Chapter 2.1 Electrons, fields, and radiation

Information is transmitted through the Internet as electromagnetic energy: electrical current in wires, radio waves in space, or flashes of light in optical fibers. The ability to quickly move large amounts of data over long distances is an essential characteristic of the Internet. This requires high transmission speeds, and only electromagnetic signals, which propagate at velocities near 3 x 10^8 m/s (186,000 miles/sec), can do this.

Rate of information transfer

The rate of information transfer is measured in bits per second. In a simple code using an electrical signal, 1.0 volts might signify a bit in the "1" state, while 0.0 volts means it is in the "0" state. After a set time (the clock period or one clock tick), the next bit is coded in the same way. Thus the bit series: 101010... would be represented as a series of one volt pulses lasting one clock tick separated by one clock tick when the voltage was zero. If one clock tick is one microsecond, we can transmit one million bits a second (1 Mbps). If the code used four voltage levels, say 0, 1, 2, or 3 volts, then the rate of information transfer would be two bits per clock tick. The rate of information transfer is limited by the speed at which the signal can be changed and the ability to measure the intensity of the pulses.

Latency

However, the rate of information transfer is very different from the physical velocity of information which is calculated from the number of milliseconds it takes a bit of information to travel, for example, from San Francisco to New York. The time lag between the transmission and the receipt of information is called the latency of the system. For a 1 km (0.6 miles) long wire, the latency is slightly more than three microseconds, since the signal will travel almost at the speed of light. Thus if we used the wire to send a message from San Francisco to New York, the latency would be about 15 milliseconds. Of course the signal would never be able to actually make that journey. Relay amplifiers would be needed along the route to make up for all the losses in the wire, and each amplifier would add some time to the total latency. On the Internet, latencies between these two cities can be several hundred milliseconds. Most of that time is spent at the branch points, or nodes of the network, where specialized computers called routers decide the next path the message will take.

Why would latency ever be important; does it really matter if it takes 100 or 1000 milliseconds to send an e-mail from San Francisco to New York? As we will see later, some protocols involve sending a segment of a long message to the sender and waiting for a response that proves that the correct (uncorrupted) segment was in fact received. Then the next segment is transmitted. Thus the time necessary to send the entire message is indeed dependent on the latency.

Electromagnetic fields and radiation

Electricity and magnetism are closely linked; an electrical current in a conductor produces a magnetic field and a changing magnetic field produces an electrical current in a conductor. A changing electromagnetic field moves through space as a series of waves. Ocean waves are periodic changes in the height of the water surface, sound waves are periodic changes in air pressure, and electromagnetic waves are periodic variations in electric and magnetic fields. Electromagnetic waves do not penetrate conducting material, e.g. copper, but they cause free electrons at the surface to oscillate, generating alternating currents. Electromagnetic waves easily move through insulators, e.g. glass, and as they do they cause the bound electrons to oscillate with the wave.

The concept of a wave implies a medium, for it is the periodic changes in the medium that constitutes the wave. But electromagnetic waves do not depend on alternating currents or vibrating electrons for their existence, since they also travel quite well through a complete vacuum. In fact, the velocity of electromagnetic radiation is the highest in a vacuum.[1]

Pipes that carry the information

A simple wire

The simplest way to transmit electrical signals is to use a metal wire. For example, let's use a copper wire: 1 mm (0.04 in) diameter covered with a layer of non-conducting plastic for insulation. Two paths are needed for any electrical circuit; current moves out along one path and returns on the other. Thus we will use a pair of these wires. At the far end we attach a device across the pair that can measure the voltage between the two wires. Now, at our end we apply a voltage between the pair and see what happens.

Metals are typically hard and dense, and they conduct heat and electricity well. These properties result from the atomic structure of the tiny crystals that make up the wire. Some of the electrons originally associated with individual copper nuclei are delocalized in the crystal, and thus are shared among neighboring atoms. These electrons will move when we apply the voltage between the two wires. Since electrons

[1] What is the "medium" that carries the electromagnetic waves? It is the electromagnetic field. As Richard Feynman wrote [The Feynman Lectures on Physics, vol II, 1964, pg 1-9]:

> "The best way is to use the abstract field idea. That it is abstract is unfortunate, but necessary. The attempts to try to represent the electric field as the motion of some kind of gear wheels, or in terms of lines, or of stresses in some kind of material have used up more effort of physicists than it would have taken simply to get the right answers about electrodynamics."

It is possible to gain a deeper insight into the electromagnetic field using concepts from the field of Quantum Electrodynamics. However, for the purposes of this book exploration in this direction would be a digression.

have a negative charge, and like charges repel, they will move away in the wire with the negative (relative to the other wire) voltage and toward us in the other wire. If we wait for a while, a second will be much more than enough, the device at the far end will have detected the voltage, and a constant current (flow of electrons) will be moving down and back through the wires. The amount of current (measured in amperes) is inversely proportional to the total resistance (measured in ohms) of the circuit; one volt pushes one ampere through a circuit having a resistance of one ohm. The flow of electrons through a wire is often compared to the flow of water through a pipe. The flow rate (amperes) multiplied by the pressure (voltage) is the energy transmitted per unit time, or power (watts).

However, we have set up this circuit to transmit information, not power. We want to transmit as much information as possible per unit time and will do this by changing the voltage very rapidly, for example on and off for different lengths of time. This on-off pattern will carry the information in some sort of code that we have yet to design. We will assume the detector at the far end is ideal, e.g. it can measure changes in current with no time delay, which allows us to focus on the performance of the wire itself. To invent a rational code we need to know how a rapidly changing applied voltage, going from zero to maximum and back to zero in progressively smaller times, is transmitted down the wire. Of course by merely asking this question we are suggesting that a very short pulse may not be transmitted as well by the wires as a longer pulse, and thus there is a maximum rate at we should switch the voltage on and off.

The first nanosecond

The progress of the pulse along the negative wire is illustrated in Figure emField. The voltage applied at the ends of the wires (only the negative end is shown in the Figure) generates an electric field represented by the red arrows. By convention the arrows go from negative to positive voltage, thus the electrons, which are negatively charged, move in the opposite direction as the field arrows. Immediately after application of the voltage electrons start to move away from the end, and this current (a current is moving electrons) generates an increasing, circular magnetic field that radiates out from the wire at the speed of light. The electric field is along the axis of the wire and thus perpendicular to the magnetic field. These expanding fields move out along the wire at close to the speed of light. The electrons at the surface of the wire are the first to move, but after a few microseconds electron flow is uniform from center to surface.

