the reradiation enable EM waves to "bend" around objects. In order to calculate the field at a point in (or near) the "shadow" of an object, we can use Huygens' principle to find accurate numerical results. This is a short version of some advanced electromagnetics. See [24] Chapter 9 for a more detailed treatment.

See Figure 4.13 in Rappaport. The Fresnel-Kirchoff parameter ν is given by,

$$\nu = h\sqrt{\frac{2(d_1 + d_2)}{\lambda d_1 d_2}} \quad (26)$$

depends on the geometry and the frequency, and is unitless:

- d_1, d_2, distance along line-of-sight path from TX or RX to obstruction

- h, screening height

In short, we have a normalized vertical axis at the knife edge. The top of the knife edge is at position ν – below ν, there is a perfect conductor, and above ν, there is free space. We assume the knife edge is infinitely narrow. For our point of interest beyond the knife edge, Huygens' principle has us consider (sum) the effect of the secondary reradiators along the vertical axis, above the knife edge. The summation is actually an integral, and is taken from ν to ∞, and is called the complex Fresnel integral,

$$F(\nu) = \frac{1+j}{2}\int_{\nu}^{\infty} \exp\left[-\frac{j\pi t^2}{2}\right] dt \tag{27}$$

One can calculate this integral with some knowledge of complex analysis, or even Matlab, but there is no easy analytical function that comes out of the solution. We typically use a table or a plot. The dB magnitude of the power loss from the Fresnel integral, $20\log_{10} F(\nu)$, which we call the knife-edge diffraction GAIN, is given by Figure 4.14 in Rappaport, and is copied in Figure 15.

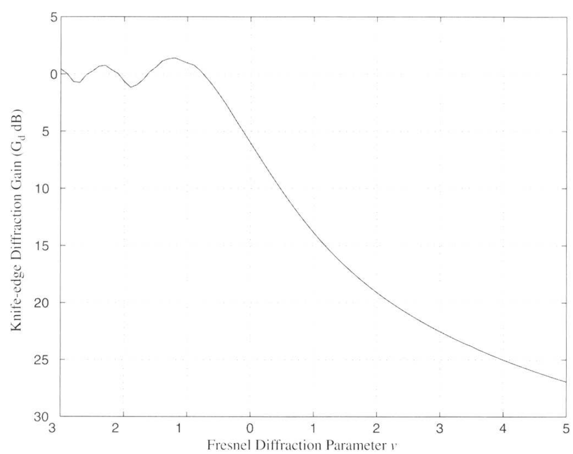


Figure 4.14 Knife-edge diffraction gain as a function of Fresnel diffraction parameter ν.

Figure 15: Knife-edge diffraction gain in dB.

Expressions exist for the multiple knife edge diffraction problem – when multiple obstructions block a propagating wave. However, these are largely computed via numerical analysis, so we won't elaborate on them.

12 Rough Surface Scattering

When we discussed reflection, we said a wave was impinging on a flat surface. But most surfaces (*e.g.*, walls, ground) are *rough*, not flat. When is a surface considered rough? When the maximum height "protuberance" from the surface, h, is greater than h_c,

$$h_c = \frac{\lambda}{8\sin\theta_i} \tag{28}$$

where θ_i is, again, the angle of incidence.

Scattering has two effects important to us:

1. Rough surface scattering reduces the power in the reflected wave.

2. Scattering causes additional multipath to be received in directions other than the specular direction (recall $\theta_r = \theta_i$).

For 1., if the surface is rough, then the reflected wave has reflection coefficient multiplied by ρ_s, so that $\Gamma_{rough} = \rho_S \Gamma$. Multiple expressions exist to compute ρ_S. Two given in the book are:

$$\begin{aligned}
\rho_S &= \exp\left[-8\left(\frac{\pi\sigma_h \sin\theta_i}{\lambda}\right)^2\right] \\
\rho_S &= \exp\left[-8\left(\frac{\pi\sigma_h \sin\theta_i}{\lambda}\right)^2\right] I_0\left[8\left(\frac{\pi\sigma_h \sin\theta_i}{\lambda}\right)^2\right]
\end{aligned}$$

where σ_h is the standard deviation of the height of the rough surface. The second expression is considered to be more accurate.

For 2., scattering is a very useful and impactful phenomena. Scattering is the basis of radar, weather systems, and passive RFID tags. Engineers have put much effort into understanding the phenomena of EM wave scattering.

Similar to Huygens' principal, the wave field at the scatterer is assumed to become a secondary re-radiator. However, the object (the scatterer) is assumed to absorb the incident power, and then re-radiate it (The re-radiation is not assumed to be occurring from the free space near the object). The object is considered a new point source, where power is "received" and "retransmitted". *Essentially we have two links, one from the transmitter to the scatterer, and one away from the scatterer to the end receiver.*

Typical airplane and weather radar is monostatic, *i.e.*, the TX and RX are co-located. In some *bistatic* wireless comm systems (and more robust airplane radar systems) the TX and RX are not in the same place. The *bistatic radar equation* describes the received power P_r in a scattered wave at a receiver that is not necessarily at the transmitter location. In linear and dB terms,

$$\begin{aligned}
P_r &= \frac{P_t G_t G_r \sigma_{RCS} \lambda^2}{(4\pi)^3 d_t^2 d_r^2} \\
P_r(\text{dBW}) &= P_t(\text{dBW}) + G_t(\text{dB}) + G_r(\text{dB}) + \sigma_{RCS}(\text{dBmeter}^2) \\
&\quad +20\log_{10}\lambda - 30\log_{10}4\pi - 20\log_{10} d_t - 20\log_{10} d_r
\end{aligned}$$

Note that there are two $1/d^2$ terms, one corresponding to each "link" described above. The σ_{RCS} term has units of meter^2 and is the *radar cross section* of the scatterer. It is an area, like an antenna effective aperture, that describes how much power is absorbed by the scatterer to be reradiated. Also, note that in the dB expression, the dB meter^2 units cancel because there are two dB meter^2 terms on top (σ_{RCS} and $20\log_{10}\lambda^2$) and two dB meter^2 terms on bottom ($20\log_{10} d_t$ and $20\log_{10} d_r$).

Lecture 9

Today: (1) Multipath Fading

13 Multipath Fading

We've talked about physics, that is, how wave propagation and its interaction with the environment causes reflection, transmission, diffraction, and scattering. Many individual propagating waves arrive at the receiver, these waves are called *multipath components*, or collectively, *multipath*. These multipath cause *fading* effects (changes in the received power) grouped together and called *multipath fading*. There are many kinds of multipath fading.

The challenges caused by multipath fading in wireless communication systems are one the most significant challenges addressed by wireless engineers today. Engineers have developed a variety of modulation and diversity schemes in order to counteract the negative influence of multipath fading. And, we are developing methods which take advantage of multipath in particular ways as a benefit for communication systems. All this to say, understanding of the fundamentals of fading is a critical skill for wireless engineers.

We're first going to talk about received power when mutiple multipath signals are reiceved at the receiver. Then, we'll present the spatial and temporal characteristics of multipath.

13.1 Multipath

We've been talking about the EM field. Specifically, we've presented expressions for the E-field E_b when we send a sine at frequency f_c through the channel, as

$$E_b = E_0 \cos(2\pi f_c t + \theta)$$

where E_0 is positive real number. In other words,

$$E_b = \mathbb{R}\left[E_0 e^{j2\pi f_c t + j\theta}\right] = \mathbb{R}\left[e^{j2\pi f_c t} E_0 e^{j\theta}\right]$$

The above expressions are called bandpass representation. When we want to write a simpler expression, we write the complex *baseband-equivalent representation*:

$$E = E_0 e^{j\theta}$$

and we can "translate" any complex baseband-equivalent signal into its bandpass signal by applying

$$E_b = \mathbb{R}\left[e^{j2\pi f_c t} E\right]$$

The simpler expression E has more easily viewable magnitude (amplitude) E_0 and phase (angle) θ.

The voltage received by the antenna is proportional to E and has the same complex baseband or real-valued baseband representation. For example, $V = \alpha E V_0 e^{j\theta}$, for some constant α.

As we discussed, many such multipath wave components arrive at the receiver. *They add together as voltages.* DO NOT add the powers of the multipath together – there is no such physical antenna that add together the powers of multipath. (If you find a way, patent it quick!)

Let's say there are M multipath components, numbered 0 through $M - 1$. (Rappaport uses N as the number of components, don't confuse with the frequency reuse factor.) Component i has amplitude V_i and phase θ_i. Then the total voltage at the receiver antenna will be:

$$V_{TOT} = \sum_{i=0}^{M-1} V_i e^{j\theta_i}$$

The received power is then proportional to $|V_{TOT}|^2$ (by a resistance).

We did this for the two-ray model, remember? The magnitude $|E_{TOT}|$ was a sum of two cosine terms, each with a different phase, which we needed to add together, before taking the magnitude squared. It would be a good exercise to re-derive Section 4.6 using phase angles rather than cosines.

13.2 Temporal

Let's expand on where these phase angles come from. Recall that $V_i e^{j\theta_i}$ is the representation of $V_i \cos(2\pi f_c t + \theta_i)$. If $V_i \cos(2\pi f_c t)$ is transmitted from the transmitter antenna, how do the phases of the multipath components behave with respect to each other? Well, each component has its own path length. It really did travel that length. And EM waves all travel at the same speed – $c = 3 \times 10^8$ m/s. So some waves arrive later than others. Let τ_i denote the time delay of arrival for multipath i relative to the transmit time. It is d_i/c, where d_i is the length of component i. What happens when a function is delayed by τ_i? We replace t with $t - \tau_i$ in the function. So $V_i \cos(2\pi f_c(t - \tau_i))$ is received. Well, not the full story – reflections and diffractions also cause phase changes (we discussed specifics for reflection in Section 4.5). Really, $V_i \cos(2\pi f_c(t - \tau_i) + \phi_i)$ is received, where ϕ_i is the sum of all phase changes caused by the physical propagation phenomena. We've been using baseband notation, what is the complex baseband notation? It is $V_i e^{j(-2\pi f_c \tau_i + \phi_i)}$.

So what is the total received voltage from all multipath?

$$V_{TOT} = \sum_{i=0}^{M-1} V_i e^{j(-2\pi f_c \tau_i + \phi_i)} \tag{29}$$

In other words, $\theta_i = -2\pi f_c \tau_i + \phi_i$. We've now written it in terms of its temporal delay, τ_i. Note that V_{TOT} has incidentally become a function of frequency f_c.

13.3 Channel Impulse Response

What we have in (29) is a frequency response as a function of frequency f_c. The equation can show the frequency response at *any* frequency. Let $f = f_c$, maybe that makes it clearer. So (29) is a frequency-domain representation of the total voltage. Let's convert to the time domain. How? Using the inverse Fourier transform:

$$\mathfrak{F}\left\{\sum_{i=0}^{M-1} V_i e^{j(-2\pi f \tau_i + \phi_i)}\right\} = \sum_{i=0}^{M-1} V_i \mathfrak{F}\left\{e^{j(-2\pi f \tau_i + \phi_i)}\right\} = \sum_{i=0}^{M-1} V_i e^{j\phi_i} \delta(\tau - \tau_i)$$

What this says is that in the *time delay domain*, the arrivals of multipath i occurs at delay τ_i.

This leads to how we frame the channel: as an echo-causing filter, with an impulse response that is a sum of time-delayed impulses. Let $s(t)$ be the transmitted signal and $r(t)$ be the received signal. Then

$$r(t) = \frac{1}{2} s(t) \star h(\tau)$$

where $h(\tau)$ is called the *channel impulse response*, and is given by

$$h(\tau) = \sum_{i=0}^{M-1} a_i e^{j\phi_i} \delta(\tau - \tau_i) \tag{30}$$

The a_i are proportional to V_i but are unitless – the units are contained in $s(t)$, which has units of Volts. The amplitude $|a_i|$ is the amplitude gain in that path; the squared magnitude $|a_i|^2$ is the power gain in that path. We often plot the squared magnitude of $h(\tau)$ in the dB domain and call it the *power delay profile*. This is what Rappaport calls $P(\tau)$. He shows some examples in Figures 5.4 and 5.5.

We've also measured many of these in my lab. For example, Figure 16 shows three examples.

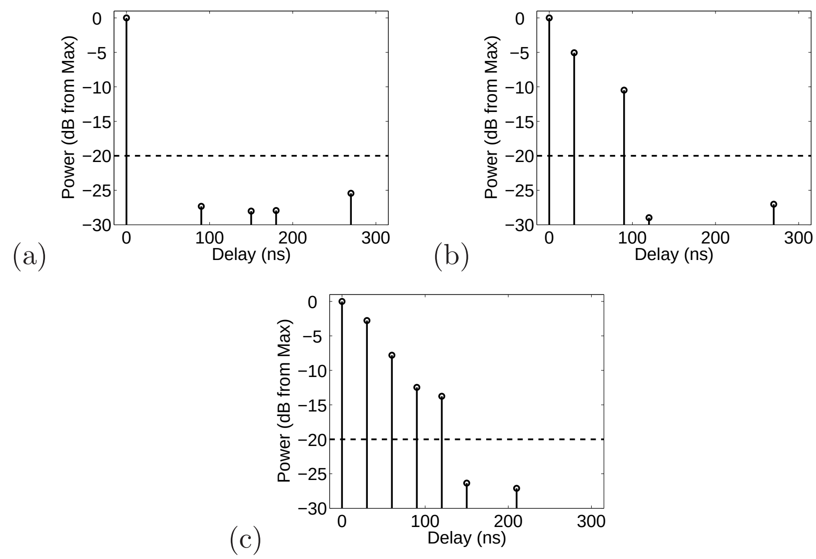


Figure 16: Measured power delay profiles (power gain normalized to the power gain of the maximum power path) in Salt Lake City (a) in residential area W of U. campus; (b) 4th S commercial district; (c) Main St., downtown [17].

13.4 Received Power

What happens to the transmitted signal $s(t)$ when it is convolved with $h(\tau)$? Well, many copies show up, at different time delays:

$$s(t) = \sum_{i=0}^{M-1} a_i e^{j\phi_i} s(\tau - \tau_i)$$

For example, what if a rectangular pulse was sent (in Digital Communications, many symbols look like a pulse)? Let $s(t) = \text{rect}\left[\frac{t}{T_b b} - \frac{1}{2}\right]$. In this case,

$$s(t) = \sum_{i=0}^{M-1} a_i e^{j\phi_i} \text{rect}\left[\frac{t-\tau_i}{T_s} - \frac{1}{2}\right]$$

where T_s is the symbol duration. Essentially, we have versions of the rect pulse piling on top of each other.

CASE 1: $\tau_i << T_s$: If the τ_is are small compared to T_s,

$$s(t) \approx \left(\sum_{i=0}^{M-1} a_i e^{j\phi_i}\right) \text{rect}\left[\frac{t}{T_s} - \frac{1}{2}\right]$$

then we have all of the pulses adding together, in a phasor sum. The sum might be constructive or destructive. But it acts on the whole pulse.

CASE 2: τ_i aren't small compared to T_s: We will have intersymbol interference, and will need an equalizer in our receiver.

Note T_s is designed by us (the engineer). Why not make T_s big? Because the symbol rate is $1/T_s$! You slow down data rate when increasing T_s.

6325 Only: Study the proof in Section 5.2.1 of $E_\theta\,[P_{WB}]$.

13.5 Time Dispersion Parameters

There are lots of τ_i so lets provide a number with more specificity, a single number that summarizes the size of the time dispersion: the RMS delay spread, σ_τ,

$$\begin{aligned}
\sigma_\tau &= \sqrt{\bar{\tau^2} - \bar{\tau}^2} \\
\bar{\tau} &= \frac{\sum_i |a_i|^2 \tau_i}{\sum_i |a_i|^2} = \frac{\sum_i P(\tau_i)\tau_i}{\sum_i P(\tau_i)} \\
\bar{\tau^2} &= \frac{\sum_i |a_i|^2 \tau_i^2}{\sum_i |a_i|^2} = \frac{\sum_i P(\tau_i)\tau_i^2}{\sum_i P(\tau_i)}
\end{aligned}$$

One good rule of thumb is that the receiver doesn't need an equalizer if $\sigma_\tau \leq 0.1 T_s$.

Lecture 10

Today: (1) Fade Distribution, (2) Doppler, (3) Fading Demo

13.6 Review from Lecture 9

We showed how the complex baseband representation of the total received voltage can be written as a sum of (complex) voltages due to the individual *multipath components* arriving at the receiver.

$$V_{TOT} = \sum_{i=0}^{M-1} V_i e^{j(-2\pi f_c \tau_i + \phi_i)} \tag{31}$$

The received power is $P_r = |V_{TOT}|^2$. Depending on the phases of these multipath components, the magnitude $|V_{TOT}|$ may be destructive or constructive (or somewhere in between). As the receiver moves, or as the center frequency changes, the magnitude $|V_{TOT}|$ will increase and decrease as the phases of the paths change from adding constructively to destructively, and back.

The phases $\exp\{-j2\pi f_c \tau_i\}$ change at different rates for different i:

- Different multipath have different delays τ_i. If we change the frequency by Δf_c, the phase change is proportional to $\Delta f_c \tau_i$. So the phase changes on longer paths more quickly than on shorter paths.
- As the receiver moves, the time delays τ_i change at different rates, as a function of the angle-of-arrival. We discuss this in the Doppler fading section below.

We discussed how the wireless channel can be written as a linear filter, with an impulse response, which we called $h(\tau)$. The received signal is the convolution of the transmitted signal with the channel's impulse response. If the channel includes many paths at wide delays, the RMS delay spread will be high, and as a result, the symbol rate in any digital communication system must be low.

14 Fade Distribution

With multipath fading so random, and so significant, how can we design a reliable system? This section describes what we can quantify about multipath fading that then allows us to design reliable wireless links.

To proceed with this quantification, we need probabilistic analysis, which requires us to move from considering purely "specular" multipath to combinations of specular and "diffuse" multipath:

- *Specular multipath*: What we've been talking about: individual multipath components, of which we said there were M total, each with its own amplitude and phase.
- *Diffuse multipath*: multipath which are each infinitely low in power, but there are infinitely many of them. Typically used to simulate the many many multipath that do arrive due to scattering and diffraction. It is easier to talk about fading when we can lump lots and lots of multipath into this "diffuse" camp, rather than specifying each individual multipath and how much power each one has.

14.1 Rayleigh Fading

Main idea: all multipath are diffuse. These diffuse multipath have small amplitude and random phase. Using the central limit theorem, and some probability (ECE 5510), we can come up with a pdf of the magnitude of the received voltage, $r = |V_{TOT}|$.

Here, the name of the random variable is R and the value r represents some number of interest.

$$f_R(r) = \begin{cases} \frac{r}{\sigma^2}\exp\left(-\frac{r^2}{2\sigma^2}\right), & r \geq 0 \\ 0, & o.w. \end{cases} \quad (32)$$

where σ is a parameter that can be set as needed. For example, if you know the mean value of the signal magnitude is r_{mean}, then you would set $\sigma = \sqrt{\frac{2}{\pi}} r_{mean}$.

The most important function for the communication system designer is the cumulative distribution function (CDF). It tells us the probability that the envelope will fall below some value r.

$$F_R(r) = P\left[R \leq r\right] = \begin{cases} 1-\exp\left(-\frac{r^2}{2\sigma^2}\right), & r \geq 0 \\ 0, & o.w. \end{cases} \quad (33)$$

<u>Def'n:</u> *Fade Margin*
the additional power loss which is included in the link budget in order to keep the received signal power higher than the minimum, with high probability.

Example: Fade margin design: Rayleigh case
Assume that it is acceptable to have the received power fall below the receiver sensitivity 1% of the time, and that the channel is Rayleigh. What fade margin (in dB) is required?
Solution: We are given $P\left[R \leq r\right] = 0.01$. We want to find the threshold in terms of the mean received signal level, so setting $r_{mean} = 1$, we have that $\sigma = 0.7979$. Then, solving (33) for r,

$$\begin{aligned} \exp\left(-\frac{r^2}{2\sigma^2}\right) &= 0.99 \\ r &= \sqrt{-2\sigma^2 \ln 0.99} = 0.1131 \end{aligned} \quad (34)$$

If the envelope falls to 0.1131, then the power falls to $(0.1131)^2 = 0.0128$, which is equal to -18.9 dB. Since the mean value is 1, or 0 dB, the required fade margin is $0 - (-18.9) = 18.9$ dB.

14.2 Ricean fading

Main idea: One specular path with amplitude $V_0 = A$ plus diffuse power.

The Rice distribution is named after its creator, Stephen O. Rice. I have his 1945 paper in my office if you want to use it for a project. Note that people just can't agree on the spelling of the adjective made out of his name, whether it is "Rician" or "Ricean". Google gives me 249k results for "Rician" and 217k results for "Ricean". Why not pick one? Who knows.

The pdf of Ricean fading is

$$f_R(r) = \begin{cases} \frac{r}{\sigma^2} \exp\left(-\frac{r^2+A^2}{2\sigma^2}\right) I_0\left(\frac{Ar}{\sigma^2}\right), & r \geq 0 \\ 0, & o.w. \end{cases} \quad (35)$$

where A is the peak amplitude of the specular (dominant) signal, and $I_0(\cdot)$ is the modified Bessel function of the first kind and zeroth order. Note Matlab has a `besseli(0, x)` function. The Ricean K-factor is defined as the ratio between the power in the specular components to the power in the diffuse components,

$$K = \frac{A^2}{2\sigma^2}, \quad \text{or } K(\text{dB}) = 10 \log_{10} \frac{A^2}{2\sigma^2}. \quad (36)$$

Note that for $K = 0$ or $K(dB) = -\infty$, the Ricean pdf is the same as the Rayleigh pdf.

