Transducer Devices & their Technological Applications

Imelda Hollenbeck

First Edition, 2012

ISBN 978-81-323-1086-0

Published by:
College Publishing House
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter 1

Antenna

Short wave "curtain" antenna (Moosbrunn, Austria)

An **antenna** (or **aerial**) is a transducer that transmits or receives electromagnetic waves. In other words, antennas convert electromagnetic radiation into electric current, or vice versa. Antennas generally deal in the transmission and reception of radio waves, and are a necessary part of all radio equipment. Antennas are used in systems such as radio and television broadcasting, point-to-point radio communication, wireless LAN, cell phones, radar, and spacecraft communication. Antennas are most commonly employed in air or outer space, but can also be operated under water or even through soil and rock at certain frequencies for short distances.

Physically, an antenna is an arrangement of one or more conductors, usually called *elements* in this context. In transmission, an alternating current is created in the elements by applying a voltage at the antenna terminals, causing the elements to radiate an electromagnetic field. In reception, the inverse occurs: an electromagnetic field from another source induces an alternating current in the elements and a corresponding voltage at the antenna's terminals. Some receiving antennas (such as parabolic and horn types) incorporate shaped reflective surfaces to collect the radio waves striking them and direct or focus them onto the actual conductive elements.

Some of the first rudimentary antennas were built in 1888 by Heinrich Hertz (1857–1894) in his pioneering experiments to prove the existence of electromagnetic waves predicted by the theory of James Clerk Maxwell. Hertz placed the emitter dipole in the focal point of a parabolic reflector. He published his work and installation drawings in *Annalen der Physik und Chemie* (vol. 36, 1889).

Terminology

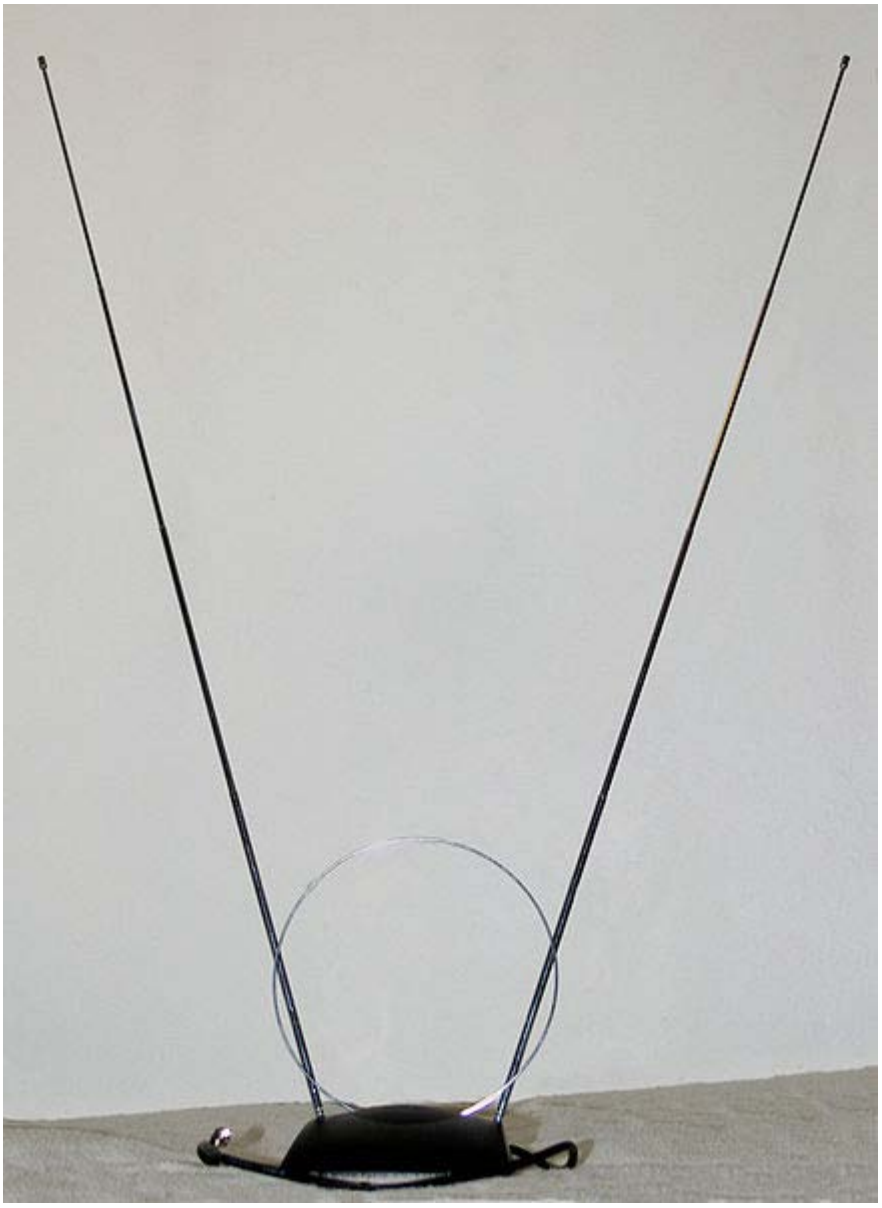
The words *antenna* (plural: *antennas*) and *aerial* are used interchangeably; but usually a rigid metallic structure is termed an antenna and a wire format is called an aerial. In the United Kingdom and other British English speaking areas the term aerial is more common, even for rigid types. The noun *aerial* is occasionally written with a diaeresis mark—*aërial*—in recognition of the original spelling of the adjective *aërial* from which the noun is derived.

The origin of the word *antenna* relative to wireless apparatus is attributed to Guglielmo Marconi. In 1895, while testing early radio apparatuses in the Swiss Alps at Salvan, Switzerland in the Mont Blanc region, Marconi experimented with early wireless equipment. A 2.5 meter long pole, along which was carried a wire, was used as a radiating and receiving aerial element. In Italian a tent pole is known as *l'antenna centrale,* and the pole with a wire alongside it used as an aerial was simply called *l'antenna.* Until then wireless radiating transmitting and receiving elements were known simply as aerials or terminals. Marconi's use of the word *antenna* (Italian for *pole*) would become a popular term for what today is uniformly known as the *antenna*.

A Hertzian antenna is a set of terminals that does not require the presence of a ground for its operation (versus a Tesla antenna which is grounded). A loaded antenna is an active antenna having an elongated portion of appreciable electrical length and having

additional inductance or capacitance directly in series or shunt with the elongated portion so as to modify the standing wave pattern existing along the portion or to change the effective electrical length of the portion. An antenna grounding structure is a structure for establishing a reference potential level for operating the active antenna. It can be any structure closely associated with (or acting as) the ground which is connected to the terminal of the signal receiver or source opposing the active antenna terminal.

In colloquial usage, the word *antenna* may refer broadly to an entire assembly including support structure, enclosure (if any), etc. in addition to the purely functional components.



"Rabbit ears" dipole antenna for television reception

Cell phone base station antennas

Parabolic antenna for communicating with spacecraft, Canberra, Australia

Yagi antenna used for mobile military communications station, Dresden, Germany, 1955

"Super Turnstile" type transmitting antenna for VHF low band television broadcasting station, Germany.

Overview

Antennas have practical uses for the transmission and reception of radio frequency signals such as radio and television. In air, those signals travel very quickly and with a very low transmission loss. The signals are absorbed when moving through more conductive materials, such as concrete walls or rock. When encountering an interface, the waves are partially reflected and partially transmitted through.

A common antenna is a vertical rod a quarter of a wavelength long. Such antennas are simple in construction, usually inexpensive, and both radiate in and receive from all horizontal directions (omnidirectional). One limitation of this antenna is that it does not radiate or receive in the direction in which the rod points. This region is called the antenna blind cone or null.

There are two fundamental types of antenna directional patterns, which, with reference to a specific two dimensional plane (usually horizontal [parallel to the ground] or vertical [perpendicular to the ground]), are either:

1. Omni-directional (radiates equally in all directions), such as a vertical rod (in the horizontal plane) or
2. Directional (radiates more in one direction than in the other).

In colloquial usage "omnidirectional" usually refers to all horizontal directions with reception above and below the antenna being reduced in favor of better reception (and thus range) near the horizon. A "directional" antenna usually refers to one focusing a narrow beam in a single specific direction such as a telescope or satellite dish, or, at least, focusing in a sector such as a 120° horizontal fan pattern in the case of a panel antenna at a cell site.

All antennas radiate some energy in all directions in free space but careful construction results in substantial transmission of energy in a preferred direction and negligible energy radiated in other directions. By adding additional *elements* (such as rods, loops or plates) and carefully arranging their length, spacing, and orientation, an antenna with desired directional properties can be created.

An antenna array is two or more simple antennas combined to produce a specific directional radiation pattern. In common usage an array is composed of active elements, such as a linear array of parallel dipoles fed as a "broadside array". A slightly different feed method could cause this same array of dipoles to radiate as an "end-fire array". Antenna arrays may be built up from any basic antenna type, such as dipoles, loops or slots.

The directionality of the array is due to the spatial relationships and the electrical feed relationships between individual antennas. Usually all of the elements are active (electrically fed) as in the log-periodic dipole array which offers modest gain and broad bandwidth and is traditionally used for television reception. Alternatively, a superficially similar dipole array, the Yagi-Uda Antenna (often abbreviated to "Yagi"), has only one active dipole element in a chain of parasitic dipole elements, and a very different performance with high gain over a narrow bandwidth.

An active element is electrically connected to the antenna terminals leading to the receiver or transmitter, as opposed to a parasitic element that modifies the antenna pattern without being connected directly. The active element(s) couple energy between the electromagnetic wave and the antenna terminals, thus any functioning antenna has at least

one active element. A careful arrangement of parasitic elements, such as rods or coils, can improve the radiation pattern of the active element(s). Directors and reflectors are common parasitic elements.

An antenna lead-in is the medium, for example, a transmission line or feed line for conveying the signal energy between the signal source or receiver and the antenna. The antenna feed refers to the components between the antenna and an amplifier.

An antenna counterpoise is a structure of conductive material most closely associated with ground that may be insulated from or capacitively coupled to the natural ground. It aids in the function of the natural ground, particularly where variations (or limitations) of the characteristics of the natural ground interfere with its proper function. Such structures are usually connected to the terminal of a receiver or source opposite to the antenna terminal.

An antenna component is a portion of the antenna performing a distinct function and limited for use in an antenna, as for example, a reflector, director, or active antenna.

An electromagnetic wave refractor is a structure which is shaped or positioned to delay or accelerate transmitted electromagnetic waves, passing through such structure, an amount which varies over the wave front. The refractor alters the direction of propagation of the waves emitted from the structure with respect to the waves impinging on the structure. It can alternatively bring the wave to a focus or alter the wave front in other ways, such as to convert a spherical wave front to a planar wave front (or vice-versa). The velocity of the waves radiated have a component which is in the same direction (director) or in the opposite direction (reflector) as that of the velocity of the impinging wave.

A *director* is a parasitic element, usually a metallic conductive structure, which re-radiates into free space impinging electromagnetic radiation coming from or going to the active antenna, the velocity of the re-radiated wave having a component in the direction of the velocity of the impinging wave.

A reflector is a parasitic element, usually a metallic conductive structure (e.g., screen, rod or plate), which re-radiates back into free space impinging electromagnetic radiation coming from or going to the active antenna. The velocity of the returned wave has a component in a direction opposite to the direction of the velocity of the impinging wave. The reflector modifies the radiation of the active antenna.

An antenna coupling network is a passive network (which may be any combination of a resistive, inductive or capacitive circuit(s)) for transmitting the signal energy between the active antenna and a source (or receiver) of such signal energy.

Typically, antennas are designed to operate in a relatively narrow frequency range. The design criteria for receiving and transmitting antennas differ slightly, but generally an antenna can receive and transmit equally well. This property is called reciprocity.

Parameters

There are several critical parameters affecting an antenna's performance that can be adjusted during the design process. These are resonant frequency, impedance, gain, aperture or radiation pattern, polarization, efficiency and bandwidth. Transmit antennas may also have a maximum power rating, and receive antennas differ in their noise rejection properties. All of these parameters can be measured through various means.

Resonant frequency

The "*resonant frequency*" and "*electrical resonance*" is related to the electrical length of an antenna. The electrical length is usually the physical length of the wire divided by its velocity factor (the ratio of the speed of wave propagation in the wire to c_0, the speed of light in a vacuum). Typically an antenna is tuned for a specific frequency, and is effective for a range of frequencies that are usually centered on that resonant frequency. However, other properties of an antenna change with frequency, in particular the radiation pattern and impedance, so the antenna's resonant frequency may merely be close to the center frequency of these other more important properties.

Antennas can be made resonant on harmonic frequencies with lengths that are fractions of the target wavelength; this resonance gives much better coupling to the electromagnetic wave, and makes the aerial act as if it were physically larger.

Some antenna designs have multiple resonant frequencies, and some are relatively effective over a very broad range of frequencies. The most commonly known type of wide band aerial is the logarithmic or log periodic, but its gain is usually much lower than that of a specific or narrower band aerial.

Gain

Gain as a parameter measures the efficiency of a given antenna with respect to a given norm, usually achieved by modification of its directionality. An antenna with a low gain emits radiation with about the same power in all directions, whereas a high-gain antenna will preferentially radiate in particular directions. Specifically, the **Gain**, **Directive gain** or **Power gain** of an antenna is defined as the ratio of the intensity (power per unit surface) radiated by the antenna in a given direction at an arbitrary distance divided by the intensity radiated at the same distance by a hypothetical isotropic antenna.

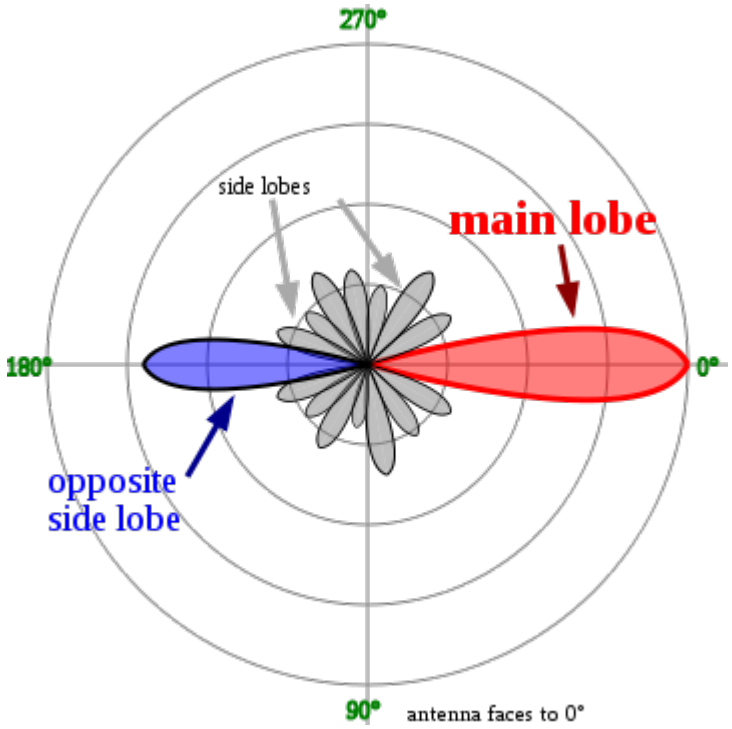
The gain of an antenna is a passive phenomenon - power is not added by the antenna, but simply redistributed to provide more radiated power in a certain direction than would be transmitted by an isotropic antenna. If an antenna has a gain greater than one in some directions, it must have a gain less than one in other directions, since energy is conserved by the antenna. An antenna designer must take into account the application for the antenna when determining the gain. High-gain antennas have the advantage of longer range and better signal quality, but must be aimed carefully in a particular direction. Low-gain antennas have shorter range, but the orientation of the antenna is relatively

inconsequential. For example, a dish antenna on a spacecraft is a high-gain device that must be pointed at the planet to be effective, whereas a typical Wi-Fi antenna in a laptop computer is low-gain, and as long as the base station is within range, the antenna can be in any orientation in space. It makes sense to improve horizontal range at the expense of reception above or below the antenna. Thus most antennas labelled "omnidirectional" really have some gain.

In practice, the half-wave dipole is taken as a reference instead of the isotropic radiator. The gain is then given in **dBd** (decibels over **d**ipole):

> NOTE: **0 dBd = 2.15 dBi**. It is vital in expressing gain values that the reference point be included. Failure to do so can lead to confusion and error.

Radiation pattern



polar plots of the horizontal cross sections of a (virtual) Yagi-Uda-antenna. Outline connects points with 3db field power compared to an ISO emitter.

The radiation pattern of an antenna is the geometric pattern of the relative field strengths of the field emitted by the antenna. For the ideal isotropic antenna, this would be a sphere. For a typical dipole, this would be a toroid. The radiation pattern of an antenna is typically represented by a three dimensional graph, or polar plots of the horizontal and vertical cross sections. The graph should show sidelobes and backlobes, where the antenna's gain is at a minimum or maximum.