Figure emField

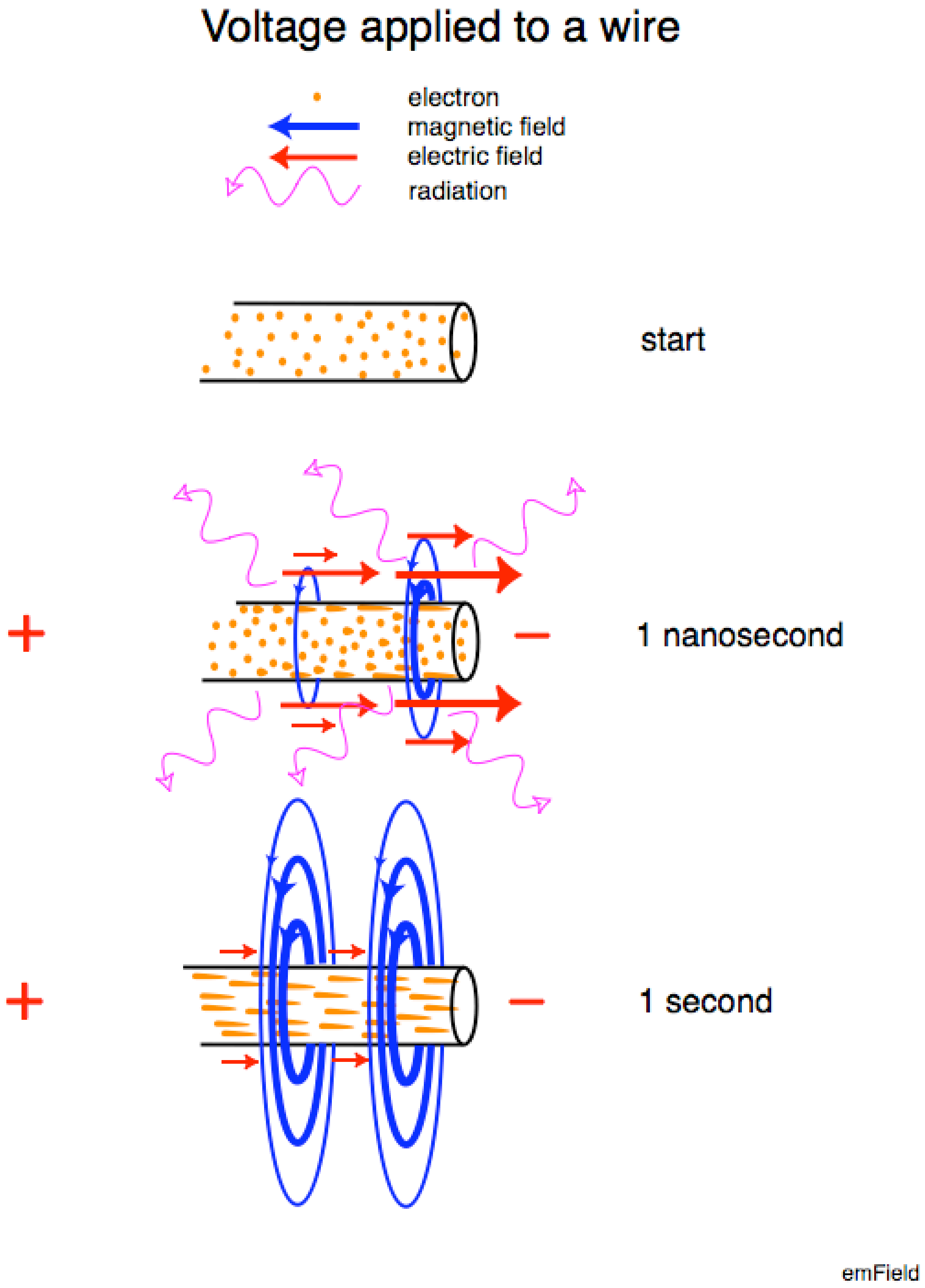


Figure emField. A negative voltage is applied to the right end of a wire, the positive voltage is applied to the left end, some distance away. As the free electrons start to move toward the positive end of the wire a magnetic field is established and an electromagnetic field radiates out from the wire. When the electric current has reached a steady state, there is an electric and a magnetic field, but since they are constant, no radiation is produced.

The electromagnetic field moves along the wire faster than the electrons themselves, and it is the speed of this field that determines the time required for the instrument at the far end of the twisted pair to detect the pulse. The electrons in the wire are similar to a long column of dominos, each domino upright on edge. After the first domino is pushed and falls against the next, a disturbance propagates along the entire column as all fall down. Each domino moves only a short distance, but the disturbance moves down the entire column. However, each domino must hit the next, and thus the disturbance moves at the average speed of the tops of the dominos. In the wire, the disturbance is not propagated by each electron hitting the next; rather, the electrons are coupled by the electromagnetic field. Thus it is the speed at which this field propagates that defines the speed of the electrical pulse. The speed is a function of the diameter of the wire and the nature of the insulation around it, but is only slightly less than the speed of light.

Of course no electromagnetic energy radiates from the wire before the pulse starts, when there is neither an electric or magnetic field. However, there is also no radiation after one second when the electric current, and thus the magnetic field, are both maximum. It is only when the electric and magnetic fields are changing, in the middle panel of Figure emField, that there is radiation. It is the change with time of the electric and magnetic fields that causes radiation. To really understand the propagation of the pulse along the wire we need to know more about electromagnetic fields, and to do this we need differential equations.

Differential equations

Most of the basic laws of physics are expressed as differential equations, which describe relationships between changes in the values of variables, not just relationships between their values, as do "normal" equations. Differential equations contain one or more derivatives, which are the rates that one variable changes with respect to another variable. In many cases the derivative of a variable is the rate of change with time. Velocity is the rate of change of position with time; thus velocity is a derivative. The derivative of a derivative is called a second derivative. Thus, acceleration, which is the rate of change with time of the velocity, is a second derivative. As these examples illustrate, derivatives are parts of everyday life, even though we may not call them derivatives.

The description of differential calculus by Gottfried Leibniz in 1684 allowed mathematicians and scientists to create and solve differential equations, and was thus a great intellectual achievement with huge practical impact. Isaac Newton published the same methods, but with different notation, in 1687. The earlier publication of clearly precursor methods allowed Newton to win a bitter legal dispute on his priority in this invention, at least in the Royal Society of England.

One of the most fundamental laws of physics is $F = m\ a$, force equals the mass times the acceleration[2]. If the acceleration is not constant, this is a differential equation[3]. A main motivation for the development of calculus by Newton, and its first dramatic use, was to use this equation to calculate the shapes of the orbits of the planets around the sun.

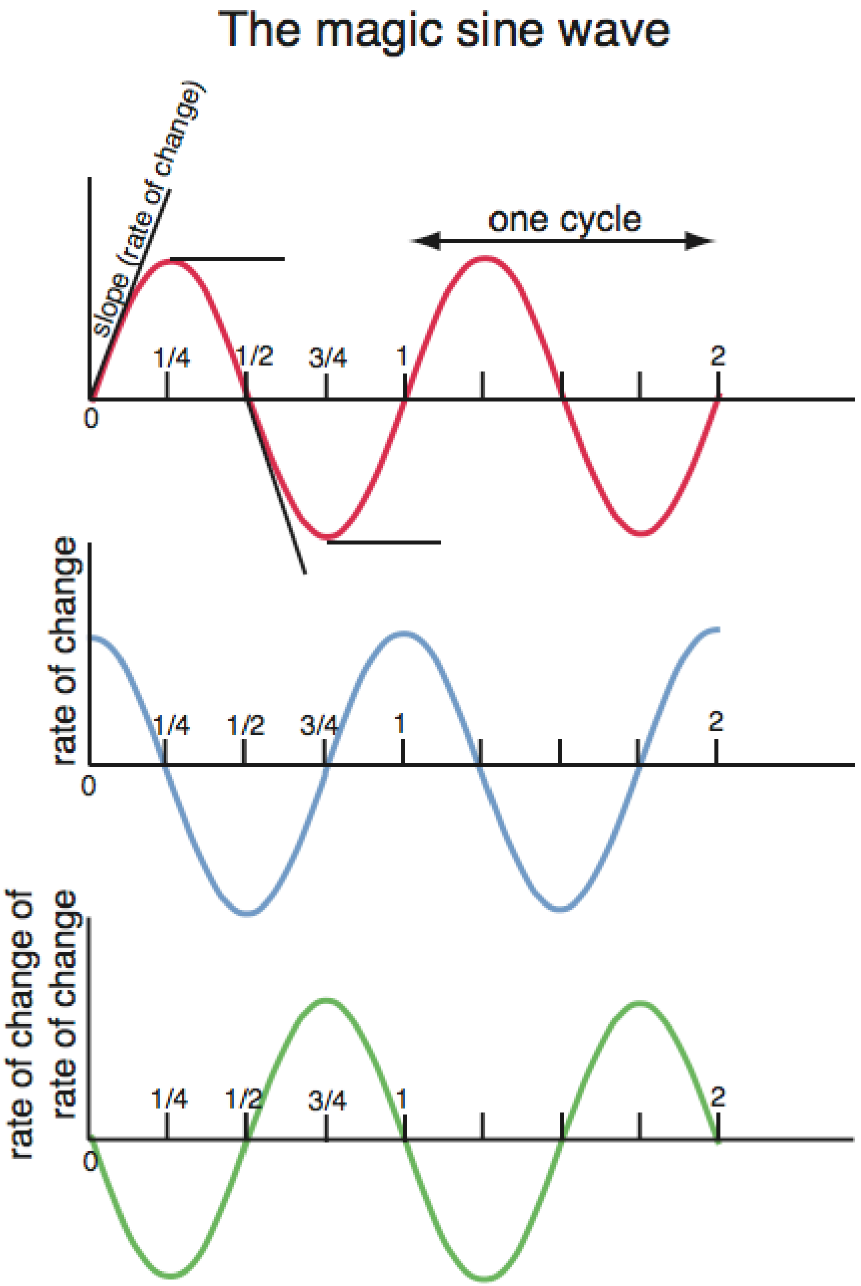
The magical sine wave

Now that we are familiar with the language of derivatives, we can more easily talk about the shapes of curves and appreciate the dramatic properties of the sine curve, the red wave at the top of Figure sin. It probably looks familiar, since it is the shape of a wave moving along the surface of a still pond, the shape of the string of a musical instrument that is producing a fundamental tone, the displacement versus time of the swinging pendulum of a clock.

[2] It is common for mathematicians and scientists to indicate multiplication of two quantities by just writing the two quantities next to each other (separated by a space), omitting the “x” that often indicates multiplication. This has the advantage of freeing “x” for use as a variable, but can cause confusion if variables are assigned symbols with two or more letters.

[3] When m is constant but a varies, the equation is usually written $F = m\ d^2\ x/dt^2$, where x is the position and the notation $d^2\ /dt^2$ indicates the change of (the change of x with respect to time) with respect to time.

Figure sin


The magic sine wave
slope (rate of change)
one cycle
0
1/4
1/2
3/4
1
2
rate of change
0
1/4
1/2
3/4
1
2
rate of change of rate of change
0
1/4
1/2
3/4
1
2

sin

Figure sin. The three curves are a sine wave, it's rate of change, and the rate of change of it’s rate of change. The rate of change of a curve is the slope of a line tangent to the curve. The slopes of the first curve at three times are indicated by the short black lines at the start, one quarter, one half, and three quarters of a cycle. The second and third curves turn out to be sine waves also, that’s what makes the sine so magical.