There isn't an analytical expression for the CDF, analogous to (33), unfortunately. So we need to use a table or a figure. The Rappaport book only has a CDF plot for $K = 6$ dB. I am including in Figure 17 a more complete Figure. Keep this figure with your notes, it will be included in Exam 2. It also includes Rayleigh, because Rayleigh is the same as Ricean when $K = -\infty$.

Example: Fade margin design: Ricean case
Assume that it is acceptable to have the received power fall below the receiver sensitivity 1% of the time, and that the channel is Ricean, with K-factor 3 dB, 9 dB, or 15 dB. What fade margin (in dB) is required?
Solution: From the chart, on the horizontal line from 1%, the fade margin would be -15.5 dB, -7.0 dB, and -3 dB, for K-factor of 3 dB, 9 dB, and 15 dB, respectively.

15 Doppler Fading

So far we've talked about fading without movement. A static link has a fading loss. If you change center frequency on a static link, you see frequency-dependent fading. But for a link with a static center frequency but with TX and/or RX in motion, (1) the fading loss changes over time, and (2) the frequency shifts. Why is this?

In lecture 9, we came up with the top expression for the complex baseband received voltage:

$$\begin{aligned} V_{TOT} &= \sum_{i=0}^{M-1} V_i e^{j(-2\pi f_c \tau_i + \phi_i)} \\ V_{TOT} &= \sum_{i=0}^{M-1} V_i e^{j(-2\pi d_i/\lambda_c + \phi_i)} \end{aligned}$$

The second expression is rewritten with $d_i = c\tau_i$, where d_i is the distance multipath component i travels. (Is this the straight-line from the TX to RX? Or the length of the line following the

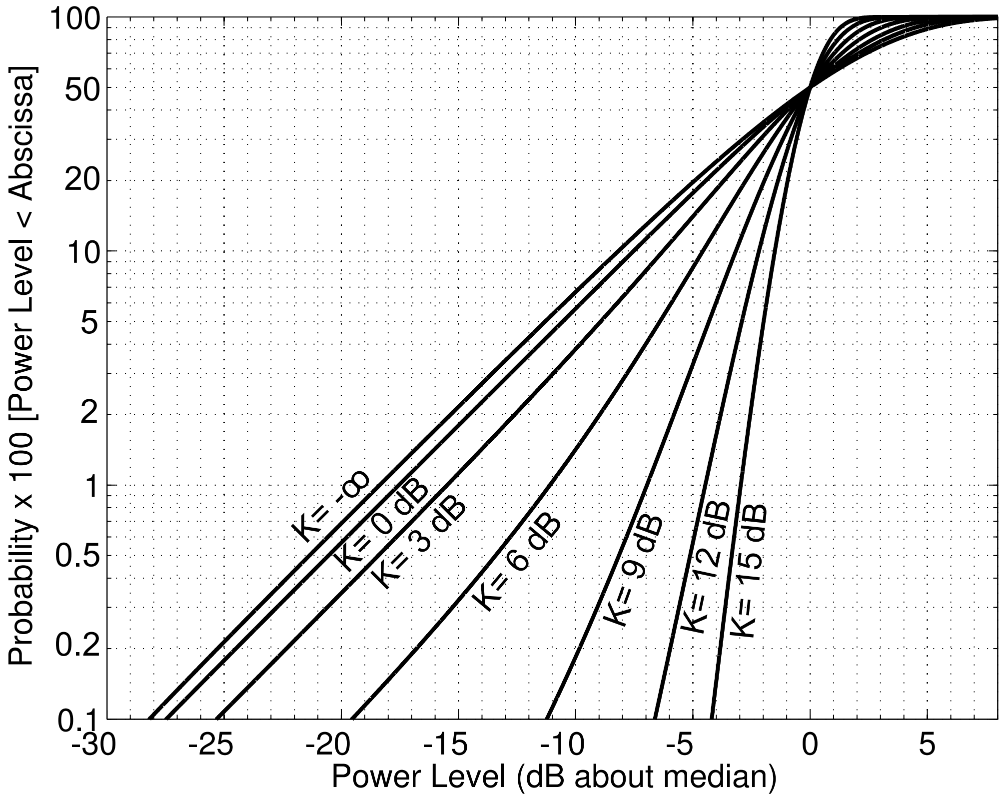


Figure 17: Ricean CDF for various K-factors. Note "Power Level" is $20\log_{10}\frac{r}{r_{median}}$, and the y-axis is the probability $\times 100$ (percentage) that the power level (fading gain compared to the median) is less than the value on the abscissa.

path? Answer: the latter.) This showed us the frequency dependence of the fading channel gain, which is $20\log_{10}|V_{TOT}|$. Now, let's talk about what happens when the receiver is moving. Motion causes the time delays to change, because the distance that the wave must travel is becoming either shorter or longer. How much shorter or longer?

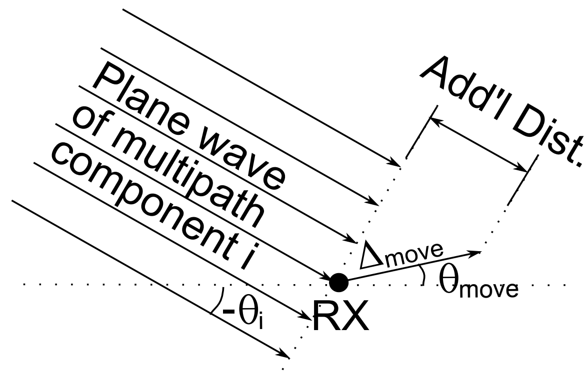


Figure 18: The plane wave of a multipath component arrives at the receiver (•). Based on the difference between the angle of movement θ_{move} and the angle of arrival of the multipath component θ_i, the multipath component distance di increases by $\cos(\theta_{move} - \theta_i)$ multiplied by the distance of travel of the receiver.

Let the angle of arrival of component i be θ_i, like the one shown in Figure 18. (Actually, consider that multipath components will arrive with different θ_i.) Let's assume I move Δ_{move} meters in the direction θ_{move}. We assume that these waves are effectively plane waves in the local (small) area near the antenna. Thus the only thing that changes when I move the antenna to a new position is that the wave is lengthened (or shortened) by a factor of the distance I moved, multiplied by the cosine of the angle in between θ_i and my direction of motion θ_{move}. After my movement of Δ_{move} meters in the direction θ_{move}, my V_{TOT} becomes:

$$V_{TOT} = \sum_{i=0}^{M-1} V_i \exp\left\{ j \left(-\frac{2\pi}{\lambda_c}[d_i + \Delta_{move}\cos(\theta_{move} - \theta_i)] + \phi_i \right) \right\}$$

15.1 One Component

First, keep things simple by considering only one arriving multipath.

$$V_{TOT} = V_0 \exp\left\{ j \left(-\frac{2\pi}{\lambda_c}[d_0 + \Delta_{move}\cos(\theta_{move} - \theta_0)] + \phi_0 \right) \right\}$$

Let's consider when the antenna is moving at a constant velocity v meters per second. Then, at time t, we have $\Delta_{move} = vt$. So,

$$V_{TOT} = V_0 \exp\left\{ j \left(-\frac{2\pi}{\lambda_c}[d_0 + vt\cos(\theta_{move} - \theta_0)] + \phi_0 \right) \right\}$$

There's a lot going on here, but we actually have not just a phase, but a complex sinusoid of,

$$f_d = -\frac{v\cos(\theta_{move} - \theta_0)}{\lambda_c} \tag{37}$$

This is called the *Doppler shift.* When one sends a frequency f_c, the received frequency of the signal has shifted up (or down, if f_d is negative) by this factor.

1. What happens when I move such that $\theta_i = \theta_{move}$?
2. What happens when I move such that $\theta_i = \theta_{move} + 90$ degrees?
3. What happens when I move such that $\theta_i = \theta_{move} + 180$ degrees?

The worst case is called the maximum Doppler frequency, $f_m = |f_d| = v/\lambda_c$.

Example: What are the maximum Doppler frequencies for a mobile in a vehicle on I-15, at 850 and 1950 MHz?

Solution: Given 1 mile per hour = 0.447 meters per second, let's say a maximum speed of 80 miles/hr, which gives $v = 35.8$ m/s. Then $f_m = 101$ Hz, or 232 Hz, for 850 or 1950 MHz, respectively.

15.2 Many Components

Now, each component contributes a complex sinusoid of frequency $f_d = -\frac{v\cos(\theta_{move}-\theta_i)}{\lambda_c}$ to the sum V_{TOT}. The effect is *frequency spreading.* That is, for each frequency that is sent, many frequencies are received. In the frequency domain, the frequency content of the signal (say, $S(\omega)$) is convolved with the Doppler spectrum $V_{TOT}(\omega)$).

If there are diffuse multipath coming equally from all directions, then the power spectrum of V_{TOT} can be determined to be:

$$S(f) = \frac{1.5}{\pi f_m \sqrt{1 - \left(\frac{f-f_c}{f_m}\right)^2}}$$

which is shown in Figure 5.20 on page 219 in Rappaport.

15.3 System Design Implications

The Doppler spread is said to have an effect when the symbol duration is long compared to $1/f_m$. Specifically, let the coherence time T_C be defined as $T_C = 0.423\frac{1}{f_m}$. Rappaport says that if $T_s > T_C$, then you have "fast fading", and demodulation will be difficult, because the channel changes during one symbol period. Instead, one should design systems with $T_s << T_C$, (for "slow fading") or equivalently, a symbol rate $1/T_s$ that is much greater than $1/T_C$. Typically, this is not a problem, except for some OFDM systems which have long T_s.

See Figure 5.15 on page 211 in Rappaport for an example of Rayleigh fading. How often does the power level drop cross a horizontal "threshold" line? This is the *level crossing rate* (LCR). For the Clarke AOA spectrum model, the average number of level crossings per second is

$$N_R = \sqrt{2\pi} f_m \rho e^{-\rho^2},$$

where $\rho = r/R_{rms}$, and R_{rms} is the root-mean square (RMS) average signal amplitude, and r is the amplitude "level" you are using. For example, to check crossings that are 10 dB down from the RMS received amplitude, use $\rho = 10^{-10/10} = 0.1$. Or, to check crossings that are 30 dB down from the RMS received *power*, use $\rho = 10^{-30/20} = 0.032$. Another useful statistic is how long the signal stays below this threshold when it goes below it. This is the *average fade duration*, denoted $\bar{t}_{fade}$ (or $\bar{\tau}$ in Rappaport). Again, for the Clarke AOA spectrum model,

$$\bar{t}_{fade} = \frac{e^{\rho^2} - 1}{\rho f_m \sqrt{2\pi}}$$

Example: Fading statistics
Let $1/T_s = 1MHz$, and use the solution for 1950 MHz highway driving above. What is the average fade duration and level crossing rate for a fade that is 20 dB lower in power than the RMS received power?

Lecture 11

Today: (1) Intro to Digital Communications

16 Digital Communications: Overview

My six word summary of digital communications: **Use linear combinations of orthogonal waveforms**.

16.1 Orthogonal Waveforms

My "engineering" definition of a set of orthogonal waveforms: They are waveforms that can be separated at the receiver.

Def'n: *Orthogonal*
Two real-valued waveforms (finite-energy functions) $\phi_1(t)$ and $\phi_2(t)$ are *orthogonal* if

$$\int_{-\infty}^{\infty} \phi_1(t)\phi_2(t)dt = 0$$

Two complex-valued waveforms $\phi_1(t)$ and $\phi_2(t)$ are orthogonal if

$$\int_{-\infty}^{\infty} \phi_1(t)y^*(t)dt = 0$$

where $y^*(t)$ is the complex conjugate of $\phi_2(t)$.

Def'n: *Orthogonal Set*
N waveforms $\phi_1(t), \ldots, \phi_N(t)$ are mutually orthogonal, or form an orthogonal set, if every pair of waveforms $\phi_i(t)$, $\phi_j(t)$, for $i \neq j$, is orthogonal.

Example: Sine and Cosine

Let

$$\begin{aligned}
\phi_1(t) &= \begin{cases} \cos(2\pi t), & 0 < t \le 1 \\ 0, & o.w. \end{cases} \\
\phi_2(t) &= \begin{cases} \sin(2\pi t), & 0 < t \le 1 \\ 0, & o.w. \end{cases}
\end{aligned}$$

Are $\phi_1(t)$ and $\phi_2(t)$ orthogonal?
Solution: Using $\sin 2x = 2\cos x \sin x$,

$$\begin{aligned}
\int_{-\infty}^{\infty} \phi_1(t)\phi_2(t)dt &= \int_0^1 \cos(2\pi t)\sin(2\pi t)dt \\
&= \int_0^1 \frac{1}{2}\sin(4\pi t)dt \\
&= \frac{-1}{8\pi}\cos(4\pi t)\Big|_0^1 = \frac{-1}{8\pi}(1-1) = 0
\end{aligned}$$

So, yes, $\phi_1(t)$ and $\phi_2(t)$ are orthogonal.

Example: Frequency Shift Keying
Assume $T_s >> 1/f_c$, and show that these two are orthogonal.

$$\begin{aligned}
\phi_0(t) &= \begin{cases} \cos(2\pi f_c t), & 0 \le t \le T_s \\ 0, & o.w. \end{cases} \\
\phi_1(t) &= \begin{cases} \cos\left(2\pi\left[f_c + \frac{1}{T_s}\right]t\right), & 0 \le t \le T_s \\ 0, & o.w. \end{cases}
\end{aligned}$$

Solution: The integral of the product of the two must be zero. Checking, and using the identity for the product of two cosines,

$$\begin{aligned}
&\int_0^{T_s} \cos(2\pi f_c t)\cos\left(2\pi\left[f_c + \frac{1}{T_s}\right]t\right)dt \\
&= \frac{1}{2}\int_0^{T_s} \cos(2\pi t/T_s)\,dt + \int_0^{T_s} \cos(4\pi f_c t + 2\pi t/T_s)\,dt \\
&= \frac{1}{2}\left[\frac{T_s}{2\pi}\sin(2\pi t/T_s)\Big|_0^{T_s} + \left[\frac{1}{2\pi(2f_c+1/T_s)}\sin(2\pi(2f_c+1/T_s)t)\Big|_0^{T_s}\right.\right.
\end{aligned}$$

The second term has a $\frac{1}{2\pi(2f_c+1/T_s)}$ constant out front. Because f_c is very high, this term will be very very low. The sine term is limited to between -1 and +1 so it will not cause the second term to be large. So we will approximate this second term as zero.

$$\int_{-\infty}^{\infty} \phi_1(t)\phi_2(t)dt \approx \frac{1}{2}\frac{T_s}{2\pi}[\sin(2\pi) - \sin(0)] = 0$$

Thus the two different frequency waveforms are orthogonal.

16.2 Linear Combinations

What is a linear combination of orthogonal waveforms? Well, consider the orthogonal set $\phi_1(t), \ldots, \phi_N(t)$. A linear combination $s_i(t)$ is

$$s_i(t) = a_{i,1}\phi_1(t) + a_{i,2}\phi_2(t) + \cdots + a_{i,N}\phi_N(t) = \sum_{i=k}^{N} a_{i,k}\phi_k(t)$$

We also call the a linear combination a *symbol*. We use subscript i to indicate that it's not the only possible linear combination (or symbol). In fact, we will use M different symbols, so $i = 1, \ldots, M$, and we will use $s_1(t), \ldots, s_M(t)$.

We represent the ith symbol (linear combination of the orthogonal waveforms), $s_i(t)$, as a vector for ease of notation:

$$\mathbf{a}_i = [a_{i,1}, a_{i,2}, \ldots, a_{i,N}]^T$$

The superscript T is for transpose – $\mathbf{a}_i$ is a column vector. Vectors are easy to deal with because they can be plotted in vector space, to show graphically what is going on. We call the plot of all possible $\mathbf{a}_i$, that is, for $i = 1, \ldots M$, the *constellation diagram*. Some examples are shown in Figure 19.

16.3 Using M Different Linear Combinations

Here is how a transmitter *uses* the different linear combinations to convey digital bits to the receiver. First, consider that there are M different symbols for the TX to chose from. Each symbol is described by a $\log_2 M$-length bit sequence. For example, if there are 8 possible combinations, we would label them 000, 001, 011, 010, 110, 111, 101, 100.

The transmitter knows which $\log_2 M$-bit sequence it wants to send. It picks the symbol that corresponds to that bit sequence, let's call it symbol i. Then it sends $s_i(t)$.

If the receiver is able to determine that symbol i was sent, it will correctly receive those $\log_2 M$ bits of information. In this example, it will receive three bits of information.

16.4 Reception

At a receiver, because we've used an orthogonal set, we can determine how much of each waveform was sent. This can be done using a bank of matched filters. In short, we can recover an attenuated $\mathbf{a}_i$ plus noise. We might write $\mathbf{x} = g\mathbf{a}_i + n_i$ where g is the total gain ($g \ll 1$) introduced by the channel and antennas, and n_i is the random additive noise added in at the receiver. Assume we have receiver gain that multiplies the received filter by $1/g$ (to cancel out g). If the resulting $\hat{\mathbf{a}}_i = \mathbf{x}/g$ is close enough to the actual transmitted signal $\mathbf{a}_i$, then the correct bit string is received.

How does the receiver compute the $\hat{\mathbf{a}}_i$? It does this using a matched filter receiver. A matched filter receiver has one bank for each orthogonal waveform (the $\phi_1(t), \ldots, \phi_N(t)$. Bank k has a filter that quantifies how much of $\phi_k(t)$ is present in the received signal. Because the waveforms are orthogonal, the job is simple. An analogy would be to a really good division of labor between members of a team – filter k comes up with the answer for waveform k, without being distracted by the other waveforms which may or may not be present in the received signal.

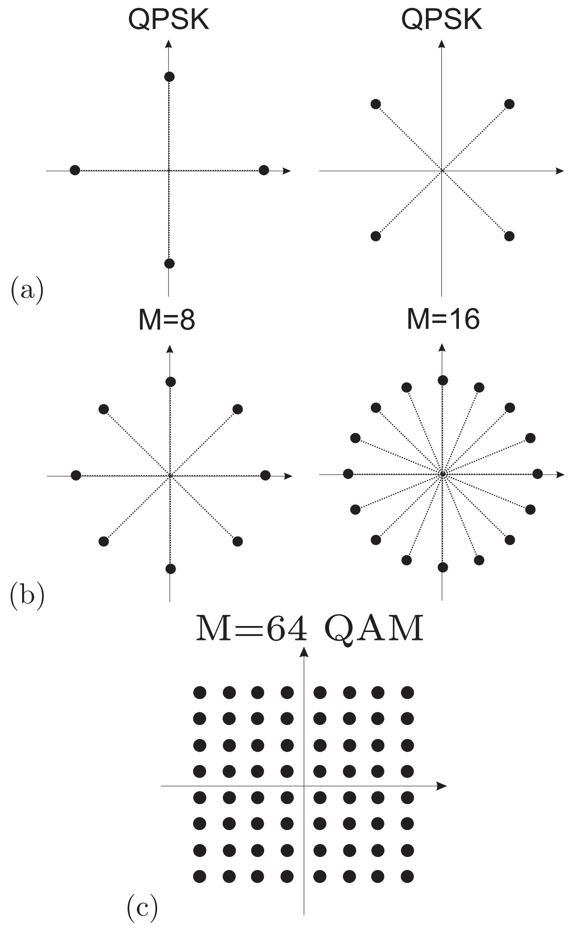


Figure 19: Signal constellations for (a) $M = 4$ PSK (a.k.a. BPSK), (b) $M = 8$ and $M = 16$ PSK, and (c) 64-QAM.

16.5 How to Choose a Modulation

A *digital modulation* is the choice of: (1) the linear combinations $\mathbf{a}_1, \ldots, \mathbf{a}_M$ and, (2) the orthogonal waveforms $\phi_1(t), \ldots, \phi_N(t)$. We will choose a digital modulation as a tradeoff between the following characteristics:

1. Bandwidth efficiency: How many bits per second (bps) can be sent per Hertz of signal bandwidth. Thus the bandwidth efficiency has units of bps/Hz.

2. Power efficiency: How much energy per bit is required at the receiver in order to achieve a desired fidelity (low bit error rate). We typically use S/N or E_s/N_0 or E_b/N_0 as our figure of merit.

3. Cost of implementation: Things like symbol and carrier synchronization, and linear transmitters, require additional device cost, which might be unacceptable in a particular system.

16.6 Intersymbol Interference and Bandwidth

To send a symbol, we will send a finite-duration waveform. This waveform requires some bandwidth. We can find this bandwidth by taking the Fourier transform of the waveform.

For example, what is the bandwidth required to send a rectangular pulse? Let $s_0(t) = \text{rect}(t/T_s)$. It has

$$S_0(f) = T_s \frac{\sin(\pi f T_s)}{\pi f T_s}$$

This has frequency content at all frequencies, *i.e.*, infinite bandwidth!

Can we send a new symbol every T_s and still not have symbols interfering with each other at the receiver? Yes, we can! We mentioned orthogonality. Well, **two symbols sent at different times must also be orthogonal**, so that we can send information at one time, and then at the next, without interacting with each other.

As an example, again, rectangular pulses:

$$\begin{aligned} p_0(t) &= \begin{cases} 1, & 0 < t < T_s \\ 0, & o.w. \end{cases} \\ p_1(t) &= \begin{cases} 1, & T_s < t < 2T_s \\ 0, & o.w. \end{cases} \\ p_2(t) &= \begin{cases} 1, & 2T_s < t < 3T_s \\ 0, & o.w. \end{cases} \end{aligned} \tag{38}$$

form an orthogonal set. The problem is, again, the infinite bandwidth.

Can we reduce the bandwidth and still have zero intersymbol interference? Yes. There is a method to come up with pulse shapes that are orthogonal to other versions of itself that are delayed by integer multiples of T_s. For more information, look in the book at Section 6.6.1, which discusses the Nyquist zero intersymbol interference (zero ISI) criterion.

