Practical Guide to Radio-Frequency Analysis and Design

© EETech Media, LLC. All rights reserved.

ALL ABOUT CIRCUITS

Practical Guide to Radio-Frequency Analysis and Design

Table of Contents

Introduction

The first engineered radio transmission system was implemented in the last decade of the nineteenth century. Since then, wireless communication has spread to all corners of the globe, profoundly influencing commerce, transportation, scientific investigation, warfare, and daily life. And yet, we are entering a new era of wireless connectivity—an era in which high-speed radio transceivers will be woven into the very fabric of the human experience.

The design and analysis of radio-frequency (RF) systems is considered a difficult discipline within the field of electrical engineering. Perhaps this is true, or perhaps electrical engineers simply do not receive adequate education in RF circuitry, which does indeed diverge from the techniques and components that we regularly encounter when working in the digital and low-frequency-analog domains. In any case, there surely are many students and technical professionals, within the EE community and beyond, who have not had access to introductory material that provides a firm and comprehensive foundation for future development of RF proficiency. This textbook is an attempt to remedy that situation. As the title indicates, the book emphasizes practice over theory—favoring approachable, engaging explanations that promote comprehension and a spirit of inquiry—and is intended to guide the reader toward successful analysis and design of wireless-communication systems that operate within the radio spectrum.

It is our hope that you find this book useful, interesting, and perhaps even enjoyable. We sincerely believe that it offers something unique to the engineering community and will prove to be a valuable resource for the next generation of RF engineers.

CHAPTER 1

Introduction to RF Principles and Components

- What Is RF and Why Do We Use It?
- Learning to Live in the Frequency Domain
- The RF Engineer’s Guide to the Decibel
- Passive Components in RF Circuits
- Active Components in RF Circuits

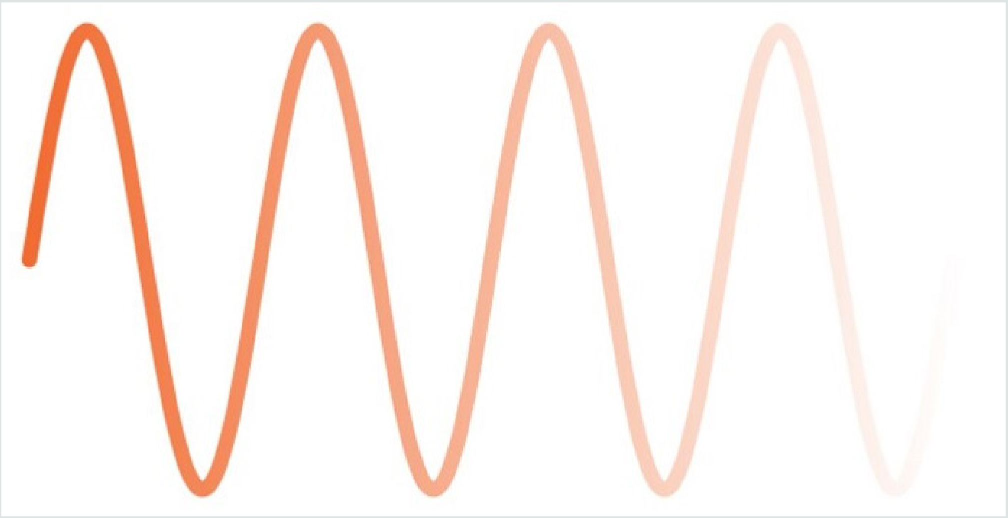
What Is RF and Why Do We Use It?

Learn about electromagnetic radiation and why it is so useful for wireless communication.

When we think of electricity, we naturally think of wires. From high-voltage transmission lines to tiny traces on a printed circuit board, wires are still the fundamental means of transferring electrical energy from one location to another.

But history has consistently demonstrated that human beings are rarely, if ever, satisfied with the fundamental way of doing things, and thus we should not be surprised to learn that the proliferation of electricity was followed by widespread efforts to free electrical functionality from the constraints of physical interconnections.

There are various ways to incorporate “wireless” functionality into an electrical system. One of these is the use of electromagnetic radiation, which is the basis for RF communication. However, it’s important to recognize that electromagnetic radiation is not unique in its ability to extend electrical circuitry into the wireless domain. Anything that can travel through a nonconductive material—mechanical motion, sound waves, heat—could be used as a (perhaps crude) means of converting electrical energy into information that does not rely on conductive interconnections.



Carefully manipulated sinusoidal voltage (or current) signals are the foundation of the modern wireless age.

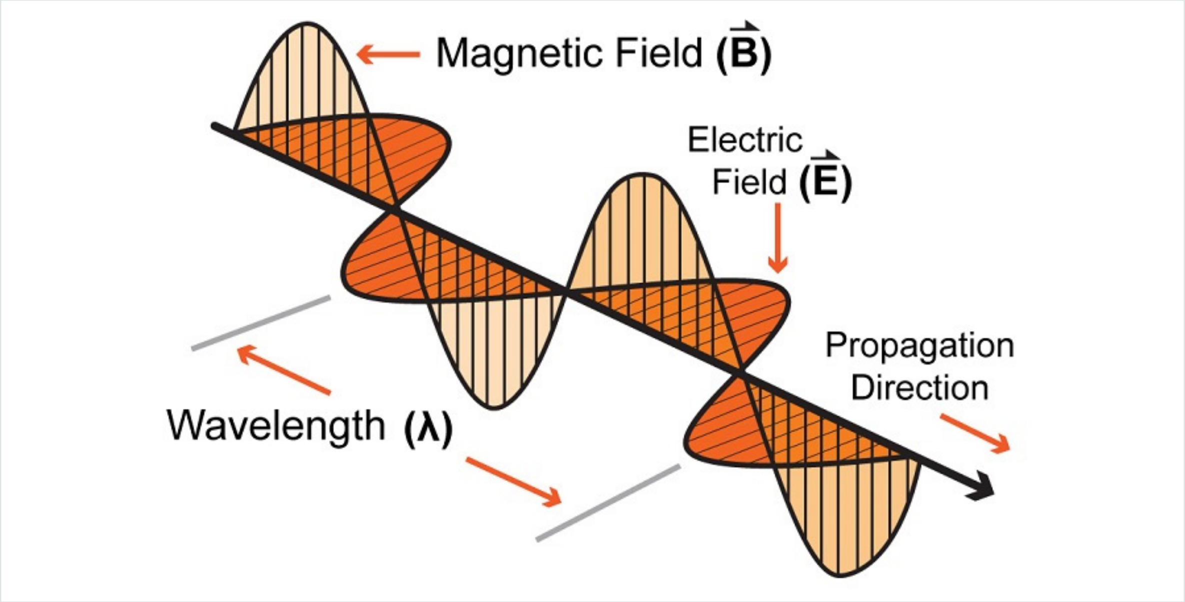
With this in mind, we can ask ourselves the more relevant questions: Why is electromagnetic radiation the preferred method? Why are other types of wireless communication of such secondary importance? Before we answer these questions, let’s make sure we understand what electromagnetic radiation is.

Fields and Waves

You could spend years studying the details of electromagnetism. Fortunately, you don't need that sort of expertise to successfully design and implement RF circuits. But you do need to have a basic idea of the mysterious energy being emitted from your device's antenna.

As the name implies, electromagnetic radiation involves both electric fields and magnetic fields. If you have voltage —such as the voltage across the impedance of an antenna—you have an electric field (from a mathematical standpoint, electric field is proportional to the spatial rate of change of voltage). If you have electric current—such as the current passing through the impedance of an antenna—you have a magnetic field (the strength of the field is proportional to the magnitude of the current).

The electric and magnetic fields are present even if the magnitude of the voltage or current is constant. However, these fields would not propagate. If we want a wave that will *propagate* out into the universe, we need *changes* in voltage and current.



The electric and magnetic components of an electromagnetic wave are represented as perpendicular sinusoids.

The key to this propagation phenomenon is the self-sustaining relationship between the electric and magnetic components of electromagnetic radiation. A changing electric field generates a magnetic field, and a changing magnetic field generates an electric field. This mutual regeneration is manifested as a distinct entity, namely, an electromagnetic wave. Once generated, this wave will travel outward from its source, careening day after day, at the speed of light, toward the depths of the unknown.

Creating EMR vs. Controlling EMR

Designing an entire RF communication system is not easy. However, it is extremely easy to generate electromagnetic radiation (EMR), and in fact you generate it even when you don't want to. Any time-varying signal in any circuit will generate EMR, and this includes digital signals. In most cases this EMR is simply noise. If it's not causing any trouble, you can ignore it. In some cases it can actually interfere with other circuitry, in which case it becomes EMI (electromagnetic interference).

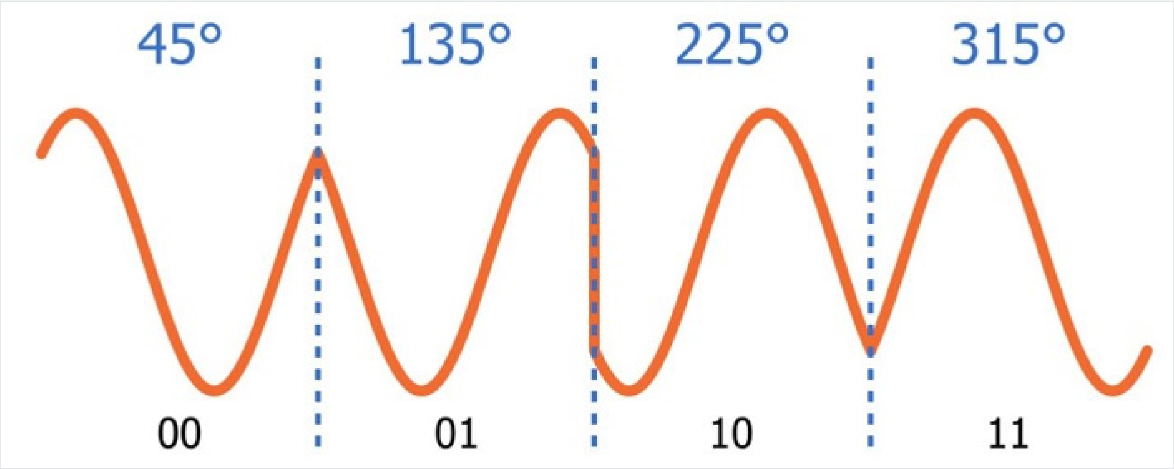
We see, then, that RF design is not about merely generating EMR; rather, RF design is the art and science of generating and manipulating and interpreting EMR in a way that allows you to reliably transfer meaningful information between two circuits that have no direct electrical connection.

Why EMR?

Now let's return to the question of why EMR-based systems are so common compared to other forms of wireless communication. In other words, why does "wireless" almost always refer to RF when various other phenomena can transfer information without the aid of wires? There are a few reasons:

Agility

EMR is a natural extension of the electrical signals used in wired circuits. Time-varying voltages and currents generate EMR whether you want them to or not, and furthermore, that EMR is a precise representation of the AC components of the original signal.



Each portion of this intricate QPSK waveform transfers two bits of digital information.

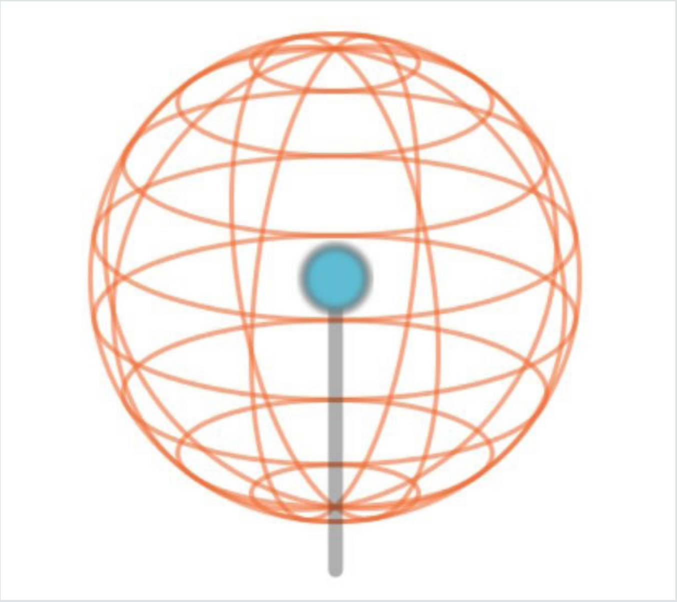
Let's consider an extreme (and completely impractical) counterexample: a heat-based wireless communication system. Imagine that a room contains two separate devices. The transmitter device heats up the room to a certain temperature based on the message it wants to send, and the receiver device measures and interprets the ambient temperature. This is a sluggish, awkward system because the temperature of the room cannot precisely follow the variations of an intricate electrical signal. EMR, on the other hand, is highly responsive. Transmitted RF signals can faithfully reproduce even the complex, high-frequency waveforms used in state-of-the-art wireless systems.

Speed

In AC-coupled systems, the rate at which data can be transferred depends on how quickly a signal can experience variations. In other words, a signal must be *doing something*—such as increasing and decreasing in amplitude—in order to convey information. It turns out that EMR is a practical communication medium even at very high frequencies, which means that RF systems can achieve extremely high rates of data transfer.

Range

The pursuit of wireless communication is closely linked to the pursuit of long-distance communication; if the transmitter and receiver are in close proximity, it is often simpler and more cost-effective to use wires. Though the strength of an RF signal decreases according to the inverse-square law, EMR—in conjunction with modulation techniques and sophisticated receiver circuitry—still has a remarkable ability to transfer usable signals over long distances.



The intensity of EMR decreases exponentially as the emitted energy propagates outward in all directions.

No Line of Sight Needed

The only wireless communication medium that can compete with EMR is light; this is perhaps not too surprising, since light is actually very-high-frequency EMR. But the nature of optical transmission highlights what is perhaps the definitive advantage offered by RF communication: a clear line of sight is not required.

Our world is filled with solid objects that block light—even very powerful light. We have all experienced the intense brightness of the summer sun, yet that intensity is greatly reduced by nothing more than a thin piece of fabric. In contrast, the lower-frequency EMR used in RF systems passes through walls, plastic enclosures, clouds, and—though it may seem a bit strange—every cell in the human body. RF signals are not completely unaffected by these materials and, in some cases, significant attenuation can occur. But compared to light, (lower-frequency) EMR goes just about anywhere.

Summary

- “RF” refers to the use of electromagnetic radiation for transferring information between two circuits that have no direct electrical connection.
- Time-varying voltages and currents generate electromagnetic energy that propagates in the form of waves. We can wirelessly transfer analog and digital data by manipulating and interpreting these waves.
- EMR is the dominant form of wireless communication. One alternative is the use of light (such as in fiber optics), but RF is much more versatile because lower-frequency EMR is not blocked by opaque objects.

Learning to Live in the Frequency Domain

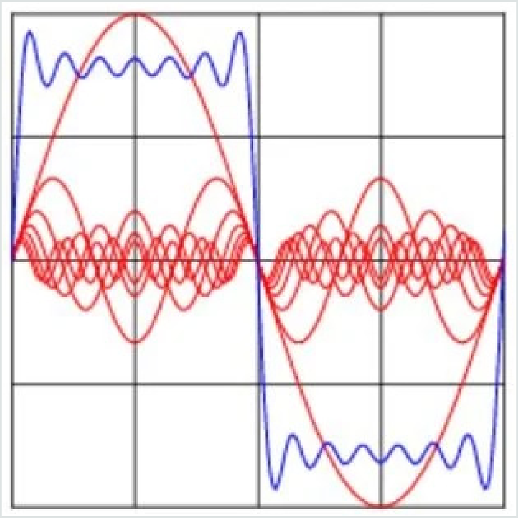
What is the frequency domain? And why is it so valuable for RF design, analysis, and testing?

Perhaps one of the most fundamental steps in the process of gaining proficiency in RF design is learning to think in the *frequency domain*. For most of us, the vast majority of our early experience with electrical circuits and signals remains within the context of voltages and currents that are either static or dynamic with respect to time. For example, when we measure the voltage of a battery with a multimeter, we have a static quantity, and when we look at a sinusoidal voltage on an oscilloscope, we have a time-varying quantity.

RF, on the other hand, is a world of frequencies. We do not send static voltages to antennas, and the oscilloscope is usually not an effective tool for capturing and visualizing the types of signal manipulation that are involved in wireless communication. Indeed, we can say that the time domain is simply not a convenient place for the design and analysis of RF systems. We need a different paradigm.

Fourier

The Fourier transform is the mathematical path that leads to this alternative paradigm, because it provides a precise method of describing a signal according to its *frequency content.*



This diagram shows some of the frequency content (red) in a square wave (blue).

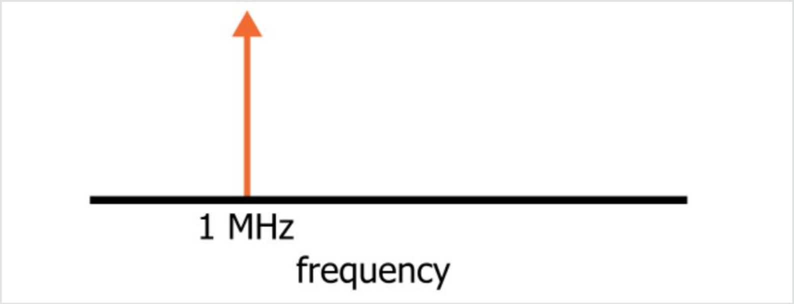
In the context of RF, the Fourier transform can take extremely complex signal variations and translate them into frequency-domain components that are far more informative than the original time-domain waveform.

The details involved in computing the Fourier transform or the discrete Fourier transform (DFT) are not trivial; however, this is not something we have to worry about at this point. You can understand and employ frequency-domain techniques even if you know very little about the underlying mathematical procedures.

The Fourier transform produces expressions that reveal a signal's frequency content, and the DFT produces corresponding numerical data. However, in the context of practical engineering, a graphical representation is often much more convenient. Eventually these frequency-domain plots become as normal and intuitive as an oscilloscope trace.

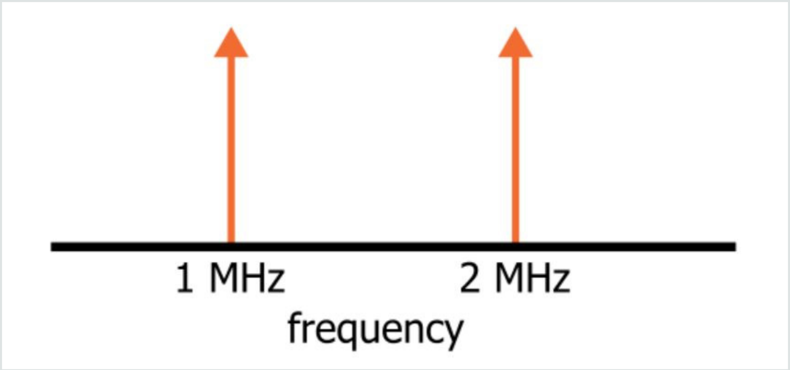
The "Spectrum"

A frequency-domain plot is referred to as a spectrum. The idealized spectrum for a 1 MHz sinusoid is as follows:

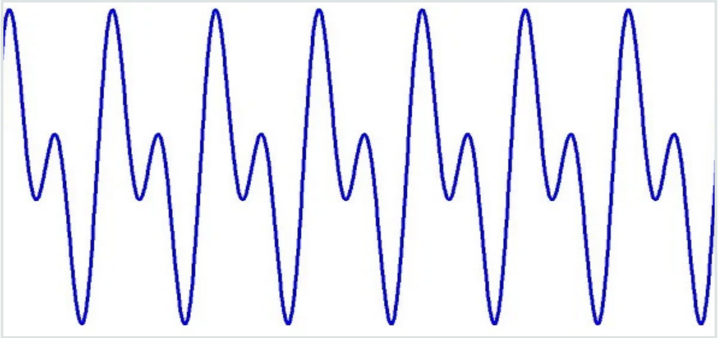


The vertical arrow indicates that a certain amount of "energy" is present at 1 MHz. The line portion of the arrow is so thin because this idealized signal has absolutely no other frequency components—all the energy is concentrated exactly at 1 MHz.

If we used a summing circuit to combine this perfect 1 MHz sinusoid with a perfect 2 MHz sinusoid, the spectrum would be the following:



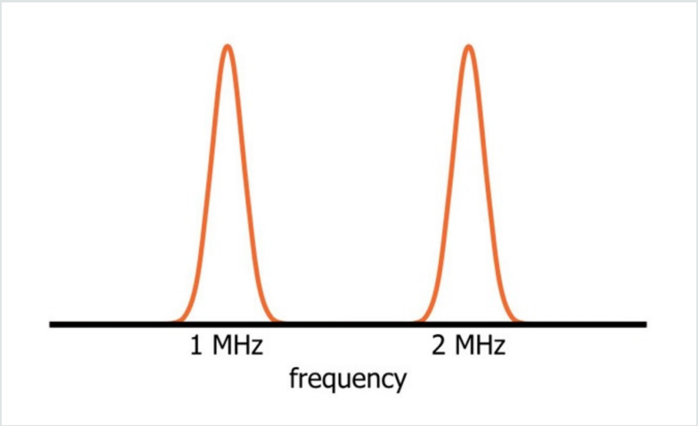
This frequency-domain plot provides very clear data regarding the frequency characteristics of our new signal. If you are primarily interested in the non-instantaneous frequency-related behavior of your circuit, the spectrum gives you the information you need. In contrast, the time-domain waveform is not straightforward:



It is far less obvious that this trace is the result of adding one sinusoidal quantity of frequency *f* to a second sinusoidal quantity of frequency 2*f*.

Ideal vs. Real

The thin-vertical-arrow frequency components shown above are mathematical constructs; real-world spectral measurements look more like this:



Why the discrepancy? First of all, the resolution of the measurement system is limited, and such limitations inherently compromise whatever “ideal” qualities might be present in the original signal. But even if we had an infinitely accurate measuring device, the spectrum would differ from the mathematical version because of noise.

The only type of signal that could produce the “pure” spectral components shown in the previous section is a perfect sinusoid—i.e., no noise and no variations in period or amplitude. Any deviation from the characteristics of a perfect sinusoid would introduce additional frequency components.

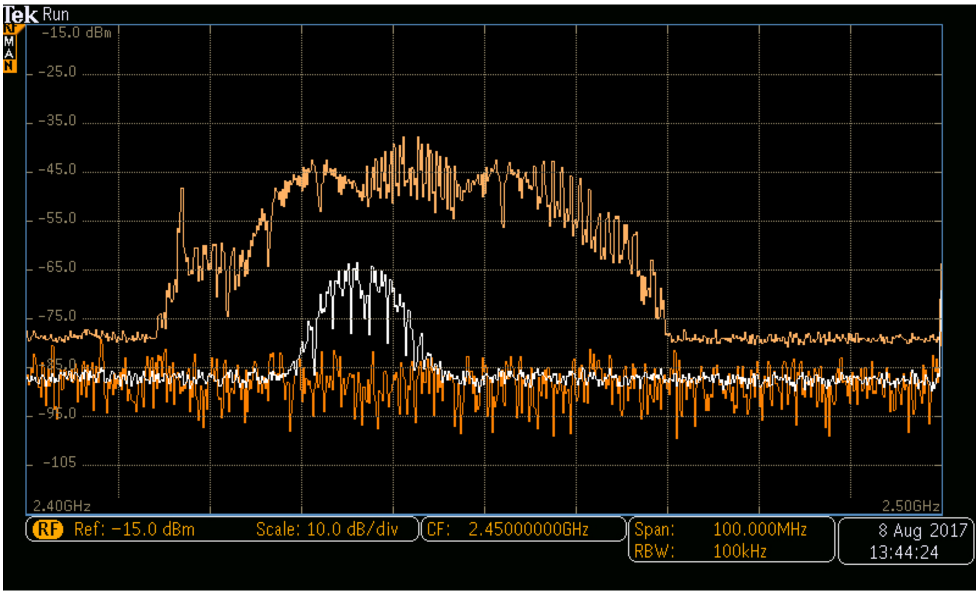
An intuitive example is phase noise: It is not practical to expect a real-world oscillator to always produce the exact same frequency; inevitably there will be (hopefully small) variations in the actual duration of a cycle, and this is called phase noise. If you collect data covering one thousand cycles and then perform spectral analysis, you are effectively averaging the frequency content of those one thousand cycles. The result will be the spectral shape shown above; the width of the waveform corresponds to the averaged deviation from the nominal frequency.

Spectral Measurements

Frequency-domain plots provide a very convenient means of discussing and analyzing RF systems. Modulation schemes, interference, harmonic distortion—even basic spectra drawn on a piece of scratch paper can really help to clarify a situation.

But we'll generally need something more sophisticated when it comes time to successfully design an RF system. More specifically, we need something that gives us the spectral characteristics of an actual signal. This is important for characterizing the functionality of an existing system, but usually the more pressing need is diagnosis and resolution—i.e., why is this device not working, and how can we fix it.

Digital oscilloscopes offer "FFT" (fast Fourier transform) functionality, and this is one way to obtain spectral measurements. However, the tool of choice for real-world frequency analysis is called a spectrum analyzer. This is a piece of test equipment that is specifically designed to accept a high-frequency input signal and display the frequency-domain representation of this signal. Acquiring a bit of hands-on experience with a spectrum analyzer is an important initial step in becoming familiar with practical aspects of RF engineering.



Ref: -15.0 dBm Scale: 10.0 dB/div CF: 2.45000000GHz Span: 100.000MHz RBW: 100kHz 8 Aug 2017 13:44:24

Summary

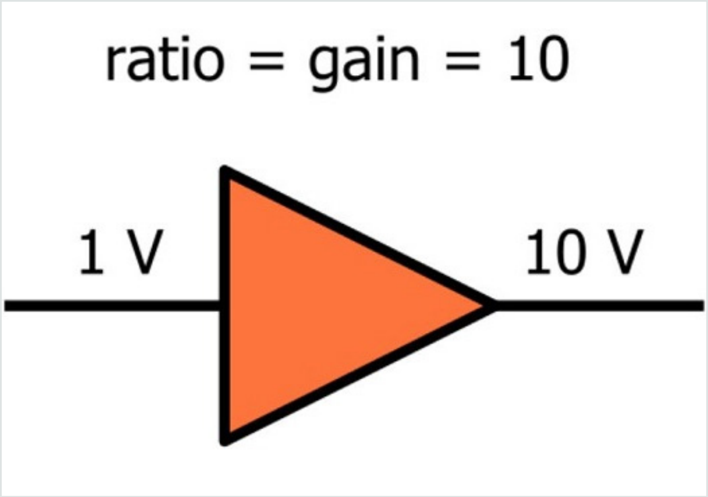
- Engineers can interact with electrical signals via the time domain or the frequency domain. In the context of RF, it is generally more productive and intuitive to work in the frequency domain.
- Frequency-domain analysis naturally suppresses details that are often of little importance in RF design and testing, and at the same time it emphasizes the characteristics that we need to focus on.
- A frequency-domain plot is referred to as a spectrum. A spectrum can conveniently convey the salient characteristics of, for example, a modulation scheme or an actual signal that is experiencing problems caused by interference.
- Theoretical spectra often consist of thin vertical arrows that correspond to idealized fixed-frequency sinusoids.
- Real-world measurement equipment and real-world RF signals are always subject to imperfections that result in a wider frequency-domain waveform.
- An essential piece of equipment for an RF design lab is the spectrum analyzer. These devices provide frequency-domain plots as well as various signal-analysis capabilities.

The RF Engineer's Guide to the Decibel

Learn about the decibel and its variants in the context of RF design and testing.

RF engineering, like all scientific disciplines and subdisciplines, involves quite a bit of specialized terminology. One of the most important words that you will need when working in the world of RF is "dB" (and some variants thereof). If you become deeply entrenched in an RF project, you may find that the word "dB" becomes as familiar to you as your own name.

As you probably know, dB stands for decibel. It's a logarithmic unit that provides a convenient way of referring to ratios, such as the ratio between the amplitudes of an input signal and an output signal.



We won't cover the generic details of decibels because they are already available on this page of the AAC Electric Circuits textbook. Instead, we will focus on practical aspects of the decibel in the specific context of RF systems.

Relative, Not Absolute

It is easy to forget that dB is a *relative* unit. You cannot say, "The output power is 10 dB."

Voltage is an absolute measurement because we always speak of a potential difference, i.e., the difference in potential between two points; usually we are referring to the potential of one node with respect to a 0 V ground node. Current is also an absolute measurement because the unit (amperes) involves a specific amount of charge with respect to a specific amount of time. In contrast, dB is a unit that involves the logarithm of a ratio between two numbers. A straightforward example is amplifier gain: If the power of the input signal is 1 W and the power of the output signal is 5 W, we have a ratio of 5:

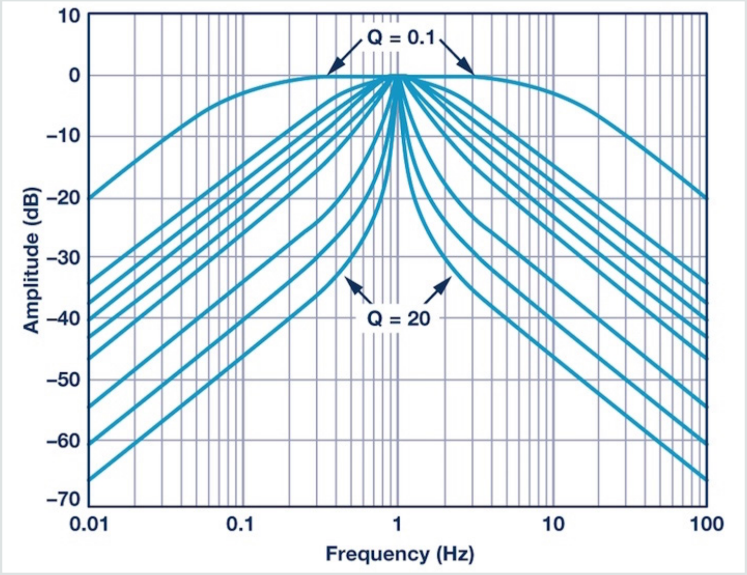
$$10\log_{10}\left(\frac{P_{OUT}}{P_{IN}}\right) = 10\log_{10}(5) \approx 7\ dB$$

Thus, this amplifier provides 7 dB of power gain—i.e., the ratio between the output signal strength and the input signal strength can be expressed as 7 dB.

Why dB?

It would certainly be possible to design and test RF systems without the use of dB, but in practice dB’s are everywhere. One advantage is that the dB scale allows us to express very large ratios without using very large numbers: a power gain of 1,000,000 is only 60 dB. Also, the total gain or loss of a signal chain is easily computed in the dB domain, because the individual dB figures are simply added (whereas multiplication would be required if we were working with ordinary ratios).

Another advantage is something that we’re familiar with from our experience with filters. RF systems revolve around frequencies and the various ways in which frequencies are generated, controlled, or affected by components and parasitic circuit elements. The dB scale is convenient in a context such as this because frequency response plots are intuitive and visually informative when the frequency axis uses a logarithmic scale and the amplitude axis uses a dB scale.



A Bode plot showing the magnitude response of different band-pass filters. Image courtesy of AnalogDialogue.

When dB Is Absolute

We've established that dB is a ratio and thus cannot describe the absolute power or amplitude of a signal. However, it would be awkward to be constantly switching back and forth between dB and non-dB values, and perhaps this is why RF engineers developed the dBm unit.

We can avoid the "ratios only" problem by simply creating a new unit that always includes a reference value. In the case of dBm, the reference value is 1 mW. Thus, if we have a 5 mW signal and we want to stay within the realm of dB, we can describe this signal as having a power of 7 dBm:

$$10\log_{10}\left(\frac{5\ mW}{1\ mW}\right) = 10\log_{10}(5) \approx 7\ dBm$$

You definitely want to familiarize yourself with the concept of dBm. This is a standard unit used in real-life RF system development, and it's very convenient when, for example, you are calculating a link budget, because gains and losses expressed in dB can simply be added to or subtracted from the output power expressed in dBm.

There is also a dBW unit; this uses 1 W for the reference value instead of 1 mW. Nowadays most RF engineers are working with relatively low-power systems, and this probably explains why dBm is more common.

More dB Variants

Two other dB-based units are dBc and dBi.

Instead of a fixed value such as 1 mW, dBc uses the strength of the carrier signal as the reference. For example, phase noise (discussed in page 2 of this chapter) is reported in units of dBc/Hz; the first part of this unit indicates that the phase noise power at a specific frequency is being measured with respect to the power of the carrier (in this case "carrier" refers to the signal strength at the nominal frequency).

An idealized point-source antenna receives a certain amount of energy from the transmitter circuit and radiates it equally in all directions. These "isotropic" antennas are considered to have zero gain and zero loss.

Other antennas, however, can be designed to concentrate radiated energy in certain directions, and in this sense an antenna can have "gain." The antenna is not actually adding power to the signal, but it *effectively* increases the transmitted power by concentrating electromagnetic radiation according to the orientation of the communication system (obviously this is more practical when the antenna designer knows the spatial relationship between the transmitter and receiver).

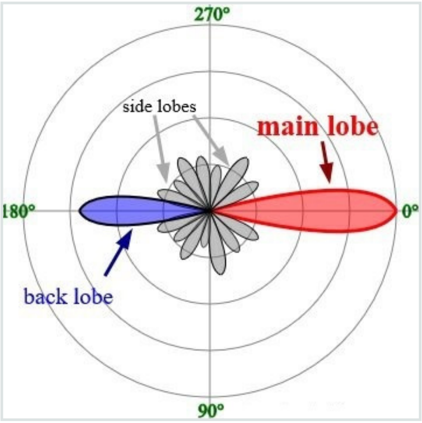


Image from Timothy Truckle (own work) [GFDL]. Here you can see the unequal distribution of radiated energy that results in gain in the forward direction (i.e., 0°).

The dBi unit allows antenna manufacturers to specify a “gain” figure that uses the ever-popular dB scale. As always, we need a ratio when we’re working with dB, and in the case of dBi, the antenna gain is given with reference to the gain of an isotropic antenna.

Some antennas (such as those accompanied by a parabolic dish) have significant amounts of gain, and thus they can make a nontrivial contribution to the range or performance of an RF system.

Summary

- The dB scale is a method of expressing ratios between two quantities. It is convenient and widely used in the context of RF design and testing.
- Though dB figures are inherently relative, absolute quantities can be expressed via the dB scale by using units that incorporate a standardized reference value.
- The most common absolute dB unit is dBm; it conveys the dB power of a signal with respect to 1 mW.
- The dBc unit expresses power with respect to the power of a related signal.
- The dBi unit expresses the gain of an antenna relative to the response of an idealized point-source antenna.

Passive Components in RF Circuits

Resistors, capacitors, temperature-compensated oscillators. . . . Learn about passive components used in RF systems.

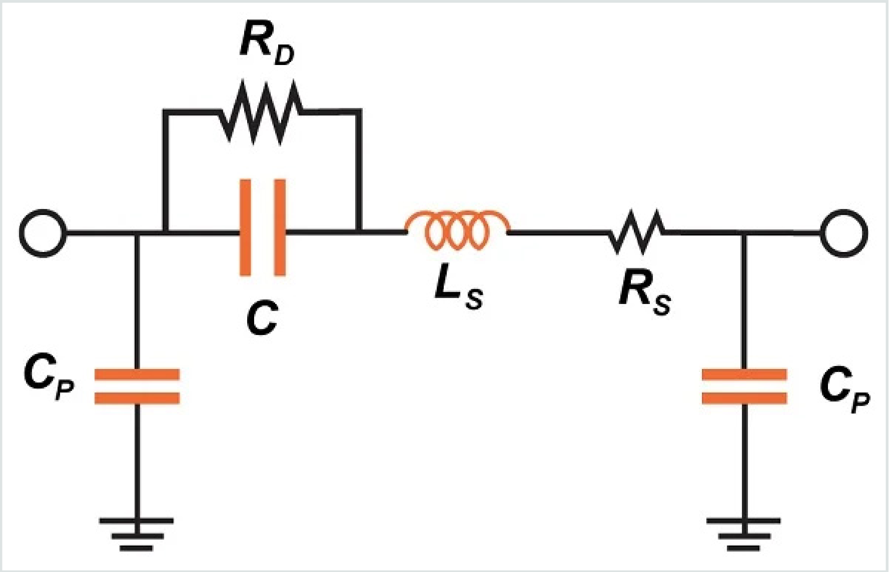
RF systems are not fundamentally different from other types of electric circuits. The same laws of physics apply, and consequently the basic components used in RF designs are also found in digital circuits and low-frequency analog circuits.

However, RF design involves a unique set of challenges and objectives, and consequently the characteristics and uses of components call for special consideration when we are operating in the context of RF. Also, some integrated circuits perform functionality that is highly specific to RF systems—they are not used in low-frequency circuits and may not be well understood by those who have little experience with RF design techniques.

We often categorize components as either active or passive, and this approach is equally valid in the realm of RF. This page discusses passive components specifically in relation to RF circuits, and the next page covers active components.

Capacitors

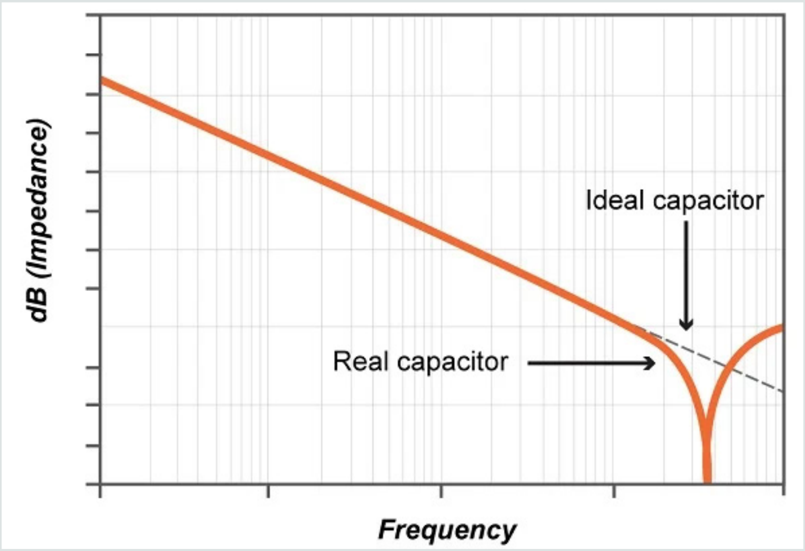
An ideal capacitor would provide exactly the same functionality for a 1 Hz signal and a 1 GHz signal. But components are never ideal, and the nonidealities of a capacitor can be quite significant at high frequencies.



A model representing the real electrical behavior of a capacitor.

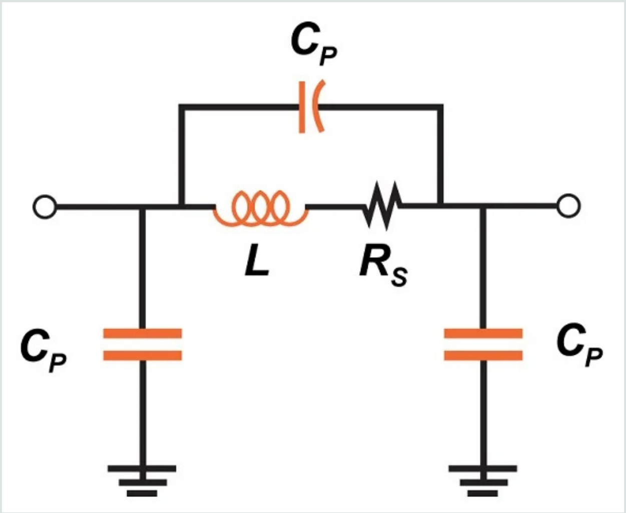
"C" corresponds to the ideal capacitor that is buried among so many parasitic elements. We have non-infinite resistance between the plates (R_D), series resistance (R_S), series inductance (L_S), and parallel capacitance (C_P) between the PCB pads and the ground plane (we're assuming surface-mount components; more on this later).

The most significant nonideality when we're working with high-frequency signals is the inductance. We expect the impedance of a capacitor to endlessly decrease as frequency increases, but the presence of the parasitic inductance causes the impedance to dip down at the self-resonant frequency and then begin to increase:



Inductors

The following is an equivalent circuit for an inductor:



A model representing the real electrical behavior of an inductor.

An ideal inductor would provide impedance that steadily increases as frequency increases, but the parallel capacitor eventually dominates the response, and the result is impedance that decreases as frequency increases. So we can see that both capacitors and inductors must be chosen carefully when they will be used in RF circuits, especially RF circuits with frequencies well above 1 GHz.

Resistors, et al.

Even resistors can be troublesome at high frequencies, because they have series inductance, parallel capacitance, and the typical capacitance associated with PCB pads.

And this brings up an important point: when you're working with high frequencies, parasitic circuit elements are everywhere. No matter how simple or ideal a resistive element is, it still needs to be packaged and soldered to a PCB, and the result is parasitics. The same applies to any other component: if it's packaged and soldered to the board, parasitic elements are present.

Crystals

The essence of RF is manipulating high-frequency signals so that they convey information, but before we manipulate we need to generate. As in other types of circuits, crystals are a fundamental means of generating a stable frequency reference.

However, in digital and mixed-signal design, it is often the case that crystal-based circuits actually do not require the precision that a crystal can provide, and consequently it's easy to become careless with regard to crystal selection. An RF circuit, in contrast, may have strict frequency requirements, and this calls for not only initial frequency precision but also frequency stability.

The oscillation frequency of an ordinary crystal is sensitive to temperature variations. The resulting frequency instability creates problems for RF systems, especially systems that will be exposed to large variations in ambient temperature. Thus, a system may require a TCXO, i.e., a temperature-compensated crystal oscillator. These devices incorporate circuitry that compensates for the crystal's frequency variations:

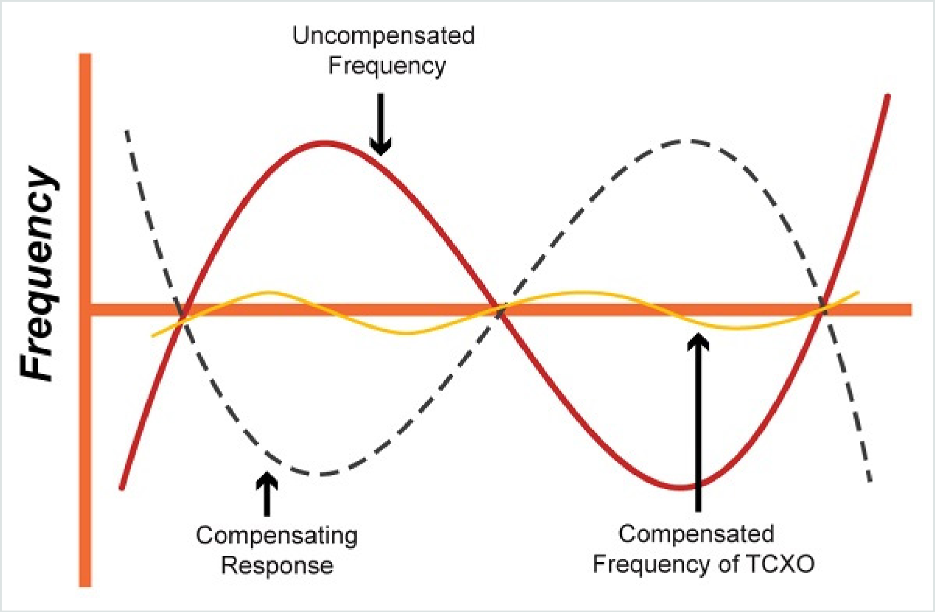
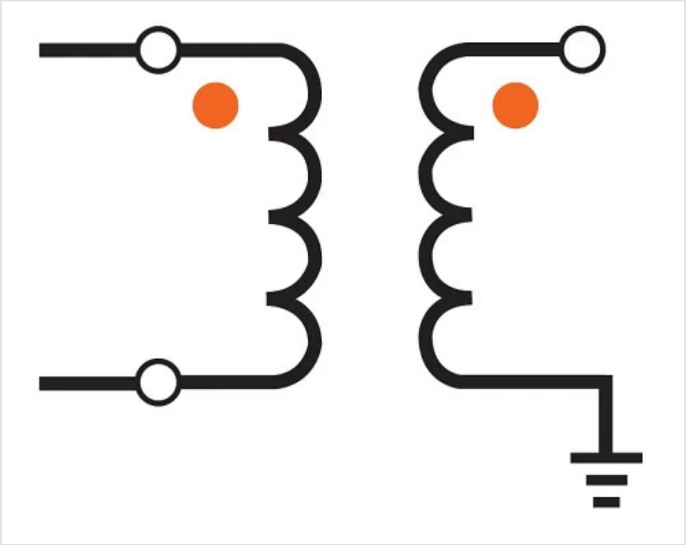


Image adapted from Crystek

Baluns

One component that is common in RF systems but rare elsewhere is the balun. The name comes from “balanced to unbalanced,” a phrase which helps us to remember that baluns are used to convert differential (i.e., balanced) signals to single-ended (i.e., unbalanced) signals, or single-ended to differential.

Baluns fall within the general category of transformers, as you can see from the circuit symbol:



A balun. In this case, the signal connected to the left-hand terminals is differential and the signal connected to the right-hand terminal is single-ended.

In lower-frequency circuits we often use amplifiers to convert between single-ended and differential, whereas in RF baluns are common. Why the difference? Well, the explanation is related to a fact that influences many RF design decisions: simple passive components are often more practical than IC-based equivalents when you're dealing with very high frequencies.

Antennas

An antenna is a passive component that is used to convert an RF electrical signal into electromagnetic radiation (EMR), or vice versa. With other components and conductors we try to minimize the effects of EMR, and with antennas we try to optimize the generation or reception of EMR with respect to the needs of the application.

Antenna science is by no means simple. Various factors influence the process of choosing or designing an antenna that is optimal for a particular application. AAC has two articles (click here and here) that provide an excellent introduction to antenna concepts.

Higher frequencies are accompanied by various design challenges, though the antenna portion of the system can actually become less problematic as frequency increases, because higher frequencies allow for the use of shorter antennas. Nowadays it is common to use either a "chip antenna," which is soldered to a PCB like typical surface-mount components, or a PCB antenna, which is created by incorporating a specially designed trace into the PCB layout.

Surface Mount vs. Through Hole

Earlier I referred to how the equivalent circuits assume that we're using surface-mount components. Through-hole components are by no means unsuitable for RF, but it's important to understand that surface-mount packaging is inherently superior when you're working with high-frequency signals.

Surface-mount technology brings various advantages, but in this case we're talking specifically about inductance: We want to minimize parasitic inductance in high-frequency circuits. Longer leads have more inductance, and consequently surface-mount packaging is preferred.

Summary

- Some components are common only in RF applications, and others must be chosen and implemented more carefully because of their nonideal high-frequency behavior.

- Passive components exhibit nonideal frequency response as a result of parasitic inductance and capacitance.
- RF applications may require crystals that are more accurate and/or stable than crystals commonly used in digital circuits.
- Baluns allow for high-frequency conversion between single-ended and differential signaling.
- Antennas are critical components that must be chosen according to the characteristics and requirements of an RF system.

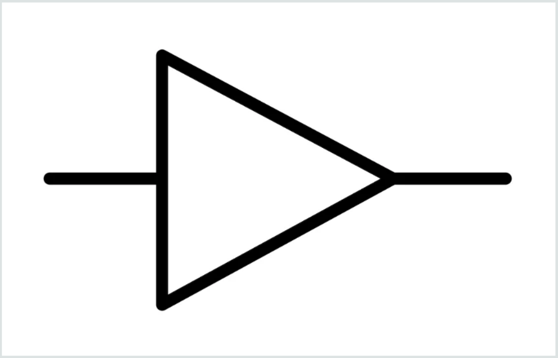
Active Components in RF Circuits

LNAs, PAs, mixers. . . . Learn about active components used in RF systems.

As with passive components, the active components used in RF circuits share many characteristics with active components typically found in lower-frequency analog systems. However, there are certain components that are highly specific to RF design. Furthermore, different semiconductor technologies are often employed to ensure that RF components maintain adequate performance at very high frequencies.

Amplifiers

Amplifier circuits, often built around an operational amplifier, are extremely common in both low-frequency and high-frequency analog design. In RF systems, there are two fundamental types of amplifiers: power amplifiers and low-noise amplifiers. The former are used to increase the power level of an RF signal prior to transmission, and the latter are used to amplify the (often very small) signals received by the antenna.



Power Amplifiers

The power amplifier, or PA, is used to increase the power level of the signal before it is sent to the antenna. A similar situation is found in audio circuits: the audio signal’s amplitude may be perfectly adequate in terms of voltage, but a power amplifier is needed to supply large amounts of current to the speaker coil. In audio, more current corresponds to more power, and this in turn corresponds to more volume. In RF, higher power means longer range.

Low-Noise Amplifiers

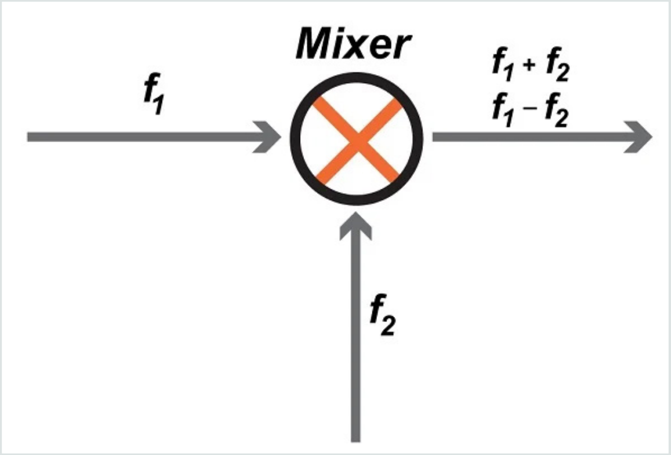
There are many non-RF applications that require low-noise amplification, but the specific phrase “low-noise amplifier” is common only in the context of RF. Actually, we usually hear the abbreviated version of the term, i.e., LNA.

The received signal delivered by an antenna can be of very low magnitude, and furthermore, it is buried in noise. This signal needs to be amplified for further processing, but it is also important to minimize further degradation of the signal-to-noise ratio. Thus, a low-noise amplifier is designed to provide high voltage gain while contributing minimal noise.

The noise performance of an LNA is quantified via the "noise figure" (NF), which corresponds to the amount of SNR degradation (in dB) created by the amplifier. Thus, an ideal amplifier would have NF = 0 dB, and as noise performance declines, NF increases.

Mixers

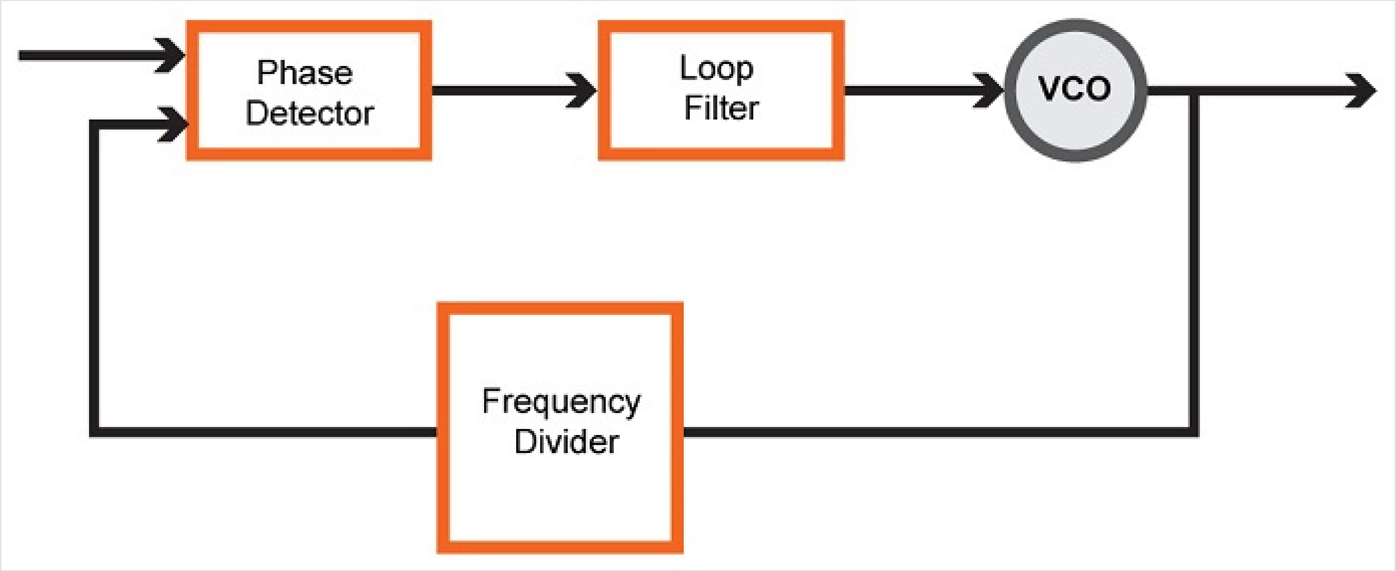
Another fundamental RF component is the mixer. This name can be misleading; an RF mixer does not *combine* signals as an audio mixer does. Rather, an RF mixer takes two input frequencies and generates a third output frequency via multiplication. In other words, a mixer performs frequency *translation.*



Mixers allow signals to be shifted to higher or lower frequencies in a way that maintains the details of the signal. For example, an information-carrying (i.e., modulated) baseband signal can be shifted to a higher frequency that is suitable for wireless transmission, and the transmitted signal will retain the important modulation details that were present in the baseband signal.

Phase-Locked Loops

The actual generation of a periodic signal is more closely related to the domain of passive components, but active components are used to manipulate these periodic signals. A phase-locked loop (PLL) is actually a system of subcomponents—at minimum, a phase detector, a low-pass filter, a voltage-controlled oscillator (VCO), and a frequency divider—that allows a wide variety of output frequencies to be generated from one input frequency.



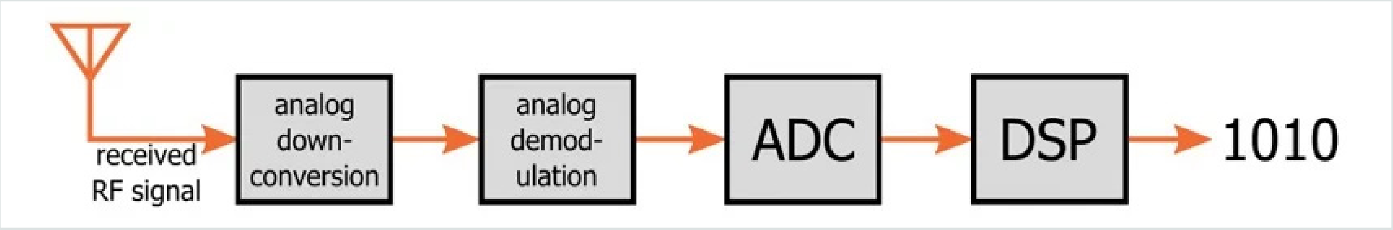
Combining a PLL with a high-precision temperature-compensated oscillator converts a highly accurate but *fixed* reference frequency into a system that can produce highly accurate yet *variable* output frequencies. An oscillator combined with a PLL is referred to as a synthesizer, i.e., a component that can generate a range of frequencies.

This ability to adjust the oscillator frequency is very important in RF design. A particular system may need to operate on different channels in order to avoid interference, and thus the oscillation circuitry must be adjustable with respect to frequency. Furthermore, the frequency spacing between adjacent channels may be relatively small, and thus the adjustments must be precise.

Data Converters

Though not standard components in the context of historical RF engineering, it is important to recognize that analog-to-digital converters (ADCs) and digital-to-analog converters (DACs) are increasingly important in many RF systems. ADCs and DACs allow RF systems to benefit from the special capabilities offered by digital-signal-processing techniques and from the general flexibility and convenience associated with software-based solutions.

The term “software-defined radio” (SDR) refers to wireless communication systems that rely on software to implement important portions of the RF signal chain. Data converters are critical components in such systems—for example, a DAC could be used to directly generate a baseband waveform, or an ADC could be used to digitize a received baseband waveform (followed by further analysis in a digital signal processor).



An example of an SDR receive path.

SDRs can introduce additional design complexity, but they also offer advantages that are particularly valuable in certain applications.

RF Semiconductors

Silicon is still the dominant material in semiconductor manufacturing. However, other materials are more compatible with the high signal frequencies present in RF systems. Three alternative materials that are used in RF semiconductors are gallium nitride (GaN), gallium arsenide (GaAs), and silicon germanium (SiGe). Specialized semiconductor technologies make it possible to fabricate devices that maintain adequate performance at extremely high frequencies, i.e., above 100 GHz.

Inside the IC

As with low-frequency devices, the fundamental active component in RF integrated circuits is the transistor. However, thus far we have used the word “component” to refer to devices that may consist of numerous transistors. It is important to understand the justification for this: Designing high-performance, high-frequency RF components is extremely challenging and not within the skill set of many RF engineers. Practical RF engineering is focused on combining these components into functional circuits and then dealing with the various complicated issues that arise.

Summary

- Active components intended for RF systems may offer specialized functionality, or they may offer standard functionality but with greater ability to maintain performance at high frequencies.
- An RF amplifier is generally categorized as a power amplifier (PA) or a low-noise amplifier (LNA). The former provides power gain in preparation for transmission, and the latter provides high voltage gain and low noise figure.

- RF mixers perform frequency translation by multiplying two input signals.
- A phase-locked loop (PLL) can be combined with an oscillator to produce a frequency synthesizer.
- ADCs and DACs are important components in some RF devices. They are increasingly common in modern wireless systems, and they are essential in software-defined radios.
- SiGe, GaAs, and GaN are specialized semiconductor materials that are superior to silicon in high-performance RF applications.

CHAPTER 2

The Electromagnetic Spectrum

- The Many Frequencies of RF Communication
- RF Transmission: Regulations, Interference, and Power Transfer
- Low-Power RF Devices and the ISM Bands

The Many Frequencies of RF Communication

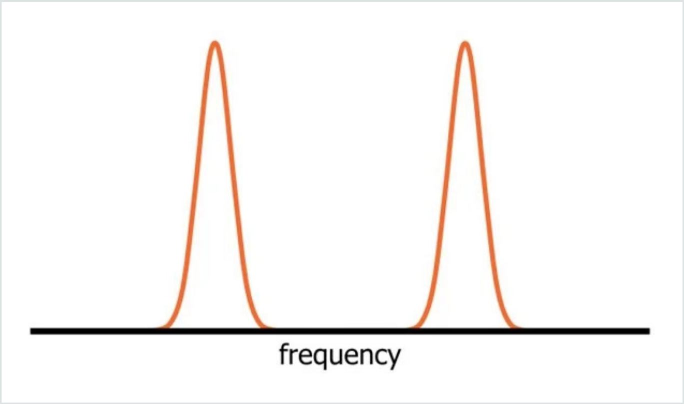
There is only one electromagnetic spectrum, but by using different carrier frequencies, numerous RF devices can coexist.

The world of RF is a world of frequencies. This is true within a single system or even a single PCB, considering that one RF design may involve signals in multiple frequency ranges. But at this point we want to look at the broad context in which a particular RF system exists; the name we give to this concept is “the electromagnetic spectrum.”

More specifically, we will discuss the portion of the electromagnetic spectrum that is commonly used for RF communication. Light is included in the electromagnetic spectrum, and so are extremely-low-frequency radio waves that have limited use in engineered systems. Light is a useful means of transmitting information, but it behaves very differently from medium-frequency electromagnetic radiation (EMR), and consequently we place it in its own category—*optical* communication instead of *wireless* communication. Low-frequency EMR has specialized uses and is also generated constantly all over the world by the power grid, but it is not a part of mainstream wireless communication.

Frequencies: Why and How

Before we discuss the various frequency categories, let’s review two fundamental issues: Why do we use so many different frequencies? And how does a designer decide which frequency is appropriate for a certain application?



Adequate separation between frequencies allows an interfering signal to be ignored.