The magic is in the way the sine curve changes as we move along it; these changes are its derivatives. Suppose the vertical distance on the graph is the elevation or depression of the surface of the water at one location relative to its undisturbed level, and the horizontal distance is time. The curve thus described the vertical movement of the surface at one location. The slope of the curve at any point is the rate at which the surface is rising or falling, its vertical velocity. At the start of the curve, the left end, the slope is greatest; the surface is moving rapidly upward. After 1/4 of a cycle, the height of the surface is at a maximum, but the slope (and thus its velocity) is zero. After another 1/4 of a cycle, the surface has fallen to its original level, and the slope (its velocity) has a large negative value. At 3/4 of a cycle, the level is at its lowest, but now the slope is again zero. And then the curve returns to its original value and slope, and the cycle repeats over and over again.

The next curve down, the blue line in Figure sin, shows the rate of change of the red sine wave, just illustrating the narrative of the previous paragraph. But now we see the magic. This curve is also a sine curve, but just shifted 1/4 of a cycle to the left! The green curve at the bottom of the Figure is the rate of change of the blue curve, i.e. the rate of change of the rate of change or acceleration. It is also a sine curve, but again shifted 1/4 of a cycle to the left. This means it is shifted 1/2 of a cycle to the left of the top curve, or just the negative of the top curve, i.e. the original sine curve times –1. The bottom line is that the second derivative of a sine wave is equal to the negative of the curve itself.

The sine wave is the only curve that has a second derivative equal to its negative value. Thus, if you start out to find a curve that has this property, you will end up with a sine wave. This means that this property can be used as a definition of the sine wave.

Now you can guess why the sine wave looks so familiar. The differential equations that define the motion and shape of a water wave, the vibration of a stretched string, and the motion of a pendulum all require that the acceleration, or second derivative, be proportional to the negative of the displacement from the equilibrium position. Thus a sine wave is a solution of all these equations and an ubiquitous expression of nature.

Another way to get the sine wave

Many people first hear the word sine when they study trigonometry. Then it is written “sin”, and is a function of an angle in a right triangle: the ratio between the length of the side opposite the angle and the length of the hypotenuse. When you make a graph of this sin versus angle, you get the sine wave in Figure sin. Two apparently different definitions give the same curve; more mathematical magic.

The sine wave is a solution to Maxwell's equations

James Clerk Maxwell published the comprehensive mathematical theory of electromagnetic radiation in 1873, unifying earlier work of Faraday and others on the properties of electric and magnetic fields. As you can guess by now, his theory was expressed as a set of differential equations, and our now familiar sine wave happened to be a solution of those equations. Maxwell’s equations make many predictions about the properties of electromagnetic waves.

Perhaps the most dramatic prediction was that wires carrying a changing electric current would radiate energy into space. The intensity of the radiation depends on the amount of current and the frequency of the wave, i.e. the number of cycles in one second. The greater the frequency, the more electromagnetic power will be radiated out from the wire per unit length. The velocity of the electromagnetic radiation in space is equal to the speed of light; in fact light is just electromagnetic radiation with a high frequency. The inverse of radiating energy into space is capturing some of the power of electromagnetic radiation using a wire. Maxwell’s equations have a great deal of symmetry, and radiation and adsorption of electromagnetic radiation by a wire is a dramatic one.

But we want to transmit square waves

The behavior of sine waves may be interesting, but if we are going to build a computer network we probably want to transmit a square wave down the wire, e.g. apply a positive voltage for one clock cycle if the bit is 1, and apply a negative voltage for one clock tick if the bit is 0, and changing between the two voltages as quickly as possible. The bit sequence 10101010… would then be a sequence of square pulses, or a square wave. We know how to use Maxwell's equations to predict the behavior of sine waves, but what do they have to say about square waves?

This turns out to be a common problem. You know a set of curves is the solution to a differential equation, but you are interested in a curve that is not one of those curves. Jean Fourier discovered a general method to solve this kind of problem, and in 1822 published its use to describe the diffusion of heat. If the temperature along a metal bar varies as a sine wave, it is easy to solve the differential equation for heat transport (yes, another differential equation) and thus predict how the temperature will change with time. But it is unlikely you would want to predict the change in temperature in a metal bar in which the temperature distribution was a sine curve. How would you even get in a situation where the temperature along the bar was a sine curve? It is more likely that, for example, you would want to know the temperature versus time along a bar that starts out with one-half of the bar at one temperature and the other half at another temperature. At least that would be easy to create; pick up a bar that is at one temperature and another bar that has a different temperature and quickly put the two ends together. In fact, this initial temperature distribution is just the square wave that we are interested in!

Fourier relied on a fundamental property of all the differential equations that we have been talking about: if curve A and curve B are both solutions, then the sum of the curves, A + B, is also a solution. Equations with solutions that can be added together and still be solutions are said to be linear. Linearity may seem to be rather dull, but Fourier turned it into dynamite when he showed how you could create any given curve by adding together a number of sine waves. Since Fourier picked sine curves which were solutions of the differential equation, the sum of these, the curve he was actually interested in, was also a solution.

Figure Fourier

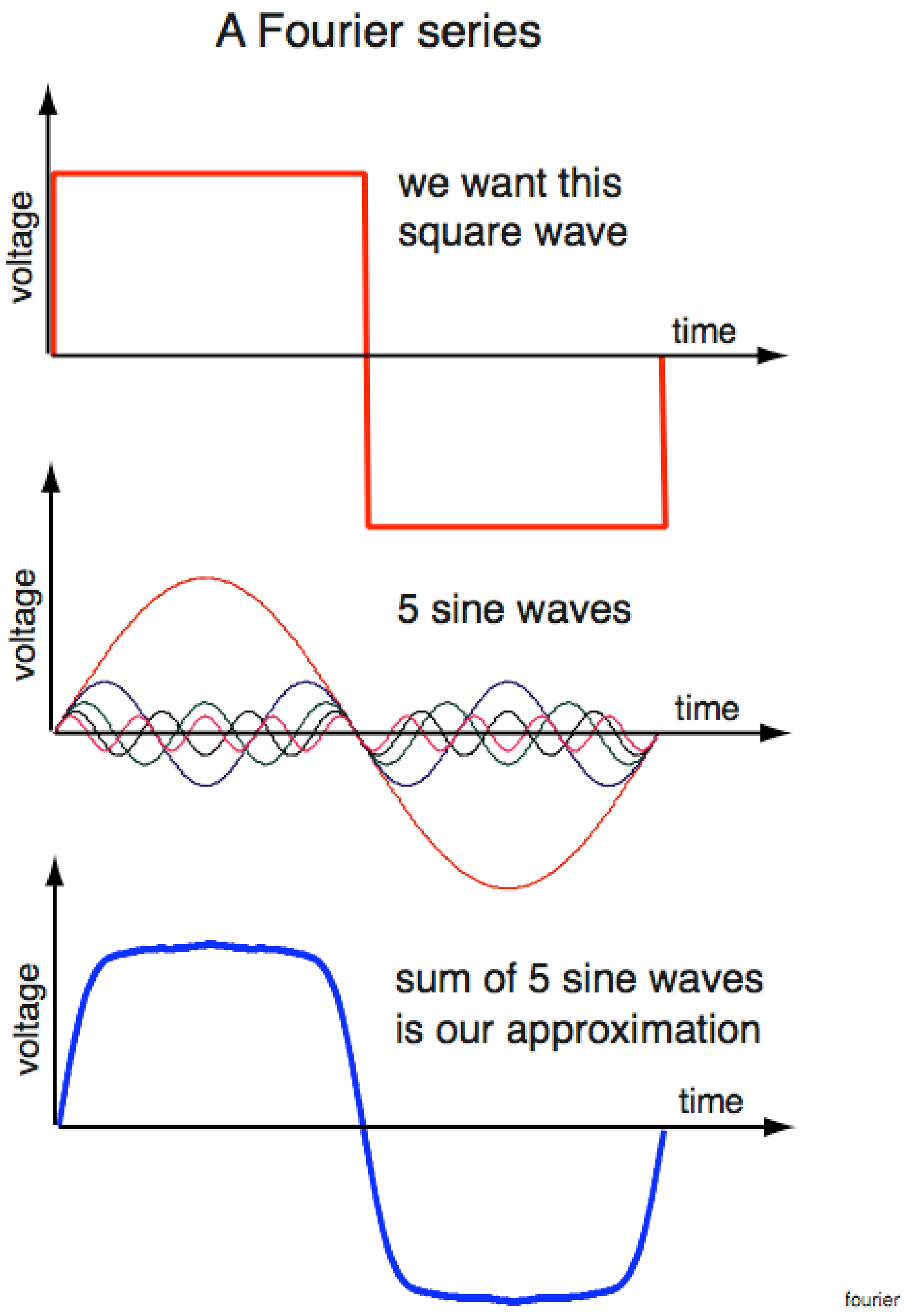


Figure Fourier. A good approximation to a square wave can be generated by adding together sine waves of increasing frequency.