Impedance

As an electro-magnetic wave travels through the different parts of the antenna system (radio, feed line, antenna, free space) it may encounter differences in impedance (E/H, V/I, etc.). At each interface, depending on the impedance match, some fraction of the wave's energy will reflect back to the source, forming a standing wave in the feed line. The ratio of maximum power to minimum power in the wave can be measured and is called the standing wave ratio (**SWR**). A SWR of 1:1 is ideal. A SWR of 1.5:1 is considered to be marginally acceptable in low power applications where power loss is more critical, although an SWR as high as 6:1 may still be usable with the right equipment. Minimizing impedance differences at each interface (impedance matching) will reduce SWR and maximize power transfer through each part of the antenna system.

Complex impedance of an antenna is related to the electrical length of the antenna at the wavelength in use. The impedance of an antenna can be matched to the feed line and radio by adjusting the impedance of the feed line, using the feed line as an impedance transformer. More commonly, the impedance is adjusted at the load (see below) with an antenna tuner, a balun, a matching transformer, matching networks composed of inductors and capacitors, or matching sections such as the gamma match.

Efficiency

Efficiency is the ratio of power actually radiated to the power put into the antenna terminals. A dummy load may have an SWR of 1:1 but an efficiency of 0, as it absorbs all power and radiates heat but not RF energy, showing that SWR alone is not an effective measure of an antenna's efficiency. Radiation in an antenna is caused by radiation resistance which can only be measured as part of total resistance including loss resistance. Loss resistance usually results in heat generation rather than radiation, and reduces efficiency. Mathematically, efficiency is calculated as radiation resistance divided by total resistance.

Bandwidth

The *bandwidth* of an antenna is the range of frequencies over which it is effective, usually centered on the resonant frequency. The bandwidth of an antenna may be increased by several techniques, including using thicker wires, replacing wires with *cages* to simulate a thicker wire, tapering antenna components (like in a feed horn), and combining multiple antennas into a single assembly and allowing the natural impedance to select the correct antenna. Small antennas are usually preferred for convenience, but there is a fundamental limit relating bandwidth, size and efficiency.

Polarization

The *polarization* of an antenna is the orientation of the electric field (E-plane) of the radio wave with respect to the Earth's surface and is determined by the physical structure of the antenna and by its orientation. It has nothing in common with antenna

directionality terms: "horizontal", "vertical" and "circular". Thus, a simple straight wire antenna will have one polarization when mounted vertically, and a different polarization when mounted horizontally. "Electromagnetic wave polarization filters" are structures which can be employed to act directly on the electromagnetic wave to filter out wave energy of an undesired polarization and to pass wave energy of a desired polarization.

Reflections generally affect polarization. For radio waves the most important reflector is the ionosphere - signals which reflect from it will have their polarization changed unpredictably. For signals which are reflected by the ionosphere, polarization cannot be relied upon. For line-of-sight communications for which polarization can be relied upon, it can make a large difference in signal quality to have the transmitter and receiver using the same polarization; many tens of dB difference are commonly seen and this is more than enough to make the difference between reasonable communication and a broken link.

Polarization is largely predictable from antenna construction but, especially in directional antennas, the polarization of side lobes can be quite different from that of the main propagation lobe. For radio antennas, polarization corresponds to the orientation of the radiating element in an antenna. A vertical omnidirectional WiFi antenna will have vertical polarization (the most common type). An exception is a class of elongated waveguide antennas in which vertically placed antennas are horizontally polarized. Many commercial antennas are marked as to the polarization of their emitted signals.

Polarization is the sum of the E-plane orientations over time projected onto an imaginary plane perpendicular to the direction of motion of the radio wave. In the most general case, polarization is elliptical, meaning that the polarization of the radio waves varies over time. Two special cases are linear polarization (the ellipse collapses into a line) and circular polarization (in which the two axes of the ellipse are equal). In linear polarization the antenna compels the electric field of the emitted radio wave to a particular orientation. Depending on the orientation of the antenna mounting, the usual linear cases are horizontal and vertical polarization. In circular polarization, the antenna continuously varies the electric field of the radio wave through all possible values of its orientation with regard to the Earth's surface. Circular polarizations, like elliptical ones, are classified as right-hand polarized or left-hand polarized using a "thumb in the direction of the propagation" rule. Optical researchers use the same rule of thumb, but pointing it in the direction of the emitter, not in the direction of propagation, and so are opposite to radio engineers' use.

In practice, regardless of confusing terminology, it is important that linearly polarized antennas be matched, lest the received signal strength be greatly reduced. So horizontal should be used with horizontal and vertical with vertical. Intermediate matchings will lose some signal strength, but not as much as a complete mismatch. Transmitters mounted on vehicles with large motional freedom commonly use circularly polarized antennas so that there will never be a complete mismatch with signals from other sources.

Transmission and reception

All of the antenna parameters are expressed in terms of a transmission antenna, but are identically applicable to a receiving antenna, due to reciprocity. Impedance, however, is not applied in an obvious way; for impedance, the impedance at the load (where the power is consumed) is most critical. For a transmitting antenna, this is the antenna itself. For a receiving antenna, this is at the (radio) receiver rather than at the antenna. Tuning is done by adjusting the length of an electrically long linear antenna to alter the electrical resonance of the antenna.

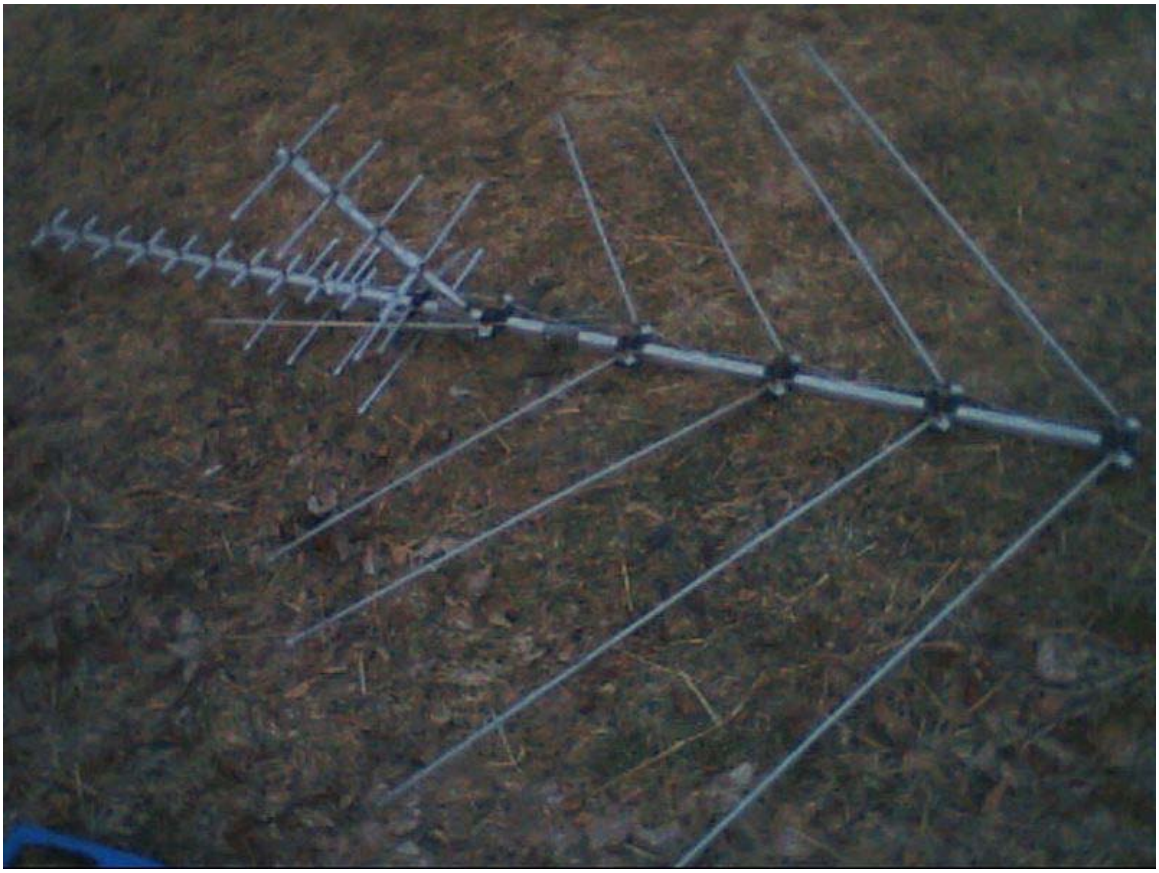
Antenna tuning is done by adjusting an inductance or capacitance combined with the active antenna (but distinct and separate from the active antenna). The inductance or capacitance provides the reactance which combines with the inherent reactance of the active antenna to establish a resonance in a circuit including the active antenna. The established resonance being at a frequency other than the natural electrical resonant frequency of the active antenna. Adjustment of the inductance or capacitance changes this resonance.

Antennas used for transmission have a maximum power rating, beyond which heating, arcing or sparking may occur in the components, which may cause them to be damaged or destroyed. Raising this maximum power rating usually requires larger and heavier components, which may require larger and heavier supporting structures. This is a concern only for transmitting antennas, as the power received by an antenna rarely exceeds the microwatt range.

Antennas designed specifically for reception might be optimized for noise rejection capabilities. An *antenna shield* is a conductive or low reluctance structure (such as a wire, plate or grid) which is adapted to be placed in the vicinity of an antenna to reduce, as by dissipation through a resistance or by conduction to ground, undesired electromagnetic radiation, or electric or magnetic fields, which are directed toward the active antenna from an external source or which emanate from the active antenna. Other methods to optimize for noise rejection can be done by selecting a narrow bandwidth so that noise from other frequencies is rejected, or selecting a specific radiation pattern to reject noise from a specific direction, or by selecting a polarization different from the noise polarization, or by selecting an antenna that favors either the electric or magnetic field.

For instance, an antenna to be used for reception of low frequencies (below about ten megahertz) will be subject to both man-made noise from motors and other machinery, and from natural sources such as lightning. Successfully rejecting these forms of noise is an important antenna feature. A small coil of wire with many turns is more able to reject such noise than a vertical antenna. However, the vertical will radiate much more effectively on transmit, where extraneous signals are not a concern.

Basic antenna models



Typical US multiband TV antenna (aerial)

There are many variations of antennas. Below are a few basic models. More can be found in Category:Radio frequency antenna types.

- The isotropic radiator is a purely theoretical antenna that radiates equally in all directions. It is considered to be a point in space with no dimensions and no mass. This antenna cannot physically exist, but is useful as a theoretical model for comparison with all other antennas. Most antennas' gains are measured with reference to an isotropic radiator, and are rated in dBi (decibels with respect to an isotropic radiator).
- The dipole antenna is simply two wires pointed in opposite directions arranged either horizontally or vertically, with one end of each wire connected to the radio and the other end hanging free in space. Since this is the simplest practical antenna, it is also used as a reference model for other antennas; gain with respect to a dipole is labeled as dBd. Generally, the dipole is considered to be omnidirectional in the plane perpendicular to the axis of the antenna, but it has deep nulls in the directions of the axis. Variations of the dipole include the folded dipole, the half wave antenna, the ground plane antenna, the whip, and the J-pole.

- The Yagi-Uda antenna is a directional variation of the dipole with parasitic elements added which are functionality similar to adding a reflector and lenses (directors) to focus a filament light bulb.
- The random wire antenna is simply a very long (at least one quarter wavelength) wire with one end connected to the radio and the other in free space, arranged in any way most convenient for the space available. Folding will reduce effectiveness and make theoretical analysis extremely difficult. (The added length helps more than the folding typically hurts.) Typically, a random wire antenna will also require an antenna tuner, as it might have a random impedance that varies non-linearly with frequency.
- The horn is used where high gain is needed, the wavelength is short (microwave) and space is not an issue. Horns can be narrow band or wide band, depending on their shape. A horn can be built for any frequency, but horns for lower frequencies are typically impractical. Horns are also frequently used as reference antennas.
- The parabolic antenna consists of an active element at the focus of a parabolic reflector to reflect the waves into a plane wave. Like the horn it is used for high gain, microwave applications, such as satellite dishes.
- The patch antenna consists mainly of a square conductor mounted over a groundplane. Another example of a planar antenna is the tapered slot antenna (TSA), as the Vivaldi-antenna.

Practical antennas

Very common "rabbit ears" set-top antenna

Although any circuit can radiate if driven with a signal of high enough frequency, most practical antennas are specially designed to radiate efficiently at a particular frequency. An example of an inefficient antenna is the simple Hertzian dipole antenna, which radiates over wide range of frequencies and is useful for its small size. A more efficient variation of this is the half-wave dipole, which radiates with high efficiency when the signal wavelength is twice the electrical length of the antenna.

One of the goals of antenna design is to minimize the reactance of the device so that it appears as a resistive load. An "antenna inherent reactance" includes not only the distributed reactance of the active antenna but also the natural reactance due to its location and surroundings (as for example, the capacity relation inherent in the position of the active antenna relative to ground). Reactance diverts energy into the reactive field, which causes unwanted currents that heat the antenna and associated wiring, thereby wasting energy without contributing to the radiated output. Reactance can be eliminated by operating the antenna at its resonant frequency, when its capacitive and inductive reactances are equal and opposite, resulting in a net zero reactive current. If this is not possible, compensating inductors or capacitors can instead be added to the antenna to cancel its reactance as far as the source is concerned.

Once the reactance has been eliminated, what remains is a pure resistance, which is the sum of two parts: the ohmic resistance of the conductors, and the radiation resistance. Power absorbed by the ohmic resistance becomes waste heat, and that absorbed by the radiation resistance becomes radiated electromagnetic energy. The greater the ratio of radiation resistance to ohmic resistance, the more efficient the antenna.

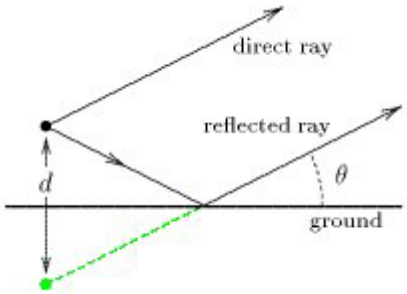
Effect of ground

Antennas are typically used in an environment where other objects are present that may have an effect on their performance. Height above ground has a very significant effect on the radiation pattern of some antenna types.

At frequencies used in antennas, the ground behaves mainly as a dielectric. The conductivity of ground at these frequencies is negligible. When an electromagnetic wave arrives at the surface of an object, two waves are created: one enters the dielectric and the other is reflected. If the object is a conductor, the transmitted wave is negligible and the reflected wave has almost the same amplitude as the incident one. When the object is a dielectric, the fraction reflected depends (among others things) on the angle of incidence. When the angle of incidence is small (that is, the wave arrives almost perpendicularly) most of the energy traverses the surface and very little is reflected. When the angle of incidence is near 90° (grazing incidence) almost all the wave is reflected.

Most of the electromagnetic waves emitted by an antenna to the ground below the antenna at moderate (say < 60°) angles of incidence enter the earth and are absorbed (lost). But waves emitted to the ground at grazing angles, far from the antenna, are almost totally reflected. At grazing angles, the ground behaves as a mirror. Quality of reflection

depends on the nature of the surface. When the irregularities of the surface are smaller than the wavelength reflection is good.



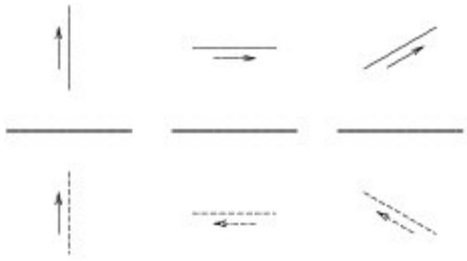
The wave reflected by earth can be considered as emitted by the image antenna

This means that the receptor "sees" the real antenna and, under the ground, the image of the antenna reflected by the ground. If the ground has irregularities, the image will appear fuzzy.

If the receiver is placed at some height above the ground, waves reflected by ground will travel a little longer distance to arrive to the receiver than direct waves. The distance will be the same only if the receiver is close to ground.

In the drawing at right, we have drawn the angle θfar bigger than in reality. Distance between the antenna and its image is d.

The situation is a bit more complex because the reflection of electromagnetic waves depends on the polarization of the incident wave. As the refractive index of the ground (average value $\simeq 2$) is bigger than the refractive index of the air ($\simeq 1$), the direction of the component of the electric field parallel to the ground inverses at the reflection. This is equivalent to a phase shift of πradians or 180°. The vertical component of the electric field reflects without changing direction. This sign inversion of the parallel component and the non-inversion of the perpendicular component would also happen if the ground were a good electrical conductor.



The vertical component of the current reflects without changing sign. The horizontal component reverses sign at reflection.

This means that a receiving antenna "sees" the image antenna with the current in the same direction if the antenna is vertical or with the current inverted if the antenna is horizontal.