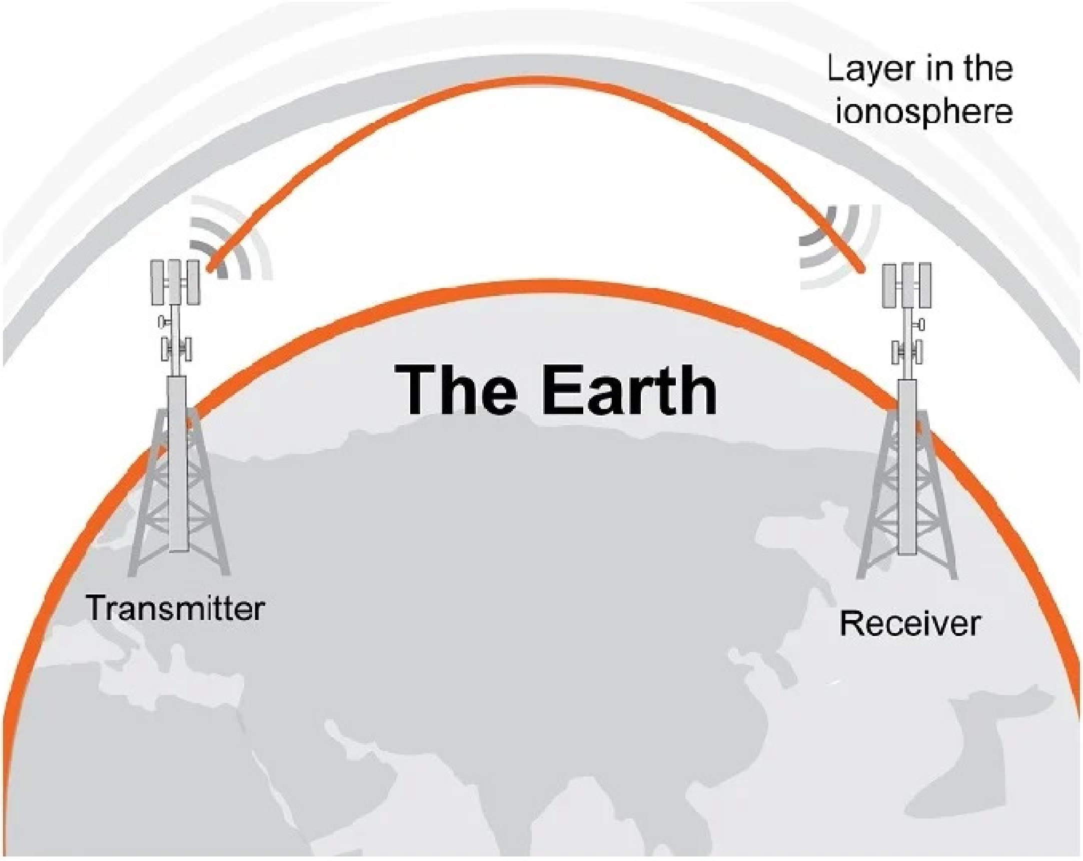
Interference

Two or more transmitters operating at the same frequency create interference, i.e., they make it difficult for a receiver device to separate the relevant RF signal from irrelevant RF signals. This problem largely disappears when different frequencies are used. EMR at one frequency does not “corrupt” EMR at a different frequency, and the irrelevant signals are easily ignored via filtering.

Of course, interference doesn't disappear just because two signals are separated by a fraction of a hertz—more frequency separation leads to less interference. Nevertheless, the use of different frequencies for different types of RF communication is amazingly effective: every day, all over the world, numerous wireless systems operate simultaneously with no significant loss of functionality.

Choosing a Frequency

The characteristics of EMR vary according to frequency. For example, extremely-low-frequency waves can effectively penetrate water and thus can be helpful when you need to communicate with a submarine. As another example, certain frequencies enable a radio signal to travel very long distances because these frequencies experience atmospheric refraction. The point is, the dominant objectives of a particular RF system heavily influence the process of choosing the operational frequency range.



Ionospheric refraction enables long-range communication.

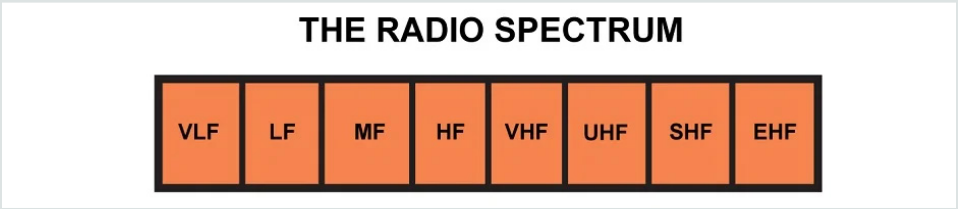
The previous paragraph mentioned examples in which frequency affects propagation characteristics. Often, though, a more important consideration is bandwidth (in analog systems) or data rate (in digital systems).

If you want to wirelessly transmit an audio signal that has frequency components as high as 10 kHz, you cannot use a 5 kHz transmitter (i.e., carrier) frequency. Frequency corresponds to the rate at which a signal can transmit information, so you cannot “fit” 10 kHz of audio information into a 5 kHz carrier. Furthermore, practical considerations require the carrier frequency to be significantly higher than the information (i.e., baseband) frequency. Thus, wider-bandwidth and higher-data-rate systems must occupy higher-frequency portions of the electromagnetic spectrum.

Frequencies of Interest

The radio spectrum—i.e., the radio-communication portion of the electromagnetic spectrum—extends from the VLF (very-low-frequency) band to the EHF (extremely-high-frequency) band, i.e., from about 3 kHz to 300 GHz. The other bands that separate VLF from EHF are:

- LF (low frequency),
- MF (medium frequency),
- HF (high frequency),
- VHF (very high frequency),
- UHF (ultra high frequency), and
- SHF (super high frequency).



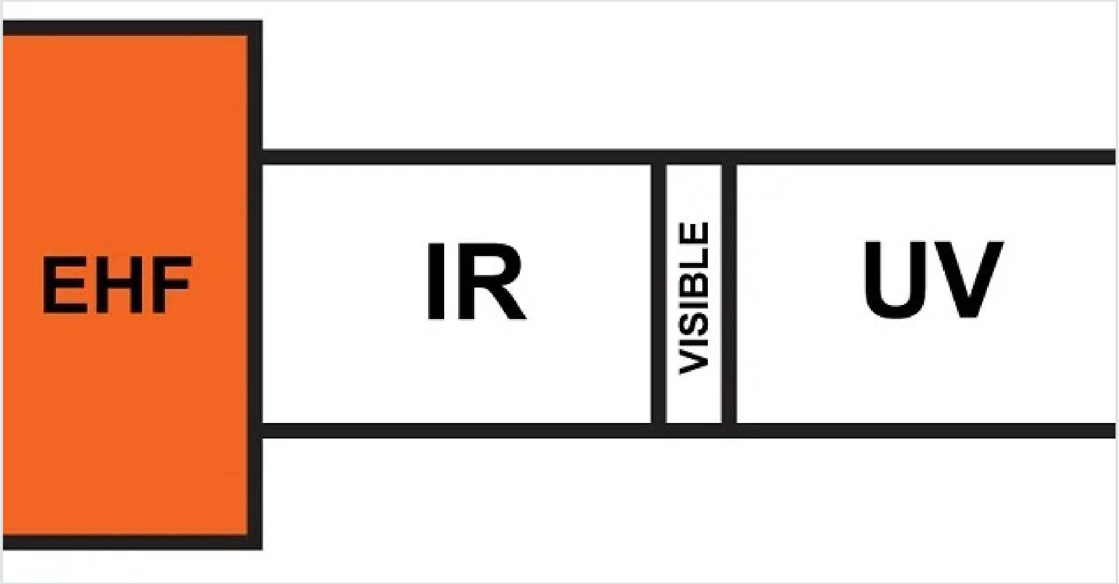
These divisions are rather arbitrary and there is no dire need to know the exact frequency ranges. It would be better to simply give some examples of wireless-communication categories that are found in different portions of the spectrum, because this will help us gain an intuitive awareness of which frequency ranges are more appropriate for certain types of systems.

- AM radio communication uses the MF band; more specifically, the carrier frequencies vary from 540 to 1600 kHz. We know from experience that AM radio has good range and is resistant to physical interference from buildings, but AM does not have a reputation for excellent audio quality.

- FM radio communication uses the VHF band, with carrier frequencies from 88.1 to 108.1 MHz. The allowable deviation from the carrier is significantly higher in FM than in AM, which means that FM signals can transfer more information per unit time than AM signals. (Keep in mind that in this context "AM" and "FM" refer to standardized categories of radio transmission, not to amplitude and frequency modulation in general.)
- Digital communication systems such as Bluetooth and some of the 802.11 protocols operate in the low-gigahertz range, more specifically, at frequencies near 2.4 GHz. These are generally short-range systems, but they offer reliable communication and the high carrier frequency enables high data rates. These protocols can be used by devices that are very small yet provide relatively long battery life.
- Satellites—obviously representing an application in which long range is important—tend to operate at very high frequencies. At the lower end of this range (1–2 GHz) is the L band, which is used by GPS satellites. The C band (4–8 GHz) is used, for example, by satellite TV networks. The Ku band, which extends to the impressive frequency of 18 GHz, is employed for various satellite applications and is an important part of the communication equipment on the International Space Station.

From EMR to Light

The satellite frequencies mentioned above mostly remain within the SHF section of the radio spectrum. The EHF band serves as the transition between radio waves and optical waves; EHF signals are more seriously obstructed by the gases and moisture in the atmosphere, and this reminds us of optical radiation and its inability to penetrate opaque objects. Signals with frequencies above those of the EHF band are classified as infrared radiation, not as radio waves:



Summary

- The electromagnetic spectrum refers to the range of EMR frequencies present in the universe. This spectrum is divided and subdivided into different frequency bands.
- The general section that is relevant to RF communication is referred to as the radio spectrum, and the radio spectrum is divided into eight bands.
- Interference among separate radio systems can be avoided by using different carrier frequencies.
- Bandwidth and propagation requirements influence the choice of carrier frequency, and in turn the carrier frequency influences the characteristics of a particular system.
- The highest-frequency band within the radio spectrum represents the transition from signals that behave more like radio waves to signals that behave more like optical waves.

RF Transmission: Regulations, Interference, and Power Transfer

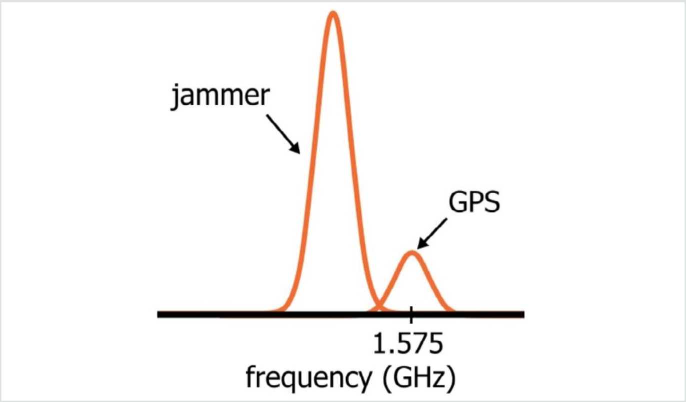
Learn about how to transfer maximum power from your amplifier to your antenna, and how to estimate this power using an oscilloscope.

An important characteristic of RF technology is the following: it is relatively easy for one person to impede, or even thoroughly ruin, another person's wireless communication. Radio waves travel through the air and are available to everyone, including those who—intentionally or accidentally—are transmitting signals that could be described as *interference.*

First, it's important to understand that you cannot "destroy" or "damage" radio signals that have already been transmitted. Nonetheless, the effect of interference can be equivalent to destroying an original signal because it compromises the receiver's ability to extract the important information contained in this signal. In other words, the information is still present, but *with respect to a particular receiver* it has, in practice, ceased to exist.

Interference is a constant challenge in RF design, and the proliferation of wireless devices is not making the situation any simpler. There are various ways of making a system resistant to interference, and these will be discussed later in the textbook. Most of this interference is simply due to the fact that non-communicating devices must often utilize similar carrier frequencies.

However, there is also such a thing as deliberate interference. This is called *jamming;* the goal is to broadcast a signal that in one way or another prevents other wireless systems from maintaining successful communication. Jamming is an important tactic in modern warfare, and in daily life it's a nuisance (or worse) and is completely illegal.



This spectrum depicts a strong signal intended to jam a GPS device.

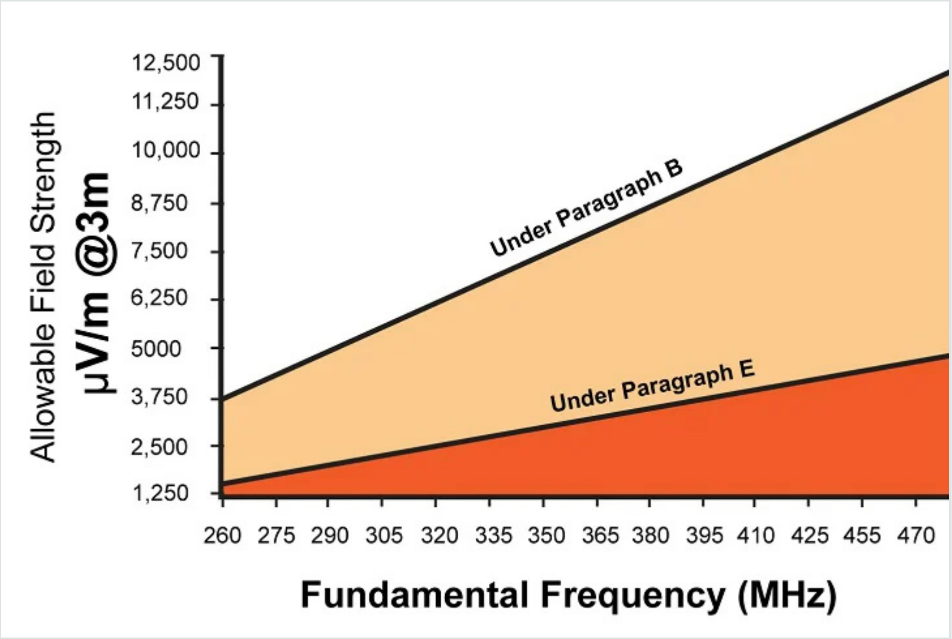
Regulations

It may initially seem strange that the government would regulate wireless transmissions—can we really impose laws on something as intangible as electromagnetic radiation? But the jamming example makes it clear that the absence of regulations would lead to serious problems. Strict organization is required to ensure that the realm of EMR does not deteriorate into a chaotic horde of interfering signals.

In the United States, the task of maintaining order in the world of wireless communication falls to the Federal Communications Commission (FCC). Private and public organizations that want to utilize a portion of the electromagnetic spectrum must obtain permission from the FCC; this permission is referred to as a license. There are exceptions for systems that are limited in range and thus unlikely to cause a major disturbance.

Max Power

If you are interested in (legal) license-free radio transmissions, you need to know your transmit power. Even if the official regulations are presented in terms of effective range or some other metric, you should be able to determine the transmit power that is generally considered acceptable in these situations—and estimating power is easier than trying to accurately measure the system's range or the field strength at a certain distance from the antenna.



This plot gives field-strength limits (for a specific range of frequencies) based on the FCC's "Part 15" rules. Image adapted from Digi-Key

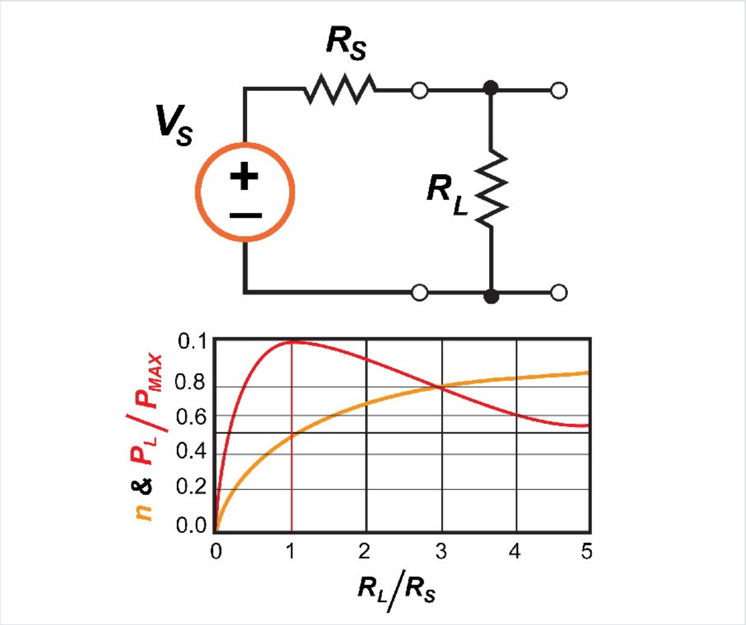
In RF and all other types of electric circuits, the power dissipated by a component is equal to the voltage across that component multiplied by the current flowing through the component. You may think of an antenna as simply a conductor and therefore as something with very little resistance. It's true that a conductor can have very low resistance at DC, but at higher frequencies an antenna has significant amounts of input impedance. We're interested in the impedance of the antenna at the specific frequencies that we are using to transmit our RF signal; we will need this information to estimate the amount of power delivered to the antenna.

Voltage Transfer vs. Power Transfer

In a typical digital or analog circuit, we wouldn't want a wire or PCB trace to have a resistance of 50 Ω. This seems like an awfully high resistance for something described as a conductor. But we have to remember that in low-frequency circuits we are typically interested in voltage transfer, i.e., we want to ensure that the voltage at an input pin is as close as possible to the voltage at the preceding output pin. To achieve good voltage transfer, we need low output impedance, low conductor impedance, and high input impedance.

However, in the output stage of an RF transmitter (or of an audio amplifier), the goal is power transfer. We don't just want to move voltage from one device to another; we want significant current flowing through the antenna, so that it has plenty of *electrical* energy that can be converted into radiated *electromagnetic* energy.

Maximum power transfer occurs when the magnitude of the load impedance is equal to the magnitude of the source impedance.



As you can see, the load power (P_L) is maximum when $R_L = R_S$. Notice, though, that efficiency (η) continues to increase beyond this point. Maximum power transfer does not correspond to maximum efficiency.

In RF circuitry, the amplifier’s output stage (and the transmission line that connects the amplifier to the antenna) will often have impedance of 50 Ω, and thus the antenna impedance must also be 50 Ω to ensure maximum power transfer. (Another important topic here is “matching networks,” which are used to improve impedance matching between an amplifier and an antenna; this will be discussed later in the textbook.)

Estimating Power

The preceding discussion explains why we can analyze an RF output stage by connecting the power amplifier to a 50 Ω oscilloscope input: most RF systems are built around 50 Ω impedances, and thus you will generally need a 50 Ω antenna impedance.

Of course, if you know the relevant voltage and impedance characteristics of your circuit, you can simply calculate the power delivered to the antenna. A SPICE simulator would be another effective approach. But if these techniques are not practical in your circumstances, or if you want empirical verification, you need to use measurement equipment.

If you have a spectrum analyzer, by all means use it. It is designed to provide exactly this sort of information. If you don’t have a spectrum analyzer, you can use an oscilloscope. Look at the RMS voltage of the signal using a 50 Ω scope input, and then calculate power as V^2 / R, where R = 50 Ω.

Summary

- Electromagnetic transmission is carefully regulated to mitigate problems associated with unintentional interference. Intentional interference, known as jamming, is illegal in the context of civilian life.
- In the United States, transmitting devices generally must be licensed by the FCC.
- License-free operation is possible under certain conditions associated with restricted transmit power.
- To achieve maximum transfer of electrical power from an amplifier to an antenna, the magnitude of the amplifier’s output impedance must match the magnitude of the antenna’s input impedance.
- Transmit power can be determined via mathematical analysis or SPICE simulation. It can also be estimated empirically using a spectrum analyzer or an oscilloscope.

Low-Power RF Devices and the ISM Bands

Digital modulation can help improve the reliability of low-power RF communication. And what exactly is an ISM band?

When considered from a historical perspective, RF systems are closely associated with high-power transmission. We imagine large antennas for AM and FM stations, long-distance military radios, and even exotic applications such as the systems used to communicate with and control spacecraft. These systems are associated with a vague idea that longer range is better, and therefore more power is better.

High-power RF is by no means unimportant or rare, but in many ways it is increasingly separated from our daily lives. Or at least we can say that it is less *noticeable* in our daily lives, because so much of our attention is now focused on small, low-power wireless devices.

Bluetooth products are examples of the battery-powered wireless devices that are increasingly common in modern life.

In systems such as these, extreme design effort is devoted to achieving acceptable performance at the lowest possible power consumption. This means that efficiency may be more important than maximum power transfer, and it also means that there may be no desire to achieve maximum range. The goal is simply to achieve *adequate* range, i.e., range that allows the device to be used for its intended purpose.

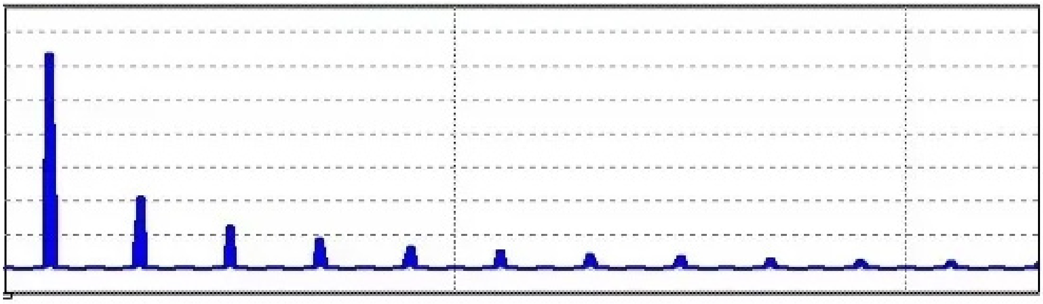
An interesting example involves hearing aids. It should come as no surprise that the human body's sensory system is designed to work with two ears; the human brain refines our ability to experience and react to sound by combining these two related sensory streams (presumably in rather complex ways).

Wearing hearing aids in both ears can help to restore this balanced perception of sound, but modern devices go a step further by actually communicating with the hearing aid in the other ear. In this way, the two hearing aids can “work together” to fine-tune their response.

This is a perfect example of an RF system that does not need to maximize range. The designers know almost exactly how much distance will separate the transmitter and receiver, and there is no realistic situation in which it would be beneficial to have longer range.

Digital vs. Analog

An important technique in low-power RF systems is digital modulation. This does not refer to actually transmitting digital (i.e., rectangular) signals; though this is not impossible, it is impractical because a rectangular wave has high harmonic content. The transmitted signal contains large amounts of energy at frequencies quite far away from the carrier frequency, and consequently it would be a source of interference.

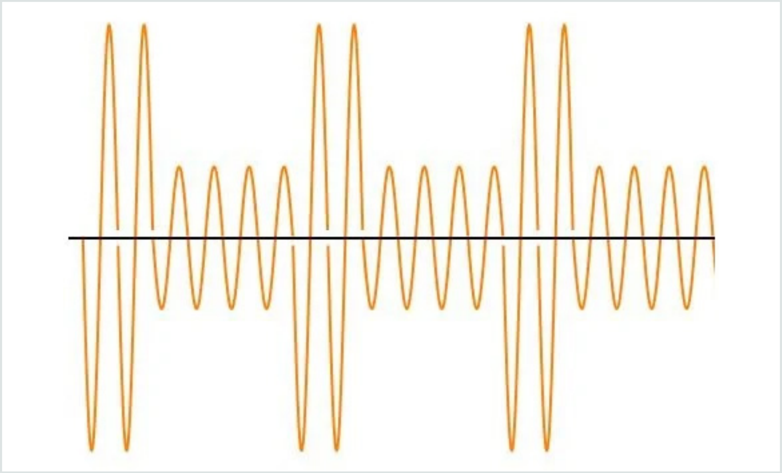


The spectrum of a square wave; there is too much energy at the harmonic frequencies.

As discussed in the previous page, the electromagnetic spectrum must remain organized to ensure that numerous unrelated devices can reliably implement wireless communication. This means that wireless transmissions must be restricted to a specific allocated frequency range, and this is not possible when using rectangular signals.

Digital Modulation

Digital modulation, then, uses sinusoidal waves, just as analog modulation does. The difference is that in a digital system the modulation of the carrier does not represent a continuous representation of the analog baseband signal. Instead, it represents digital data. The changes in the carrier wave occur in discrete sections referred to as symbols, and each symbol represents one or more bits. We will discuss digital modulation in more depth later in this textbook.



An example of digital modulation—in this case, amplitude modulation.

Digital modulation provides benefits analogous to those of typical digital communication. Because information is transferred as discrete bits instead of a continuously varying signal, transmit power can be minimized with very little loss of data—as long as the power is sufficient to enable the receiver to distinguish between a zero and a one, all of the data will be transferred successfully. Also, digital communication allows the receiver to ask the transmitter to resend specific sections of data, if, for example, transient interference caused a brief reduction in signal-to-noise ratio.

Digital RF systems, often referred to as data links, have the additional advantage of being able to evaluate their own performance in real time. An error-detection algorithm, such as a cyclic redundancy check, can be used to assess the quality of the connection. If the receiving device notices a significant increase in the frequency of bit errors, it can ask the transmitter to increase its output power. In this way the transmitter's power consumption can be optimized based on the actual performance of the data link.

The ISM Bands

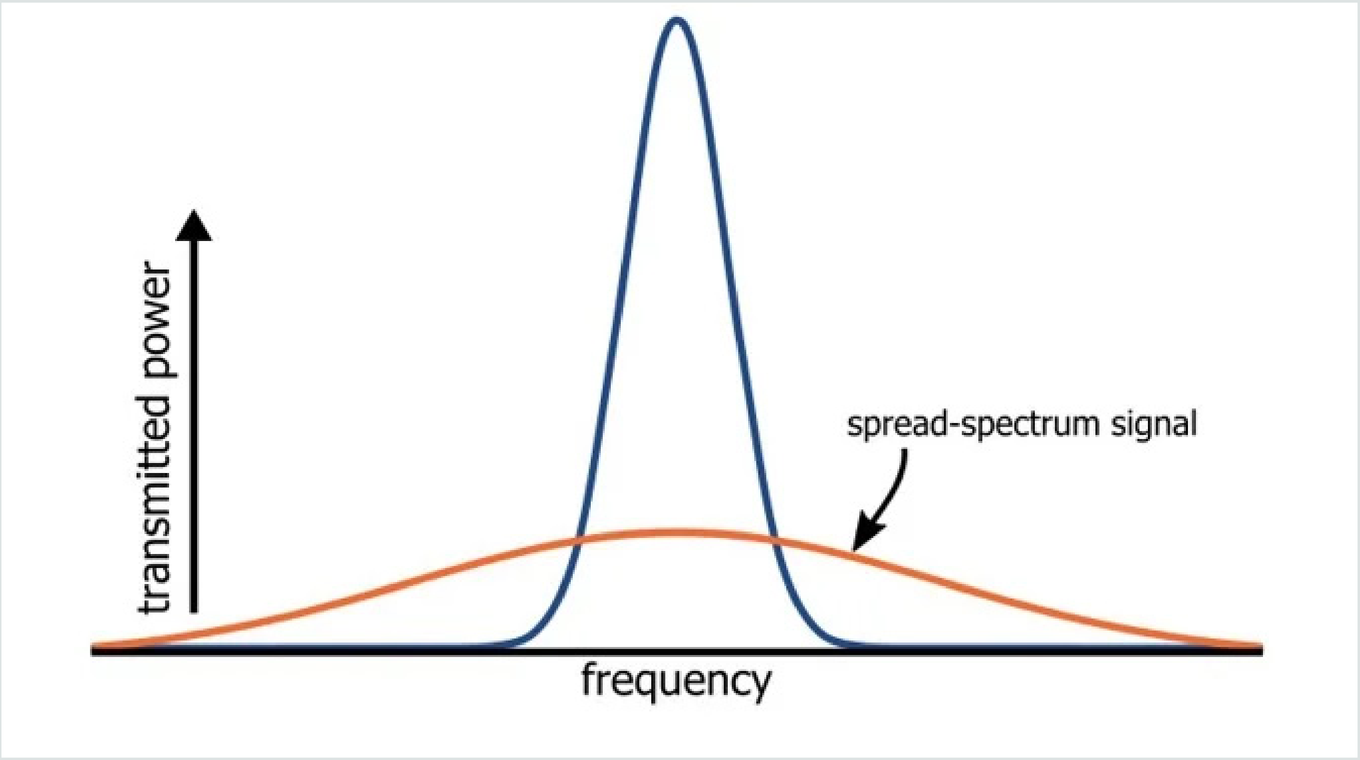
As discussed in the previous page, any organization that wants to operate an RF transmitter must obtain explicit permission from the proper regulatory agency (such as the FCC in the United States). The most notable exception to this rule is the use of the ISM bands.

ISM stands for industrial, scientific, and medical. Presumably, this reflects the original intention of the FCC, but the name is no longer relevant. The ISM bands are used by numerous devices from other product categories—Bluetooth, Wi-Fi, home security systems, radio-frequency identification (RFID), toys, cordless phones. . . .

Unlicensed vs. Unregulated

The ISM bands are *unlicensed*, but they are emphatically not unregulated. “Unlicensed” means that it is legal to develop and market an ISM-band device without obtaining explicit permission from a regulatory body. “Unregulated” would imply that you can transmit anything you want as long as you stay within the ISM frequencies, and this is not the case.

The most straightforward limitation is that of transmit power: in general, the power delivered to the antenna cannot exceed 1 W (30 dBm). However, the situation becomes more complicated when you get into details such as frequency hopping or spread-spectrum transmission.



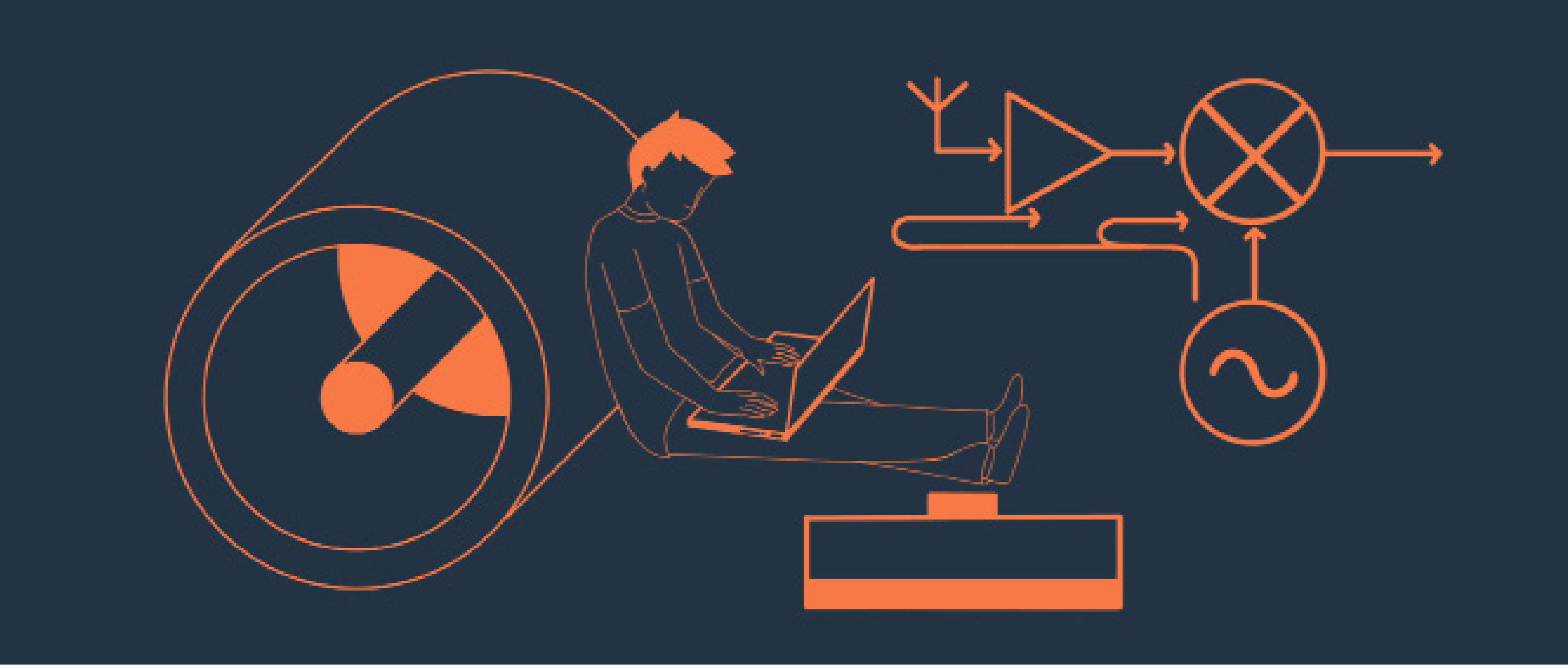
Spread-spectrum modulation; this will be discussed later in the textbook.

Also, there are restrictions on out-of-band transmitted energy—this is relevant because low-order harmonics can result in significant transmitted energy that falls outside the acceptable range of frequencies.

The most important ISM band is referred to as the 2.4 GHz band, though 2.4 GHz is actually not the center frequency; the band extends from 2.4 to 2.4835 GHz. A major advantage of this band is its worldwide availability—other ISM bands vary from one region to another, but 2.4 GHz is available for unlicensed operation all over the world.

Summary

- Low-power RF devices are increasingly common in our daily lives. Apart from the general interest in conserving energy, low-power operation increases battery life.
- Digital data transfer is an important technique in many RF systems; in low-power systems it allows for a more efficient use of battery power.
- Digital modulation refers to the use of analog waveforms to transfer digital data.
- The ISM bands are the most significant exception to typical RF licensing requirements. Numerous wireless devices utilize ISM frequencies.
- ISM-band devices do not require a license, but they must comply with the regulations that govern these bands.



CHAPTER 3

Real-Life RF Signals

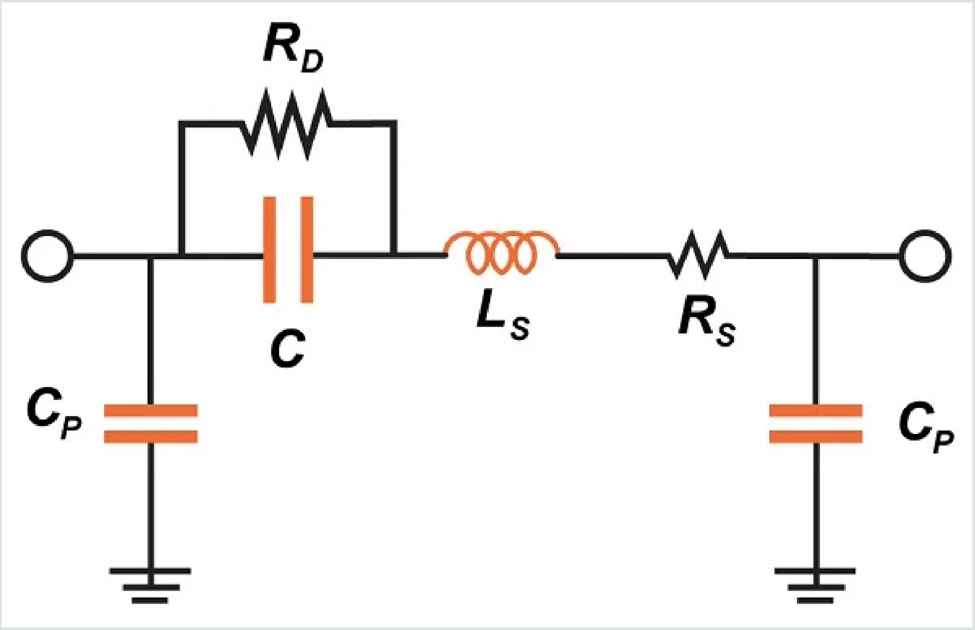
- Coupling and Leakage in RF Systems
- What Is a Transmission Line?
- Understanding Reflections and Standing Waves in RF Circuit Design
- The 50 Ω Question: Impedance Matching in RF Design

Coupling and Leakage in RF Systems

RF design and analysis requires an understanding of the complex ways in which high-frequency signals move through a real circuit.

RF design is known to be particularly challenging among the various subdisciplines of electrical engineering. One reason for this is the extreme inconsistency between theoretical electrical signals and high-frequency sinusoidal signals.

At some point we all start to realize that the idealized components and wires and signals found in theoretical circuit analysis are helpful though highly inaccurate approximations of reality. Components have tolerances and temperature dependencies and parasitic elements; wires have resistance, capacitance, and inductance; signals have noise. However, numerous successful circuits are designed and implemented with little if any consideration for these nonidealities.



The equivalent circuit model for a real "capacitor"; at very high frequencies it actually behaves like an inductor.

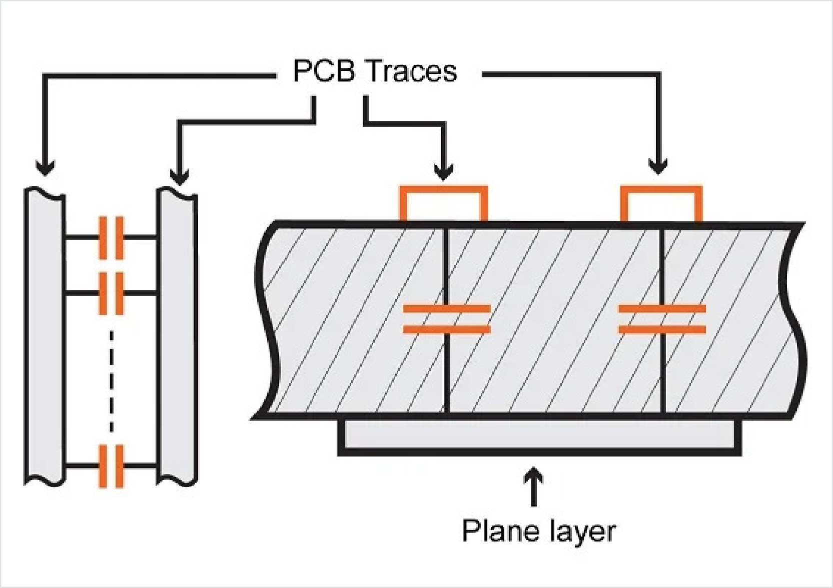
This is possible because so many circuits these days involve primarily low-frequency or digital signals. Low-frequency systems are much less subject to nonideal signal and component behavior; consequently, low-frequency circuits tend to diverge much less from the operation that we expect based on theoretical analysis. High-frequency digital systems are more subject to nonidealities, but the effects of these nonidealities are usually not prominent because digital communication is inherently robust.

A digital signal may experience significant degradation as a result of nonideal circuit behavior, but as long as the receiver can still correctly distinguish logic high from logic low, the system maintains full functionality.

In the RF world, of course, signals are neither digital nor of low frequency. Unexpected signal behavior becomes the norm, and every dB of reduced signal-to-noise ratio corresponds to reduced range, or lower audio quality, or increased bit error rate.

Capacitive Coupling

It is essential to understand that RF signals absolutely do not confine themselves to the intended conduction paths. This is particularly true in the context of printed circuit boards, where the various traces and components often have little physical separation.



Examples of parasitic capacitance.

A typical circuit diagram consists of components, wires, and the empty space in between. The assumption is that signals travel along wires and cannot pass through the empty space. In reality, though, those empty spaces are filled with capacitors. Capacitance is formed whenever two conductors are separated by an insulating material, with closer physical proximity corresponding to higher capacitance.

Capacitors block DC and present high impedance to low-frequency signals. Thus, we can more or less ignore all this unintended capacitance in the context of low-frequency design. But impedance decreases as frequency increases; at very high frequencies, a PCB is filled with relatively low-impedance conduction paths created by parasitic capacitance.

Radiated Coupling

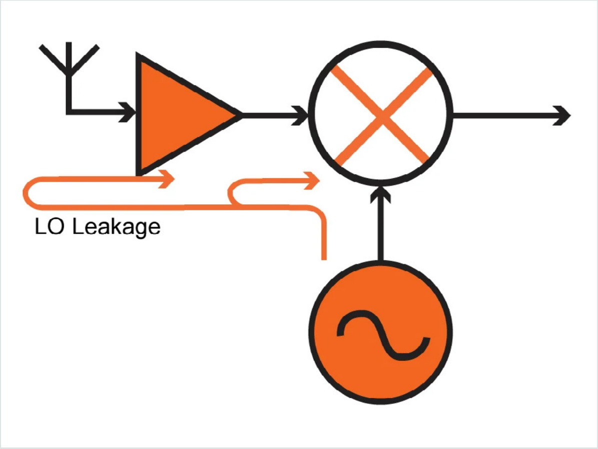
In the idealized world, every RF device has one antenna. In reality, every conductor is an antenna in the sense that it is capable of emitting and receiving electromagnetic radiation. Thus, radiated coupling provides another means by which RF signals can pass through the supposedly nonconductive empty spaces between schematic symbols.

As usual, this problem becomes more serious as frequency increases. An antenna is more effective when its length is a significant fraction of the signal wavelength, and thus PCB traces (which are usually rather short) are more problematic when high frequencies are present.

The term “radiated coupling” is more appropriate when referring to far-field effects, i.e., interference caused by electromagnetic radiation that is not in the immediate vicinity of the antenna. When the emitting and receiving conductors are separated by less than approximately one wavelength, the interaction is occurring in the near field. In this situation the magnetic field dominates, and consequently the more accurate term is “inductive coupling.”

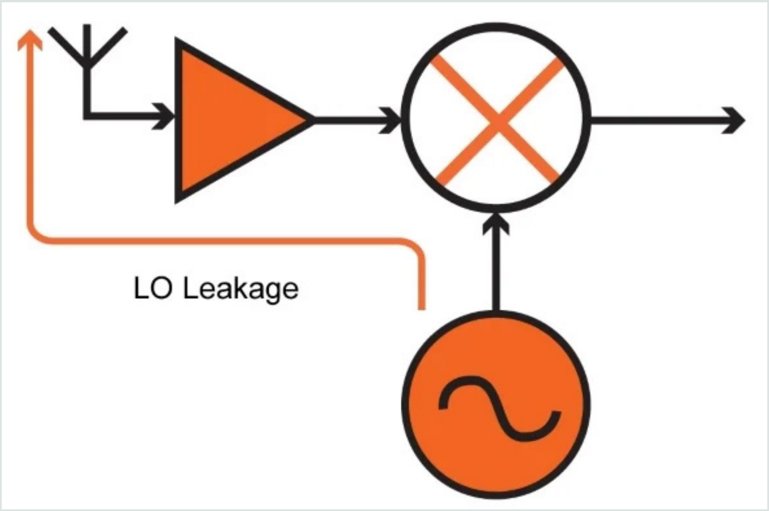
Leakage

An RF signal that is coupling into unwanted portions of a circuit is described as “leaking.” A classic example of leakage is depicted in the following diagram:



The local oscillator (LO) signal is fed directly to the LO input of the mixer; this is the intentional conduction path. At the same time, the signal finds an unintentional conduction path and manages to leak into the mixer’s other input port. Mixing two signals of identical frequency and phase results in a DC offset (the magnitude of the offset decreases toward zero as the phase difference approaches 90° or –90°). This DC offset constitutes a major design challenge with respect to receiver architectures that translate the input signal directly from the radio frequency to the baseband frequency.

Another leakage path is from a mixer through a low-noise amplifier to the antenna:



But it doesn't stop there; the LO signal could be radiated by the antenna, reflected by an external object, and then received by the same antenna. This would again produce self-mixing and the resulting DC offset, but in this case the offset would be highly unpredictable—the amplitude and polarity of the offset would be affected by the constantly changing magnitude of the reflected signal.

Transmitters and Receivers

Another situation that leads to leakage problems is when an RF device includes both a receiver and a transmitter. The transmitter portion has a power amplifier that is designed to send a strong signal to the antenna. The receiver portion is designed to amplify and demodulate signals of very small amplitude. So the transmitter provides high power, and the receiver provides high sensitivity.

You can probably see where this is going. A coupling path could allow the PA output to leak into the receive chain; even a highly attenuated PA signal could cause problems for the sensitive receiver circuitry.

Simplex, Duplex

This PA-to-receiver leakage is only a concern when the circuit must support simultaneous transmission and reception. A system composed of two such devices—called transceivers, because they can function as **trans**mitters and re**ceivers**—is referred to as full duplex. A full-duplex system enables simultaneous two-way communication.

A half-duplex system supports only non-simultaneous two-way communication, though the devices used in a half-duplex system are still transceivers because they can transmit and receive. With half-duplex devices we don’t have to worry about leakage from the PA to the receiver because the receive chain is not active during transmissions.

A one-way RF communication system is referred to as “simplex.” A very common example is AM or FM broadcasting; the station’s antenna transmits, and the car radio receives.

Summary

- Real-life electrical signals and components are more difficult to predict and analyze than their idealized counterparts; this is especially true for high-frequency analog signals.
- RF signals readily travel through unintended conduction paths created by capacitive coupling, radiated coupling, and inductive coupling.
- The movement of RF signals through unintended conduction paths is referred to as leakage.
- RF systems can be divided into three general categories
    - full duplex (simultaneous two-way communication)
    - half duplex (non-simultaneous two-way communication)
    - simplex (one-way communication)

What Is a Transmission Line?

High-frequency interconnects require special consideration because they often behave not as ordinary wires but rather as transmission lines.

In low-frequency systems, components are connected by wires or PCB traces. The resistance of these conductive elements is low enough to be negligible in most situations.

This aspect of circuit design and analysis changes dramatically as frequency increases. RF signals do not travel along wires or PCB traces in the straightforward fashion that we expect based on our experience with low-frequency circuits.

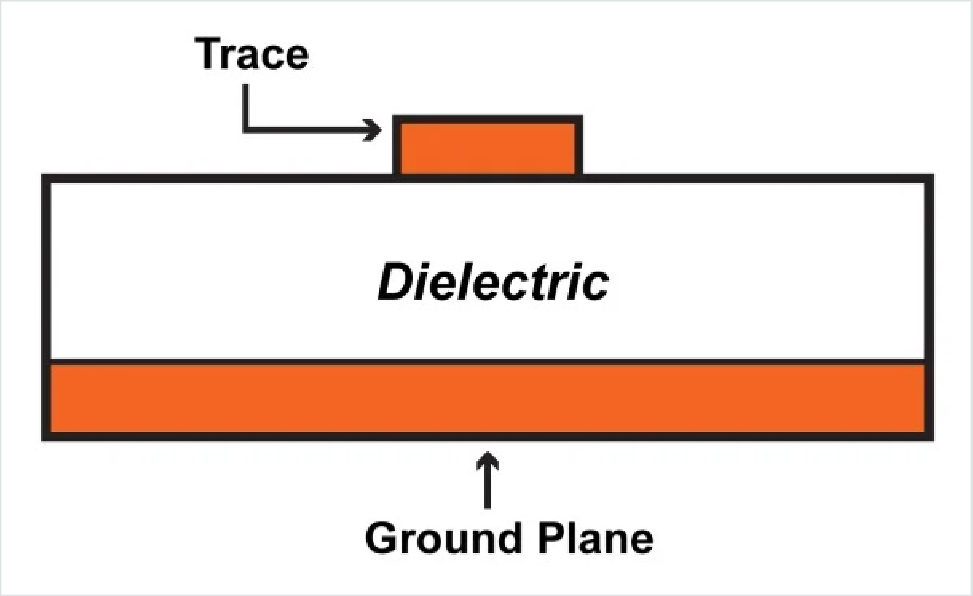
The Transmission Line

The behavior of RF interconnects is very different from that of ordinary wires carrying low-frequency signals—so different, in fact, that additional terminology is used: *a transmission line* is a cable (or simply a pair of conductors) that must be analyzed according to the characteristics of high-frequency signal propagation.

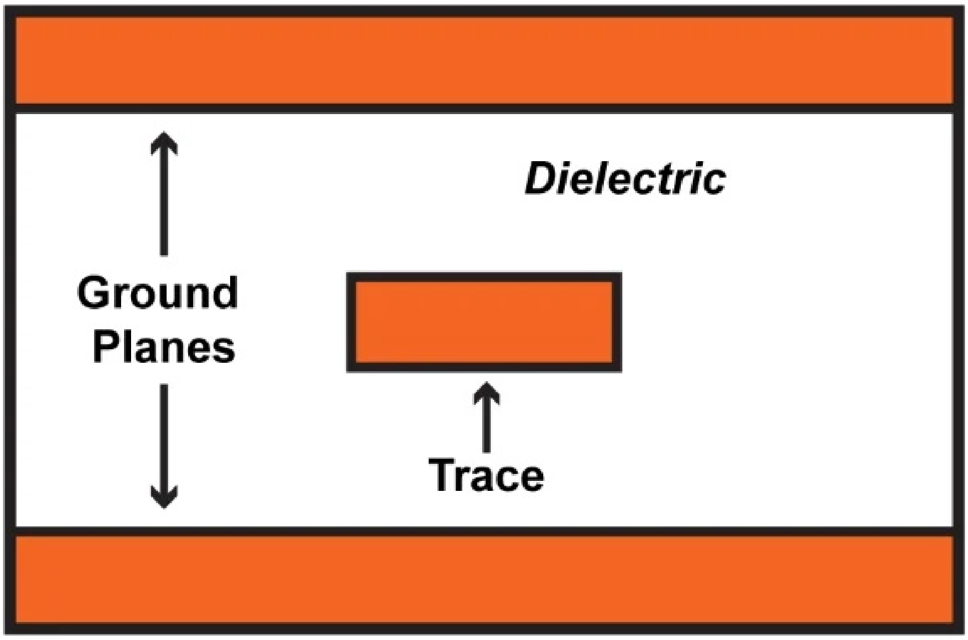
First, let's clarify two things:

Cable vs. Trace

"Cable" is a convenient but imprecise word in this context. The coaxial cable is certainly a classic example of a transmission line, but PCB traces also function as transmission lines. The "microstrip" transmission line consists of a trace and a nearby ground plane, as follows:



The “stripline” transmission line consists of a PCB trace and two ground planes:



PCB transmission lines are particularly important because their characteristics are controlled directly by the designer. When we buy a cable, its physical properties are fixed; we simply gather the necessary information from the datasheet. When laying out an RF PCB, we can easily customize the dimensions—and thus the electrical characteristics—of the transmission line according to the needs of the application.

The Transmission Line Criterion

Not every high-frequency interconnect is a transmission line; this term refers primarily to the electrical interaction between signal and cable, not to the frequency of the signal or the physical characteristics of the cable. So when do we need to incorporate transmission-line effects into our analysis?

The general idea is that transmission-line effects become significant when the length of the line is comparable to or greater than the wavelength of the signal. A more specific guideline is one-fourth of the wavelength:

- If the interconnect length is less than one-fourth of the signal wavelength, transmission-line analysis is not necessary. The interconnect itself does not significantly affect the electrical behavior of the circuit.
- If the interconnect length is greater than one-fourth of the signal wavelength, transmission-line effects become significant, and the influence of the interconnect itself must be taken into account.

Recall that wavelength is equal to propagation velocity divided by frequency:

$$\lambda = \frac{v}{f}$$

If we assume a propagation velocity of 0.7 times the speed of light, we have the following wavelengths:

1 kHz	**210 km**
1 MHz	**210 m**
1 GHz	**210 mm**
10 GHz	**21 mm**

The corresponding transmission-line thresholds are the following:

1 kHz	**52.5 km**
1 MHz	**52.5 m**
1 GHz	**52.5 mm**
10 GHz	**5.25 mm**

So for very low frequencies, transmission-line effects are negligible. For medium frequencies, only very long cables require special consideration. However, at 1 GHz many PCB traces must be treated as transmission lines, and as frequencies climb into the tens of gigahertz, transmission lines become ubiquitous.

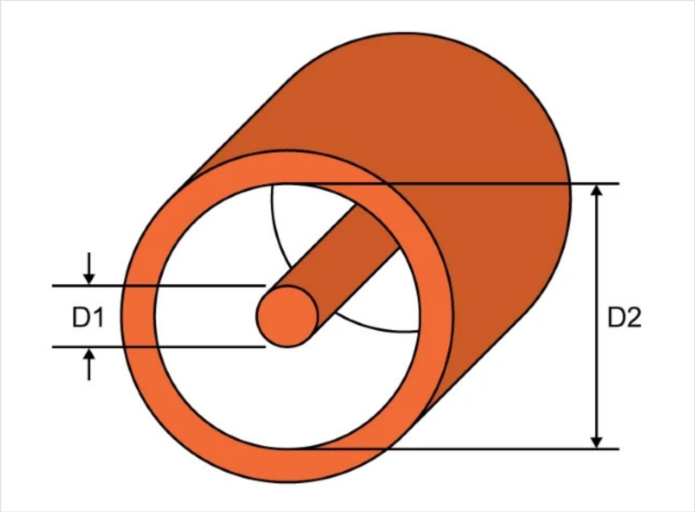
Characteristic Impedance

The most important property of a transmission line is the characteristic impedance (denoted by Z_0). Overall this is a fairly straightforward concept, but initially it can cause confusion.

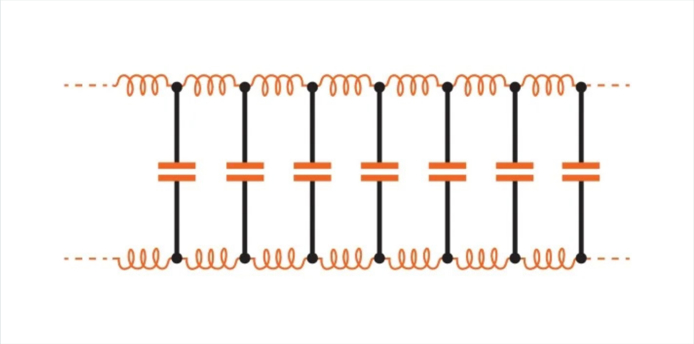
First, a note on terminology: “Resistance” refers to opposition to any flow of current; it is not dependent on frequency. “Impedance” is used in the context of AC circuits and often refers to a frequency-dependent resistance. However, we sometimes use “impedance” where “resistance” would theoretically be more appropriate; for example, we might refer to the “output impedance” of purely resistive circuit.

Thus, it’s important to have a clear idea of what we mean by “characteristic impedance.” It is not the resistance of the signal conductor inside the cable—a common characteristic impedance is 50 Ω, and a DC resistance of 50 Ω for a short cable would be absurdly high. Here are some salient points that help to clarify the nature of characteristic impedance:

- Characteristic impedance is determined by the physical properties of the transmission line; in the case of a coaxial cable, it is a function of the inner diameter (D1 in the diagram below), the outer diameter (D2), and the relative permittivity of the insulation between the inner and outer conductors.



- Characteristic impedance is not a function of cable length. It is present everywhere along the cable, because it results from the cable’s inherent capacitance and inductance.



In this diagram, individual inductors and capacitors are used to represent the distributed capacitance and inductance that is continuously present throughout the length of the cable.

- In practice, a transmission line's impedance is not relevant at DC, but a theoretical transmission line of infinite length would present its characteristic impedance even to a DC source such as a battery. This is the case because the infinitely long transmission line would perpetually draw current in an attempt to charge up its infinite supply of distributed capacitance, and the ratio of the battery voltage to the charging current would be equal to the characteristic impedance.

- The characteristic impedance of a transmission line is purely resistive; no phase shift is introduced, and all signal frequencies propagate at the same speed. Theoretically this is true only for *lossless* transmission lines—i.e., transmission lines that have zero resistance along the conductors and infinite resistance between the conductors. Obviously such lines do not exist, but lossless-line analysis is sufficiently accurate when applied to real-life low-loss transmission lines.

Reflections and Matching

The impedance of a transmission line is not intended to restrict current flow in the way that an ordinary resistor would. Characteristic impedance is simply an unavoidable result of the interaction between a cable composed of two conductors in close proximity. The importance of characteristic impedance in the context of RF design lies in the fact that the designer must match impedances in order to prevent reflections and achieve maximum power transfer. This will be discussed in the next page.

Summary

- An interconnect is considered a transmission line when its length is at least one-fourth of the signal wavelength.
- Coaxial cables are commonly used as transmission lines, though PCB traces also serve this purpose. Two standard PCB transmission lines are the microstrip and the stripline.
- PCB interconnects are typically short, and consequently they do not exhibit transmission-line behavior until signal frequencies approach 1 GHz.
- The ratio of voltage to current in a transmission line is referred to as the characteristic impedance. It is a function of the physical properties of the cable, though it is not affected by length, and for idealized (i.e., lossless) lines it is purely resistive.

Understanding Reflections and Standing Waves in RF Circuit Design

High-frequency circuit design must account for two important though somewhat mysterious phenomena: reflections and standing waves.

We know from our exposure to other branches of science that waves are associated with special types of behavior. Light waves refract when they move from one medium (such as air) into a different medium (such as glass). Water waves diffract when they encounter boats or large rocks. Sound waves interfere, resulting in periodic variations in volume (called “beats”).

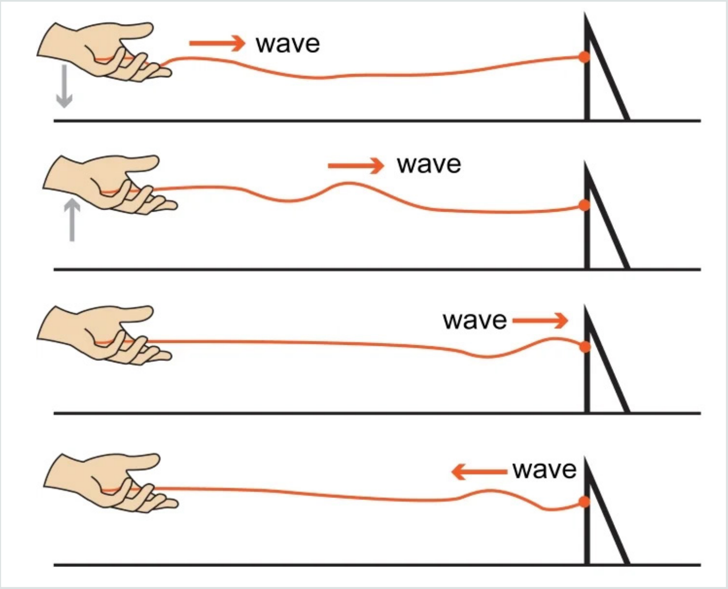
Electrical waves are also subject to behavior that we usually do not associate with electrical signals. The general lack of familiarity with the wave nature of electricity is not surprising, though, because in numerous circuits these effects are negligible or nonexistent. It is possible for a digital or low-frequency-analog engineer to work for years and design many successful systems without ever acquiring a thorough understanding of the wave effects that become prominent in high-frequency circuits.

As discussed in the previous page, an interconnect that is subject to special high-frequency signal behavior is called a transmission line. Transmission-line effects are significant only when the length of the interconnect is at least one-fourth of the signal wavelength; thus, we don’t have to worry about wave properties unless we are working with high frequencies or very long interconnects.

Reflection

Reflection, refraction, diffraction, interference—all of these classic wave behaviors apply to electromagnetic radiation. But at this point we’re still dealing with electrical signals, i.e., signals that have not yet been converted by the antenna into electromagnetic radiation, and consequently we only have to concern ourselves with two of these: reflection and interference.

We generally think of an electrical signal as a one-way phenomenon; it travels from the output of one component to the input of another component, or in other words, from a source to a load. In RF design, however, we must always be aware of the fact that signals can travel in both directions: from source to load, certainly, but also—because of reflections—from load to source.



The wave traveling along the string experiences reflection when it reaches a physical barrier.

A Water-Wave Analogy

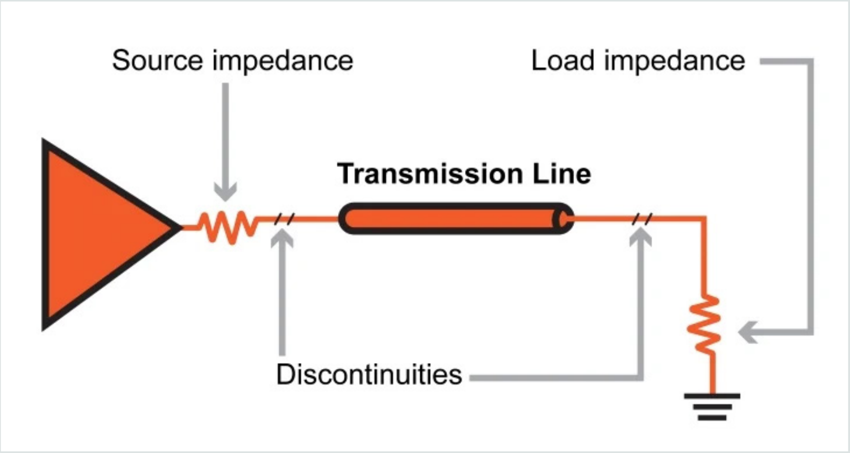
Reflections occur when a wave encounters a discontinuity. Imagine that a storm has resulted in large water waves propagating through a normally calm harbor. These waves eventually collide with a solid rock wall. We intuitively know that these waves will reflect off the rock wall and propagate back out into the harbor. However, we also intuitively know that water waves breaking onto a beach will rarely result in significant reflection of energy back out into the ocean. Why the difference?