Figure Fourier shows how a sum of sine waves of increasing frequency become progressively closer to a square wave[4]. The sine wave with the lowest frequency gives the outline of the square wave, while sine waves with higher frequencies add together to generate the steep beginning, flat middle, and sharp end of the square wave. As we use more sine waves the approximation gets better and better.

The bottom line is that if we know how sine waves of different frequencies travel down the wire, we know how a square wave moves down the wire. We know that the higher frequency sine waves radiate more energy into space. We also know that the higher frequency waves often move at slower velocities or are absorbed to a greater extent in some transmission systems. The progressive lose of the higher frequencies cause the initially sharp beginnings and ends of the square waves to become rounded. If the higher frequencies travel at a lower velocity than the lower frequencies, the square wave is smeared out. The net effect is to make it more difficult for the instrument at the receiving end of the wires to distinguish one pulse (bit) from another and to determine if the bit is a "0" or a "1".

The twisted pair

The electromagnetic field that radiates from the wire can cause trouble. The field represents a loss of energy, and thus the pulse becomes weaker as it travels down the wire. The field can be picked up by neighboring wires, generating unwanted signals in them. Twisting the two wires around each other decreases the intensity of the radiating field. To an observer some distance from the pair the direction of current flow seems to alternate along the pair, since first one wire is visible then the other, and the current is moving in opposite directions in the wires. Thus the fields radiating from the two wire partially cancel each other.

Twisted pairs of wires are commonly used to transmit information at rates between 10^5 and 10^8 bits per second (0.1 to 100 Mbps) over distances between 1 km (3300 feet) and 1 m (3.3 feet). Note the inverse relation between distance and information rate, which is the result of blurring of the pulses as the signal travels down the twisted pair of wires.

[4] Note that the size (or amplitude) of each sine wave of increasing frequency is not the same; they must be calculated, and Fourier described how to do this. Finding the best amplitudes is not really that difficult, but we don’t need to do it here to make the point that it can be done. A more pesky point is that the sums of the sine waves do not exactly converge to the values of the square wave at every value of time. The sharp corners of our idealized square wave have infinite derivatives, and this will cause the sum of the sine waves to have little spikes at the corners. However, if we round the corners of the square wave ever so slightly this trouble will disappear.

Co-ax; better than the twist

The next level of sophistication is the coaxial cable. One wire is replaced by a hollow conducting cylinder, and the other wire runs down the middle, the cylinder's axis. The cylinder can be a metal pipe filled with air or some other gas. However, for low power an adequate and far cheaper construction uses a wire braid as the outer cylinder which covers a flexible plastic insulation and a wire in the center. This is a great improvement over the twisted pair, since very little radiation is lost. However, it is more expensive than the twisted pair, and the connectors are now small pieces of plumbing. Some energy is still lost as heat due to the resistance of the wire and motion of the molecules in the insulation. Cables with a larger diameter are generally better, but more expensive. Cables with different diameters have different impedances, and the impedance of any coaxial cable is likely to be different from a twisted pair of wires. When connecting cables of different impedances you should use a matching device, a transformer. Otherwise signals moving down the cables will generate reflections at the connection (as predicted by Maxwell's equations), and these reflections can confuse the devices at the two ends of the link, since they will literally not know what is coming or going. The transformers actually carry out the same basic function as the heavy metal devices on electric power poles, but fortunately, due to the much higher frequency and lower power on the communication line are modest devices not much bigger that the coaxial cable connectors themselves.

Wireless communication

The radiation of electromagnetic energy from a wire is a pain if you are trying to communicate by that wire but a blessing if you want to communicate without one. In 1887, only a few years after Maxwell's prediction, Heinrich Hertz demonstrated that alternating currents in a wire produce electromagnetic waves that radiate out into space. Hertz was a true academic, and didn't explore commercial applications, but eight years later Guglielmo Marconi developed a radio telegraph and in 1901 transmitted messages across the Atlantic Ocean. The field of wireless communication exploded in the 1980s and 1990s, due to the demand for mobile communication and the availability of cheap microchips that could operate at very high frequencies and perform complex signal processing.

The carrier signal

There is a fundamental difference between transmitting bits of information on a wire as voltage pulses and through space (or glass fiber) as electromagnetic radiation. On a wire that can carry current, the change from a bit of value 1 to a bit of value 0 can be encoded as a change in voltage, e.g. from 1 volt to 0 volts. Thus the receiver need only measure voltage, since the changes in voltage are the data. But what corresponds to a change in voltage in a microwave signal? It could be a change in intensity of the signal, however the signal is not a voltage but instead is a rapidly changing electromagnetic field. If the microwave wavelength is 3 cm (1.2 inches) the frequency of the oscillating electromagnetic field is 10^{10} cycles per second, or 10 gigahertz. This signal is called the carrier, because it is changes in this signal that carry the information.

There are two motives for using a carrier signal that is a pure sine wave oscillation with a constant frequency. The first is the need to transmit and receive many different streams of information simultaneously. While some antennas are highly directional, most are not, and thus we are bathed in electromagnetic signals from multiple cell phones, television, radio and radar transmitters, etc. We can make sense of this cacophony because each signal oscillates at its own, unique frequency, and the receivers have circuits which block all carrier signals except the one with the correct frequency. Due to the communal nature of radio transmission, generation of signals of any significant power are strictly regulated by national and international agencies. A license to transmit limits power and specifies a single, or a narrow band of frequencies that can be used.

The second motive for a constant carrier frequency is rejection of noise. It is easy to design circuits that amplify very weak signals, however, there is a constant background of electromagnetic noise caused by man made devices and natural processes. This noise is picked up by the antenna and amplified along with the desired signal. A powerful amplifier does not solve the problem of a weak signal if the amplified noise is greater than the amplified signal. The solution for detecting weak signals is the fact that the noise is spread out over all frequencies. Thus, the total noise intensity is proportional to the band of frequencies that the receiver amplifies. If the signal that is to be received has a very constant frequency all other frequencies containing noise can discarded and the signal then easily detected[5].

Carrier modulation

There are three common methods of changing, or modulating, the carrier to transmit information. The first is amplitude modulation: the intensity of the carrier is changed to encode the message. If the message is a series of pulses that represent binary information, the carrier could be just turned on and off. This is the original technique used by Marconi to transmit Morse code. The second common method is frequency modulation: the frequency of the carrier is changed slightly to encode the message. This method is inherently more resistant to noise, which is why music sounds better on FM compared to AM receivers. A third method of modulation is to shift the phase of the carrier to encode the information.

Modulation broadens the frequency spectrum of the carrier

A sine wave has a single frequency. However, any change in the sine wave adds additional frequencies. A simple example would be to amplitude modulate a carrier with a sine wave of a much lower frequency.

[5] Some modern radio transmitters, e.g. cell phones, change frequencies rapidly within a defined band according to a predetermined code. This strategy permits more information to be transmitted within the band by multiple transmitters compared to the capacity if each transmitter was assigned a fixed frequency.