A good solution (although not the only possibility) is to use a pulse shape in the "square root raised cosine" (SRRC) class. A pulse shape is SRRC if it has frequency domain representation $P_{SRRC}(f)$,

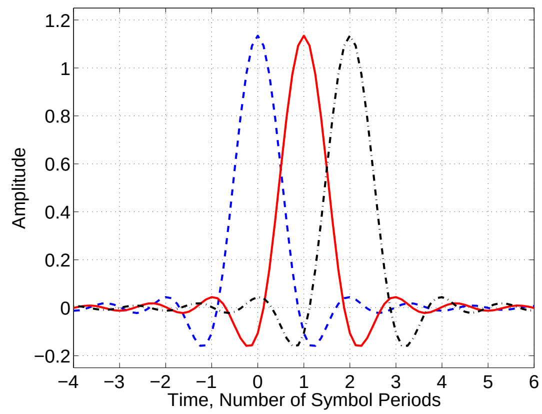
$$P_{SRRC}(f) = \begin{cases} \sqrt{T_s}, & 0 \le |f| \le \frac{1-\alpha}{2T_s} \\ \sqrt{\frac{T_s}{2}\left\{1 + \cos\left[\frac{\pi T_s}{\alpha}\left(|f| - \frac{1-\alpha}{2T_s}\right)\right]\right\}}, & \frac{1-\alpha}{2T_s} \le |f| \le \frac{1+\alpha}{2T_s} \\ 0, & o.w. \end{cases} \tag{39}$$

You can find $p(t)$ by taking the inverse Fourier transform of the above expression, and an example is shown in Figure 20.

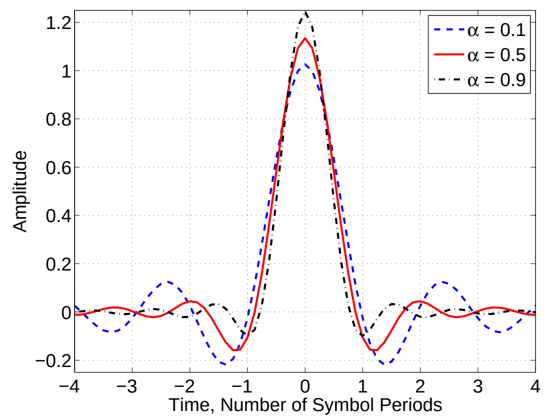
There is a parameter α in (39) which describes the width of the spectrum. Specifically, the bandwidth of signals sent using this pulse shape are $(1+\alpha)/T_s$. In other words, compared to the symbol rate, the bandwidth is multiplied by a factor of $1+\alpha$. One would like $\alpha = 0$, but that makes the receiver's work more prone to large errors when time synchronization is imperfect. Typically, the minimum α we can build robust receivers for is the range of 0.2-0.3.

Lecture 12

Today: (1) Modulation (2) Fidelity, (3) Link Budgets & Modulation



(a)



(b)

Figure 20: (a) Square root raised cosine pulses separated by T_s form an orthogonal set. These three have $\alpha = 0.5$. (b) SRRC pulses at various α parameters.

17 Modulation

Last lecture we talked about how a digital transmitter sends one of M symbols, that is, linear combinations, of a small number N of orthogonal waveforms. The TX's choice from the M possible linear combinations lets it send one of $\log_2 M$ bits per symbol. This section talks specifically about the choices for the orthogonal waveforms and the symbols

17.1 PAM

Three "types" of modulations only use one (orthogonal) waveform, $\phi_0(t) = \cos(2\pi f_c t)p(t)$, where $p(t)$ is the pulse shape. (It has nothing to be orthogonal to, except for waveforms sent during other symbol periods.) A linear combination, then, is just $a_{i,0}p(t)$, for some constant $a_{i,0}$. A few different modulations are so named by their choice of M different choices of $a_{i,0}$, for $i = 0, \ldots, M-1$.

- On-off-keying (OOK): $M = 2$, and we choose $a_{0,0} = 0$, and $a_{1,0} = 1$. When sending a "0" bit, the transmitter doesn't send anything. It only actually transmits energy when sending a "1" bit.

- Binary PAM, a.k.a. binary phase-shift-keying (BPSK): $M = 2$, and we choose $a_{0,0} = -1$, and $a_{1,0} = 1$. Now, the amplitude of the sinusoid being transmitted is switched from a +1 to a -1, or vice versa, when the bit switches. This is also a phase shift of 180^o, which is why it can be called "phase shift" keying.

- M-ary PAM: Binary PAM is extended in M-ary PAM to include M (for M equal to some power of 2) equally spaced amplitudes, centered on zero. So, for example, $M = 4$ PAM would have $a_{0,0} = -3$, $a_{1,0} = -1$, $a_{2,0} = +1$, $a_{3,0} = +3$. In general, for M-ary PAM, $a_{i,0} = 2i - M + 1$.

Note that *differential phase shift keying* (DPSK) is an implementation variant of BPSK, which we will discuss next lecture.

17.2 M-ary QAM and PSK

Many types of modulations use the following two orthogonal waveforms:

$$\begin{aligned} \phi_0(t) &= \cos(2\pi f_c t) p(t) \\ \phi_1(t) &= \sin(2\pi f_c t) p(t) \end{aligned}$$

where $p(t)$, again, is the pulse shape. These symbols $\mathbf{a}_i = [a_{i,0}, a_{i,1}]^T$ can then be plotted on a 2-D graph, typically with the $\phi_0(t)$ amplitude plotted on the horizontal axis, and the $\phi_1(t)$ amplitude plotted on the vertical axis.

There are two main "types" of modulations which use these two orthogonal waveforms:

- Phase-shift keying (PSK): PSK places all $\mathbf{a}_i$ uniformly on the unit circle. That is, $\|\mathbf{a}_i\| = 1$ for all i. For $M = 2$, this is the same as BPSK. For $M = 4$, the symbols are spaced every 90 degrees and thus the constellation looks like the four corners of a square, and is also called quadrature phase shift keying (QPSK). For $M = 8$, symbols are every 45 degrees apart.

- QAM: Generally, any modulation that uses $\phi_0(t)$ and $\phi_1(t)$ above can be considered as QAM. So PSK is a sometimes a subtype of QAM. But QAM is not limited to $\|\mathbf{a}_i\| = 1$. For example, square QAM has $M = 2^{2k}$ for some integer $k \geq 4$, where symbols are arranged in a square grid, centered at zero. Note $M = 4$ square QAM is the same as QPSK. But $M = 16$ and $M = 64$ are also common values of M for square QAM.

We drew five different constellation diagrams of PSK and QAM modulations in the Lecture 11 notes, on page 4.

Note that for QPSK or square M-QAM, you send $\log_2 M$ bits total per symbol. We can also look at the modulation as sending two independent signals, one on the in-phase (horizontal axis) and one on the quadrature (vertical axis), each sending $\frac{1}{2}\log_2 M$ bits per symbol.

Note that OQPSK and $\pi/4$ QPSK are variations on QPSK, that have equivalent fidelity performance and identical bandwidth as QPSK, but they are "constant envelope" modulations, which we will discuss next lecture.

Bandwidth: Note that for PAM, PSK, and QAM, assuming SRRC pulse shaping with rolloff factor α, the null-to-null bandwidth of the signal is $B = (1 + \alpha)/T_s$, where $1/T_s$ is the symbol rate.

17.3 FSK

Binary FSK modulations use the following orthogonal waveforms:

$$\begin{aligned} \phi_0(t) &= \cos\left(2\pi \left[f_c - n\Delta f\right] t\right) p(t) \\ \phi_1(t) &= \cos\left(2\pi \left[f_c + n\Delta f\right] t\right) p(t) \end{aligned}$$

for some positive integer n, where $\Delta f = \frac{1}{4T_s}$. We showed that for $n = 2$, that is, when the frequency difference ("offset") is $1/T_s$, that the two are orthogonal. Standard binary FSK uses $n = 2$.

Bandwidth: For Binary FSK, when a SRRC pulse is used in $p(t)$, the transmitted signal bandwidth is given by

$$B = 2n\Delta f + (1+\alpha)R_s$$

where $R_s = R_b = 1/T_b$. So, for Binary FSK,

$$B = 4\frac{1}{4T_s} + (1+\alpha)R_s = (2+\alpha)R_s$$

17.4 MSK

You can also show that orthogonality is achieved for any integer n in (40). Since putting the two sinusoids closer together reduces the bandwidth of the signal, this is a good idea for spectral efficiency. When $n = 1$ binary FSK is called *minimum shift keying (MSK)*.

But MSK is actually the same as OQPSK with a half-cosine pulse shape. This is not something we discuss until next lecture.

17.5 Receiver Complexity Options

One issue in receivers is the costs of coherent vs. non-coherent reception.

Def'n: *Coherent Reception*
A receiver is coherent if it is phase-synchronous, that is, it estimates and then uses the phase of the incoming signal.

Coherent reception requires a phase-locked loop (PLL) for each different carrier frequency in the signal. This can be a problem for FSK receivers that have high M. Also, accurate phase-locked loops can be difficult to achieve for mobile radios for cases with high Doppler.

Differential reception is another technique to aid receivers which operate without accurate phase. When Doppler is present, and the coherent receiver ends up with a consistent phase error of 180 degrees, a standard receiver will flip every subsequent bit estimate. A differential receiver encodes only the change in phase. In the same case, when the phase is flipped by 180 degrees, one bit error will be made because of the flip, but the subsequent bits will be all correct.

18 Fidelity

Now, let's compare the fidelity (probability of bit error and probability of symbol error) across different modulations and receiver complexity options. For the same energy per bit, using a different modulation would result in a different fidelity.

Recall $N = FkT_0B$. We also define $N_0 = FkT_0$, so that $N = N_0B$. The units of N_0 are Watts per Hertz, or equivalently, Joules.

The signal power at the receiver can also be written in terms of the energy per symbol or energy per bit. Since energy is power times time, the energy used to send a symbol duration T_s is $\mathcal{E}_s = ST_s$, where S is the received power P_r, and $\mathcal{E}_s$ has units Joules per symbol. To calculate the energy needed to receive one bit, we calculate $\mathcal{E}_b = ST_s/\log_2 M$ Joules per bit. To shorten this, we define $T_b = T_s/\log_2 M$ and then $\mathcal{E}_b = ST_b$ or $\mathcal{E}_b = S/R$, where $R = 1/T_b$ is the bit rate.

Name	P [**symbol error**]	P [**bit error**]
BPSK	$= \mathrm{Q}\left(\sqrt{\frac{2\mathcal{E}_b}{N_0}}\right)$	same
OOK	$= \mathrm{Q}\left(\sqrt{\frac{\mathcal{E}_b}{N_0}}\right)$	same
DPSK	$= \frac{1}{2}\exp\left(-\frac{\mathcal{E}_b}{N_0}\right)$	same
M-PAM	$= \frac{2(M-1)}{M}\mathrm{Q}\left(\sqrt{\frac{6\log_2 M}{M^2-1}\frac{\mathcal{E}_b}{N_0}}\right)$	$\approx \frac{1}{\log_2 M} P$ [symbol error]
QPSK		$= \mathrm{Q}\left(\sqrt{\frac{2\mathcal{E}_b}{N_0}}\right)$
M-PSK	$\leq 2\mathrm{Q}\left(\sqrt{2(\log_2 M)\sin^2(\pi/M)\frac{\mathcal{E}_b}{N_0}}\right)$	$\approx \frac{1}{\log_2 M} P$ [symbol error]
Square M-QAM		$\approx \frac{4}{\log_2 M}\frac{(\sqrt{M}-1)}{\sqrt{M}}\mathrm{Q}\left(\sqrt{\frac{3\log_2 M}{M-1}\frac{\mathcal{E}_b}{N_0}}\right)$
2-non-co-FSK	$= \frac{1}{2}\exp\left[-\frac{\mathcal{E}_b}{2N_0}\right]$	same
M-non-co-FSK	$= \sum_{n=1}^{M-1}\binom{M-1}{n}\frac{(-1)^{n+1}}{n+1}\exp\left[-\frac{n\log_2 M}{n+1}\frac{\mathcal{E}_b}{N_0}\right]$	$= \frac{M/2}{M-1} P$ [symbol error]
2-co-FSK	$= \mathrm{Q}\left(\sqrt{\frac{\mathcal{E}_b}{N_0}}\right)$	same
M-co-FSK	$\leq (M-1)\mathrm{Q}\left(\sqrt{\log_2 M\frac{\mathcal{E}_b}{N_0}}\right)$	$= \frac{M/2}{M-1} P$ [symbol error]

Table 1: Summary of probability of bit and symbol error formulas for several modulations.

The probability of bit error is a function of the ratio $\mathcal{E}_b/N_0$. For example, for BPSK,

$$P_{e,BPSK} = \mathrm{Q}\left(\sqrt{\frac{2\mathcal{E}_b}{N_0}}\right)$$

where $\mathrm{Q}(z)$ is the tail probability of a Gaussian r.v., given by

$$\mathrm{Q}(z) = \frac{1}{2}\mathrm{erfc}\left(\frac{z}{\sqrt{2}}\right)$$

Put this into Matlab or your calculator, it is a useful function. You might also use the table or graph on pages 645-646 of the Rappaport book.

Other symbol error rates and bit error rates for a variety of modulations is shown in Table 18.

What are the most important? In my opinion: BPSK, DPSK, QPSK, 16-QAM, 64-QAM, 2-non-co-FSK, 2-co-FSK. For the homework, you will make a table, for each modulation, listing P [bit error] formula, and bandwidth. You will also plot P [bit error] as a function of $\frac{\mathcal{E}_b}{N_0}$. Both will be very useful to have in your portfolio so that you have them available on the exam 2. An example is shown in Figure 21.

19 Link Budgets with Digital Modulation

Link budgeting was discussed in the first part of this course assuming that the required SNR was known. But actually, we deal with a required $\frac{\mathcal{E}_b}{N_0}$. Note

$$\frac{\mathcal{E}_b}{N_0} = \frac{S}{N_0}\frac{1}{R_b} = \frac{S}{N}\frac{B}{R_b}$$

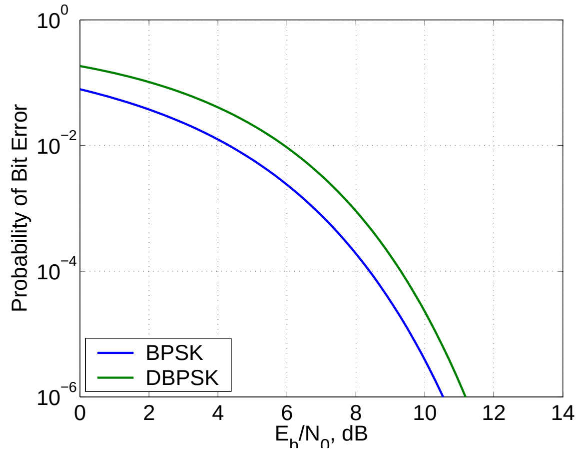


Figure 21: Comparison of probability of bit error for BPSK and Differential BPSK.

Typically the ratio of bandwidth to bit rate is known. For example, what is $\frac{B}{R_b}$ for SRRC pulse shaping?

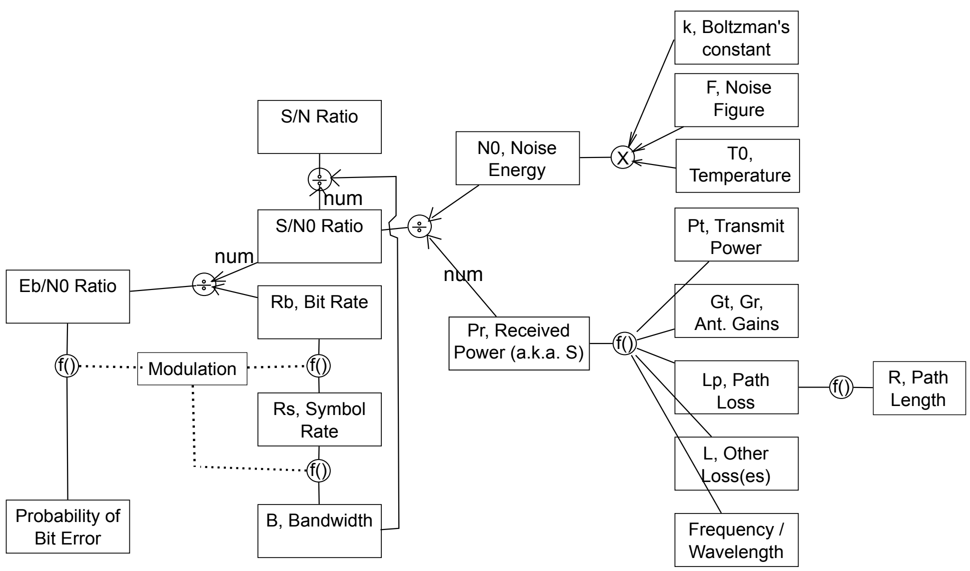


Figure 22: Link Budgeting including Modulation: This relationship graph summarizes the relationships between important variables in the link budget and the choice of modulation.

19.1 Shannon-Hartley Channel Capacity

We've talked about S/N, the signal power (S) to noise power (N) ratio. In the link budget section, we provided a S/N required for a given modulation. Here, we talk about the fundamental limits for bandwidth efficiency for a given S/N.

<u>Def'n:</u> *Bandwidth efficiency*
The bandwidth efficiency of a digital communication system is ratio of $\eta_B = R/B$, where R is the bits per second achieved on the link, and B is the signal bandwidth occupied by the signal. Bandwidth efficiency has units of bits per second per Hertz.

The limit on bandwidth efficiency is given by Claude Shannon [23], who extended work by Ralph Hartley (a UofU alum!). The Shannon-Hartley theorem is,

$$\frac{R_{max}}{B} = \log_2\left(1 + \frac{S}{N}\right)$$

where R_{max} is the maximum possible bit rate which can be achieved on the channel for the given signal to noise ratio.

Example: Maximum for various S/N (dB)
What is the maximum bandwidth efficiency of a link with $S/N(\text{dB}) = 10, 15, 20$? What maximum bit rate can be achieved per 30 kHz of spectrum (one AMPS/USDC channel)? Is this enough for 3G?
Solution: $S/N(\text{dB}) = 10, 15, 20$ translates into $S/N = 10, 31.6, 100$. Then

$$\begin{aligned}
\frac{R_{max}}{B} &= \log_2(11) = 3.46 \\
\frac{R_{max}}{B} &= \log_2(32.6) = 5.03 \\
\frac{R_{max}}{B} &= \log_2(101) = 6.66
\end{aligned}$$

With 30 kHz, multiplying, we have $R_{max} = 104, 151, 200$ kbps. No, this isn't enough for 3G cellular, which says it aims to achieve up to 14 Mbps.

Shannon's bound is great as engineers, we can come up with a quick answer for what we *cannot* do. But it doesn't necessarily tell us what we *can* achieve. The problem is that it is very difficult to get within a couple of dB of Shannon's bound. So, then we have to resort to the performance equations for particular modulation types.

19.2 Examples

Example: 200 kHz

What bit rate can be achieved on a 200 kHz if the S/N ratio is 20 dB?

Example: 1.25 MHz

What bit rate can be achieved on a 200 kHz if the S/N ratio is 12 dB?