For a vertical polarized emission antenna the far electric field of the electromagnetic wave produced by the direct ray plus the reflected ray is:

$$|E_{\perp}| = 2\,|E_{\theta_1}|\,\left|\cos\left(\tfrac{kd}{2}\sin\theta\right)\right|$$

The sign inversion for the parallel field case just changes a cosine to a sine:

$$|E_{=}| = 2\,|E_{\theta_1}|\,\left|\sin\left(\tfrac{kd}{2}\sin\theta\right)\right|$$

In these two equations:

- $E_{\theta 1}$ is the electrical field radiated by the antenna if there were no ground.
- $k=\frac{2\pi}{\lambda}$ is the wave number.
- λ is the wave length.
- d is the distance between antenna and its image (twice the height of the center of the antenna).

Radiation patterns of antennas and their images reflected by the ground. At left the polarization is vertical and there is always a maximum for $\theta=0$. If the polarization is horizontal as at right, there is always a zero for $\theta=0$.

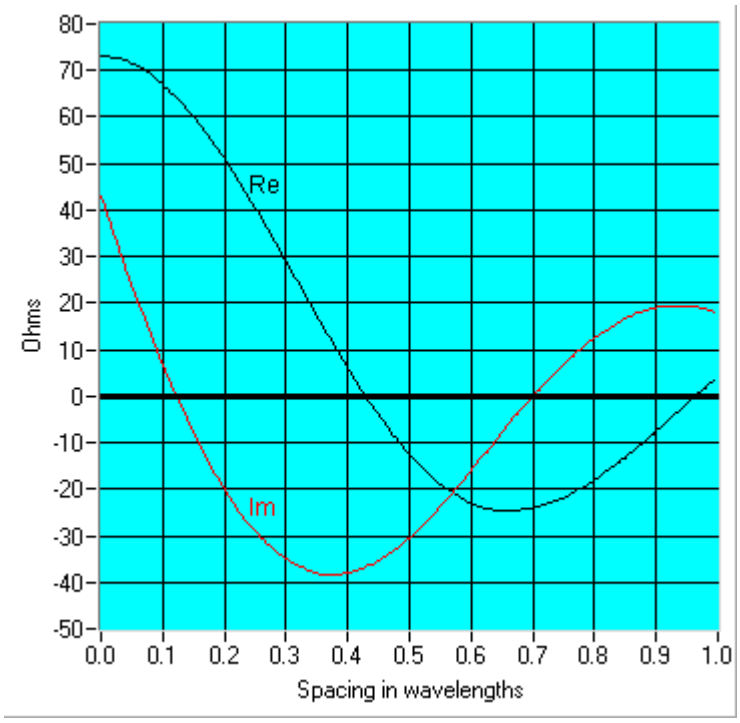
For emitting and receiving antenna situated near the ground (in a building or on a mast) far from each other, distances traveled by direct and reflected rays are nearly the same. There is no induced phase shift. If the emission is polarized vertically the two fields (direct and reflected) add and there is maximum of received signal. If the emission is polarized horizontally the two signals subtracts and the received signal is minimum. This is depicted in the image at right. In the case of vertical polarization, there is always a maximum at earth level (left pattern). For horizontal polarization, there is always a minimum at earth level. Note that in these drawings the ground is considered as a perfect mirror, even for low angles of incidence. In these drawings the distance between the antenna and its image is just a few wavelengths. For greater distances, the number of lobes increases.

Note that the situation is different–and more complex–if reflections in the ionosphere occur. This happens over very long distances (thousands of kilometers). There is not a direct ray but several reflected rays that add with different phase shifts.

This is the reason why almost all public address radio emissions have vertical polarization. As public users are near ground, horizontal polarized emissions would be poorly received. Observe household and automobile radio receivers. They all have vertical antennas or horizontal ferrite antennas for vertical polarized emissions. In cases where the receiving antenna must work in any position, as in mobile phones, the emitter and receivers in base stations use circular polarized electromagnetic waves.

Classical (analog) television emissions are an exception. They are almost always horizontally polarized, because the presence of buildings makes it unlikely that a good emitter antenna image will appear. However, these same buildings reflect the electromagnetic waves and can create ghost images. Using horizontal polarization, reflections are attenuated because of the low reflection of electromagnetic waves whose magnetic field is parallel to the dielectric surface near the Brewster's angle. Vertically polarized analog television has been used in some rural areas. In digital terrestrial television reflections are less obtrusive, due to the inherent robustness of digital signalling and built-in error correction.

Mutual impedance and interaction between antennas



Mutual impedance between parallel $\frac{\lambda}{2}$dipoles not staggered. Curves **Re** and **Im** are the resistive and reactive parts of the impedance.

Current circulating in any antenna induces currents in all others. One can postulate a **mutual impedance** Z_{12}between two antennas that has the same significance as the $j\omega M$in ordinary coupled inductors. The mutual impedance Z_{12}between two antennas is defined as:

$$Z_{12} = \frac{v_2}{i_1}$$

where $i_1$is the current flowing in antenna 1 and $v_2$is the voltage that would have to be applied to antenna 2–with antenna 1 removed–to produce the current in the antenna 2 that was produced by antenna 1.

From this definition, the currents and voltages applied in a set of coupled antennas are:

$$\begin{array}{ccccccccc} v_1 & = & i_1 Z_{11} & + & i_2 Z_{12} & + & \cdots & + & i_n Z_{1n} \\ v_2 & = & i_1 Z_{21} & + & i_2 Z_{22} & + & \cdots & + & i_n Z_{2n} \\ \vdots & & \vdots & & \vdots & & & & \vdots \\ v_n & = & i_1 Z_{n1} & + & i_2 Z_{n2} & + & \cdots & + & i_n Z_{nn} \end{array}$$

where:

- $v_i$is the voltage applied to the antenna *i*
- $Z_{ii}$is the impedance of antenna *i*
- $Z_{ij}$is the mutual impedance between antennas *i* and *j*

Note that, as is the case for mutual inductances,

$$Z_{ij} = Z_{ji}$$

This is a consequence of Lorentz reciprocity. If some of the elements are not fed (there is a short circuit instead a feeder cable), as is the case in television antennas (Yagi-Uda antennas), the corresponding $v_i$are zero. Those elements are called parasitic elements. Parasitic elements are unpowered elements that either reflect or absorb and reradiate RF energy.

In some geometrical settings, the mutual impedance between antennas can be zero. This is the case for crossed dipoles used in circular polarization antennas.

Antenna gallery

Antennas and antenna arrays

A Yagi-Uda beam antenna.

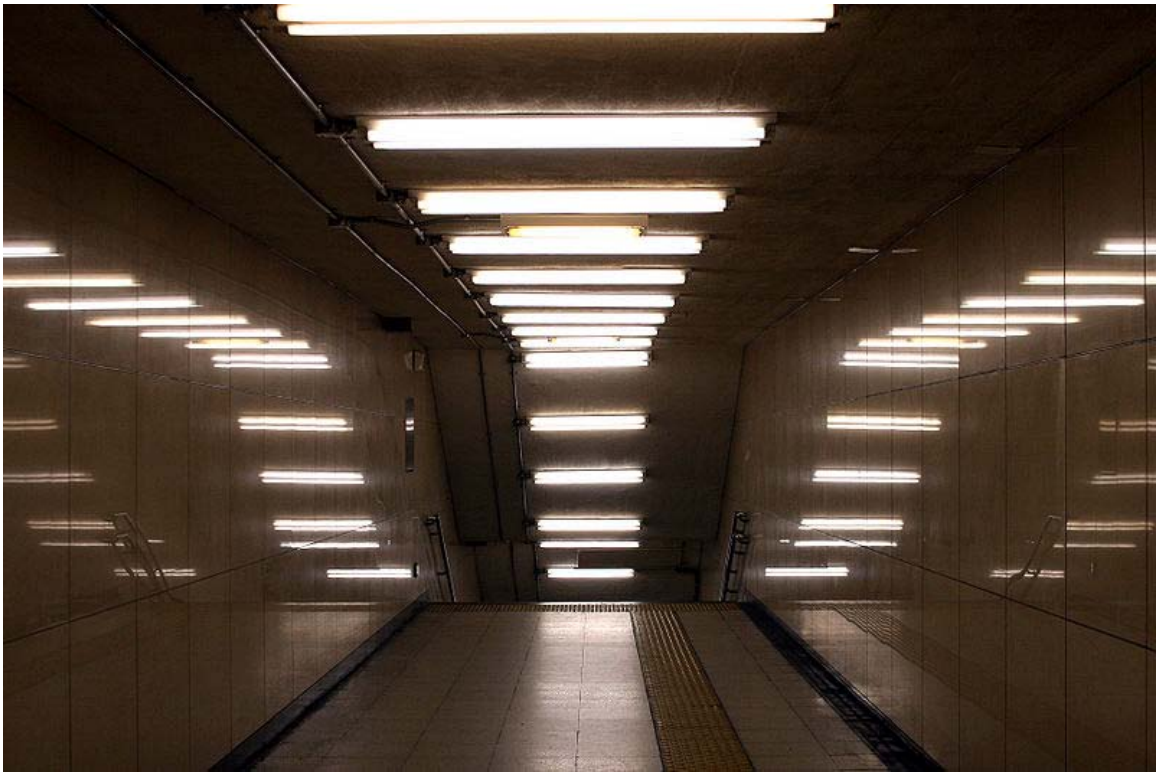
A multi-band rotary directional antenna for amateur radio use.

Rooftop TV antenna. It is actually three Yagi antennas. The longest elements are for the low band, while the medium and short elements are for the high and UHF band.

A terrestrial microwave radio antenna array.

Chapter 2

Fluorescent Lamp



Fluorescent lamps

Assorted types of fluorescent lamps. Top, two compact fluorescent lamps. Bottom, two regular tubes. Left, matchstick shown for scale.

Typical F71T12 100 W bi-pin lamp used in tanning beds. Note the (Hg) symbol indicating it contains mercury. In the US this symbol is now required on all fluorescent bulbs that contain mercury.

Inside the lamp end of a bi-pin lamp

A **fluorescent lamp** or **fluorescent tube** is a gas-discharge lamp that uses electricity to excite mercury vapor. The excited mercury atoms produce short-wave ultraviolet light that then causes a phosphor to fluoresce, producing visible light. A fluorescent lamp converts electrical power into useful light more efficiently than an incandescent lamp. Lower energy cost typically offsets the higher initial cost of the lamp. The lamp is more costly because it requires a ballast to regulate the current through the lamp.

While larger fluorescent lamps have been mostly used in commercial or institutional buildings, the compact fluorescent lamp is now available in the same popular sizes as incandescents and is used as an energy-saving alternative in homes.

History

Physical discoveries

Fluorescence of certain rocks and other substances had been observed for hundreds of years before its nature was understood. By the middle of the 19th century, experimenters had observed a radiant glow emanating from partially evacuated glass vessels through which an electrical current passed. One of the first to explain it about 1845 was the Irish

scientist Sir George G. Stokes from Cambridge University, who named the phenomenon fluorescence after fluorite, a mineral many of whose samples fluoresce strongly due to impurities. The explanation relied on the nature of electricity and light phenomena as developed by the British scientists Michael Faraday and James Clerk Maxwell in the 1840s.

Little more was done with this phenomenon until 1856 when a German glassblower named Heinrich Geissler created a mercury vacuum pump that evacuated a glass tube to an extent not previously possible. When an electrical current passed through a Geissler tube, a strong green glow on the walls of the tube at the cathode end could be observed. Because it produced some beautiful light effects, the Geissler tube was a popular source of amusement. More important, however, was its contribution to scientific research. One of the first scientists to experiment with a Geissler tube was Julius Plücker who systematically described in 1858 the luminescent effects that occurred in a Geissler tube. He also made the important observation that the glow in the tube shifted position when in proximity to an electromagnetic field. Alexandre Edmond Becquerel observed in 1859 that certain substances gave off light when they were placed in a Geissler tube. He went on to apply thin coatings of luminescent materials to the surfaces of these tubes. Fluorescence occurred, but the tubes were very inefficient and had a short operating life.

Inquiries that began with the Geissler tube continued as even better vacuums were produced. The most famous was the evacuated tube used for scientific research by William Crookes. That tube was evacuated by the highly effective mercury vacuum pump created by Hermann Sprengel. Research conducted by Crookes and others ultimately led to the discovery of the electron in 1897 by J. J. Thomson. But the Crookes tube, as it came to be known, produced little light because the vacuum in it was too good and thus lacked the trace amounts of gas that are needed for electrically stimulated luminescence.

Early discharge lamps

While Becquerel was primarily interested in conducting scientific research into fluorescence, Thomas Edison briefly pursued fluorescent lighting for its commercial potential. He invented a fluorescent lamp in 1896 that used a coating of calcium tungstate as the fluorescing substance, excited by X-rays, but although it received a patent in 1907, it was not put into production. As with a few other attempts to use Geissler tubes for illumination, it had a short operating life, and given the success of the incandescent light, Edison had little reason to pursue an alternative means of electrical illumination. Nikola Tesla made similar experiments in the 1890s, devising high frequency powered fluorescent bulbs that gave a bright greenish light, but as with Edison's devices, no commercial success was achieved.

Although Edison lost interest in fluorescent lighting, one of his former employees was able to create a gas-based lamp that achieved a measure of commercial success. In 1895 Daniel McFarlan Moore demonstrated lamps 2 to 3 meters (6.6 to 9.8 ft) in length that

used carbon dioxide or nitrogen to emit white or pink light, respectively. As with future fluorescent lamps, they were considerably more complicated than an incandescent bulb.

After years of work, Moore was able to extend the operating life of the lamps by inventing an electromagnetically controlled valve that maintained a constant gas pressure within the tube. Although Moore's lamp was complicated, expensive to install, and required very high voltages, it was considerably more efficient than incandescent lamps, and it produced a more natural light than incandescent lamps. From 1904 onwards Moore's lighting system was installed in a number of stores and offices. Its success contributed to General Electric's motivation to improve the incandescent lamp, especially its filament. GE's efforts came to fruition with the invention of a tungsten-based filament. The extended lifespan of incandescent bulbs negated one of the key advantages of Moore's lamp, but GE purchased the relevant patents in 1912. These patents and the inventive efforts that supported them were to be of considerable value when the firm took up fluorescent lighting more than two decades later.

At about the same time that Moore was developing his lighting system, another American was creating a means of illumination that also can be seen as a precursor to the modern fluorescent lamp. This was the Mercury-vapor lamp, invented by Peter Cooper Hewitt and patented in 1901 (US patent 889692). Hewitt's lamp luminesced when an electric current was passed through mercury vapor at a low pressure. Unlike Moore's lamps, Hewitt's were manufactured in standardized sizes and operated at low voltages. The mercury-vapor lamp was superior to the incandescent lamps of the time in terms of energy efficiency, but the blue-green light it produced limited its applications. It was, however, used for photography and some industrial processes.

Mercury vapor lamps continued to be developed at a slow pace, especially in Europe, and by the early 1930s they received limited use for large-scale illumination. Some of them employed fluorescent coatings, but these were primarily used for color correction and not for enhanced light output. Mercury vapor lamps also anticipated the fluorescent lamp in their incorporation of a ballast to maintain a constant current.

Cooper-Hewitt had not been the first to use mercury vapor for illumination, as earlier efforts had been mounted by Way, Rapieff, Arons, and Bastian and Salisbury. Of particular importance was the mercury vapor lamp invented by Küch in Germany. This lamp used quartz in place of glass to allow higher operating temperatures, and hence greater efficiency. Although its light output relative to electrical consumption was better than other sources of light, the light it produced was similar to that of the Cooper-Hewitt lamp in that it lacked the red portion of the spectrum, making it unsuitable for ordinary lighting.

Neon lamps

The next step in gas-based lighting took advantage of the luminescent qualities of neon, an inert gas that had been discovered in 1898. In 1909 Georges Claude, a French chemist, observed the red glow that was produced when running an electric current through a

neon-filled tube. He also discovered that argon emitted a blue glow. While neon lighting was used around 1930 in France for general illumination, it was no more energy-efficient than conventional incandescent lighting. Neon lighting came to be used primarily for eye-catching signs and advertisements. Neon lighting was relevant to the development of fluorescent lighting, however, as Claude's improved electrode (patented in 1915) overcame "sputtering", a major source of electrode degradation. Sputtering occurred when ionized particles struck an electrode and tore off bits of metal. Although Claude's invention required electrodes with a lot of surface area, it showed that a major impediment to gas-based lighting could be overcome.

The development of the neon light also was significant for the last key element of the fluorescent lamp, its fluorescent coating. In 1926 Jacques Risler received a French patent for the application of fluorescent coatings to neon light tubes. The main use of these lamps, which can be considered the first commercially successful fluorescents, was for advertising, not general illumination. This, however, was not the first use of fluorescent coatings. As has been noted above, Edison used calcium tungstate for his unsuccessful lamp. Other efforts had been mounted, but all were plagued by low efficiency and various technical problems. Of particular importance was the invention in 1927 of a low-voltage "metal vapor lamp" by Friedrich Meyer, Hans-Joachim Spanner, and Edmund Germer, who were employees of a German firm in Berlin. A German patent was granted but the lamp never went into commercial production.