Figure Modulation

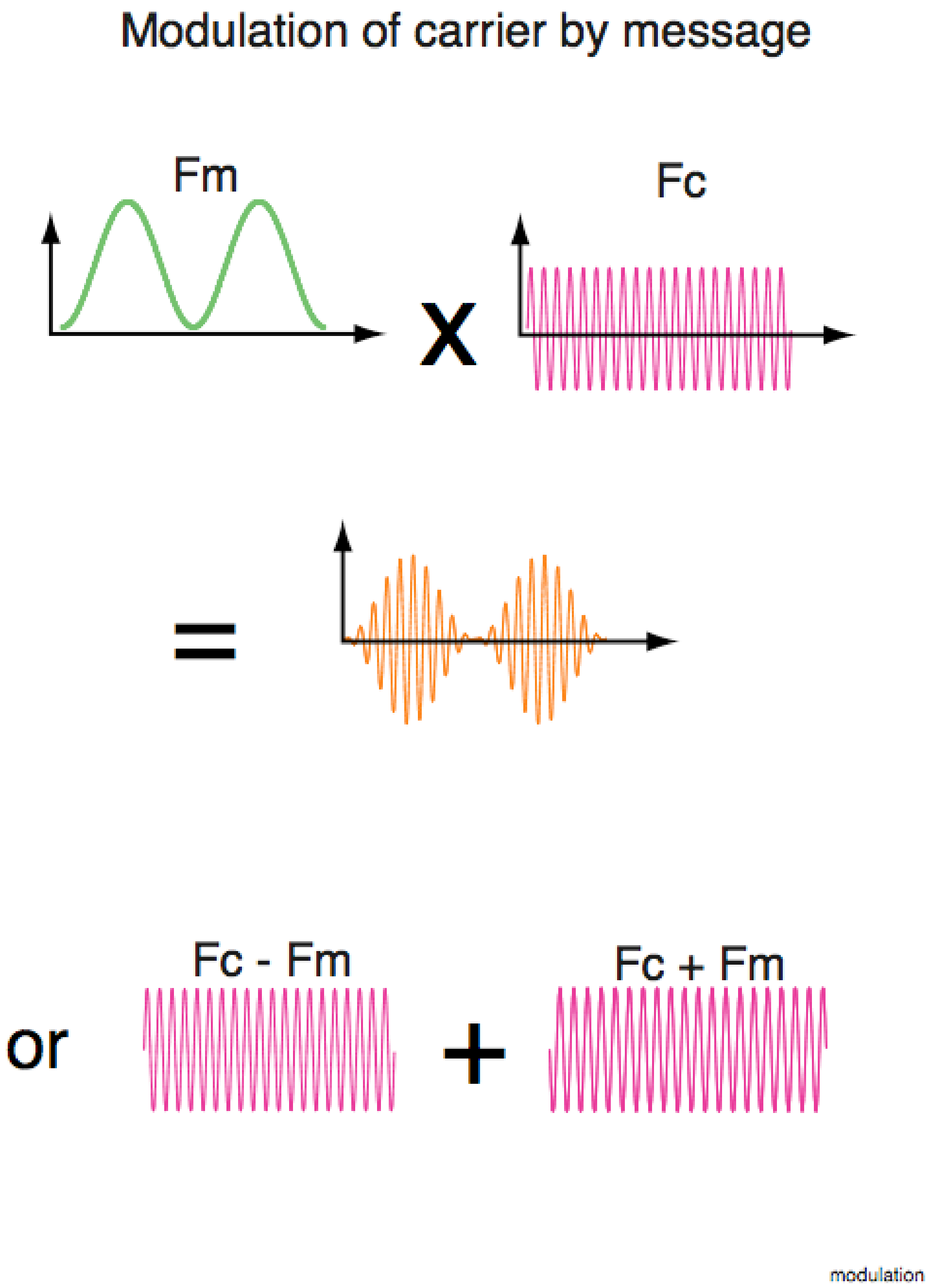


Figure modulation. The green wave, Fm, is used to modulate the red carrier wave, Fc. Modulating a carrier is equivalent to multiplying the value of the modulating signal with the value of the carrier, thus the X between the two signals. The result is orange carrier that changes in intensity with the frequency of the modulating signal. However, this orange signal is equivalent to the sum of two carrier signals with frequencies equal to the carrier signal plus and minus the modulation signal.

At the top of Figure modulation we see the message signal, a sine wave with frequency Fm, and the carrier sine wave, with a much higher frequency, Fc. To produce the transmitted signal we multiply the carrier by the message at each time point. The product, the modulated carrier, is the regular series of hourglass shaped oscillations in the third panel. This curve looks very much like a sine wave, but it isn't.

Here comes the shocker. This hour-glassed shaped curve, the product of two sine waves, is also the sum of two sine waves, one with a frequency equal to the carrier minus the message frequency and one with a frequency equal to the carrier plus the message frequency. How the sum of these two wave gives the hourglass shaped wave can be appreciated by adding the two waves together going from left to right, i.e. as time passes. At the beginning the waves cancel each other, giving a zero signal. However, as we proceed they become in phase and then add giving a wave twice as strong. As we move again to the right they become out of phase and cancel.

When a message modulates a carrier, the resulting signal contains the sums and the differences of all the frequencies present in the message and the carrier frequency. In general the message will contain many frequencies, and thus the modulated carrier will occupy a band of frequencies, centered on the original carrier frequency, with a width twice as wide as the highest frequency in the message. These new frequencies are called sidebands. and the upper and lower sidebands carry the same information. In a single sideband (SSB), system one sideband and the carrier is removed before transmission, but that's as far as you can go without loosing some of the information in the message.

Information rate, bandwidth, and noise

Now we can see why there is a fundamental relation between bandwidth and the rate at which information can be transmitted. If information is transmitted rapidly, the message signal must be changing rapidly. A rapidly changing signal contains higher frequencies. A carrier modulated by higher frequencies generates a wider band of frequencies and the receiver must accept this wider band to obtain the message.

Assuming the receiver is well designed, the limiting factor is noise. The amount of noise is proportional to the bandwidth the receiver must have to receive the signal. As the bandwidth of the receiver increases, the effective power of the transmitter must be increased so that the signal to noise ratio (SNR) remains constant. The effective power can be increased by increasing the power of the transmitter or increasing the directivity (the gain) of the transmitter or receiver antenna. It's obvious why increasing the directivity of the transmitter antenna will increase the effective power; the same power is just focused into a smaller zone so the power per unit area, which is all the receiver sees, is increased. On the other side of the transmission, increasing the directivity of the receiver antenna does not increase the power received, but since noise is arriving at all directions, but an antenna which accepts radiation from a smaller zone receives less noise, and thus the SNR is increased.

A shortwave radio experiment

These relationships can be illustrated by the following armchair experiment. We have a short wave receiver which can be tuned to different carrier frequencies, and we set it to 11 MHz. It is likely that we will hear only noise (if not we adjust the frequency slightly until that is all we hear). Our receiver has another dial that adjusts the range of frequencies that we can hear, the bandwidth. We first adjust it to 0.1 KHz, and hear a very small static signal. Then we increase it to 1 KHz and observe that the amount of noise increases 10 fold. At 10 KHz the noise again increases 10 fold. To a good approximation, the noise is uniformly distributed across the frequency spectrum, so the more frequencies we allow through our receiver, the more noise we hear.

We turn the bandwidth dial back to 0.1 KHz and adjust the carrier frequency dial to exactly 10.0000 MHz. Now we will hear time signals transmitted by the US National Institute of Standards and Technology (NIST). The time signals are a regular series of "clicks" transmitted at the beginning of each second. The 30th click is omitted, and toward the end of each minute the clicks stop. If we listen very carefully, we will hear a faint rumble during this pause, but not much else. Consistent with our first experiment with this receiver, the reception will be almost static free, and thus the time clicks will be very clear.

Now open up the bandwidth to 10 KHz. The static will increase dramatically, but if we have a good receiver and antenna we should still be able to distinguish the time clicks at most locations in the world at most times, but not always, everywhere. When the ticks stop at the end of the minutes we will now be entertained by a person, usually first a woman, then a man, who will tell us the time of day (in Universal Mean Time). Occasionally, all the time ticks for a minute will be preempted by another person who will give us the location of major storms or tell us about the status of radio navigation stations or Global Position Satellites.

We can now make some practical rules. If we have a weak communication link (high noise and low signal strength) we can decrease bandwidth and reduce noise, and then we will be able to detect slow time clicks, because they represent information transmission at a low data rate or frequency. We may also be able use a small transmitter to send a simple ID code up to a satellite 40,000 Km (25,000 miles) above us, if the ID code is sent slowly. This strategy is used by Emergency Beacon Floats carried by ships. If it takes several seconds, it's still fast enough to save our life. One of the slowest communication rates is used by an extremely low frequency NIST system broadcasting a carrier of 20 kHz which is modulated slowly to automatically synchronize clocks once every day.

If we have a better communication link (lower noise or higher signal strength) we will be able to increase the bandwidth and then we can transmit and receive voices, and even distinguish men from women. The human voice requires frequencies from about 300 - 3,000 Hz (the bandwidth of a typical telephone) to be clearly understood, although that bandwidth isn't exactly high fidelity. A ten fold better link allows us to transmit simple still images and we can surf the Internet. Ten to a hundred fold higher still and we can see full video.

The electromagnetic frequency spectrum

The range of frequencies of electromagnetic radiation that we come into contact with on a daily basis in enormous. Radiations having these different frequencies have different names and properties. In Figure emSpectrum frequencies and their corresponding wavelengths are cataloged.