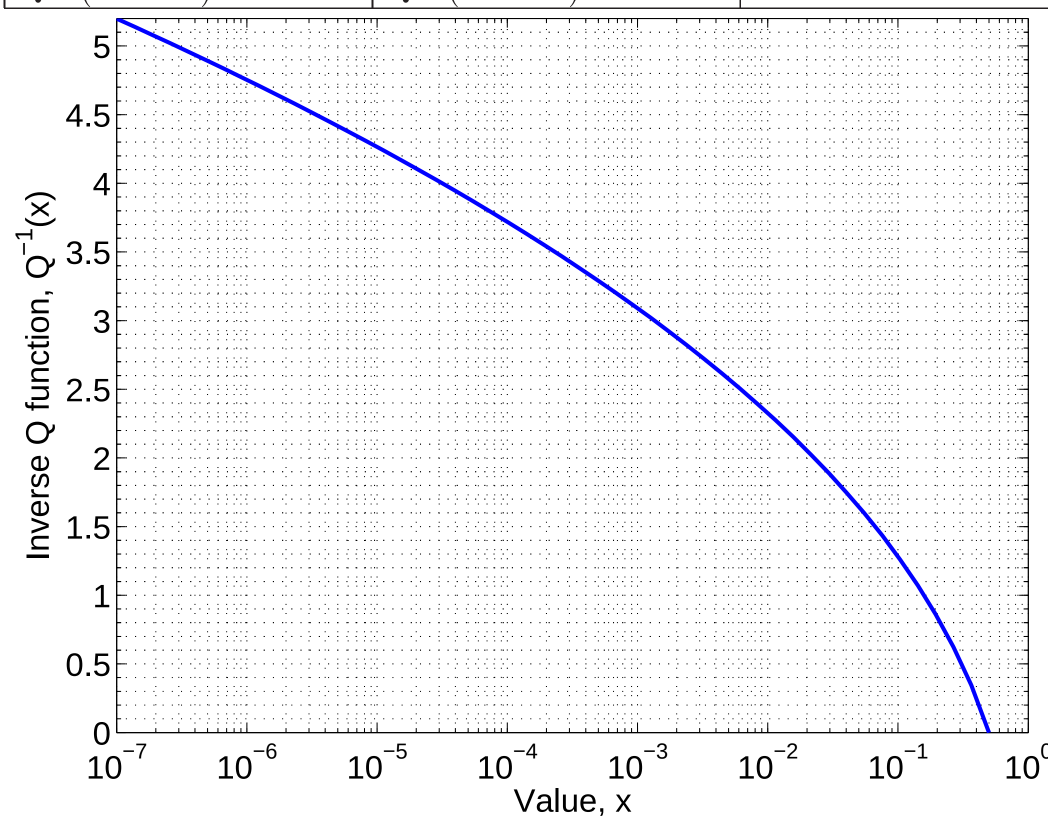
Lecture 13

Today: (1) Implementation Costs

19.3 Q-Function and Inverse

TABLE OF THE $Q^{-1}(\cdot)$ FUNCTION:

$Q^{-1}(1\times10^{-6})=4.7534$	$Q^{-1}(1\times10^{-4})=3.719$	$Q^{-1}(1\times10^{-2})=2.3263$
$Q^{-1}(1.5\times10^{-6})=4.6708$	$Q^{-1}(1.5\times10^{-4})=3.6153$	$Q^{-1}(1.5\times10^{-2})=2.1701$
$Q^{-1}(2\times10^{-6})=4.6114$	$Q^{-1}(2\times10^{-4})=3.5401$	$Q^{-1}(2\times10^{-2})=2.0537$
$Q^{-1}(3\times10^{-6})=4.5264$	$Q^{-1}(3\times10^{-4})=3.4316$	$Q^{-1}(3\times10^{-2})=1.8808$
$Q^{-1}(4\times10^{-6})=4.4652$	$Q^{-1}(4\times10^{-4})=3.3528$	$Q^{-1}(4\times10^{-2})=1.7507$
$Q^{-1}(5\times10^{-6})=4.4172$	$Q^{-1}(5\times10^{-4})=3.2905$	$Q^{-1}(5\times10^{-2})=1.6449$
$Q^{-1}(6\times10^{-6})=4.3776$	$Q^{-1}(6\times10^{-4})=3.2389$	$Q^{-1}(6\times10^{-2})=1.5548$
$Q^{-1}(7\times10^{-6})=4.3439$	$Q^{-1}(7\times10^{-4})=3.1947$	$Q^{-1}(7\times10^{-2})=1.4758$
$Q^{-1}(8\times10^{-6})=4.3145$	$Q^{-1}(8\times10^{-4})=3.1559$	$Q^{-1}(8\times10^{-2})=1.4051$
$Q^{-1}(9\times10^{-6})=4.2884$	$Q^{-1}(9\times10^{-4})=3.1214$	$Q^{-1}(9\times10^{-2})=1.3408$
$Q^{-1}(1\times10^{-5})=4.2649$	$Q^{-1}(1\times10^{-3})=3.0902$	$Q^{-1}(1\times10^{-1})=1.2816$
$Q^{-1}(1.5\times10^{-5})=4.1735$	$Q^{-1}(1.5\times10^{-3})=2.9677$	$Q^{-1}(1.5\times10^{-1})=1.0364$
$Q^{-1}(2\times10^{-5})=4.1075$	$Q^{-1}(2\times10^{-3})=2.8782$	$Q^{-1}(2\times10^{-1})=0.84162$
$Q^{-1}(3\times10^{-5})=4.0128$	$Q^{-1}(3\times10^{-3})=2.7478$	$Q^{-1}(3\times10^{-1})=0.5244$
$Q^{-1}(4\times10^{-5})=3.9444$	$Q^{-1}(4\times10^{-3})=2.6521$	$Q^{-1}(4\times10^{-1})=0.25335$
$Q^{-1}(5\times10^{-5})=3.8906$	$Q^{-1}(5\times10^{-3})=2.5758$	$Q^{-1}(5\times10^{-1})=0$
$Q^{-1}(6\times10^{-5})=3.8461$	$Q^{-1}(6\times10^{-3})=2.5121$	
$Q^{-1}(7\times10^{-5})=3.8082$	$Q^{-1}(7\times10^{-3})=2.4573$	
$Q^{-1}(8\times10^{-5})=3.775$	$Q^{-1}(8\times10^{-3})=2.4089$	
$Q^{-1}(9\times10^{-5})=3.7455$	$Q^{-1}(9\times10^{-3})=2.3656$	



Example: Range Comparison
Consider a wireless LAN system at $f_c = 900$ MHz with P_t =1 W, $G_t = 2$, $G_r = 1.6$, a receiver noise figure $F = 8$, a fade margin of 18 dB, SRRC pulse shaping with $\alpha = 0.25$, and path loss given by the log-distance model with $n_p = 3.2$ after a reference distance 1 meter (and free space up to 1 m).

1. Use DPSK with 1 Mbps, with a P[bit error] of 10^{-3} or 10^{-6}.
2. Use 64-QAM with 8 Mbps, with a P[bit error] of 10^{-3} or 10^{-6}.

Solution: For DPSK, $P\left[\text{bit error}\right] = \frac{1}{2}\exp\left(-\frac{\mathcal{E}_b}{N_0}\right)$. So for $P\left[\text{bit error}\right] = 10^{-3}$,

$$\frac{\mathcal{E}_b}{N_0} = -\ln(2P\left[\text{bit error}\right]) = 6.2$$

The bit rate is $R_b = 10^6$. So

$$\frac{S}{N_0} = \frac{\mathcal{E}_b}{N_0}R_b = 6.2 \times 10^6 1/\text{s}$$

The $N_0 = kT_0F$, in our case,

$$N_0 = kT_0F = 1.38 \times 10^{-23}\text{J/K } 290\text{K } 8 = 3.2 \times 10^{-20}\text{J}$$

So S (which is the received power, the same thing as P_r),

$$P_r = S = \frac{\mathcal{E}_b}{N_0}R_bN_0 = 6.2 \times 10^6 1/\text{s } 3.2 \times 10^{-20}\text{J} = 1.99 \times 10^{-13}\text{W}$$

or $P_r(dBW) = -127.0$ dBW. Then using the path loss equations,

$$P_0(1m) = 10\log_{10}\left(P_tG_tG_r\left(\frac{\lambda}{4\pi(1m)}\right)^2\right) = -26.5 \text{ dBW} \tag{40}$$

Then, we can formulate the link budget:

$$\begin{aligned}
P_r &= P_0(1m) - 10n\log_{10}\frac{d}{1m} - \text{Fade Margin} \\
-127dBW &= -26.5dBW - 32\log_{10}\frac{d}{1m} - 18 \\
-82.5dBW &= -32\log_{10}\frac{d}{1m} \\
d &= 1m \times 10^{82.5/32} \approx 380m
\end{aligned} \tag{41}$$

For $P\left[\text{bit error}\right] = 10^{-6}$,

$$\frac{\mathcal{E}_b}{N_0} = -\ln(2P\left[\text{bit error}\right]) = 13.1$$

So ,

$$P_r = S = \frac{\mathcal{E}_b}{N_0}R_bN_0 = 13.1 \times 10^6 1/\text{s } 3.2 \times 10^{-20}\text{J} = 4.2 \times 10^{-13}\text{W}$$

or $P_r(dBW) = -123.7$ dBW. From the link budget:

$$\begin{aligned} -123.7dBW &= -26.5dBW - 32\log_{10}\frac{d}{1m} - 18 \\ -79.2dBW &= -32\log_{10}\frac{d}{1m} \\ d &= 1m \times 10^{79.2/32} \approx 300m \end{aligned} \quad (42)$$

For 64-QAM:

$$\Pr\text{bit error} = \frac{4}{\log_2 M}\frac{(\sqrt{M}-1)}{\sqrt{M}}\mathrm{Q}\left(\sqrt{\frac{3\log_2 M}{M-1}\frac{\mathcal{E}_b}{N_0}}\right)$$

with $M = 64$, $\log_2 64 = 6$ and $\sqrt{64} = 8$, so

$$\Pr\text{bit error} = \frac{7}{12}\mathrm{Q}\left(\sqrt{\frac{2}{7}\frac{\mathcal{E}_b}{N_0}}\right)$$

So

$$\frac{\mathcal{E}_b}{N_0} = \frac{7}{2}\left[Q^{-1}\left(\frac{12}{7}\Pr\text{bit error}\right)\right]^2$$

For $\Pr$ bit error $= 10^{-3}$, we need the Q^{-1} of argument $\frac{12}{7} \times 10^{-3} = 1.7 \times 10^{-3}$. From the Q-inverse figure, we get approximately 2.9. Then $\frac{\mathcal{E}_b}{N_0} = 3.5(2.9)^2 = 29.4$.

For $\Pr$ bit error $= 10^{-6}$, we need the Q^{-1} of argument 1.7×10^{-6}. From the Q-inverse figure, we get approximately 4.65. Then $\frac{\mathcal{E}_b}{N_0} = 3.5(4.65)^2 = 75.7$.

We could complete this problem by using these two different values of $\frac{\mathcal{E}_b}{N_0}$ in the link budget part of the solution given above for DPSK, but this is left out of these notes.

20 Implementation Costs

20.1 Power Amplifiers and Constant Envelope

Recall that the PAM modulation format uses waveforms that are orthogonal at different time periods. For time period k, we can write $\phi_0(t - kT_s) = p(t - kT_s)\cos(2\pi f_c t)$, for a pulse shape $p(t)$ and a carrier frequency f_c. The total transmitted signal is a linear combination of symbols sent at different symbol periods:

$$s(t) = \sum_k a_0^{(k)} p(t - kT_s)\cos(2\pi f_c t) = \cos(2\pi f_c t)\sum_k a_0^{(k)} p(t - kT_s) \quad (43)$$

where $a_0^{(k)}$ is the amplitude of ϕ_0 during symbol k. This cosine term itself has constant power (power itself for a wave is a short time-average of the squared value). So the power at time t is proportional to the square of the $\sum_k a^{(k)} p(t - kT_s)$ term. Consider Figure 23, which shows $s(t)$ for binary PAM when three symbols are sent in a row ($k = 0, 1, 2$). The "envelope" of the signal is the dashed line skirting the maximum of the signal. The "power" is the envelope squared. Note that the power is very close to zero between symbols, a value much smaller than the maximum power.

A metric to quantify the envelope changes is the peak-to-average power ratio (PAPR). The PAPR is defined as

$$PAPR = \frac{\text{Maximum Power}}{\text{Average Power}}$$

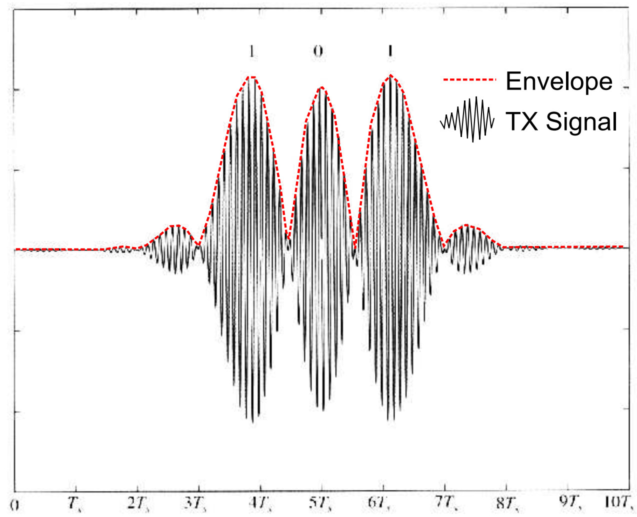


Figure 6.19 Raised cosine filtered ($\alpha = 0.5$) pulses corresponding to 1, 0, 1 data stream for a BPSK signal. Notice that the decision points (at $4T_s$, $5T_s$, $6T_s$) do not always correspond to the maximum values of the RF waveform.

Figure 23: Transmitted signal and signal envelope for an example BPSK signal. Modified from Rappaport Figure 6.19.

where "Envelope" is the power in transmitted signal $s(t)$. Of course, the maximum is always greater than the average, so the PAPR ≥ 1. But when the max power is not very large compared to the average power, then the PAPR is close to 1, and we say the signal has a constant envelope.

A power amplifier at the transmitter to take the transmitted signal and amplify it to the desired output power (*e.g.*, 500 mW for cellular). Power amplifiers waste some energy, and are rated by their efficiency. Amplifiers that can deal with *any* signal you send to it are about 50% efficient (Class A amplifiers). However, the most efficient (Class C) amplifiers (90% efficient) require a *constant envelope* modulation. Designers thus tend to choose constant envelope modulations for battery-powered transmitters.

In the next section, we compare QPSK (which is not constant envelope) with O-QPSK (which is considered constant envelope).

20.1.1 Offset QPSK

For QPSK, we used orthogonal waveforms:

$$\begin{aligned} \phi_0(t) &= \cos(2\pi f_c t) p(t) \\ \phi_1(t) &= \sin(2\pi f_c t) p(t) \end{aligned}$$

Essentially, there are two 90 degree offset sinusoids multiplying the same pulse shape. After T_s, these sinusoids will multiply $p(t - T_s)$. The effect is similar to Figure 23, but with both cosine and sine terms lined up. Following the notation of (43),

$$s(t) = \cos(2\pi f_c t) \sum_k a_0^{(k)} p(t - kT_s) + \sin(2\pi f_c t) \sum_k a_1^{(k)} p(t - kT_s)$$

We can look at the complex baseband signal as

$$\tilde{s}(t) = \sum_k a_0^{(k)} p(t - kT_s) + i \sum_k a_1^{(k)} p(t - kT_s)$$

The magnitude $|\tilde{s}(t)|$ is the envelope of signal $s(t)$ and $|\tilde{s}(t)|^2$ is its power.

Whenever a_0 switches sign between symbol $k-1$ and symbol k, the linear combination of the pulses will go through zero. When this happens for both a_0 and a_1, the envelope and power go through zero. This means that a linear amplifier (class A) is needed. See Figure 24(b) to see this graphically.

For offset QPSK (OQPSK), we delay the quadrature $T_s/2$ with respect to the in-phase. Now, our orthogonal waveforms are:

$$\begin{aligned}\phi_0(t) &= \cos(2\pi f_c t) p(t) \\ \phi_1(t) &= \sin(2\pi f_c t) p\left(t - \frac{T_s}{2}\right)\end{aligned}$$

Now, the complex baseband signal is,

$$\tilde{s}(t) = \sum_k a_0^{(k)} p(t - kT_s) + i \sum_k a_1^{(k)} p(t - [k + 1/2]T_s)$$

Now, even when both a_0 and a_1 switch signs, we don't see a zero envelope. When a_0 switches sign, the real part will go through zero, but the imaginary part is at a maximum of the current $p(t)$ pulse and is not changing. See Figure 24(d). The imaginary part doesn't change until $T_s/2$ later.

At the receiver, we just need to delay the sampling on the quadrature half of a sample period with respect to the in-phase signal. The new transmitted signal takes the same bandwidth and average power, and has the same $\frac{\mathcal{E}_b}{N_0}$ vs. probability of bit error performance. However, the envelope $|s(t)|$ is largely constant. See Figure 24 for a comparison of QPSK and OQPSK.

There are some disadvantages of OQPSK: Because the amplitudes of both ϕ_0 and ϕ_1 do not change at the same time, timing synchronization at the receiver can be (somewhat) more difficult [20]. In addition, it is difficult to implement differential decoding with OQPSK.

20.1.2 Other Modulations

What are other modulations that do or do not have a constant envelope?

1. $\pi/4$ QPSK: see the Rapppaport book – alternate between the standard QPSK constellation in Figure **??** (a), and a $\pi/4$ radians rotated version of the same constellation for the next symbol period. The envelope will not go as close to zero as QPSK, but is still not considered constant envelope. Used in IS-54.

2. FSK: FSK modulations DO have a constant envelope.

3. M-QAM: Square M-QAM modulations DO NOT have a constant envelope, and are generally even worse than QPSK in that they need a linear amplifier.

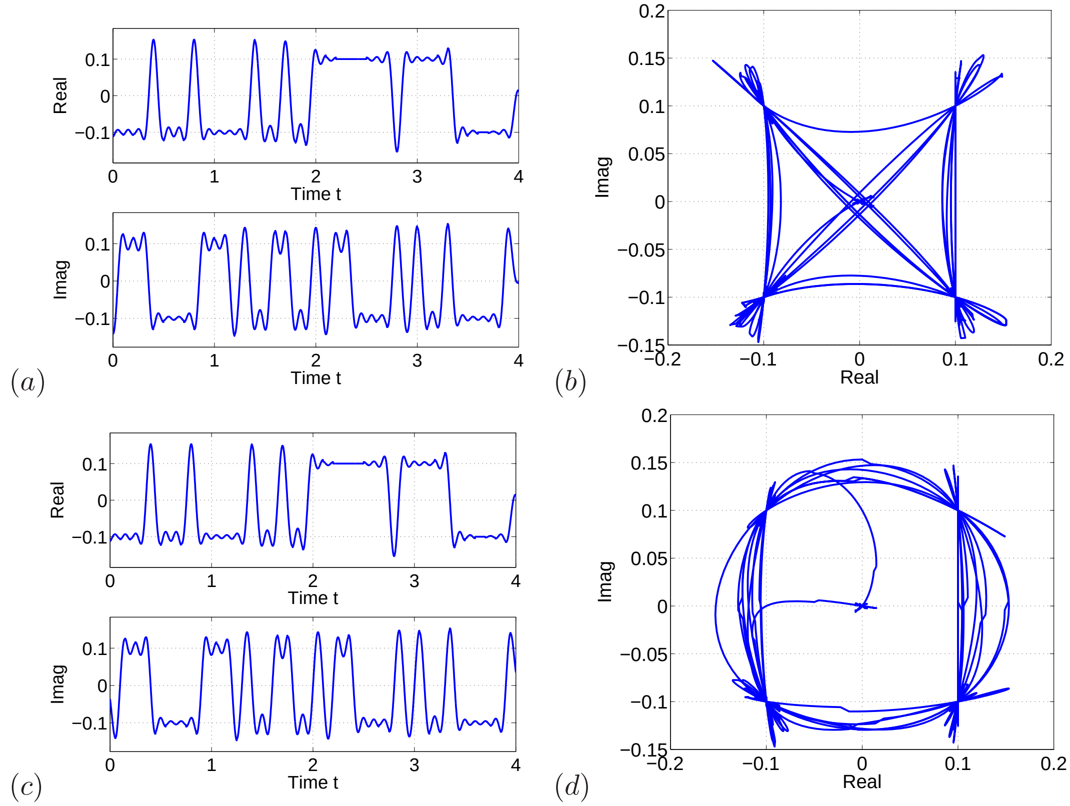


Figure 24: Matlab simulation of (a-b) QPSK and (c-d) O-QPSK, showing the (d) largely constant envelope of OQPSK, compared to (b) that for QPSK.

20.2 Synchronization

There are two main kinds of synchronization which we require in a receiver:

1. Phase: To multiply by the correct sine or cosine, we need to know the phase of the incoming signal. Typically, we track this with a phase-locked loop (PLL). PLLs can be difficult to implement.

2. Symbol Timing: One must synchronize to the incoming signal to know when to sample the symbol value. Errors in sampling time introduce inter-symbol interference (ISI). Using a lower α makes a receiver more sensitive to symbol synch errors.

In this section, we present two different ways to simplify phase synchronization.

20.2.1 Energy Detection of FSK

FSK reception can be done via a matched filter receiver, in which there is a multiply and integrate with each orthogonal waveform (in this case, sinusoids at different frequencies). This is shown in Figure 25. This is called "coherent reception" of FSK. For binary FSK, there must be 2 PLLs, which have more difficulty, because the received signal contains each frequency sinusoid only half of the time. FSK is popular in inexpensive radios; so two PLLs might be significant enough additional cost to preclude their use.

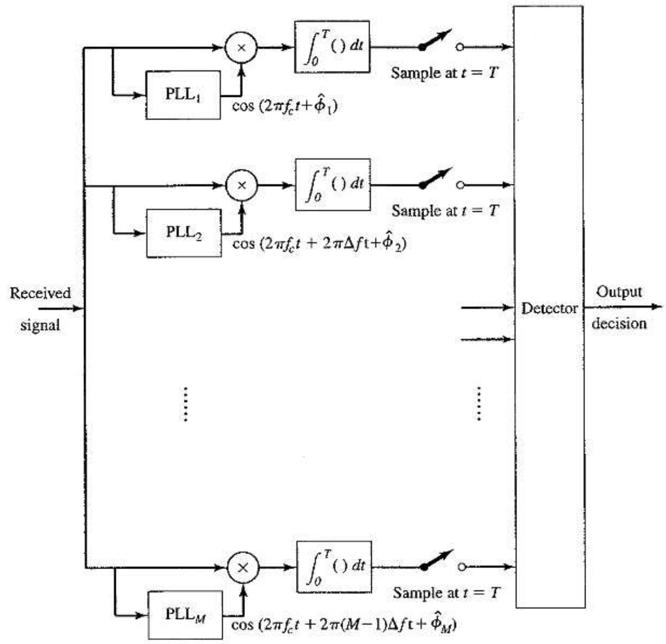


Figure 25: Phase-coherent demodulation of M-ary FSK signals, from Proakis & Salehi [22], Figure 7.46.

A non-coherent binary FSK receiver avoids PLLs altogether by simply computing the energy $\mathcal{E}_{f_k}$ at each frequency f_k, for $k = 0, 1$. It decides the bit by seeing which energy, $\mathcal{E}_{f_0}$ or $\mathcal{E}_{f_1}$, is higher.

To calculate the energy, it still must multiply and integrate with a cosine and sine at each frequency. That's because if it just multiplies with the cosine, for example, and the phase of the received signal makes it a sine, then the multiply and integrate will result in zero (sine and cosine are orthogonal, after all). But if you find the energy in the cosine, x_k^I, and the energy in the cosine, x_k^Q, then the total energy at frequency k is

$$\mathcal{E}_{f_k} = (x_k^I)^2 + (x_k^Q)^2$$

is calculated for each frequency f_k, $k = 0, 1$.

The downside of non-coherent reception of FSK is that the probability of bit error increases. The optimal, coherent FSK receiver has performance $\mathrm{Q}\left(\sqrt{\frac{\mathcal{E}_b}{N_0}}\right)$. The non-coherent binary FSK receiver has probability of bit error of $\frac{1}{2}\exp\left[-\frac{\mathcal{E}_b}{2N_0}\right]$. There is about a 1 dB difference between the two – to achieve the same probability of bit error, the non-coherent receiver must have about 1 dB more received power.