Waves transfer energy. When water waves are propagating through open water, this energy is simply moving. When the wave reaches a discontinuity, though, the smooth movement of energy is interrupted; in the case of a beach or a rock wall, wave propagation is no longer possible. But what happens to the energy that was being transferred by the wave? It cannot disappear; it must be either absorbed or reflected. The rock wall does not absorb the wave energy, so reflection occurs—the energy continues propagating in wave form, but in the opposite direction. The beach, however, allows the wave energy to dissipate in a more gradual and natural way. The beach absorbs the wave's energy, and thus minimal reflection occurs.

From Water to Electrons

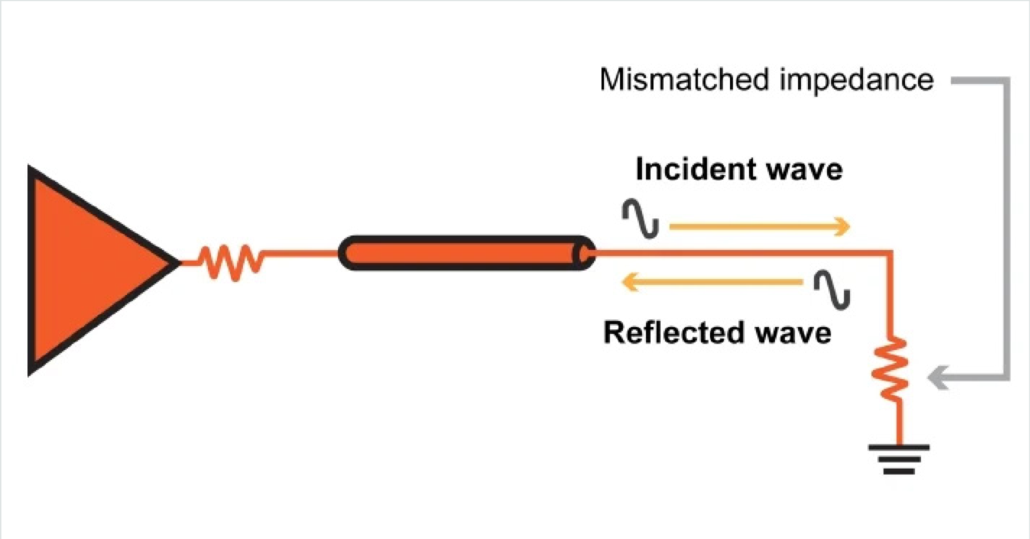
Electrical circuits also present discontinuities that affect wave propagation; in this context, the critical parameter is impedance. Imagine an electrical wave traveling down a transmission line; this is equivalent to the water wave in the middle of the ocean.

The wave and its associated energy are propagating smoothly from source to load. Eventually, though, the electrical wave reaches its destination: an antenna, an amplifier, etc.



We know from a previous page that maximum power transfer occurs when the magnitude of the load impedance is equal to the magnitude of the source impedance. (In this context "source impedance" can also refer to the characteristic impedance of a transmission line.) With matched impedances, there really is no discontinuity, because the load can absorb all of the wave's energy. But if the impedances are not matched, only some of the energy is absorbed, and the remaining energy is reflected in the form of an electrical wave traveling in the opposite direction.

The amount of reflected energy is influenced by the seriousness of the mismatch between source and load impedance. The two worst-case scenarios are an open circuit and a short circuit, corresponding to infinite load impedance and zero load impedance, respectively. These two cases represent a complete discontinuity; no energy can be absorbed, and consequently all the energy is reflected.

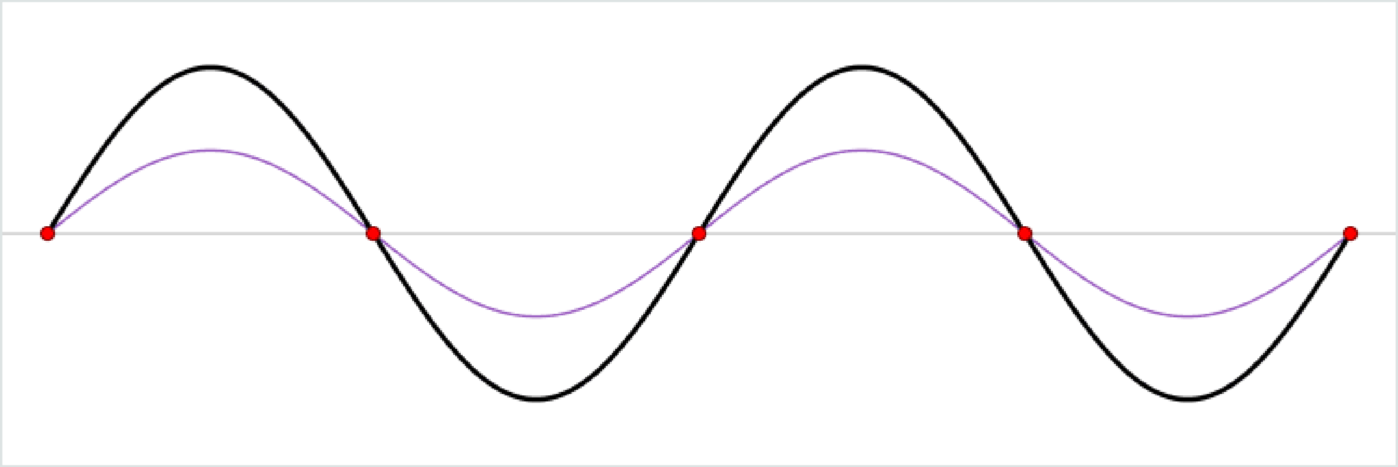


The Importance of Matching

If you've even been involved in RF design or testing, you know that impedance matching is a common topic of discussion. We now understand that impedances must be matched to prevent reflections, but why so much concern about reflections?

The first problem is simply efficiency. If we have a power amplifier connected to an antenna, we don't want half of the output power to be reflected back to the amplifier. The whole point is to generate electrical power that can be converted into electromagnetic radiation. In general, we want to move power from source to load, and this means that reflections must be minimized.

The second issue is a bit more subtle. A continuous signal transferred through a transmission line to a mismatched load impedance will result in a continuous reflected signal. These incident and reflected waves pass each other, going in opposite directions. Interference results in a *standing wave*, i.e., a stationary wave pattern equal to the sum of the incident and reflected waves. This standing wave really does create peak-amplitude variations along the physical length of the cable; certain locations have higher peak amplitude, and other locations have lower peak amplitude.



Standing waves result in voltages that are higher than the original voltage of the transmitted signal, and in some cases the effect is severe enough to cause physical damage to cables or components.

Summary

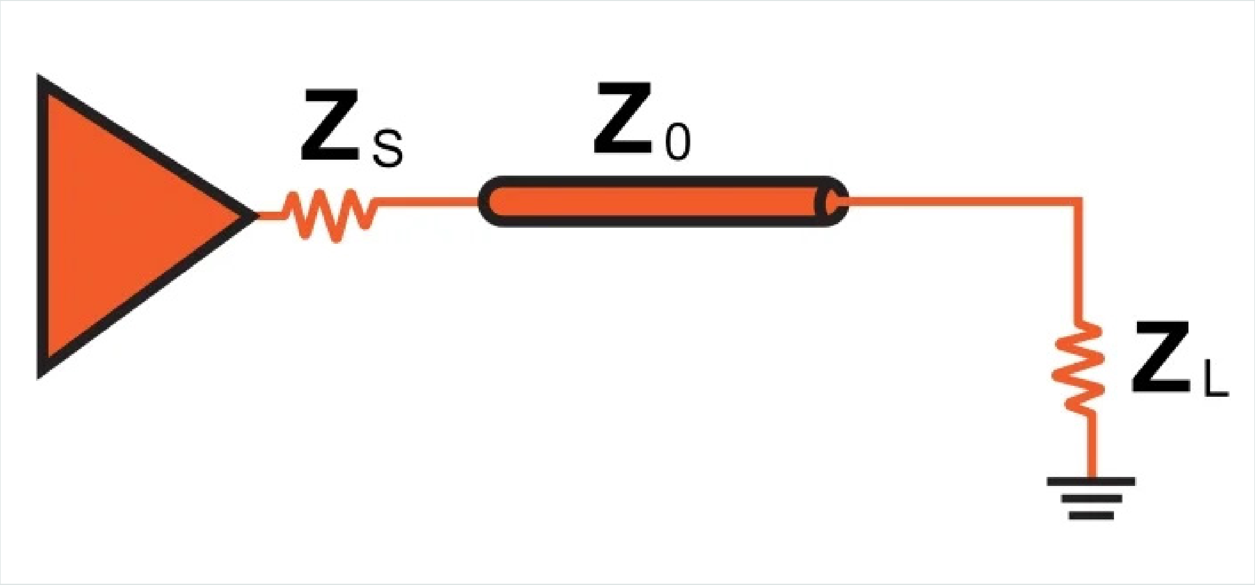
- Electrical waves are subject to reflection and interference.

- Water waves reflect when they reach a physical obstruction such as a stone wall. Similarly, electrical reflection occurs when an AC signal encounters an impedance discontinuity.
- We can prevent reflection by matching the load impedance to the characteristic impedance of the transmission line. This allows the load to absorb the wave energy.
- Reflections are problematic because they reduce the amount of power that can be transferred from source to load.
- Reflections also lead to standing waves; the high-amplitude portions of a standing wave can damage components or cables.

The 50 Ω Question: Impedance Matching in RF Design

Impedance matching is a fundamental aspect of RF design and testing; the signal reflections caused by mismatched impedances can lead to serious problems.

Matching seems like a trivial exercise when you're dealing with a theoretical circuit composed of an ideal source, a transmission line, and a load.



Let's assume that the load impedance is fixed. All we need to do is include a source impedance (Z_S) equal to Z_L and then design the transmission line so that its characteristic impedance (Z_0) is also equal to Z_L.

But let's consider for a moment the difficulty of implementing this scheme throughout a complex RF circuit consisting of numerous passive components and integrated circuits. The RF design process would be seriously unwieldy if engineers had to modify every component and specify the dimensions of every microstrip according to the one impedance chosen as the basis for all the others.

Also, this assumes that the project has already reached the PCB stage. What if we want to test and characterize a system using discrete modules, with off-the-shelf cables as interconnects? Compensating for mismatched impedances is even more impractical under these circumstances.

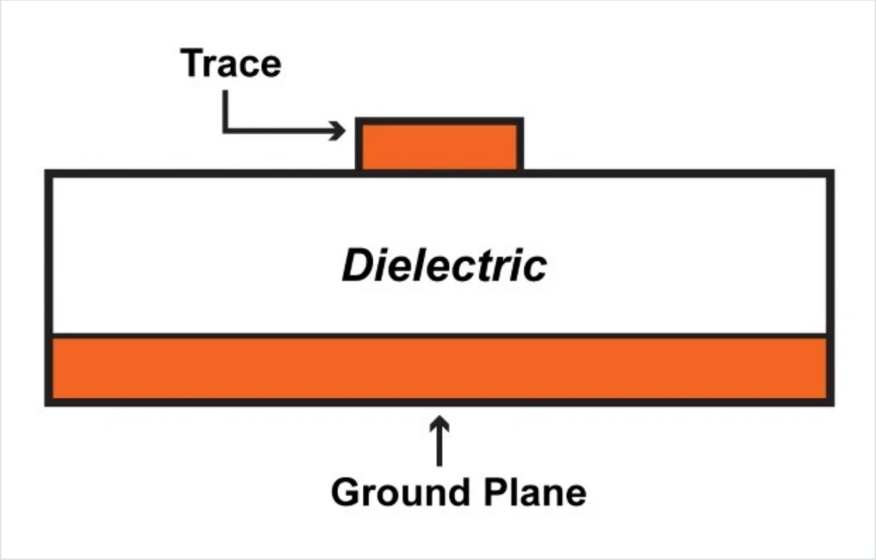
The solution is simple: choose a standardized impedance that can be used in numerous RF systems, and ensure that components and cables are designed accordingly. This impedance has been chosen; the unit is ohms, and the number is 50.

Fifty Ohms

The first thing to understand is that there is nothing intrinsically special about a 50 Ω impedance. This is not a fundamental constant of the universe, though you might get the impression that it is if you spend enough time around RF engineers. It is not even a fundamental constant of electrical engineering—remember, for example, that simply changing the physical dimensions of a coaxial cable will alter the characteristic impedance.

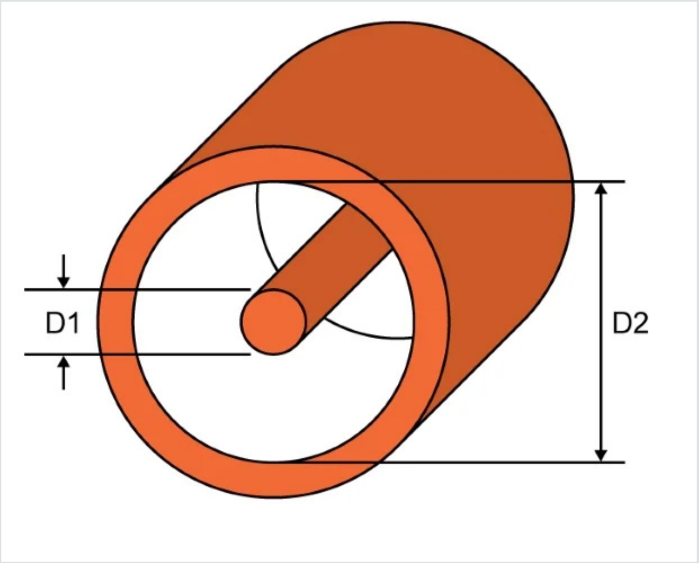
Nevertheless, 50 Ω impedance is very important, because it is the impedance around which most RF systems are designed. It is difficult to determine exactly why 50 Ω became the standardized RF impedance, but it's reasonable to assume that 50 Ω was found to be a good compromise in the context of early coaxial cables.

The important issue, of course, is not the origin of the specific value but rather the benefits of having this standardized impedance. Achieving a well-matched design is vastly simpler because manufacturers of ICs, fixed attenuators, antennas, and so forth can build their parts with this impedance in mind. Also, PCB layout becomes more straightforward because so many engineers have the same goal, namely, to design microstrips and striplines that have a characteristic impedance of 50 Ω.



According to this app note from Analog Devices, you can create a 50 Ω microstrip as follows: 1-ounce copper, 20-mil-wide trace, 10-mil separation between trace and ground plane (assuming FR-4 dielectric).

Before we move on, let's be clear that not every high-frequency system or component is designed for 50 Ω. Other values could be chosen, and in fact 75 Ω impedance is still common. The characteristic impedance of a coaxial cable is proportional to the natural log of the ratio of the outer diameter (D2) to the inner diameter (D1).



This means that more separation between the inner conductor and outer conductor corresponds to a higher impedance. Greater separation between the two conductors also leads to lower capacitance. Thus, 75 Ω coax has lower capacitance than 50 Ω coax, and this makes 75 Ω cable more suitable for high-frequency digital signals, which require low capacitance in order to avoid excessive attenuation of the high-frequency content associated with the rapid transitions between logic low and logic high.

Reflection Coefficient

Considering how important impedance matching is in RF design, we shouldn't be surprised to find that there is a specific parameter used to express the quality of a match. It is called the reflection coefficient; the symbol is Γ (the Greek capital letter gamma). It is the ratio of the complex amplitude of the reflected wave to the complex amplitude of the incident wave. However, the relationship between incident wave and reflected wave is determined by the source (Z_S) and load (Z_L) impedances, and thus it is possible to define the reflection coefficient in terms of these impedances:

$$\Gamma = \frac{Z_L - Z_S}{Z_L + Z_S}$$

If the "source" in this case is a transmission line, we can change the Z_S to Z_0.

$$\Gamma = \frac{Z_L - Z_0}{Z_L + Z_0}$$

In a typical system, the magnitude of the reflection coefficient is a number between zero and one. Let's look at three mathematically straightforward situations to help us understand how the reflection coefficient corresponds to actual circuit behavior:

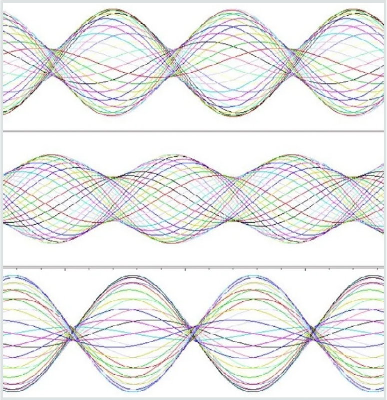
- If the match is perfect ($Z_L = Z_0$), the numerator is zero, and thus the reflection coefficient is zero. This makes sense because perfect matching results in no reflection.
- If the load impedance is infinite (i.e., an open circuit), the reflection coefficient becomes infinity divided by infinity, which is one. A reflection coefficient of one corresponds to full reflection, i.e., all of the wave energy is reflected. This makes sense because a transmission line connected to an open circuit corresponds to a complete discontinuity (see the previous page)—the load cannot absorb any energy, so it must all be reflected.
- If the load impedance is zero (i.e., a short circuit), the magnitude of the reflection coefficient becomes Z_0 divided by Z_0. Thus we again have $|\Gamma| = 1$, which makes sense because a short circuit also corresponds to a complete discontinuity that cannot absorb any of the incident wave energy.

VSWR

Another parameter used to describe impedance matching is the voltage standing wave ratio (VSWR). It is defined as follows:

$$VSWR = \frac{1+|\Gamma|}{1-|\Gamma|}$$

VSWR approaches impedance matching from the perspective of the resulting standing wave. It conveys the ratio of the highest standing-wave amplitude to the lowest standing-wave amplitude. This video can help you visualize the relationship between impedance mismatch and the amplitude characteristics of the standing wave, and the following diagram conveys standing-wave amplitude characteristics for three different reflection coefficients.

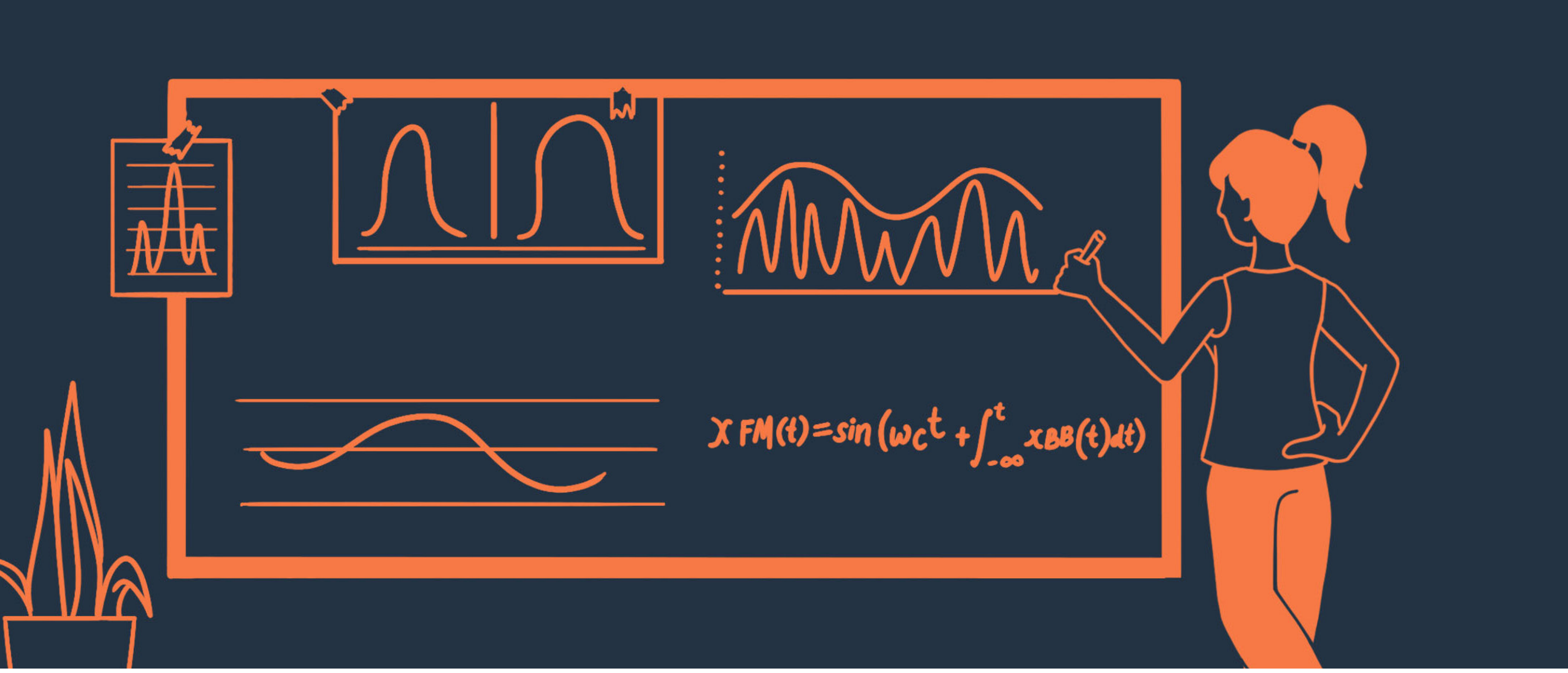


More impedance mismatch leads to a greater difference between the highest-amplitude and lowest-amplitude locations along the standing wave. Image used courtesy of the Interferometrist [CC BY-SA 4.0]

VSWR is commonly expressed as a ratio. A perfect match would be 1:1, meaning that the peak amplitude of the signal is always the same (i.e., there is no standing wave). A ratio of 2:1 indicates that reflections have resulted in a standing wave with a maximum amplitude that is twice as large as its minimum amplitude.

Summary

- The use of a standardized impedance makes RF design much more practical and efficient.
- Most RF systems are built around 50 Ω impedance. Some systems use 75 Ω; this latter value is more appropriate for high-speed digital signals.
- The quality of an impedance match can be expressed mathematically by the reflection coefficient (Γ). A perfect match corresponds to Γ = 0, and a complete discontinuity (in which all the energy is reflected) corresponds to Γ = 1.
- Another way of quantifying the quality of an impedance match is the voltage standing wave ratio (VSWR).



CHAPTER 4

Radio Frequency Modulation

- The Many Types of Radio Frequency Modulation
- Amplitude Modulation in RF: Theory, Time Domain, Frequency Domain
- Frequency Modulation: Theory, Time Domain, Frequency Domain
- Phase Modulation: Theory, Time Domain, Frequency Domain
- Digital Modulation: Amplitude and Frequency
- Digital Phase Modulation: BPSK, QPSK, DQPSK
- Comparing and Contrasting Amplitude, Frequency, and Phase Modulation

The Many Types of Radio Frequency Modulation

RF communication is built upon a simple concept: by continually altering the characteristics of a sinusoid, we can use it to transfer information.

At this point, we have covered a variety of important concepts that serve as a foundation for the successful design and analysis of real-world RF circuits and systems. We are now ready to explore a fundamental aspect of RF engineering: modulation.

What Is Modulation?

The general meaning of the verb “to modulate” is “to modify, to regulate, to vary,” and this captures the essence of modulation even in the specialized context of wireless communication. To modulate a signal is simply to intentionally modify it, but of course, this modification is done in a very specific way because the goal of modulation is data transfer.

We want to transfer information—ones, and zeros if we’re dealing with digital data, or a sequence of continuously varying values if we are working in the analog realm. But the restrictions imposed by wireless communication do not allow us to express this information in the typical way; instead, we have to devise a new “language,” or you might think of it as a code, that allows us to convey the same information but within the constraints of an electromagnetic-radiation-based system. More specifically, we need a language that is compatible with high-frequency sinusoidal signals, because such signals constitute the only practical means of “carrying” information in a typical RF system.

This high-frequency sinusoid that is used to carry information is called, appropriately, the carrier. It’s a helpful name because it reminds us that the purpose of an RF system is not to generate and transmit a high-frequency sinusoid. Rather, the purpose is to transfer (lower-frequency) information, and the carrier is simply the means that we must use to move this information from an RF transmitter to an RF receiver.

Modulation Schemes

In verbal communication, the human body generates sound waves and modifies—or modulates—them so as to produce a wide variety of vowels and consonants. Intelligent use of these vowels and consonants results in the transfer of information from the speaker to the listener. The system according to which the sound waves are modulated is called a language.

In RF communication, the situation is very similar. A device modulates electrical waves according to a predefined system called a modulation scheme (or modulation technique). Just as there are many human languages, there are many ways in which a carrier can be modulated.

Sophisticated modulation schemes help modern RF systems to achieve increased range and improved immunity to interference.

It is possible that certain human languages are especially effective in conveying certain types of information; to take an example from the ancient world, perhaps Greek was better for philosophy and Latin was better for codifying laws. There is no doubt, however, that reliable communication is possible with any properly developed language, *as long as the speaker and the listener both know it.* The same is true for RF systems. Each modulation scheme has its advantages and disadvantages, but all can provide excellent wireless communication if the fundamental requirement is fulfilled—i.e., the receiver must be able to understand what the transmitter is saying.

Amplitude, Frequency, Phase

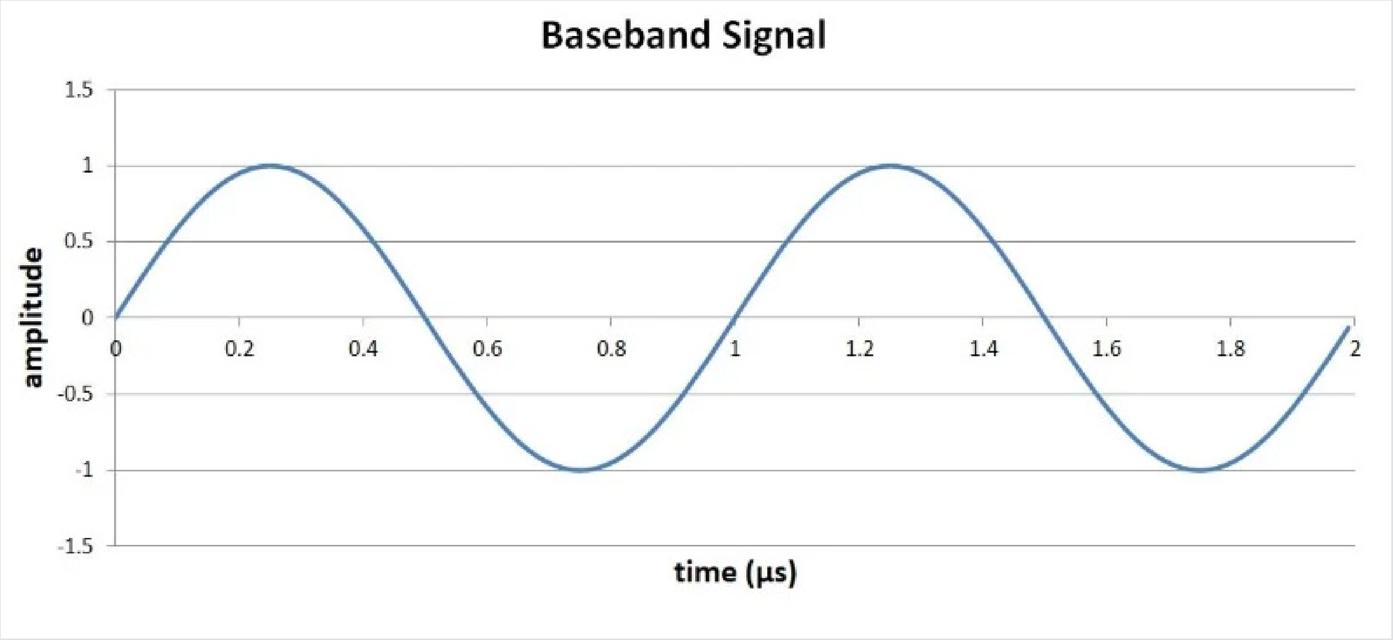
A basic sinusoid is a simple thing. If we ignore DC offset, it can be completely characterized with only two parameters: amplitude and frequency. We also have phase, which comes into play when we consider the initial state of the sinusoid, or when changes in wave behavior allow us to contrast one portion of the sinusoid with a preceding portion. Phase is also relevant when comparing two sinusoids; this aspect of sinusoidal phase has become very important because of the widespread use of quadrature (or “IQ”) signals in RF systems. We’ll look at IQ concepts later in the textbook.

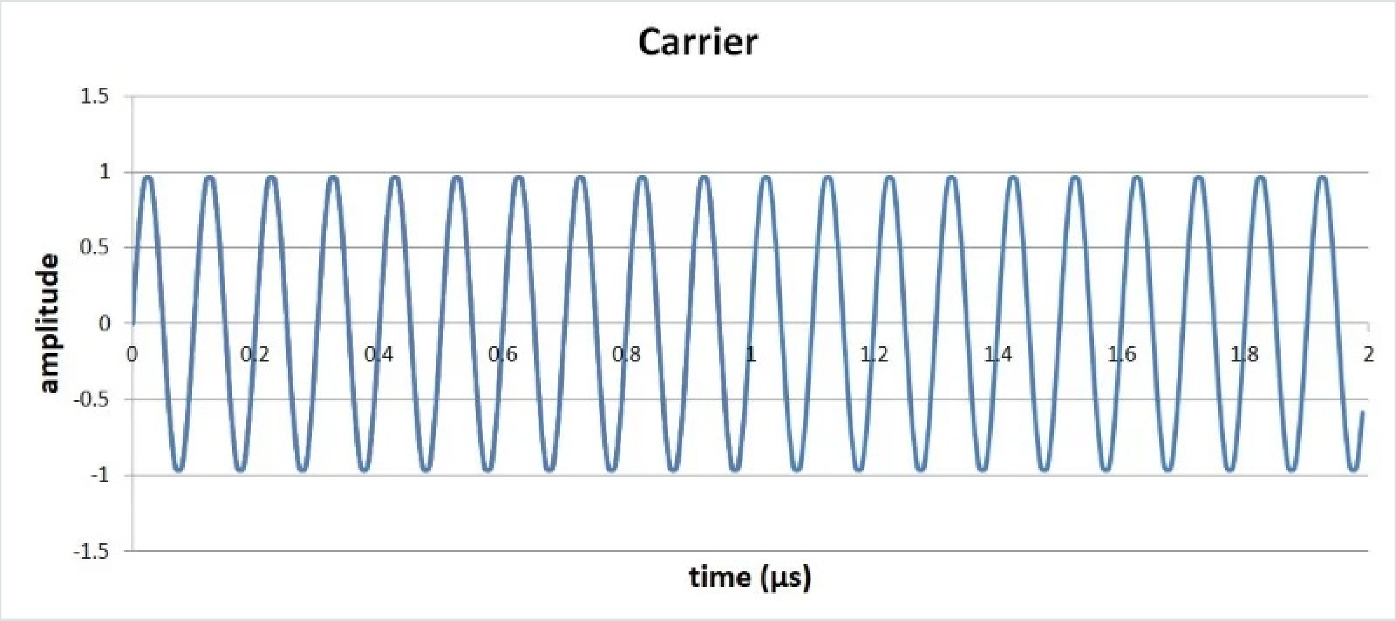
As discussed above, modulation is modification, and we can modify only what is already present. Sinusoids have amplitude, frequency, and phase, and thus it should come as no surprise that modulation schemes are categorized as amplitude modulation, frequency modulation, or phase modulation. (Actually, it is possible to bridge these categories by combining amplitude modulation with frequency or phase modulation.) Within each category we have two subcategories: analog modulation and digital modulation.

Amplitude Modulation (AM)

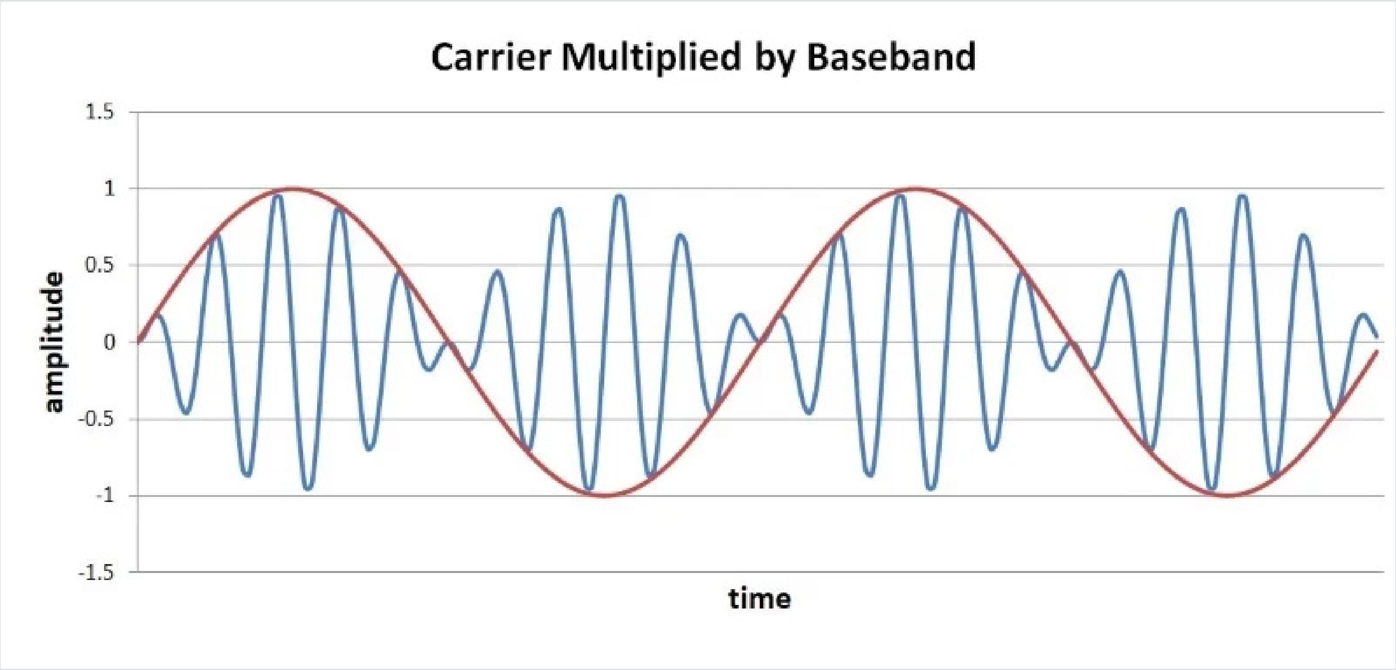
Analog AM consists of multiplying a continuously varying sinusoidal carrier by an offset version of a continuously varying information (aka baseband) signal. By “offset version” I mean that the amplitude of the baseband signal is always greater than or equal to zero.

Let’s assume that we have a 10 MHz carrier and a 1 MHz baseband waveform:





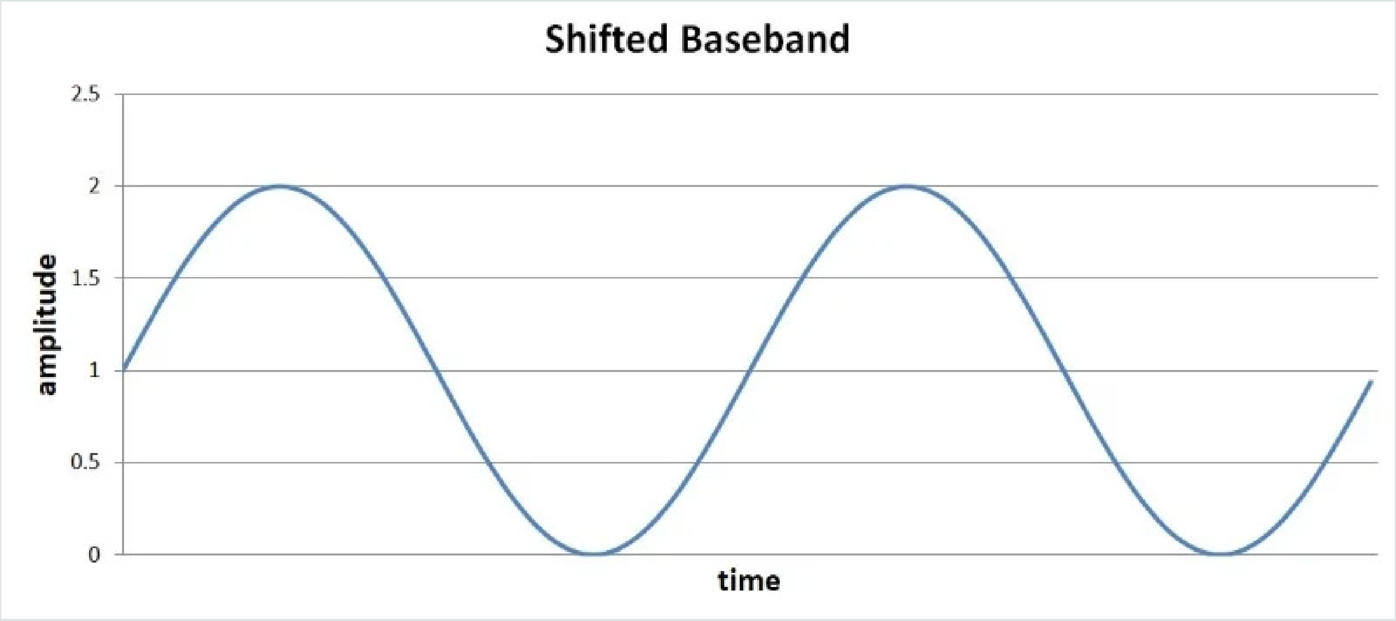
If we multiply these two signals, we get the following (incorrect) waveform:



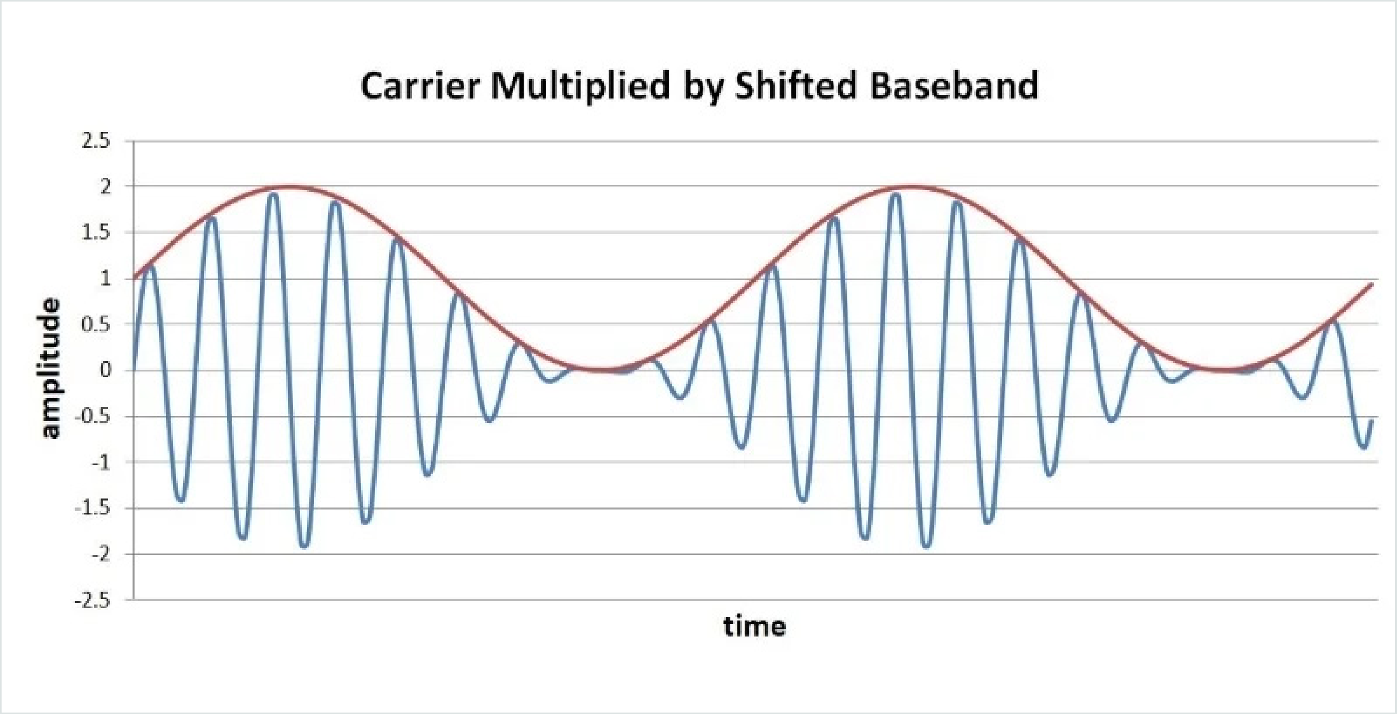
You can clearly see the relationship between the baseband signal (red) and the amplitude of the carrier (blue).

But we have a problem: If you look only at the amplitude of the carrier, how can you determine if the baseband value is positive or negative? You can't—and, consequently, amplitude demodulation will not extract the baseband signal from the modulated carrier.

The solution is to shift the baseband signal so that it varies from 0 to 2 instead of -1 to 1:



If we multiply the shifted baseband signal by the carrier, we have the following:



Now the amplitude of the carrier can be mapped directly to the behavior of the baseband signal.

The most straightforward form of digital AM applies the same mathematical relationship to a baseband signal whose amplitude is either 0 or 1. The result is referred to as “on-off keying” (OOK): when the information signal is logic zero, the carrier’s amplitude is zero (= “off”); when the information signal is logic one, the carrier is at full amplitude (= “on”).

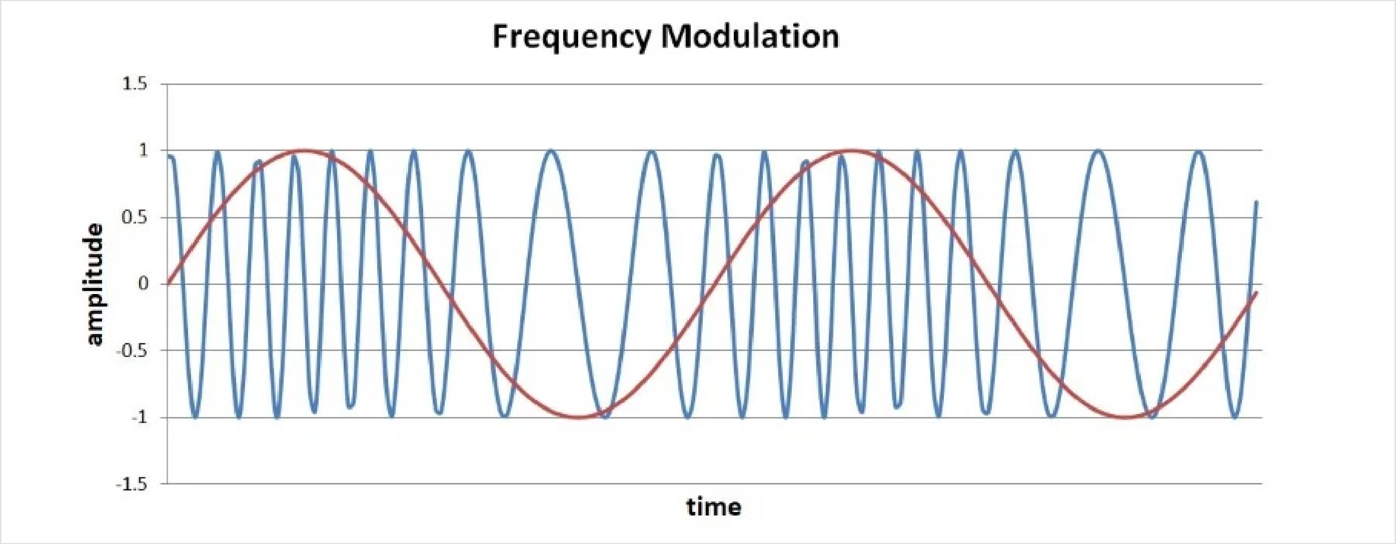
Frequency Modulation (FM) and Phase Modulation (PM)

FM and PM are closely related because frequency and phase are closely related. This is not so obvious if you think of frequency as the number of full cycles per second—what does cycles per second have to do with the position of the sinusoid at a given moment during its cycle? But it makes more sense if you consider the instantaneous frequency, i.e., the frequency of a signal at a given moment. (It is undoubtedly paradoxical to describe a frequency as instantaneous—but, in the context of practical signal processing, we can safely ignore the complicated theoretical details associated with this concept.)

In a basic sinusoid, the value of the instantaneous frequency is the same as that of the "normal" frequency. The analytical value of instantaneous frequency arises when we are dealing with signals that have a time-varying frequency, i.e., the frequency is not a constant value but rather a function of time, written as ω(t). In any event, the important point for our current discussion regarding the close relationship between frequency and phase is the following: instantaneous angular frequency is the derivative, with respect to time, of phase. So if you have an expression φ(t) that describes the time-varying behavior of the signal's phase, the rate of change (with respect to time) of φ(t) gives you the expression for instantaneous angular frequency:

$$\omega(t) = \frac{\mathrm{d}\phi(t)}{\mathrm{d}t}$$

We'll take a closer look at frequency and phase modulation later in this chapter. For now let's conclude with the following plot, which applies the mathematical relationship for frequency modulation to the baseband and carrier signals used above:



Summary

- Modulation refers to the process of carefully modifying an existing signal so that it can transfer information.
- In the context of RF, the existing signal is called the carrier, and the information is contained in the baseband signal.
- There are many different modulation schemes, meaning that there are different ways of incorporating baseband information into a sinusoidal carrier wave.
- Modulation involves modification of a carrier’s amplitude, frequency, or phase, and it can be used to transfer analog signals or digital data.

Amplitude Modulation in RF: Theory, Time Domain, Frequency Domain

Learn about the most straightforward way of encoding information in a carrier waveform.

We have seen that RF modulation is simply the intentional modification of the amplitude, frequency, or phase of a sinusoidal carrier signal. This modification is performed according to a specific scheme that is implemented by the transmitter and understood by the receiver. Amplitude modulation—which of course is the origin of the term "AM radio"—varies the amplitude of the carrier according to the instantaneous value of the baseband signal.

The Math

The mathematical relationship for amplitude modulation is simple and intuitive: you multiply the carrier by the baseband signal. The frequency of the carrier itself is not altered, but the amplitude will vary constantly according to the baseband value. (However, as we will see later, the amplitude variations introduce new frequency characteristics.) The one subtle detail here is the need to shift the baseband signal; we discussed this in the previous page. If we have a baseband waveform that varies between –1 and +1, the mathematical relationship can be expressed as follows:

$$x_{AM} = x_C(1 + x_{BB})$$

where x_{AM} is the amplitude-modulated waveform, x_C is the carrier, and x_{BB} is the baseband signal. We can take this a step further if we consider the carrier to be an endless, constant-amplitude, fixed-frequency sinusoid. If we assume that the carrier amplitude is 1, we can replace x_C with $\sin(\omega_C t)$.

$$x_{AM}(t) = \sin(\omega_C t)(1 + x_{BB}(t))$$

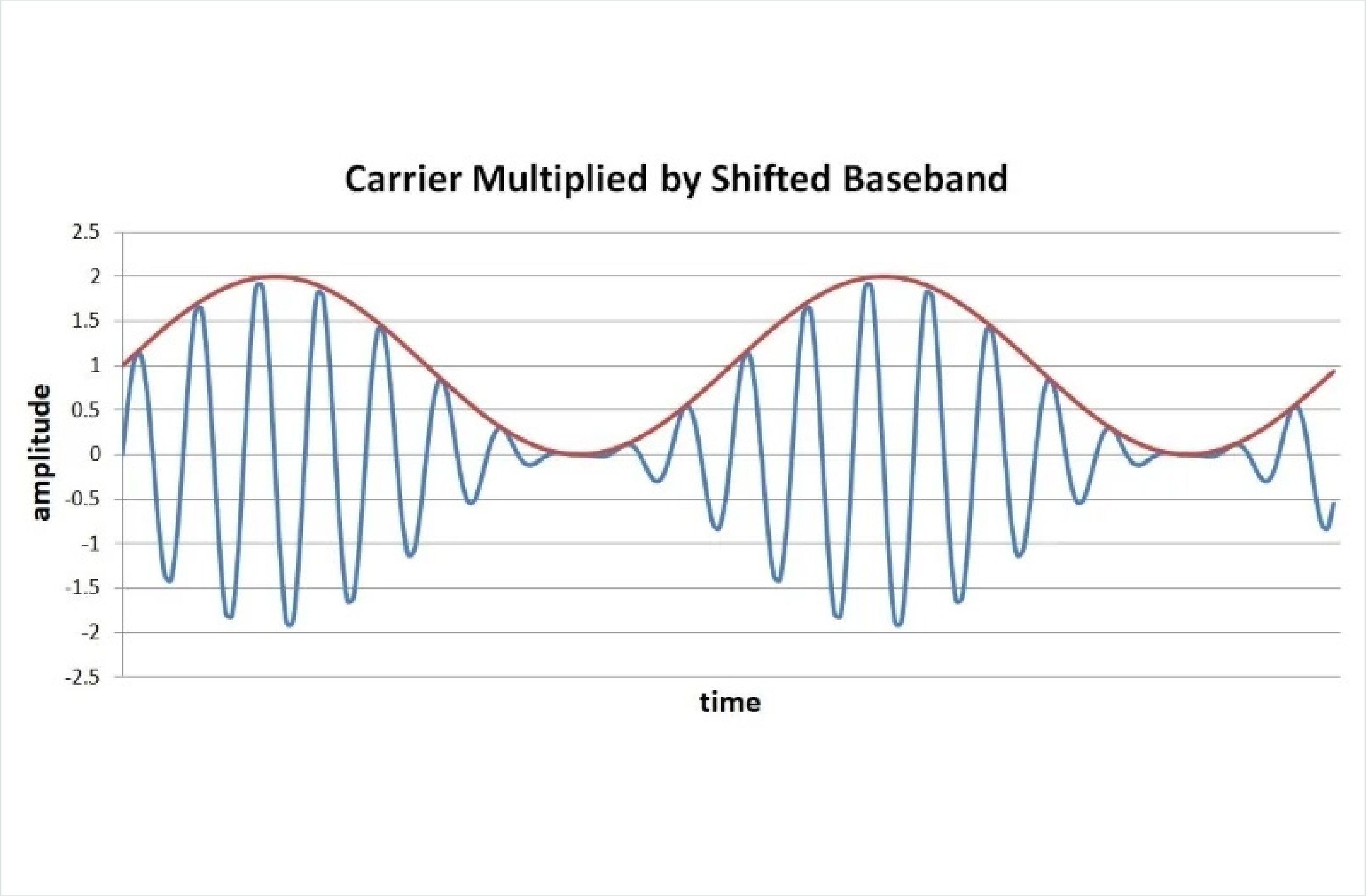
So far so good, but there is one problem with this relationship: you have no control over the "intensity" of the modulation. In other words, the baseband-change-to-carrier-amplitude-change relationship is fixed. We cannot, for example, design the system such that a small change in the baseband value will create a large change in the carrier amplitude. To address this limitation, we introduce m, known as the modulation index.

$$x_{AM}(t) = \sin(\omega_C t)(1 + m x_{BB}(t))$$

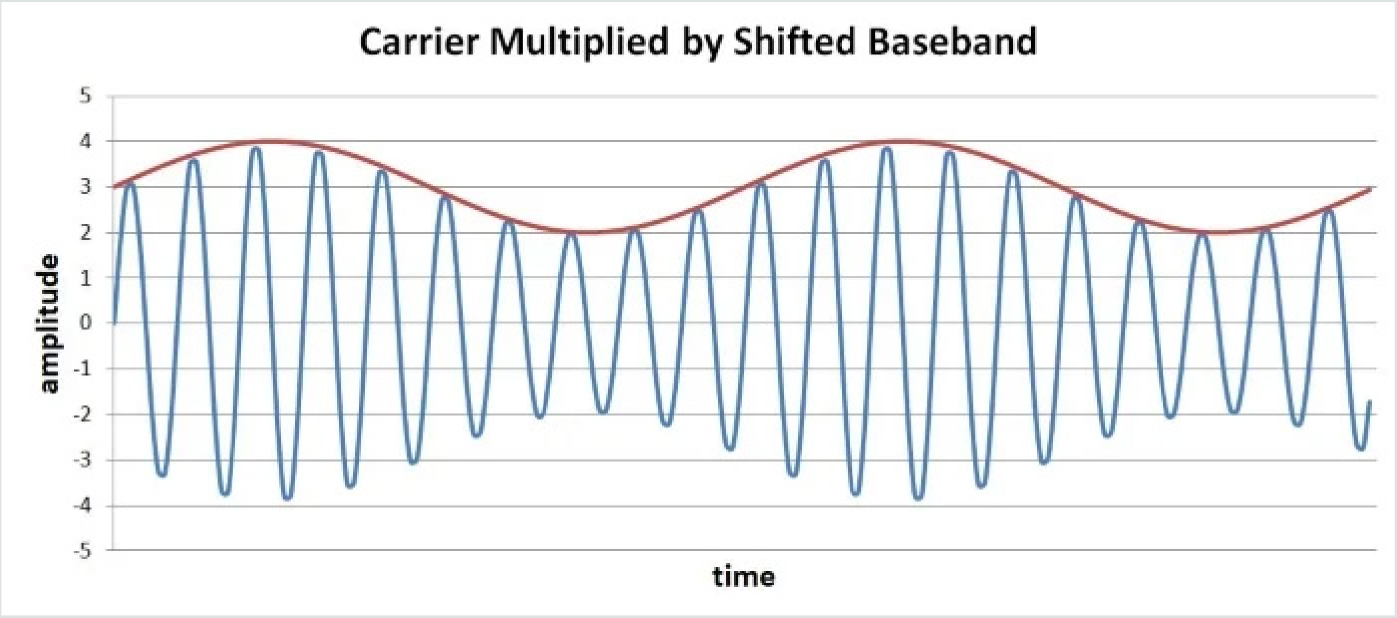
Now, by varying m we can control the intensity of the baseband signal's effect on the carrier amplitude. Notice, however, that m is multiplied by the original baseband signal, not the shifted baseband. Thus, if x_{BB} extends from –1 to +1, any value of m greater than 1 will cause ($1 + mx_{BB}$) to extend into the negative portion of the y-axis—but this is exactly what we were trying to avoid by shifting it upwards in the first place. So remember, if a modulation index is used, the signal must be shifted based on the maximum amplitude of mx_{BB} , not $x_{BB.}$

The Time Domain

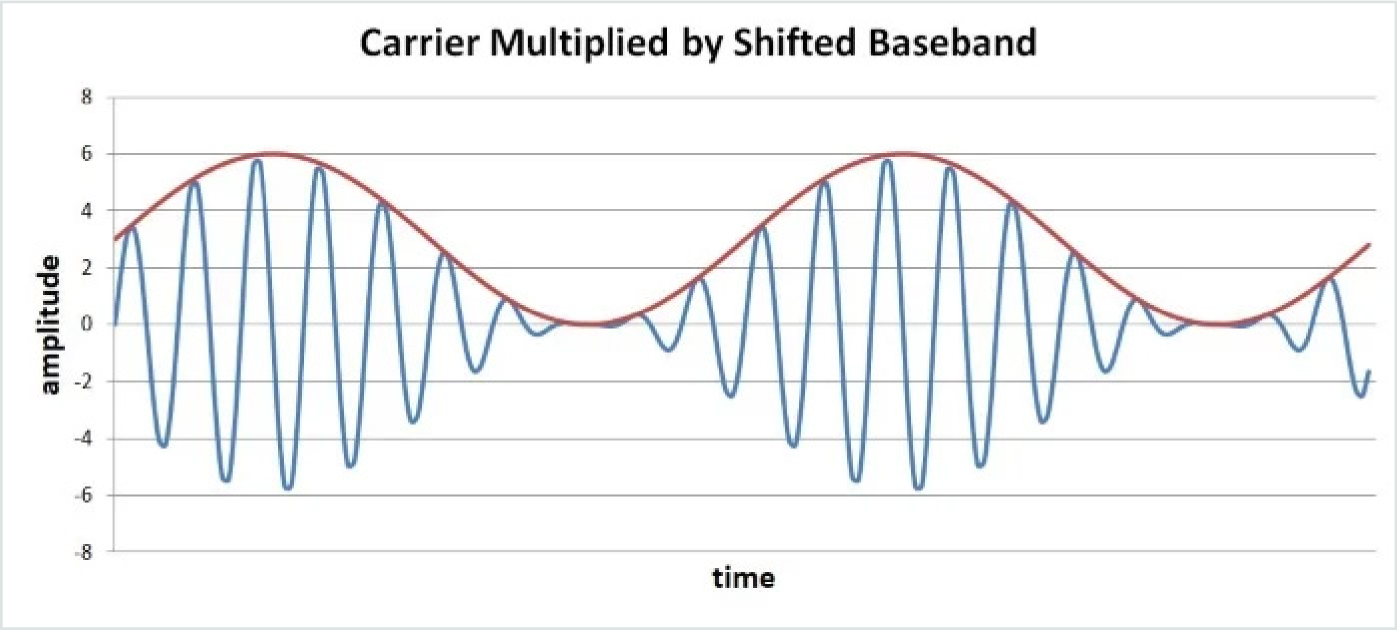
We looked at AM time-domain waveforms in the previous page. Here was the final plot (baseband in red, AM waveform in blue):



Now let's look at the effect of the modulation index. Here is a similar plot, but this time I shifted the baseband signal by adding 3 instead of 1 (the original range is still –1 to +1).



Now we will incorporate a modulation index. The following plot is with m = 3.



The carrier's amplitude is now "more sensitive" to the varying value of the baseband signal. The shifted baseband does not enter the negative portion of the y-axis because I chose the DC offset according to the modulation index.

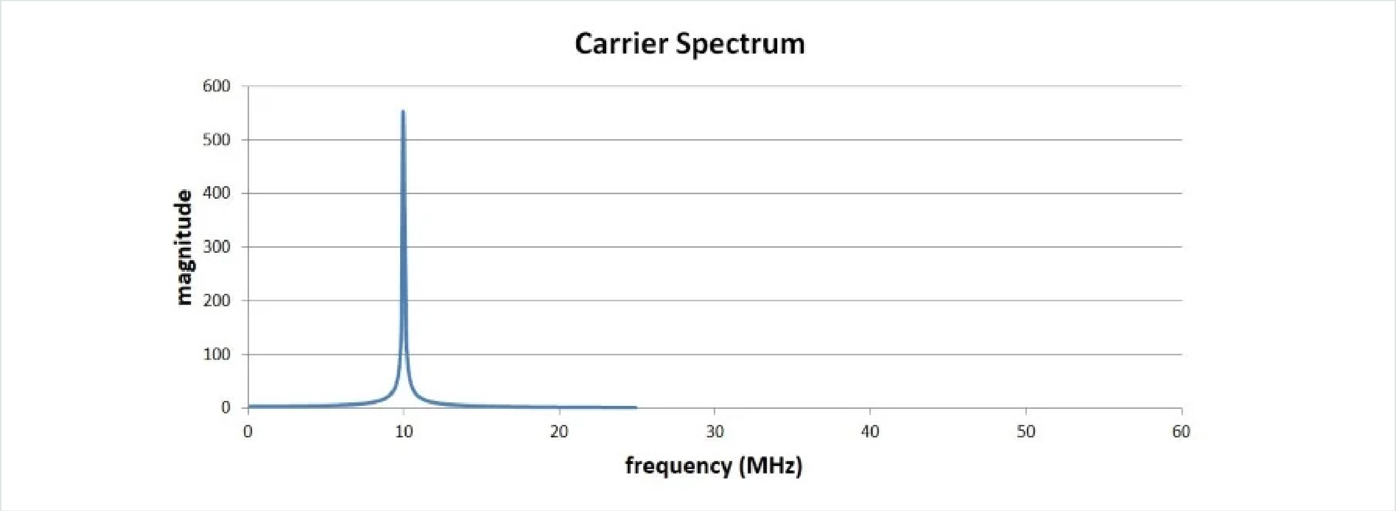
You might be wondering about something: How can we choose the correct DC offset without knowing the exact amplitude characteristics of the baseband signal? In other words, how can we ensure that the baseband waveform's negative swing extends exactly to zero? Answer: You don't need to. The previous two plots are equally valid AM waveforms; the baseband signal is faithfully transferred in both cases.