Commercialization of fluorescent lamps

All the major features of fluorescent lighting were in place at the end of the 1920s. Decades of invention and development had provided the key components of fluorescent lamps: economically manufactured glass tubing, inert gases for filling the tubes, electrical ballasts, long-lasting electrodes, mercury vapor as a source of luminescence, effective means of producing a reliable electrical discharge, and fluorescent coatings that could be energized by ultraviolet light. At this point, intensive development was more important than basic research.

In 1934, Arthur Compton, a renowned physicist and GE consultant, reported to the GE lamp department on successful experiments with fluorescent lighting at General Electric Co., Ltd. in Great Britain (unrelated to General Electric in the United States). Stimulated by this report, and with all of the key elements available, a team led by George E. Inman built a prototype fluorescent lamp in 1934 at General Electric's Nela Park (Ohio) engineering laboratory. This was not a trivial exercise; as noted by Arthur A. Bright, "A great deal of experimentation had to be done on lamp sizes and shapes, cathode construction, gas pressures of both argon and mercury vapor, colors of fluorescent powders, methods of attaching them to the inside of the tube, and other details of the lamp and its auxiliaries before the new device was ready for the public."

In addition to having engineers and technicians along with facilities for R&D work on fluorescent lamps, General Electric controlled what it regarded as the key patents covering fluorescent lighting, including the patents originally issued to Hewitt, Moore,

and Küch. More important than these was a patent covering an electrode that did not disintegrate at the gas pressures that ultimately were employed in fluorescent lamps. This invention had been created by Albert W. Hull of GE's Schenectady Research Laboratory, and was registered as US patent 1790153.

While the Hull patent gave GE a basis for claiming legal rights over the fluorescent lamp, a few months after the lamp went into production the firm learned of a U.S. patent application that had been filed in 1927 for the aforementioned "metal vapor lamp" invented in Germany by Meyer, Spanner, and Germer. The patent application indicated that the lamp had been created as a superior means of producing ultraviolet light, but the application also contained a few statements referring to fluorescent illumination. Efforts to obtain a U.S. patent had met with numerous delays, but were it to be granted, the patent might have caused serious difficulties for GE. At first, GE sought to block the issuance of a patent by claiming that priority should go to one of their employees, Leroy J. Buttolph, who according to their claim had invented a fluorescent lamp in 1919 and whose patent application was still pending. GE also had filed a patent application in 1936 in Inman's name to cover the "improvements" wrought by his group. In 1939 GE decided that the claim of Meyer, Spanner, and Germer had some merit, and that in any event a long interference procedure was not in their best interest. They therefore dropped the Buttolph claim and paid $180,000 to acquire the Meyer, et al. application, which at that point was owned by a firm known as Electrons, Inc. The patent (US patent 2182732) was duly awarded in December 1939. This patent, along with the Hull patent, put GE on what seemed to be firm legal ground, although it faced years of legal challenges from Sylvania Electric Products, Inc., which claimed infringement on patents that it held.

Even though the patent issue would not be completely resolved for many years, General Electric's strength in manufacturing and marketing the bulb gave it a pre-eminent position in the emerging fluorescent light market. Sales of "fluorescent lumiline lamps" commenced in 1938 when four different sizes of tubes were put on the market used in fixtures manufactured by three leading corporations, two based in New York City. During the following year GE and Westinghouse publicized the new lights through exhibitions at the New York World's Fair and the Golden Gate International Exposition in San Francisco. Fluorescent lighting systems spread rapidly during World War II as wartime manufacturing intensified lighting demand. By 1951 more light was produced in the United States by fluorescent lamps than by incandescent lamps.

Principles of operation

The fundamental means for conversion of electrical energy into radiant energy in a fluorescent lamp relies on inelastic scattering of electrons. An incident electron collides with an atom in the gas. If the free electron has enough kinetic energy, it transfers energy to the atom's outer electron, causing that electron to temporarily jump up to a higher energy level. The collision is 'inelastic' because a loss of energy occurs.

This higher energy state is unstable, and the atom will emit an ultraviolet photon as the atom's electron reverts to a lower, more stable, energy level. Most of the photons that are

released from the mercury atoms have wavelengths in the ultraviolet (UV) region of the spectrum predominantly at wavelengths of 253.7 nm and 185 nm. These are not visible to the human eye, so they must be converted into visible light. This is done by making use of fluorescence. Ultraviolet photons are absorbed by electrons in the atoms of the lamp's interior fluorescent coating, causing a similar energy jump, then drop, with emission of a further photon. The photon that is emitted from this second interaction has a lower energy than the one that caused it. The chemicals that make up the phosphor are chosen so that these emitted photons are at wavelengths visible to the human eye. The difference in energy between the absorbed ultra-violet photon and the emitted visible light photon goes toward heating up the phosphor coating.

When the light is turned on, the electric power heats up the cathode enough for it to emit electrons. These electrons collide with and ionize noble gas atoms inside the bulb surrounding the filament to form a plasma by a process of impact ionization. As a result of avalanche ionization, the conductivity of the ionized gas rapidly rises, allowing higher currents to flow through the lamp.

Construction

Close-up of the cathodes of a germicidal lamp (an essentially similar design that uses no fluorescent phosphor, allowing the electrodes to be seen.)

A fluorescent lamp tube is filled with a gas containing low pressure mercury vapor and argon, xenon, neon, or krypton. The pressure inside the lamp is around 0.3% of atmospheric pressure. The inner surface of the bulb is coated with a fluorescent (and often slightly phosphorescent) coating made of varying blends of metallic and rare-earth phosphor salts. The bulb's electrodes are typically made of coiled tungsten and usually referred to as cathodes because of their prime function of emitting electrons. For this, they are coated with a mixture of barium, strontium and calcium oxides chosen to have a low thermionic emission temperature.

The unfiltered ultraviolet glow of a germicidal lamp is produced by a low pressure mercury vapor discharge (identical to that in a fluorescent lamp) in an uncoated fused quartz envelope.

Fluorescent lamp tubes are typically straight and range in length from about 100 millimeters (3.9 in) for miniature lamps, to 2.43 meters (8.0 ft) for high-output lamps. Some lamps have the tube bent into a circle, used for table lamps or other places where a more compact light source is desired. Larger U-shaped lamps are used to provide the same amount of light in a more compact area, and are used for special architectural purposes. Compact fluorescent lamps have several small-diameter tubes joined in a bundle of two, four, or six, or a small diameter tube coiled into a spiral, to provide a high amount of light output in little volume.

Light-emitting phosphors are applied as a paint-like coating to the inside of the tube. The organic solvents are allowed to evaporate, then the tube is heated to nearly the melting point of glass to drive off remaining organic compounds and fuse the coating to the lamp tube. Careful control of the grain size of the suspended phosphors is necessary; large grains, 35 micrometers or larger, lead to weak grainy coatings, whereas too many small particles 1 or 2 micrometers or smaller leads to poor light maintenance and efficiency. Most phosphors perform best with a particle size around 10 micrometers. The coating must be thick enough to capture all the ultraviolet light produced by the mercury arc, but not so thick that the phosphor coating absorbs too much visible light. The first phosphors were synthetic versions of naturally occurring fluorescent minerals, with small amounts of metals added as activators. Later other compounds were discovered, allowing differing colors of lamps to be made.

Electrical aspects of operation

Different ballasts for fluorescent and discharge lamps

Fluorescent lamps are negative differential resistance devices, so as more current flows through them, the electrical resistance of the fluorescent lamp drops, allowing even more current to flow. Connected directly to a constant-voltage power supply, a fluorescent lamp would rapidly self-destruct due to the uncontrolled current flow. To prevent this, fluorescent lamps must use an auxiliary device, a ballast, to regulate the current flow through the tube.

The terminal voltage across an operating lamp varies depending on the arc current, tube diameter, temperature, and fill gas. A fixed part of the voltage drop is due to the electrodes. A general lighting service T12 48 inch (1200 mm) lamp operates at 430 mA, with 100 volts drop. High output lamps operate at 800 mA, and some types operate up to 1500 mA. The power level varies from 10 watts per foot (33 watts per meter) to 25 watts per foot (82 watts per meter) of tube length for T12 lamps.

The simplest ballast for alternating current use is an inductor placed in series, consisting of a winding on a laminated magnetic core. The inductance of this winding limits the flow of AC current. This type is still used, for example, in 120 volt operated desk lamps using relatively short lamps. Ballasts are rated for the size of lamp and power frequency. Where the mains voltage is insufficient to start long fluorescent lamps, the ballast is often a step-up autotransformer with substantial leakage inductance (so as to limit the current flow). Either form of inductive ballast may also include a capacitor for power factor correction.

230 V ballast for 18–20 W

Many different circuits have been used to operate fluorescent lamps. The choice of circuit is based on mains voltage, tube length, initial cost, long term cost, instant versus non-instant starting, temperature ranges and parts availability, etc.

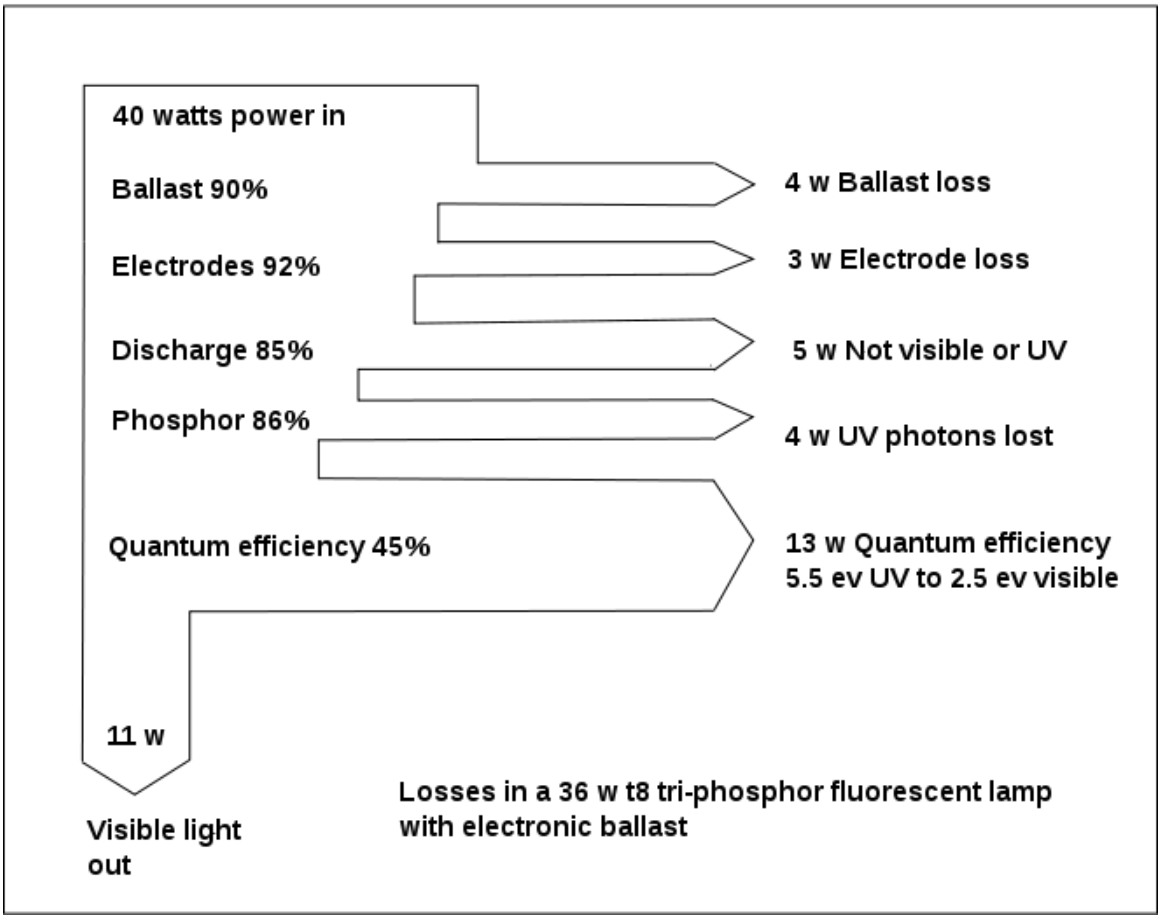
Fluorescent lamps can run directly from a DC supply of sufficient voltage to strike an arc. The ballast must be resistive, and would consume about as much power as the lamp. When operated from DC, the starting switch is often arranged to reverse the polarity of the supply to the lamp each time it is started; otherwise, the mercury accumulates at one end of the tube. Fluorescent lamps are (almost) never operated directly from DC for those reasons. Instead, an inverter converts the DC into AC and provides the current-limiting function as described below for electronic ballasts.

Effect of temperature

The light output and performance of fluorescent lamps is critically affected by the temperature of the bulb wall and its effect on the partial pressure of mercury vapor within the lamp. Each lamp contains a small amount of mercury, which must vaporize to support the lamp current and generate light. At low temperatures the mercury is in the form of dispersed liquid droplets. As the lamp warms, more of the mercury is in vapor form. At higher temperatures, self-absorption in the vapor reduces the yield of UV and visible light. Since mercury condenses at the coolest spot in the lamp, careful design is required to maintain that spot at the optimum temperature, around 40 °C.

By using an amalgam with some other metal, the vapor pressure is reduced and the optimum temperature range extended upward; however, the bulb wall "cold spot" temperature must still be controlled to prevent migration of the mercury out of the amalgam and condensing on the cold spot. Fluorescent lamps intended for higher output will have structural features such as a deformed tube or internal heat-sinks to control cold spot temperature and mercury distribution. Heavily loaded small lamps, such as compact fluorescent lamps, also include heat-sink areas in the tube to maintain mercury vapor pressure at the optimum value.

Losses



A Sankey diagram of energy losses in a fluorescent lamp. In modern designs, the biggest loss is the quantum efficiency of converting high-energy UV photons to lower-energy visible light photons.

The efficiency of fluorescent lighting owes much to the fact that low pressure mercury discharges emit about 65% of their total light in the 254 nm line (another 10–20% of the light is emitted in the 185 nm line). The UV light is absorbed by the bulb's fluorescent coating, which re-radiates the energy at longer wavelengths to emit visible light. The blend of phosphors controls the color of the light, and along with the bulb's glass prevents the harmful UV light from escaping.

Only a fraction of the electrical energy input into a lamp gets turned into useful light. The ballast dissipates some heat; electronic ballasts may be around 90% efficient. A fixed voltage drop occurs at the electrodes. Some of the energy in the mercury vapor column is also dissipated, but about 85% is turned into visible and ultraviolet light.

Not all the UV energy on the phosphor gets converted into visible light. In a modern lamp, for every 100 incident photons of UV impacting the phosphor, only 86 visible light photons are emitted (a quantum efficiency of 86%). The largest single loss in modern lamps is due to the lower energy of each photon of visible light, compared to the energy of the UV photons that generated them. Incident photons have an energy of 5.5 electron volts but produces visible light photons with energy around 2.5 electron volts, so only 45% of the UV energy is used. If a so-called "two-photon" phosphor could be developed, this would improve the efficiency but much research has not yet found such a system.

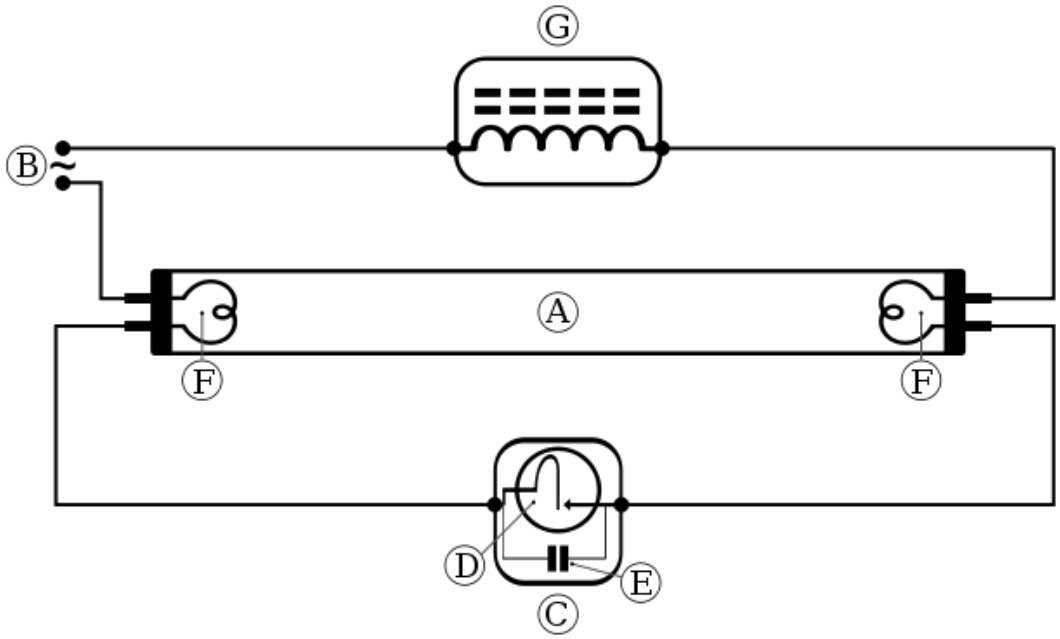
Cold cathode lamps

Most fluorescent lamps use electrodes that operate in thermionic emission mode, meaning they are operated at a high enough temperature for the chosen material (normally a special coating) to liberate electrons across to the gas-fill by heat.