Figure emSpectrum

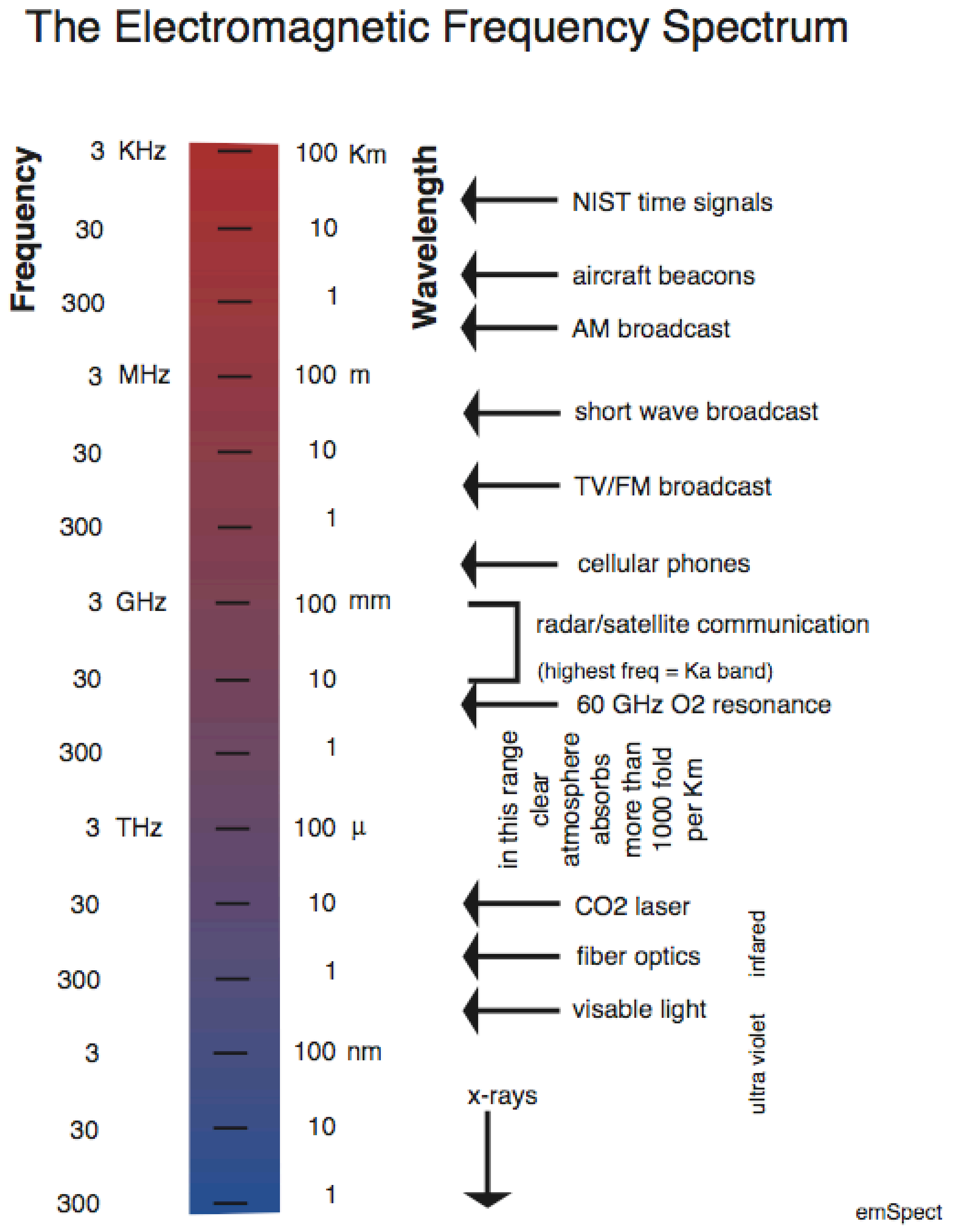


Figure emSpectrum. The low frequencies at the red end of the spectrum have long wavelengths, while the high frequencies at the blue end of the spectrum have short wavelengths. Frequency, wavelength, and uses of electromagnetic radiation.

Systems using the lower frequencies, below 3 MHz, are characteristically non-directional. The radiation follows the curvature of the earth for several hundred kilometers, greatly exceeding the "line of sight" path. This can be an advantage if you are broadcasting to everyone everywhere in that region.

Frequencies from 3 to 30 MHz are historically called the "short wave" band. Antennas of modest size are directional, and the radiation is partially reflected from ionized layers of the atmosphere. Multiple reflection between these layers and the earth permit signals to travel thousands of Km, although transmission range depends on the time of day, season of the year, and the activity of sunspots.

In the range 30 MHz to 30 GHz, radiation becomes very directional with even small antennas. The noise level, due to electrical devices and static discharge, is low at these frequencies. Since the total bandwidth is huge compared to the lower frequency bands, the amount of information that can be sent using these frequencies is large. In the early days of radio (before 1960) it was difficult or impossible to transmit at these frequencies, but now efficient solid state devices are available. If you know where you want to go, and there is a clear "line of sight" path, these are the frequencies to use. Transmission to and from satellites uses these frequencies.

The next band, extending over a hundred fold range in frequencies, is unusable for communication because the radiation is strongly absorbed by molecules in the air.

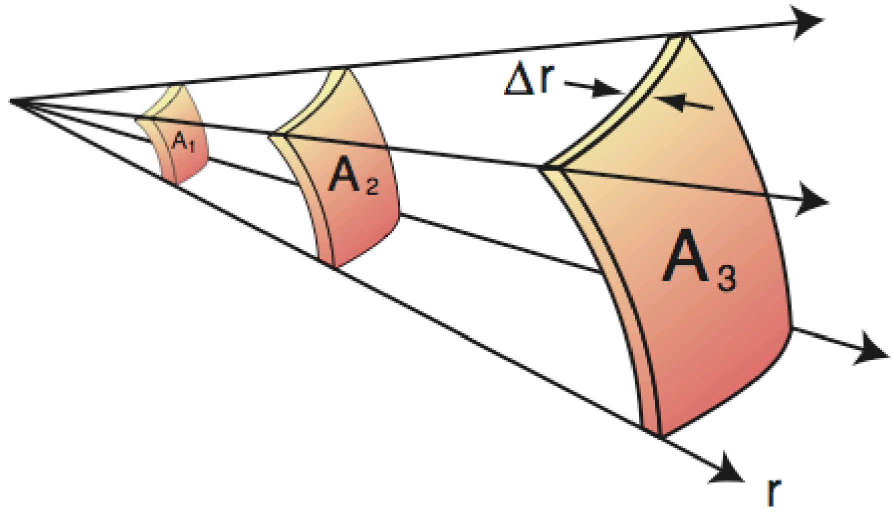
The optical frequencies start at 30 THz, a wavelength of 10 microns. You can't see radiation with a wave length of 10 microns, it's in the far infrared, and is perceived as heat. The CO_2 laser, used to cut steel, produces radiation of this wavelength. Radiation with wavelengths of 1 to 3 microns is used to transmit information through optical fibers and read CDs. Light is perceived as red at about 0.6 microns while blue light has a wavelength of 0.45 microns.

The $1/r^2$ law

Radiation may leave the transmitter antenna and propagate almost equally in all directions or with another antenna the radiation may be focused into a narrow beam. However, after leaving any antenna, the intensity of the signal will eventually decrease in proportion to the square of the distance. Thus, if the distance to the receiver increases by a factor of two, the intensity of the radiation will decrease by a factor of four. This fundamental $1/r^2$ law is a consequence of the constant speed of electromagnetic radiation and conservation of energy. As illustrated in Figure rSquare a pulse of energy will occupy a volume of constant thickness (proportional to the duration of the pulse) as it travels away from the transmitter. Since the area of the propagation pulse increases as the square of the distance, the energy density (intensity per unit volume) must decrease with the square of the distance for the total energy to remain constant.

Figure rSquare

The $1 / r^2$ law



$$\Delta r = c\,\Delta t = \text{constant}$$

$$A = k\,r^2$$

rSquare

Figure rSquare. A pulse of radiation occupies a zone of constant thickness as it moves away from the antenna because its speed is constant. However, the area of the zone , and thus its volume, increases as the square of the distance. Thus the energy density must decrease with the square of the distance to keep the total energy constant.

Antennas

The antenna that transmits the signal may be a simple wire, a network of wires, or a metal dish. The function of the antenna is to focus the energy into a pattern or beam, so the intensity is higher than would be the case if the energy were radiated out equally in all directions. The $1/r^2$ law still holds, but the intensity will be higher if the energy is concentrated in a useful direction. The definition of a useful direction is very dependent on the specific use of the system. However, it is typically not useful to send energy down into the earth or upward into space, unless you are communicating with a satellite. A directional antenna can be as beneficial for the receiver as the transmitter, since it reduces the noise that enters the receiver. Powerful amplifiers are fairly easy to make, and thus it is typically the ratio of the signal to noise that limits the effective transmission of information.

To achieve a narrow beam and thus a high intensity, you need a large antenna. However, “large” is defined with respect to the wave length of the radiation. A small wave length means that you can make a strongly directional antenna that is small in absolute dimensions, i.e. a dish with a diameter of 100 wavelengths will give the same directional pattern what ever the actual wavelength is. Thus, one appeal of using short wavelengths is that powerful antennas can be made that have modest absolute size.

An antenna is directional, whether it is a system of wires or a parabolic dish, because radiation coming from various parts of the antenna is in phase (and thus reinforces itself) in the desired direction, while it is out of phase (and thus cancels itself) in all other directions. The most simple antenna is a dipole, shown in Figure antenna. Collections of dipoles can produce a highly directional antenna. If fact, a microwave dish can be considered a surface of dipoles, with currents in the metal surface induced by the radiation that is focused by the dish.