Any DC offset that remains after demodulation is easily removed by a series capacitor. (The next chapter will cover demodulation.)

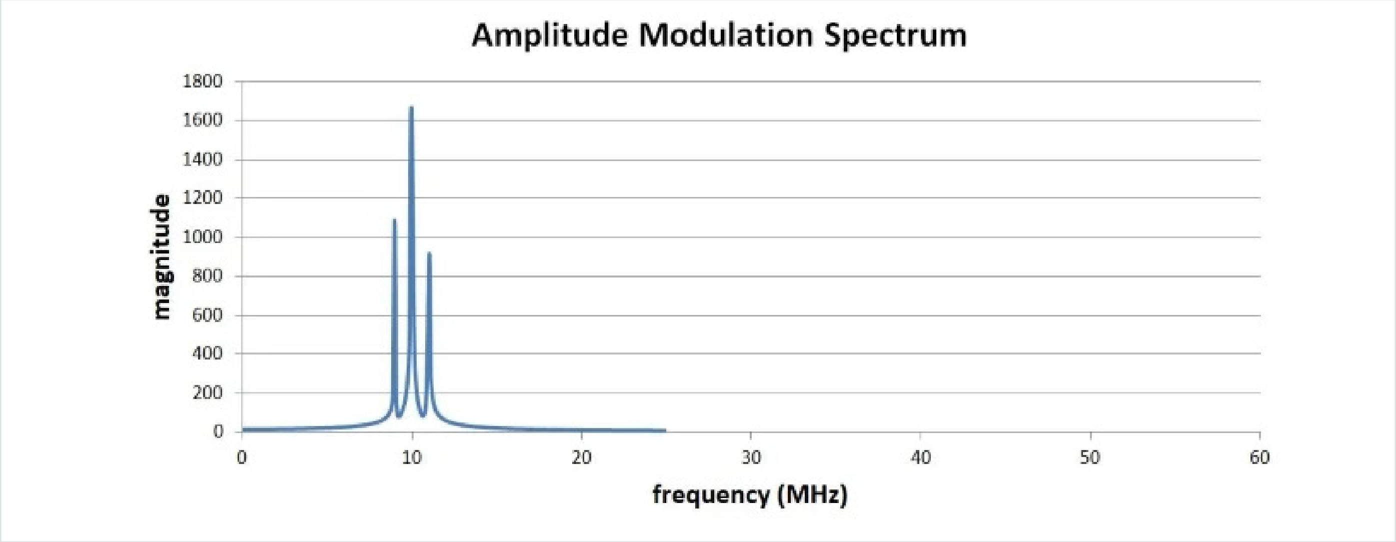
The Frequency Domain

As discussed in the second page of this textbook, RF development makes extensive use of frequency-domain analysis. We can inspect and evaluate a real-life modulated signal by measuring it with a spectrum analyzer, but this means that we need to know what the spectrum should look like.

Let's start with the frequency-domain representation of a carrier signal:

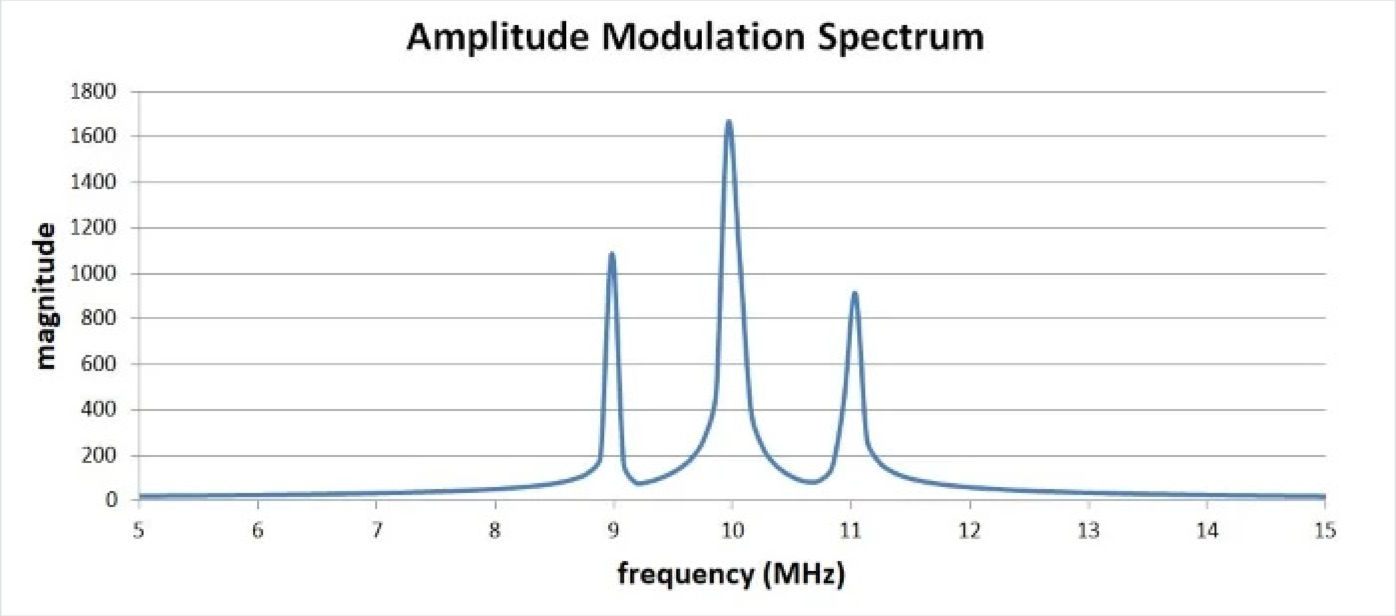


This is exactly what we expect for the unmodulated carrier: a single spike at 10 MHz. Now let's look at the spectrum of a signal created by amplitude modulating the carrier with a constant-frequency 1 MHz sinusoid.



Here you see the standard characteristics of an amplitude-modulated waveform: the baseband signal has been shifted according to the frequency of the carrier. You could also think of this as “adding” the baseband frequencies onto the carrier signal, which is indeed what we’re doing when we use amplitude modulation—the carrier frequency remains, as you can see in the time-domain waveforms, but the amplitude variations constitute new frequency content that corresponds to the spectral characteristics of the baseband signal.

If we look more closely at the modulated spectrum, we can see that the two new peaks are 1 MHz (i.e., the baseband frequency) above and 1 MHz below the carrier frequency:

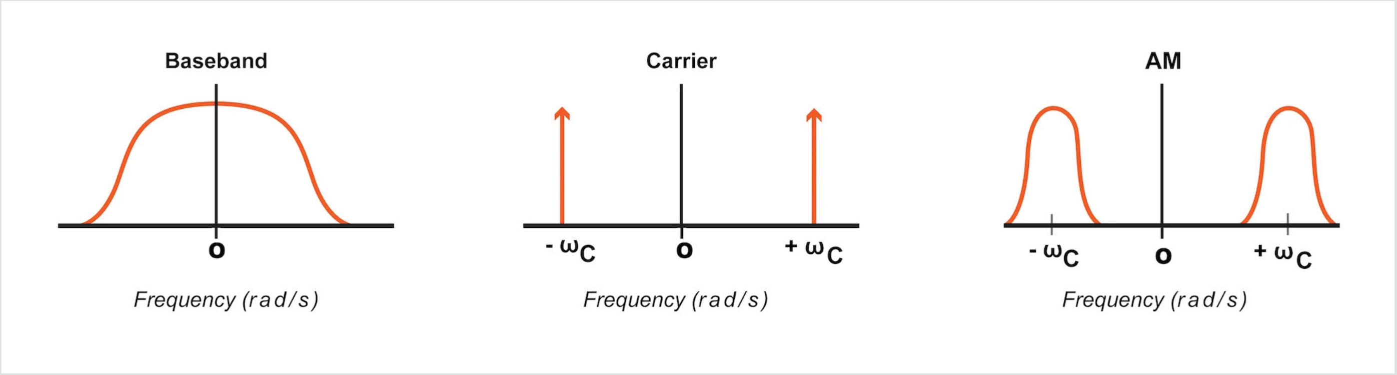


(In case you’re wondering, the asymmetry is an artifact of the calculation process; these plots were generated using real data, with limited resolution. An idealized spectrum would be symmetrical.)

Negative Frequencies

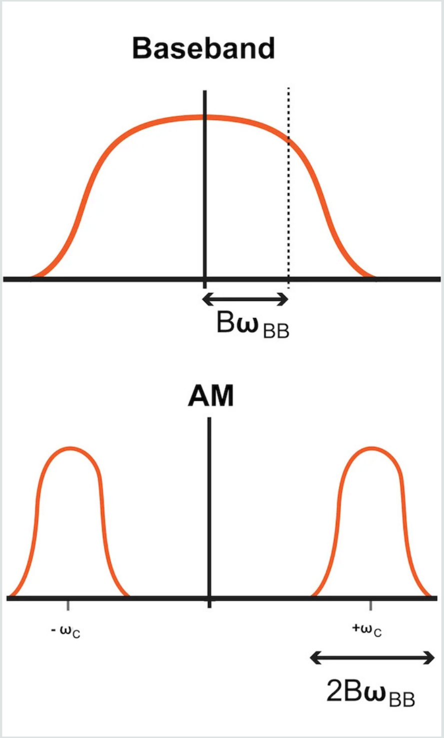
To summarize, then, amplitude modulation translates the baseband spectrum to a frequency band centered around the carrier frequency. There is something we need to explain, though: Why are there two peaks—one at the carrier frequency plus the baseband frequency, and another at the carrier frequency minus the baseband frequency? The answer becomes clear if we simply remember that a Fourier spectrum is symmetrical with respect to the y-axis; even though we often display only the positive frequencies, the negative portion of the x-axis contains corresponding negative frequencies. These negative frequencies are easily ignored when we’re dealing with the original spectrum, but it is essential to include the negative frequencies when we are shifting the spectrum.

The following diagram should clarify this situation.



As you can see, the baseband spectrum and the carrier spectrum are symmetrical with respect to the y-axis. For the baseband signal, this results in a spectrum that extends continuously from the positive portion of the x-axis to the negative portion; for the carrier, we simply have two spikes, one at $+\omega_C$ and one at $-\omega_C$. And the AM spectrum is, once again, symmetrical: the translated baseband spectrum appears in the positive portion and the negative portion of the x-axis.

And here's one more thing to keep in mind: amplitude modulation causes the bandwidth to increase by a factor of 2. We measure bandwidth using only the positive frequencies, so the baseband bandwidth is simply BW_{BB} (see the diagram below). But after translating the entire spectrum (positive and negative frequencies), all the original frequencies become positive, such that the modulated bandwidth is $2BW_{BB.}$



Summary

- Amplitude modulation corresponds to multiplying the carrier by the shifted baseband signal.
- The modulation index can be used to make the carrier amplitude more (or less) sensitive to the variations in the value of the baseband signal.
- In the frequency domain, amplitude modulation corresponds to translating the baseband spectrum to a band surrounding the carrier frequency.
- Because the baseband spectrum is symmetrical with respect to the y-axis, this frequency translation results in a factor-of-2 increase in bandwidth.

Frequency Modulation: Theory, Time Domain, Frequency Domain

Though less intuitive than amplitude modulation, frequency modulation is still a fairly straightforward method of wireless data transmission.

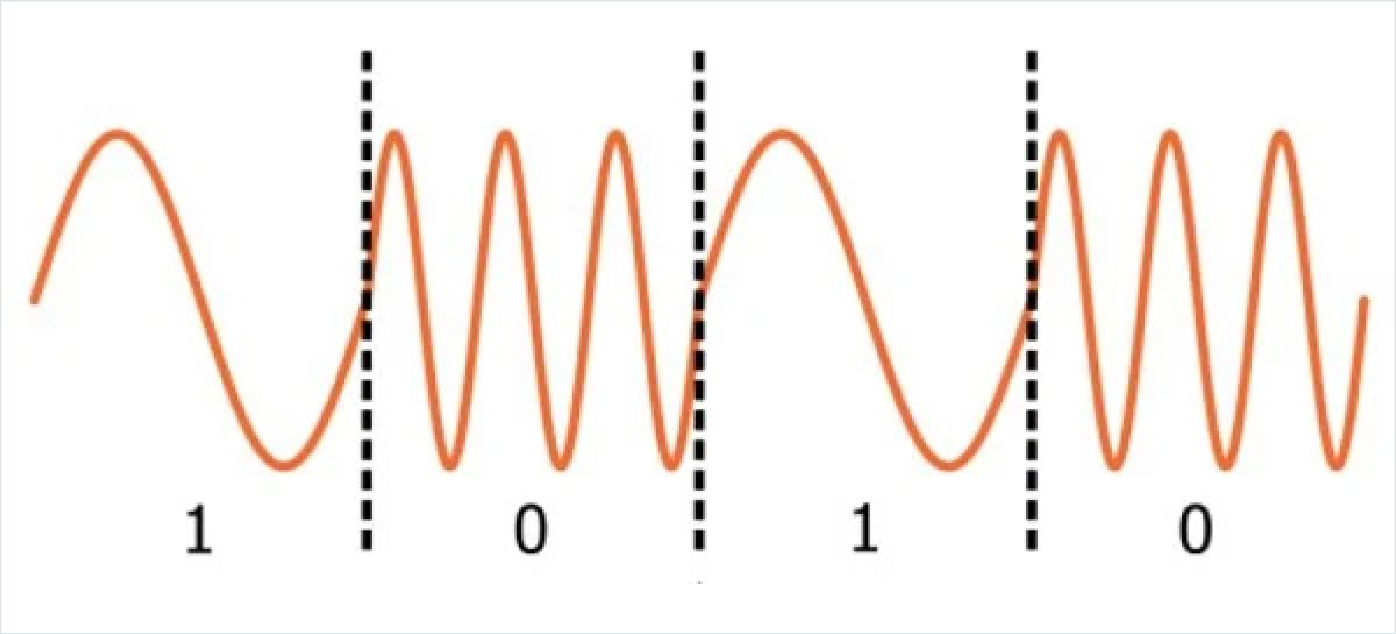
We are all at least vaguely familiar with frequency modulation—it's the origin of the term "**FM** radio." If we think of frequency as something that has an instantaneous value, rather than as something that consists of several cycles divided by a corresponding period of time, we can continuously vary frequency in accordance with the instantaneous value of a baseband signal.

The Math

In the first page of this chapter, we discussed the paradoxical quantity referred to as instantaneous frequency. If you find this term unfamiliar or confusing, go back to that page and read through the "Frequency Modulation (FM) and Phase Modulation (PM)" section. You may still be a bit unsure, though, and that's understandable—the idea of an instantaneous frequency violates the basic principle according to which "frequency" indicates *how frequently* a signal completes a full cycle: ten times per second, a million times per second, or whatever it may be.

We won't attempt any sort of thorough or comprehensive treatment of instantaneous frequency as a mathematical concept. (If you're determined to explore this issue in depth, here is an academic paper that should help.) In the context of FM, the important thing is to realize that instantaneous frequency follows naturally from the fact that the frequency of the carrier varies *continuously* in response to the modulating wave (i.e., the baseband signal). The instantaneous value of the baseband signal influences the frequency at a particular moment, not the frequency of one or more complete cycles.

Actually, though, this is true only of analog FM; with digital FM, one bit corresponds to a discrete number of cycles. This leads to the interesting situation in which the older technology (analog FM) is less intuitive than the newer technology (digital FM, also called frequency shift keying, or FSK).



You don't need to ponder instantaneous frequency in order to understand digital frequency modulation.

As in the previous page, we will write the carrier as sin(ω_Ct). It already has a frequency (namely, ω_C), so we will use the term *excess frequency* to refer to the frequency component contributed by the modulation procedure. This term is slightly misleading, though, because "excess" implies a higher frequency, whereas modulation can result in a carrier frequency that is higher or lower than the nominal carrier frequency. In fact, this is why frequency modulation (in contrast to amplitude modulation) does not require a shifted baseband signal: Positive baseband values increase the carrier frequency and negative baseband values decrease the carrier frequency. Under these conditions demodulation is not a problem, because all baseband values map to a unique frequency.

Anyways, back to our carrier signal: sin(ω_Ct). If we add the baseband signal (x_{BB}) to the quantity inside the parentheses, we are making the excess *phase* linearly proportional to the baseband signal. But we're looking for frequency modulation, not phase modulation, so we want the excess frequency to be linearly proportional to the baseband signal. We know from the first page of this chapter that we can obtain frequency by taking the derivative, with respect to time, of phase. Thus, if we want the *frequency* to be proportional to x_{BB}, we have to add not the baseband signal itself but rather the integral of the baseband signal (because taking the derivative cancels out the integral, and we are left with x_{BB} as the excess frequency).

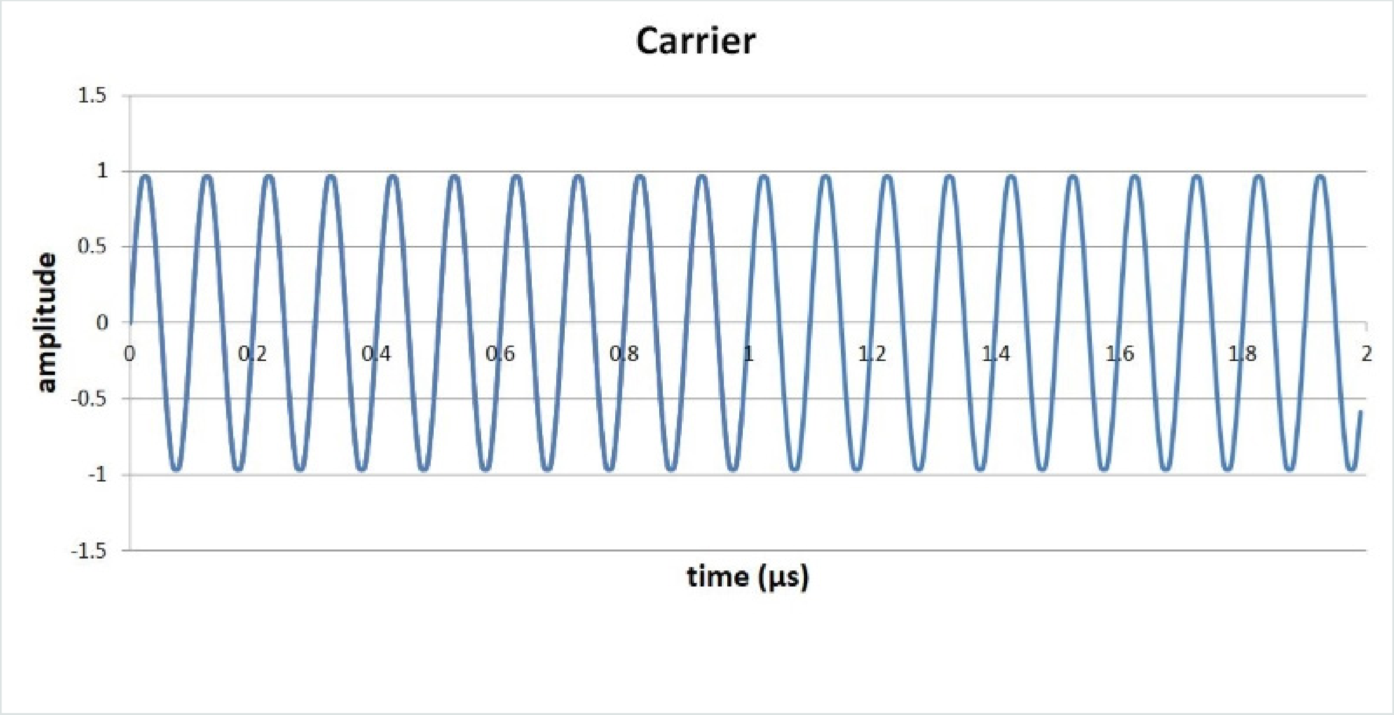
$$x_{FM}(t) = \sin(\omega_C t + \int_{-\infty}^{t} x_{BB}(t)dt)$$

The only thing we need to add here is the modulation index, m. In the previous page we saw that the modulation index can be used to make the carrier's amplitude variations more or less sensitive to the baseband-value variations. It's function in FM is equivalent: the modulation index allows us to fine-tune the intensity of the change in frequency that is produced by a change in the baseband value.

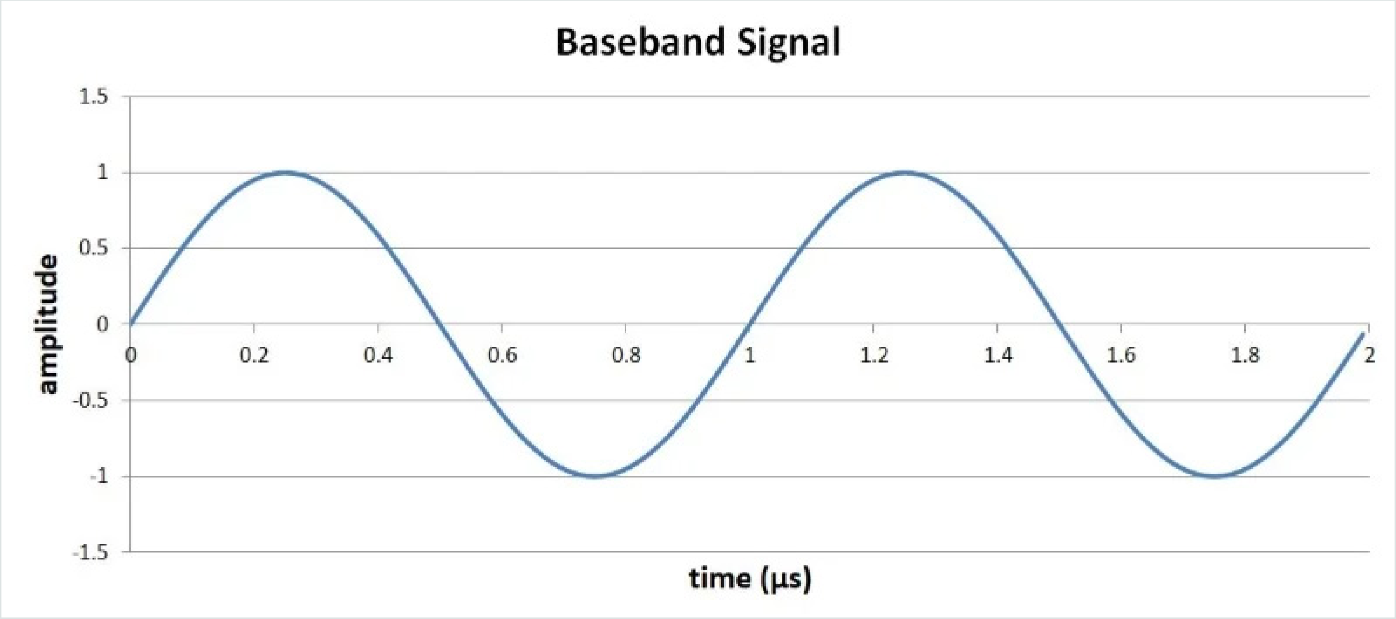
$$x_{FM}(t) = \sin(\omega_C t + m \int_{-\infty}^{t} x_{BB}(t)dt)$$

The Time Domain

Let's look at some waveforms. Here is our 10 MHz carrier:



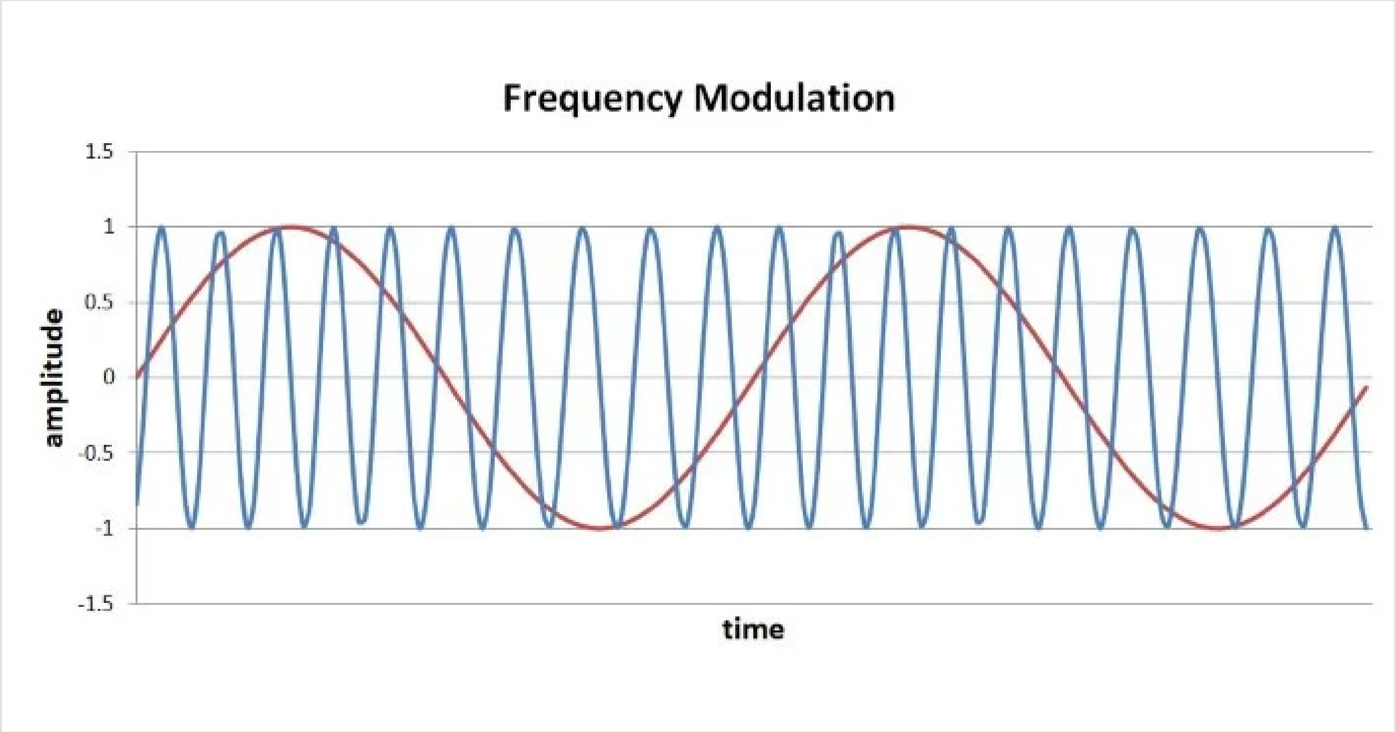
The baseband signal will be a 1 MHz sine wave, as follows:



The FM waveform is generated by applying the formula given above. The integral of sin(x) is –cos(x) + C. The constant C is not relevant here, so we can use the following equation to compute the FM signal:

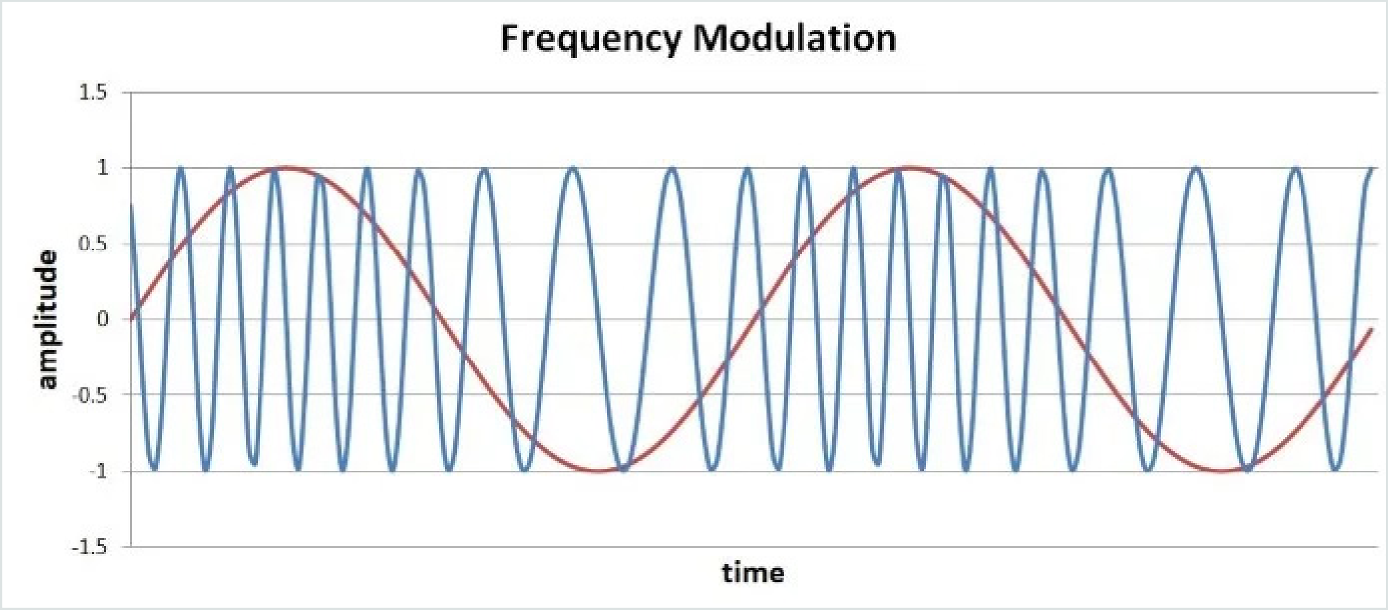
$$x_{FM}(t) = \sin((10 \times 10^6 \times 2\pi t) - \cos(1 \times 10^6 \times 2\pi t))$$

Here is the result (the baseband signal is shown in red):



It almost seems that the carrier hasn't changed, but if you look closely, the peaks are slightly closer together when the baseband signal is near its maximum value. So we do have frequency modulation here; the problem is that the baseband variations are not producing enough carrier-frequency variation. We can easily remedy this situation by increasing the modulation index. Let's use m = 4:

$$x_{FM}(t) = \sin((10 \times 10^6 \times 2\pi t) - 4\cos(1 \times 10^6 \times 2\pi t))$$

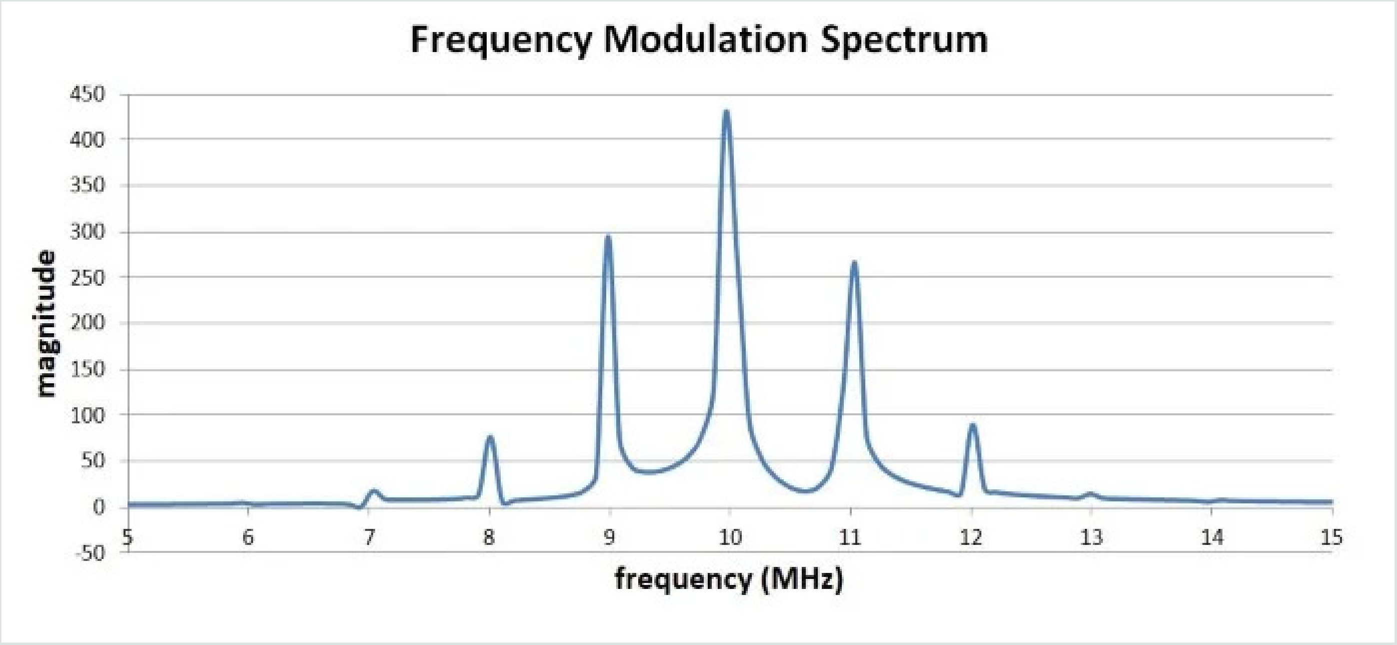


Now we can see more clearly how the frequency of the modulated carrier continuously tracks the instantaneous baseband value.

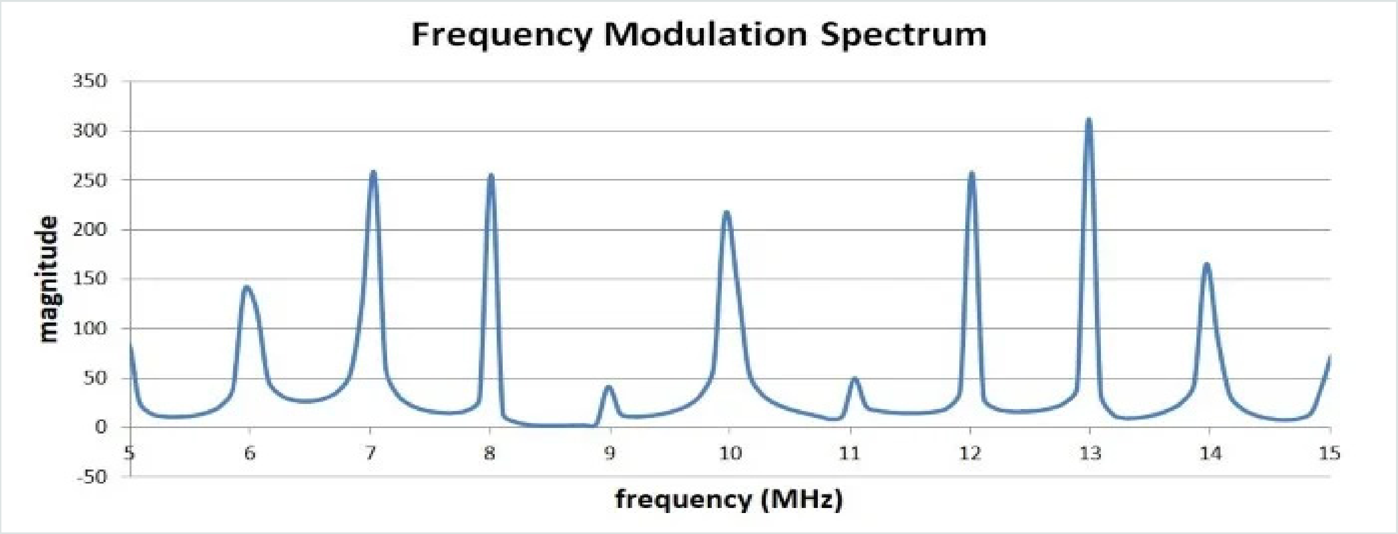
The Frequency Domain

AM and FM time-domain waveforms for the same baseband and carrier signals look very different. It is interesting, then, to find that AM and narrowband FM produce similar changes in the frequency domain. (Narrowband FM involves a limited modulating bandwidth and allows for easier analysis.) In both cases a low-frequency spectrum (including the negative frequencies) is translated to a band that extends above and below the carrier frequency. With AM, the baseband spectrum itself is shifted upwards. With FM, it is the spectrum of the integral of the baseband signal that appears in the band surrounding the carrier frequency.

For the single-baseband-frequency, m-equals-1 modulation shown above, we have the following:



The next spectrum is with m = 4:



This demonstrates very clearly that the modulation index influences the frequency content of the modulated waveform. Spectral analysis with frequency modulation is more complicated than it is with amplitude modulation; it is difficult to predict the bandwidth of frequency-modulated signals.

Summary

- The mathematical representation of frequency modulation consists of a sinusoidal expression with the integral of the baseband signal added to the argument of the sine or cosine function.

- The modulation index can be used to make the frequency deviation more sensitive or less sensitive to variations in the baseband value.
- Narrowband frequency modulation results in a translation of the spectrum of the integral of the baseband signal to a band surrounding the carrier frequency.
- An FM spectrum is influenced by the modulation index as well as by the ratio of the amplitude of the modulating signal to the frequency of the modulating signal.

Phase Modulation: Theory, Time Domain, Frequency Domain

Phase modulation is similar to frequency modulation and is an important technique in digital communication systems.

We have all heard of AM radio and FM radio. But phase modulation seems to be in a different category—"**PM** radio" is by no means a common term. It turns out that phase modulation is more relevant in the context of digital RF. In a way, though, we can say that PM radio is as common as FM radio simply because there is little difference between phase modulation and frequency modulation. FM and PM are best considered as two closely related variants of *angle modulation*, where "angle" refers to the modification of the quantity passed to a sine or cosine function.

The Math

We saw in the previous page that frequency modulation is achieved by adding the integral of the baseband signal to the argument of a sine or cosine function (where the sine or cosine function represents the carrier):

$$x_{FM}(t) = \sin(\omega_C t + \int_{-\infty}^{t} x_{BB}(t)dt)$$

You will recall, though, that we introduced frequency modulation by first discussing phase modulation: adding the baseband signal itself, rather than the integral of the baseband signal, causes the phase to vary according to the baseband value. Thus, phase modulation is actually a bit simpler than frequency modulation.

$$x_{PM}(t) = \sin(\omega_C t + x_{BB}(t))$$

As with frequency modulation, we can use the modulation index to make the phase variations more sensitive to the changes in the baseband value:

$$x_{PM}(t) = \sin(\omega_C t + m x_{BB}(t))$$

The similarity between phase modulation and frequency modulation becomes clear if we consider a single-frequency baseband signal. Let's say that $x_{BB}(t) = \sin(\omega_{BB}t)$. The integral of sine is negative cosine (plus a constant, which we can ignore here)—in other words, the integral is simply a time-shifted version of the original signal. Thus, if we perform phase modulation and frequency modulation with this baseband signal, the only difference in the modulated waveforms will be the alignment between the baseband value and the variations in the carrier; the variations themselves are the same. This will be more clear in the next section, where we'll look at some time-domain plots.

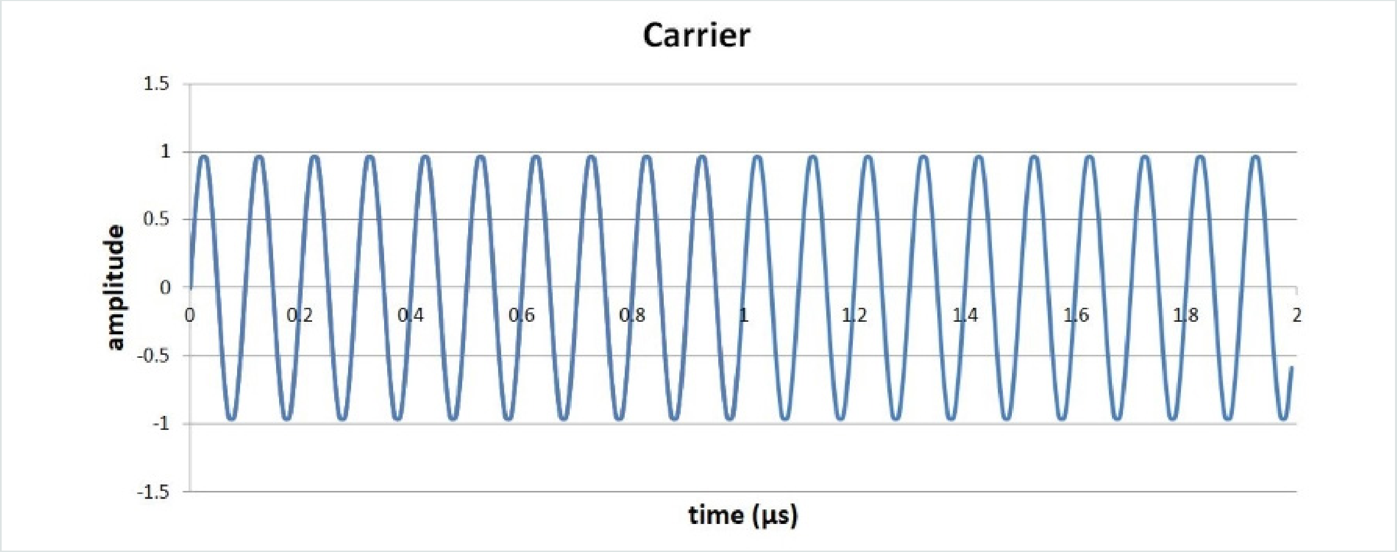
It's important to keep in mind that we're dealing with instantaneous phase, just as frequency modulation is based on the concept of instantaneous frequency. The term "phase" is rather vague. One familiar meaning refers to the initial state of a sinusoid; for example, a "normal" sine wave begins with a value of zero and then increases toward its maximum value. A sine wave that begins at a different point in its cycle has a phase offset. We can also think of phase as a specific portion of a full waveform cycle; for example, at a phase of π/2, a sinusoid has completed one-fourth of its cycle.

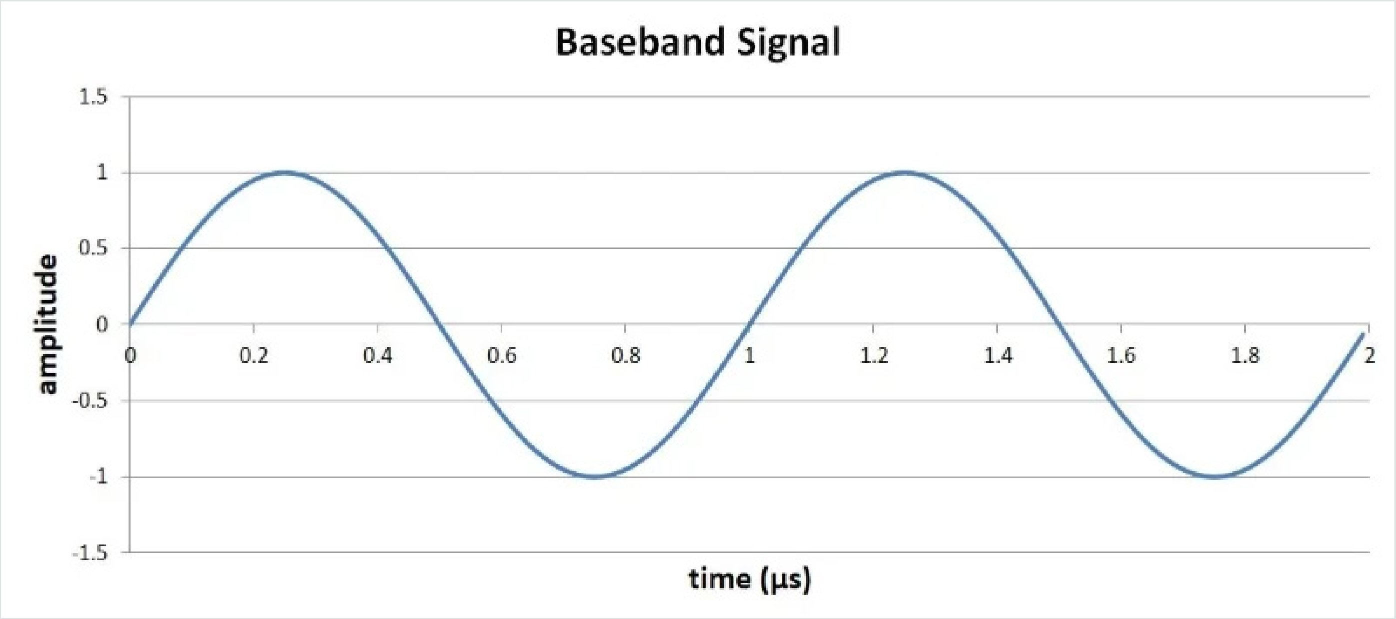
These interpretations of "phase" don't help us very much when we're dealing with a phase that continuously varies in response to a baseband waveform. Rather, we use the concept of *instantaneous* phase, i.e., the phase at a given moment, which corresponds to the value passed (at a given moment) to a trigonometric function. We can think of these continuous variations in instantaneous phase as "pushing" the carrier value farther from or closer to the preceding state of the waveform.

One more thing to keep in mind: Trig functions, including sine and cosine, operate on angles. Changing the argument of a trig function is equivalent to changing the angle, and this explains why both FM and PM are described as angle modulation.

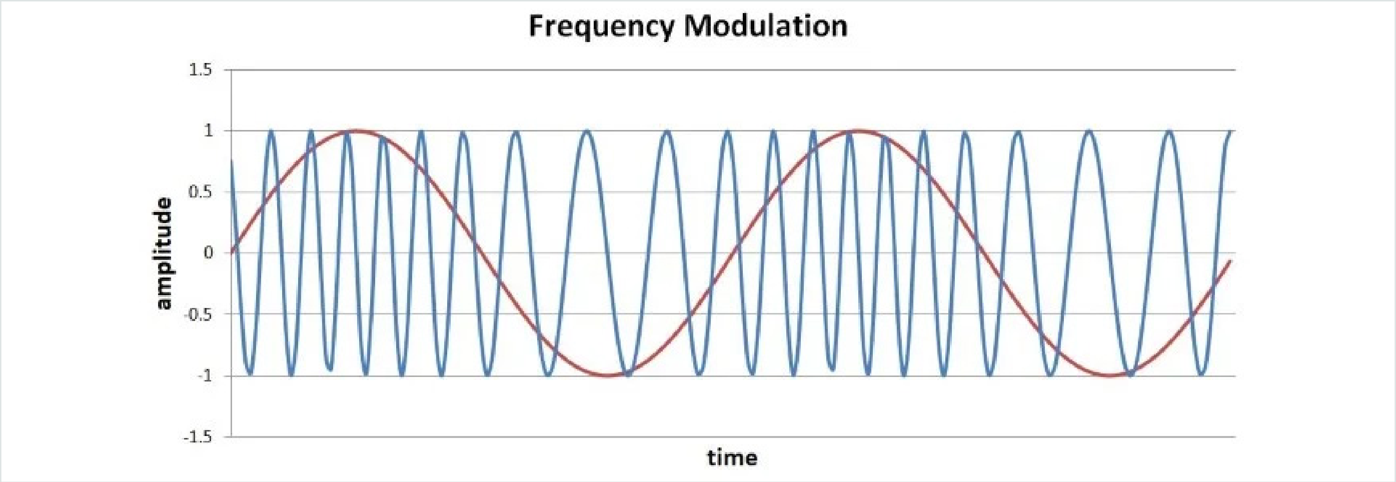
The Time Domain

We'll use the same waveforms that we used for the FM discussion, i.e., a 10 MHz carrier and a 1 MHz sinusoidal baseband signal:





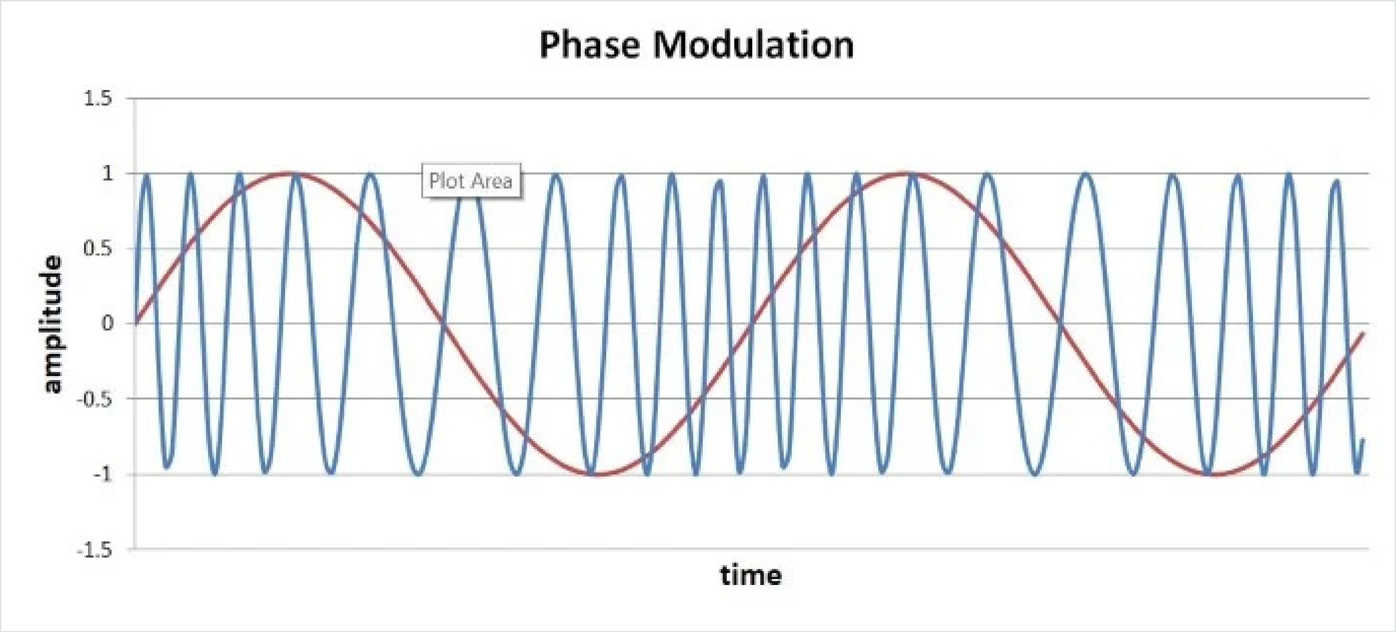
Here is the FM waveform (with m = 4) that we saw in the previous page:



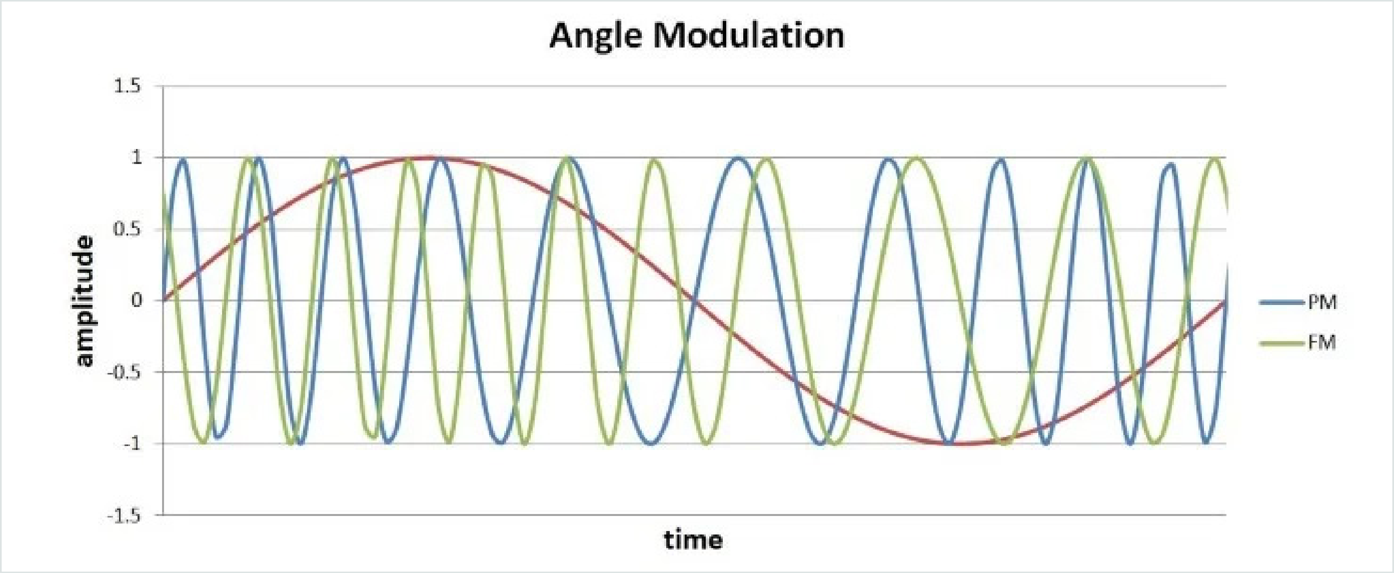
We can compute the PM waveform by using the following equation, where the signal added to the argument of the carrier wave uses positive sine (i.e., the original signal) instead of negative cosine (i.e., the integral of the original signal).

$$x_{PM}(t) = \sin((10 \times 10^6 \times 2\pi t) + \sin(1 \times 10^6 \times 2\pi t))$$

Here is the PM plot:



Before we discuss this, let’s look also at a plot that shows the FM waveform and the PM waveform:

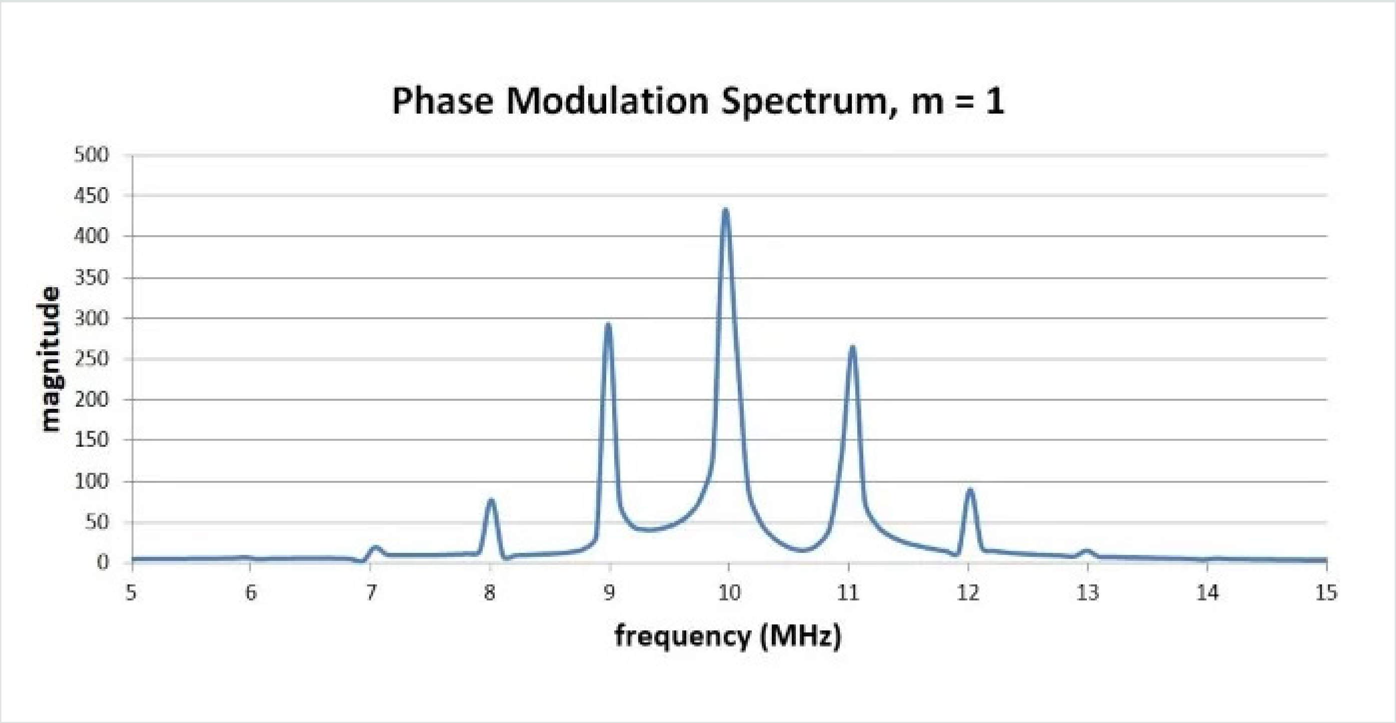


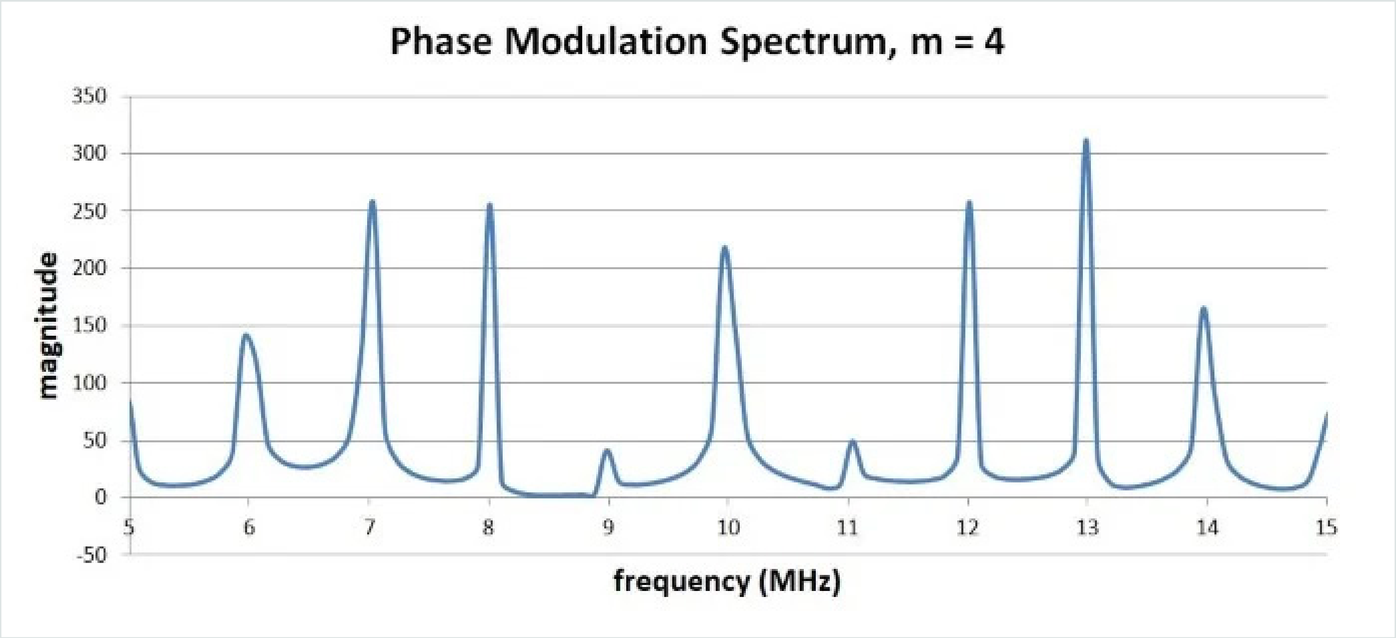
The first thing that comes to mind here is that from a visual standpoint, FM is more intuitive than PM—there is a clear visual connection between the higher- and lower-frequency sections of the modulated waveform and the higher and lower baseband values. With PM, the relationship between the baseband waveform and the behavior of the carrier is perhaps not immediately apparent. However, after a bit of inspection we can see that the PM carrier frequency corresponds to the *slope* of the baseband waveform; the highest-frequency sections occur during the steepest positive slope of x_{BB}, and the lowest-frequency sections occur during the steepest negative slope.

This makes sense: Recall that frequency (as a function of time) is the derivative of phase (as a function of time). With phase modulation, the slope of the baseband signal governs how quickly the phase changes, and the rate at which the phase changes is equivalent to frequency. So in a PM waveform, high baseband slope corresponds to high frequency, and low baseband slope corresponds to low frequency. With frequency modulation, we use the integral of x_{BB}, which has the effect of shifting the high- (or low-) frequency carrier sections to the baseband values *following* the high- (or low-) slope portions of the baseband waveform.