However, there are also tubes that operate in cold cathode mode, whereby electrons are liberated only by the level of potential difference provided. This doesn't mean the electrodes are cold (and indeed, they can be very hot), but it does mean they are operating below their thermionic emission temperature. Because cold cathode lamps have no thermionic emission coating to wear out they can have much longer lives than is commonly available with thermionic emission tubes. This quality makes them desirable for maintenance-free long-life applications (such as LCD backlight displays). Sputtering of the electrode may still occur, but electrodes can be shaped (e.g. into an internal cylinder) to capture most of the sputtered material so it isn't lost from the electrode.

Cold cathode lamps are generally less efficient than thermionic emission lamps because the cathode fall voltage is much higher. The increased fall voltage results in more power dissipation at tube ends, which doesn't contribute to light output. However, this is less significant with longer tubes. The increased power dissipation at tube ends also usually means cold cathode tubes have to be run at a lower loading than their thermionic emission equivalents. Given the higher tube voltage required anyway, these tubes can easily be made long, and even run as series strings. They are better suited for bending into special shapes for lettering and signage, and can also be instantly switched on or off.

Starting



A *preheat* fluorescent lamp circuit using an automatic starting switch. A: Fluorescent tube, B: Power (+220 volts), C: Starter, D: Switch (bi-metallic thermostat), E: Capacitor, F: Filaments, G: Ballast

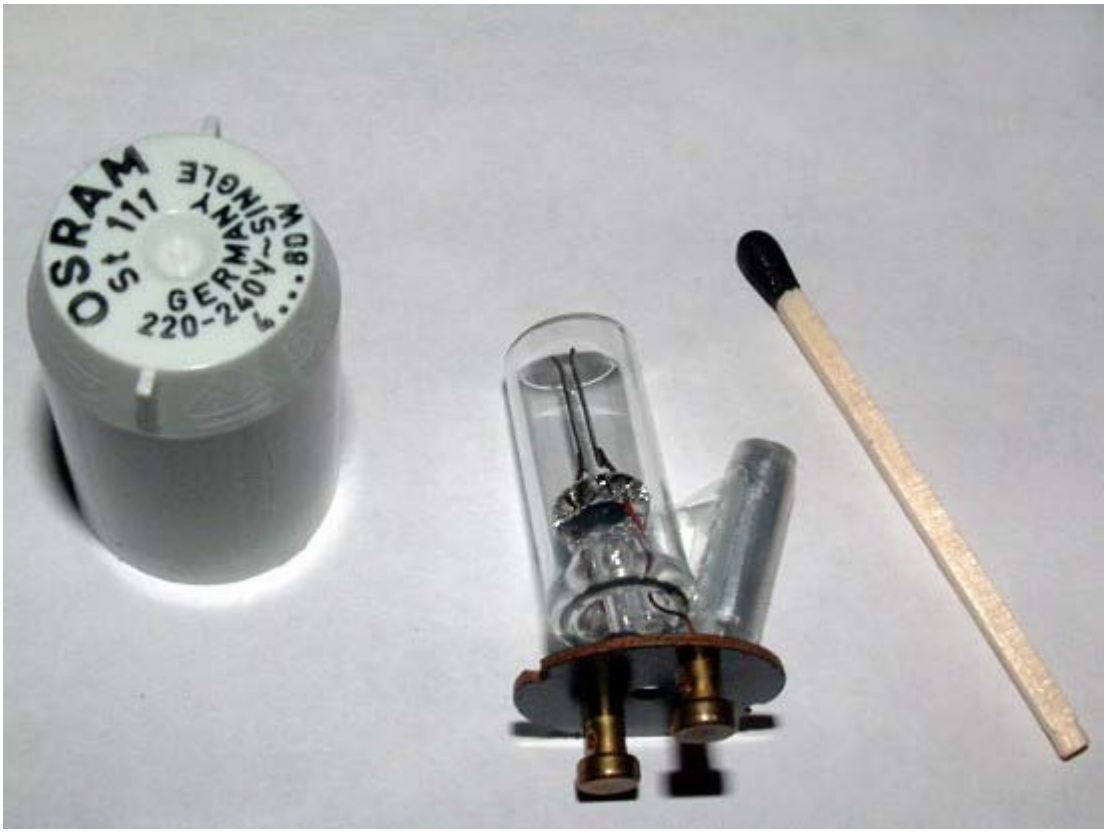
Starting a preheat lamp. The automatic starter switch flashes orange each time it attempts to start the lamp.

The mercury atoms in the fluorescent tube must be ionized before the arc can "strike" within the tube. For small lamps, it does not take much voltage to strike the arc and starting the lamp presents no problem, but larger tubes require a substantial voltage (in the range of a thousand volts).

Switchstart/preheat

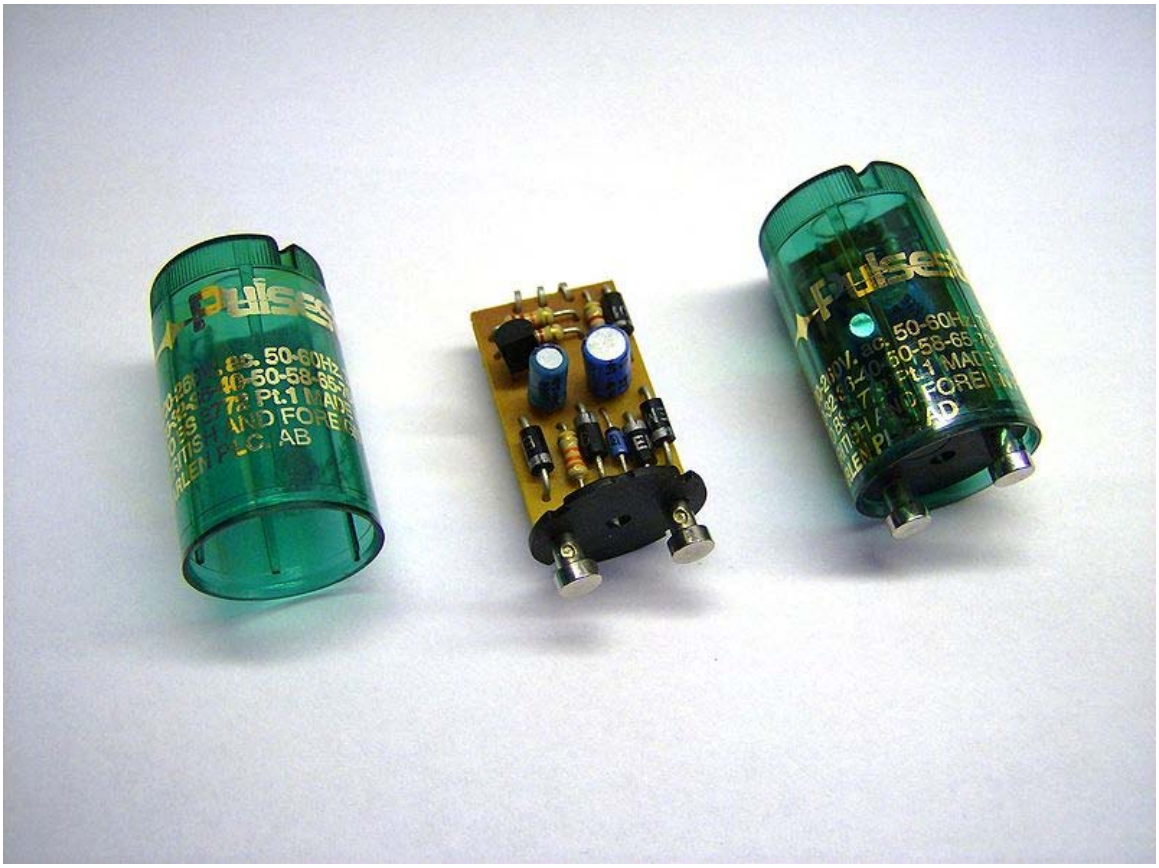
This technique uses a combination filament/cathode at each end of the lamp in conjunction with a mechanical or automatic switch that initially connect the filaments in series with the ballast and thereby preheat the filaments prior to striking the arc. Note that in North America, this is referred to as *Preheat*. Elsewhere this is referred to as *Switchstart*.

These systems are standard equipment in 200–240 V countries (and for 100–120 V lamps up to about 30 watts), and generally use a glow starter. Before the 1960s, four-pin thermal starters and manual switches were also used. Electronic starters are also sometimes used with these electromagnetic ballast lamp fittings.



A *preheat* fluorescent lamp "starter" (automatic starting switch)

The automatic glow starter shown in the photograph to the left consists of a small gas-discharge tube, containing neon and/or argon and fitted with a bi-metallic electrode. The special bi-metallic electrode is the key to the automatic starting mechanism.



Electronic fluorescent lamp starters.

When power is first applied to the lamp circuit, a glow discharge will appear over the electrodes of the starter. This glow discharge will heat the gas in the starter and cause the bi-metallic electrode to bend towards the other electrode. When the electrodes touch, the two filaments of the fluorescent lamp and the ballast will effectively be switched in series to the supply voltage. This causes the filaments to glow and emit electrons into the gas column by thermionic emission. In the starter's tube, the touching electrodes have stopped the glow discharge, causing the gas to cool down again. The bi-metallic electrode also cools down and starts to move back. When the electrodes separate, the inductive kick from the ballast provides the high voltage to start the lamp. The starter additionally has a capacitor wired in parallel to its gas-discharge tube, in order to prolong the electrode life.

Once the tube is struck, the impinging main discharge then keeps the cathode hot, permitting continued emission without the need for the starter to close. The starter does not close again because the voltage across the lit tube is insufficient to start a glow discharge in the starter.

Tube strike is reliable in these systems, but glow starters will often cycle a few times before allowing the tube to stay lit, which causes undesirable flashing during starting. (The older thermal starters behaved better in this respect.)

If the tube fails to strike, or strikes but then extinguishes, the starting sequence is repeated. With automated starters such as glow starters, a failing tube will cycle endlessly, flashing as the lamp quickly goes out because emission is insufficient to keep the lamp current high enough to keep the glow starter open. This causes flickering, and runs the ballast at above design temperature. Some more advanced starters time out in this situation, and do not attempt repeated starts until power is reset. Some older systems used a thermal over-current trip to detect repeated starting attempts. These require manual reset.

Electronic starters use a more complex method to preheat the cathodes of a fluorescent lamp. They commonly use a specially designed semiconductor switch. They are programmed with a predefined preheat time to ensure that the cathodes are fully heated and reduce the amount of sputtered emission mix to prolong the life of the lamp. Electronic starters contain a series of capacitors that are capable of producing a high voltage pulse of electricity across the lamp to ensure that it strikes correctly. Electronic starters only attempt to start a lamp for a short time when power is initially applied and will not repeatedly attempt to restrike a lamp that is dead and cannot sustain an arc. This eliminates the re-striking of a lamp and the cycle of flashing that a failing lamp installed with a glow starter can produce. Electronic starters have also been developed that are capable of striking the fluorescent tube within 0.3 seconds, which gives a virtually instant start.

Instant start

In some cases, a high voltage is applied directly: *instant start* fluorescent tubes simply use a high enough voltage to break down the gas and mercury column and thereby start arc conduction. These tubes can be identified by a single pin at each end of the tube. The lamp holders have a "disconnect" socket at the low-voltage end to isolate the ballast and prevent electric shock. Low-cost lighting fixtures with an integrated electronic ballast use instant start on preheat lamps, even if it reduces the lamp lifespan.

Rapid start

Newer *rapid start* ballast designs provide filament power windings within the ballast; these rapidly and continuously warm the filaments/cathodes using low-voltage AC. No inductive voltage spike is produced for starting, so the lamps must be mounted near a grounded (earthed) reflector to allow the glow discharge to propagate through the tube and initiate the arc discharge. In some lamps a "starting aid" strip of grounded metal is attached to the outside of the lamp glass.

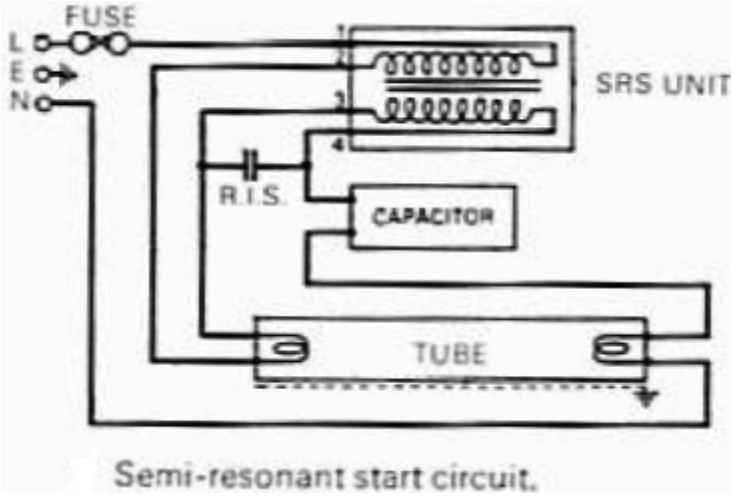
A rapid-start "iron" (magnetic) ballast continually heats the cathodes at the ends of the lamps. This ballast runs two F40T12 lamps in series.

Quick-start

Quick-start ballasts use a small auto-transformer to heat the filaments when power is first applied. When an arc strikes, the filament heating power is reduced and the tube will start within half a second. The auto-transformer is either combined with the ballast or may be a separate unit. Tubes need to be mounted near an earthed metal reflector in order for them to strike. Quick-start ballasts were more common in commercial installations because of lower maintenance as no starter switches need to be replaced. They are also used in domestic installations due to the virtually instant start. Quick-start ballasts are only used on 240 V circuits and are designed for use with the older less efficient T12 tubes, T8 retrofits will not start when used with quick-start ballasts.

Semi-resonant start

A 65W Semi-resonant lamp starting



A circuit diagram of a semi-resonant start fluorescent lamp.

Semi-resonant start was invented by Thorn Lighting for use with T12 fluorescent tubes. This method uses a double wound transformer and a capacitor. With no arc current, the transformer and capacitor ring at mains frequency and generate about twice mains voltage across the tube, and a small electrode heating current. This tube voltage is too low to strike the arc with cold electrodes, but as the electrodes heat up to thermionic emission temperature, the tube striking voltage reduces below that of the ringing voltage, and the arc strikes. As the electrodes heat, the lamp slowly, over 3-5 seconds, reaches full

brightness. As the arc current increases and tube voltage drops, the circuit provides current limiting.

Semi-resonant start was mainly used in commercial installations because of their higher initial cost. There are no starter switches to be replaced and cathode damage is reduced during starting. Due to the high open circuit tube voltage, this starting method was particularly good for starting tubes in cold locations. Additionally, the circuit power factor is almost 1, and no additional power factor correction is needed in the lighting installation. As the design requires that twice the mains voltage must be lower than the cold-cathode striking voltage (or the tubes would erroneously instant-start), this design can only be used with 5ft and longer tubes on 240V mains. Semi-resonant start fixtures are generally incompatible with energy saving T8 retrofit tubes, because such tubes have a higher starting voltage than T12 lamps and may not start reliably,especially in low temperatures. Recent proposals in some countries to phase out T12 tubes will reduce the application of this starting method.

Electronic ballasts

Electronic ballast for fluorescent lamp, 2x58W

Electronic ballasts and different compact fluorescent lamps

Starting a lamp that has an electronic ballast.

Electronic ballasts employ transistors to alter mains voltage frequency into high-frequency AC while also regulating the current flow in the lamp. These ballasts take advantage of the higher efficacy of lamps operated with higher-frequency current. Efficacy of a fluorescent lamp rises by almost 10% at a frequency of 10 kHz, compared to efficacy at normal power frequency. When the AC period is shorter than the relaxation time to de-ionize mercury atoms in the discharge column, the discharge stays closer to

optimum operating condition. Electronic ballasts typically work in rapid start or instant start mode. Electronic ballasts are commonly supplied with AC power, which is internally converted to DC (with Bridge rectifier and Reservoir capacitor) and then back to a variable frequency AC waveform. Depending upon the capacitance and the quality of constant-current pulse-width-modulation, this can largely eliminate modulation at 100 or 120 Hz.