Figure antenna

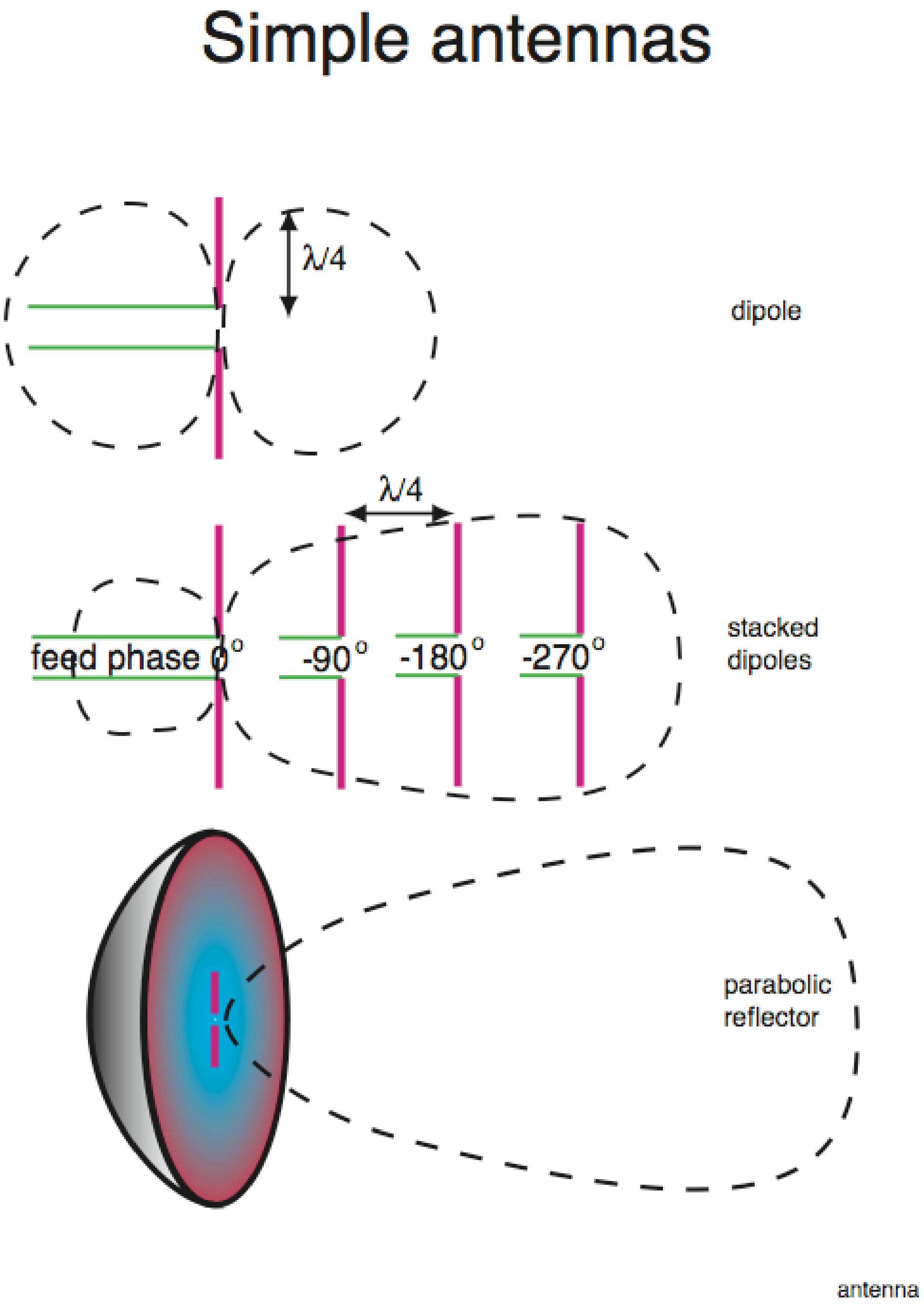


Figure antenna. The most simple antenna is a dipole. The two active elements are the red wires, each a quarter of a wavelength long and fed by the green wires. By stacking several dipoles the directivity can be increased and the pattern made unidirectional . A parabolic reflecting dish with a diameter ten or one hundred wavelengths can produce a very directional beam of radiation. The black dashed lines indication intensity.

The phase relationships between the different parts of an antenna, which define the radiation pattern, are usually due to the geometry of the system, i.e. they are a consequence of the antenna's physical structure. However, in some advanced systems, the relative phase of the signal in different parts of the antenna are controlled by active components, amplifiers or oscillators that shift the phase of the signal in response to a command signal. In this way the direction of the beam can be changed without physically moving the structure. Such systems find use in the radar systems on aircraft carriers. The antenna are typically four large flat panels on the superstructure. The radar beams can be directed in an arbitrarily rapid and complex pattern with no physical movement of the antenna itself. Similar systems are used in some communication satellites. Signals can then be beamed down to various locations on the earth in rapid sequence with no mechanical motion of the satellite itself.

Satellite links

A satellite can be used to form a link between two ground based stations. Microwaves are beamed to the satellite by one station and the information is retransmitted by the satellite down to the other earth station. The satellites are typically geostationary, which means that they appear stationary to the observer on earth. However, they are actually orbiting the earth around its equator, in the same direction that the earth moves, exactly once every 24 hours. The stable orbit with this rotation period has a radius of about 40,000 Km (25,000 miles). The ground stations must use a medium power transmitter and large antenna to send and receive a high data rate signal over this distance. A small, hand held device does not have the effective power to send a high data rate signal to a geo-stationary satellite. However, the small emergency beacons typically carried on ships and airplanes are able to sent a short message at a low bit rate to a geostationary satellite. The message consists of a beacon ID and its latitude and longitude, if the beacon is connected to a GPS receiver. This is enough to initiate a search and rescue mission.

Much higher data rates are possible between a small portable device and a satellite if the satellite is only a thousand Km above the earth. However, now the satellite is moving rapidly across the sky, and communication can only be maintained for a few minutes. The obvious, but technically demanding solution is to have many satellites in overlapping orbits which can communicate with each other to provide a continually changing but unbroken path from one site on earth to another. The first commercial system of this design, Iridium, used 66 satellites and became operational in 1998. It provides voice and low speed data links using hand held devices only slightly larger than the ubiquitous cellular phone. Unfortunately for investors in Iridium, the number of users of the system were far less than needed to pay for its operation. However, satellite systems using similar technology but capable of higher data rates were still being planned at of 2001. Fiber optic links are replacing satellites for high volume point to point data transmission since they carry data at a much higher rate.

Optical fiber

A carrier can only be modulated by frequencies considerably lower than its own frequency. Thus, if we need high data transmission rates we need high carrier frequencies. Light is electromagnetic radiation having a very high frequency. Light can be easily directed and controlled by mirrors and lenses, however, it is absorbed or scattered by water droplets, dust, and other atmospheric contaminants and disturbances. In addition, it is often convenient to be able to bend a communication link around corners and pass it through complex structures, e.g. buildings. The optical fiber used for light transmission is a analog to a wire used for electrical transmission. An optical fiber is a long cylindrical tube which transmits light. Obviously the fiber must be highly transparent, and is thus glass or a clear plastic.

Transparent materials are insulators and have only bound electrons

Electromagnetic radiation can only penetrate a conductor for a few wavelengths. The oscillating fields induce currents in the conductor that quickly cancel out the very fields that induced them. Thus copped wire or even a thin copper film is effectively opaque to light, which has wavelengths of 0.3 to 3 μ. However, light can travel long distances through an insulator, which is a material that has no free electrons. In order for the light to not be scattered as it travels along its path, the molecular structure of the material must be uniform at the scale of a wavelength. Glass and some plastics are thus ideal conductors for light. It is easy to decide if a material is a possibility for an optical fiber; it appears clear and transparent.

The propagation of pulse of light down a glass fiber has some similarities to the movement of an electrical pulse down a wire. The advancing wave of light causes bound electrons in the material to vibrate, and these vibrations in turn generate their own electromagnetic field that reinforces the original field and causes electrons further along the path to vibrate. Thus, a transparent material is not just a passive hole for the radiation to pass through, rather the vibrating electrons of the material actively participate in transmitting the energy. These bound electrons however play a different role than the free electrons that generate a net current down a conductor. The molecular structure of the glass determines how the electrons will vibrate, which in turn defines the speed of light in the glass.

Index of refraction and the speed of light

A narrow beam of light directed downward into water will bend as it enters the surface (unless it is exactly perpendicular to the surface). The path of the light can easily be seen if the air contains a little dust or smoke and the water is not completely clear. This same phenomenon causes the image of a straw inserted into a glass of water to appear to be bent where it passes through the surface. Glass, like water, also bends light, and this is the basis of a lens. Glass bends light and thus glass lenses are able to focus light and produce images. The bending of light is called refraction, and the degree to which a beam of light is bent when it moves from on material into another is a measure of the difference in the index of refraction of the two materials.

The reason a beam of light bends when is passes from a material with one index of refraction into a material with another index of refraction is that the speed of light is different in the two materials. The larger the index of refraction the slower the velocity of light in the material. The index of refraction in a vacuum is defined to be 1.000, and the value for air is very similar. The index of refraction of water is about 1.3, while a typical glass has an index of refraction of 1.5.

Thick fiber

When the angle at which a beam of light hits an interface between materials of different refractive indexes is smaller than a critical value, all the light is reflected from the boundary. This reflection allows transmission of light inside a thick optical fiber; the light just bounces from wall to wall.