The Frequency Domain

The preceding time-domain plots demonstrate what was said previously: frequency modulation and phase modulation are quite similar. It is not surprising, then, that PM's effect in the frequency domain is similar to that of FM. Here are spectra for phase modulation with the carrier and baseband signals used above:





Summary

- Phase modulation is calculated by adding the baseband signal to the argument of a sine or cosine function that represents the carrier.
- The modulation index makes the phase variations more or less sensitive to the behavior of the baseband signal.
- The frequency-domain effects of phase modulation are similar to those of frequency modulation.
- Analog phase modulation is not common; however, digital phase modulation is widely used.

Digital Modulation: Amplitude and Frequency

Though based on the same concepts, digital-modulation waveforms look quite different from their analog counterparts.

Though far from extinct, analog modulation is simply incompatible with a digital world. We no longer focus our efforts on moving analog waveforms from one place to another. Rather, we want to move *data:* wireless networking, digitized audio signals, sensor measurements, and so forth. To transfer digital data, we use digital modulation.

We have to be careful, though, with this terminology. “Analog” and “digital” in this context refer to the type of information being transferred, not to the basic characteristics of the actual transmitted waveforms. Both analog and digital modulation use smoothly varying signals; the difference is that an analog-modulated signal is demodulated into an analog baseband waveform, whereas a digitally modulated signal consists of discrete modulation units, called *symbols*, that are interpreted as digital data.

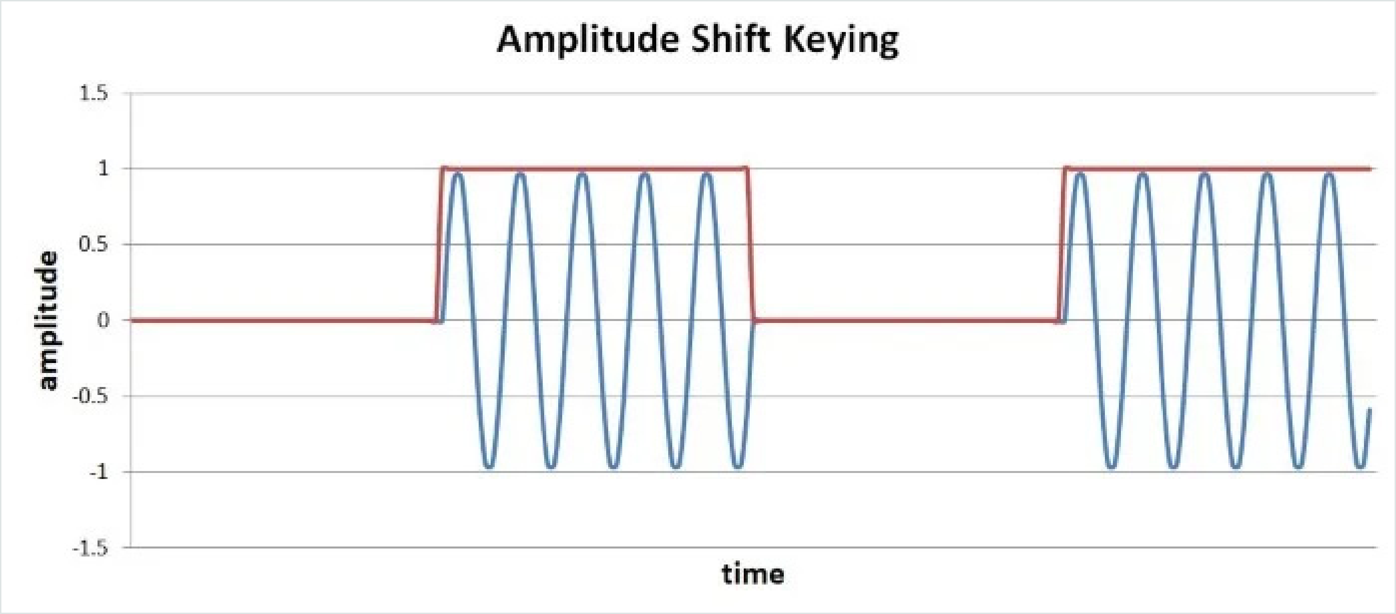
There are analog and digital versions of the three modulation types. Let’s start with amplitude and frequency.

Digital Amplitude Modulation

This type of modulation is referred to as amplitude shift keying (ASK). The most basic case is “on-off keying” (OOK), and it corresponds almost directly to the mathematical relationship discussed in the page dedicated to [[analog amplitude modulation]]: If we use a digital signal as the baseband waveform, multiplying the baseband and the carrier results in a modulated waveform that is normal for logic high and “off” for logic low. The logic-high amplitude corresponds to the modulation index.

Time Domain

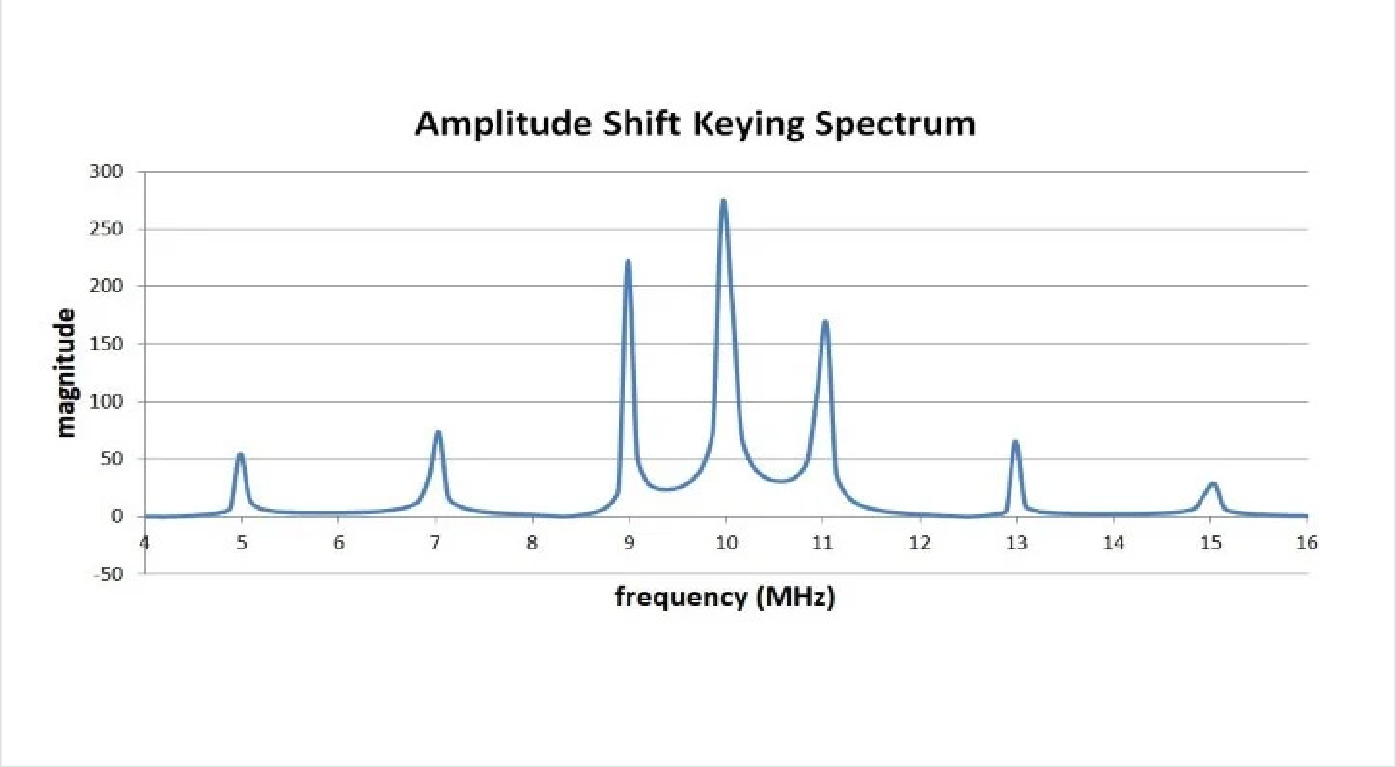
The following plot shows OOK generated using a 10 MHz carrier and a 1 MHz digital clock signal. We’re operating in the mathematical realm here, so the logic-high amplitude (and the carrier amplitude) is simply dimensionless “1”; in a real circuit you might have a 1 V carrier waveform and a 3.3 V logic signal.



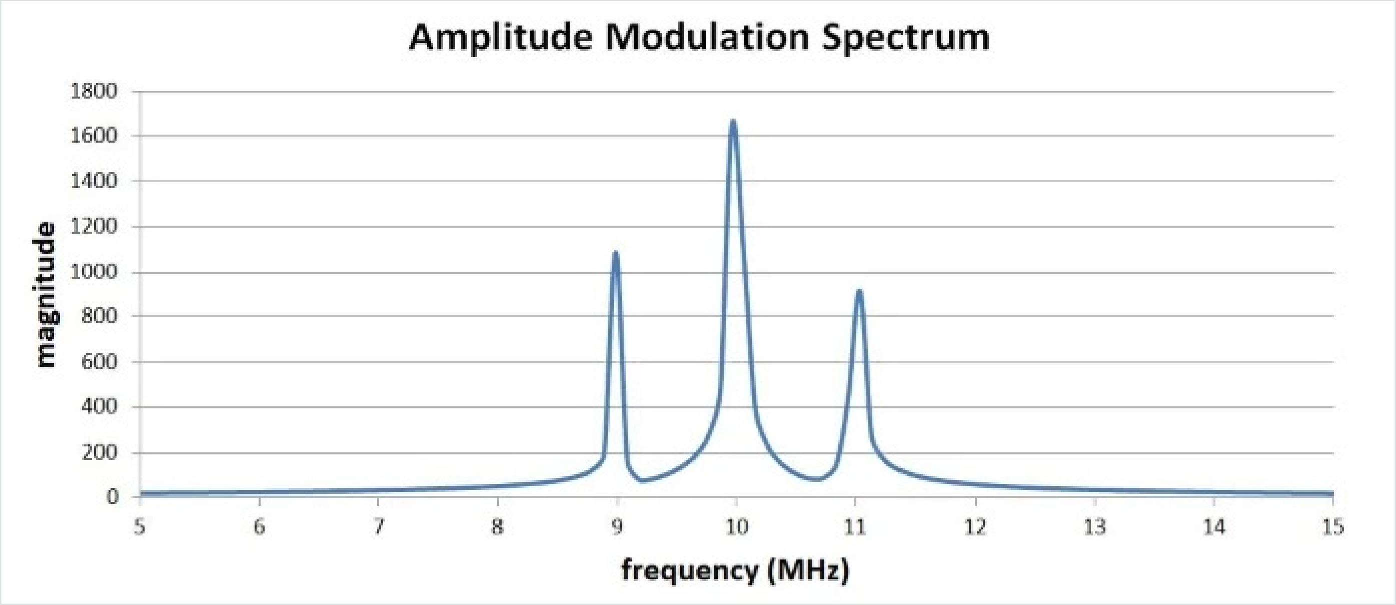
You may have noticed one inconsistency between this example and the mathematical relationship discussed in the [[Amplitude Modulation]] page: we didn't shift the baseband signal. If you're dealing with a typical DC-coupled digital waveform, no upward shifting is necessary because the signal remains in the positive portion of the y-axis.

Frequency Domain

Here is the corresponding spectrum:



Compare this to the spectrum for amplitude modulation with a 1 MHz sine wave:



Most of the spectrum is the same—a spike at the carrier frequency (f_C) and a spike at f_C plus the baseband frequency and f_C minus the baseband frequency. However, the ASK spectrum also has smaller spikes that correspond to the 3rd and 5th harmonics: The fundamental frequency (f_F) is 1 MHz, which means that the 3rd harmonic (f_3) is 3 MHz and the 5th harmonic (f_5) is 5 MHz. So we have spikes at f_C plus/minus f_F, f_3, and f_5. And actually, if you were to expand the plot, you would see that the spikes continue according to this pattern.

This makes perfect sense. A Fourier transform of a square wave consists of a sine wave at the fundamental frequency along with decreasing-amplitude sine waves at the odd harmonics, and this harmonic content is what we see in the spectrum shown above.

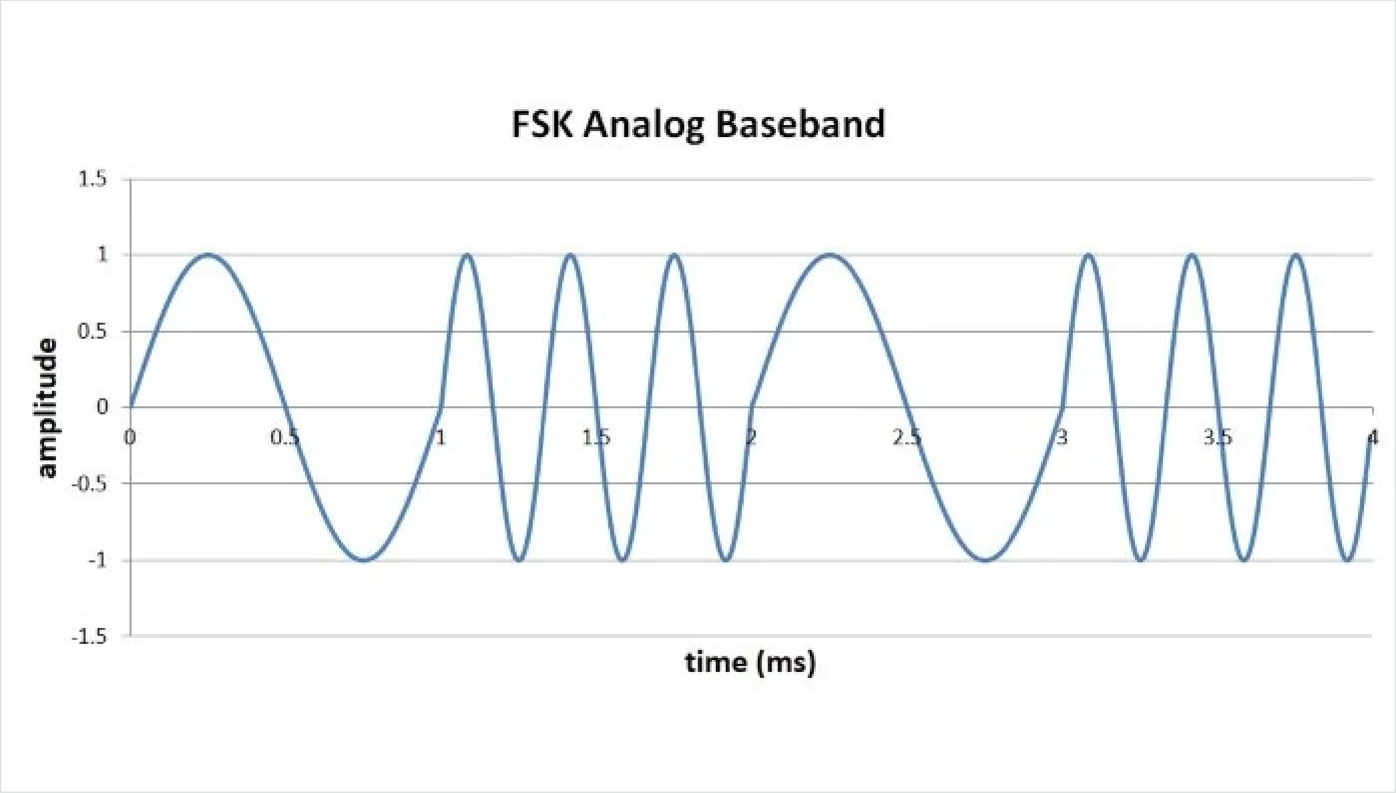
This discussion leads us to an important practical point: abrupt transitions associated with digital modulation schemes produce (undesirable) higher-frequency content. We have to keep this in mind when we consider the actual bandwidth of the modulated signal and the presence of frequencies that could interfere with other devices.

Digital Frequency Modulation

This type of modulation is called frequency shift keying (FSK). For our purposes it is not necessary to consider a mathematical expression of FSK; rather, we can simply specify that we will have frequency f_1 when the baseband data is logic 0 and frequency f_2 when the baseband data is logic 1.

Time Domain

One method of generating the ready-for-transmission FSK waveform is to first create an analog baseband signal that switches between f_1 and f_2 according to the digital data. Here is an example of an FSK baseband waveform with f_1 = 1 kHz and f_2 = 3 kHz. To ensure that a symbol is the same duration for logic 0 and logic 1, we use one 1 kHz cycle and three 3 kHz cycles.

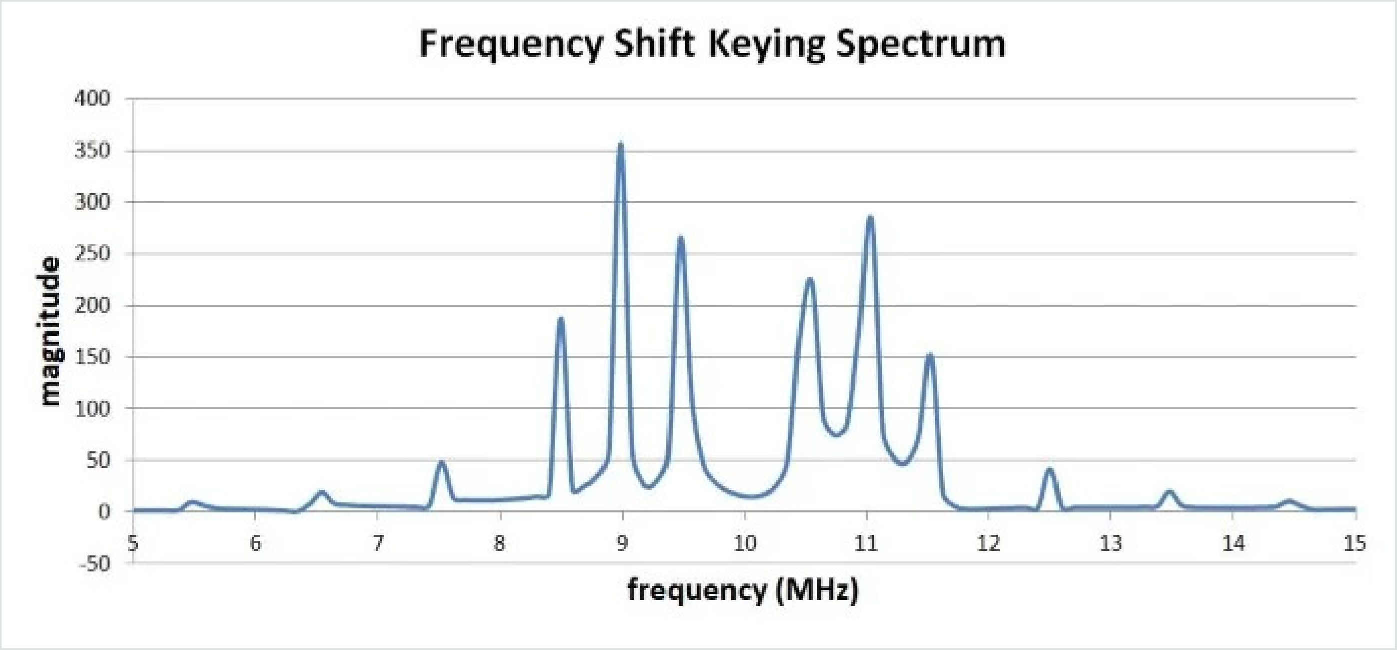


The baseband waveform is then shifted (using a mixer) up to the carrier frequency and transmitted. This approach is particularly handy in software-defined-radio systems: the analog baseband waveform is a low-frequency signal, and thus it can be generated mathematically then introduced into the analog realm by a DAC. Using a DAC to create the high-frequency transmitted signal would be much more difficult.

A more conceptually straightforward way to implement FSK is to simply have two carrier signals with different frequencies (f_1 and f_2); one or the other is routed to the output depending on the logic level of the binary data. This results in a final transmitted waveform that switches abruptly between two frequencies, much like the baseband FSK waveform above except that the difference between the two frequencies is much smaller in relation to the average frequency. In other words, if you were looking at a time-domain plot, it would be difficult to visually differentiate the f_1 sections from the f_2 sections because the difference between f_1 and f_2 is only a tiny fraction of f_1 (or f_2).

Frequency Domain

Let's look at the effects of FSK in the frequency domain. We'll use our same 10 MHz carrier frequency (or average carrier frequency in this case), and we'll use ±1 MHz as the deviation. (This is unrealistic, but convenient for our current purposes.) So the transmitted signal will be 9 MHz for logic 0 and 11 MHz for logic 1. Here is the spectrum:



Note that there is no energy at the "carrier frequency." This is not surprising, considering that the modulated signal is never at 10 MHz. It is always at either 10 MHz minus 1 MHz or 10 MHz plus 1 MHz, and this is precisely where we see the two dominant spikes: 9 MHz and 11 MHz.

But what about the other frequencies present in this spectrum? Well, FSK spectral analysis is not particularly straightforward. We know that there will be additional Fourier energy associated with the abrupt transitions between frequencies. It turns out that FSK results in a sinc-function type of spectrum for each frequency, i.e., one is centered on f_1 and the other is centered on f_2. These account for the additional frequency spikes seen on either side of the two dominant spikes.

Summary

- Digital amplitude modulation involves varying the amplitude of a carrier wave in discrete sections according to binary data.
- The most straightforward approach to digital amplitude modulation is on-off keying.

- With digital frequency modulation, the frequency of a carrier or a baseband signal is varied in discrete sections according to binary data.
- If we compare digital modulation to analog modulation, we see that the abrupt transitions created by digital modulation result in additional energy at frequencies farther from the carrier.

Digital Phase Modulation:BPSK, QPSK, DQPSK

Digital phase modulation is a versatile and widely used method of wirelessly transferring digital data.

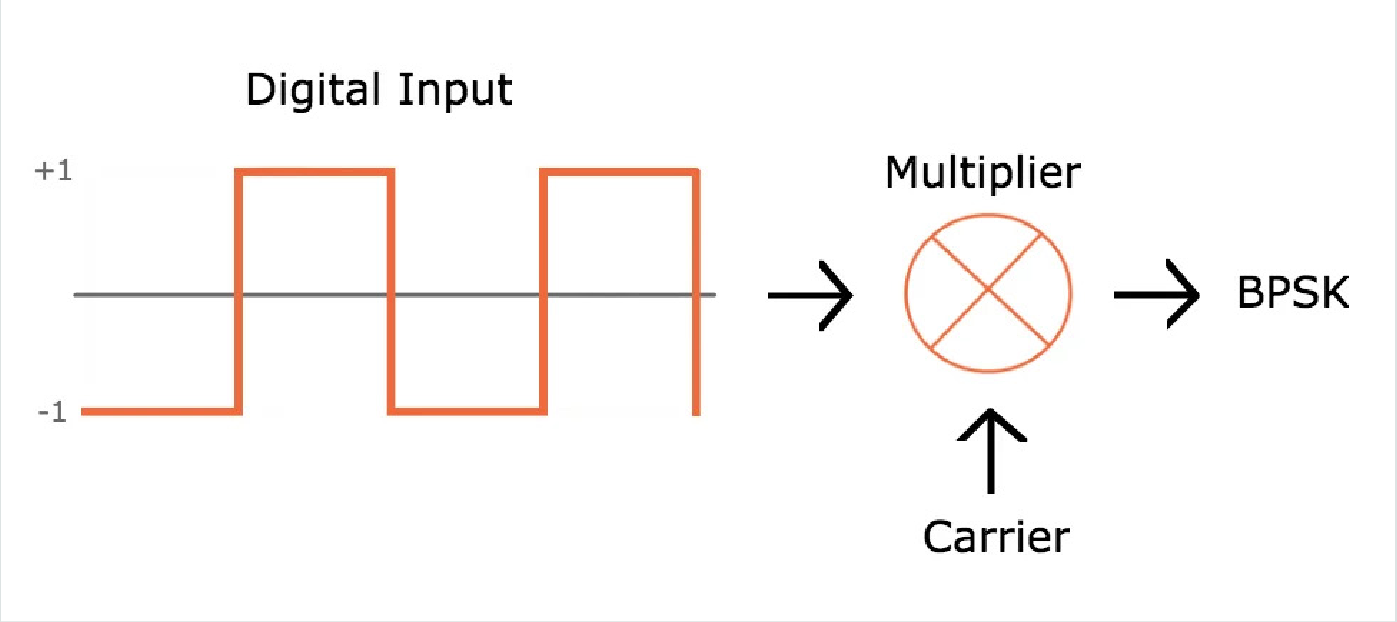
In the previous page, we saw that we can use discrete variations in a carrier's amplitude or frequency as a way of representing ones and zeros. It should come as no surprise that we can also represent digital data using phase; this technique is called phase shift keying (PSK).

Binary Phase Shift Keying

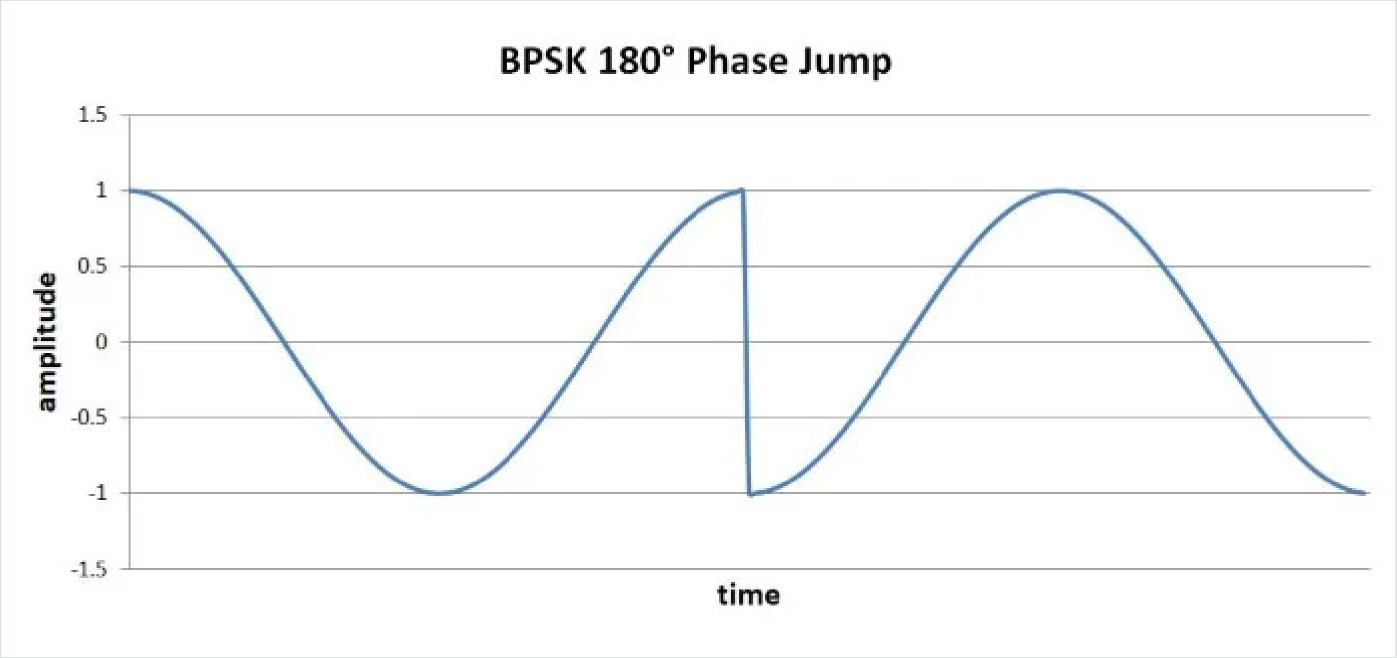
The most straightforward type of PSK is called binary phase shift keying (BPSK), where "binary" refers to the use of two phase offsets (one for logic high, one for logic low).

We can intuitively recognize that the system will be more robust if there is greater separation between these two phases—of course it would be difficult for a receiver to distinguish between a symbol with a phase offset of 90° and a symbol with a phase offset of 91°. We only have 360° of phase to work with, so the maximum difference between the logic-high and logic-low phases is 180°. But we know that shifting a sinusoid by 180° is the same as inverting it; thus, we can think of BPSK as simply inverting the carrier in response to one logic state and leaving it alone in response to the other logic state.

To take this a step further, we know that multiplying a sinusoid by negative one is the same as inverting it. This leads to the possibility of implementing BPSK using the following basic hardware configuration:

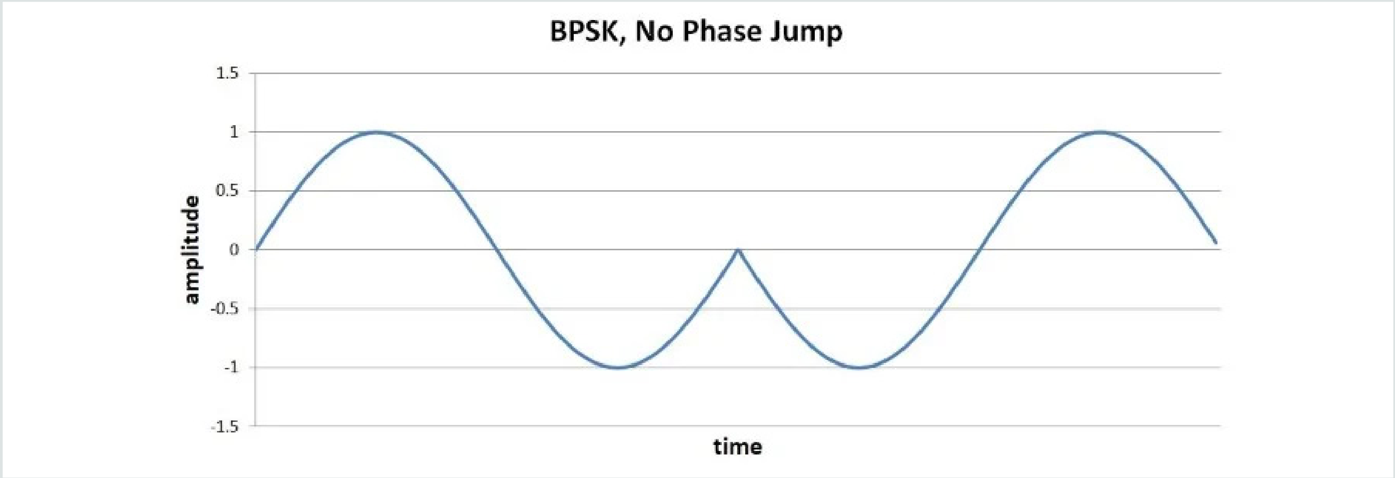


However, this scheme could easily result in high-slope transitions in the carrier waveform: if the transition between logic states occurs when the carrier is at its maximum value, the carrier voltage has to rapidly move to the minimum voltage.



High-slope events such as these are undesirable because they generate higher-frequency energy that could interfere with other RF signals. Also, amplifiers have limited ability to produce high-slope changes in output voltage.

If we refine the above implementation with two additional features, we can ensure smooth transitions between symbols. First, we need to ensure that the digital bit period is equal to one or more complete carrier cycles. Second, we need to synchronize the digital transitions with the carrier waveform. With these improvements, we could design the system such that the 180° phase change occurs when the carrier signal is at (or very near) the zero-crossing.

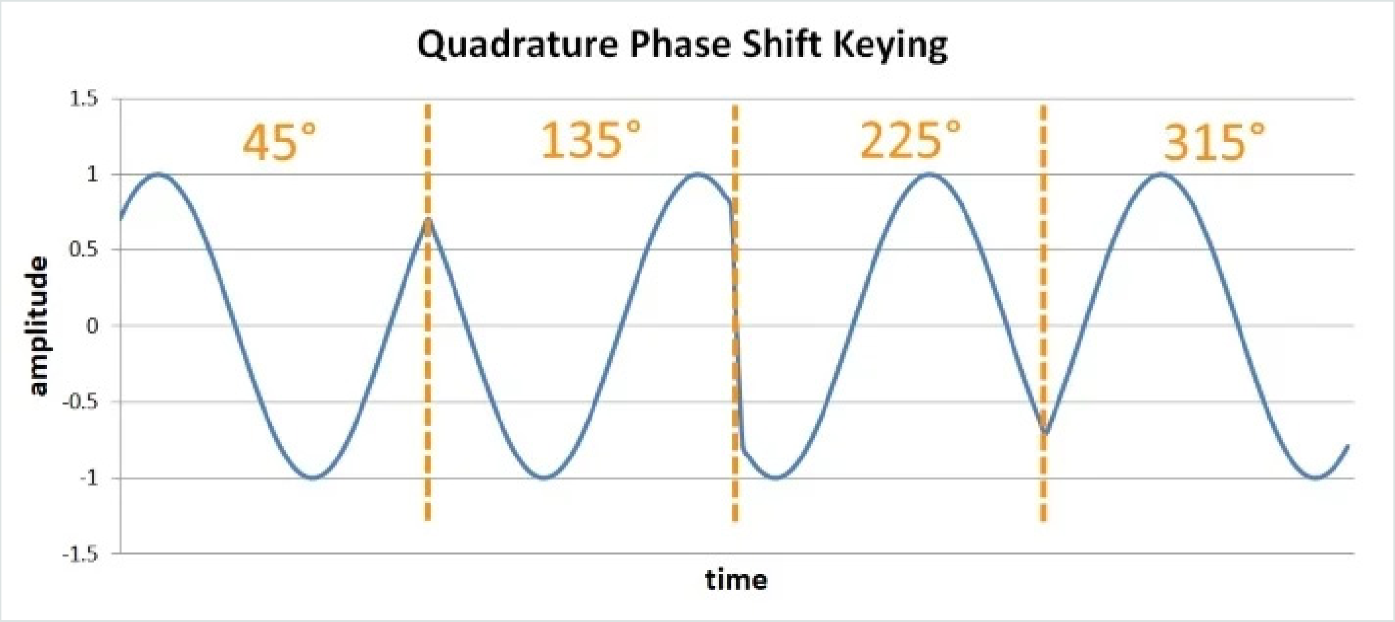


QPSK

BPSK transfers one bit per symbol, which is what we're accustomed to so far. Everything we've discussed with regard to digital modulation has assumed that the carrier signal is modified according to whether a digital voltage is logic low or logic high, and the receiver constructs digital data by interpreting each symbol as either a 0 or a 1.

Before we discuss quadrature phase shift keying (QPSK), we need to introduce the following important concept: There is no reason why one symbol can transfer only one bit. It's true that the world of digital electronics is built around circuitry in which the voltage is at one extreme or the other, such that the voltage always represents one digital bit. But RF is not digital; rather, we're using *analog* waveforms to transfer *digital* data, and it is perfectly acceptable to design a system in which the analog waveforms are encoded and interpreted in a way that allows one symbol to represent two (or more) bits.

QPSK is a modulation scheme that allows one symbol to transfer two bits of data. There are four possible two-bit numbers (00, 01, 10, 11), and consequently we need four phase offsets. Again, we want maximum separation between the phase options, which in this case is 90°.



The advantage is higher data rate: if we maintain the same symbol period, we can double the rate at which data is moved from transmitter to receiver. The downside is system complexity. (You might think that QPSK is also significantly more susceptible to bit errors than BPSK, since there is less separation between the possible phase values. This is a reasonable assumption, but if you go through the math it turns out that the error probabilities are actually very similar.)

Variants

QPSK is, overall, an effective modulation scheme. But it can be improved.

Phase Jumps

Standard QPSK guarantees that high-slope symbol-to-symbol transitions will occur; because the phase jumps can be ±90°, we can't use the approach described for the 180° phase jumps produced by BPSK modulation.

This problem can be mitigated by using one of two QPSK variants. Offset QPSK, which involves adding a delay to one of two digital data streams used in the modulation process, reduces the maximum phase jump to 90°. Another option is π/4-QPSK, which reduces the maximum phase jump to 135°. Offset QPSK is thus superior with respect to reducing phase discontinuities, but π/4-QPSK is advantageous because it is compatible with differential encoding (discussed in the next subsection).

Another way to deal with symbol-to-symbol discontinuities is to implement additional signal processing that creates smoother transitions between symbols. This approach is incorporated into a modulation scheme called minimum shift keying (MSK), and there is also an improvement on MSK known as Gaussian MSK.

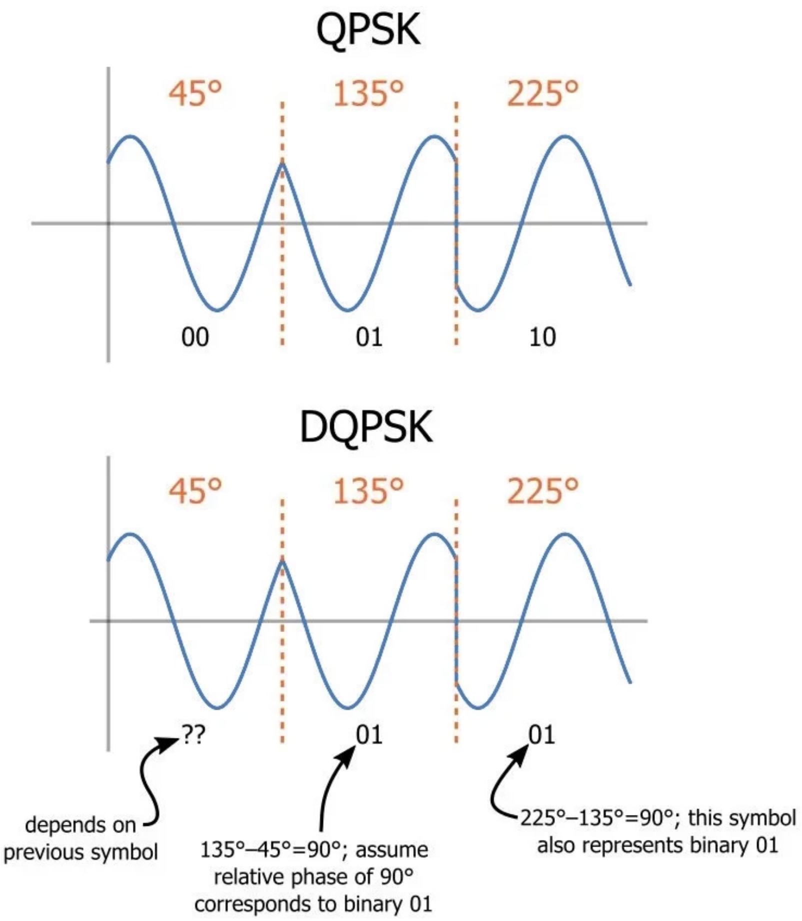
Differential Encoding

Another difficulty is that demodulation with PSK waveforms is more difficult than with FSK waveforms. Frequency is "absolute" in the sense that frequency changes can always be interpreted by analyzing the signal variations with respect to time. Phase, however, is relative in the sense that it has no universal reference—the transmitter generates the phase variations with reference to a point in time, and the receiver might interpret the phase variations with reference to a separate point in time.

The practical manifestation of this is the following: If there are differences between the phase (or frequency) of the oscillators used for modulation and demodulation, PSK becomes unreliable. And we have to assume that there will be phase differences (unless the receiver incorporates carrier-recovery circuitry).

Differential QPSK (DQPSK) is a variant that is compatible with noncoherent receivers (i.e., receivers that don't synchronize the demodulation oscillator with the modulation oscillator). Differential QPSK encodes data by producing a certain phase shift *relative to the preceding symbol.*

By using the phase of the preceding symbol in this way, the demodulation circuitry analyzes the phase of a symbol using a reference that is common to the receiver and the transmitter.


QPSK
45°
135°
225°
00
01
10
DQPSK
45°
135°
225°
??
01
01
depends on previous symbol
135°–45°=90°; assume relative phase of 90° corresponds to binary 01
225°–135°=90°; this symbol also represents binary 01

Summary

- Binary phase shift keying is a straightforward modulation scheme that can transfer one bit per symbol.
- Quadrature phase shift keying is more complex but doubles the data rate (or achieves the same data rate with half the bandwidth).
- Offset QPSK, π/4-QPSK, and minimum shift keying are modulation schemes that mitigate the effects of high-slope symbol-to-symbol voltage changes.
- Differential QPSK uses the phase difference between adjacent symbols to avoid problems associated with a lack of phase synchronization between the transmitter and receiver.

Comparing and Contrasting Amplitude, Frequency, and Phase Modulation

How do the different modulation schemes compare in terms of performance and applications? Let's take a look.

It's important to understand the salient characteristics of the three types of RF modulation. But this information doesn't exist in isolation—the goal is to design real systems that effectively and efficiently meet the performance objectives. Thus, we need to have a general idea of which modulation scheme is appropriate for a particular application.

Amplitude Modulation

Amplitude modulation is straightforward in terms of implementation and analysis. Also, AM waveforms are fairly easy to demodulate. Overall, then, AM can be viewed as a simple, low-cost modulation scheme. As usual, though, simplicity and low cost are accompanied by performance compromises—we never expect the easier, cheaper solution to be the best one.

It may not be accurate to describe AM systems as "rare," since countless vehicles all over the world include AM receivers. However, the applications of analog amplitude modulation are currently quite limited, because AM has two significant disadvantages.

In addition to AM radio broadcasting, analog amplitude modulation is used in civil aviation.

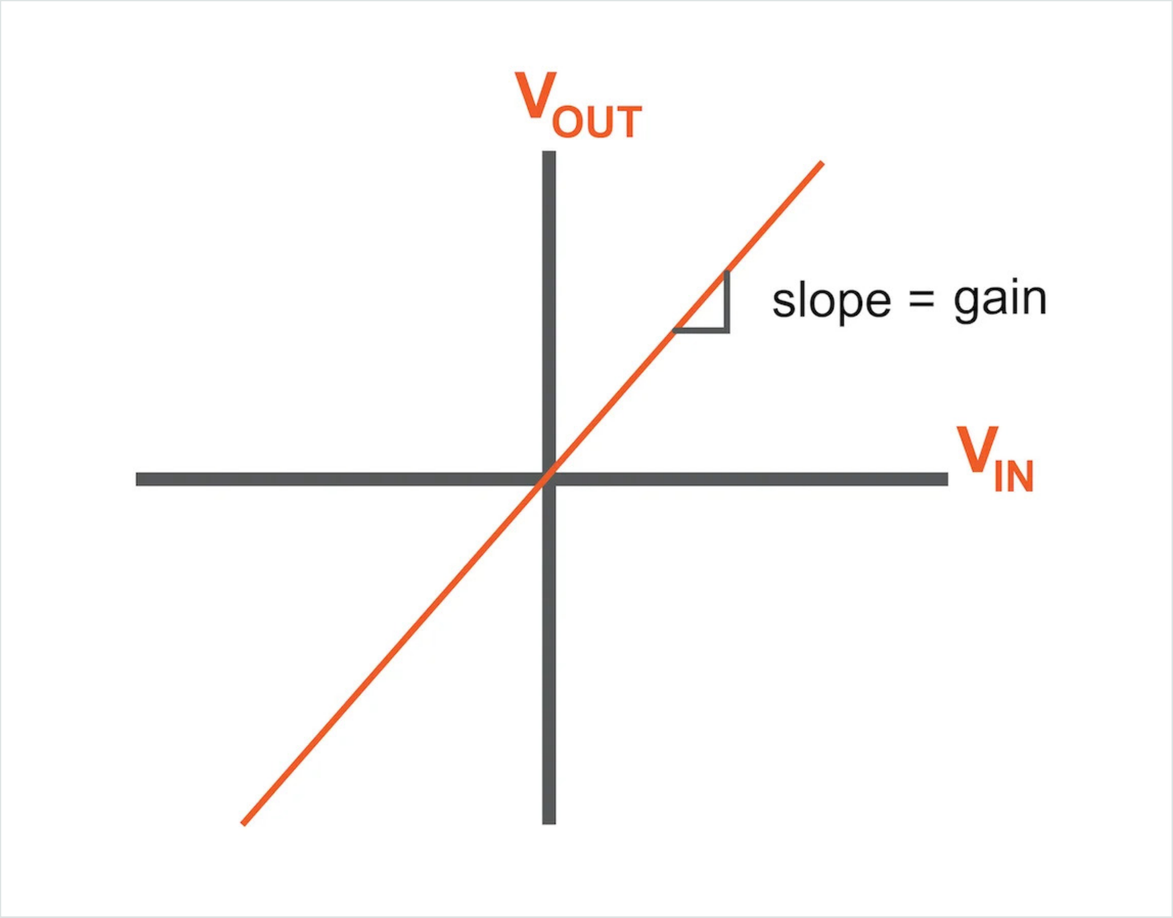
Amplitude Noise

Noise is a perpetual difficulty in wireless communication systems. In a certain sense, the quality of an RF design can be summarized by the signal-to-noise ratio of the demodulated signal: less noise in the received signal means higher quality output (for analog systems) or fewer bit errors (for digital systems). Noise is always present, and we always have to recognize it as a fundamental threat to the overall performance of the system.

Noise—random electrical noise, interference, electrical and mechanical transients—operates on the magnitude of a signal. In other words, noise can create amplitude modulation. This is a problem, since the random amplitude modulation resulting from noise cannot be distinguished from the intentional amplitude modulation performed by the transmitter. Noise is a problem for any RF signal, but AM systems are particularly susceptible.

Amplifier Linearity

One of the primary challenges in the design of RF power amplifiers is linearity. (More specifically, it is difficult to achieve both high efficiency and high linearity.) A linear amplifier applies a certain fixed gain to the input signal; in graphical terms, the transfer function of a linear amplifier is simply a straight line, with the slope corresponding to the gain.



A straight line represents the response of a perfectly linear amplifier: the output voltage is always the input voltage multiplied by a fixed gain.

Real-life amplifiers always have some degree of nonlinearity, meaning that the gain applied to the input signal is affected by the characteristics of the input signal. The result of nonlinear amplification is distortion, i.e., the creation of spectral energy at harmonic frequencies.

We can also say that nonlinear amplification is a form of amplitude modulation. If the gain of an amplifier varies according to the frequency of the input signal, or according to external factors such as temperature or power-supply conditions, the transmitted signal is experiencing unintended (and undesirable) amplitude modulation. This is a problem in AM systems because the spurious amplitude modulation interferes with the intentional amplitude modulation.

Any modulation scheme that incorporates amplitude variations is more susceptible to the effects of nonlinearity. This includes both ordinary analog amplitude modulation and the widely used digital schemes known collectively as quadrature amplitude modulation (QAM).

Angle Modulation

Frequency and phase modulation encode information in the temporal characteristics of the transmitted signal, and consequently they are robust against amplitude noise and amplifier nonlinearity. The frequency of a signal cannot be changed by noise or distortion. Additional frequency content may be added, but the original frequency will still be present. Noise does, of course, have negative effects on FM and PM systems, but the noise is not directly corrupting the signal characteristics that were used to encode the baseband data.

As mentioned above, power-amplifier design involves a trade-off between efficiency and linearity. Angle modulation is compatible with low-linearity amplifiers, and these low-linearity amplifiers are more efficient in terms of power consumption. Thus, angle modulation is a good choice for low-power RF systems.

Bandwidth

The frequency-domain effects of amplitude modulation are more straightforward than those of frequency and phase modulation. This can be considered an advantage of AM: it's important to be able to predict the bandwidth occupied by the modulated signal.

However, the difficulty of predicting the spectral characteristics of FM and PM is more relevant to the theoretical portion of the design. If we focus on practical considerations, angle modulation could be considered advantageous because it can translate a given baseband bandwidth to a somewhat smaller (compared to AM) transmission bandwidth.

Frequency vs. Phase

Frequency modulation and phase modulation are closely related; nevertheless, there are situations in which one is a better choice than the other. The differences between the two are more pronounced with digital modulation.

Analog Frequency and Phase Modulation

As we saw in the page on phase modulation, when the baseband signal is a sinusoid, a PM waveform is simply a shifted version of a corresponding FM waveform. It's not surprising, then, that there are no major FM vs. PM pros and cons related to spectral characteristics or noise susceptibility.

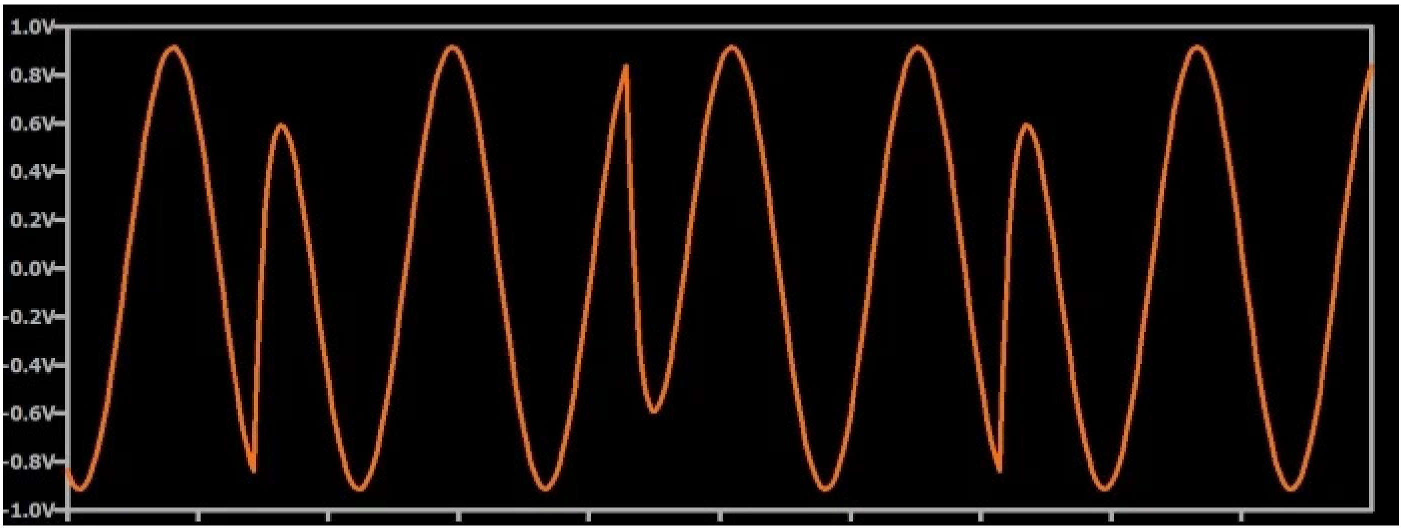
However, analog FM is much more common than analog PM, and the reason is that FM modulation and demodulation circuitry is more straightforward. For example, frequency modulation can be accomplished with something as simple as an oscillator built around an inductor and a voltage-controlled capacitor (i.e., a capacitor that experiences capacitance variations in response to the voltage of a baseband signal).

Digital Frequency and Phase Modulation

The differences between PM and FM become quite significant when we enter the realm of digital modulation. The first consideration is bit error rate. Obviously the bit error rate of any system will depend on various factors, but if we mathematically compare a binary PSK system to an equivalent binary FSK system, we find that binary FSK needs significantly more transmit energy to achieve the same bit error rate. This is an advantage of digital phase modulation.

But ordinary digital PM also has two significant disadvantages.

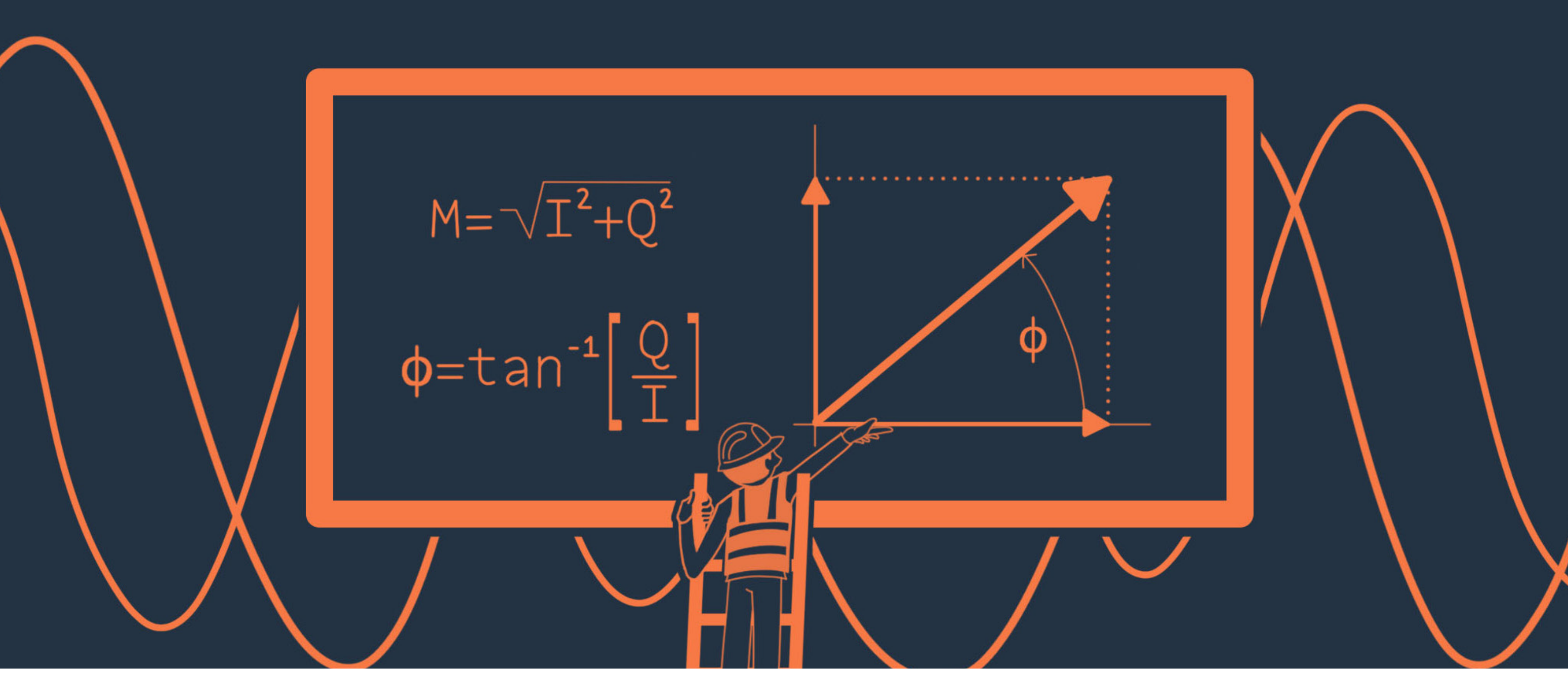
- As discussed in the digital phase modulation page, ordinary (i.e., non-differential) PSK is not compatible with noncoherent receivers. FSK, in contrast, does not require coherent detection.
- Ordinary PSK schemes, especially QPSK, involve abrupt phase changes that result in high-slope signal variations, and high-slope sections of the waveform decrease in amplitude when the signal is processed by a low-pass filter. These amplitude variations combined with nonlinear amplification lead to a problem called spectral regrowth. To mitigate spectral regrowth we can either use a more linear (and thus less efficient) power amplifier or implement a specialized version of PSK. Or we can switch to FSK, which doesn't require abrupt phase changes.



Here you can see amplitude variations caused by low-pass filtering a PSK signal.

Summary

- Amplitude modulation is simple, but it is susceptible to noise and requires a high-linearity power amplifier.
- Frequency modulation is less susceptible to amplitude noise and can be used with higher-efficiency, lower-linearity amplifiers.
- Digital phase modulation offers better theoretical performance in terms of bit error rate than digital frequency modulation, but digital FM is advantageous in low-power systems because it does not require a high-linearity amplifier.



CHAPTER 5

Radio Frequency Demodulation

- How to Demodulate an AM Waveform
- How to Demodulate an FM Waveform
- How to Demodulate Digital Phase Modulation
- Understanding I/Q Signals and Quadrature Modulation
- Understanding Quadrature Demodulation
- Quadrature Frequency and Phase Demodulation

How to Demodulate an AM Waveform

Learn about two circuits that can extract the original information from an amplitude-modulated carrier signal.

At this point we know that modulation refers to intentionally modifying a sinusoid such that it can carry lower-frequency information from a transmitter to a receiver. We also have covered many details related to the different methods—amplitude, frequency, phase, analog, digital—of encoding information in a carrier wave.

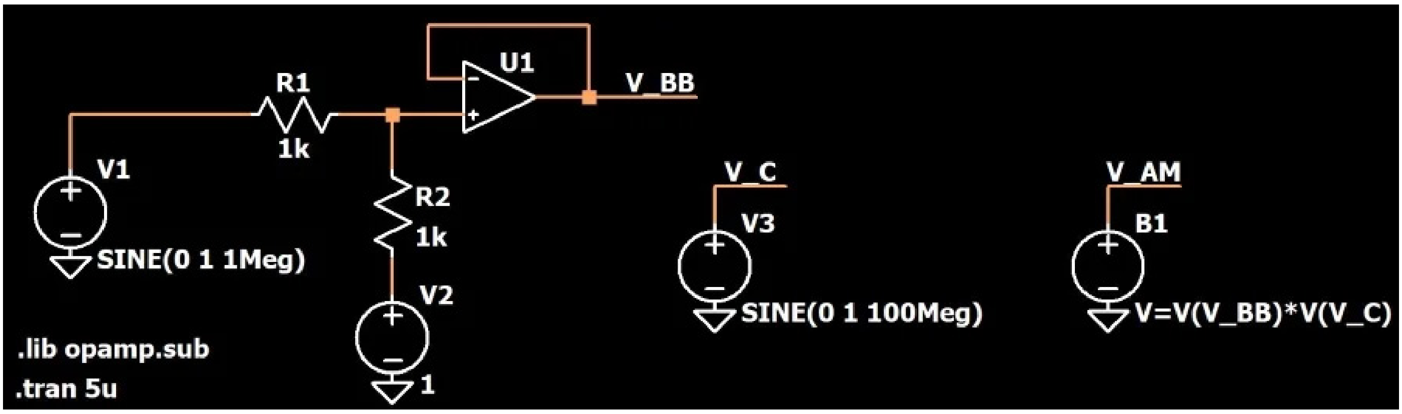
But there is no reason to integrate data into a transmitted signal if we cannot extract that data from the received signal, and this is why we need to study demodulation. Demodulation circuitry ranges from something as simple as a modified peak detector to something as complex as coherent quadrature downconversion combined with sophisticated decoding algorithms performed by a digital signal processor.

Creating the Signal

We'll use LTspice to study techniques for demodulating an AM waveform. But before we demodulate we need something that is modulated.

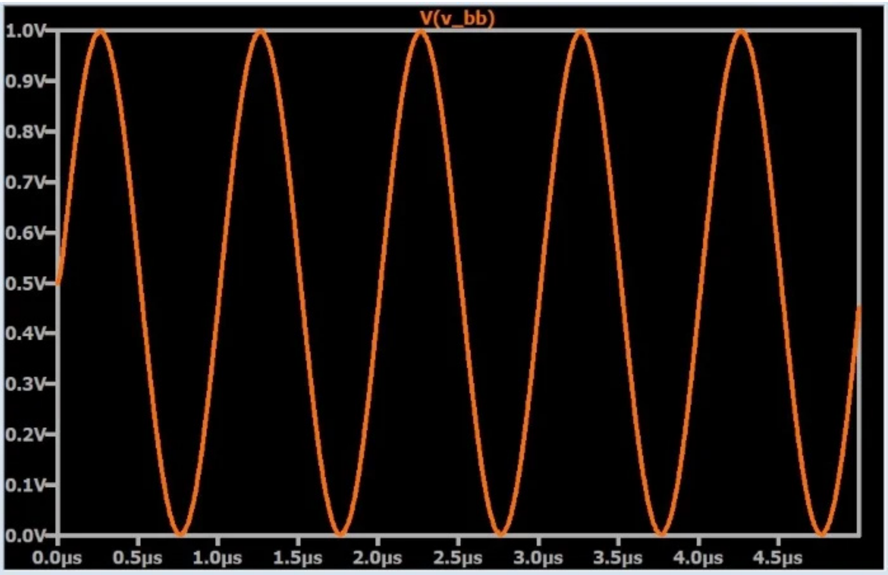
In the AM modulation page, we saw that four things are needed to generate an AM waveform. First, we need a baseband waveform and a carrier waveform. Then we need a circuit that can add an appropriate DC offset to the baseband signal. And finally, we need a multiplier, since the mathematical relationship corresponding to amplitude modulation is multiplying the shifted baseband signal by the carrier.

The following LTspice circuit will generate an AM waveform.