Low cost ballasts mostly contain only a simple oscillator and series resonant LC circuit. When turned on, the oscillator starts, and the LC circuit charges. After a short time the voltage across the lamp reaches about 1 kV and the lamp ignites. The process is too fast to preheat the cathodes, so the lamp instant-starts in cold cathode mode. The cathode filaments are still used for protection of the ballast from overheating if the lamp does not ignite. A few manufacturers use positive temperature coefficient (PTC) thermistors to disable instant starting and give some time to preheat the filaments.

More complex electronic ballasts use programmed start. The output AC frequency is started above the resonance frequency of the output circuit of the ballast; and after the filaments are heated, the frequency is rapidly decreased. If the frequency approaches the resonant frequency of the ballast, the output voltage will increase so much that the lamp will ignite. If the lamp does not ignite, an electronic circuit stops the operation of the ballast.

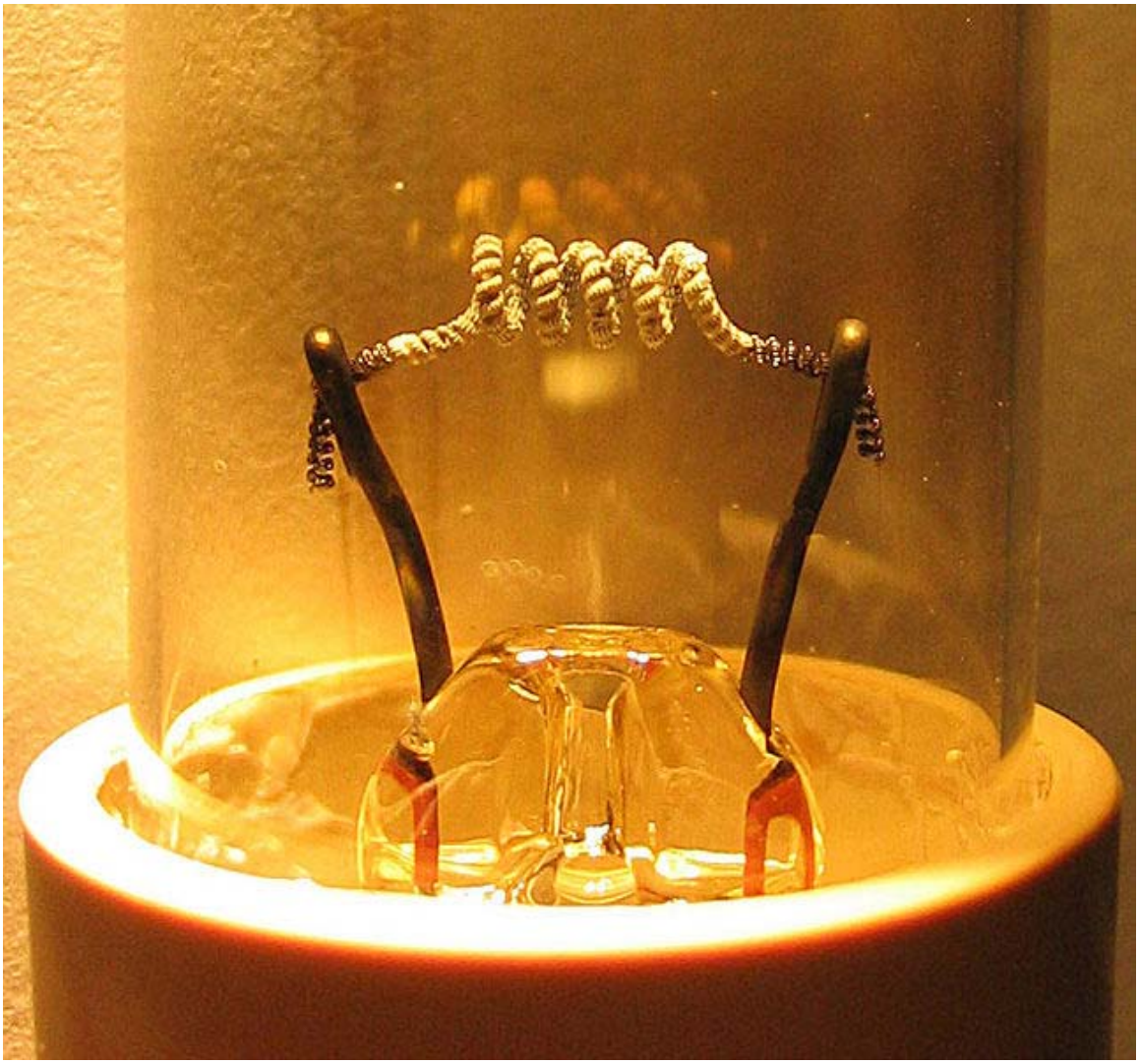
Nowadays, many electronic ballasts are controlled by a PIC microcontroller or similar, and these are sometimes called digital ballasts. Digital ballasts can apply quite complex logic to lamp starting and operation. This enables functions such as testing for broken electrodes and missing tubes before attempting to start, auto detect tube replacement, and auto detection of tube type, such that a single ballast can be used with several different tubes, even those that operate at different arc currents, etc. Once such fine grained control over the starting and arc current is achievable, features such as dimming, and having the ballast maintain a constant light level against changing sunlight contribution are all easily included in the embedded microcontroller software, and can be found in various manufacturers' products.

Since introduction in the 1990s, high frequency ballasts have been used in general lighting fixtures with either rapid start or pre-heat lamps. These ballasts convert the incoming power to an output frequency in excess of 20 kHz. This increases lamp efficiency. These are used in several applications, including new generation tanning lamp systems, whereby a 100 watt lamp (e.g., F71T12BP) can be lit using 65 to 70 watts of actual power while obtaining the same luminous flux (measured in lumens) as magnetic ballasts. These ballasts operate with voltages that can be almost 600 volts, requiring some consideration in housing design, and can cause a minor limitation in the length of the wire leads from the ballast to the lamp ends.

End of life

The end of life failure mode for fluorescent lamps varies depending how they are used and their control gear type. Often the light will turn pink, with black burns on the ends of the bulb, due to sputtering. The lamp may also flicker at a noticeable rate.

Emission mix



Closeup of the filament on a low pressure mercury gas discharge lamp showing white thermionic emission mix coating on the central portion of the coil acting as hot cathode. Typically made of a mixture of barium, strontium and calcium oxides, the coating is sputtered away through normal use, often eventually resulting in lamp failure.

The "emission mix" on the tube filaments/cathodes is necessary to enable electrons to pass into the gas via thermionic emission at the tube operating voltages used. The mix is slowly sputtered off by bombardment with electrons and mercury ions during operation,

but a larger amount is sputtered off each time the tube is started with cold cathodes. The method of starting the lamp has a significant impact on this. Lamps operated for typically less than 3 hours each switch-on will normally run out of the emission mix before other parts of the lamp fail. The sputtered emission mix forms the dark marks at the tube ends seen in old tubes. When all the emission mix is gone, the cathode cannot pass sufficient electrons into the gas fill to maintain the discharge at the designed tube operating voltage. Ideally, the control gear should shut down the tube when this happens. However, some control gear will provide sufficient increased voltage to continue operating the tube in cold cathode mode, which will cause overheating of the tube end and rapid disintegration of the electrodes and their support wires until they are completely gone or the glass cracks, wrecking the low pressure gas fill and stopping the gas discharge.

Burned filaments

The filaments can burn at the end of the lamp's lifetime, opening the circuit and losing the capability to heat up. Both filaments lose function as they are connected in series, with just a simple switch start circuit a broken filament will render the lamp completely useless. Filaments rarely burn or fail open circuit unless the filament becomes depleted of emitter and the control gear is able to supply a high enough voltage across the tube to operate it in cold cathode mode. Some digital electronic ballasts are capable of detecting broken filaments and can still strike an arc with one or both filaments broken providing there is still sufficient emitter.

Ballast electronics

This may occur in compact fluorescent lamps with integral electrical ballasts or in linear lamps. Ballast electronics failure is a somewhat random process that follows the standard failure profile for any electronic device. There is an initial small peak of early failures, followed by a drop and steady increase over lamp life. Life of electronics is heavily dependent on operating temperature—it typically halves for each 10 °C temperature rise. The quoted average life of a lamp is usually at 25 °C ambient (this may vary by country). The average life of the electronics at this temperature is normally greater than this, so at this temperature, not many lamps will fail due to failure of the electronics. In some fittings, the ambient temperature could be well above this, in which case failure of the electronics may become the predominant failure mechanism. Similarly, running a compact fluorescent lamp base-up will result in hotter electronics, which can cause shorter average life (particularly with higher power rated ones). Electronic ballasts should be designed to shut down the tube when the emission mix runs out as described above. In the case of integral electronic ballasts, since they never have to work again, this is sometimes done by having them deliberately burn out some component to permanently cease operation.

In most CFLs the filaments are connected in series, with a small capacitor between them. The discharge, once lit, is in parallel to the capacitor and presents a lower-resistance path, effectively shorting the capacitor out. One of the most common failure modes of cheap

lamps is caused by underrating this capacitor (using lower-voltage, lower-cost part), which is very stressed during operation, leading to its premature failure.

Phosphor

The phosphor drops off in efficiency during use. By around 25,000 operating hours, it will typically be half the brightness of a new lamp (although some manufacturers claim much longer half-lives for their lamps). Lamps that do not suffer failures of the emission mix or integral ballast electronics will eventually develop this failure mode. They still work, but have become dim and inefficient. The process is slow, and often only becomes obvious when a new lamp is operating next to an old one.

Loss of mercury

Like in all mercury-based gas-filled tubes, mercury is slowly absorbed into glass, phosphor, and tube electrodes throughout the lamp life, where it can no longer function. Newer lamps now have just enough mercury to last the expected life of the lamp. Loss of mercury will take over from failure of the phosphor in some lamps. The failure symptoms are similar, except loss of mercury initially causes an extended run-up time to full light output, and finally causes the lamp to glow a dim pink when the mercury runs out and the argon base gas takes over as the primary discharge.

Subjecting the tube to asymmetric waveforms, where the total current flow through the tube does not cancel out and the tube effectively operates under a DC bias, causes asymmetric distribution of mercury ions along the tube due to cataphoresis. The localized depletion of mercury vapor pressure manifests as pink luminescence of the base gas in the vicinity of one of the electrodes, and the operating lifetime of the lamp may be dramatically shortened. This can be an issue with some poorly designed inverters.

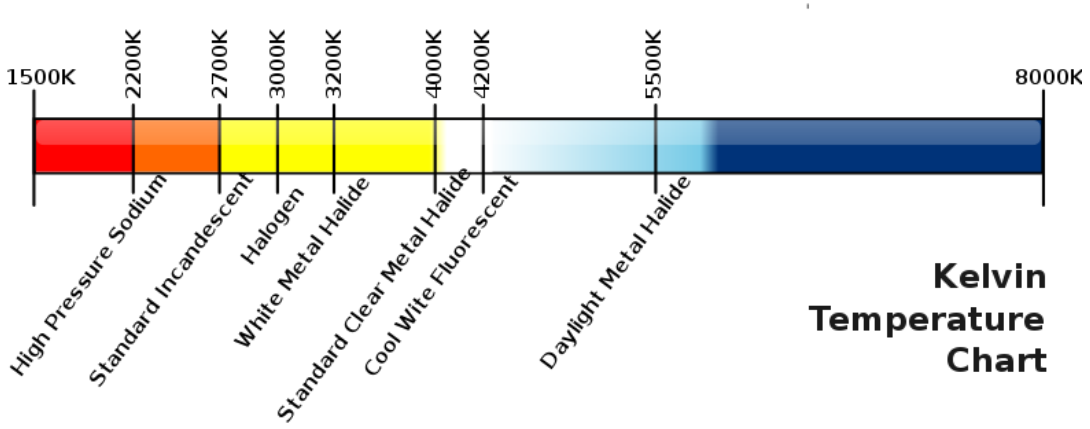
The same effect can be observed with new tubes. Mercury is present in the form of an amalgam and takes some time to be liberated in sufficient amount. New lamps may initially glow pink for several seconds after startup. This period is minimized after about first 100 hours of operation.

Phosphors and the spectrum of emitted light

Light from a fluorescent tube lamp diffracted by a CD shows the individual bands of color.

The spectrum of light emitted from a fluorescent lamp is the combination of light directly emitted by the mercury vapor, and light emitted by the phosphorescent coating. The spectral lines from the mercury emission and the phosphorescence effect give a combined spectral distribution of light that is different from those produced by incandescent sources. The relative intensity of light emitted in each narrow band of wavelengths over the visible spectrum is in different proportions compared to that of an incandescent source. Colored objects are perceived differently under light sources with differing spectral distributions. For example, some people find the color rendition produced by some fluorescent lamps to be harsh and displeasing. A healthy person can sometimes appear to have an unhealthy skin tone under fluorescent lighting. The extent to which this phenomenon occurs is related to the light's spectral composition, and may be gauged by its color rendering index (CRI).

Color temperature



The color temperature of different electric lamps

Correlated color temperature (CCT) is a measure of the "shade" of whiteness of a light source, again by comparison with a blackbody. Typical incandescent lighting is 2700 K, which is yellowish-white. Halogen lighting is 3000 K. Fluorescent lamps are manufactured to a chosen CCT by altering the mixture of phosphors inside the tube. Warm-white fluorescents have CCT of 2700 K and are popular for residential lighting. Neutral-white fluorescents have a CCT of 3000 K or 3500 K. Cool-white fluorescents have a CCT of 4100 K and are popular for office lighting. Daylight fluorescents have a CCT of 5000 K to 6500 K, which is bluish-white.

High CCT lighting generally requires higher light levels. At dimmer illumination levels, the human eye perceives lower color temperatures as more natural, as related through the Kruithof curve. So, a dim 2700 K incandescent lamp appears natural and a bright 5000 K lamp also appears natural, but a dim 5000 K fluorescent lamp appears too pale. Daylight-type fluorescents look natural only if they are very bright.

Color rendering index

Color rendering index (CRI) is a measure of how well colors can be perceived using light from a source, relative to light from a reference source such as daylight or a blackbody of the same color temperature. By definition, an incandescent lamp has a CRI of 100. Real-life fluorescent tubes achieve CRIs of anywhere from 50 to 99. Fluorescent lamps with low CRI have phosphors that emit too little red light. Skin appears less pink, and hence "unhealthy" compared with incandescent lighting. Colored objects appear muted. For example, a low CRI 6800 K halophosphate tube (an extreme example) will make reds appear dull red or even brown. Since the eye is relatively less efficient at detecting red light, an improvement in color rendering index, with increased energy in the red part of the spectrum, may reduce the overall luminous efficacy.

Lighting arrangements use fluorescent tubes in an assortment of tints of white. Sometimes this is because of the lack of appreciation for the difference or importance of differing tube types. Mixing tube types within fittings can improve the color reproduction of lower quality tubes.

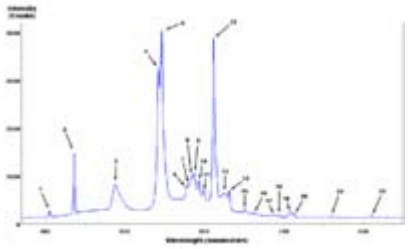
Phosphor composition

Some of the least pleasant light comes from tubes containing the older halophosphate type phosphors (chemical formula $Ca_5(PO_4)_3(F, Cl):Sb^{3+}, Mn^{2+}$). This phosphor mainly emits yellow and blue light, and relatively little green and red. In the absence of a reference, this mixture appears white to the eye, but the light has an incomplete spectrum. The CRI of such lamps is around 60.

Since the 1990s, higher quality fluorescent lamps use either a higher CRI halophosphate coating, or a *triphosphor* mixture, based on europium and terbium ions, that have emission bands more evenly distributed over the spectrum of visible light. High CRI halophosphate and triphosphor tubes give a more natural color reproduction to the human eye. The CRI of such lamps is typically 82–100.

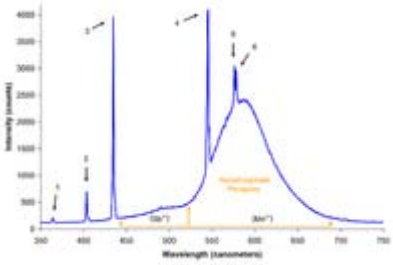
Fluorescent lamp spectra

Typical fluorescent lamp with "rare earth" phosphor



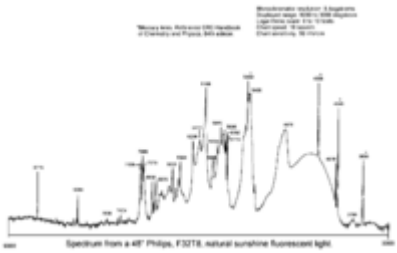
A typical "cool white" fluorescent lamp utilizing two rare earth doped phosphors, Tb^{3+}, Ce^{3+}:$LaPO_4$ for green and blue emission and Eu:Y_2O_3 for red. For an explanation of the origin of the individual peaks.
Note that several of the spectral peaks are directly generated from the mercury arc. This is likely the most common type of fluorescent lamp in use today.