20.2.2 Differential PSK

In BPSK, the bit is encoded by the sign of the sinusoid. But the sign is just a 180 degree phase shift in the sinusoid. So when we receive the signal and lock our PLL to the signal, we'll see the phase changing 180 degrees, but we won't know which one is positive and which one is negative. One thing that we can do is to encode the bit using the change of phase. That is, change the phase

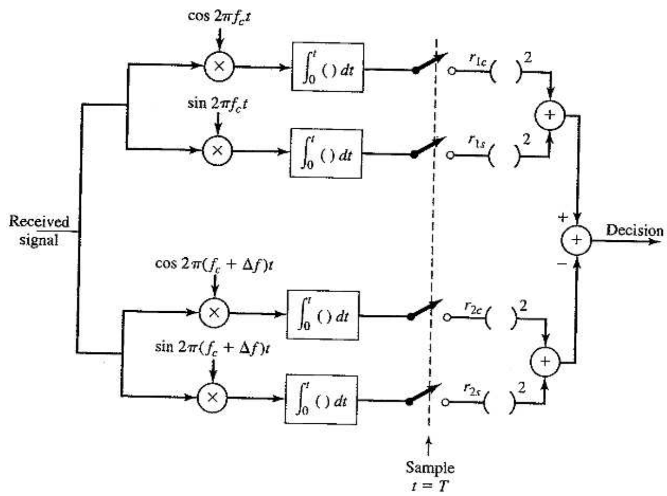


Figure 26: Demodulation and square-law detection of binary FSK signals, from Proakis & Salehi [22], Figure 7.49.

180 degrees to indicate a "1" bit, and keep the phase the same to indicate a "0" bit. This is called differential encoding. In the RX, we'll keep track of the phase of the last symbol and compare it to the phase of the current symbol. When that change is closer to 180 degrees than 0 degrees we'll decide that a "1" bit was sent, or alternatively, if that change is closer to 0 degrees than 180 degrees, we'll decide that a "1" bit was sent. This is called differential phase shift keying (DPSK) instead of BPSK. The downside of BPSK is that once we're off by 180 degrees, every bit will be decided in error. Instead, DPSK never becomes 100% wrong, it degrades gracefully with phase (and thus frequency) offset.

The downside of DPSK is that the probability of bit error increases. The optimal BPSK receiver has performance $Q\left(\sqrt{\frac{2\mathcal{E}_b}{N_0}}\right)$. The DPSK receiver has probability of bit error of $\frac{1}{2}\exp\left[-\frac{\mathcal{E}_b}{N_0}\right]$. There is about a 1 dB difference between the two – to achieve the same probability of bit error, the DPSK receiver must have about 1 dB more received power.

Lecture 14

Today: (1) OFDM

21 Multi-carrier Modulation

The basic idea for multi-carrier modulation is to divide the total available bandwidth B into N different subchannels. In each subchannel, we'll use some particular modulation, and send all signals on all subchannels simultaneously. Considering orthogonal waveforms, we'll use both the sin and the cosine function at many different frequencies.

The benefit is that with narrower bandwidth subchannels, we can achieve flat fading on each subchannel. Remember, if the bandwidth is low enough, we don't need an equalizer, because the channel experiences flat fading. (Q: What was the rule given in the Rappaport book given T_s and

σ_τ to determine if fading is flat or frequency selective?)

So we will set N high enough so that the symbol rate in each subchannel (which is B/N wide) will be high enough to "qualify" as flat fading. How do you determine the symbol period T_N to make the signal fit in B/N bandwidth? (Here we are calling the multicarrier modulation symbol period T_N to match the notation in the Goldsmith handout). Recall the bandwidth (now B/N) is given by $\frac{1}{T_N}(1+\alpha)$ for rolloff factor α so now $T_N = \frac{N}{B}(1+\alpha)$. The Goldsmith book denotes T_s as the symbol period that a single carrier system would have used for the same total bandwidth B, that is, $T_s = (1+\alpha)/B$.

If we use $T_N \geq 10\sigma_\tau$ then now we can choose N as

$$N \geq 10\sigma_\tau B \frac{1}{1+\alpha}$$

Now, bit errors are not an 'all or nothing' game. In frequency multiplexing, there are N parallel bitstreams, each of rate $\log_2(M)/T_s$, where $\log_2(M)/T_s$ is the bit rate on each subchannel. As a first order approximation, a subchannel either experiences a high SNR and makes no errors; or is in a severe fade, has a very low SNR, and experiences a BER of 0.5 (the worst bit error rate!). If γ is the probability that a subchannel experiences a severe fade, the overall probability of error will be 0.5γ.

Frequency multiplexing is typically combined with channel coding designed to correct a small percentage of bit errors.

Note that the handout uses T_N for the symbol period in multicarrier modulation systems. They reserve T_s to be the symbol period of the single-carrier system that had the bandwidth B, that is, T_N/N. Worse, the OFDM symbol period is called T, which is slightly more than T_N, as we describe below. Please make a note.

21.1 OFDM

In the previous section we made no restriction on the frequencies of the subcarriers. Well, we know that to have orthogonal sinusoids at different frequencies, we need a particular condition on Δf between the two frequencies.

21.1.1 Orthogonal Waveforms

In FSK, we use a single basis function at each of different frequencies. In QAM, we use two basis functions at the same frequency. OFDM is the combination:

$$\begin{aligned}
\phi_{0,I}(t) &= p(t)\cos(2\pi f_c t) \\
\phi_{0,Q}(t) &= p(t)\sin(2\pi f_c t) \\
\phi_{1,I}(t) &= p(t)\cos(2\pi f_c t + 2\pi \Delta f t) \\
\phi_{1,Q}(t) &= p(t)\sin(2\pi f_c t + 2\pi \Delta f t) \\
&\vdots \\
\phi_{N-1,I}(t) &= p(t)\cos(2\pi f_c t + 2\pi (N-1) \Delta f t) \\
\phi_{N-1,Q}(t) &= p(t)\sin(2\pi f_c t + 2\pi (N-1) \Delta f t)
\end{aligned}$$

where $\Delta f = \frac{1}{T_N}$, and $p(t)$ is the pulse shape (typically a SRRC pulse). These are all orthogonal functions! Note we have $2N$ basis functions here. But, we can transmit much more information than possible in M-ary FSK for $M = 2N$, because we use an arbitrary linear combination of the $\phi()$ rather than sending only one orthogonal waveform at a time. In other words, rather than transmitting on one of the M carriers at a given time (like FSK) we transmit information in parallel on all N channels simultaneously. An example state space diagram for $N = 3$ and PAM on each channel is shown in Figure 27.

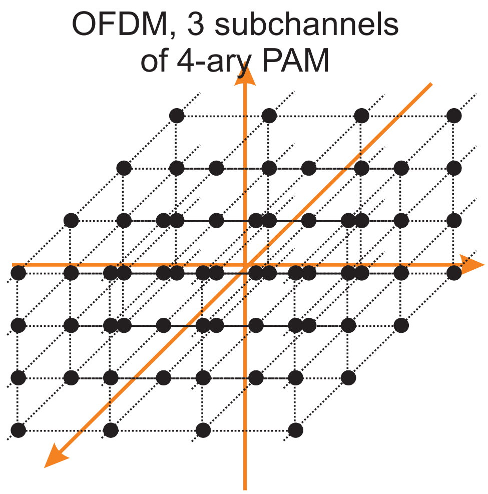


Figure 27: Signal space diagram for $N = 3$ subchannel OFDM with 4-PAM on each channel.

Note that the subchannels overlap. They are separated by $1/T_N$, but the bandwidth of any subcarrier is $(1+\alpha)/T_N$, because of the square root raised cosine (SRRC) pulse. Nevertheless, they are orthogonal so they can be separated at the receiver.

21.1.2 Fourier Transform Implementation

The signal on subchannel k might be represented as:

$$x_k(t) = p(t)\left[a_{k,I}(t)\cos(2\pi f_k t) + a_{k,Q}(t)\sin(2\pi f_k t)\right]$$

The complex baseband signal of the sum of all N subchannel signals might then be represented as

$$\begin{aligned} x_l(t) &= p(t)\mathbb{R}\left\{\sum_{k=1}^{N}(a_{k,I}(t) + ja_{k,Q}(t))e^{j2\pi k\Delta f t}\right\} \\ x_l(t) &= p(t)\mathbb{R}\left\{\sum_{k=1}^{N} A_k(t)e^{j2\pi k\Delta f t}\right\} \qquad (44) \end{aligned}$$

where $A_k(t) = a_{k,I}(t) + ja_{k,Q}(t)$. Does this look like an inverse discrete Fourier transform? If yes, than you can see why it might be possible to use an IFFT and FFT to generate the complex baseband signal.

FFT implementation: There is a particular implementation of the transmitter and receiver that use FFT/IFFT operations. This avoids having N independent transmitter chains and receiver chains. The FFT implementation (and the speed and ease of implementation of the FFT in hardware) is why OFDM is currently so popular.

21.1.3 Cyclic Prefix

However, because the FFT implementation assumes that the data is periodic, and it is not, we must add a cyclic prefix at the start of each block of N symbols. The cyclic prefix is length μT_s and is an overhead (doesn't contain data) and thus increases the symbol period. Thus for OFDM, we calculate the OFDM symbol period $T = T_N + \mu T_s$. This reduces our overall data rate because the symbol period is longer, and no new data is sent during this μT_s cyclic prefix period. The actual data rate will be the total bit rate multiplied by $\frac{N}{N+\mu}$. We determine μ to be longer than the expected delay spread σ_τ. So, $\mu \geq \sigma_\tau / T_s$.

21.1.4 Problems with OFDM

Note that one of the big problems with OFDM is that it requires linear amplifiers, because of its high peak-to-average ratio (PAR). The Goldsmith handout does a great job of proving that the peak to average ratio is approximately N, the number of subcarriers, so for reasonable N, you would need a linear amplifier.

OFDM is also sensitive to timing and frequency offset errors. In a mobile radio system corrupted by Doppler spread, each carrier is spread into the next one, causing the subcarriers to no longer be orthogonal to each other.

21.1.5 Examples

Example: 802.11a
IEEE 802.11a uses OFDM with 52 subcarriers. Four of the subcarriers are reserved for pilot tones, so effectively 48 subcarriers are used for data. Each data subcarrier can be modulated in different ways. One example is to use 16 square QAM on each subcarrier (which is 4 bits per symbol per subcarrier). The symbol rate in 802.11a is 250k/sec. Thus the bit rate is

$$250 \times 10^3 \frac{\text{OFDM symbols}}{\text{sec}} 48 \frac{\text{subcarriers}}{\text{OFDM symbol}} 4 \frac{\text{coded bits}}{\text{subcarrier}} = 48 \frac{\text{Mb (coded)}}{\text{sec}}$$

Example: Outdoor OFDM System Design
Design an OFDM system to operate in a 20 MHz bandwidth at 1300 MHz in outdoor environments with large delay spreads of $\sigma_\tau = 4\mu$s. Determine: (1) Does we need multicarrier modulation, and if so, why? (2) The number of subcarriers. (3) The overhead of the cyclic prefix. (4) The symbol period. (5) The data rate of the system assuming $M = 4$ and coding with rate 1/2 is used to correct bit errors.
Solution:

1. A single carrier system would have symbol period $T_s = 1/20$MHz or $T_s = 50$ ns. The delay spread, 4 microseconds, is huge compared to 0.05 microseconds, so we definitely need multicarrier modulation to mitigate the intersymbol interference.

2. We want $B_N = B/N \ll 0.1/\sigma_\tau$, that is, $\ll 25$kHz. Note we also want N to be a power of 2 for ease of implementation via the FFT. $B = 20$ MHz divided by 25 kHz is 800, so we'd choose $N = 1024$, which results in a subcarrier bandwidth B_N of 20 MHz /1024 = 19.5 kHz.

3. Calculating, μ, the overhead of the cyclic prefix, is $\mu = \sigma_\tau / T_s = 80$.

4. The symbol period for OFDM is or $T_N + \mu T_s = N/B + \mu/B = (N + \mu)/B = (1024 + 80)50$ ns $= 55.2\mu$s.

5. Each carrier generates $\log_2 4 = 2$ bits per symbol period. There are 1024 carriers, for a total of 2048 bits per symbol. Since one symbol equals 55.2×10^{-6}s, the raw data rate is $2048/(55.2 \times 10^{-6}) = 37.1$ Mbps. Half of those bits (for rate 1/2 coding) are code bits, so the real data rate is $111/2 = 18.5$ Mbps.

Lecture 15

Today: (1) FEC Coding (2) Error Detection

22 Forward Error Correction Coding

The material for this section comes largely from Jeff Frolik's MUSE channel coding video, available at:

- `http://www.uvm.edu/~muse/CTA.html`

Def'n: *Forward error correction coding or channel coding*
Adding redundancy to our data at the transmitter with the purpose of detecting and correcting errors at the receiver.

The transmitter takes in data bits and puts out coded bits. Our notation is that for each k data bits input to the FEC operator, the FEC operation will produce $n > k$ coded bits out.

22.1 Block vs. Convolutional Coding

Def'n: (k, n) *Block Code*
A (k, n) block code inputs k-bits which are accumulated (via serial-to-parallel conversion) in a k-length vector $\mathbf{d}$. Block encoding multiplies $\mathbf{d}$ by a $k \times n$ generator matrix, G, to output a n-length bit vector $\mathbf{c}$. Block decoding then multiplies the received vector $\mathbf{r}$ by the syndrome matrix S to determine if any errors occurred and determine which (if any) bits were in error out of the n sent.

The syndrome is just a rearrangement of the transpose of the generator matrix, as shown by example below.

In contrast, a convolutional code is a "running" code. For encoding, bits are input into what is effectively a binary filter, the output bits are dependent on the current and past bits.

Compare the advantages and disadvantages:

- Block code: Advantages: Better for data that is not coming in large streams (bursty data sources, <1000 bits), *e.g.*, wireless sensor networks. Pretty simple computation. Not the best one can do in terms of improving energy efficiency / removing errors. Block codes are used in CDs, DVDs, and disk drives.

- Convolutional codes: Advantages: Best for very large data streams. More energy efficient than block codes when you have large streams of data. Convolutional codes are used in: deep space communication (Voyager program), satellite and terrestrial digital video broadcasting. Disadvantages: Computational complexity increases exponentially in the length of the code. Andrew Viterbi (founder of Qualcomm) is credited with the optimal decoder, called the Viterbi algorithm.

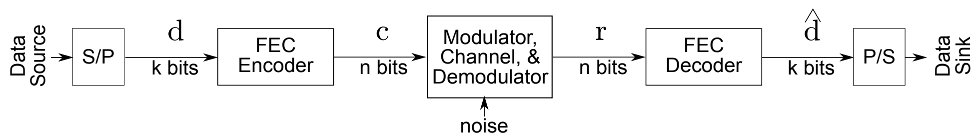


Figure 28: Block diagram of a communication system emphasizing the forward error correction (FEC) "block coding" encoder and decoder.

22.2 Block Code Implementation

Let the input be denoted $\mathbf{d}$, a k-bit vector. Let the output be $\mathbf{c}$, a n-bit vector. Let G be the generator matrix. Then

$$\mathbf{c} = \mathbf{d}G$$

Thus the G matrix has size $k \times n$. This operation is done modulo-2. That is, multiply all of the pairs, sum them, and then take the mod 2 of the result. That is, if the sum of the products is even, the answer is 0, if the sum is odd, the answer is 1.

Def'n: *Systematic*
The first k bits of the n bits output, are the same as the k bits in $\mathbf{d}$.

Example: (6, 3) systematic block code which can correct one bit error
Let G be given by:

$$G = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

Encode the data bits $\mathbf{d} = [1, 1, 1]$.
Solution: $\mathbf{c} = [1, 1, 1, 0, 0, 0]$

Example: Reception
You receive $\mathbf{r} = [1, 1, 1, 0, 0, 1]$, that is, what you received has an error in the last bit compared to $\mathbf{c}$ (the coded bits that were sent through the channel). What was is the block decoder's estimate of the transmitted data?
Solution: At the receiver, multiply by the syndrome

$$S = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Compute: $\mathbf{r}S = [0, 0, 1]$.

Look at all of the rows of the syndrome. The row number of the syndrome S that matches the output $\mathbf{r}S$, is the same as the number of the bit that was in error. If $\mathbf{r}S$ is all zeros, that indicates that there were no errors. Since the sixth bit was in error, instead of $[1, 1, 1, 0, 0, 1]$, we know the correct coded bits were $[1, 1, 1, 0, 0, 0]$.

Finally, because it is a systematic code, we know the first three bits are the data bits. The receiver will just drop the last three bits.

Example: (7, 4) Block Code

1. Encode $\mathbf{d} = [0, 1, 1, 0]$ with the $(7, 4)$ block code with generator,

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

2. If $\mathbf{r} = [0, 1, 1, 0, 1, 1, 1]$ is received, and S is given as below, what would the receiver determine to be the demodulated bits?

$$S = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

3. If $\mathbf{r} = [0, 0, 0, 1, 1, 1, 0]$ is received, what would the receiver determine to be the demodulated bits?

4. If $\mathbf{r} = [1, 0, 0, 1, 1, 1, 0]$ is received, what would the receiver determine to be the demodulated bits?

5. If $\mathbf{r} = [1, 1, 0, 1, 1, 1, 0]$ is received, what would the receiver determine to be the demodulated bits?

Solution: (1) I get $\mathbf{c} = [0, 1, 1, 0, 1, 1, 0]$. (2) Then, multiplying $[0, 1, 1, 0, 1, 1, 1]S$, I get $[0, 0, 1]$, which is the same as the 7th row, which says that the last row was incorrectly received, and so the 7th bit was incorrect. Thus the correct four bits sent were $[0, 1, 1, 0]$. (3) I get $\mathbf{r}S = [0, 0, 0]$ which means no bits were received in error, so the four data bits sent were $[0, 0, 0, 1]$. (4) I get $\mathbf{r}S = [1, 1, 1]$ which means that the first bit was received in error, so the four data bits sent were $[0, 0, 0, 1]$. (5) I get $\mathbf{r}S = [1, 0, 0]$ which means that the receiver thinks the fifth bit was received in error, so the receiver would guess the four data bits were $[1, 1, 0, 1]$.

22.3 Performance and Costs

Using a (7,4) block code, we can correct a single error in the 7 bits. But we need to increase the number of bits to send, which then requires more energy. So when using channel coding, we reduce the transmit power when comparing uncoded and coded transmissions. Still, the probability of bit error goes down for equal $\frac{\mathcal{E}_b}{N_0}$(dB). Equivalently, we can achieve the same bit error rate at 1 dB lower $\frac{\mathcal{E}_b}{N_0}$. This value, 1 dB, is the *coding gain*. In our link budgets, coding goes in the Gains column, added in with the antenna gains.

However, coding requires sending additional bits. So, in a coded system, there is always a ratio of data bits to coded bits, r, called the *code rate*. In the (7,4) block code it is $r =$ 4 data bits /7 coded bits. For a fixed bandwidth, this reduces the achievable data rate by r. For a fixed data rate, it increases the bandwidth by a factor of $1/r$.

23 Error Detection via CRC

Besides trying to correct the data, we can also simply try to detect the error. (One or both error correction and error detection can be done.) If the check fails, we might request the data again. For error detection, we have the following packet structure:

1. Header: Information about who the data is meant for, what purpose data is being sent, etc.
2. Payload: Data that we need to communicate. May be of various lengths.
3. Footer: Where the CRC is included. The CRC is always the same length, regardless of the payload length.

The TX calculates the cyclic redundancy check (CRC) from the header and payload and puts it in the footer. The RX then, once it has all of the header and footer, calculates the CRC based on the received header and footer, and compares it to the one it received. If they match, the RX decides the data was received correctly.

23.1 Generation of the CRC

A y-bit CRC is described by a y-order polynomial. The y refers to the maximum exponent in the CRC polynomial. For example, a 1st order CRC is $c(x) = x + 1$. The data we want to send, $\mathbf{d}$, can also be described as a polynomial. For example $\mathbf{d} = [0, 1, 0, 1]$ corresponds to the polynomial $d(x) = 0x^3 + 1x^2 + 0x + 1x^0 = x^2 + 1$. To determine the polynomial, just multiply the (row) bit vector with the column vector $[x^3, x^2, x^1, 1]^T$, or, in general, for n-length data string, multiply by $[x^{n-1}, x^n, \ldots, 1]^T$.

To find the CRC bit, we divide the two polynomials, $d(x) \div c(x)$ (modulo-2) and the CRC bit is the remainder. So in fact the CRC is $d(x) \mod c(x)$. In this example, the CRC is zero.

Then, what we send through the channel is the desired bits appended by the CRC bit, in this example, $[0, 1, 0, 1, 0]$.

Example: CRC for 5 bit payload
Let $\mathbf{d} = [1, 1, 0, 1, 1]$. What is the CRC bit for $c(x) = x + 1$?
Solution: When dividing $d(x) = x^4 + x^3 + x + 1$ by $c(x) = x + 1$, you get a remainder of 0. So 0 should be appended.

For the 1-bit CRC check, there is also a much easier way – to determine the 1-bit CRC, just check if the data (including CRC bit) has an even number of 1s. The one-bit CRC is also called the parity check. Next, we consider a 4th order (four bit) CRC.

Example: CRC calculation with a 4-bit CRC
Let $c(x) = x^4 + x + 1$. Let $d(x) = x^6 + x^3 + x^2 + x + 1$. Because the $c(x)$ polynomial is 4th order, there are four CRC bits. Calculate the CRC bits in this case.
Solution: The solution is $[0, 0, 0, 1]$. See my written solutions on the last page.

23.2 Performance and Costs

Using a CRC, we add a few additional bits to the packet, fewer than a FEC code would add. This allows us to detect an error, but not correct it. However, it is often important to ensure that the data received did not have any errors, beyond any reasonable doubt. For example, for cell phone audio, it may be okay to make a bit error, and have it cause some temporary noise in the audio signal; but when transferring a file, it may be critical not to have the file show up with incorrect data, and not know that it was corrupted. A y-bit CRC will have a missed detection rate around 2^-y.