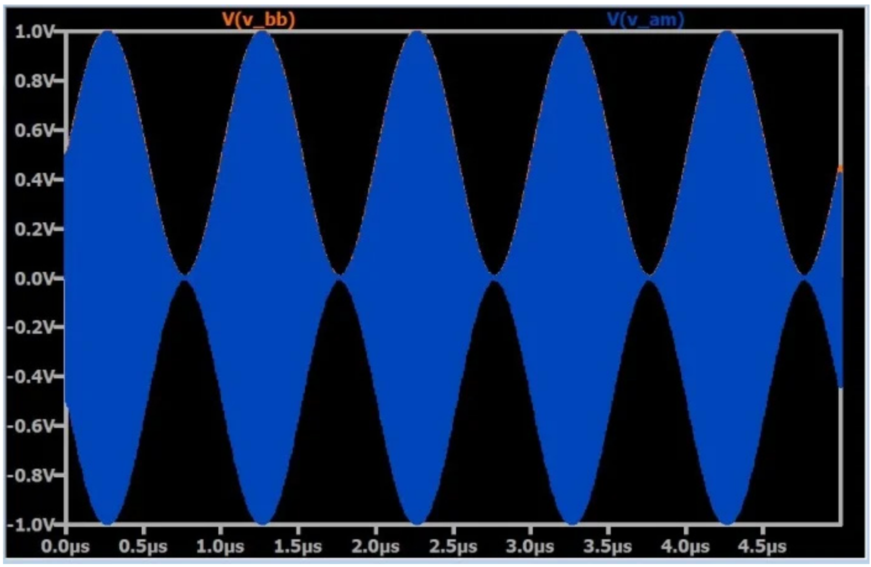


- V1 is a 1 MHz sine wave voltage source that provides the original baseband signal.
- V3 produces a 100 MHz sine wave for the carrier.
- The op-amp circuit is a level shifter (it also reduces the input amplitude by half). The signal coming from V1 is a sine wave that swings from –1 V to +1 V, and the output of the op-amp is a sine wave that swings from 0 V to +1 V.
- B1 is an "arbitrary behavioral voltage source." Its "value" field is a formula rather than a constant; in this case the formula is the shifted baseband signal multiplied by the carrier waveform. In this way B1 can be used to perform amplitude modulation.

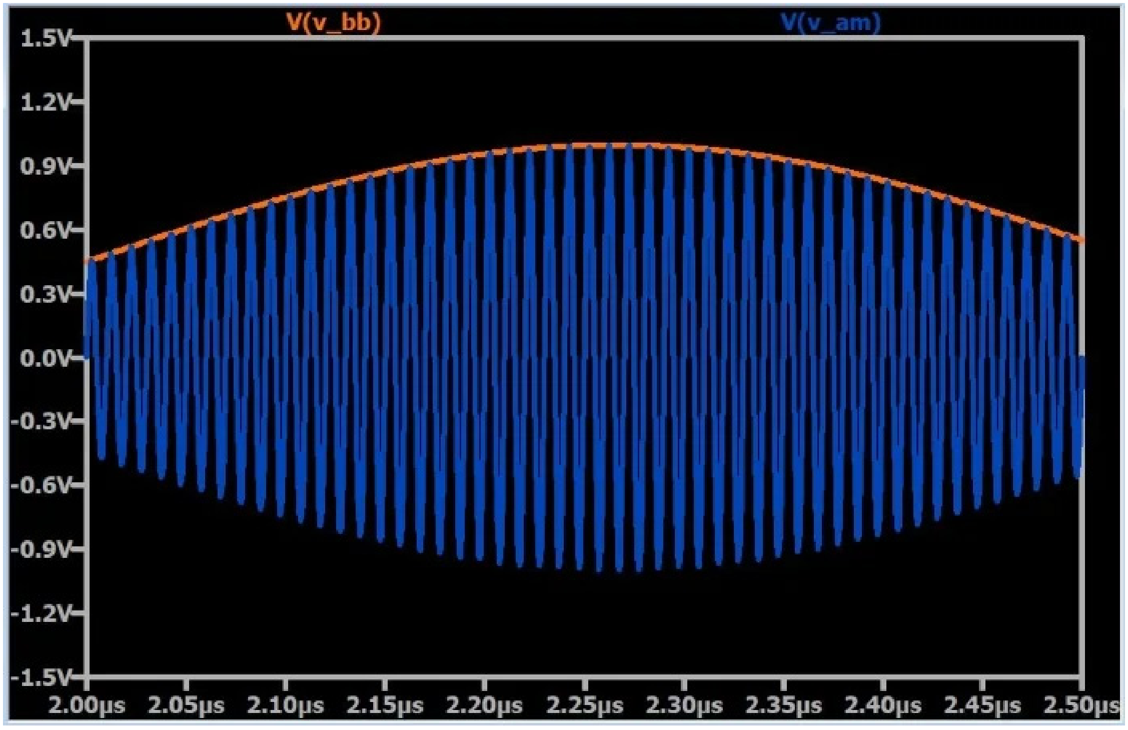
Here is the shifted baseband signal:



And here you can see how the AM variations correspond to the baseband signal (i.e., the orange trace that is mostly obscured by the blue waveform):



Zooming in reveals the individual cycles of the 100 MHz carrier frequency.

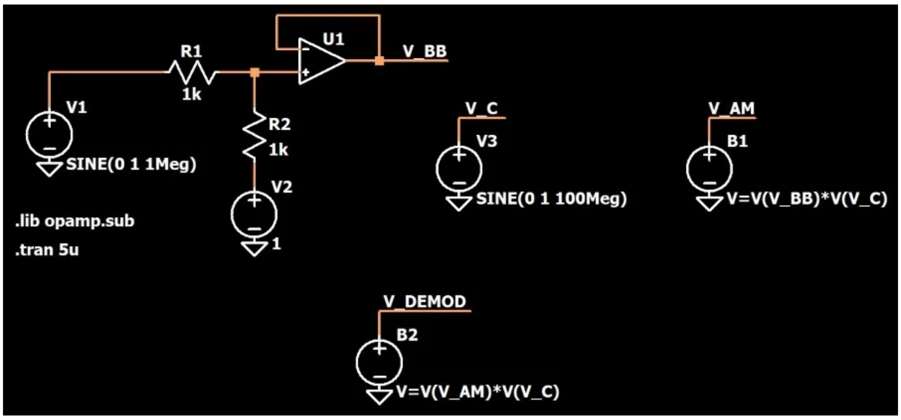


Demodulation

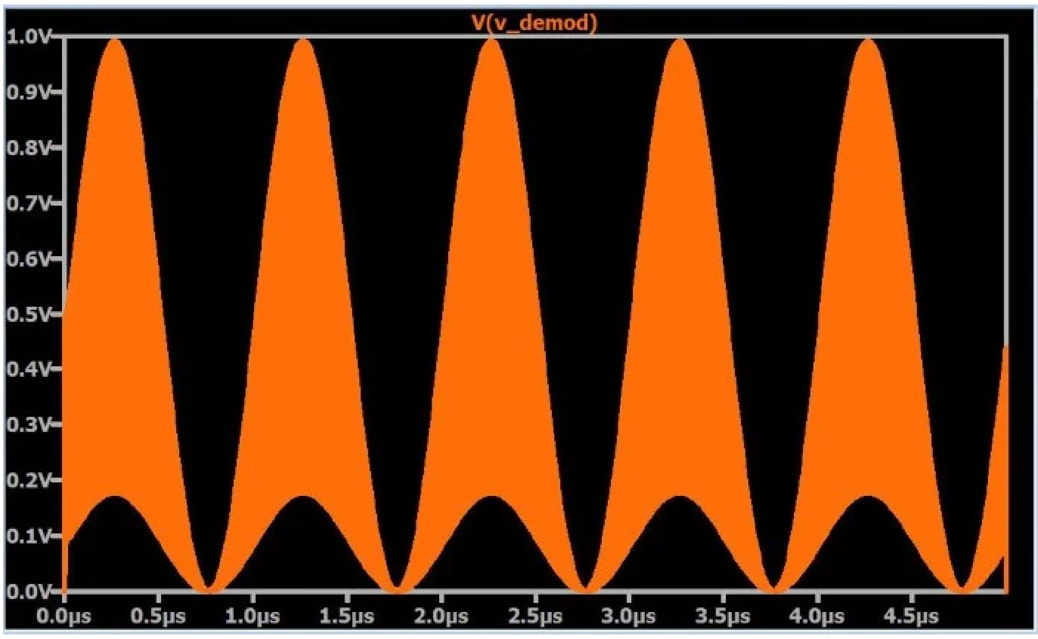
As discussed in the AM modulation page, the multiplication operation used to perform amplitude modulation has the effect of transferring the baseband spectrum to a band surrounding the positive carrier frequency (+f_C) and the negative carrier frequency (–f_C). Thus, we can think of amplitude modulation as shifting the original spectrum upward by f_C and downward by f_C. It follows, then, that multiplying the modulated signal by the carrier frequency will transfer the spectrum back to its original position—i.e., it will shift the spectrum downward by f_C such that it is once again centered around 0 Hz.

Option 1: Multiplication and Filtering

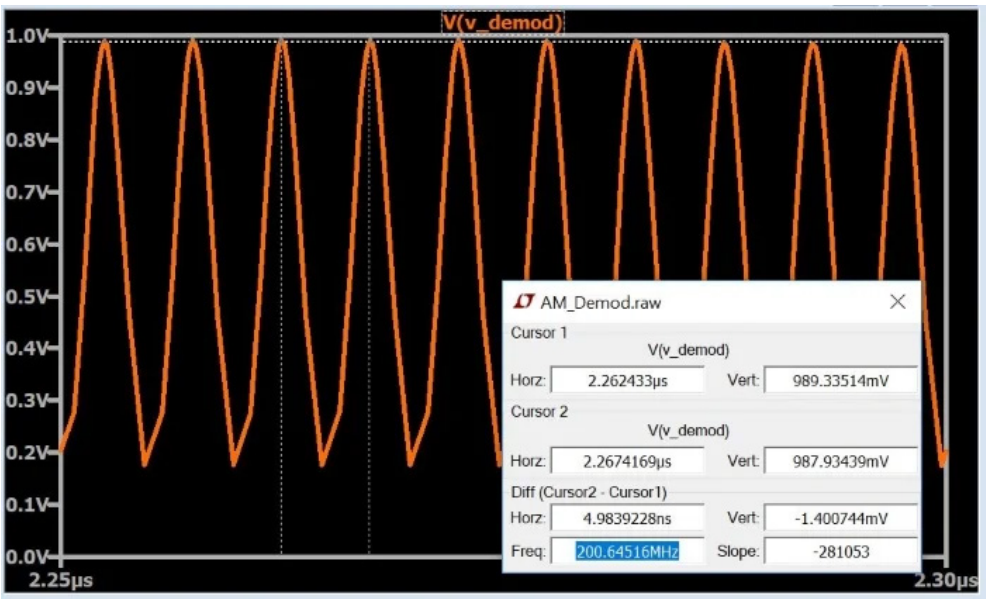
The following LTspice schematic includes a demodulating arbitrary behavioral voltage source; B2 multiplies the AM signal by the carrier.



And here is the result:



This definitely does not look correct. If we zoom in, we see the following:



AM_Demod.raw

Cursor 1
V(v_demod)
Horz: 2.262433µs Vert: 989.33514mV

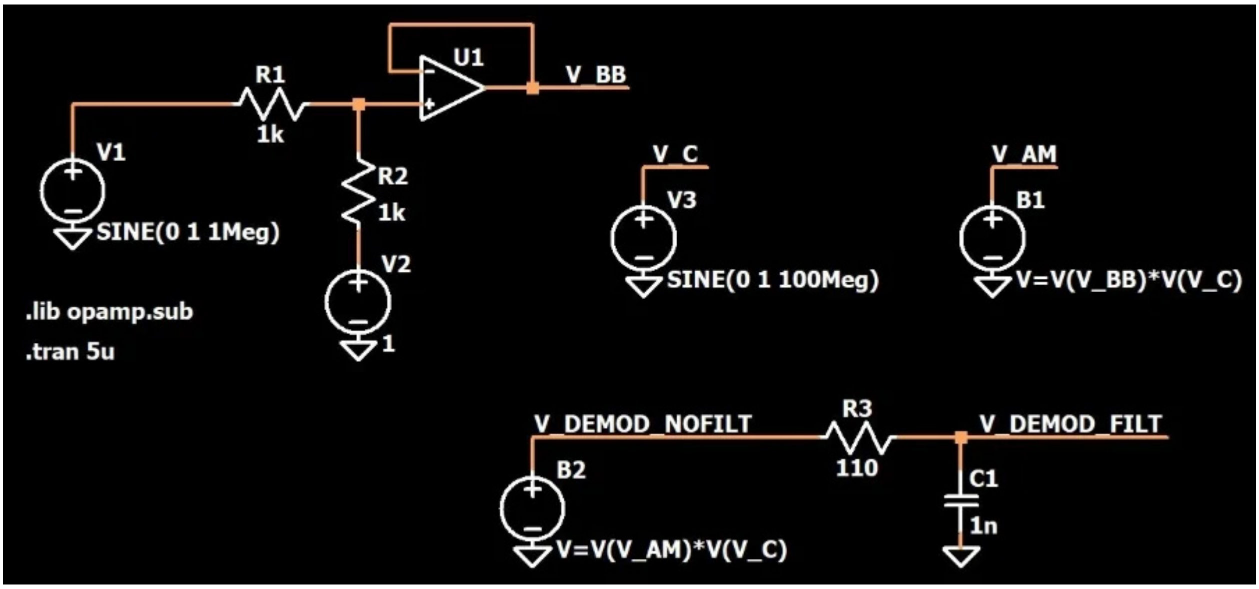
Cursor 2
V(v_demod)
Horz: 2.2674169µs Vert: 987.93439mV

Diff (Cursor2 - Cursor1)
Horz: 4.9839228ns Vert: -1.400744mV
Freq: 200.64516MHz Slope: -281053

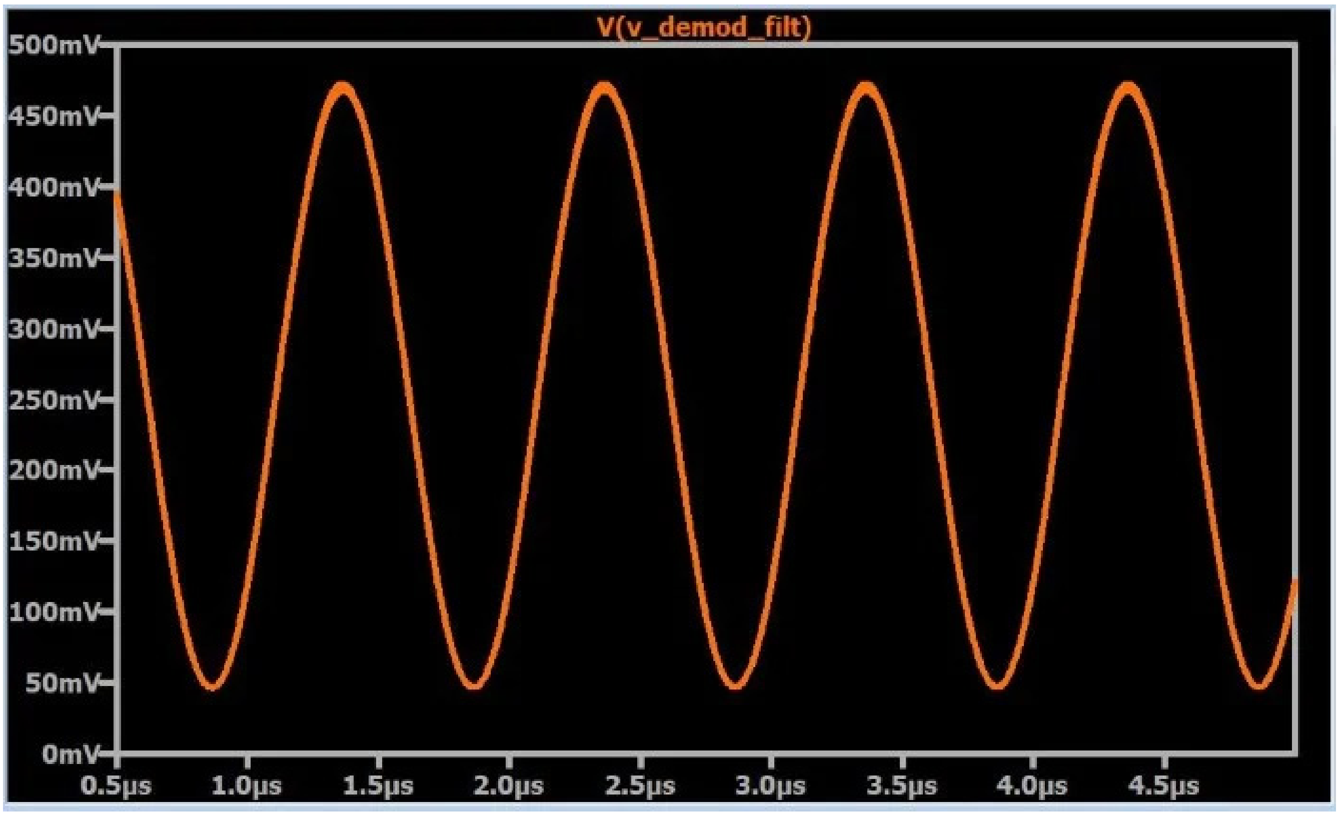
And this reveals the problem. After amplitude modulation, the baseband spectrum is centered around +f_C. Multiplying the AM waveform by the carrier shifts the baseband spectrum down to 0 Hz, **but it also shifts it up to 2f_C** (in this case 200 MHz), because (as stated above) multiplication moves the existing spectrum up by fC and down by f_C.

It is clear, then, that multiplication alone is not sufficient for proper demodulation. What we need is multiplication and a low-pass filter; the filter suppresses the spectrum that was shifted up to 2f_C.

The following schematic includes an RC low-pass filter with a cutoff .frequency of ~1.5 MHz.



And here is the demodulated signal:



This technique is actually more complicated than it appears because the phase of the receiver's carrier-frequency waveform must be synchronized with the phase of the transmitter's carrier. This is discussed further in page 5 of this chapter (Understanding Quadrature Demodulation).

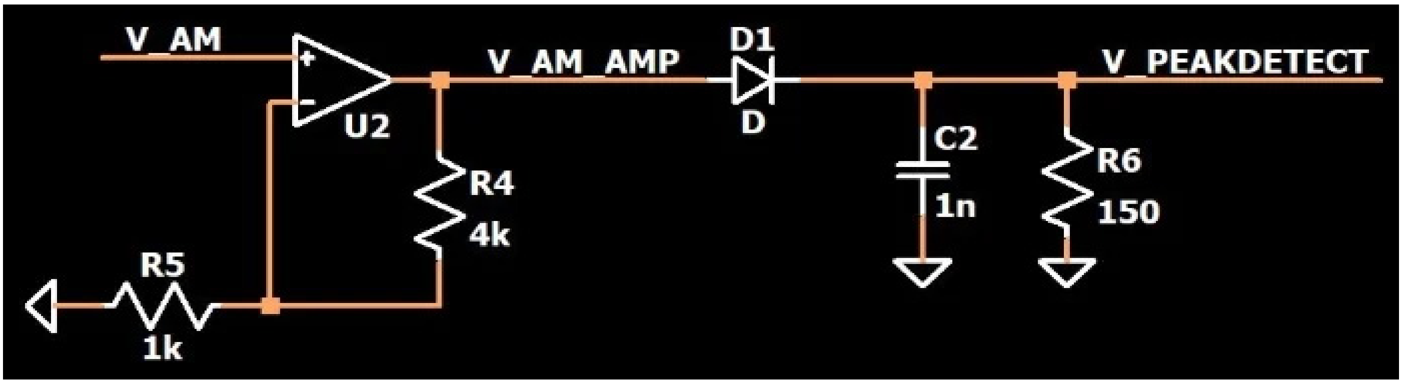
Option 2: Peak Detector

As you can see above in the plot that shows the AM waveform (in blue) and the shifted baseband waveform (in orange), the positive portion of the AM "envelope" matches the baseband signal. The term "envelope" refers to the carrier's variations in sinusoidal amplitude (as opposed to the variations in the instantaneous value of the waveform itself). If we could somehow extract the positive portion of the AM envelope, we could reproduce the baseband signal without using a multiplier.

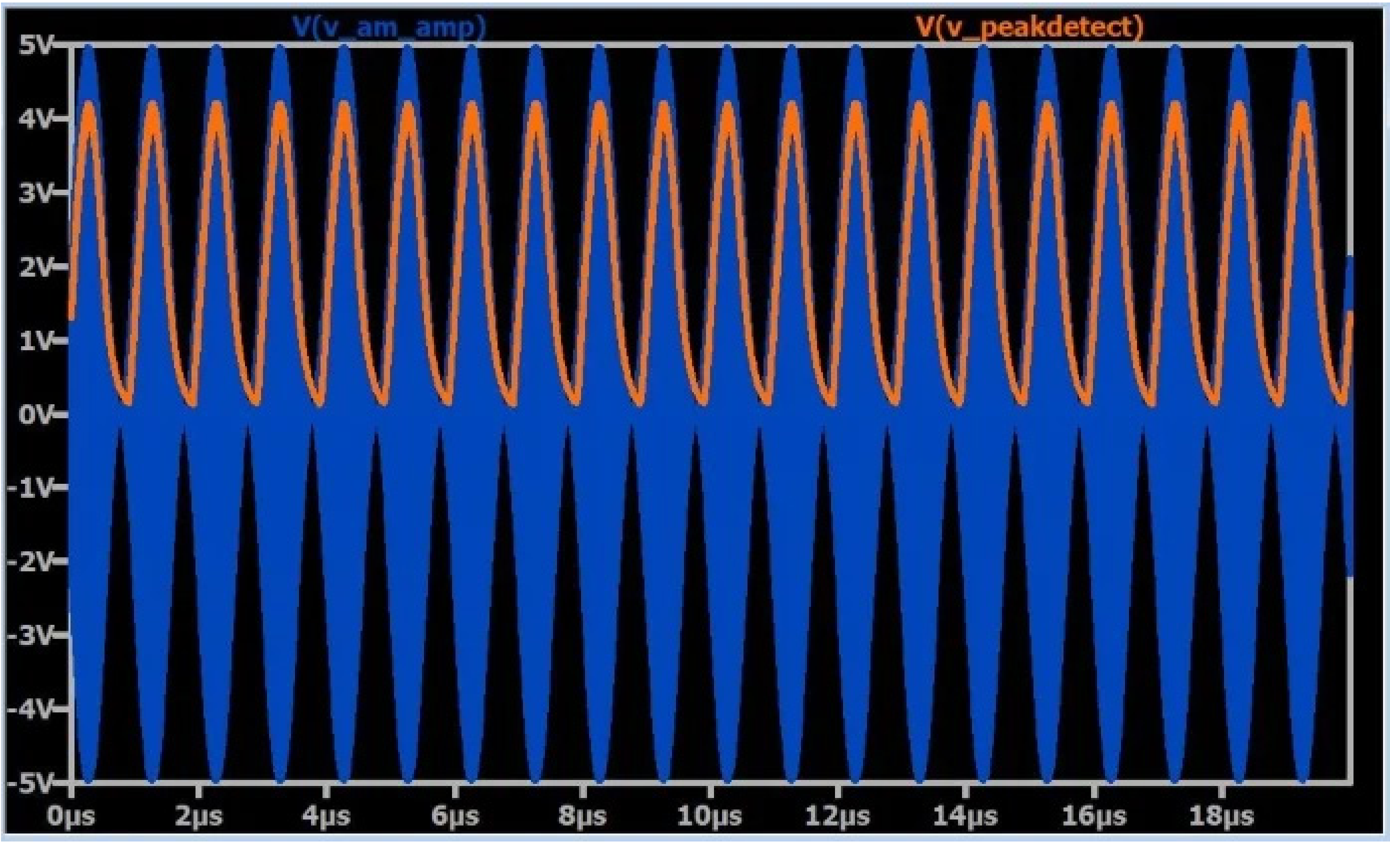
It turns out that it is quite easy to convert the positive envelope into a normal signal. We start with a peak detector, which is just a diode followed by a capacitor. The diode conducts when the input signal is at least ~0.7 V above the voltage on the capacitor, and otherwise it acts like an open circuit. Thus, the capacitor maintains the peak voltage: if the current input voltage is lower than the capacitor voltage, the capacitor voltage doesn't decrease because the reverse-biased diode prevents discharge.

However, we don't want a peak detector that will retain the peak voltage for a long period of time. Instead, we want a circuit that retains the peak relative to the high-frequency variations of the carrier waveform, but does not retain the peak relative to the lower-frequency variations of the envelope. In other words, we want a peak detector that holds the peak only for a short period of time. We accomplish this by adding parallel resistance that allows the capacitor to discharge. (This type of circuit is called a "leaky peak detector," where "leaky" refers to the discharge path provided by the resistor.) The resistance is chosen such that the discharge is **slow enough** to smooth out the carrier frequency and **fast enough** to *not* smooth out the envelope frequency.

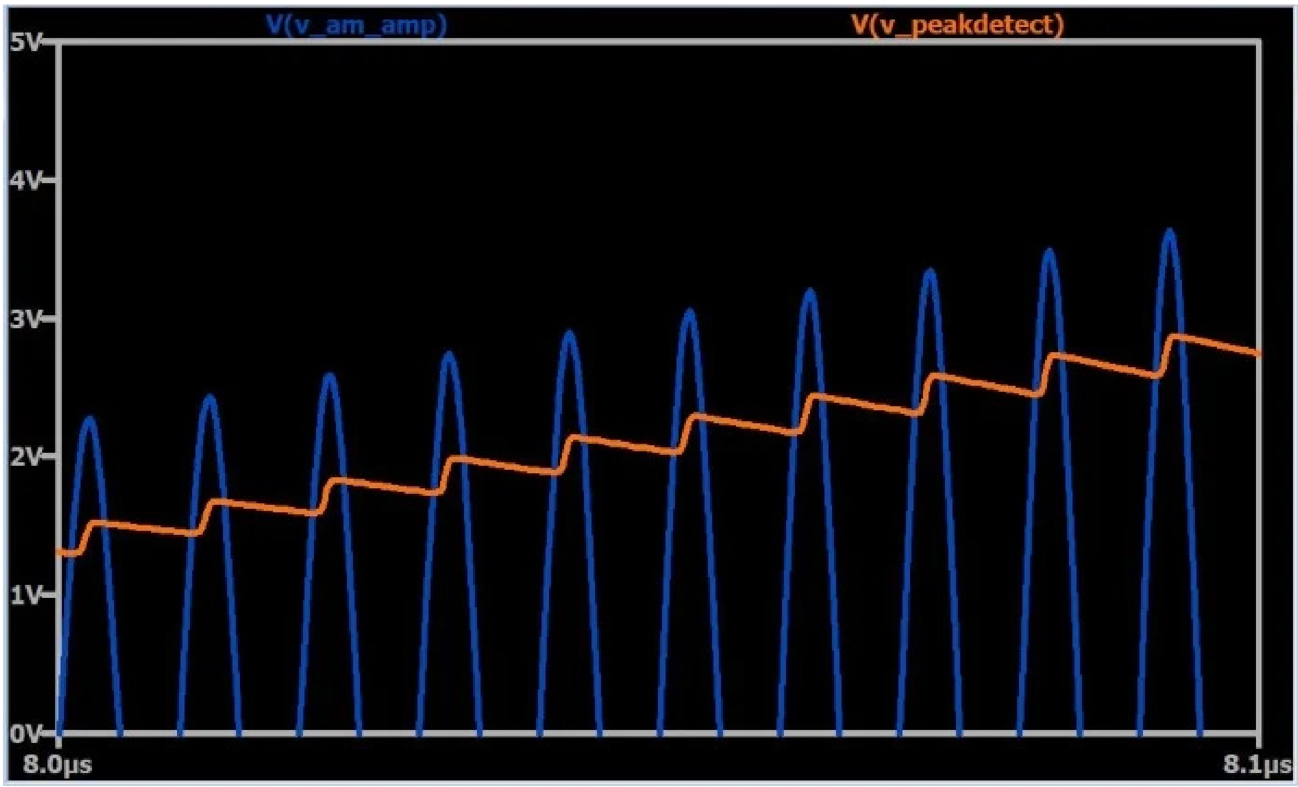
Here is an example of a leaky peak detector for AM demodulation:



Note that I have amplified the AM signal by a factor of five in order to make the peak detector's input signal larger relative to the diode's forward voltage. The following plot conveys the general result that we are trying to achieve with the leaky peak detector.



The final signal exhibits the expected charge/discharge characteristic:



A low-pass filter could be used to smooth out these variations.

Summary

- In LTspice, an arbitrary behavioral voltage source can be used to create an AM waveform.
- AM waveforms can be demodulated using a multiplier followed by a low-pass filter.
- A simpler (and lower-cost) approach is to use a leaky peak detector, i.e., a peak detector with parallel resistance that allows the capacitor to discharge at an appropriate rate.

How to Demodulate an FM Waveform

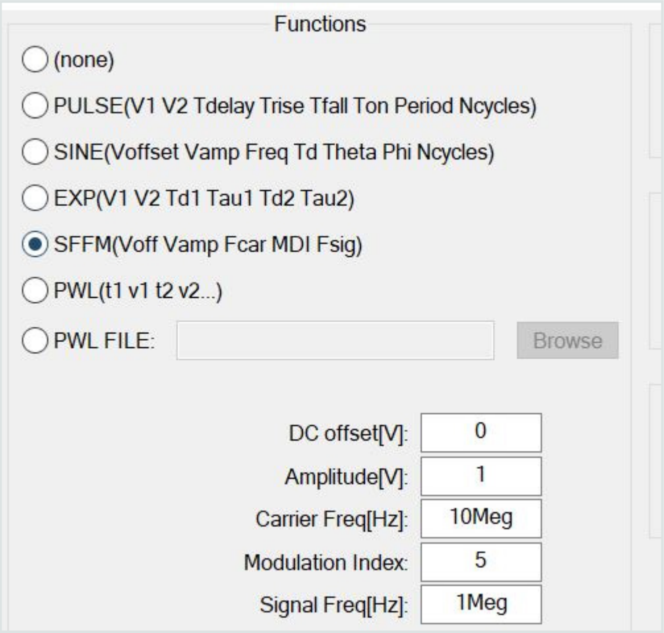
Learn about two techniques for recovering the baseband signal from a frequency-modulated carrier.

Frequency modulation offers improved performance over amplitude modulation, but it is somewhat more difficult to extract the original information from an FM waveform. There are a few different ways to demodulate FM; in this page we'll discuss two. One of these is quite straightforward, and the other is more complex.

Creating the Signal

As in How to Demodulate an AM Waveform, we'll use LTspice to explore FM demodulation, and once again we need to first perform frequency modulation so that we have something to demodulate. If you look back at the page on analog frequency modulation, you will see that the mathematical relationship is less straightforward than that of amplitude modulation. With AM, we simply added an offset and then performed ordinary multiplication. With FM, we need to add continuously varying values to the quantity inside a sine (or cosine) function, and furthermore, these continuously varying values are not the baseband signal but rather the integral of the baseband signal.

Consequently, we can't generate an FM waveform using an arbitrary behavioral voltage source and a simple mathematical relationship, as we did with AM. It turns out, though, that it is actually easier to generate an FM signal. We simply use the SFFM option for a normal voltage source:

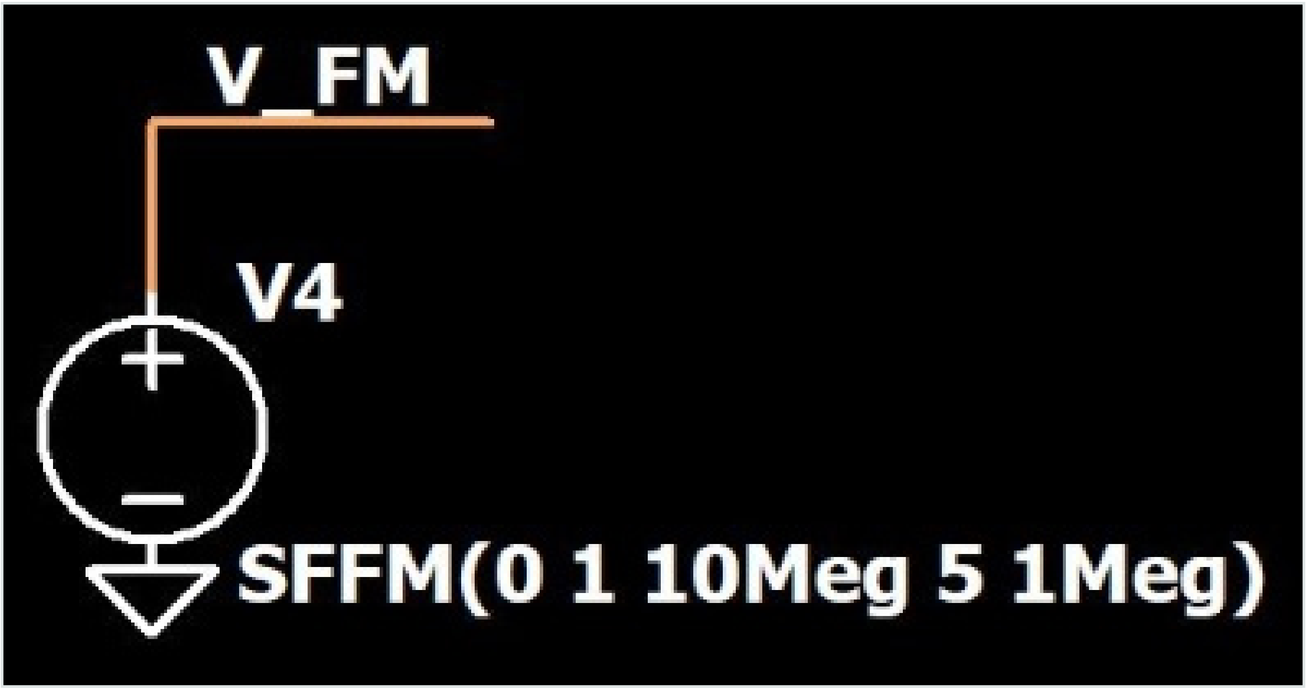


Functions

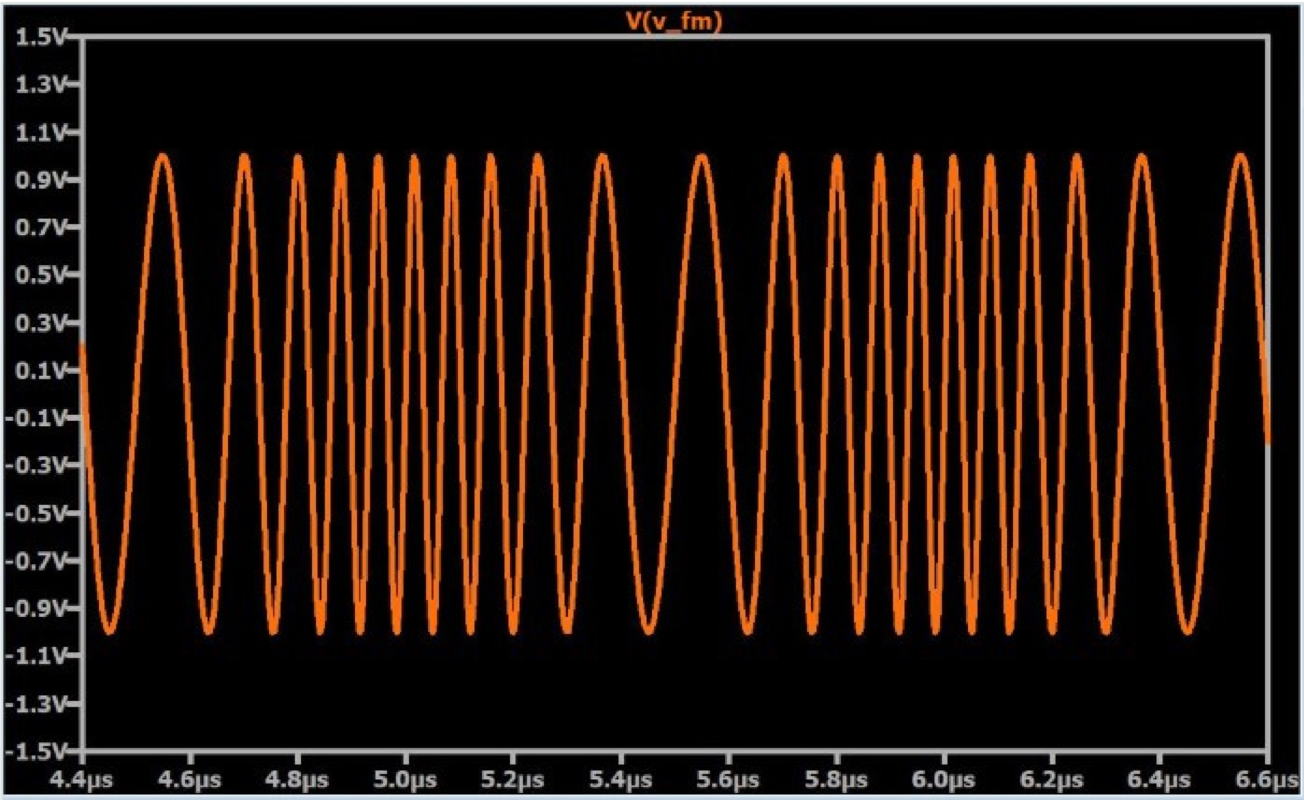
○ (none)
○ PULSE(V1 V2 Tdelay Trise Tfall Ton Period Ncycles)
○ SINE(Voffset Vamp Freq Td Theta Phi Ncycles)
○ EXP(V1 V2 Td1 Tau1 Td2 Tau2)
● SFFM(Voff Vamp Fcar MDI Fsig)
○ PWL(t1 v1 t2 v2...)
○ PWL FILE:

DC offset[V]:	0
Amplitude[V]:	1
Carrier Freq[Hz]:	10Meg
Modulation Index:	5
Signal Freq[Hz]:	1Meg

The following “circuit” is all we need for creating an FM waveform consisting of a 10 MHz carrier and a 1 MHz sinusoidal baseband signal:



Note that the modulation index is five; a higher modulation index makes it easier to see the frequency variations. The following plot shows the waveform created by the SFFM voltage source.

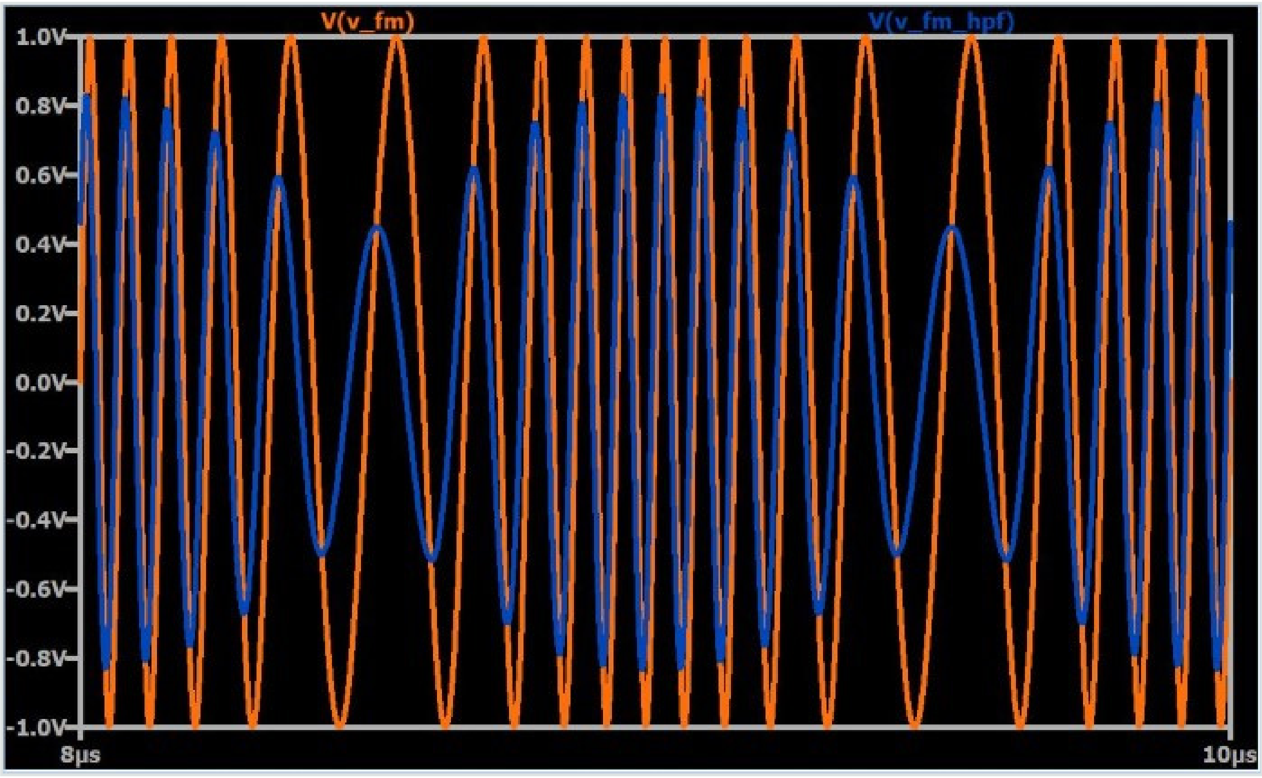


Demodulation: The High-Pass Filter

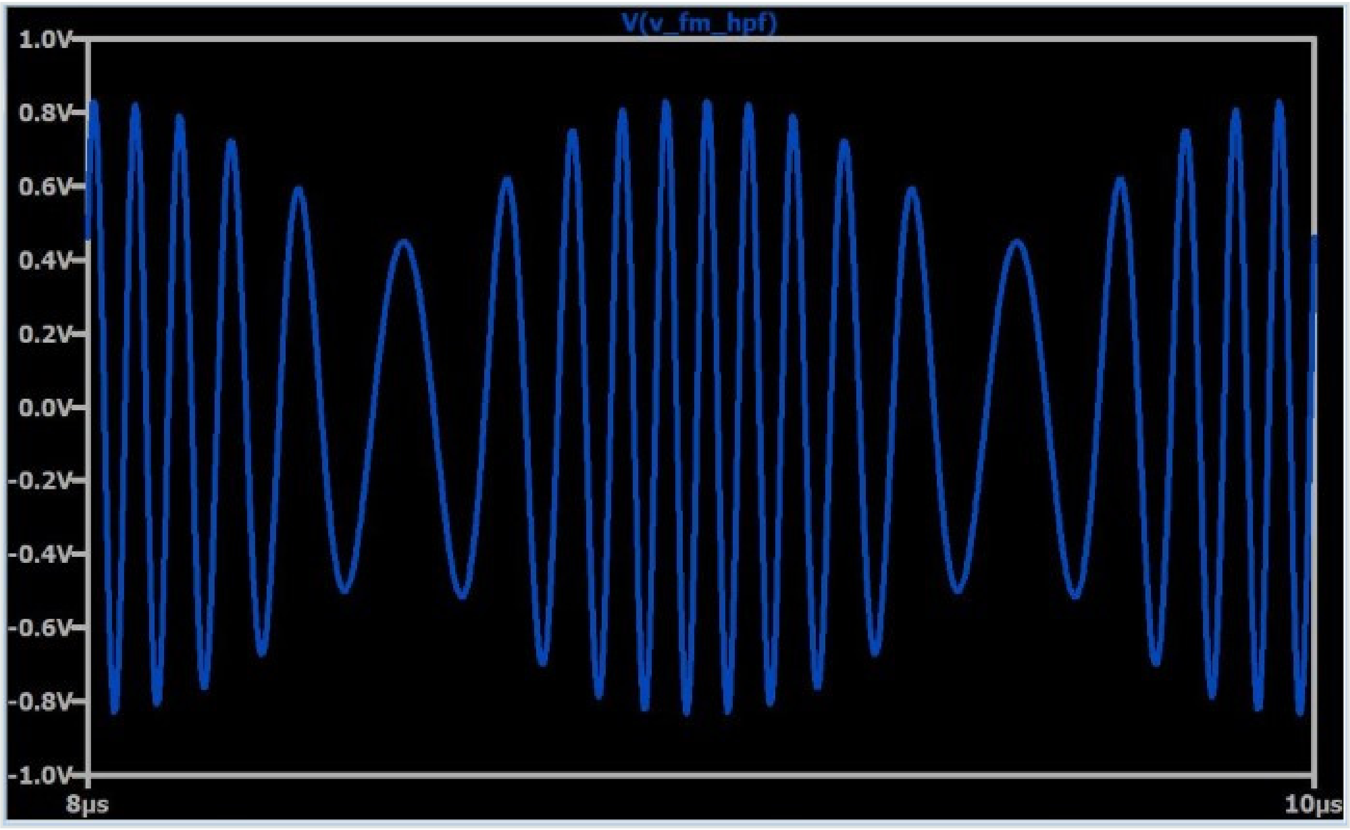
The first demodulation technique that we'll look at begins with a high-pass filter. We'll assume that we're dealing with narrowband FM (which is discussed briefly in this page). We need to design the high-pass filter such that the attenuation will vary significantly within a frequency band whose width is twice the bandwidth of the baseband signal. Let's explore this concept more thoroughly.

The received FM signal will have a spectrum that is centered around the carrier frequency. The width of the spectrum is approximately equal to twice the bandwidth of the baseband signal; the factor of two results from the shifting of the positive and negative baseband frequencies (as discussed here), and it is "approximately" equal because the integration applied to the baseband signal can affect the shape of the modulated spectrum. Thus, the lowest frequency in the modulated signal is approximately equal to the carrier frequency minus the highest frequency in the baseband signal, and the highest frequency in the modulated signal is approximately equal to the carrier frequency plus the highest frequency in the baseband signal.

Our high-pass filter needs to have a frequency response that causes the lowest frequency in the modulated signal to be attenuated significantly more than the highest frequency in the modulated signal. If we apply this filter to an FM waveform, what will be the result? It will be something like this:



This plot shows both the original FM waveform and the high-pass-filtered waveform, for purposes of comparison. The next plot shows just the filtered waveform, so that you can see it more clearly.



By applying the filter, we have turned frequency modulation into amplitude modulation. This is a convenient approach to FM demodulation, because it allows us to benefit from envelope-detector circuitry that has been developed for use with amplitude modulation. The filter used to produce this waveform was nothing more than an RC high-pass with a cutoff frequency approximately equal to the carrier frequency.

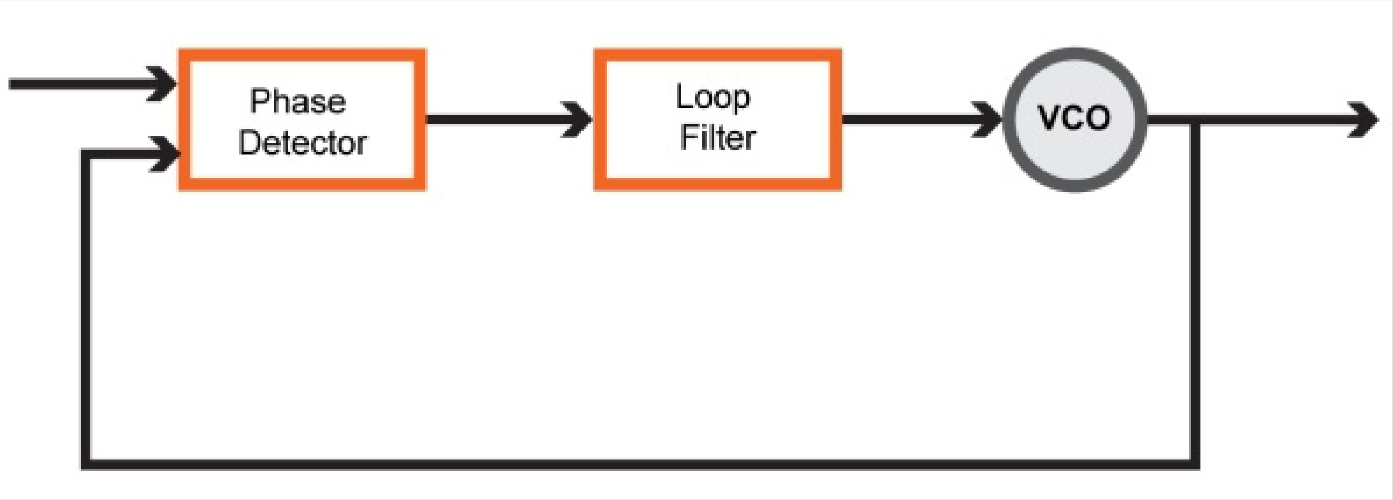
Amplitude Noise

The simplicity of this demodulation scheme naturally makes us think that it is not the highest-performance option, and in fact this approach does have a major weakness: it is sensitive to amplitude variations. The transmitted signal will have a constant envelope because frequency modulation does not involve changes to the amplitude of the carrier, but the received signal will *not* have a constant envelope because amplitude is inevitably affected by error sources.

Consequently, we cannot design an acceptable FM demodulator simply by adding a high-pass filter to an AM demodulator. We also need a limiter, which is a circuit that mitigates amplitude variations by restricting the received signal to a certain amplitude. The existence of this simple and effective remedy for amplitude variations enables FM to maintain its greater (compared to AM) robustness against amplitude noise: We cannot use a limiter with AM signals because restricting the amplitude corrupts the information encoded in the carrier. FM, on the other hand, encodes all of the information in the temporal characteristics of the transmitted signal.

Demodulation: The Phase-Locked Loop

A phase-locked loop (PLL) can be used to create a complex but high-performance circuit for FM demodulation. A PLL can “lock onto” the frequency of an incoming waveform. It does this by combining a phase detector, a low-pass filter (aka “loop filter”), and a voltage-controlled oscillator (VCO) into a negative-feedback system, as follows:



After the PLL has locked, it can create an output sinusoid that follows frequency variations in the incoming sinusoid. This output waveform would be taken from the output of the VCO. In an FM-demodulator application, however, we don’t need an output sinusoid that has the same frequency as the input signal. Instead, we use the output from the loop filter as a demodulated signal. Let’s look at why this is possible.

The phase detector produces a signal that is proportional to the phase difference between the incoming waveform and the output of the VCO. The loop filter smooths this signal, which then becomes the control signal for the VCO. Thus, if the frequency of the incoming signal is constantly increasing and decreasing, the VCO control signal has to increase and decrease accordingly to ensure that the VCO output frequency remains equal to the input frequency. In other words, the output of the loop filter is a signal whose amplitude variations correspond to the input-frequency variations. This is how a PLL accomplishes frequency demodulation.

Summary

- In LTspice, a frequency-modulated sinusoid can be generated by using the SFFM option for standard voltage sources.
- A simple and effective FM demodulation technique involves a high-pass filter (for FM-to-AM conversion) followed by an AM demodulator.
- A high-pass-filter-based FM demodulator is preceded by a limiter to prevent amplitude variations from contributing error to the demodulated signal.
- A phase-locked loop can be used to achieve high-performance FM demodulation. The use of integrated-circuit PLLs makes this approach less complex than it might seem.

How to Demodulate Digital Phase Modulation

Learn about how to extract the original digital data from a phase-shift-keying waveform.

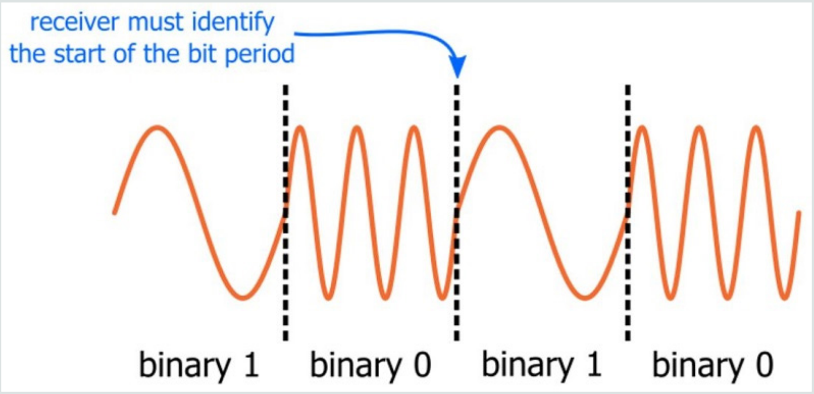
In the previous two pages we discussed systems for performing demodulation of AM and FM signals that carry analog data, such as (non-digitized) audio. Now we are ready to look at how to recover original information that has been encoded via the third general type of modulation, namely, phase modulation.

However, analog phase modulation is not common, whereas digital phase modulation is very common. Thus, it makes more sense to explore PM demodulation in the context of digital RF communication. We'll explore this topic using binary phase shift keying (BPSK); however, it's good to be aware that quadrature phase shift keying (QPSK) is more relevant to modern wireless systems.

As the name implies, binary phase shift keying represents digital data by assigning one phase to binary 0 and a different phase to binary 1. The two phases are separated by 180° to optimize demodulation accuracy—more separation between the two phase values makes it easier to decode the symbols.

Multiply and Integrate—and Synchronize

A BPSK demodulator consists primarily of two functional blocks : a multiplier and an integrator. These two components will produce a signal that corresponds to the original binary data. However, synchronization circuitry is also needed, because the receiver must be able to identify the boundary between bit periods. This is an important difference between analog demodulation and digital demodulation, so let's take a closer look.



This diagram shows a frequency-shift-keying signal, but the same concept applies to digital phase modulation and digital amplitude modulation.

In analog demodulation, the signal doesn't really have a beginning or an end. Imagine an FM transmitter that is broadcasting an audio signal, i.e., a signal that continuously varies according to the music. Now imagine an FM receiver that is initially turned off. The user can power up the receiver at any moment in time, and the demodulation circuitry will begin extracting the audio signal from the modulated carrier. The extracted signal can be amplified and sent to a speaker, and the music will sound normal. The receiver has no idea if the audio signal represents the beginning or end of a song, or if the demodulation circuitry starts functioning at the beginning of a measure, or right on the beat, or in between two beats. It doesn't matter; each instantaneous voltage value corresponds to one exact moment in the audio signal, and the sound is re-created when all of these instantaneous values occur in succession.

With digital modulation, the situation is completely different. We're not dealing with instantaneous amplitudes but rather a *sequence of amplitudes* that represents one discrete piece of information, namely, a number (one or zero). Each sequence of amplitudes—called a symbol, with a duration equal to one bit period—must be distinguished from the preceding and following sequences: If the broadcaster (from the above example) were using digital modulation and the receiver powered up and started demodulating at a random point in time, what would happen? Well, if the receiver happened to start demodulating in the middle of a symbol, it would be trying to interpret half of one symbol and half of the following symbol. This would, of course, lead to errors; a logic-one symbol followed by a logic-zero symbol would have an equal chance of being interpreted as a one or a zero.

Clearly, then, synchronization must be a priority in any digital RF system. One straightforward approach to synchronization is to precede each packet with a predefined "training sequence" consisting of alternating zero symbols and one symbols (as in the above diagram). The receiver can use these one-zero-one-zero transitions to identify the temporal boundary between symbols, and then the rest of the symbols in the packet can be interpreted properly simply by applying the system's predefined symbol duration.

The Effect of Multiplication

As mentioned above, a fundamental step in PSK demodulation is multiplication. More specifically, we multiply an incoming BPSK signal by a reference signal with frequency equal to the carrier frequency. What does this accomplish? Let's look at the math; first, the product identify for two sine functions:

$$\sin(x) * \sin(y) = \frac{1}{2}(\cos(x - y) - \cos(x + y))$$

If we turn these generic sine functions into signals with a frequency and phase, we have the following:

$$\sin(\omega_C + \theta_1) * \sin(\omega_C + \theta_2) = \frac{1}{2}[\cos(\omega_C + \theta_1 - (\omega_C + \theta_2)) - \cos(\omega_C + \theta_1 + \omega_C + \theta_2)]$$

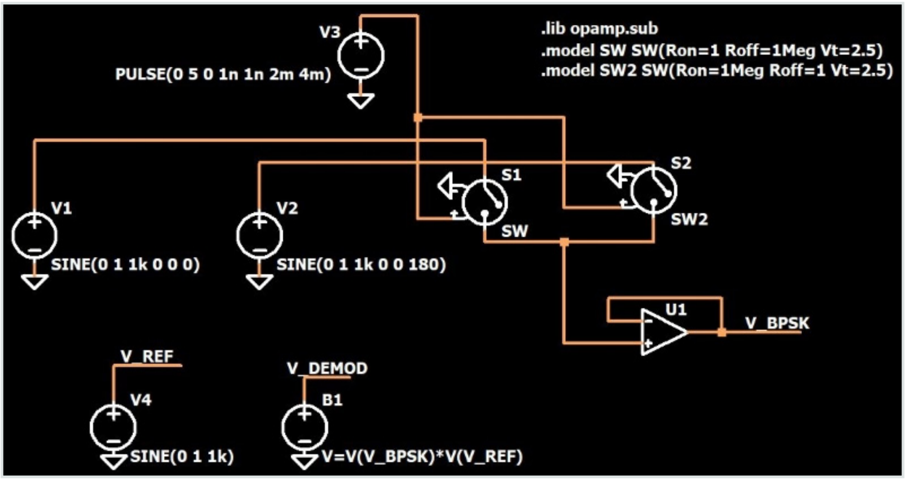
Simplifying, we have:

$$\sin(\omega_C + \theta_1) * \sin(\omega_C + \theta_2) = \frac{1}{2}[\cos(\theta_1 - \theta_2) - \cos(2\omega_C + (\theta_1 + \theta_2))]$$

So when we multiply two sinusoids of equal frequency but different phase, the result is a sinusoid of double the frequency *plus an offset* that depends on the difference between the two phases. The offset is the key: If the phase of the received signal is equal to the phase of the reference signal, we have cos(0°), which equals 1. If the phase of the received signal is 180° different from the phase of the reference signal, we have cos(180°), which is –1. Thus, the output of the multiplier will have a positive DC offset for one of the binary values and a negative DC offset for the other binary value. This offset can be used to interpret each symbol as a zero or a one.