An older style halophosphate phosphor fluorescent lamp



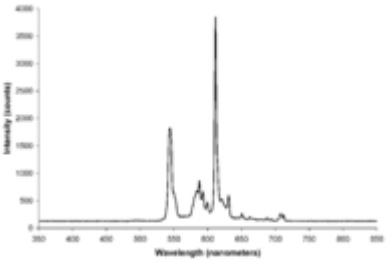
Halophosphate phosphors in these lamps usually consist of trivalent antimony and divalent manganese doped calcium halophosphate ($Ca_5(PO_4)_3(Cl, F):Sb^{3+}, Mn^{2+}$). The color of the light output can be adjusted by altering the ratio of the blue emitting antimony dopant and orange emitting manganese dopant. The color rendering ability of these older style lamps is quite poor. Halophosphate phosphors were invented by A.H. McKeag *et al.* in 1942.

"Natural sunshine" fluorescent light



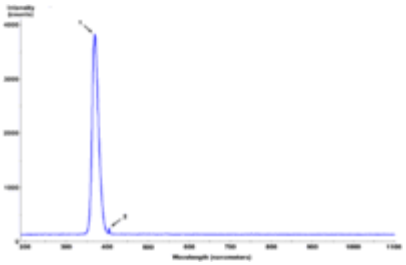
An explanation of the origin of the peaks is on the image page.

Yellow fluorescent lights



The spectrum is nearly identical to a normal fluorescent bulb except for a near total lack of light below 500 nanometers. This effect can be achieved through either specialized phosphor use or more commonly by the use of a simple yellow light filter. These lamps are commonly used as lighting for photolithography work in cleanrooms and as "bug repellent" outdoor lighting (the efficacy of which is questionable).

Spectrum of a "blacklight" bulb



There is typically only one phosphor present in a blacklight bulb, usually consisting of europium-doped strontium fluoroborate, which is contained in an envelope of Wood's glass.

Applications

Fluorescent light bulbs come in many shapes and sizes. The compact fluorescent light bulb (CFL) is becoming more popular. Many compact fluorescent lamps integrate the auxiliary electronics into the base of the lamp, allowing them to fit into a regular light bulb socket.

In US residences, fluorescent lamps are mostly found in kitchens, basements, or garages, but schools and businesses find the cost savings of fluorescent lamps to be significant and rarely use incandescent lights. Tax incentives and environmental awareness result in higher use in places such as California.

In other countries, residential use of fluorescent lighting varies depending on the price of energy, financial and environmental concerns of the local population, and acceptability of the light output. In East and Southeast Asia it is very rare to see incandescent bulbs in buildings anywhere.

Some countries are encouraging the phase-out of incandescent light bulbs and substitution of incandescent lamps with fluorescent lamps or other types of energy-efficient lamps.

The newest fluorescent lamps can be used to grow indoor plants to maturity. These lamps are marketed as High-Output T5 Fluorescents. The T8 and T12 predecessors can be used to rear seedlings, but are not powerful enough for mature plant growth.

In addition to general lighting, special fluorescent lights are often used in stage lighting for film and video production. They are cooler than traditional halogen light sources, and use high-frequency ballasts to prevent video flickering and high color-rendition index bulbs to approximate daylight color temperatures.

Advantages

Luminous efficacy

Fluorescent lamps convert more of the input power to visible light than incandescent lamps. A typical 100 watt tungsten filament incandescent lamp may convert only 2% of its power input to visible white light, whereas typical fluorescent lamps convert about 22% of the power input to visible white light.

The efficacy of fluorescent tubes ranges from about 16 lumens per watt for a 4 watt tube with an ordinary ballast to over 100 lumens per watt with modern electronic ballast, commonly averaging 50 to 67 lm/W overall. Most compact fluorescents above 13 watts with integral electronic ballasts achieve about 60 lm/W. Lamps are rated by lumens after 100 hours of operation. For a given fluorescent tube, a high-frequency electronic ballast gives about 10% efficacy improvement over an inductive ballast. It is necessary to include the ballast loss when evaluating the efficacy of a fluorescent lamp system; this can be about 25% of the lamp power with magnetic ballasts, and around 10% with electronic ballasts.

Fluorescent lamp efficacy is dependent on lamp temperature at the coldest part of the lamp. In T8 lamps this is in the center of the tube. In T5 lamps this is at the end of the tube with the text stamped on it. The ideal temperature for a T8 lamp is 25 °C (77 °F) while the T5 lamp is ideally at 35 °C (95 °F).

Life

Typically a fluorescent lamp will last between 10 to 20 times as long as an equivalent incandescent lamp when operated several hours at a time.

The higher initial cost of a fluorescent lamp is usually more than compensated for by lower energy consumption over its life. The longer life may also reduce lamp replacement costs, providing additional saving especially where labour is costly.

Therefore they are widely used by businesses and institutions, but not as much by households.

Lower luminosity

Compared with an incandescent lamp, a fluorescent tube is a more diffuse and physically larger light source. In suitably designed lamps, light can be more evenly distributed without point source of glare such as seen from an undiffused incandescent filament; the lamp is large compared to the typical distance between lamp and illuminated surfaces.

Lower heat

About two-thirds to three-quarters less heat is given off by fluorescent lamps compared to an equivalent installation of incandescent lamps. This greatly reduces the size, cost, and energy consumption of air-conditioning equipment.

Disadvantages

Frequent switching

If the lamp is installed where it is frequently switched on and off, it will age rapidly. Under extreme conditions, its lifespan may be much shorter than a cheap incandescent lamp. Each start cycle slightly erodes the electron-emitting surface of the cathodes; when all the emission material is gone, the lamp cannot start with the available ballast voltage. Fixtures intended for flashing of lights (such as for advertising) will use a ballast that maintains cathode temperature when the arc is off, preserving the life of the lamp.

Health and safety issues

If a fluorescent lamp is broken, a very small amount of mercury can contaminate the surrounding environment. About 99% of the mercury is typically contained in the phosphor, especially on lamps that are near their end of life. The broken glass is usually considered a greater hazard than the small amount of spilled mercury. The EPA recommends airing out the location of a fluorescent tube break and using wet paper towels to help pick up the broken glass and fine particles. Any glass and used towels should be disposed of in a sealed plastic bag. Vacuum cleaners can cause the particles to become airborne, and should not be used.

Ultraviolet emission

Fluorescent lamps emit a small amount of ultraviolet (UV) light. A 1993 study in the US found that UV exposure from sitting under fluorescent lights for eight hours is equivalent to only one minute of sun exposure. Very sensitive individuals may experience a variety of health problems relating to light sensitivity that is aggravated by artificial lighting.

UV light can affect sensitive paintings, especially watercolors and many textiles. Valuable art work must be protected from light by additional glass or transparent acrylic sheets put between the lamp(s) and the painting.

Ballast

Magnetic single-lamp ballasts have a low power factor

Fluorescent lamps require a ballast to stabilize the current through the lamp, and to provide the initial striking voltage required to start the arc discharge. This increases the cost of fluorescent light fixtures, though often one ballast is shared between two or more lamps. Electromagnetic ballasts with a minor fault can produce an audible humming or buzzing noise. Magnetic ballasts are usually filled with a tar-like potting compound to reduce emitted noise. Hum is eliminated in lamps with a high-frequency electronic ballast. Energy lost in magnetic ballasts can be significant, on the order of 10% of lamp input power. Electronic ballasts reduce this loss.

Power quality and radio interference

Simple inductive fluorescent lamp ballasts have a power factor of less than unity. Inductive ballasts include power factor correction capacitors. Simple electronic ballasts may also have low power factor due to their rectifier input stage.

Fluorescent lamps are a non-linear load and generate harmonic currents in the electrical power supply. The arc within the lamp may generate radio frequency noise, which can be conducted through power wiring. Suppression of radio interference is possible. Very good suppression is possible, but adds to the cost of the fluorescent fixtures.

Operating temperature

Fluorescent lamps operate best around room temperature. At much lower or higher temperatures, efficiency decreases. At below-freezing temperatures standard lamps may not start. Special lamps may be needed for reliable service outdoors in cold weather. In applications such as road and railway signalling, fluorescent lamps which do not generate as much heat as incandescent lamps may not melt snow and ice build up around the lamp, leading to reduced visibility.

Lamp shape

Fluorescent tubes are long, low-luminance sources compared with high pressure arc lamps and incandescent lamps. However, low luminous intensity of the emitting surface is useful because it reduces glare. Lamp fixture design must control light from a long tube instead of a compact globe.

The compact fluorescent lamp (CFL) replaces regular incandescent bulbs. However, some CFLs will not fit some lamps, because the harp (heavy wire shade support bracket) is shaped for the narrow neck of an incandescent lamp. CFLs tend to have a wide housing for their electronic ballast close to the bulb's base, and so may not fit some lamps.

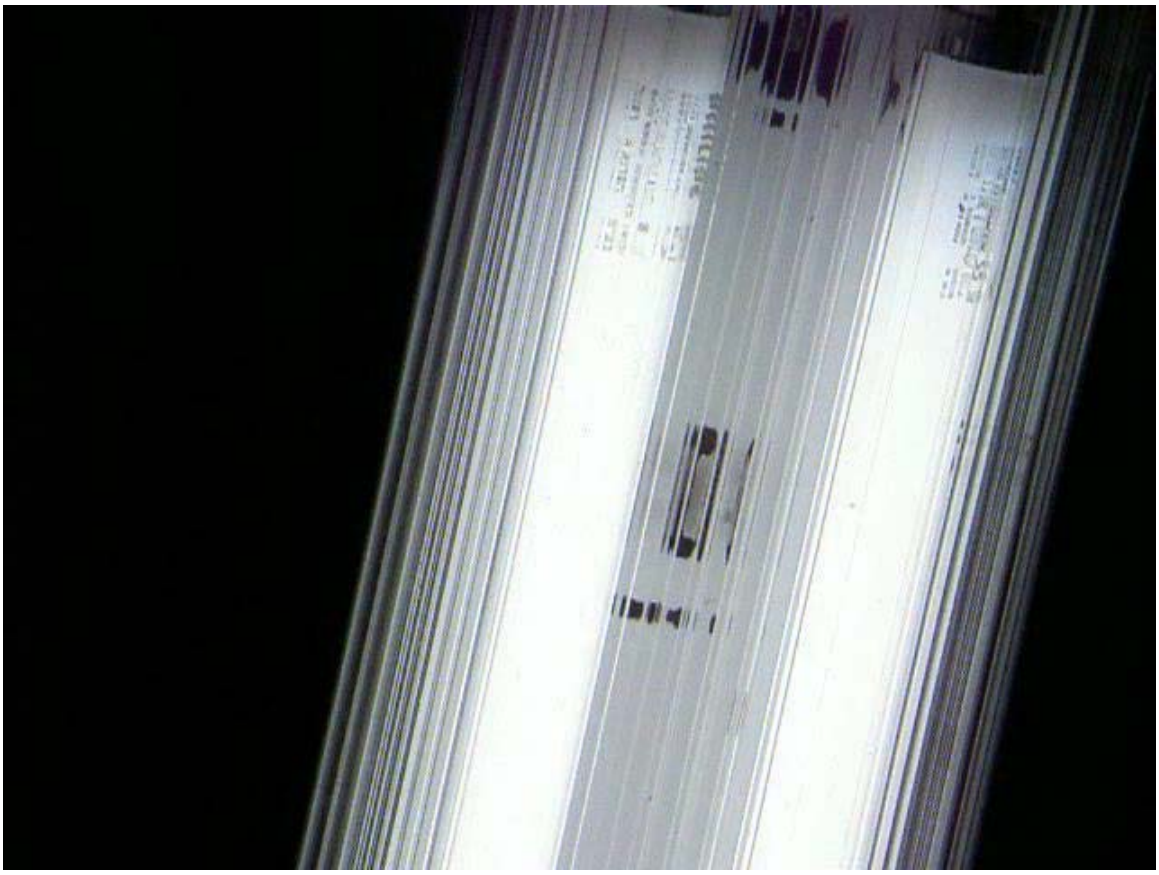
Flicker problems

The "beat effect" problem created when shooting photos or film under standard fluorescent lighting.

Fluorescent lamps using a magnetic mains frequency ballast do not give out a steady light; instead, they flicker at twice the supply frequency. This results in fluctuations not only with light output but color temperature as well, which may pose problems for photography and people who are sensitive to the flicker. Even among persons not sensitive to light flicker, a stroboscopic effect can be noticed, where something spinning at just the right speed may appear stationary if illuminated solely by a single fluorescent lamp. This effect is eliminated by paired lamps operating on a lead-lag ballast. Unlike a true strobe lamp, the light level drops in appreciable time and so substantial "blurring" of the moving part would be evident.

In some circumstances, fluorescent lamps operated at mains frequency can also produce flicker at the mains frequency (50 or 60 Hz) itself, which is noticeable by more people. This can happen in the last few hours of tube life when the cathode emission coating at one end is almost run out, and that cathode starts having difficulty emitting enough electrons into the gas fill, resulting in slight rectification and hence uneven light output in positive and negative going mains cycles. Mains frequency flicker can also sometimes be emitted from the very ends of the tubes, if each tube electrode produces slightly different

light output pattern on each half-cycle. Flicker at mains frequency is more noticeable in the peripheral vision than it is in the center of gaze.



New fluorescent lamps may show a twisting spiral pattern of light in a part of the lamp. This effect is due to loose cathode material and usually disappears after a few hours of operation.

Electromagnetic ballasts may also cause problems for video recording as there can be a 'beat effect' between the periodic reading of a camera's sensor and the fluctuations in intensity of the fluorescent lamp.

Fluorescent lamps using high-frequency electronic ballasts do not produce visible light flicker, since above about 5 kHz, the excited electron state half-life is longer than a half cycle, and light production becomes continuous. Operating frequencies of electronic ballasts are selected to avoid interference with infrared remote controls. Poor quality (or failing) electronic ballasts may have insufficient reservoir capacitance or have poor regulation, thereby producing considerable 100/120 Hz modulation of the light.

Dimming

Lighting Systems can save up to 85% compared to fittings with magnetic ballasts

Fluorescent light fixtures cannot be connected to the same dimmer switch used for incandescent lamps. Two effects are responsible for this: the waveshape of the voltage emitted by a standard phase-control dimmer interacts badly with many ballasts, and it becomes difficult to sustain an arc in the fluorescent tube at low power levels. Dimming installations require a compatible dimming ballast. These systems keep the cathodes of the fluorescent tube fully heated even as the arc current is reduced, promoting easy thermionic emission of electrons into the arc stream. CFLs are available that work in conjunction with a suitable dimmer.

Disposal and recycling

The disposal of phosphor and particularly the toxic mercury in the tubes is an environmental issue. Governmental regulations in many areas require special disposal of fluorescent lamps separate from general and household wastes. For large commercial or

industrial users of fluorescent lights, recycling services are available in many nations, and may be required by regulation. In some areas, recycling is also available to consumers.

Lamp sizes and designations

Systematic nomenclature identifies mass-market lamps as to general shape, power rating, length, color, and other electrical and illuminating characteristics.

Other fluorescent lamps

Black lights
: Blacklights are a subset of fluorescent lamps that are used to provide near ultraviolet light (at about 360 nm wavelength). They are built in the same fashion as conventional fluorescent lamps but the glass tube is coated with a phosphor that converts the short-wave UV within the tube to long-wave UV rather than to visible light. They are used to provoke fluorescence (to provide dramatic effects using blacklight paint and to detect materials such as urine and certain dyes that would be invisible in visible light) as well as to attract insects to bug zappers. So-called *blacklite blue* lamps are also made from more expensive deep purple glass known as Wood's glass rather than clear glass. The deep purple glass filters out most of the visible colors of light directly emitted by the mercury-vapor discharge, producing proportionally less visible light compared with UV light. This allows UV-induced fluorescence to be seen more easily (thereby allowing blacklight posters to seem much more dramatic). The blacklight lamps used in bug zappers do not require this refinement so it is usually omitted in the interest of cost; they are called simply *blacklite* (and not blacklite blue).

Tanning lamps
: The lamps used in tanning beds contain a different phosphor blend (typically 3 to 5 or more phosphors) that emits both UVA and UVB, provoking a tanning response in most human skin. Typically, the output is rated as 3% to 10% UVB (5% most typical) with the remaining UV as UVA. These are mainly F71, F72 or F73 HO (100 W) lamps, although 160 W VHO are somewhat common. One common phosphor used in these lamps is lead-activated barium disilicate, but a europium-activated strontium fluoroborate is also used. Early lamps used thallium as an activator, but emissions of thallium during manufacture were toxic.

Grow lamps
: Grow lamps contain phosphor blends that encourage photosynthesis, growth, and/or flowering in plants, algae, photosynthetic bacteria, and other light-dependent organisms. These often emit light in the red and blue color range, which is absorbed by chlorophyll and used for photosynthesis in plants.