Most systems use both an FEC and CRC to increase the energy efficiency (with the FEC) and to make sure that received data is correct (with the CRC).

Lecture 16

Today: (1) Spread Spectrum

24 Spread Spectrum

"Spread spectrum" is the use of a much wider bandwidth than necessary in a radio communications system, in order to achieve other objectives.

For example, frequency modulation (FM) historically was considered to be widely spread in spectrum (broadcast FM radio uses 200 kHz to send audio, which is inherently a 3-10 kHz bandwidth signal). However, FM was used because it could mitigate the effects of multipath fading, and could achieve better signal quality at low SNR.

The two major types of spread spectrum modulation are:

1. Frequency Hopping Spread Spectrum (FH-SS)
2. Direct-Sequence Spread Spectrum (DS-SS)

Both types of SS use *pseudo-random codes*, periodic sequences of zeros and ones. These are not actually random – a computer generates them using (deterministic) binary feedback logic. But, an average person would look at the sequence and judge them to be random. So they are called pseudo-random.

24.1 FH-SS

FH-SS pseudo-randomly changes center frequency each "hopping period", T_h. Bluetooth is a FH-SS system, which achieves a (coded) bit rate of 1 Mbps (potentially up to 3 Mbps), but uses 80 MHz of spectrum, in 79 different center frequencies, with a hopping period $T_h = 1/1600$ s/hop. While at each center frequency, the modulation is similar to things we have been talking about, *e.g.*, FSK, DPSK, QPSK, etc.

Three benefits of FH-SS are:

1. *Interference avoidance*: There may be significant interference at a few of the center frequencies. But even if we totally lose all bits during hops to those few frequencies, we will be able to recover using the bits received during successful (non-interfered) hops. We also avoid being an interferer to someone else's signal for too long.

2. *Multiple Access*: Two devices can occupy the same spectrum and operate without coordinating medium access at all. Their transmissions will "collide" some small fraction of the time, but (hopefully) not often enough to cause failure.

3. *Stealth*: There is an advantage to switching randomly among frequencies when an eavesdropper doesn't know your hopping pattern – they will not be able to easily follow your signal. This was the original reason for the first use of FH-SS (it was discovered by actor and inventor Hedy Lamarr, in 1940 [18] as a means for covert military communication. In the patent, she describes FH-SS by analogy to piano playing).

24.2 DS-SS

We will see that DS-SS has the same three advantages as FH-SS, but for different reasons.

Direct-sequence spread spectrum simply uses a pulse shape that has a wide bandwidth. This pulse shape is known as a pseudo-noise signal, which is, essentially, itself a BPSK-modulated signal, but which "pseudo-random" data that is known to both transmitter and receiver.

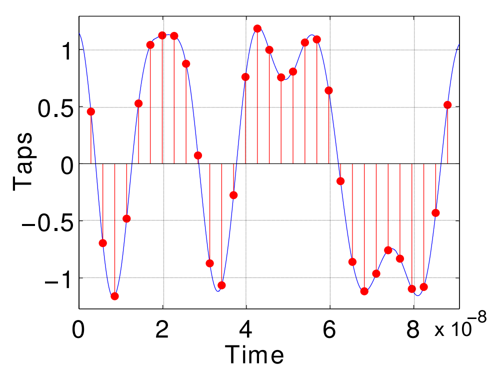


Figure 29: Pseudo-noise ("Barker") code used in 802.11b DS-SS modulation.

Figure 29 shows the PN signal used in 802.11b. Can you draw on top of Figure 29 what the pulse shape would be if it were not DS-SS, and instead, simply modulated BPSK, with a SRRC pulse shape? You can see that the bandwidth of the DS-SS signal will be very large compared to that of the BPSK signal.

We can describe the modulated signal, sent by user k, as (from Rappaport):

$$s_k(t) = a_k(t)p_k(t)\cos(2\pi f_c t)$$

where $a_k(t)$ is the bit signal, +1 or -1, which indicates the data bits to be sent. For example, the Barker code uses bits $[+1, 1, +1, +1, 1, +1, +1, +1, 1, 1, 1]$. As before, we have a pulse shape $p_k(t)$, however, in this case, $p_k(t)$ is not simply a SRRC pulse. It is a high bandwidth pseudo-noise signal. Essentially, $p_k(t)$ is a BPSK-modulated signal itself. The "bits" of the PN signal are called "chips" to distinguish them from bits. We denote the number of chips in $p_k(t)$ as PG, it is the number of chips per bit. It is called PG because it is also called the processing gain. The period of each chip is denoted T_c, so

$$PG = \frac{T_s}{T_c} = \frac{R_c}{R_s}$$

Where $R_c = 1/T_c$ is the chip rate. For example, 802.11b has a chip rate of 11 M (chips per second) and a symbol rate of 1 M (symbols per second).

Note that the chips do not necessarily need to be the same each time. In IS-95 (also called code-division multiple access (CDMA) by Qualcomm), the "short code" has $PG = 2^{15} = 32768$. There are not 32768 chips per bit, though – there are 64 chips per bit. The PN code generator just provides a source of chips that are taken 64 at a time to produce the pulse shape for each data symbol. In the IS-95 case, $PG = 64$.

Incidentally, also in IS-95 is a long code that has length $2^{42} - 1$. The long code is different for every mobile. The output chips are xor-ed with the long code to make the signal hard to eavesdrop and makes it unique to the particular mobile.

The bandwidth of a DS-SS symbol, when chips have the SRRC shape, is

$$B = (1 + \alpha)R_c$$

Which is then PG times the bandwidth of the BPSK signal would have been as a narrowband signal.

Recall that the SNR required to demodulate a signal is given by:

$$SNR = \frac{\mathcal{E}_b}{N_0}\frac{R_b}{B}$$

So with DS-SS, the SNR is lowered by a factor of PG

$$SNR = \frac{\mathcal{E}_b}{N_0}\frac{R_b}{(1+\alpha)R_c} = \frac{\mathcal{E}_b}{N_0}\frac{1}{(1+\alpha)PG}$$

However, If you thought this signal was just a plain-old BPSK signal, *i.e.*, didn't know the PN signal, you'd need the regular SNR $\frac{\mathcal{E}_b}{N_0}\frac{1}{1+\alpha}$, which is PG times higher. This makes us understand advantage #3 of DS-SS: Stealth. Knowing the PN signal allows one to demodulate the signal with PG times less SNR than an eavesdropper could. If the PG is high enough, the signal would be extremely difficult to detect at all, but could still be used by the intended receiver.

Advantage #1: Reception of DS-SS uses the same principle as discussed earlier – the received signal is correlated with the known PN signal. What if a narrowband interference signal was also in the received signal? Well, this narrowband signal would effectively be spread, when it is multiplied by the PN signal in the receiver. In contrast, the desired signal is de-spread (becomes narrowband again) due to correlation with the PN signal. The spread interference can then be partially filtered out using a narrowband filter. See Figure 6.50 in Rappaport (page 333).

Advantage #2: Further, DS-SS can be designed for some particular benefits for multiple access. These relate to the near-orthogonality of the particular PN codes used in DS-SS. In short, some sets of PN signals are nearly orthogonal or completely orthogonal *to each other*; and some PN signals have the property that the PN signal is orthogonal *to itself* at a different time delay. This is where the term, code-division multiple access (CDMA) comes from.

First, consider sets of PN signals orthogonal to each other. One example is the set of Gold codes (used in the GPS network). Gold code signals are nearly orthogonal with each other.

Another example is the Walsh-Hadamard (WH) sequence set, used in IS-95 (CDMA). The WH-64 sequences, shown in Figure 30(c), are used in IS-95. The 64 WH signals are exactly orthogonal to each other. These signals provide a means on the downlink to send 64 simultaneous user signals, and yet have each mobile perform a correlation operation that completely zeros out the signal sent from the BS to all 63 other mobiles. When one mobile correlates with its signal, $p_k(t)$, it has zero correlation with the other 63 WH signals.

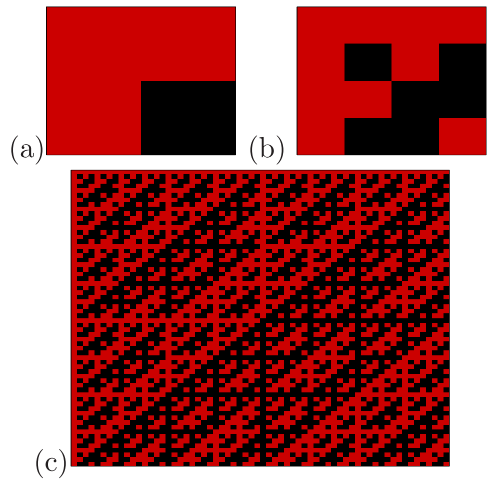


Figure 30: Walsh Hadamard sequences are in the rows of the above images. The signal is +1 during black squares, and -1 during red squares: (a) two WH-2 sequences in two rows, (b) four WH-4 sequences in four rows, and (c) sixty-four WH-64 sequences in 64 rows.

Note that WH sequences are not used on the uplink, because they are only orthogonal if time-synchronized. Mobiles aren't able to time-synchronize with each other very well.

Second, consider the autocorrelation of a PN signal (the correlation of a signal with itself at different time delays). The autocorrelation is defined as

$$R_p(\tau) = \int_0^{T_s} a_k(t) p_k(t) p_k(t-\tau) dt$$

First assume that the data signal $a_k(t) = 1$. If $\tau = 0$, the value of $R_p(\tau)$ is simply the energy in the PN signal $p_k(t)$ over a duration T_s. For τ a multiple of T_s, that is, the period of $p_k(t)$, we get the same value, *i.e.*, $R_p(nT_s) = R_p(0)$. For in between values of τ, (say, $T_c < \tau < T_s - T_c$, PN signals have a nearly (but not quite) zero autocorrelation. Generally, for these τ, $R_p(\tau) \approx -R_p(0)/PG$. Figure 32(top plot) shows an example for the Barker code used in 802.11b.

Now, if the data was modulated, the exact $-R_p(0)/PG$ is gone, but we still do have $|R_p(\tau)| \leq R_p(0)/PG$. Notice that the autocorrelation becomes strongly negative when the bit sent was a -1.

24.3 PN code generation

PN code sequences are generated by binary logic gates called "linear feedback shift registers" (LFSR). A properly designed k-state LFSR produces a $2^k - 1$ length PN code sequence. Figure 32(a) and (b) show 4 and 5 state LFSRs. Each state of $s_1, \ldots, s_k$ is either zero or one. Note that we should never have all zero states. Assume that we start out with states that are not all zero. Then, at each time, we compute the mod-2 addition specified, and the output of the adder goes

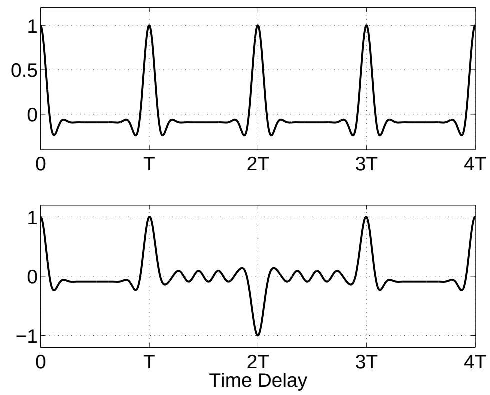


Figure 31: Autocorrelation of "Barker" code used in 802.11b DS-SS modulation, (a) with all symbol values $= +1$, that is, $a_k(t) = 1$, and (b) with data modulation.

into s_1 in the next time step. Similarly, we shift states, so that state s_i always takes the value that state s_{i-1} had in the previous time step. The output is always drawn from s_k, the rightmost state.

Example: 4-state LFSR
Calculate the first 6 outputs of the LFSR given that the initial states $[s_1, s_2, s_3, s_4]$ are [0, 1, 0, 1].
Solution:

Step	s_1	s_2	s_3	s_4	Output
0	0	1	0	1	n/a
1	1	0	1	0	1
2	1	1	0	1	0
3	0	1	1	0	1
4	0	0	1	1	0
5	1	0	0	1	1
6	0	1	0	0	1
7	0	0	1	0	0
8	0	0	0	1	0
9	1	0	0	0	1
10	1	1	0	0	0
11	1	1	1	0	0
12	1	1	1	1	0
13	0	1	1	1	1
14	1	0	1	1	1
15	0	1	0	1	1

Lecture 17

Today: (1) Medium Access Intro, (2) Packet Radio, (3) 802.11 MAC

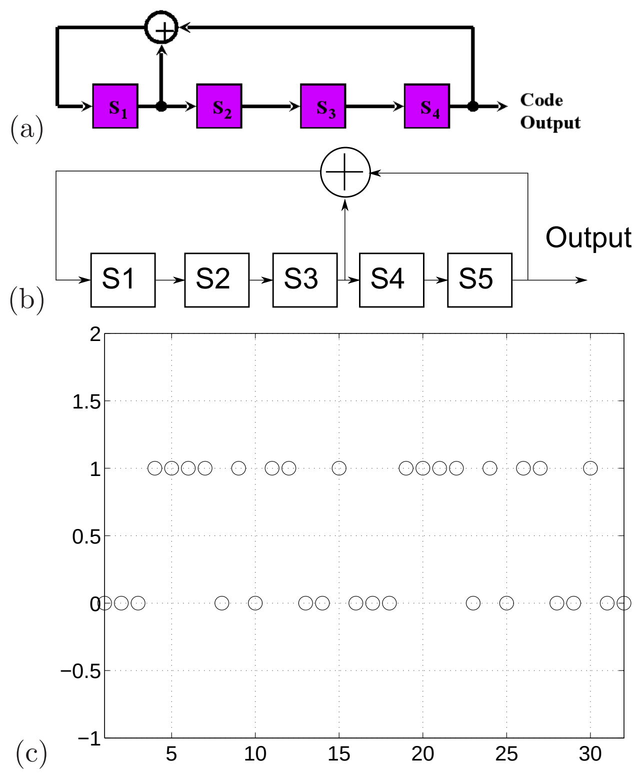


Figure 32: Linear feedback shift register (LFSR) generation of an (a) 4-stage, or (b) 5-stage, maximal-length PN code; and (c) the generated code sequence for the 4-stage LFSR.

25 Medium Access Control

In this lecture, we describe some of the ways in which wireless networks can accommodate many users at the same time, even on the same channel. This is called medium access control (MAC). It is also called multiple access control, because we are designing a protocol for multiple users to share one channel (the "medium"). We have talked about multiple access throughout this semester already

- Frequency-division multiple access (FDMA): we have used distinct frequency channels to allow many users in a cellular communication system.
- Time-division multiple access (TDMA): users of the same frequency channel are assigned different "slots" in time.
- Space-division multiple access (SDMA): because of the cellular concept, we can assign users separated in space (in different cells or sectors) the same channel and still ensure that they do not interfere with each other.
- Code-division multiple access (CDMA): We can assign users distinct code channels, that is, have them use orthogonal pulse shapes that allow them, at the receiver, to be separated from any other user's signal on a different code channel.

In this lecture, we will talk specifically about a variation of TDMA called "packet radio". Packet radio allows multiple users to use a shared channel sometimes to send a "packet" of data. Compared to cellular TDMA, we are now talking about data communications systems which only sometimes need to send data. If we were to reserve each user a regular slot (like TDMA does), most of the time this slot would go unused. By having users send data only when they need to, we hope to have a more efficient use of this shared channel.

This lecture is about **control** of medium access, *i.e.*, MAC. What can control of the channel do for the efficiency of packet radio? What are some of the problems experienced in MAC protocols?

We will in this lecture:

- Learn how aloha and slotted aloha packet radio networks work, and their capacities.
- Learn what the "hidden terminal problem" is.
- Learn how 802.11 networks schedule traffic, and their capacities.

The first topic motivates the problem with packet radio, which then motivates why it can be more efficient to employ scheduling in a MAC protocol.

1. Some control: An 802.11 (WiFi) access point also exercises control over the users that communicate with it. But it has no control over where in space the next access point is that is also using the same channel. So problems ensue.
2. No control: Some wireless network protocols do not attempt to exercise control over when users offer traffic to the channel. In these cases, when two users transmit at the same time, their signals collide and can both be lost (*i.e.*, not recoverable at the intended receiver). Actually, depending on the signal to interference ratio (SIR), a signal may be able to be received despite the fact that another interfering signal overlaps with it. Further, as we saw in the spread spectrum lecture, some modulation methods make it easier to recover the desired signal when it overlaps with other interfering signals.

26 Packet Radio

Let's first discuss the reading [1]. I assign this reading because Dr. Abramson and the University of Hawaii addressed the design of a data communication system from scratch, like you are assigned to do for the semester design project. They did this 40 years ago, but we still use the analysis that they did in wireless networking systems today, because they did such a good job of presenting the fundamental tradeoffs in the design of packet radio systems.

The problem that Dr. Abramson had was that the University of Hawaii wanted to connect its campuses on different islands to a central server (called the "menehune" in the paper), which itself was connected to the ARPANET. However using telephone line connections just was such an inefficient and expensive method to do so. Essentially, it was expensive to reserve a dedicated telephone channel for each computer terminal in the Hawaii system. It ended up being much cheaper to use radio communication, on a single channel with a higher bit rate (24 kbps!).

26.1 Aloha

This protocol is called "aloha" after the Hawaiian word. In this single channel, each terminal, when it had data, would just transmit a fixed duration (τ-length in time) "packet" of data. If two terminals happened to send packets that overlapped in time, you might have them collide and get neither packet, but that if this happened, both terminals would just retransmit later. There is a reverse channel on which the receiver acknowledged any packet from any sender that was (correctly) received. This positive acknowledgement is abbreviated as "ACK".

Let the average number of data packets per second (from all senders) be r. Again, τ is the duration of a packet. Then the average total utilization of the channel is $r\tau$. A utilization of $r\tau = 1$ would be perfect utilization. **The result of [1] is that the utilization of the channel is, at most, 18.4%**. At a higher r (and thus utilization), the number of collisions increases faster in a way that reduces the rate of successfully received packets. This maximum utilization is also referred to as the capacity of an aloha packet radio system.

Note that in this protocol, the receiver exercises very little control. It is the senders who decide when to transmit.

We also have a formula for the maximum number of active terminals (or users) sending packets to the server. This maximum is k_{max},

$$k_{max} = \frac{1}{2e\lambda\tau}$$

where e is the base of the natural logarithm, λ is the average rate at which each terminal sends packets to the server (in packets per second), and τ is the packet duration. Note that we use "active terminals" to describe terminals that send packets right now at the average rate of λ per second. There might be other "inactive terminals" which are not sending packets at all at this time.

26.2 Slotted Aloha

Abramson also studied a variation of the aloha system in which terminals were synchronized and could agree on non-overlapping "slots", that is, times during which packets could be transmitted [2]. That is, rather than transmitting at just *any* time, a terminal would wait for the next slot's start time. By having all packets transmitted starting at the slot start time, there will be fewer collisions. Abramson showed that this simple modification allows the maximum utilization to double, to 36.8%. The maximum number of active terminals becomes $k_{max} = \frac{1}{e\lambda\tau}$, also double the result for regular aloha.

The only problem is that, now, each terminal must be synchronized to a clock. This may be easy or difficult depending on the synchronization requirements.

Slotted aloha is slightly more controlled, in the sense that each terminal must be synced to the same clock, and must start transmitting only at the start of a slot.

Example: How many active users could you support?
How many active users could you support on a 1 MHz channel using a packet radio system? Take, in this example, a QPSK system with SRRC pulse shaping with $\alpha = 0.5$. For this example, use the same packet rate and duration used in the original aloha system – 704 bits per packet, and a packet generation rate of 1 per minute. Calculate both for aloha and slotted aloha.
Solution: The bit rate is 1 MHz $(\log_2 4)/(1 + 0.5) = 1.333$ MHz. For 704 bits, this would take a duration of 0.53 ms. Using aloha,

$$k_{max} = \frac{1}{2e\lambda\tau} = \frac{60}{2e(0.0021)} \approx 20.9 \times 10^3$$

For slotted aloha, the max number of users would double to 41.8×10^3.

27 CSMA-CA

27.1 Carrier Sensing

One more simple idea in packet radio is *carrier sensing*. Carrier sensing refers to a transceiver's use of its receiver to sense whether or not there is currently another packet being transmitted on the channel, *i.e.*, "if the channel is busy", before transmitting its own packet. If it detects transmission, it will wait until the transmission is completed, until starting its own transmission. This is called "carrier sense multiple access" (CSMA).

A minor modification of CSMA is called CSMA-collision avoidance, or CSMA-CA. This means that, when a CSMA transceiver detects the channel to be busy, rather than just waiting for the other transmission to be completed, it *also* waits a random time interval before trying to retransmit.

27.2 Hidden Terminal Problem

What is a problem of relying on one terminal to know when the channel is busy? Since signal power is spatially varying, one terminal may be transmitting to the server successfully, even while a second terminal at a third location is not able to hear that the channel is busy, as shown in Figure 33.

The hidden terminal problem is ubiquitous in packet radio networks. This makes carrier sensing systems less useful than you would otherwise imagine.

27.3 802.11 DCF

Our 802.11 networks operate using a basic CSMA-CA protocol which is named the "distributed control function" (DCF). They use CSMA-CA, and when a collision occurs, for collision avoidance, a terminal waits a random amount of time in a procedure called exponential backoff.

This material supplements and draws heavily from [5].

The job of a terminal's transmitter is to send its packets. When it has a packet to send, it:

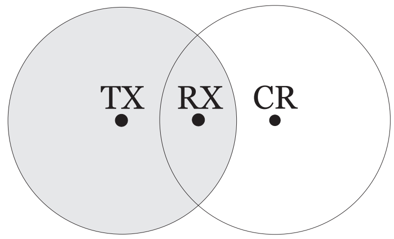


Figure 33: Terminal TX is transmitting to server RX. Terminal CR cannot hear TX's signal, so even if it uses carrier sensing, when it has data to send, it would transmit and cause interference at terminal CR.