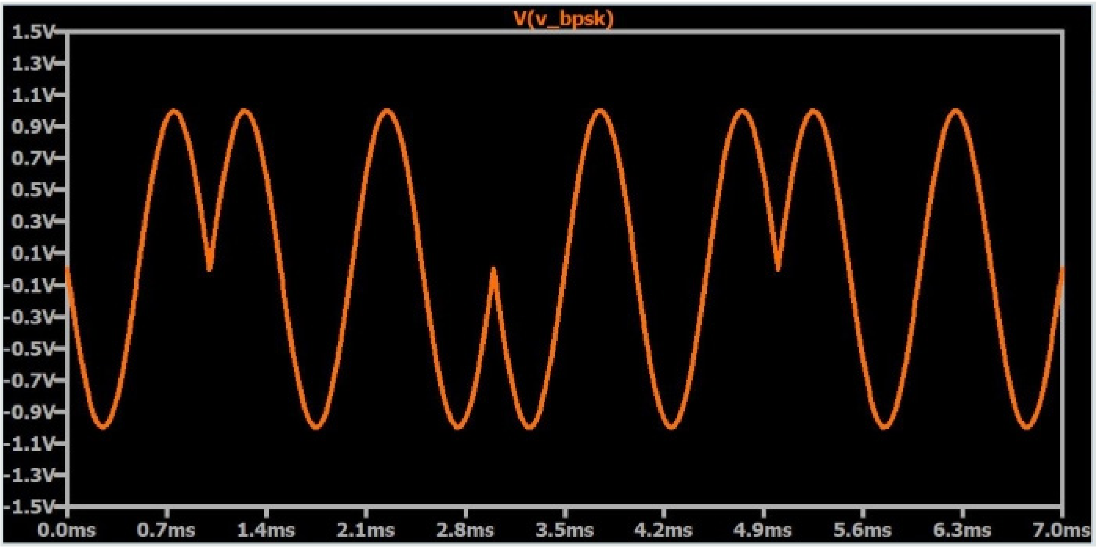
Simulation Confirmation

The following BPSK modulation-and-demodulation circuit shows you how you can create a BPSK signal in LTspice:

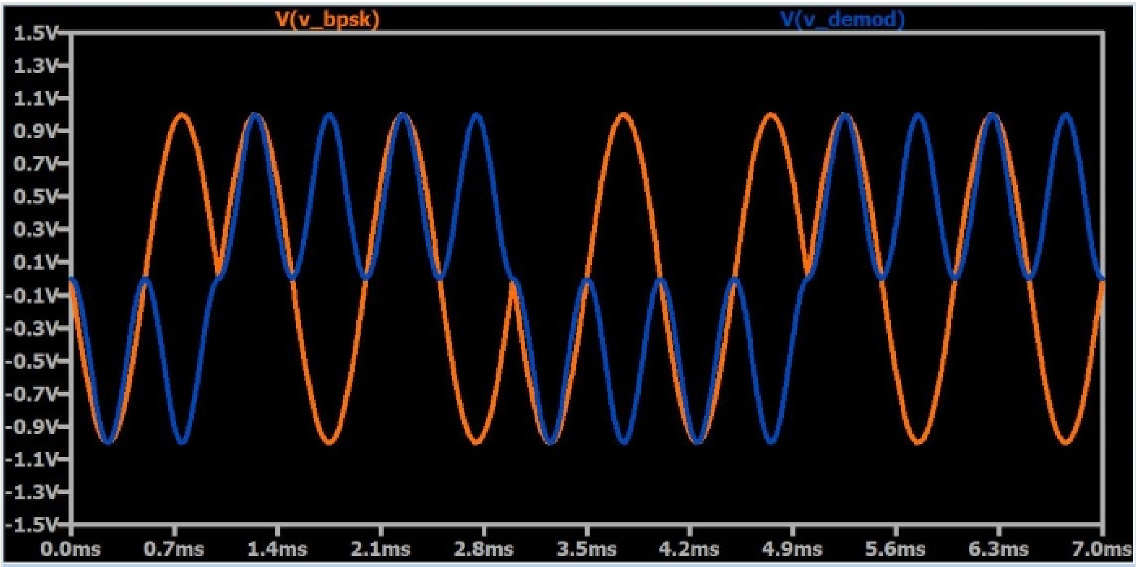


```
.lib opamp.sub
.model SW SW(Ron=1 Roff=1Meg Vt=2.5)
.model SW2 SW(Ron=1Meg Roff=1 Vt=2.5)
```

Two sine sources (one with phase = 0° and one with phase = 180°) are connected to two voltage-controlled switches. Both switches have the same square-wave control signal, and the on and off resistances are configured such that one is open while the other is closed. The “output” terminals of the two switches are tied together, and the op-amp buffers the resulting signal, which looks like this:



Next, we have a reference sinusoid (V4) with frequency equal to the frequency of the BPSK waveform, and then we use an arbitrary behavioral voltage source to multiply the BPSK signal by the reference signal. Here is the result:

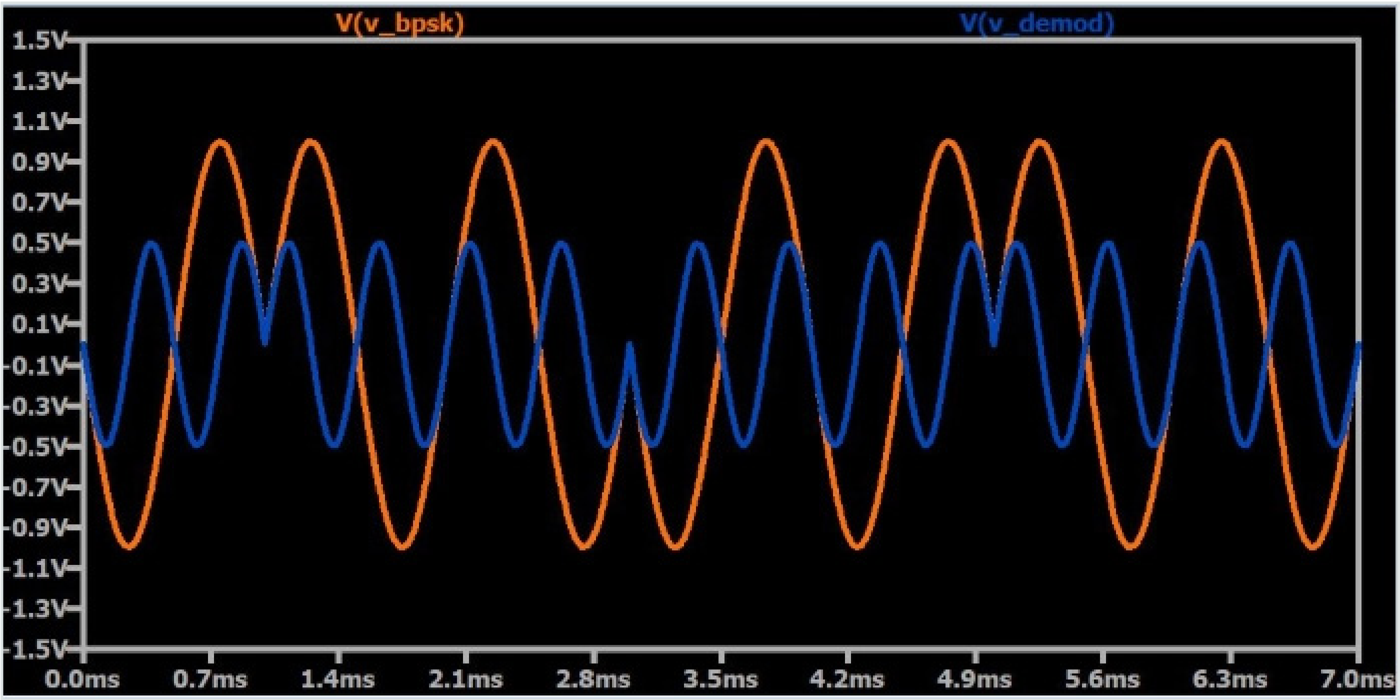


As you can see, the demodulated signal is double the frequency of the received signal, and it has a positive or negative DC offset according to the phase of each symbol. If we then integrate this signal with respect to each bit period, we will have a digital signal that corresponds to the original data.

Coherent Detection

In this example, the phase of the receiver's reference signal is synchronized with the phase of the incoming modulated signal. This is easily accomplished in a simulation; it's significantly more difficult in real life. Furthermore, as discussed in this page under "Differential Encoding," ordinary phase shift keying cannot be used in systems that are subject to unpredictable phase differences between transmitter and receiver. For example, if the receiver's reference signal is 90° out of phase with the transmitter's carrier, the phase difference between the reference and the BPSK signal will always be 90°, and cos(90°) is 0. Thus, the DC offset is lost, and the system is completely nonfunctional.

This can be confirmed by changing the phase of the V4 source to 90°; here is the result:



Summary

- Digital demodulation requires bit-period synchronization; the receiver must be able to identify the boundaries between adjacent symbols.
- Binary-phase-shift-keying signals can be demodulated via multiplication followed by integration. The reference signal used in the multiplication step has the same frequency as the transmitter's carrier.
- Ordinary phase-shift-keying is reliable only when the phase of the receiver's reference signal can maintain synchronization with the phase of the transmitter's carrier.

Understanding I/Q Signals and Quadrature Modulation

Learn about “I/Q” signals, how they are used, and why they are advantageous in RF systems.

This chapter would not be complete without a page on quadrature demodulation. However, before we explore quadrature demodulation, we need to at least briefly discuss quadrature modulation. And before we discuss quadrature modulation, we need to understand I/Q signals.

In-Phase and Quadrature

The term “I/Q” is an abbreviation for “in-phase” and “quadrature.” Unfortunately, we already have a terminology problem. First of all, “in-phase” and “quadrature” have no meaning on their own; phase is relative, and something can only be “in phase” or “out of phase” with reference to another signal or an established reference point. Furthermore, we now have the word “quadrature” applied to both a signal and the modulation/demodulation techniques associated with that signal.

In any event, “in-phase” and “quadrature” refer to two sinusoids that have the same frequency and are 90° out of phase. By convention, the I signal is a cosine waveform, and the Q signal is a sine waveform. As you know, a sine wave (without any additional phase) is shifted by 90° relative to a cosine wave. Another way to express this is that the sine and cosine waves are in *quadrature*.

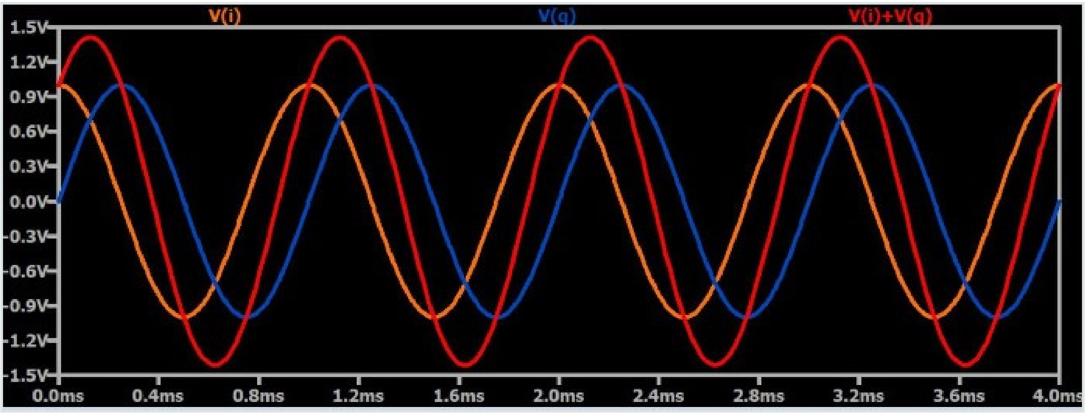
The first thing to understand about I/Q signals is that they are always amplitude-modulated, not frequency- or phase-modulated. However, I/Q amplitude modulation is different from the AM technique discussed in Chapter 4: in an I/Q modulator, the signals that modulate the I/Q sinusoids are not shifted such that they are always positive. In other words, I/Q modulation involves multiplying I/Q waveforms by modulating signals that can have negative voltage values, and consequently the “amplitude” modulation can result in a 180° phase shift. Later in this page we’ll explore this issue in more detail.

What is so advantageous about amplitude-modulating two sinusoids that are 90° out of phase? Why are I/Q modulation and demodulation so widespread? Read on.

Summing I and Q

I and Q signals on their own are not very interesting. The interesting thing happens when I and Q waveforms are added. It turns out that any form of modulation can be performed simply by varying the amplitude—only the amplitude—of I and Q signals, and then adding them together.

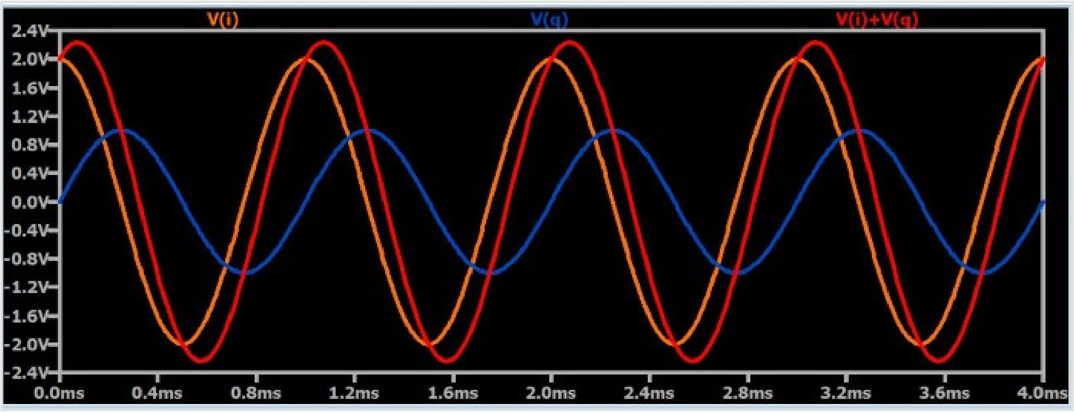
If you take I and Q signals of equal amplitude and add them, the result is a sinusoid with a phase that is exactly between the phase of the I signal and the phase of the Q signal.

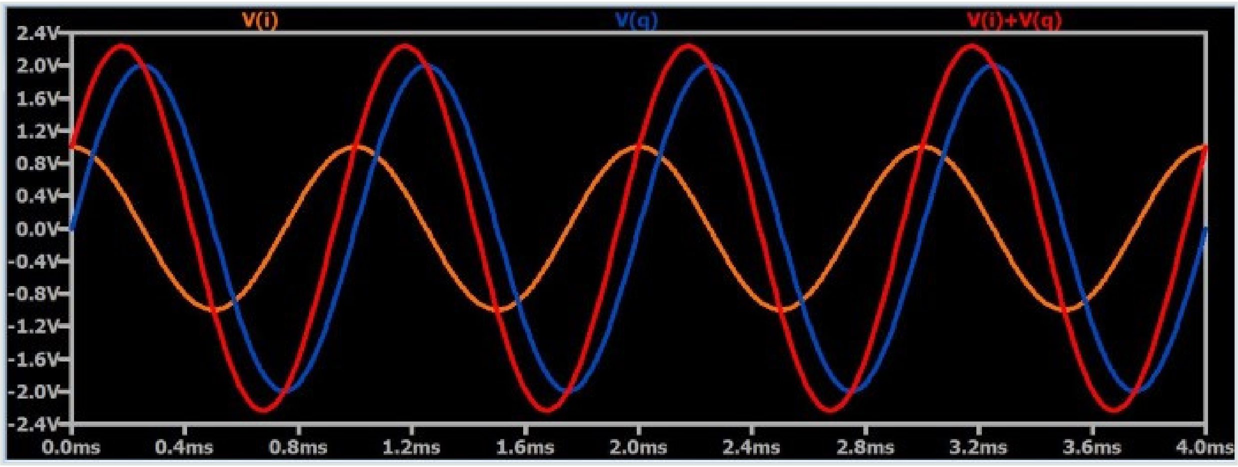


In other words, if you consider the I waveform to have a phase of 0° and the Q waveform to have a phase of 90°, the summation signal will have a phase of 45°. If you want to use these I and Q signals to create an amplitude-modulated waveform, you simply amplitude modulate the individual I and Q signals. Obviously a signal will increase or decrease in amplitude if it is created by adding together two signals that are both increasing or both decreasing in amplitude. However, you must ensure that the amplitude modulation applied to the I signal is identical to the amplitude modulation applied to the Q signal, because if they are not identical, you will have phase shift. And that brings us to the next property of I/Q signaling.

From Amplitude to Phase

Phase modulation, in the form of phase shift keying, is an important technique in modern RF systems, and phase modulation can be conveniently achieved by varying the *amplitude* of I/Q signals. Consider the following plots:



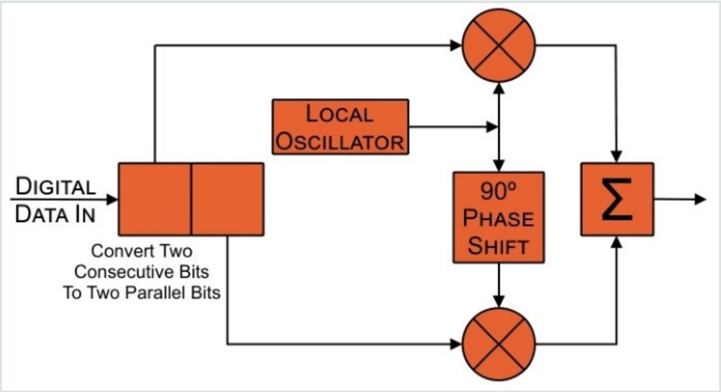


As you can see, increasing the amplitude of one of the waveforms relative to the other causes the summation signal to shift toward the higher-amplitude waveform. This makes intuitive sense: if you eliminated the Q waveform, for example, the summation would shift all the way over to the phase of the I waveform, because (obviously) adding the I waveform to zero will result in a summation signal that is identical to the I waveform.

It would seem from the above discussion that I/Q signaling can only be used to shift a signal 90° (i.e., 45° in each direction): if the Q amplitude is reduced to zero, the summation goes all the way to the I phase; if the I amplitude is reduced to zero, the summation goes all the way to the Q phase. How, then, could we use I/Q signals to create (for example) quadrature phase shift keying (QPSK), which uses phase values covering a range of 270°? We'll discuss this in the next section.

Quadrature Modulation

The term "quadrature modulation" refers to modulation that is based on the summation of two signals that are in quadrature. In other words, it is I/Q-signal-based modulation. We'll use QPSK as an example of how quadrature modulation works, and in the process we'll see how amplitude modulation of I/Q signals can produce phase shifts beyond 90°.



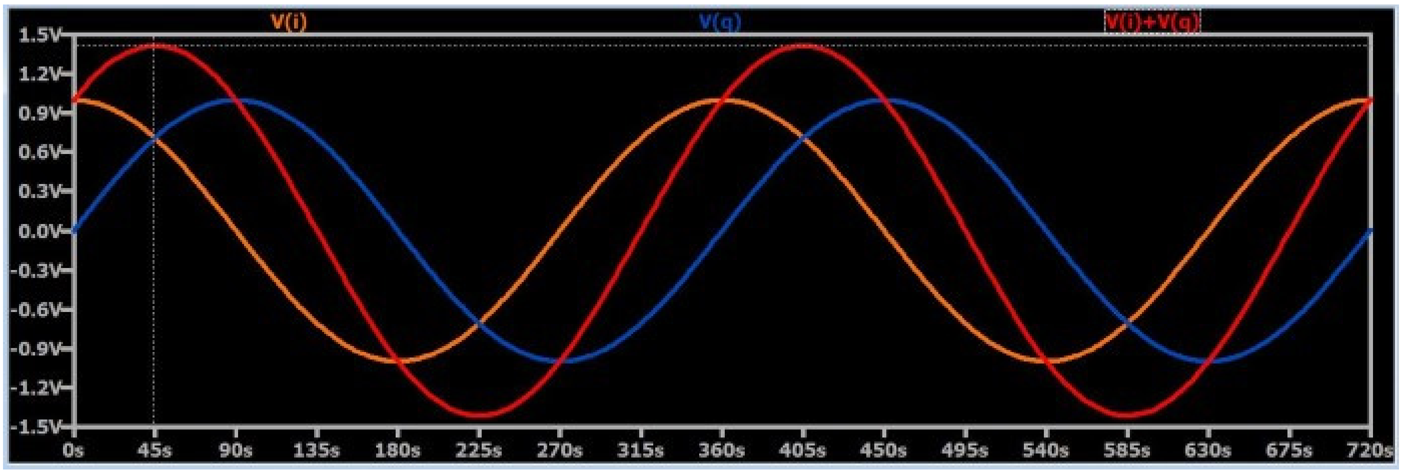
This is a basic block diagram for a QPSK modulator. First, the digital data stream is processed so that two consecutive bits become two parallel bits. Both of these bits will be transmitted simultaneously; in other words, as mentioned in this page, QPSK allows one symbol to transfer two bits. The local oscillator generates the carrier sinusoid. The local oscillator signal itself becomes the I carrier, and a 90° phase shift is applied to create the Q carrier. The I and Q carriers are multiplied by the I and Q data streams, and the two signals resulting from these multiplications are summed to produce the QPSK-modulated waveform.

The I and Q data streams are amplitude-modulating the I and Q carriers, and as explained above, these individual amplitude modulations can be used to produce phase modulation in the final signal. If the I and Q data streams were typical digital signals extending from ground to some positive voltage, we would be applying on-off keying to the I and Q carriers, and our phase shift would be restricted to 45° in either direction. However, if the I and Q data streams are *bipolar* signals—i.e., if they swing between a negative voltage and a positive voltage—our "amplitude modulation" is actually inverting the carrier whenever the input data is logic low (because the negative input voltage multiplied by the carrier results in inversion). This means that we will have four I/Q states:

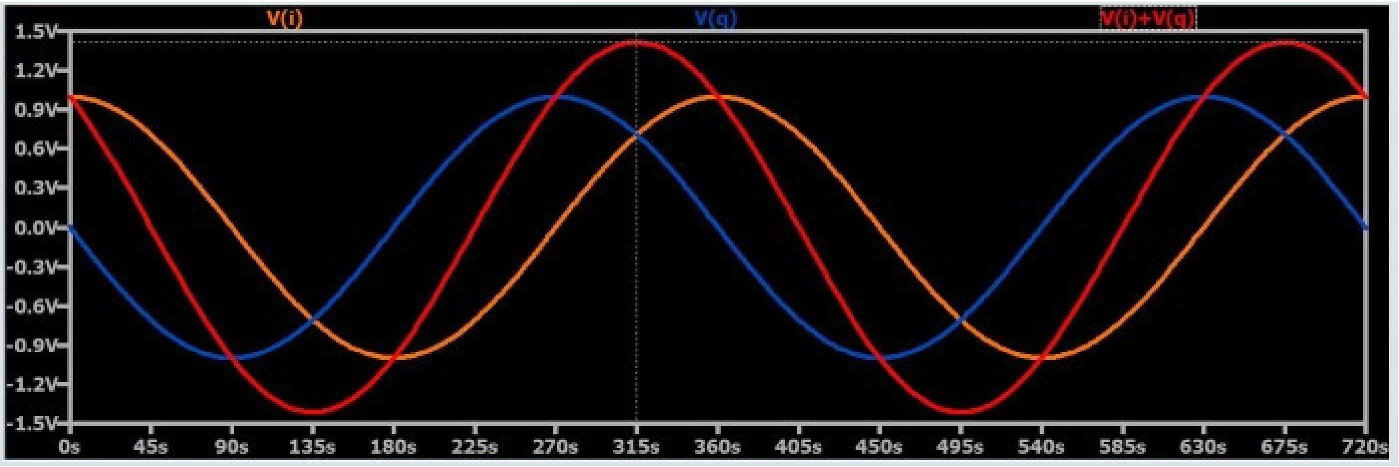
- I normal and Q normal
- I normal and Q inverted
- I inverted and Q normal
- I inverted and Q inverted

What will summation produce in each of these cases? (Note that in the following plots the frequency of the waveforms is chosen such that the number of seconds on the x-axis is the same as the phase shift in degrees.)

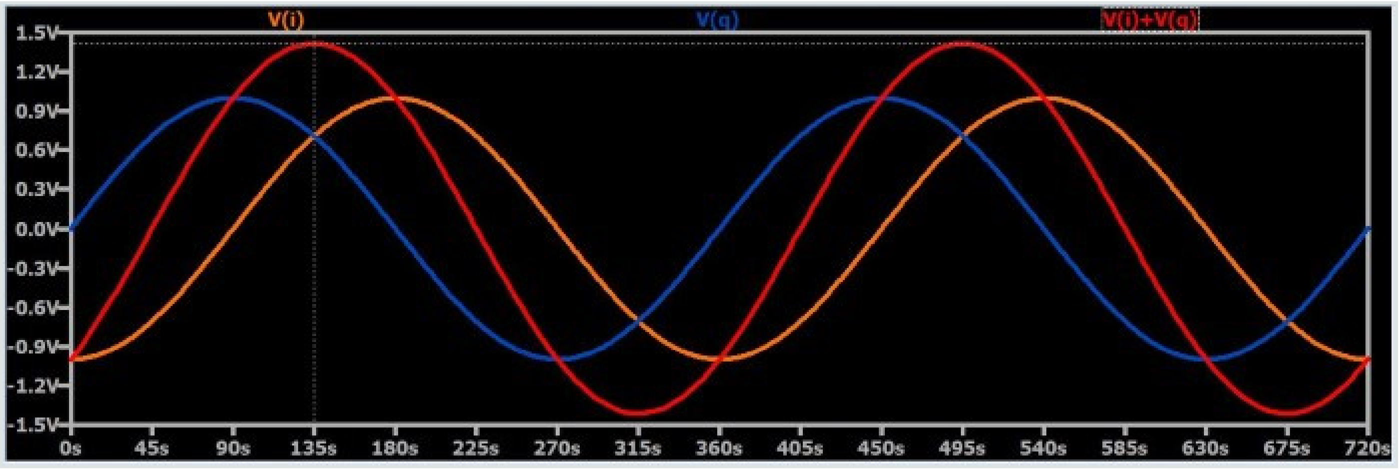
I Normal and Q Normal



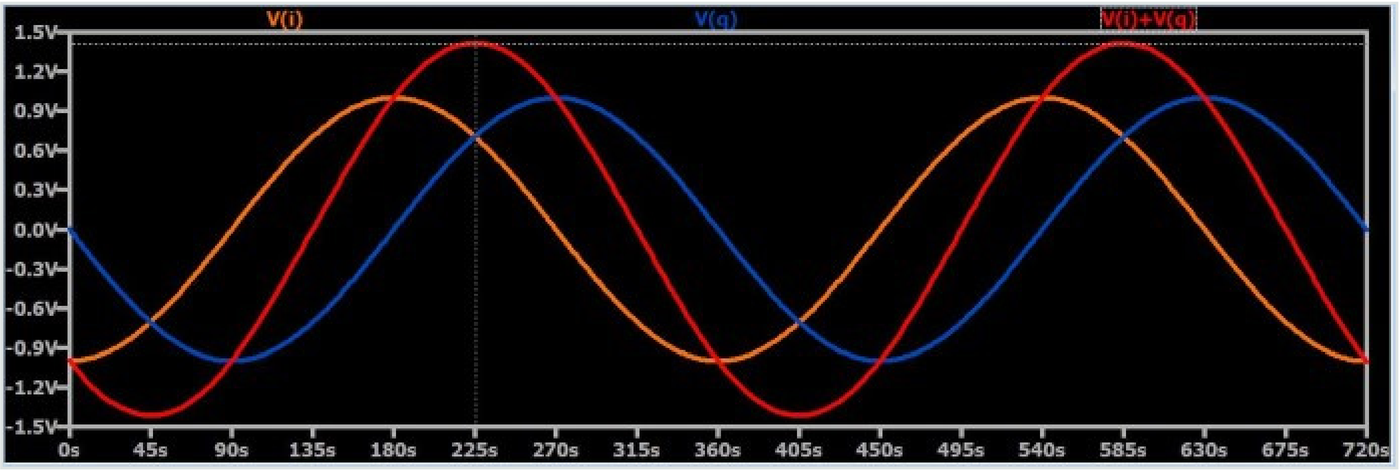
I Normal and Q Inverted



I Inverted and Q Normal



I Inverted and Q Inverted



As you can see, summation in these four cases produces exactly what we want to have in a QPSK signal: phase shifts of 45°, 135°, 225°, and 315°.

Summary

- I/Q signaling refers to the use of two sinusoids that have the same frequency and a relative phase shift of 90°.
- Amplitude, phase, and frequency modulation can be performed by summing amplitude-modulated I/Q signals.
- Quadrature modulation refers to modulation that involves I/Q signals.
- Quadrature phase shift keying can be accomplished by adding I and Q carriers that have been individually multiplied, in accordance with the incoming digital data, by +1 or –1.

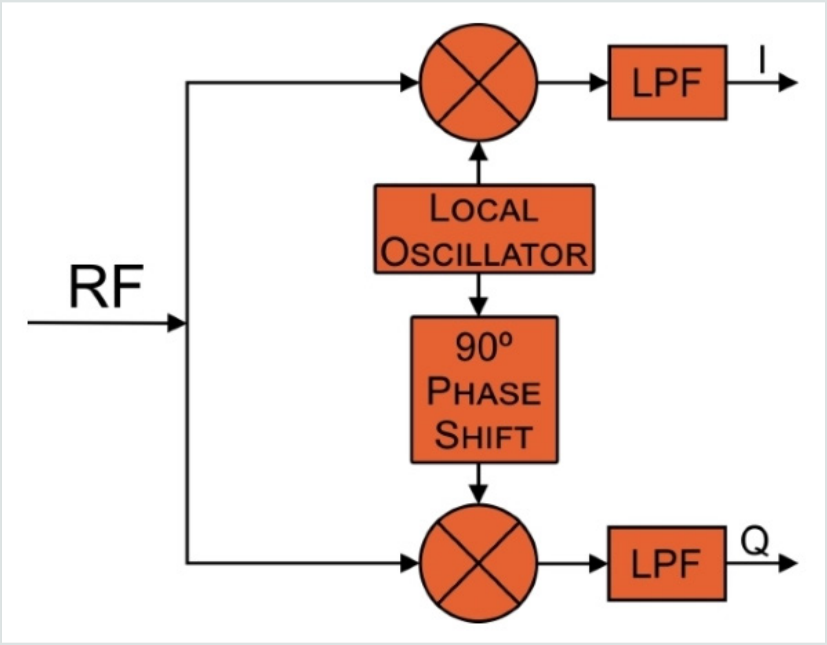
Understanding Quadrature Demodulation

This page explains what quadrature demodulation is and provides insight into the nature of I/Q signals.

If you have read the previous page, you know what I/Q signals are and how quadrature (i.e., I/Q-signal-based) modulation is accomplished. In this page we'll discuss quadrature demodulation, which is a versatile technique for extracting information from amplitude-, frequency-, and phase-modulated waveforms.

Converting to I and Q

The following diagram conveys the basic structure of a quadrature demodulator.



You will readily notice that the system is similar to a quadrature modulator in reverse. The RF signal is multiplied by the local oscillator signal (for the I channel) and the local oscillator shifted by 90° (for the Q channel). The result (after low-pass-filtering, which will be explained shortly) are I and Q waveforms that are ready for further processing.

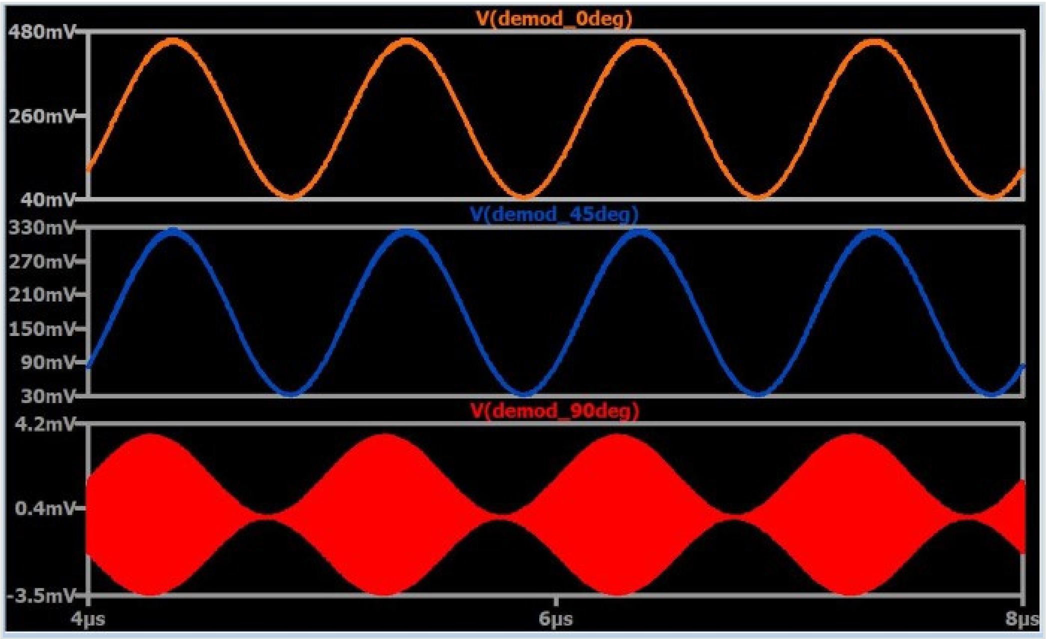
In quadrature modulation, we use baseband I/Q signals to create an amplitude-, frequency-, or phase-modulated waveform that will be amplified and transmitted. In quadrature demodulation, we are converting the existing modulation into the corresponding I/Q baseband signals. It's important to understand that the received signal could be from any sort of transmitter—quadrature demodulation is not limited to signals that were originally created through quadrature modulation.

The low-pass filters are needed because the quadrature multiplication applied to the received signal is no different from the multiplication employed in, for example, an ordinary AM demodulator. The received spectrum will be shifted downward and upward by the carrier frequency (f_C); thus, a low-pass filter is needed to suppress the high-frequency content associated with the spectrum centered around $2f_C$.

If you've read the page on amplitude demodulation, the preceding paragraph may have caused you to realize that a quadrature demodulator is actually composed of two amplitude demodulators. Of course, you cannot apply ordinary amplitude demodulation to a frequency-modulated signal; there is no information encoded in the FM signal's amplitude. But quadrature (amplitude) demodulation *can* capture the frequency-encoded information—this is simply the (rather interesting) nature of I/Q signals. By using two amplitude demodulators driven by carrier-frequency sinusoids with a 90° phase difference, we generate two different baseband signals that together can convey information encoded via changes in the received signal's frequency or phase.

Quadrature Amplitude Demodulation

As mentioned in the first page of this chapter, How to Demodulate an AM Waveform, one approach to amplitude demodulation involves multiplying the received signal by a carrier-frequency reference signal, and then low-pass-filtering the result of this multiplication. This method provides higher performance than AM demodulation that is built around a leaky peak detector. However, this approach has a serious weakness: the result of the multiplication is affected by the phase relationship between the transmitter's carrier and the receiver's carrier-frequency reference signal.

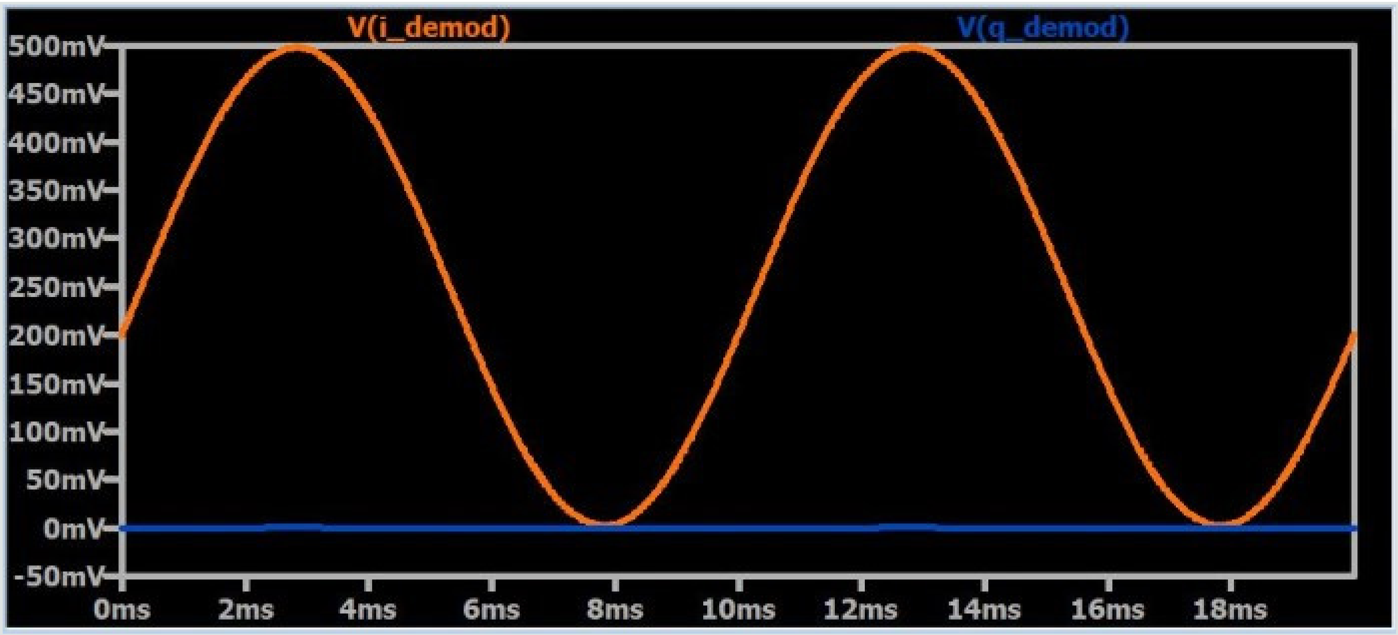


These plots show the demodulated signal for three values of transmitter-to-receiver phase difference. As the phase difference increases, the amplitude of the demodulated signal decreases. The demodulation procedure has become nonfunctional at 90° phase difference; this represents the worst-case scenario—i.e., the amplitude begins to increase again as the phase difference moves away (in either direction) from 90°.

One way to remedy this situation is through additional circuitry that synchronizes the phase of the receiver's reference signal with the phase of the received signal. However, quadrature demodulation can be used to overcome the absence of synchronization between transmitter and receiver. As was just pointed out, the worst-case phase discrepancy is ±90°. Thus, if we perform multiplication with two reference signals separated by 90° of phase, the output from one multiplier compensates for the decreasing amplitude of the output from the other multiplier. In this scenario the worst-case phase difference is 45°, and you can see in the above plot that a 45° phase difference does not result in a catastrophic reduction in the amplitude of the demodulated signal.

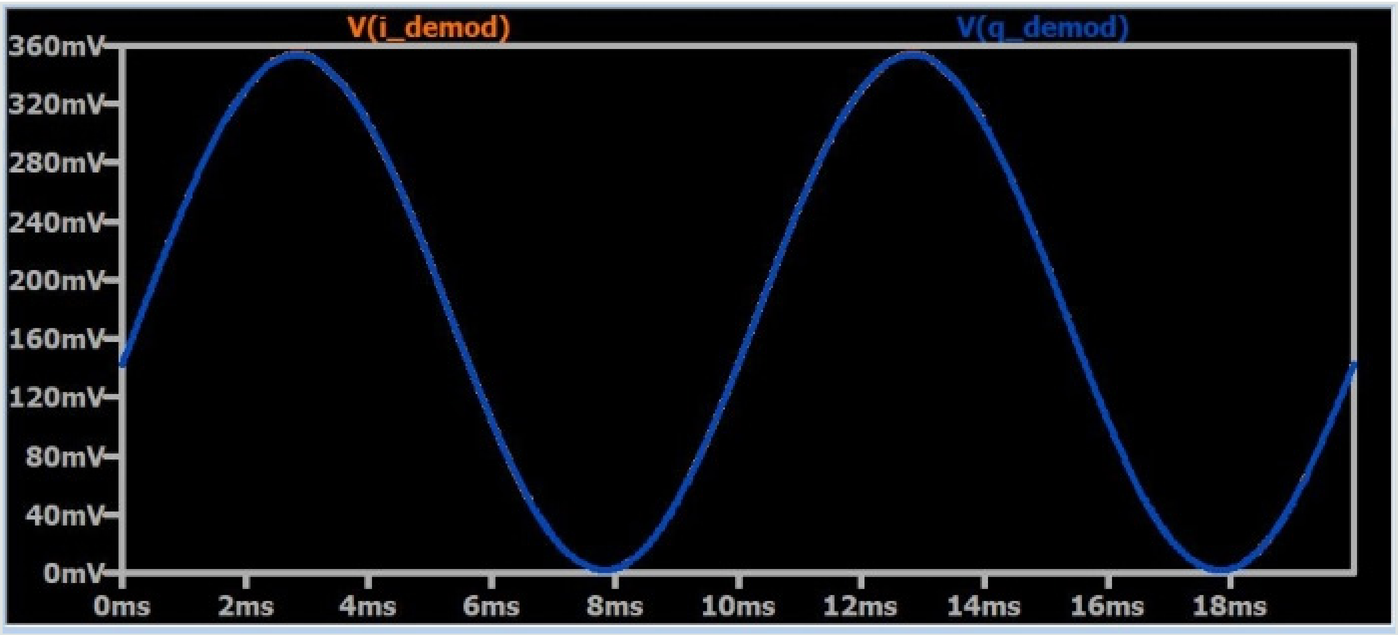
The following plots demonstrate this I/Q compensation. The traces are demodulated signals from the I and Q branches of a quadrature demodulator.

Transmitter phase = 0°

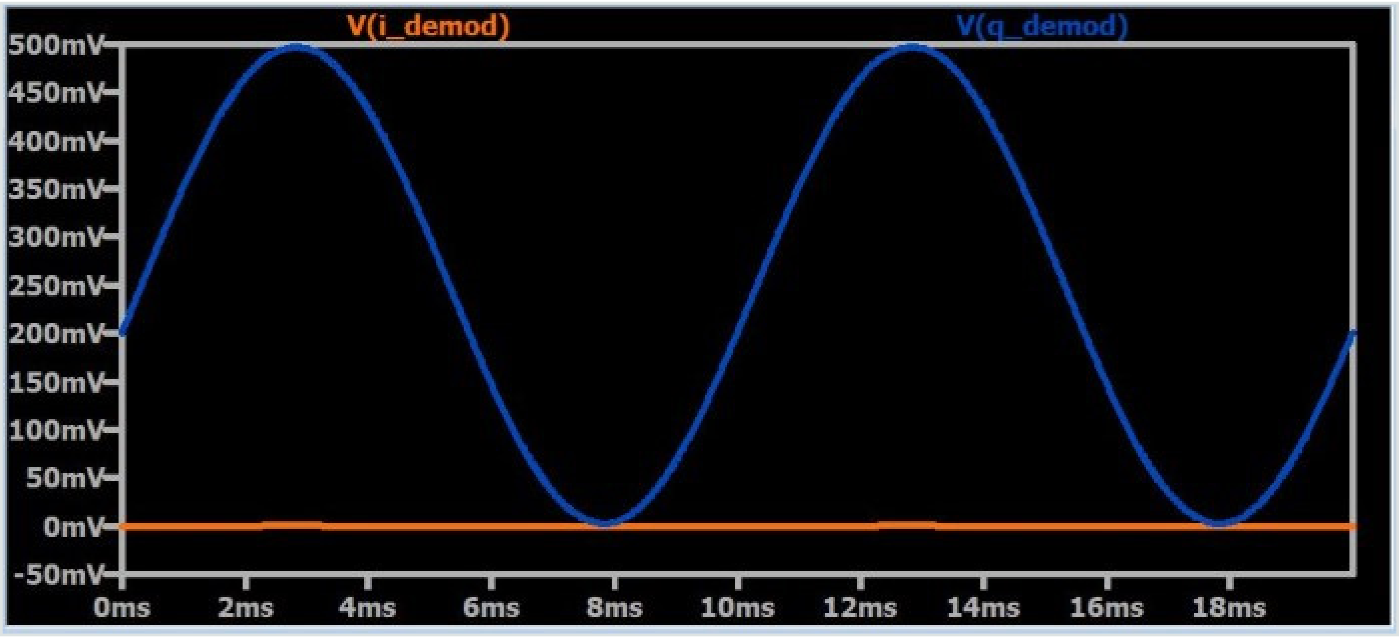


Transmitter phase = 45°

(the orange trace is behind the blue trace—i.e., the two signals are identical)

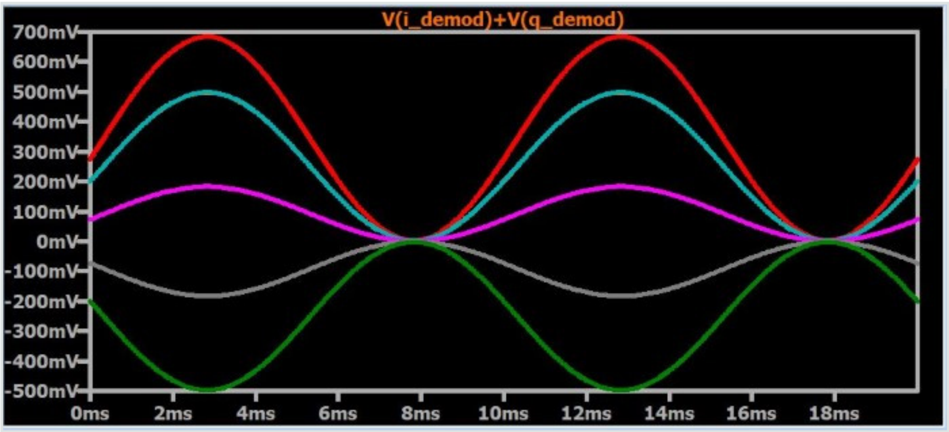


Transmitter phase = 90°

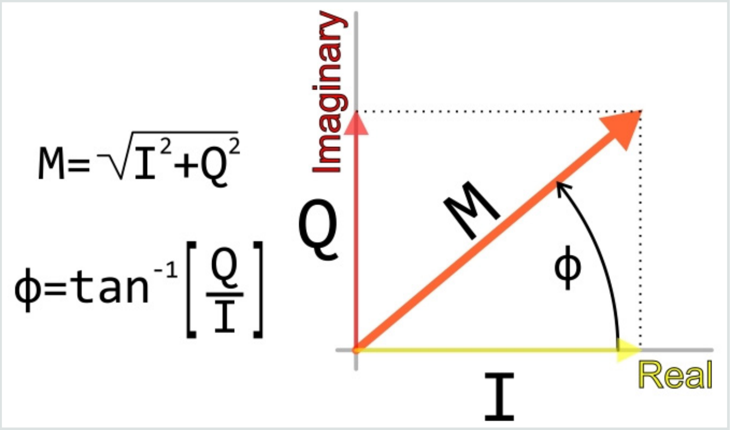


Constant Amplitude

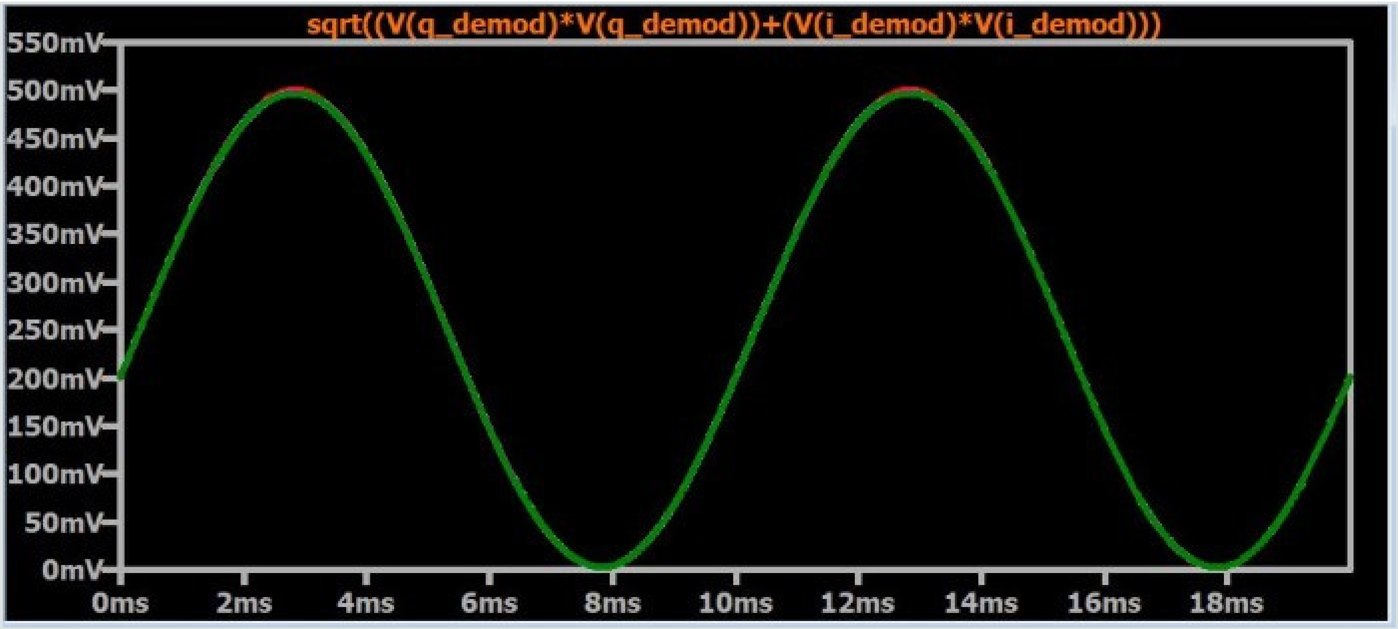
It would be convenient if we could combine the I and Q versions of the demodulated signal into one waveform that maintains a constant amplitude regardless of the phase relationship between transmitter and receiver. Your first instinct might be to use addition, but unfortunately it's not that simple. The following plot was generated by repeating a simulation in which everything is the same except the phase of the transmitter's carrier. The phase value is assigned to a parameter that has seven distinct values: 0°, 30°, 60°, 90°, 120°, 150°, and 180°. The trace is the sum of the demodulated I waveform and the demodulated Q waveform.



As you can see, addition is certainly not the way to produce a signal that is not affected by variations in the transmitter-to-receiver phase relationship. This is not surprising if we remember the mathematical equivalence between I/Q signaling and complex numbers: the I and Q components of a signal are analogous to the real and imaginary parts of a complex number. By performing quadrature demodulation, we obtain real and imaginary components that correspond to the magnitude and phase of the baseband signal. In other words, I/Q demodulation is essentially *translation:* we are translating from a magnitude-plus-phase system (used by a typical baseband waveform) to a Cartesian system in which the I component is plotted on the x-axis and the Q component is plotted on the y-axis.



To obtain the magnitude of a complex number, we can't simply add the real and imaginary parts, and the same applies to I and Q signal components. Instead, we have to use the formula shown in the diagram, which is nothing more than the standard Pythagorean approach to finding the length of the hypotenuse of a right triangle. If we apply this formula to the I and Q demodulated waveforms, we can obtain a final demodulated signal that is not affected by phase variations. The following plot confirms this: the simulation is the same as the previous one (i.e., seven different phase values), but you see only one signal, because all the traces are identical.



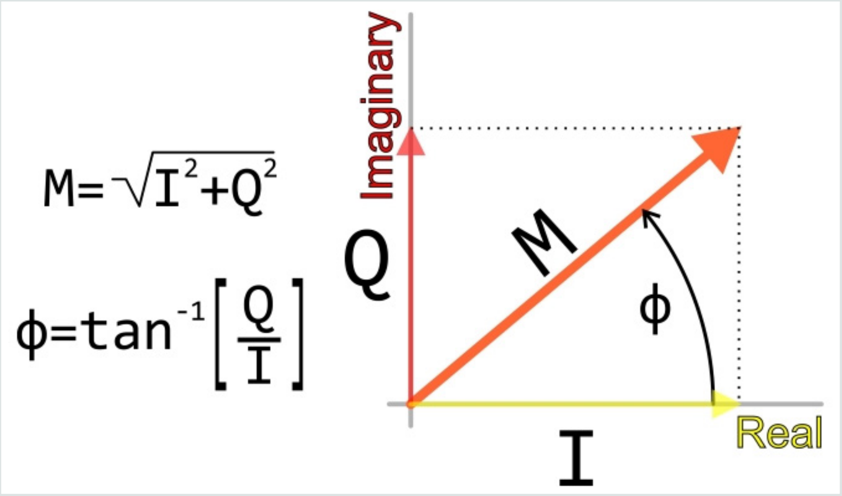
Summary

- Quadrature demodulation uses two reference signals separated by 90° of phase, along with two multipliers and two low-pass filters, to generate I and Q demodulated waveforms.
- Quadrature demodulation can be used to make an AM demodulator that is compatible with lack of phase synchronization between transmitter and receiver.
- The I and Q waveforms resulting from quadrature demodulation are equivalent to the real and imaginary parts of a complex number.

Quadrature Frequency and Phase Demodulation

This page explores the use of quadrature demodulation with frequency- and phase-modulated signals.

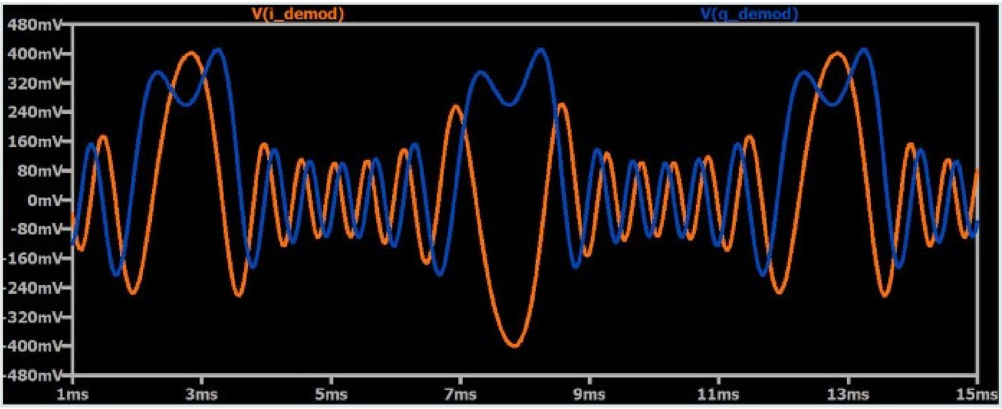
From the previous page we know that quadrature demodulation produces two baseband waveforms that, when taken together, convey the information that was encoded into the carrier of the received signal. More specifically, these I and Q waveforms are equivalent to the real and imaginary parts of a complex number. The baseband waveform contained in the modulated signal corresponds to a magnitude-plus-phase representation of the original data, and quadrature demodulation converts that magnitude-plus-phase representation into I and Q signals that correspond to a Cartesian representation.



It is perhaps not very surprising that we can use quadrature demodulation to demodulate AM signals, considering that a quadrature demodulator is simply two amplitude demodulators driven by carrier-frequency reference signals that have a 90° phase difference. However, one of the most important characteristics of quadrature demodulation is its universality. It works not only with amplitude modulation but also with frequency and phase modulation.

Quadrature Frequency Demodulation

First let’s look at the I and Q waveforms that are produced when we apply quadrature demodulation to frequency modulation. The received FM waveform is a 100 kHz carrier modulated by a 100 Hz sinusoid. We’re using the same quadrature demodulator that was used in the AM simulation; it has two arbitrary behavioral voltage sources for performing the multiplication, and each voltage source is followed by a two-pole low-pass filter (the cutoff frequency is ~1 kHz). You can refer to the page on How to Demodulate an FM Waveform for information on how to create an FM signal in LTspice.

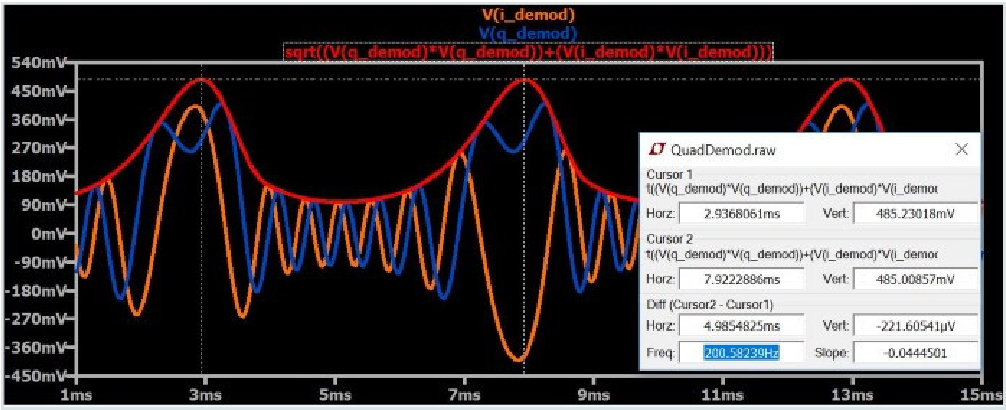


Perhaps the common reaction to this plot would be confusion. What do these odd-looking signals have to do with the constant-frequency sinusoid that should result from the demodulation process? First let's make two observations:

- Clearly, the frequency of the I and Q signals is not constant. You may find this a bit confusing at first, since we know that I/Q modulation involves the *amplitude* modulation of quadrature carriers. Why is the frequency changing as well? It's essential to remember that these I/Q signals correspond to the *modulating* signals, not to the quadrature sinusoids that would be added together in a quadrature modulator. The frequency of the *modulated* quadrature carriers does not change, but the baseband waveforms that serve as the amplitude-modulating signals do not necessarily have constant frequency.
- Though we cannot intuitively interpret the information in this plot, we can see that the signals exhibit periodic variations and that these variations correspond to the period (=10 ms) of the 100 Hz baseband signal.

Finding the Angle

Now that we have I/Q signals, we need to somehow process them into a normal demodulated waveform. Let's first try the approach that we used with amplitude modulation: use a bit of math to extract the magnitude data.



Cursor 1
t((V(q_demod)*V(q_demod))+(V(i_demod)*V(i_democ
Horz: 2.9368061ms Vert: 485.23018mV

Cursor 2
t((V(q_demod)*V(q_demod))+(V(i_demod)*V(i_democ
Horz: 7.9222886ms Vert: 485.00857mV

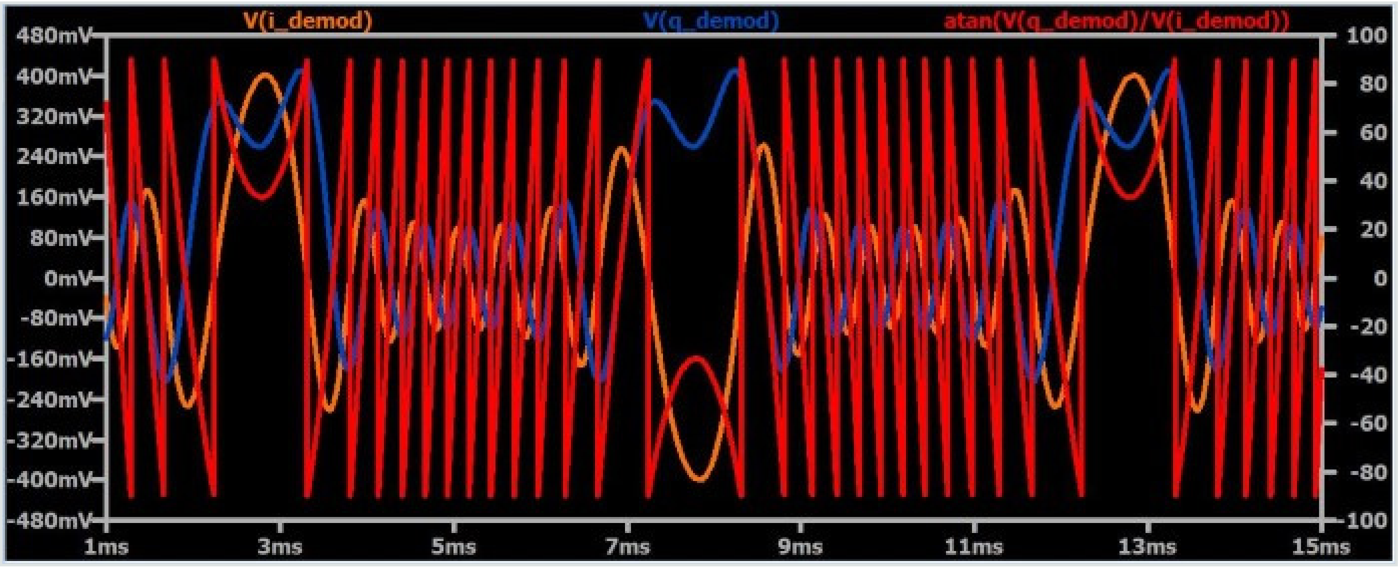
Diff (Cursor2 - Cursor1)
Horz: 4.9854825ms Vert: -221.60541µV
Freq: 200.58239Hz Slope: -0.0444501

Clearly this didn't work: the magnitude signal (the red trace) doesn't look like a sinusoid, and the frequency is incorrect (200 Hz instead of 100 Hz). After further consideration, though, this is not surprising. The original data is characterized by magnitude and phase; when we apply the $\surd(I_2 + Q_2)$ computation, we are extracting the magnitude. The trouble is, the original data was not encoded in the magnitude of the carrier—it was encoded in the angle (remember that frequency modulation and phase modulation are two forms of angle modulation).