Figure opticFiber

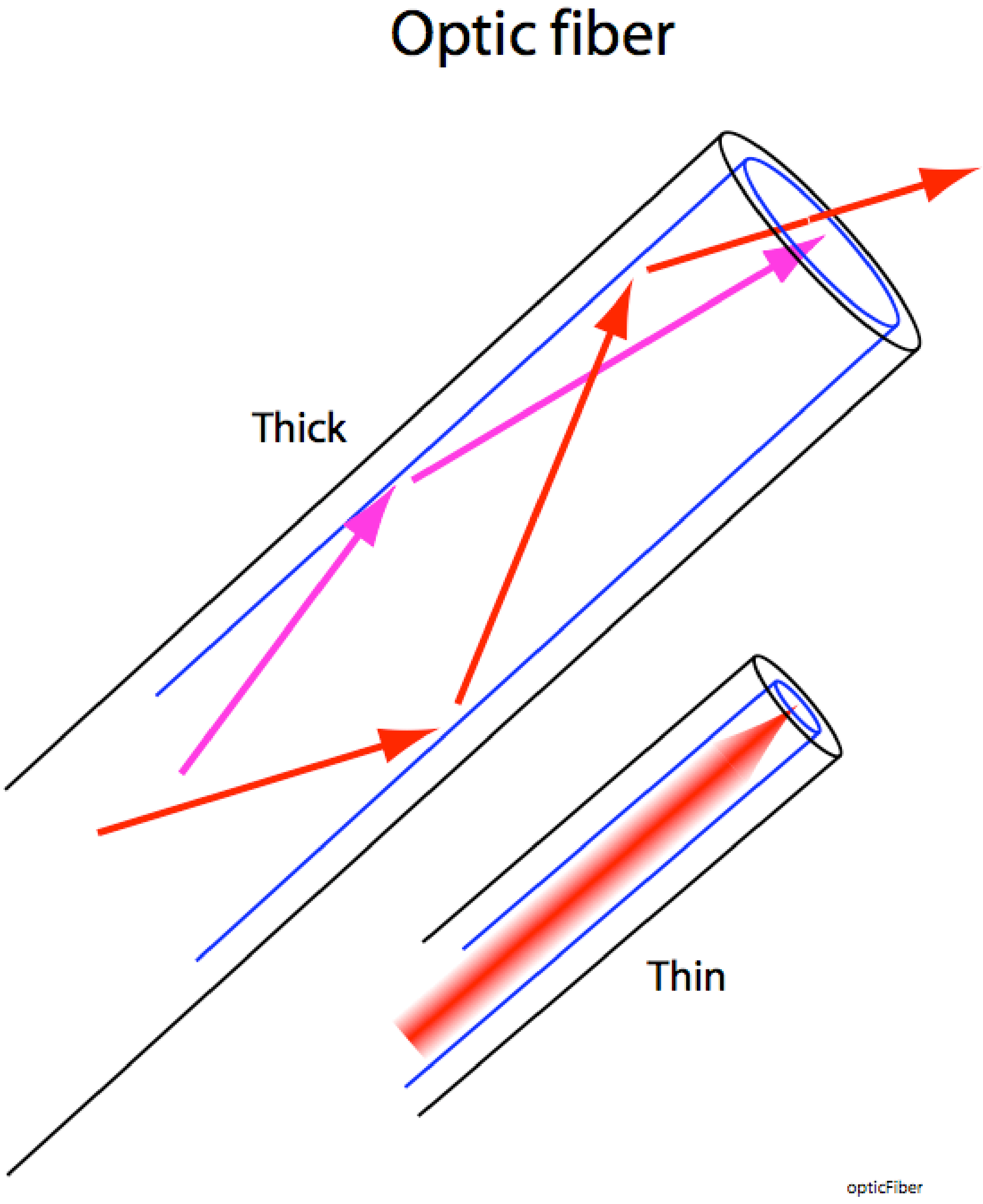


Figure opticFiber. Multiple beams move down a thick fiber by reflection from the surface. Since there are multiple paths, each with a different ratio of total length to distance down the fiber, the signals eventually become out of phase and interfere with each other. A thin fiber only transmits a single mode with a single path length, thus avoiding this problem.

Thick is relative, here it only means that the transparent core is on the order of dozens of wavelengths in diameter. With this thickness the energy propagating down the fiber can be considered a collection of individual rays of light, each with its own characteristic angle with the axis of the fiber. However, when any of the beams hit the surface of the fiber, the angle must be less than the critical angle, or else energy will be lost and this path will quickly be extinguished from the collection. Each of the paths down the fiber can be considered to be a mode of transmission, and since the different modes have different total path lengths along the fiber axis, light pulses which start together at one end of the fiber and travel to the other end by different modes will arrive at different times. This is called mode dispersion. A more sophisticated version of thick fiber has an inner cylinder with a radial gradient in the index of refraction. This gradient concentrated light into the fastest mode, however, the fiber costs more to fabricate. An advantage of thick fiber is that it is cheap to manufacture and connections between segments are easy to make since the physical alignment is not very critical.

Thin fiber

Now the inner cylinder is only a few wavelengths in diameter; for light with a wavelength of 1.55 μm the diameter should be less than 4 μ. In this fiber the electromagnetic radiation can not travel in a zigzag path, careening from one surface of the cylinder to the other; it must travel in a single mode down the axis of the fiber. There is thus a single speed and a pulse of light will retain its shape for a long distance. The disadvantages are that the fiber is more expensive to manufacture and the joints require far more precise mechanical alignment.

Dispersion

In real materials the index or refraction is not a constant, it changes slightly with wavelength. This is called dispersion. In a camera with a cheap lens this means that a sharp edge will appear as a narrow rainbow in the print, i.e. the image of the edge in different colors will be bent and focused by the lens to slightly different positions, and thus will not be in register. In the camera world this effect is called chromatic aberration. It can be greatly reduced by using a compound lens made of several types of glass with different index of refraction versus wavelength curves. Clever engineering can generate lenses in which the aberrations of the different glasses almost cancel each other out, to produce a much improved image.

Even if the laser produced an absolutely pure signal of one color, as we have seen previously just the modulation of the carrier produces a band of frequencies or colors. To minimize the effects of dispersion the carrier wavelength is usually chosen to be close the value were the change of index of refraction versus wavelength is a minimum. For the common silica glass optical fibers the minimum dispersion is at 1.3 μ.

Attenuation

Glass looks clear but it absorbs some energy. For the common silica glass optical fibers the minimum attenuation is at 1.55 μ. Murphy's law says it not the same as the wavelength of lowest dispersion; bummer! A great property of silica glasses is that by adding very small amounts of other elements, typically rare earths, the electro-optical properties can be changed. In this way it is possible to change the point of minimal wavelength dispersion (dispersion shifting), to 1.55 μ; back in the saddle again.

Wavelength multiplexing

An obvious way to increase the rate of information transfer in a fiber is to use several wavelengths to transmit information. This requires methods of producing and modulating laser radiation at closely grouped wavelengths. These wavelengths must be separated and demodulated at the receiving end of the fiber link.

Solitons

In the real world, properties of materials change as the power increases. Amazingly, in some cases the change in the parameter with intensity, e.g. the change in speed with intensity, can generate a special wave called a soliton, which does not spread out as it travels down the fiber even if there is some dispersion. As an example, in 1998 scientists at the Chalmers University of Technology at Gothenburg, Sweden achieved 40 Gbps soliton transmission over a 400 Km dispersion-shifted fiber. However, it is not clear when, or even if soliton transmission will be used in a commercial system.

Optical amplifiers and more?

A communications network like the Internet is far more than transmission links. Aside from the computers that send and receive the message, there are routers at the nodes of the network that determine which link each incoming message is sent back out on, and there are amplifiers distributed along the path that periodically boost the strength of the signal to compensate for attenuation. When optical fiber was first introduced all other functions were performed electronically. Thus the light signal was converted to an electrical one, routed or amplified, and then converted back to an optical signal. The expense of making all these conversions made optical transmission economical for only the largest communication trunks, where the capacity of optical fibers for very large data transmission rates over balanced the conversion costs.

The optical amplifier eliminated one set of conversions. The amplifier is a segment of doped fiber (typically using erbium), a few meters (5 to 20 ft) in length, which is illuminated by a driver laser of a shorter wavelength than the incoming weak signal. The driver laser drives some of the electrons in the fiber to a high energy state. The incoming weak message signal triggers the energetic electrons to fall into a lower state in synchrony with the signal. The energy released by the electrons thus creates a stronger signal.

The next hurdle in replacing electronic components is routing. This process is much more difficult than amplification because at least part of the message, the header that contains the address, must be read and a logical decision (computation) made to send the message somewhere.

Chapter summary

Electrical pulses can carry information along a wire, but to send data at a high rate the pulses must be short but distinct, ideally a series of square waves. Square waves contain sine waves of high and low frequencies, and the high frequencies radiate energy out from the wire. The loss of energy in the high frequencies and differences in the speed at which the different frequencies travel down the wire, eventually degrades the shape of the electrical pulses and make it impossible to retrieve the information they carried.

Radio and optical signals move through non-conducting materials, e.g. space, air, glass. The information is transmitted by modulating a carrier signal of much higher frequency. As the rate of information transfer increases, the bandwidth of frequencies occupied by the modulated carrier also increases. Since noise increases with bandwidth, high data rates require low noise systems, which means high transmitter power or highly directional antennas or both.

The highest data rates require the highest carrier frequencies; light. Optic fiber systems become more effective as more functions can be done without converting the information into electrical signals and back to light.