1. Listens to the channel for a duration of DIFS (a constant time window called the distributed interframe space).
2. After the channel is sensed to be free for a duration of DIFS (by the receiver), the transmitter picks a random *backoff delay* X, which is an integer, picked as follows. X is chosen randomly and uniformly from the range 0 to $2^i CW_{min} - 1$ for some integer i called the backoff stage and some minimum window length CW_{min}. This backoff stage i is initially set to zero. So initially, when a collision occurs, the transmitter picks this random integer X between 0 and $CW_{min} - 1$.
3. If $X = 0$, the terminal immediately transmits.
4. If not, it continues sensing. After σ period of time (called the slot time) with the channel unoccupied, the receiver decrements X by one. Note that if the receiver hears another terminal transmit, it waits a period DIFS after the end of the transmission, and then waits σ and decrements X by one.
5. The terminal goes back to step 3.

After the transmitter sends its packet, its receiver checks for the ACK from the access point within ACK_Timeout (another constant period of time):

1. *If it does not receive the ACK within the specified time window*, it assumes that the packet collided and was lost. In this case, it increments the backoff stage i (up to a maximum of m), and starts the procedure to transmit the packet again.
2. *If it does receive the ACK*, it resets the backoff stage i to zero, and then starts the procedure to transmit the next packet (assuming there is another one).

Because there are so many acronyms, here is a table of the most important:

- ACK: positive acknowledgement
- CSMA/CA: carrier sense multiple access with collision avoidance
- DCF: distributed coordination function: the algorithm in 802.11 which is the main subject of this paper
- DIFS: distributed interframe space: how long a terminal measures the channel before determining it is idle

- MAC: medium access control
- SIFS: short inter-frame space: delay between end of reception and transmission of ACK

Notation:

- σ: slot time size (time needed for any terminal to detect a transmission)
- CW_{min}: minimum size of the contention window
- n: number of "contending terminals", *i.e.*, those offering packet traffic to the network.
- i: backoff stage, in the range $\{0, \ldots, m\}$
- m: maximum backoff stage ($CW_{max} = 2^m W$)

27.4 In-Class DCF Demo

This activity recreates a few milliseconds in the life of a 802.11 network running the DCF. For our exercise, let:

- $W = 4$, $m = 2$. Thus the maximum contention window is length 16.
- Let $\sigma = 1$, DIFS = 3, SIFS = 1.
- packet duration $P = 20$, ACK duration = 3.
- $n = 3$: terminals A, B, and C. All terminals can hear the access point. First assume that all terminals can hear each other, then assume that terminal A cannot hear terminal C, and vice versa.
- $ACK_{Timeout} = 24$.

Each person will "act out" a terminal TX or RX, or access point TX or RX. Other "actors" include the random number generator (the person who selects random numbers from 0 to $2^i CW_{min} - 1$), and the "time counter" who moves time to the next multiple of σ when the actors are ready.

The exercise starts by the random number generator presenting a random number in $\{0, \ldots, CW_{min}\}$ to each of the terminal transmitters to use as their backoff counter. Terminals are all in backoff stage 0 at the start.

Each terminal comprises a transmitter and receiver. The transmitter's job is to decrement the backoff counter whenever the receiver allows it to do so, and then transmit a packet whenever the backoff counter hits zero. The receiver's job is to sense the channel (and thus stop the backoff counter from whenever a packet is transmitted until DIFS after the ACK is finished). After transmitting a packet, or a collision, the transmitter requests a random number (ask the random number generator to pick a number out of a hat according to the backoff stage). The receiver's job is also to make sure that an ACK is received within $ACK_{Timeout}$ of the start of the transmitter sending a packet. If it is not, increment the backoff stage (up to the maximum stage m) and tell the transmitter to request a new random number for the backoff counter.

The access point receiver's job is to listen for each packet; then SIFS after the end of a packet, the access point transmitter sends an ACK. Unless, of course, two packets were transmitted at the same time, in which case, neither packet is received and an ACK is not transmitted.

Whenever your terminal or access point transmits a packet, the transmitter actor will hold up card letting the other players know that he or she is occupying the medium. For the hidden terminal simulation, terminal A should ignore terminal C, and vice versa.

Important questions:

1. How efficient is the 802.11 DCF when there is no hidden terminal problem? The paper eventually shows that a utilization rate of just above 80% is possible in this case.

2. How significant is the hidden node problem in the 802.11 DCF?

27.5 RTS/CTS

To help deal with the hidden terminal problem, a MAC protocol named *request-to-send/clear-to-send* or RTS/CTS is used. In short, a terminal doesn't just send its data — instead, it sends a very short packet with its id and an indication that it wants to send a data packet of a given duration. The destination schedules a reserved time for that transmission, and other terminals in range of the destination terminal must not transmit during that reserved time. After successful reception, the destination also sends a final ACK to acknowledge receipt.

Lecture 18

Today: Diversity: (1) Types, (2) Combining

28 Diversity

Diversity is the use of multiple channels to increase the signal to noise ratio in the presence of random fading losses. The big picture idea of diversity is "don't put all of your eggs in one basket".

For fading channels, we know that there is a finite probability that a signal power will fall below any given fade margin. For a Rayleigh channel, we showed that to have the signal above the required SNR 99% of the time, we needed to include a fade margin of 18.9 dB. This is a big "loss" in our link budget. For example, if we didn't need to include the fade margin in the link budget, we could multiply the path length by a factor of $10^{18.9/20} \approx 10$ (in free space); or increase the number of bits per symbol in the modulation significantly higher so that we can achieve higher bit rate for the same bandwidth.

There are several physical means to achieve multiple channels, and to get the received power on those channels to be nearly independent. Each has its advantages and disadvantages.

After this lecture, you should have three critical skills:

1. Understand what is meant by space, polarization, frequency, multipath, and time diversity, and the benefits and drawbacks of implementing each diversity method in a wireless communications system.

2. Understand how to combine the signals from multiple channels, including scanning, selection, equal gain, and maximal ratio combining methods.

3. Know the effect on system design: Be able to calculate the probability of outage or required fade margin when using a particular diversity combining scheme, assuming Rayleigh fading.

28.1 Methods for Channel Diversity

28.1.1 Space Diversity

Space diversity at a receiver is the use of multiple antennas across space. Because multipath fading changes quickly over space (see lecture notes on fading rate, and Doppler fading), the signal amplitude on the antennas can have a low correlation. The low correlation typically comes at separation distances of more than half the wavelength. The Jakes model (equal power from all angles) says that the correlation coefficient at $\lambda/2$ is exactly zero; however, in reality, this is not true. The actual angular power profile (multipath power vs. angle) determines the actual correlation coefficient. In general, we either accept that the correlation coefficient is not perfectly zero, or we separate the antennas further than $\lambda/2$. Or, add some gain pattern diversity, *i.e.*, using different gain patterns to help decorrelate the signals.

The problems with space diversity are most importantly that for consumer radios, we want them to be small; and multiple antennas means that the device will be larger. This is fine when space is not a big concern – for base stations, or for laptops or even access points. Another problem is (except for scanning combining, see below) that a receiver with multiple antennas must have one RF chain (downconverter, LNA, filter) per antenna. The benefits of space diversity are that no additional signal needs to be transmitted, and no additional bandwidth is required.

Space diversity could be used at a transmitter, by changing the transmit antenna until the receiver SNR is high enough. In the past, this requires some closed loop control, and so was less common. MIMO has changed that. We will cover MIMO in lecture 19, including showing how Alamouti coding can be used to use the multiple antennas at a transmitter to achieve space diversity.

28.1.2 Polarization Diversity

Polarization diversity is the use of two antennas with different polarizations. We know that reflection coefficients are different for horizontal and vertically polarized components of the signal. Scattered and diffracted signal amplitudes and phases also are different for opposite polarizations. Thus we can consider one polarized signal, which is the sum of the amplitudes and phases of many reflected, scattered, and diffracted signals, to be nearly uncorrelated with the other polarized signal.

The advantages of polarization diversity is that the two antennas don't need to be spaced $\lambda/2$ apart. It may be combined with space diversity so to further reduce the correlation coefficient between the signal received at two antennas. Polarization diversity, like space diversity, doesn't require any additional bandwidth or signal transmission from the transmitter.

The disadvantages are simply that there can be only two channels – vertical and horizontal (or equivalently, right-hand and left-hand circular) polarizations. It may require two receiver RF chains (again, unless a scanning combiner is used).

28.1.3 Frequency Diversity

Frequency diversity uses multiple transmissions in different frequency bands. This doesn't typically mean transmitting exactly the same thing on multiple different bands (which would require multiple times more bandwidth!). Frequency division multiplexing (FDM) or orthogonal FDM (OFDM) are the typical examples, which divide the data into N different bands. Error correction coding is used so that some percent of errors can be corrected, so if a certain percent of the bands experience deep fades, and all of that data is lost, the data can still be recovered during decoding. Frequency

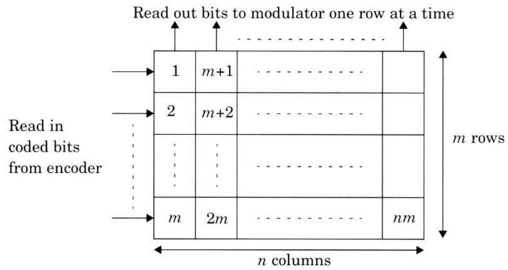


Figure 7.17 Block interleaver where source bits are read into columns and read out as n-bit rows.

bands in FDM or OFDM are typically correlated – each band needs to be in frequency flat fading so that equalization does not need to be used – but this means that bands right next to each other still have significant positive fading correlation.

FH-SS is another frequency diversity example. FH-SS may experience deep fades (and interference) on some center frequencies among its hopping set, but it is unlikely to lose more than a percentage of its data. It also uses error correction coding.

Frequency diversity methods can also be set to control which frequency bands/ channels the transmitter uses, to remove the bands that are in deep fades. Again, this requires closed loop control.

Advantages of frequency diversity are that only one antenna, and one RF chain, is needed. A disadvantage is that, because some of the transmit power is used to send data in bands that are in deep fades, the power efficiency is less compared to space diversity, in which the transmitter sends all of its power in one channel.

28.1.4 Multipath diversity

Multipath diversity is the capturing of multipath signals into independent channels. In DS-SS, a rake receiver achieves multipath diversity by isolating multipath components separately from each other based on their differing time delays. If one time delay group fades, another time delay group may not fade. These "fingers" of the rake receiver do not require different RF chains (an advantage compared to space diversity) and benefit most when the multipath channel is the worst, for example, in urban areas, or in mountain canyons. The disadvantage of DS-SS is the large frequency band required – for example, 20 MHz for 802.11b, or 1.25 MHz for IS-95 (cellular CDMA). There is also significant computational complexity in the receiver, although standard ICs now exist to do this computation for these common commercial devices.

The Rappaport book calls this "time diversity", but I think it is confusing – perhaps "multipath diversity" or even "multipath time delay diversity" are better names.

28.1.5 Time Diversity

Time diversity is the use of a changing channel (due to motion of the TX or RX) at different times (μs or ms). For example, one might send the same data at multiple different times, but this would require multiple times the transmit power, and reduce the data rate possible on one channel. This incurs additional latency (delay). Packet retransmissions (e.g., TCP) can be viewed as time diversity. However, it is used in almost all common commercial systems in the form of "interleaving". Interleaving takes an incoming coded bitstream and spreads the bits across a transmitted packet in a known pattern. An example interleaver used by a transmitter is shown in

Figure 7.17 in the Rappaport book. In the receiver, the inverse interleaving operation is performed. This way, a burst of (multiple sequential) coded bit errors caused by the channel are spread across the packet by the interleaver. Error correction codes are more effective when errors are not grouped together (recall our block coding and decoding – we assumed at most one error per 6 or 7 received coded bits). In general, coding methods correct a few out of each group of coded bits received, but not more.

Interleaving's only disadvantage is additional latency – you need to receive the entire block of coded bits before they can be put in order and decoded (and then converted into an audio signal, for example). For different applications, latency requirements are different. Voice communications are typically the most latency-sensitive, and even cell phone voice data is interleaved.

The disadvantage is that temporal correlation can be very long for most applications, even for vehicular communications.

28.2 Diversity Combining

In the previous section, we described how we might achieve M different (nearly) independent channels. In this section, we discuss what to do with those independent signals once we get them. These are called combining methods. *We need them for space, polarization, and multipath diversity methods.* For frequency diversity (FDM and OFDM) combining is done by the FDM or OFDM receiver using all frequency band signals. For time diversity (interleaving) we described the combining above. We only want one bitstream, so somehow we need to combine the channels' signals together. Here are some options, in order of complexity:

1. *Scanning Combiner*: Scan among the channels, changing when the current SNR goes below the threshold.
2. *Selection Combiner*: Select the maximum SNR channel's signal and use only it.
3. *Equal Gain Combiner*: Co-phase the signals and then add them together.
4. *Maximal Ratio Combiner*: The optimal solution in terms of SNR – co-phase and weight (multiply) each signal by the square root of its signal to noise ratio (SNR), and then add them together.

"Co-phase the signals" means that we need to multiply signals by $e^{j\phi_i}$ for some constant phase angle ϕ_i on channel i, so that the (otherwise random) phases of the signals on the different channels line up. If we don't co-phase the signals before combining them, we end up with the same multipath fading problem we've always had - signals sometimes add together destructively. You should be prepared to describe any of these combining methods, and discuss its effect on the fade margin required for a link.

28.2.1 Selection Combining

Let's say that we have M statistically independent channels. This independence means that one channel's fading does not influence, or is not correlated in any way with, another channel's fading.

Let's assume that each channel is Rayleigh with identical mean SNR Γ. At any given instant, the SNR on channel i is denoted γ_i. Based on the Rayleigh assumption for γ_i, it has a CDF of:

$$P\left[\gamma_i \le \gamma\right] = 1 - e^{-\gamma/\Gamma}$$

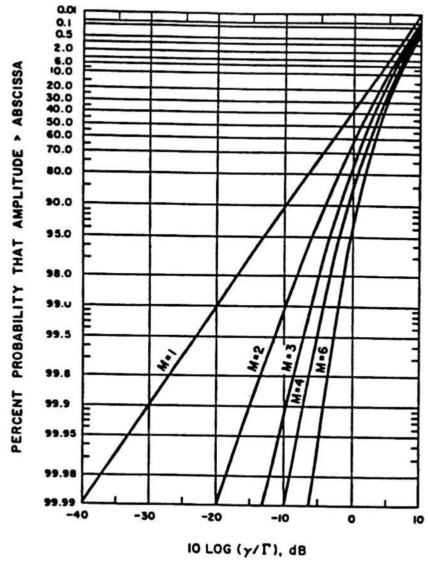


Figure 7.11 Graph of probability distributions of *SNR* = γ threshold for *M* branch selection diversity. The term Γ represents the mean *SNR* on each branch [from [Jak71] © IEEE].

Figure 34: Rappaport Figure 7.11, the impact of selection combining.

This means that the probability that the SNR on channel i is less than the threshold γ is given by $1 - e^{-\gamma/\Gamma}$, where again, Γ is the mean SNR for the channel. In past lectures, we showed that we can determine a fade margin for a single channel ($M = 1$) based on this equation. For example, setting the probability of being less than the threshold to 1%,

$$\begin{aligned} 0.01 &= 1 - e^{-\gamma/\Gamma} \\ 0.99 &= e^{-\gamma/\Gamma} \\ \gamma &= \Gamma(-\ln 0.99) = \Gamma(0.0101) = \Gamma(\text{dB}) - 19.98(\text{dB}) \end{aligned} \tag{45}$$

Thus compared to the mean SNR on the link, we need an additional 20 dB of fade margin (this is slightly less when we use the median SNR).

In contrast, in selection combining, we only fail to achieve the threshold SNR when *all* channels are below the threshold SNR. Put in another way, if any of the channels achieve good enough SNR, we'll select that one, and then our SNR after the combiner will be good enough. What is the probability all of the M channels will fail to achieve the threshold SNR γ? All M channels have to have SNR below γ. The probability is the product of each one:

$$P\left[\gamma_i < \gamma, \forall i = 1, \ldots, M\right] = \left[1 - e^{-\gamma/\Gamma}\right] \cdots \left[1 - e^{-\gamma/\Gamma}\right] = \left[1 - e^{-\gamma/\Gamma}\right]^M$$

Example: What is the required fade margin when assuming Rayleigh fading and $M = 2$ independent channels, for a 99% probability of being above the receiver threshold?

Again, set 0.01 equal, this time, to $\left[1 - e^{-\gamma/\Gamma}\right]^2$, so

$$\begin{aligned} 0.1 &= 1 - e^{-\gamma/\Gamma} \\ 0.9 &= e^{-\gamma/\Gamma} \\ \gamma &= \Gamma(-\ln 0.9) = \Gamma(0.1054) = \Gamma(\text{dB}) - 9.77(\text{dB}) \end{aligned} \tag{46}$$

So the fade margin has gone down to less than 10 dB, a reduction in fade margin of 10 dB!

As M increases beyond 2, you will see diminishing returns. For example, for $M = 3$, the required fade margin improves to 6.15 dB, a reduction of 3.6 dB, which isn't as great as the reduction in fade margin due to changing M from 1 to 2.

28.2.2 Scanning Combining

Selection combining assumes we know all signal amplitudes so that we can take the maximum. Scanning combining is a simplification which says that we only have one receiver, so we can only know the signal to noise ratio on one channel at a time. But we can switch between them when one channel's SNR drops too low. We can often achieve nearly the same results using a scanning combiner as with selection combining.

28.2.3 Equal Gain Combining

Here, we simply co-phase the signals and then add them together. The outage probability improves compared to selection combining. Denoting the SNR of the summed signal as γ_Σ, an analytical expression for the outage probability given Rayleigh fading is [11, p. 216],

$$P\left[\gamma_\Sigma < \gamma\right] = 1 - e^{-2\gamma/\Gamma} - \sqrt{\pi\gamma/\Gamma}\, e^{-\gamma/\Gamma}\left(1 - 2\mathrm{Q}\left(\sqrt{2\gamma/\Gamma}\right)\right)$$

where $\mathrm{Q}(\cdot)$ is the tail probability of a zero-mean unit-variance Gaussian random variable, as we've used before to discuss bit error probabilities of modulations.

28.2.4 Maximal Ratio Combining

For maximal ratio combining, we still co-phase the signals. But then, we weight the signals according to their SNR. The intuition is that some channels are more reliable than others, so we should "listen" to their signal more than others – just like if you hear something from multiple friends, you probably will not weight each friend equally, because you know who is more reliable than others.

The outage probability improves compared to equal gain combining. Denoting the SNR of the summed signal as γ_Σ, an analytical expression for the outage probability given Rayleigh fading is [11, p. 214],

$$P\left[\gamma_\Sigma < \gamma\right] = 1 - e^{-\gamma/\Gamma} \sum_{k=1}^{M} \frac{(\gamma/\Gamma)^{k-1}}{(k-1)!}.$$

Lecture 19

Today: (1) Capacity, (2) MIMO

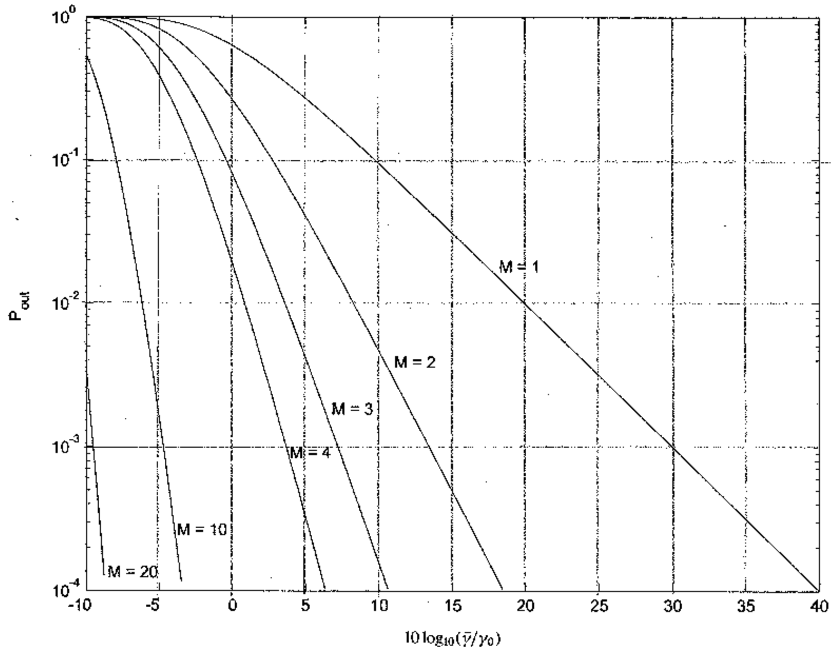


Figure 7.5: P_{out} for maximal-ratio combining with i.i.d. Rayleigh fading.

Figure 35: Goldsmith Figure 7.5, the impact of maximal ratio combining.

29 Shannon-Hartley Bandwidth Efficiency

Ralph V. L. Hartley[1] worked as a researcher in radio telephony for the Western Electric Company, and later, at Bell Laboratories, where he developed relationships useful for determining the capacity of band-limited communication channels published in 1928 [13]. Claude Shannon, known as "the father of information theory", worked at Bell Labs starting in 1940. In addition to his work on cryptography, Shannon extended the result of Hartley to provide a unifying approach to finding the capacity of digital communications systems in "A mathematical theory of communications" in 1948 [23]. Shannon's result provides a fundamental relationship between bandwidth, B, error-free bit rate, R_b, and signal to noise ratio, S/N [23]:

$$R_b \leq C = B \log_2 \left(1 + \frac{S}{N}\right) \tag{47}$$

We can use this *Shannon-Hartley bound* to develop a bound on the bandwidth efficiency.