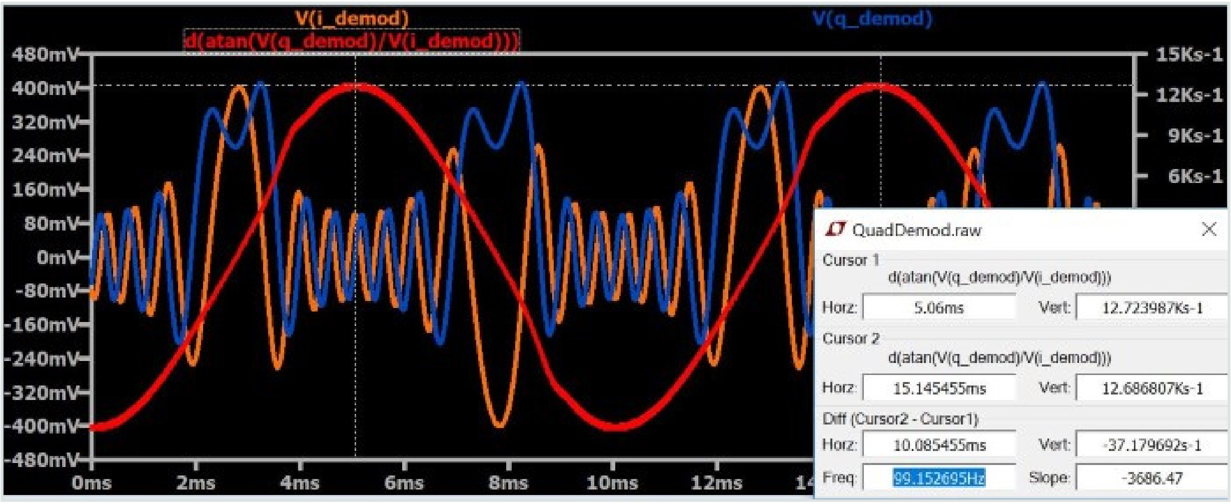
So let's try a different computation. Let's extract the angle of the I/Q data rather than the magnitude. As shown in the right-triangle diagram above, we can do this by applying the following equation:

$$\phi = \arctan\left(\frac{Q}{I}\right)$$

Here is the result:



This doesn't look good, but we are actually getting close. The red trace represents the instantaneous phase of the original data. (Note that the trace seems more erratic than it really is because the angle is jumping from –90° to +90°, or vice versa). Frequency modulation, though based on phase, does not encode information *directly* in the phase of the carrier. Rather, it encodes information in the instantaneous frequency of the carrier, and instantaneous frequency is the derivative of instantaneous phase. So what happens if we take the derivative of the red trace?



QuadDemod.raw

Cursor 1
d(atan(V(q_demod)/V(i_demod)))
Horz: 5.06ms Vert: 12.723987Ks-1

Cursor 2
d(atan(V(q_demod)/V(i_demod)))
Horz: 15.145455ms Vert: 12.686807Ks-1

Diff (Cursor2 - Cursor1)
Horz: 10.085455ms Vert: -37.179692s-1
Freq: 99.152695Hz Slope: -3686.47

As you can see, we have now recovered a waveform that is sinusoidal and has the same frequency as the original baseband signal.

How to Design an Arctangent Circuit

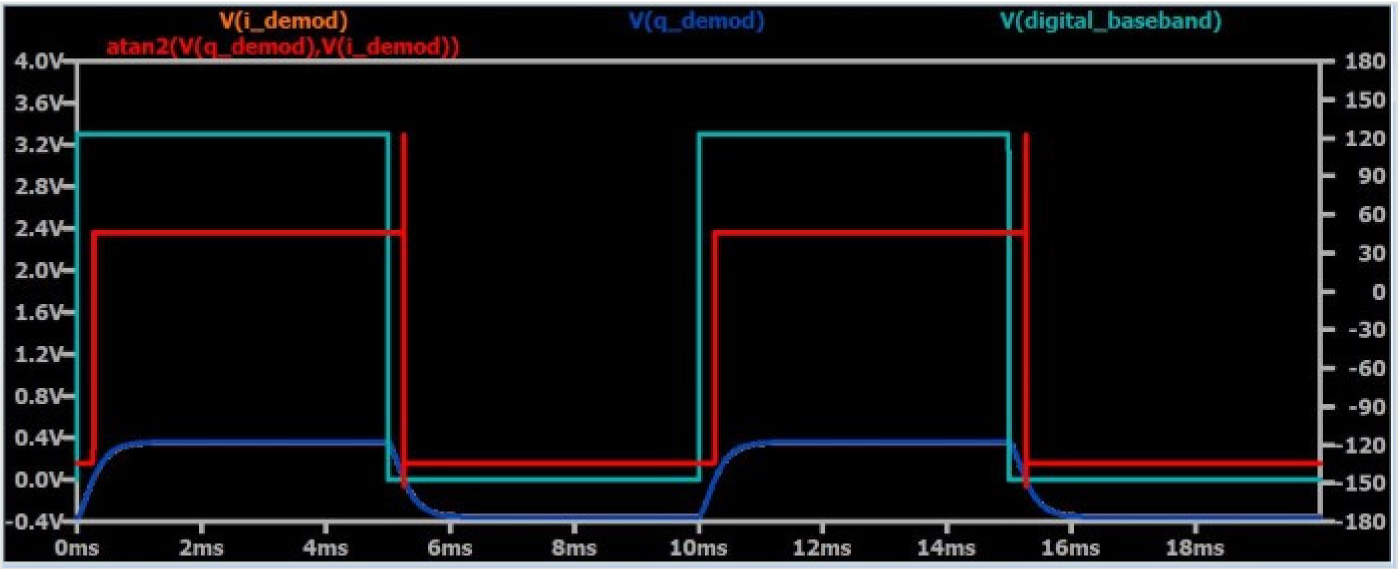
At this point you might be wondering why anyone would want to bother with I/Q demodulation. How in the world would anyone design a circuit that generates an output signal corresponding to the derivative of the arctangent of two input signals? Well, to answer the question posed in the title of this section, you digitize the signals and compute the arctangent in firmware or software. And this brings us to an important point: Quadrature demodulation is especially advantageous in the context of software-defined radios.

A software-defined radio (SDR) is a wireless communication system in which significant portions of the transmitter and/or receiver functionality are implemented via software. Quadrature demodulation is highly versatile and enables a single receiver to almost instantaneously adapt to different types of modulation. The I/Q output signals, however, are far less straightforward than a normal baseband signal produced by standard demodulator topologies. This is why a quadrature demodulator and a digital signal processor form such a high-performance receiver system: the digital signal processor can readily apply complicated mathematical operations to the I/Q data produced by the demodulator.

Quadrature Phase Demodulation

The same general considerations that we discussed in the context of quadrature frequency demodulation apply also to quadrature phase demodulation. However, to recover the original data we take the arctangent of (Q/I) rather than the derivative of the arctangent of (Q/I), because the baseband signal is encoded *directly* in the carrier’s phase rather than in the derivative of the phase (i.e., the frequency).

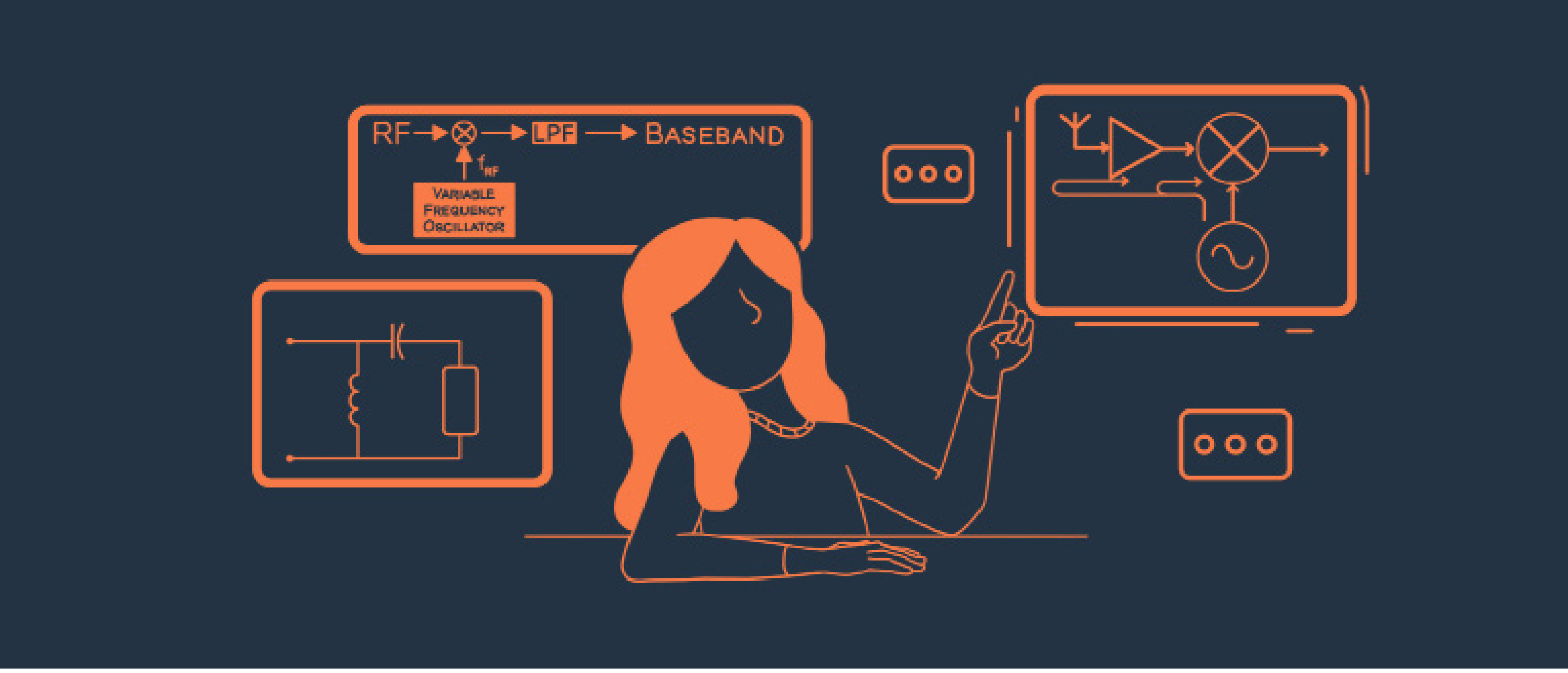
The following plot was generated by applying quadrature demodulation to a phase-shift-keying waveform consisting of a 100 kHz carrier and a 100 Hz digital baseband signal that causes the carrier's phase to change by 180° according to whether the signal is logic high or logic low. As you can see, the red trace (whose value corresponds to the phase of the received waveform) reproduces the logic transitions in the baseband signal.



Notice that the red trace is computed via the "atan2" function. Standard arctangent is limited to two quadrants (i.e., 180°) of the Cartesian plane. The atan2 function looks at the individual polarities of the input values in order to produce angles covering all four quadrants.

Summary

- Quadrature demodulation can extract angle information that is relevant to both frequency modulation and phase modulation.
- Radio systems can use a digital signal processor (in conjunction with an analog-to-digital converter) to apply mathematical analysis to I/Q waveforms.
- Baseband phase can be obtained by taking the arctangent of the ratio of Q to I; an "atan2" function is needed if the system must be able to reproduce the full 360° of phase.
- Baseband frequency can be obtained by taking the derivative of the arctangent of the ratio of Q to I.



CHAPTER 6

Selected Topics

- The Benefits of an Intermediate Frequency in RF Systems
- Image Rejection and Direct-Conversion Receivers
- Understanding Matching Networks
- Understanding Spread-Spectrum RF Communication

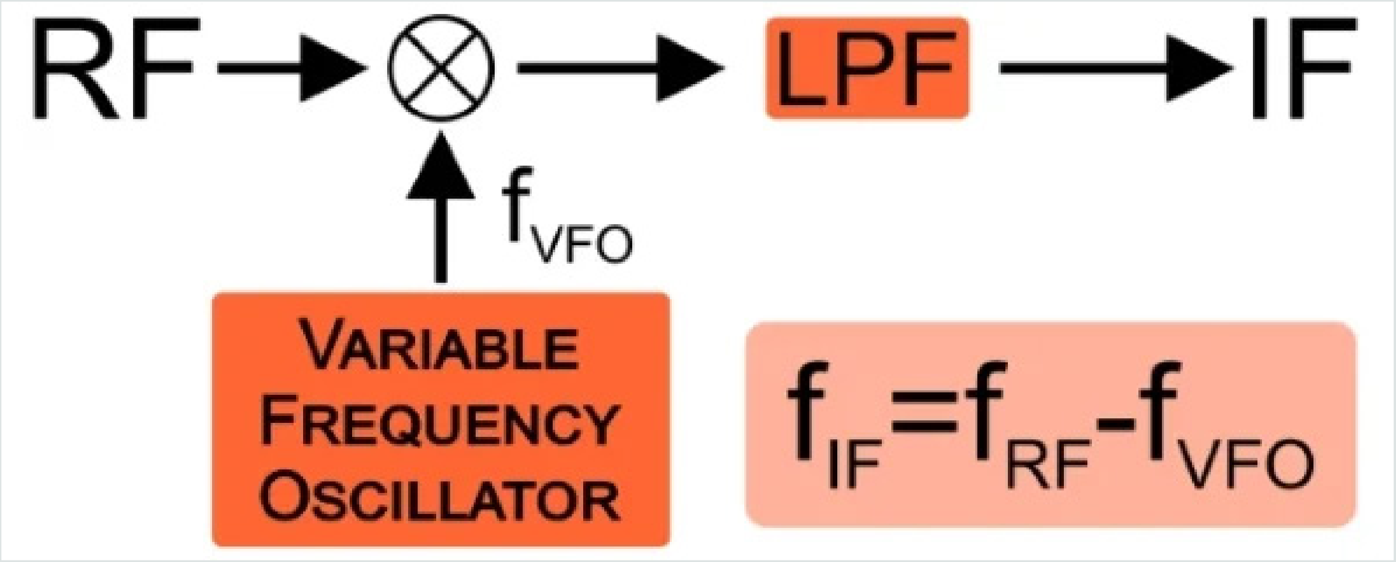
The Benefits of an Intermediate Frequency in RF Systems

Learn about “IF”—a widespread and advantageous technique used in many wireless systems.

Thus far, we have discussed RF signals in terms of two frequency bands: the baseband and the RF band. This approach provides a straightforward conceptual framework in which RF circuits are fundamentally a means of transforming a lower-frequency information signal into a higher-frequency transmitted signal, or a higher-frequency received signal into a lower-frequency information signal. This model is not incorrect, and the lessons learned so far are completely relevant to systems that have an “intermediate frequency” signal in addition to baseband and RF signals.

What Is IF?

The abbreviation “IF” refers to an intermediate frequency itself or, more generally, to intermediate-frequency-based techniques. As the name implies, an intermediate frequency is somewhere between the baseband frequency and the carrier frequency. IF circuitry can be incorporated into both transmitters and receivers, though the benefits of IF techniques are more relevant to receivers. We’ll discuss IF in the context of RF receiver design, but as you’re reading keep in mind that these beneficial characteristics could apply to transmitters as well.



Perhaps you have heard the word “heterodyne” or “superheterodyne.” These terms refer to an RF receiver that incorporates an intermediate frequency. IF techniques were developed during the first half of the twentieth century, and nowadays IF-based systems are very common.

Many Carriers, One IF

One of the more intuitive advantages of an IF is the ability to design a receiver in which more of the circuitry can be designed for one unchanging frequency band. Thus far we have assumed that the receiver can be designed for one unchanging transmitter frequency, but anyone who has used a car radio should understand that this is far from realistic. In fact, one of the most familiar characteristics of an RF receiver is that it can convey to the user information from only one station (for radio) or only one channel (for television)—in other words, it can be tuned for different carrier frequencies, and this tuning process allows it to select one of the transmitted signals and ignore all the others.

If a tunable receiver does not use an intermediate frequency, all of the high-frequency circuitry must be compatible with the full range of possible carrier frequencies; this is undesirable, because it is easier to design RF components and circuits that are optimized for a small range of signal frequencies. Also, tuning would require several knobs, because multiple subcircuits would need to be adjusted according to the selected frequency. A heterodyne receiver first shifts the received spectrum down to a band centered on the intermediate frequency, and then the remaining circuitry is optimized for this frequency range.

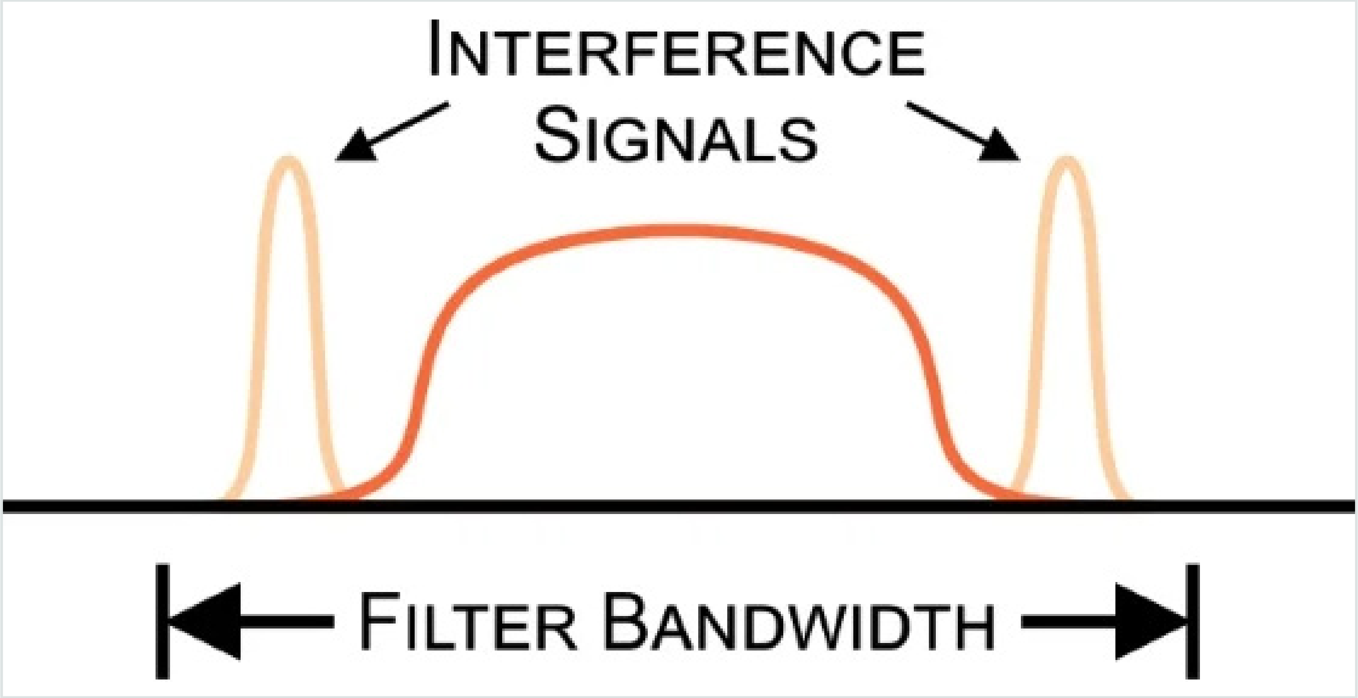
Minimizing High-Frequency Processing

Another intuitive advantage of an IF-based receiver architecture is the reduced number of components that must operate at the high—sometimes very high—frequency of the received signal. Everything becomes more difficult as frequencies climb into the gigahertz range: transistors have less gain, passive components become increasingly different from their idealized low-frequency models, transmission-line effects become more prominent.

Of course, we will always have at least a few components that are compatible with the received carrier frequency: we need a mixer that performs the conversion from RF to IF, and the mixer might be preceded by a low-noise amplifier and an image-reject filter (the image-rejection issue is discussed in the next page). But the IF approach allows us to perform only the most necessary processing in the RF band.

Lower Q

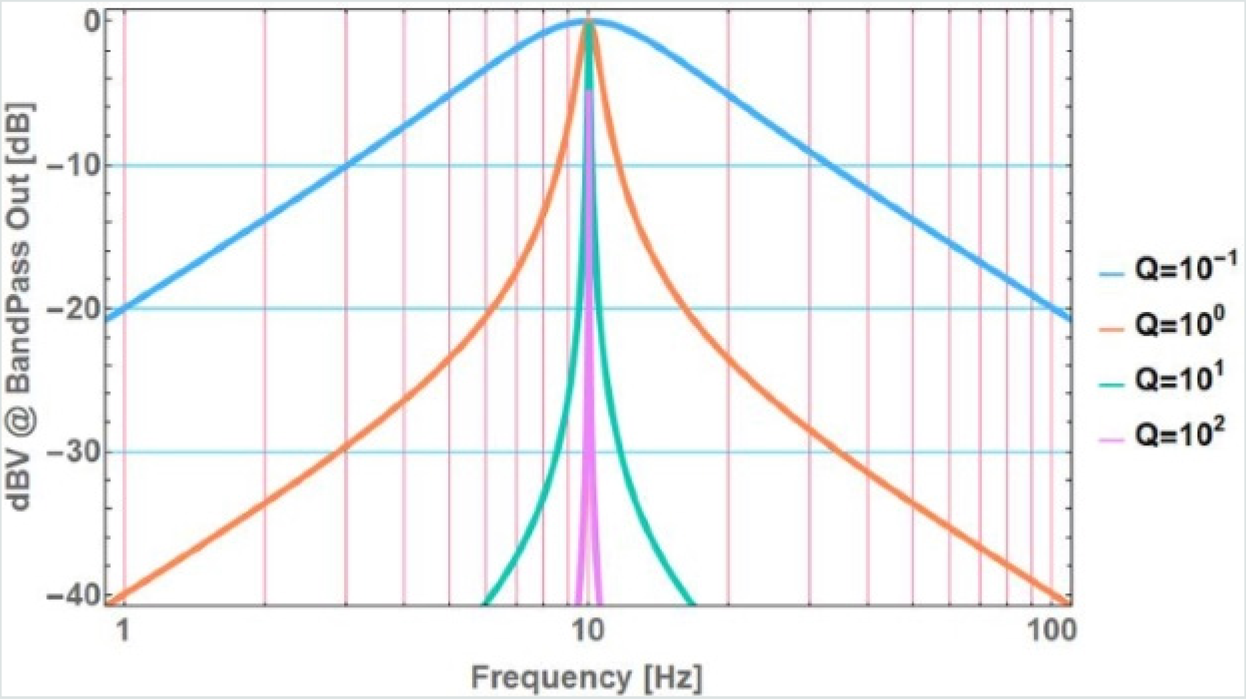
Filtering is a common requirement in all types of RF systems, but some situations place especially high demands on filter circuits. Consider the following scenario: A receiver must extract the information from a narrowband RF signal that is accompanied by strong interfering signals with frequency close to the edges of the spectrum of the desired signal.



A band-pass filter with insufficient Q may fail to suppress interfering signals.

A band-pass filter is used to suppress these interfering signals so that they don't corrupt the demodulated data; however, designing an effective band-pass filter under these circumstances is not easy.

The issue is the Q factor, which corresponds to how selective the band-pass filter is. For example:



A combination of high-frequency operation and narrow bandwidth requires a very high Q, and eventually we reach a point at which it is simply not feasible to design a band-pass filter with sufficient selectivity. The Q factor of a band-pass filter is defined as follows:

$$Q = \frac{center\ frequency}{bandwidth}$$

Thus, we can see that a straightforward way to decrease the required Q is to lower the center frequency, and IF techniques allow us to do exactly that. The width of the signal's spectrum does not change, but the center frequency is shifted down to the intermediate frequency.

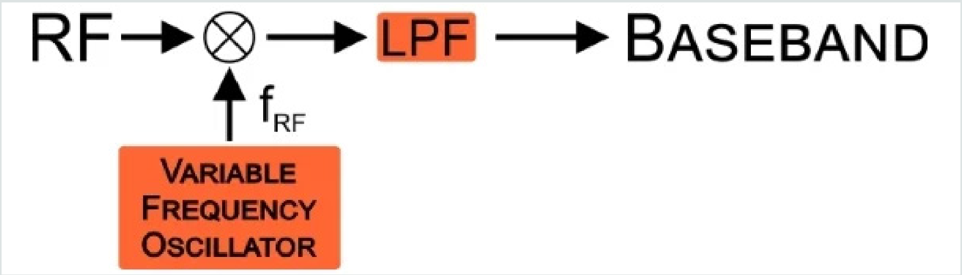
Simpler Quadrature Demodulation

We know from the previous chapter that quadrature demodulation is an important technique in modern RF systems. The mathematical relationships that govern quadrature demodulation and I/Q signal processing always assume a perfect 90° phase shift. But perfection is not so easily achieved in real life, and quadrature circuitry is no exception. Deviations from the idealized 90° phase difference, as well as amplitude mismatches between the I and Q channels, lead to errors in the demodulated data.

This may seem like an issue with quadrature modulation in general; what is the connection to IF receivers? It turns out that these error sources are more prominent in non-IF architectures, because the I/Q separation occurs at higher frequencies and because more post-separation amplification and filtering components are required.

Why Not Convert Directly to Baseband?

If an IF receiver must include high-frequency circuitry for performing the frequency translation from RF to IF, why not simply use the baseband frequency instead of an intermediate frequency?



A receiver that shifts the signal down to the baseband instead of the IF is referred to as a direct-conversion (or homodyne, or zero-IF) architecture. Are the traditional benefits of an intermediate frequency still—i.e., in the context of modern RF systems—sufficient reason for choosing IF over a direct-conversion approach? The answer to this question is somewhat complex, and it goes beyond the topics presented in this page. In the next page we'll explore more details regarding IF receivers, and we'll also discuss the heterodyne vs. direct-conversion issue.

Summary

- Many RF systems incorporate an intermediate frequency (IF) that is lower than the carrier frequency and higher than the baseband frequency. An IF-based receiver is known as a heterodyne receiver.
- The use of an IF simplifies the design of tunable receivers and reduces the number of components that must be compatible with high frequencies.
- IF architectures simplify the design of bandpass filters because the reduced center frequency results in a lower Q-factor requirement.
- An IF-based system allows for a more robust implementation of quadrature demodulation.

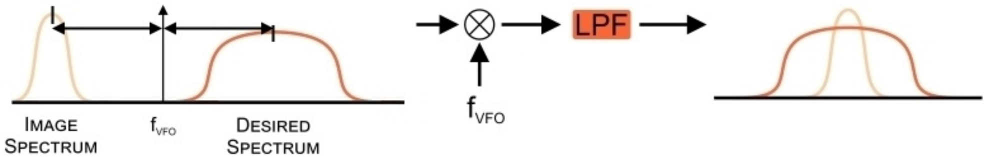
Image Rejection and Direct-Conversion Receivers

In this page we'll discuss the "image" problem in IF-based receivers, and we'll also look at an alternative approach that eliminates this complication.

In the previous page we explored benefits associated with the use of an intermediate frequency (IF). However, IF architectures entail a serious disadvantage, and in fact this disadvantage is a significant motivating factor in the development of direct-conversion-based alternatives.

The Image

An IF-based receiver uses a variable-frequency oscillator (VFO) signal to shift a received spectrum down to an equivalent spectrum centered around the intermediate frequency; the shifting is accomplished via multiplication. However, this multiplication operation affects not only the received spectrum but also whatever spectrum is located symmetrically with respect to the VFO frequency. In other words, multiplication will shift one spectrum *down* to the IF and another spectrum *up* to the IF.



As you can see, the image spectrum and the desired spectrum are both present in the IF signal that will be demodulated. In this diagram we can easily distinguish one from the other, but such is not the case in a real-life circuit—the frequency information in the desired spectrum is corrupted by the frequency information in the image spectrum.

This symmetrically located image spectrum is a serious impediment to reliable IF-based reception. Why? Because the image spectrum is (presumably) not under the control of the wireless system that you're designing, and consequently it could be anything, including a signal that is much stronger than the desired signal. Thus, if we don't do something to mitigate the effects of the image, the quality of the system's reception will be dependent upon the unpredictable behavior of the signals near the image frequency.

Image Rejection

To mitigate the effects of the image spectrum, heterodyne receivers use image-reject filters. These are placed before the mixer, such that the image spectrum is suppressed before the mixer shifts it to the intermediate frequency. This is an effective solution, but there are two complications.

The Trade-Off

The image-reject filter won't be very useful if it attenuates the desired spectrum and the image spectrum. Thus, the filter's response must transition from low attenuation at the desired band to high attenuation at the image band. As with any filter, rapid transitions from passband to stopband are challenging, and thus it will be easier to design an image-reject filter if there is a large frequency separation between the desired band and the image band.

However, the separation between the desired band and the image band is proportional to the intermediate frequency (more specifically, it is twice the intermediate frequency). This means that more separation corresponds to a higher IF. This is not catastrophic, but we have to remember that we want an intermediate frequency to be significantly more convenient, from a signal-processing perspective, than the high frequency used for RF transmission. If we increase the intermediate frequency too much, the difficulties created by the higher IF may outweigh the benefits of improved image rejection. Thus, image-rejection filtering entails a fundamental trade-off between image suppression and the desire to maintain a lower intermediate frequency.

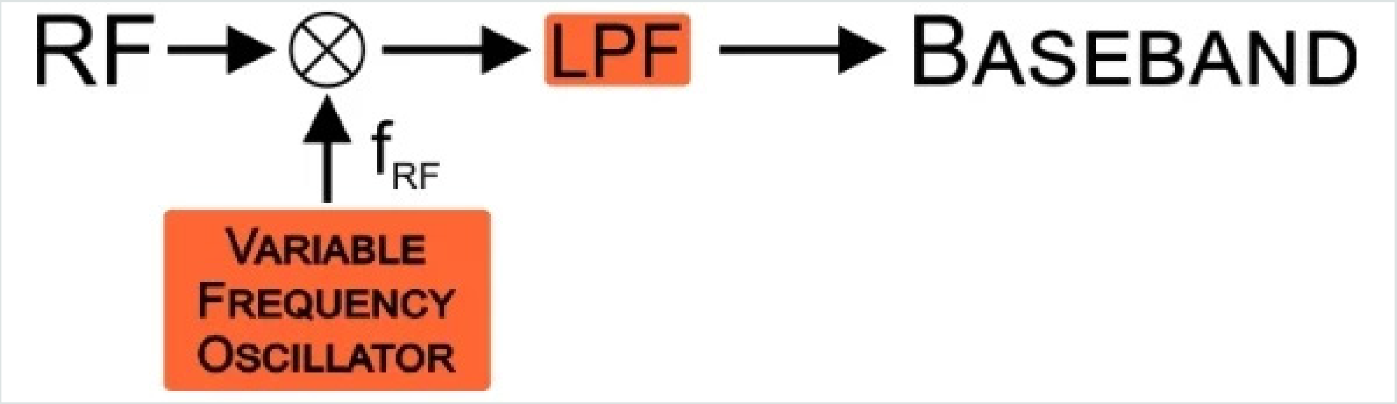
Integration, or Lack Thereof

Image rejection is typically accomplished by means of a filter that is not incorporated into an integrated circuit. In other words, image-reject filters consume PCB area and design time, and in the context of modern electronics, both of these resources are valuable and in short supply.

Companies often attempt to minimize the time involved in bringing a new product to the production phase, and an important way to reduce development time is to avoid custom design whenever possible—in other words, to use tested, characterized, proven integrated circuits instead of newly designed external circuits. Regarding PCB area, it should come as no surprise that miniaturization is a major goal in various electronics industries, and the only way to achieve extreme size reductions is through integrated-circuit technology. Thus, heterodyne receivers that rely on image-reject filters are fundamentally problematic with respect to the inescapable realities of modern electronic design.

A Possible Solution: Direct Conversion

As mentioned in the previous page, a direct-conversion receiver shifts the received signal all the way to baseband instead of to an intermediate frequency. In other words, the frequency of the VFO is always equal to the center frequency of the desired spectrum:



This approach includes one very important benefit—it eliminates the image problem. In the direct-conversion scheme, there is no image spectrum: the desired spectrum is centered around the VFO frequency, and no spectrum can be symmetrical with respect to the VFO frequency when the desired center frequency and the VFO frequency are equal.

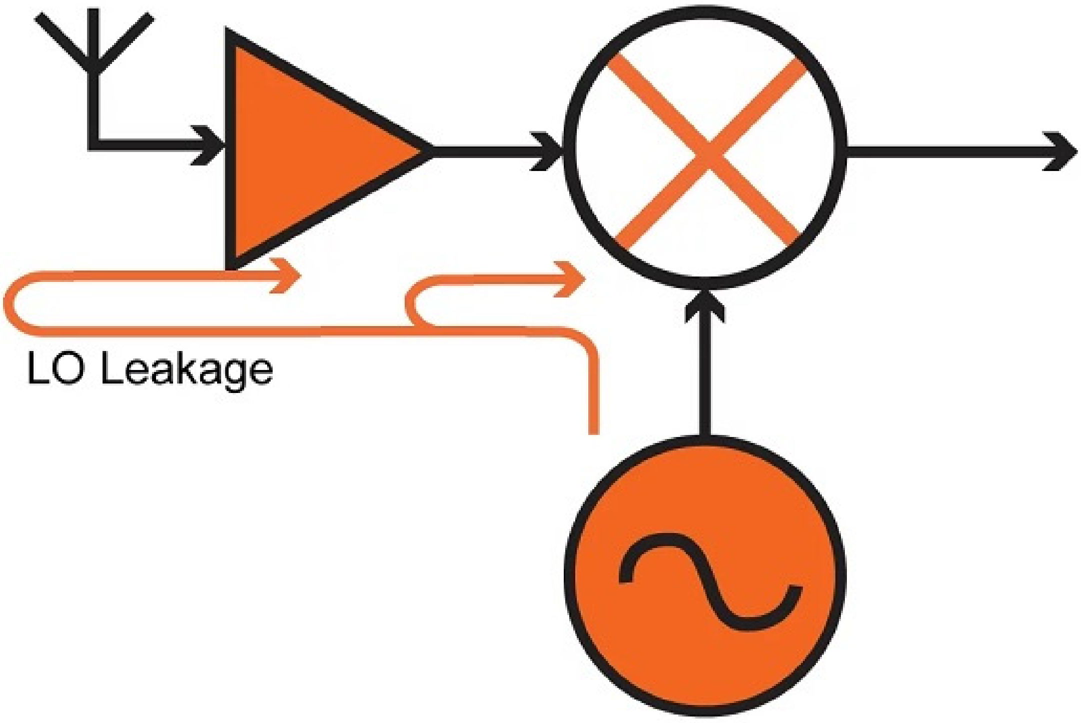
Another benefit of direct conversion is simply an extension of the benefits associated with IF-based architectures. An intermediate frequency facilitates signal processing because it is significantly lower than the transmission frequency, but signal processing can be even easier when the “intermediate” frequency is 0 Hz—i.e., when the received spectrum is shifted directly to baseband.

Direct-conversion immediately seems like a superior alternative: it’s conceptually simpler, there’s no image to corrupt the received spectrum, low-frequency signal processing replaces intermediate-frequency signal processing, and the absence of an image-reject filter allows for expanded use of IC technology. Why, then, does anyone even consider an IF-based architecture? Well, it turns out that there are several significant disadvantages associated with direct conversion. Here we will discuss only the disadvantage that is perhaps the most serious.

DC Offset

RF receivers are sensitive to DC signal components because the amplitude of received signals is often extremely small. These small-amplitude signals create the need for high-gain amplification, but high-gain amplification can rapidly break down when a signal has a nontrivial DC offset, because multiplication of the offset saturates the amplifier.

Mixers readily create DC offsets, because multiplying a sinusoid by another sinusoid with the same frequency and phase creates a non-varying signal component. Back in Chapter 3, we discussed the complications caused by the fact that RF signals do not stay in their intended conduction paths. Rather, their high frequency allows them to "leak" into portions of the circuit where we don't want them. The problem of DC offset creation is a perfect example of this difficulty: the local oscillator signal leaks into other portions of the circuit in such a way that it is present in both of the mixer's inputs, and the result is a DC offset in the output signal.



A direct-conversion receiver must implement some sort of DC offset cancellation, and this is not a particularly easy task; filtering is generally not feasible because the filter would suppress portions of the desired spectrum, which has been shifted down to the band around DC. A heterodyne receiver, on the other hand, can easily remove DC offsets via filtering because there is plenty of frequency separation between DC and the IF band.

Summary

- IF downconversion causes an image spectrum to corrupt the desired spectrum.
- The image spectrum can be suppressed via filtering, but the filtering approach to image rejection involves an important design trade-off and prevents monolithic integration.
- Direct conversion eliminates the image problem, but it has various disadvantages.
- A particularly challenging characteristic of the direct-conversion architecture is its susceptibility to DC offsets generated by the mixer.

Understanding Matching Networks

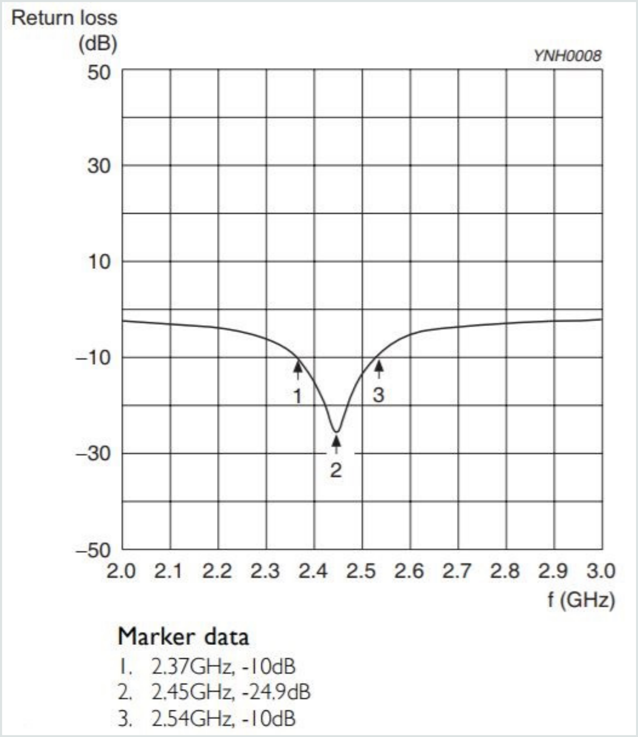
Learn about why matching networks are used and how they are designed.

Back in Chapter 3, we discussed characteristic impedance, transmission lines, and impedance matching. We know that transmission lines have a characteristic impedance and we know that this impedance is an important factor in RF circuitry, because impedances must be matched to prevent standing waves and to ensure efficient transfer of power from source to load. And even if we don't need to treat a particular conductor as a transmission line, we still have source and load impedances that need to be matched.

Also in Chapter 3, we saw that impedance matching is greatly simplified by the use of standardized impedance values (the most common being 50 Ω). Manufacturers design their components or interconnects for 50 Ω input and output, and in many cases an engineer does not have to take any specific action to achieve matched impedances.

However, there are situations in which impedance matching requires additional circuitry. For example, consider an RF transmitter composed of a power amplifier (PA) and an antenna. The manufacturer can design the PA for 50 Ω output impedance, but the impedance of an antenna will vary according to its physical characteristics as well as the characteristics of the surrounding materials.

Also, the antenna's impedance is not constant relative to signal frequency. Thus, a manufacturer could design an antenna that has 50 Ω impedance at one specific frequency, but you might have a nontrivial mismatch if you use the antenna at a different frequency. The following plot is taken from the datasheet for a ceramic surface-mount antenna intended for 2.4–2.5 GHz systems. The curve corresponds to the ratio of reflected power to incident power. You can see that the quality of the impedance match deteriorates rapidly as the signal frequency moves away from 2.45 GHz.



Plot taken from this datasheet.

The Matching Network

If your RF circuit contains components that do not have matched impedances, you have two options: modify one of the components, or add circuitry that corrects the mismatch. Nowadays the first option is generally not practical; it would be difficult indeed to adjust impedance by physically modifying an integrated circuit or a manufactured coaxial cable. Fortunately, though, the second option is perfectly adequate. The additional circuit is called a matching network or an impedance transformer. Both names are helpful in understanding the fundamental concept: a matching network enables proper impedance matching by transforming the impedance relationship between source and load.

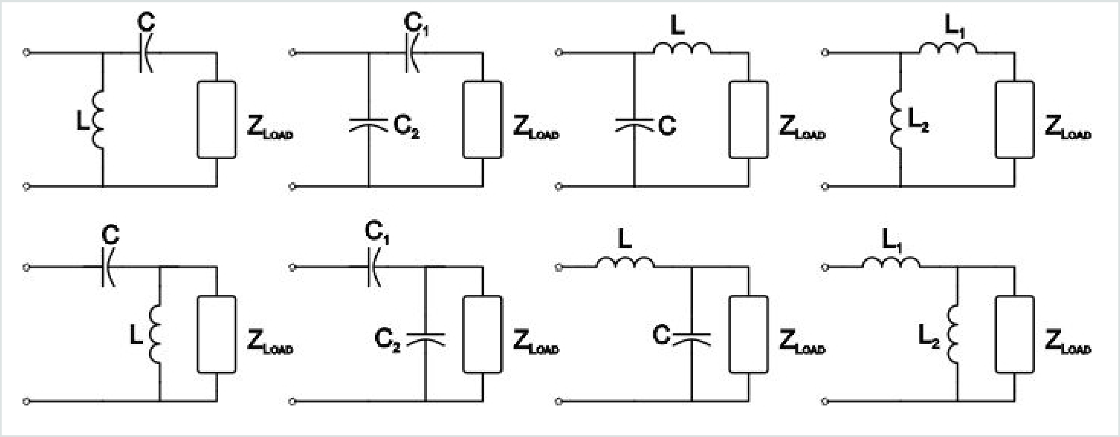
The design of matching networks is not particularly simple, and it's not something that we will discuss thoroughly in a textbook such as this one. Nonetheless, we can consider some of the basic principles, and we'll also take a look at a fairly straightforward example. Here are some salient points to keep in mind:

- A matching network is connected between a source and a load, and its circuitry is usually designed such that it transfers almost all power to the load while presenting an input impedance that is equal to the complex conjugate of the source's output impedance. Alternatively, you can think of a matching network as transforming the output impedance of the source such that it is equal to the complex conjugate of the load impedance.
    - (In real-life circuits the source impedance often has no imaginary part, and thus we don't need to always refer to the complex conjugate. We can simply say that the load impedance must equal the source impedance, because the complex conjugate isn't relevant when the impedance is purely real.)
- Typical matching networks (referred to as "lossless" networks) use only reactive components, i.e., components that store energy rather than *dissipate* energy. This characteristic follows naturally from the purpose of a matching network, namely, to enable maximum power transfer from source to load. If the matching network contained components that dissipate energy, it would consume some of the power that we are trying to deliver to the load. Thus, matching networks use capacitors and inductors, and not resistors.
- It is difficult to design a wideband matching network. This is not surprising when we remember that the matching network is composed of reactive components: the impedance of inductors and capacitors is dependent on frequency; thus, changing the frequency of the signals passing through the matching network can cause it to be less effective.

The L Network

- The most straightforward matching-network topology is called the L network. This

- refers to eight different L-shaped circuits composed of two capacitors, two inductors, or one capacitor and one inductor. The following diagram shows the eight L-network configurations:

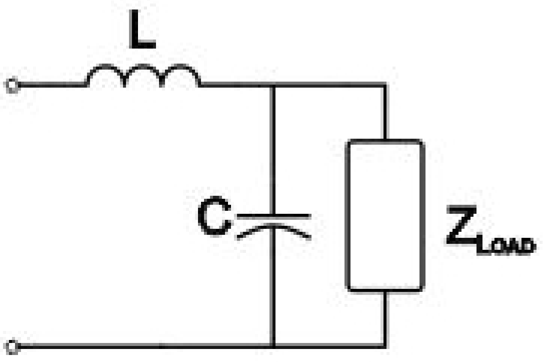


The L network is simple and effective, but it is not suitable for wideband applications. We also have to keep in mind that inductors and capacitors exhibit seriously nonideal behavior at high frequencies (as discussed in Page 4 of Chapter 1), and thus the behavior of the L network will be less predictable as frequencies climb into the gigahertz range.

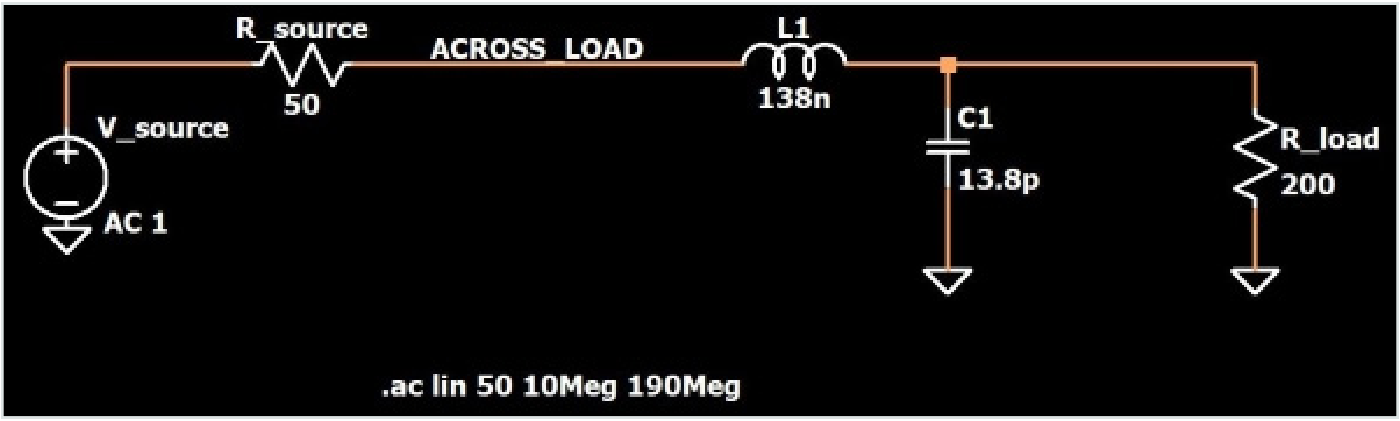
It is certainly valuable to understand the concepts involved in manually calculating matching-network values based on the source and load impedances, though this is more of an academic or intellectual exercise in an age when calculator tools can readily accomplish this task. We won't go through a calculation example here, but we will use a simulation to explore the effects of a matching network.

An Example

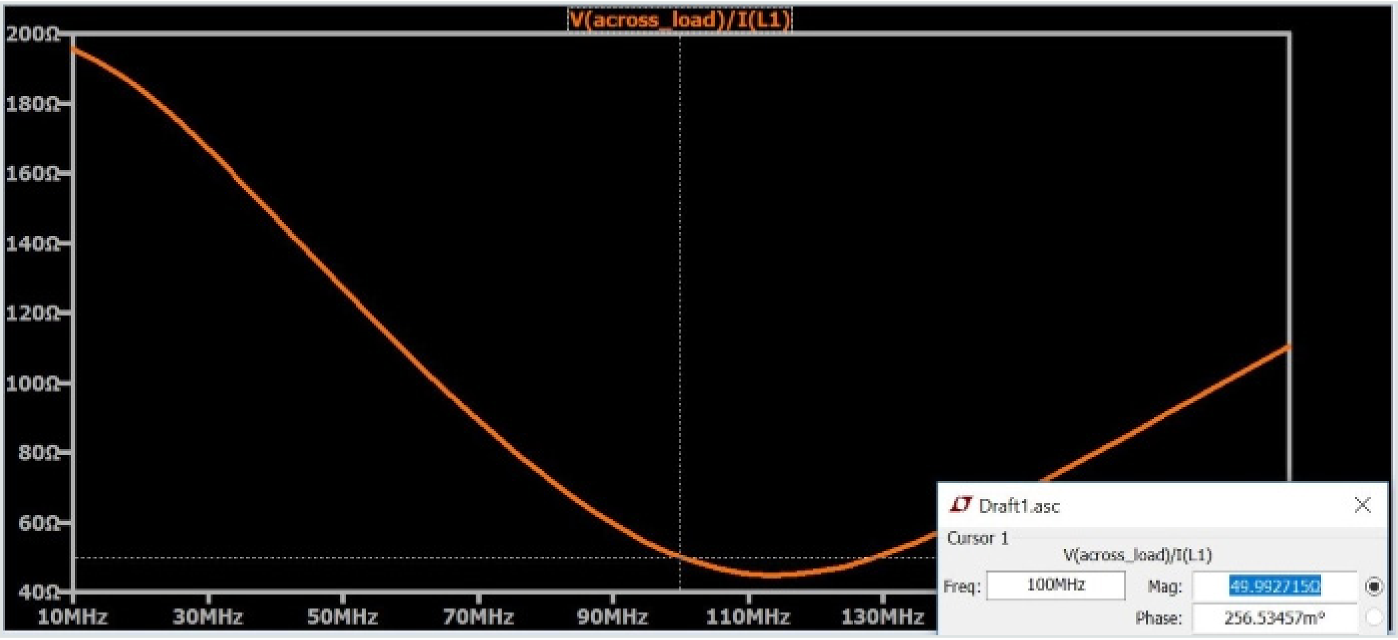
Let's say that we have a source impedance of 50 Ω and an antenna impedance of 200 Ω, and we are operating at 100 MHz. We'll use an L network consisting of an inductor followed by a capacitor:



AAC's L-network design tool gives the following values for the inductor and capacitor: 138 nH and 13.8 pF. This means that our impedance-matched circuit looks like this:



To evaluate the efficacy of the matching network, we can run a simulation and then plot the voltage across the load divided by the current flowing into the load, which is equal to the input impedance. (In this case the current flowing into the load is the current through inductor L1.) An AC analysis is particularly helpful because we can see how the effect of the matching network changes with frequency. The following plot is for a simulation with a frequency range of 10 MHz to 190 MHz (i.e., 90 MHz above and below the frequency for which the matching network was designed). Here are the results:



V(across_load)/I(L1)

Freq: 100MHz Mag: 49.9927150 ●

Phase: 256.53457m° ○

As you can see, at 100 MHz the load is very closely matched to the 50 Ω source impedance, despite the fact that the original load has an impedance of 200 Ω. However, we said above that the L network is not a wideband topology, and the simulation certainly confirms this: the input impedance changes rapidly as the signal frequency moves away from 100 MHz.

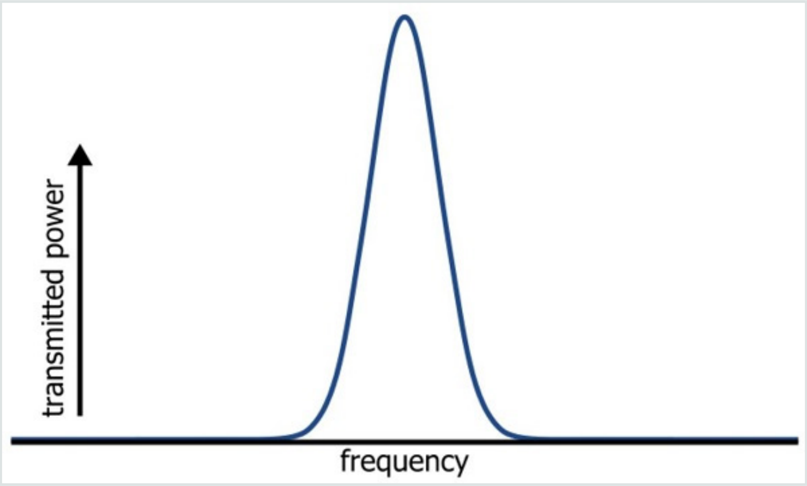
Summary

- A matching network, also called an impedance transformer, is used to create matched impedance between a source and a load (for example, between a power amplifier and an antenna).
- Lossless matching networks consist of reactive components only; resistive components are avoided because they would dissipate power, whereas the matching network is intended to facilitate the *transfer* of power from source to load.
- A straightforward, narrowband matching-network topology is the L network. It consists of two reactive components.
- Calculator tools can be used to quickly design a matching network based on the source impedance, load impedance, and signal frequency.

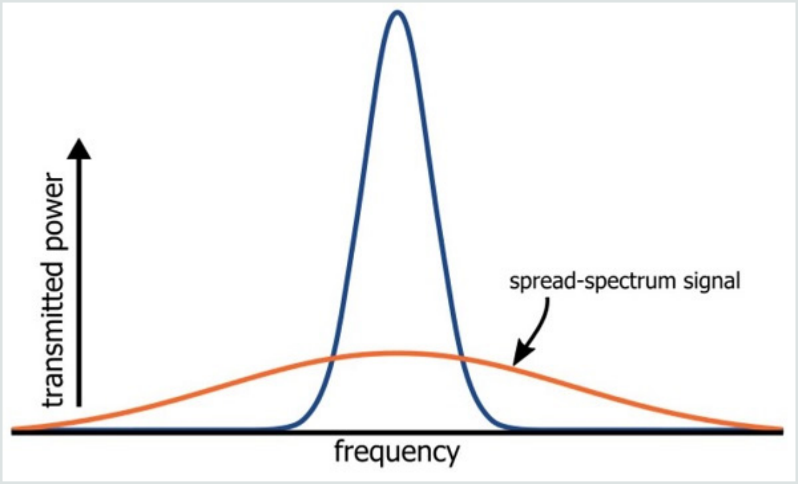
Understanding Spread-Spectrum RF Communication

Learn about an interesting technique that can make wireless communication more robust and reliable.

Throughout this book, we have visualized RF signals as both waveforms in the time-domain and distinct shapes in the frequency domain. These shapes are often rather tall and narrow, indicating that a large amount of transmitted energy is concentrated in a relatively small range of frequencies:



It turns out that we can also produce a very different kind of spectrum, namely, one that is wider and shorter. In other words, we can change an RF circuit such that it generates a signal whose transmitted power is spread out over a wider range of frequencies. These signals are appropriately described as "spread spectrum":



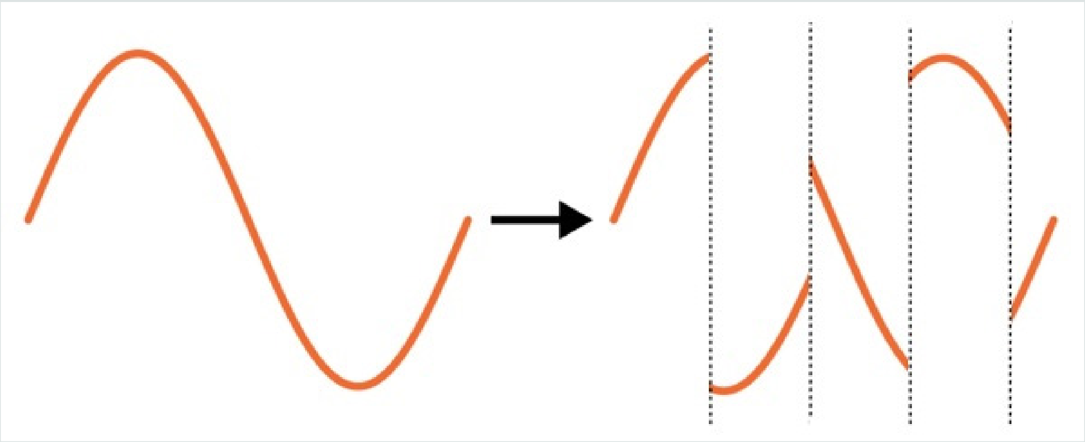
It's important to understand that the total transmitted power is not reduced. What changes is the *peak* power, because the transmitted power is distributed over a wider frequency band.

How to Spread a Spectrum

Changing a standard RF signal to a spread-spectrum signal is a process of exchanging peak power for bandwidth. However, we don't need to take any specific action to decrease the peak power: If everything else in the circuit remains the same, the overall transmitted power will also remain the same. All we need to do is increase the bandwidth of the signal, and the natural result will be the distribution of the available RF energy into a wider, shorter spectrum like the one shown above.

We know from previous pages that the bandwidth of an RF waveform corresponds to the highest frequencies present in the baseband signal. In the amplitude modulation page, we saw how the positive and negative frequencies of the baseband spectrum are shifted upward to form a symmetrical spectrum centered on the carrier frequency. Thus, if we use higher frequencies in the baseband signal, the bandwidth of the modulated signal will increase. We can spread a spectrum, then, by incorporating higher frequencies into the baseband signal. But how do we accomplish this without changing the baseband information?

The solution is something called a spreading sequence, also known as a pseudo-noise (PN) code or a pseudo-random-noise (PRN) code. This is a digital sequence that is intended to resemble a random succession of ones and zeros. The baseband signal is multiplied by the spreading sequence, except that a logic zero is treated like a negative one, such that the waveform is unchanged during the logic-high periods and inverted during the logic-low periods. The following diagram demonstrates this process:



As you can see, the frequency of the PN code (referred to as the "chip rate," because each pulse is called a "chip") is higher than the frequency of the baseband signal. We can intuitively recognize that chopping up the baseband signal in this way will introduce higher-frequency energy, and actually the factor by which the spectrum is spread is equal to the ratio of the chip rate to the baseband data rate.

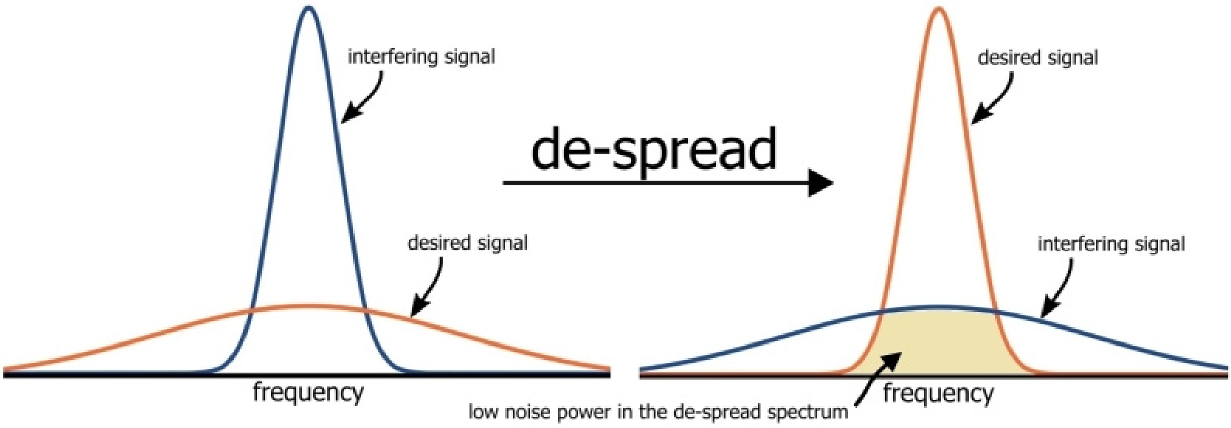
How to De-Spread a Spectrum

We have now increased the bandwidth of the baseband signal and, consequently, that of the transmitted signal, but the information appears to be seriously altered. How do we recover the data that was originally encoded in the baseband waveform? Actually, it's quite simple (at least in theory): all we need to do is multiply the received waveform by the same PN code. The sections that the transmitter multiplied by one will be multiplied by one again (i.e., they remain unchanged), and the sections that were inverted will be inverted again (i.e., returned to their original state).

Why Spread a Spectrum?

The procedure described thus far may seem like unnecessary complexity, but in certain situations it is well worth the effort. The fundamental benefit might be aptly described as "selectivity": spread-spectrum communication gives the receiver a greater ability to select the desired signal from among the various other signals that might be present in the relevant frequency band.

This selectivity results from an interesting effect of multiplying the received signals by the PN code: This receiver multiplication will de-spread *only the desired signal,* or more specifically, only the signal that was originally multiplied by the same PN code. If the unwanted signal has a narrowband (i.e., non-spread) spectrum, the PN code will spread it. If the unwanted signal has a spread spectrum that was created with a different PN code, the inverted and non-inverted sections will not align with the ones and zeros, and thus it will not be restored to its original state.



The same concept applies to wireless systems in which multiple devices must share a limited range of available frequencies. Such systems can employ various methods of minimizing problems associated with interference, and spread-spectrum communication is one of them.

The same concept applies to wireless systems in which multiple devices must share a limited range of available frequencies. Such systems can employ various methods of minimizing problems associated with interference, and spread-spectrum communication is one of them. Different devices can share the same band and their spectra can overlap; the receiver selects the desired signal by means of the PN code, which will de-spread only the desired signal.

Frequency Hopping

The spread-spectrum technique that we have discussed so far is called direct-sequence spread spectrum (DSSS). An alternative approach is to maintain the narrowband nature of the transmitted signals but to periodically change the carrier frequency. This is called frequency hopping, and it achieves a similar reduction in peak power if the transmissions are averaged over time. Like DSSS, it offers improved resistance to interference because an interfering signal will no longer corrupt the desired signal after the communicating devices have switched to a new carrier frequency.

Summary

- A spread-spectrum signal can be created by multiplying the existing baseband signal by a spreading sequence, also known as a PN code.
- The original signal is recovered by multiplying the spread-spectrum signal by the same PN code.
- A spread-spectrum signal has lower peak power, equal total power, and wider bandwidth. In other words, the available transmission power is distributed over a wider range of frequencies.
- Spread-spectrum techniques make a system more robust against jamming and interference.
- Similar results can be obtained by periodically changing the signal’s carrier frequency; this approach is called frequency hopping.

ALL ABOUT CIRCUITS

Practical Guide to Radio-Frequency Analysis and Design

© EETech Media, LLC. All rights reserved