Def'n: *Bandwidth efficiency*
The bandwidth efficiency of a digital communication system is ratio of $\eta_B = R_b/B$, where R_b is the bits per second achieved on the link, and B is the signal bandwidth occupied by the signal. Bandwidth efficiency has units of bits per second per Hertz.

The limit on bandwidth efficiency is a direct result of the Shannon-Hartley capacity formula,

[1]Hartley received the A.B. degree from the University of Utah in 1909

and is given by,

$$\begin{aligned} \frac{R_b}{B} &\leq \log_2\left(1+\frac{S}{N}\right) \\ \frac{R_b}{B} &\leq \log_2\left(1+\frac{R_b}{B}\frac{\mathcal{E}_b}{N_0}\right) \end{aligned} \tag{48}$$

Note that this latter expression can't be solved for $\frac{R_b}{B}$ directly – one needs to find a plot of the $\frac{R_b}{B}$ possible for any given $\frac{\mathcal{E}_b}{N_0}$. I have done this using Matlab's "fsolve" function, and a plot of the result is in Figure 36.

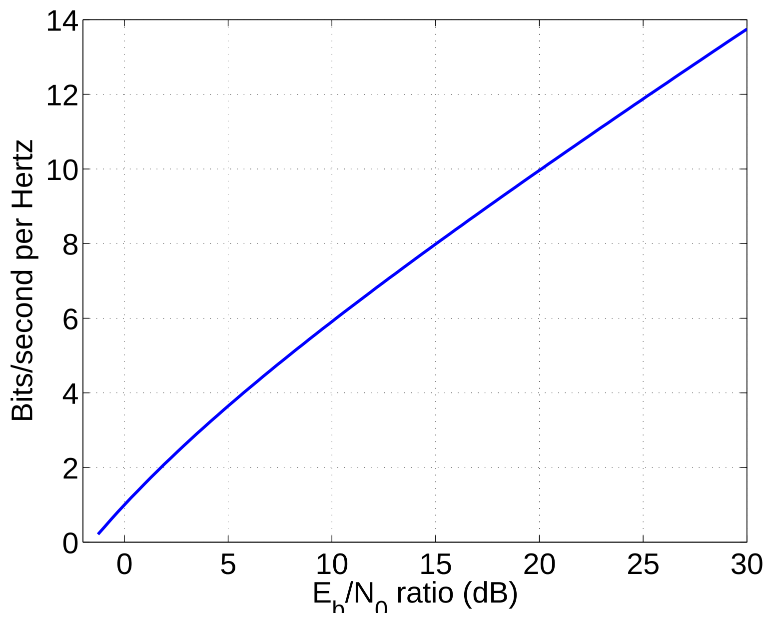


Figure 36: The maximum bandwidth efficiency, in bits per second per Hertz, vs. $\frac{\mathcal{E}_b}{N_0}$ in dB. This graph is the solution to (47).

Example: Maximum for various $\frac{\mathcal{E}_b}{N_0}$ (dB)

What is the maximum bandwidth efficiency of a link with $\frac{\mathcal{E}_b}{N_0}$(dB) $= 10, 15, 20$? What maximum bit rate can be achieved per 30 kHz of spectrum (one AMPS/USDC channel)? Is this enough for 3G?

Solution: Using Figure 36, the $\frac{R_{max}}{B} \leq$ 6, 8, and 10 bps/Hz. for $\frac{\mathcal{E}_b}{N_0}$(dB) $= 10, 15, 20$. With 30 kHz, multiplying, we have $R_b \leq$ 180, 240, 300 kbps. No, this isn't enough for 3G cellular, which is supposed to achieve up to 14 Mbps.

Shannon's bound is great as engineers, we can come up with a quick answer for what we *cannot* do. But it doesn't necessarily tell us what we *can* achieve. Our probability of bit error equations for particular modulation types is the opposite – it gives us achievable bit error rates, but no idea on whether one can do better for the particular $\frac{\mathcal{E}_b}{N_0}$ we have on the link.

Note that achieving (approximately) linear improvements in bps/Hz requires linear increases in dB energy. That is, in the example above, one needs to add 5 dB into the link budget for every extra 2 bps/Hz we need. This (5 dB increase) translates to multiplying the linear transmit power by 3.2 each time we want 2 bps/Hz more.

It is very difficult to increase bit rate to handle the exponentially increasing demand from users simply by increasing power. MIMO is a solution to this limitation – by adding additional channels, via multiple antennas at a transceiver, we can multiply the achievable bandwidth efficiency.

30 MIMO

Multiple-input multiple output (MIMO) is a particular type of space and/or polarization diversity in which both the transmitter and receiver may use multiple antennas.

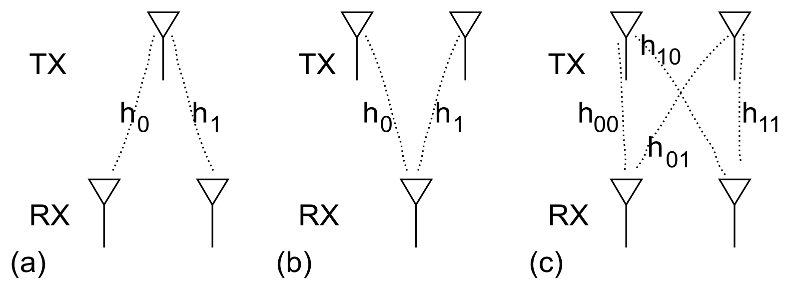


Figure 37: Transmit and receive space diversity schemes: (a) traditional space diversity with receiver combining, called single input multiple output (SIMO); (b) transmit diversity, which may use Alamouti's scheme, called multiple input single output (MISO); (c) 2×2 multiple input multiple output (MIMO).

30.1 Revisit Maximal Ratio Combining

We're going to introduce MIMO by comparing it to maximal ratio combining (MRC) using a receiver with two antennas (space diversity), as shown in Figure 37(a). Let's assume that the transmitter sends symbol voltage s. Assume the channel from TX to RX antenna 0 experiences total channel power gain α_0^2 (or voltage gain α_0), and the channel from TX to RX antenna 1 experiences total channel power gain α_1^2 (or voltage gain α_1).

We said that in MRC, the received signals r_0 and r_1 are multiplied by the square root of the SNRs. Multiplying received signals by a constant doesn't help (it amplifies the noise as much as the signal) so it really matters to multiply them by different numbers. Here, those numbers turn out to be α_0 and α_1.

$$r_{MRC} = \alpha_0 r_0 + \alpha_1 r_1$$

If the transmitted symbol voltage was s, then because the channel voltage gains were α_0 and α_1, we must have that $r_0 = \sqrt{P_t}\alpha_0 s + n_0$ and $r_1 = \sqrt{P_t}\alpha_1 s + n_1$, where n_0, n_1 are the additive noise introduced by the channel. (There would have been phase shifts as well, but remember that co-phasing is done prior to MRC.) So,

$$r_{MRC} = \sqrt{P_t}\left[\alpha_0^2 + \alpha_1^2\right] s + \alpha_0 n_0 + \alpha_1 n_1$$

In the case when we had only one receive antenna, we would have received either r_1 or r_0. In comparison, the noise terms are multiplied by α_0 or α_1, but the signal is multiplied by the sum of $\alpha_0^2 + \alpha_1^2$. If one α_i fades, we don't lose the entire signal s.

30.2 Alamouti code

The MIMO started gaining steam in 1998, from two different results, one from Bell Labs, where they had built an experimental MIMO system they called V-BLAST [28], and a simple transmit diversity scheme from S. M. Alamouti now called the *Alamouti scheme* [3]. The Alamouti scheme is a simple way to achieve a performance similar to MRC using 2×1 MISO (two transmit antennas and a single receiver), like the system shown in Figure 37(b). The advantage is that in some cases, the transmitter is more able to have multiple antennas, while the receiver is more limited in size (for example, cellular communications on the downlink).

Alamouti presented a simple scheme that sends two symbols simultaneously, but takes two symbol periods to do so, and over the two transmit antennas. Denote these two symbols s_0 and s_1. The idea is, first transmit s_0 out of antenna 0 and s_1 out of antenna 1. At the receiver, assuming the channels are $h_0 = \alpha_0 e^{j\theta_0}$ and $h_1 \alpha_1 e^{j\theta_1}$, will be

$$r_0 = s_0 \alpha_0 e^{j\theta_0} + s_1 \alpha_1 e^{j\theta_1} \tag{49}$$

Then, during the subsequent symbol period, send $-s_1^*$ out of antenna 0 and s_0^* out of antenna 1, where the superscript $*$ is used to denote complex conjugate. During the second symbol period the receiver will see

$$r_1 = -s_1^* \alpha_0 e^{j\theta_0} + s_0^* \alpha_1 e^{j\theta_1} \tag{50}$$

Note this assumes the channel was the same during the second symbol period as during the first.

The "magic" happens when we combine r_0 and r_1 in the following way to come up with estimates of s_0 and s_1. We form:

$$\begin{aligned} \tilde{s}_0 &= h_0^* r_0 + h_1 r_1^* \\ \tilde{s}_1 &= h_1^* r_0 - h_0 r_1^* \end{aligned}$$

Plugging in for h_0 and h_1,

$$\begin{aligned} \tilde{s}_0 &= \alpha_0 e^{-j\theta_0} r_0 + \alpha_1 e^{j\theta_1} r_1^* \\ \tilde{s}_1 &= \alpha_1 e^{-j\theta_1} r_0 - \alpha_0 e^{j\theta_0} r_1^* \end{aligned}$$

Plugging in for r_0 and r_1 as given in (49) and (50), respectively,

$$\begin{aligned} \tilde{s}_0 &= \alpha_0 e^{-j\theta_0}(s_0 \alpha_0 e^{j\theta_0} + s_1 \alpha_1 e^{j\theta_1}) + \alpha_1 e^{j\theta_1}(-s_1 \alpha_0 e^{-j\theta_0} + s_0 \alpha_1 e^{-j\theta_1}) \\ \tilde{s}_1 &= \alpha_1 e^{-j\theta_1}(s_0 \alpha_0 e^{j\theta_0} + s_1 \alpha_1 e^{j\theta_1}) - \alpha_0 e^{j\theta_0}(-s_1 \alpha_0 e^{-j\theta_0} + s_0 \alpha_1 e^{-j\theta_1}) \end{aligned}$$

Simplifying,

$$\begin{aligned} \tilde{s}_0 &= \alpha_0^2 s_0 + s_1 \alpha_0 \alpha_1 e^{j(\theta_1 - \theta_0)} - s_1 \alpha_0 \alpha_1 e^{j(\theta_1 - \theta_0)} + \alpha_1^2 s_0 \\ \tilde{s}_1 &= \alpha_1^2 s_1 + s_0 \alpha_0 \alpha_1 e^{j(\theta_0 - \theta_1)} - s_0 \alpha_0 \alpha_1 e^{j(\theta_0 - \theta_1)} + \alpha_0^2 s_1 \end{aligned}$$

The middle terms cancel out in each case, so finally,

$$\begin{aligned} \tilde{s}_0 &= (\alpha_0^2 + \alpha_1^2) s_0 \\ \tilde{s}_1 &= (\alpha_1^2 + \alpha_0^2) s_1 \end{aligned}$$

In short, in two symbol periods, we've managed to convey two symbols of information. Each symbol arrives with approximately the same signal amplitude that we would have had in the maximal ratio combining case.

Notes:

1. This is a two-by one code, that is, it works for two transmit antennas and one receive antenna. This code has been generalized for $n \times m$ MIMO systems, and called "space-time block codes", by Tarokh et. al. [25]. These can send more symbols in less time – in k symbol periods, you can send more than k symbols.

2. If you transmit out of two antennas, you would in general need twice as much power as the receiver diversity case, which had one transmit antenna. So generally we compare the two when using the same total transmit power, *i.e.*, cut the power in half in the transmitter diversity case. The performance is thus 3 dB worse than the receiver MRC diversity case.

3. The Alamouti and space-time block codes are not optimal. Space-time coding is the name of the general area of encoding information the multiple channels. One better-performing scheme is called space-time trellis coding. But the decoding complexity of space-time trellis codes increases exponentially as a function of the spectral efficiency [14, p377] [25].

30.3 MIMO Channel Representation

In general for MIMO, we have multiple (N_t) transmitters and multiple (N_r) receivers. We refer to the system as a (N_t, N_r) or $N_t \times N_r$ MIMO system. Figure 37(c) shows the channels for a $(2, 2)$ MIMO system. For the channel between transmitter k and receiver i, we denote the "channel voltage gain" as $h_{i,k}$. This gain is a complex number, with real and imaginary parts. Recall that the phase of a multipath component changes with distance, frequency, due to reflections, etc. The channel power gain would be $|h_{i,k}|^2$. The received voltage signal at i, just from transmitter k, is $s_k h_{i,k}$, where s_k is what was transmitted from antenna k.

To keep all these numbers organized, we use vectors and matrices. The transmitted signal from antennas $1, \ldots, N_t$ is $\mathbf{s}$,

$$\mathbf{s} = [s_1, \ldots, s_{N_t}]^T$$

and the channel gain matrix H is given as

$$H = \begin{bmatrix} h_{1,1} & h_{2,1} & \cdots & h_{N_t,1} \\ h_{1,2} & h_{2,2} & \cdots & h_{N_t,2} \\ \vdots & \vdots & \ddots & \vdots \\ h_{1,N_r} & h_{2,N_r} & \cdots & h_{N_t,N_r} \end{bmatrix} \tag{51}$$

Where there are N_r rows each corresponding to the channels measured at each receiver; and N_t columns each corresponding to the channels from each transmitter.

The received signal at receiver i is a linear combination of the s_k for $k = 1, \ldots, N_t$ terms plus noise:

$$x_i = \sum_{k=1}^{N_t} h_{i,k} s_k + w_i$$

where w_i the additive noise term, and $i = 1, \ldots, N_r$. In matrix form, we can rewrite this as:

$$\mathbf{x} = H\mathbf{s} + \mathbf{w}$$

where $\mathbf{x} = [x_1, \ldots, x_{N_r}]^T$ is the received vector and $\mathbf{w} = [w_1, \ldots, w_{N_r}]^T$ is the noise vector.

30.4 Capacity of MIMO Systems

Earlier we described the Shannon-Hartley theoretical limit to the bps per Hz we can achieve on a channel. Using multiple antennas at the TX and RX increases this theoretical limit. We said that the limit on bandwidth efficiency is given as,

$$\frac{R_{max}}{B} = \log_2(1+\rho) \tag{52}$$

where R_{max} is the maximum possible bit rate which can be achieved on the channel for given signal to noise ratio ρ and bandwidth B.

In a $N_t \times N_r$ MIMO system with channel matrix H as given in (51), with $N_t \geq N_r$, the new Shannon limit on bps per Hz is [14],

$$\frac{R_{max}}{B} = E\left[\log_2\left\{\det\left(\mathrm{I}_{N_r} + \rho\frac{1}{N_t}HH^\dagger\right)\right\}\right] \tag{53}$$

where $H^\dagger$ is the complex conjugate of H (I'm copying the notation of the Haykin Moher book), and ρ is the average signal to noise ratio. Here, we assume that each channel is Rayleigh, that is each channel voltage gain $h_{i,k}$ is complex Gaussian, and all channel gains are independent from each other. This is why we need an expected value – the matrix H is filled with random variables.

To get more intuition about the bandwidth efficiency limit, consider that the term $HH^\dagger$ is a Hermitian $N_r \times N_r$ matrix with eigendecomposition $HH^\dagger = U\Lambda U^\dagger$ where U is the matrix of eigenvectors of $HH^\dagger$ and Λ is a diagonal matrix of eigenvalues λ_i for $i = 1, \ldots, N_r$. In this case, we can rewrite (53) as,

$$\frac{R_{max}}{B} = E\left[\sum_{i=1}^{N_r}\log_2\left(1+\rho\frac{\lambda_i}{N_t}\right)\right] \tag{54}$$

Compared to (52), Equation (54) is a sum of several Shannon capacities – each with effective SNR $\rho\frac{\lambda_i}{N_t}$. Recall this was the formula for $N_t \geq N_r$. For $N_r \geq N_t$, the formula is

$$\frac{R_{max}}{B} = \sum_{i=1}^{N_t} E\left[\log_2\left(1+\rho\frac{\lambda_i}{N_r}\right)\right] \tag{55}$$

These $\min(N_t, N_r)$ "channels" are called the "eigen-mode channels" of a MIMO system.

In summary, we have created $\min(N_t, N_r)$ eigen-mode channels. Results have shown that the total capacity increases approximately with $\min(N_t, N_r)$. MIMO is so attractive for current and future communication systems because it multiplies the achievable bit rate by this factor of $\min(N_t, N_r)$, without requiring additional bandwidth or signal energy.

References

[1] N. Abramson. The aloha system – another alternative for computer communications. In *Proceedings of the Fall Joint Computer Conference*, pages 281–285, 1970.

[2] N. Abramson. The throughput of packet broadcasting channels. *IEEE Transactions on Communications*, 25(1):117–128, 1977.

[3] S. M. Alamouti. A simple transmitter diversity scheme for wireless communications. *IEEE J. Select. Areas Commun.*, 16(8):1451–1458, 1998.

[4] E. Benner and A. B. Sesay. Effects of antenna height, antenna gain, and pattern downtilting for cellular mobile radio. *IEEE Transactions on Vehicular Technology*, 45(2), May 1996.

[5] G. Bianchi. Performance analysis of the IEEE 802.11 distributed coordination function. *IEEE Journal on Selected Areas in Communications*, 18(3):535–547, 2000.

[6] J. E. Brittain. Electrical engineering hall of fame: Harald T. Friis. *Proceedings of the IEEE*, 97(9):1651–1654, Sept. 2009.

[7] N. Chandran and M. C. Valenti. Three generations of cellular wireless systems. *IEEE Spectrum*, pages 32–35, Jan/Feb 2001.

[8] G. D. Durgin, T. S. Rappaport, and H. Xu. Measurements and models for radio path loss and penetration loss in and around homes and trees at 5.85 GHz. *IEEE Trans. Communications*, 46(11):1484–1496, Nov. 1998.

[9] M. Feuerstein, K. Blackard, T. Rappaport, S. Seidel, and H. Xia. Path loss, delay spread, and outage models as functions of antenna height for microcellular system design. *Vehicular Technology, IEEE Transactions on*, 43(3):487–498, Aug 1994.

[10] G. Fors. The mobile telephone: in Bell System service, 1946-1985. Retrieved Jan. 11, 2010 from `http://www.wb6nvh.com/Carphone.htm`.

[11] A. Goldsmith. *Wireless communications*. Cambridge University Press, 2005.

[12] F. Gunnarsson, M. Johansson, A. Furuskär, M. Lundevall, A. Simonsson, C. Tidestav, and M. Blomgren. Downtilted Base Station Antennas–A Simulation Model Proposal and Impact on HSPA and LTE Performance. In *IEEE 68th Vehicular Technology Conference, 2008. VTC 2008-Fall*, pages 1–5, 2008.

[13] R. V. Hartley. Transmission of information. *AT&T Bell Laboratories Technical Journal*, 1927.

[14] S. Haykin and M. Moher. *Modern Wireless Communications*. Pearson Prentice Hall, 2005.

[15] W. Jianhui and Y. Dongfeng. Antenna downtilt performance in urban environments. In *IEEE Military Communications Conference, 1996. MILCOM'96, Conference Proceedings*, volume 3, 1996.

[16] E. Johnson. Link budgets for cellular networks. In *Global Wireless Education Consortium*, 2001. (presentation slides).

[17] D. Maas, M. H. Firooz, J. Zhang, N. Patwari, and S. K. Kasera. Channel sounding for the masses: Low complexity GNU 802.11b channel impulse response estimation. *IEEE Trans. Wireless Communications*. (in review).

[18] R. Malik. Spread spectrum - secret military technology to 3G. *IEEE History of Telecommunications Contest*, pages 1–5, 2001.

[19] C. Newport, D. Kotz, Y. Yuan, R. S. Gray, J. Liu, and C. Elliott. Experimental evaluation of wireless simulation assumptions. *SIMULATION: Transactions of The Society for Modeling and Simulation International*, 83(9):643–661, September 2007.

[20] K. Pahlavan and A. H. Levesque. *Wireless information networks*. John Wiley & Sons, 2 edition, 2005.

[21] N. Patwari. *Location Estimation in Sensor Networks*. PhD thesis, University of Michigan, Ann Arbor, MI, Sept. 2005.

[22] J. G. Proakis and M. Salehi. *Communication System Engineering*. Prentice Hall, 2nd edition, 2002.

[23] C. Shannon. A mathematical theory of communications. *Bell System Technical Journal*, 27:379–423 and 623–656, 1948. Available at `http://www.cs.bell-labs.com/cm/ms/what/shannonday/paper.html`.

[24] W. L. Stutzman and G. A. Theile. *Antenna Theory and Design*. John Wiley & Sons, 1981.

[25] V. Tarokh, H. Jafarkhani, and A. Calderbank. Space-time block codes from orthogonal designs. *IEEE Transactions on Information Theory*, 45(5):1456–1467, 1999.

[26] J. Treichler. A panoramic view of personal wireless systems from 1946 through 2016. In *2009 Virginia Tech Symposium on Wireless Personal Communications*, June 2009. (keynote presentation).

[27] US Federal Communications Commission. Radio frequency safety: Information on human exposure to radiofrequency fields from cellular and PCS radio transmitters. Retrieved Jan. 14, 2010 from `http://www.fcc.gov/oet/rfsafety/cellpcs.html`.

[28] P. Wolniansky, G. Foschini, G. Golden, and R. Valenzuela. V-BLAST: an architecture for realizing very high data rates overthe rich-scattering wireless channel. In *URSI International Symposium on Signals, Systems, and Electronics*, pages 295–300, 1998.