Infrared lamps
: Lamps can be made with a lithium metaluminate phosphor activated with iron. This phosphor has peak emissions between 675 and 875 nanometers, with lesser emissions in the deep red part of the visible spectrum.

Bilirubin lamps

Deep blue light generated from a europium-activated phosphor is used in the light therapy treatment of jaundice; light of this color penetrates skin and helps in the break up of excess bilirubin.

Germicidal lamps

Germicidal lamps depend on the property that UV light kills most germs. Germicidal lamps contain no phosphor at all (technically making them gas discharge lamps rather than fluorescent) and their tubes are made of fused quartz that is transparent to the UV light emitted by the mercury discharge. The UV emitted by these tubes will kill germs and ionize oxygen to ozone. In addition it can cause eye and skin damage and should not be used or observed without eye and skin protection. Besides their uses to kill germs and create ozone, they are sometimes used by geologists to identify certain species of minerals by the color of their fluorescence. When used in this fashion, they are fitted with filters in the same way as blacklight-blue lamps are; the filter passes the short-wave UV and blocks the visible light produced by the mercury discharge. They are also used in some EPROM erasers.

Germicidal lamps have designations beginning with G (meaning 'Germicidal'), rather than F, for example G30T8 for a 30-watt, 1-inch (2.5 cm) diameter, 36-inch (91 cm) long germicidal lamp (as opposed to an F30T8, which would be the fluorescent lamp of the same size and rating).

Electrodeless lamps

Electrodeless induction lamps are fluorescent lamps without internal electrodes. They have been commercially available since 1990. A current is induced into the gas column using electromagnetic induction. Because the electrodes are usually the life-limiting element of fluorescent lamps, such electrodeless lamps can have a very long service life, although they also have a higher purchase price.

Cold-cathode fluorescent lamps (CCFL)

Cold-cathode fluorescent lamps are used as backlighting for LCD displays in personal computer and TV monitors. They are also popular with computer case modders in recent years.

Science demonstrations

Capacitive coupling with high-voltage power lines can light a lamp continuously at low intensity

Fluorescent lamps can be illuminated by means other than a proper electrical connection. These other methods however result in very dim or very short-lived illumination, and so are seen mostly in science demonstrations. Static electricity or a Van de Graaff generator will cause a lamp to flash momentarily as it discharges a high voltage capacitance. A Tesla coil will pass high frequency current through the tube, and since it has a high voltage as well, the gases within the tube will ionize and emit light. Capacitive coupling

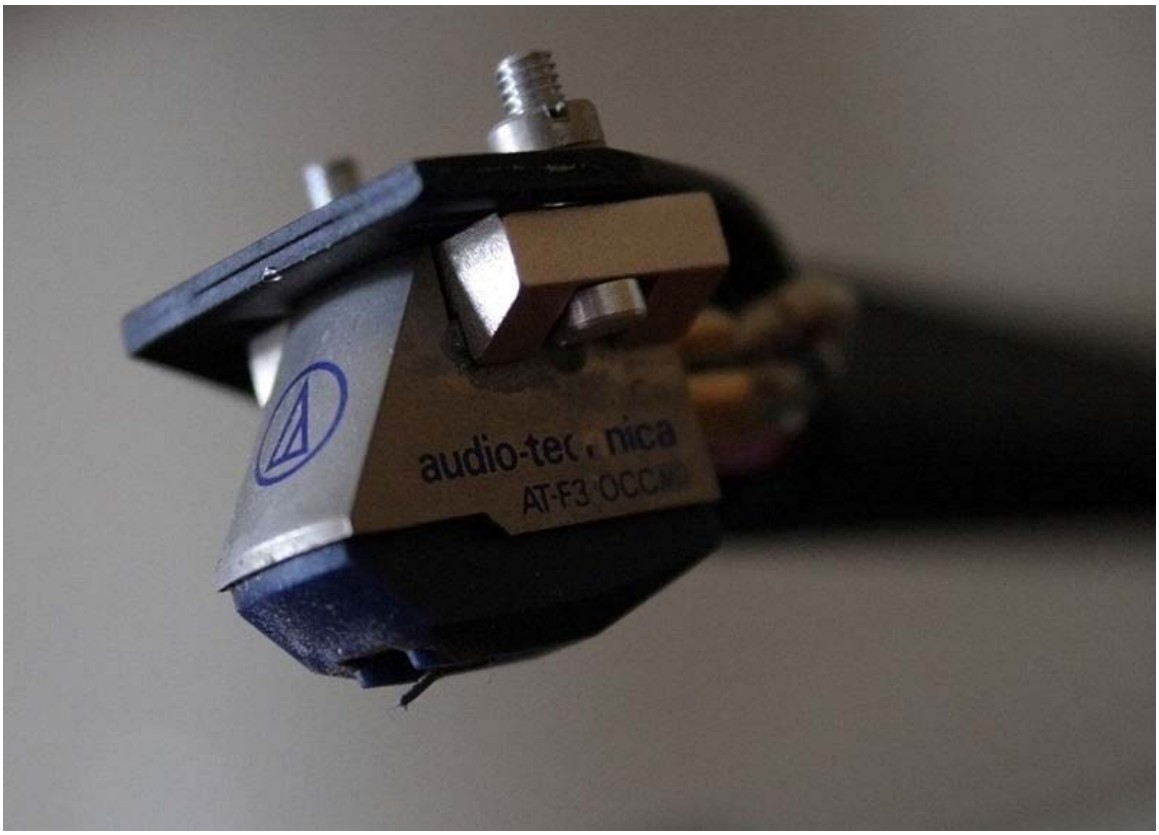
with high-voltage power lines can light a lamp continuously at low intensity, depending on the intensity of the electrostatic field.

Also, placing a bulb half way up a two-way radio antenna while transmitting will illuminate the bulb due to the RF energy.

Chapter 3

Magnetic Cartridge and Electric Motor

Magnetic cartridge



An Audio Technica AT-F3 MC cartridge

A **magnetic cartridge** is a transducer used for the playback of gramophone records on a turntable or phonograph. It converts mechanical vibrational energy from a stylus riding in a spiral record groove into an electrical signal that is subsequently amplified and then converted back to sound by a loudspeaker system.

History

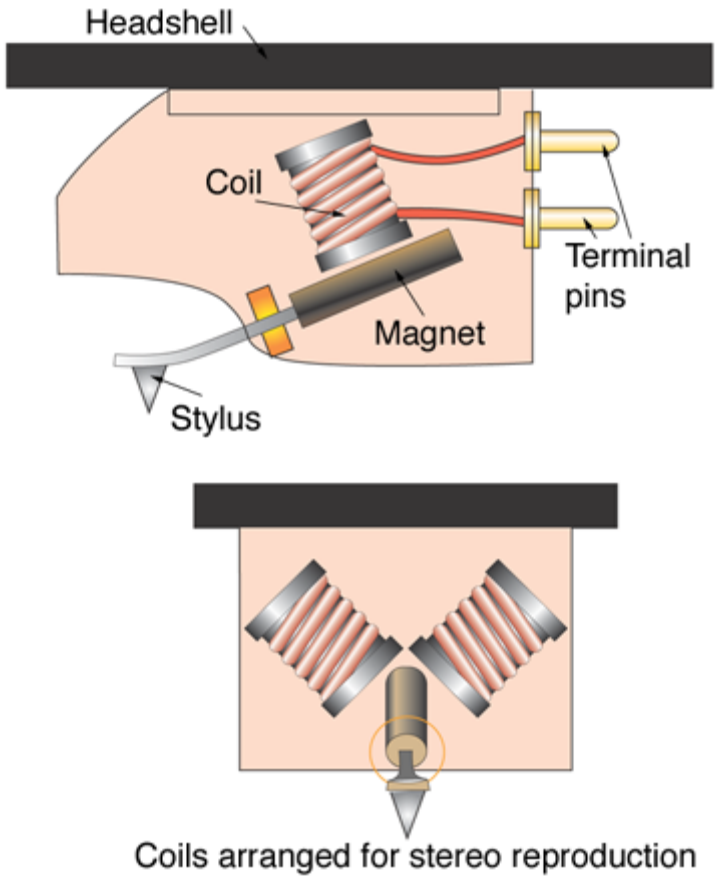
The first electric pick-ups were developed in about 1925. They used a piezo-electric crystal of quartz, stimulated by a stylus made of sapphire or diamond. The magnetic cartridge is presently the most common form of sound pickup used and came into use in the 1950s, following the introduction of magnetic cutter heads around 1945 for mastering records.

Types

In high-fidelity systems, crystal and ceramic pickups have been replaced by the **magnetic cartridge**, using either a **moving magnet** or a **moving coil**.

Compared to the crystal and ceramic pickups, the magnetic cartridge gives improved playback fidelity and reduces record wear by tracking the groove with lighter pressure. Magnetic cartridges use much lower tracking forces and thus damage the record grooves less. They also have a lower output voltage than a crystal or ceramic pickup, in the range of only a few millivolts, thus requiring greater amplification.

Moving Magnet (MM) cartridges

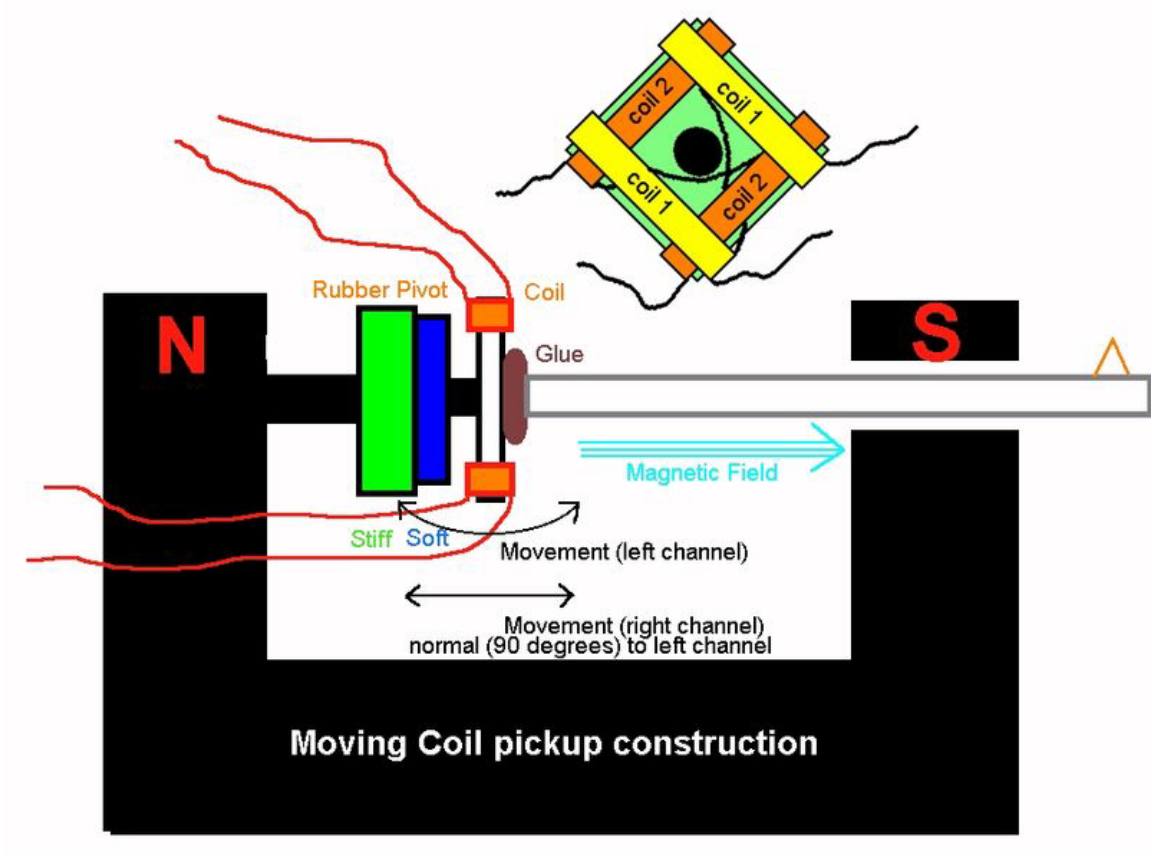


In a moving magnet cartridge, the stylus cantilever carries a tiny permanent magnet, which is positioned between two sets of fixed coils (in a stereophonic cartridge), forming a tiny electromagnetic generator. As the magnet vibrates in response to the stylus following the record groove, it induces a tiny current in the coils.

Because the magnet is small and has little mass, and is not coupled mechanically to the generator (as in a ceramic cartridge), a properly adjusted stylus follows the groove more faithfully while requiring less tracking force (the downward pressure on the stylus).

There is a sub-category. Moving iron and induced magnet types (ADC being a well known example) which have the magnet fixed and move a piece of iron or other ferrous alloy in the field of the magnet to produce the signal within the fixed coils.

Moving Coil (MC) cartridges



The MC design is again a tiny electromagnetic generator, but (unlike an MM design) with the magnet and coils reversed: the coils are attached to the stylus, and move within the field of a permanent magnet. The coils are tiny and made from very fine wire, so are even lighter than the small magnet used in an MM cartridge, thus improving the tracking ability of the cartridge. This can give extended frequency response as well as greater fidelity.

An advantage however is that moving-coil cartridges generate an even lower voltage signal than a moving-magnet type cartridge. This is because the moving coil cannot be large enough (it would be too heavy) to generate equivalent voltage levels. The resulting signal is only a few hundred microvolts, and thus more easily swamped by noise, induced hum, etc. Thus it is more challenging to design a preamplifier with the extremely low noise inputs needed for moving-coil cartridge, therefore a "step up transformer" are sometimes used instead.

Moving coil cartridges are extremely small precision instruments and are therefore generally expensive, but are frequently preferred by audiophiles due to their better performance.

Moving Micro Cross (MMC) cartridges

The MMC design was invented and patented by Bang & Olufsen. Since it often uses a special mounting, it can be mostly found in Bang & Olufsen turntables (which also cannot use another type of cartridge). Apart from being produced for the Bang & Olufsen mounting system, the SP12 and SP14 were also available in standard 1/2" mount.

The MMC cartridge is a *moving iron* design. Magnets and coils are stationary while the *micro cross* moves with the stylus, thereby varying the distances between the arms of the cross and the magnets. The design obviously offers more freedom concerning magnet and coil mass (compared to MC and MM cartridges). For example, the MMC20 (presented in 1978) uses four coils wound around the magnetic cores with 1200 turns each. Minimizing the moving mass also reduces the unavoidable wear on the records.

It is also claimed that the MMC design allows for superior channel separation, since each channel's movements appear on a separate axis.

Moving Magnet vs. Moving Coil debate

The debate as to whether MM or MC designs can ultimately produce the better sound is often heated and subjective. The distinction between the two is often blurred by cost and design considerations - i.e. can an MC cartridge requiring another step-up amplification outperform well made MM cartridges that need simpler front-end stages? Every now and then a design comes along to re-open this debate. A good example being the Linn K9 (now discontinued) - regarded by some as one of the better MM designs and competitive with MC alternatives costing more. Amongst others, Grace, ADC and Grado also manufactured notably good designs.

"Decca" Cartridges (aka "Moving Iron")

The Decca phono cartridges were a unique design, with fixed magnets and coils. The stylus shaft was composed of the diamond tip, a short piece of soft iron, and an L-shaped cantilever made of non-magnetic steel. Since the iron was placed very close to the tip (within 1 mm), the motions of the tip could be tracked very accurately. Decca engineers

called this "positive scanning". Vertical and lateral compliance was controlled by the shape and thickness of the cantilever. Decca cartridges had a reputation for being very musical; however early versions required more tracking force than competitive designs - making record wear a concern.

Stereo reproduction

One reason that magnetic cartridges superseded the crystal pick-up was the relative ease with which it could be made to reproduce stereo recordings, which were introduced in 1958. In a stereo recording, the two channels are arranged to drive the record cutter head at an angle of 45° to the vertical, effectively encoding each channel in the left and right V-shaped walls of the record groove. This system worked well, since it provided full compatibility with a monaural pick-up, so stereo records could be played on older mono equipment. To reproduce the stereo signal, the cartridge simply arranges pairs of coils at 45° to complement the cutting process. With careful design, the coils can be shielded from each other electrically and mechanically such that stereo separation is maximised.

Comparison with crystal technology

Piezoelectric crystal or ceramic pickups had a few clear advantages. They were much easier and cheaper to make and were more robust than the delicate magnetic pickup. In addition, the output voltage from a crystal pickup is relatively large, requiring less amplification, which helped improve the signal-to-noise ratio. However, the signal from a crystal is not an accurate reproduction of the recording, as there is a lot of distortion introduced. The stylus is coupled to the crystal in a fairly rigid manner, which is also not as good at following rapid changes in the record grooves, so frequency response also suffers. This requires a greater tracking force, which in turn wears the records out faster. (This has earned cheap portable record players, the nickname "portable grinding wheel", in some circles.)

By contrast, since the magnetic cartridge is not mechanically coupled, the stylus and lever arm weight can be made exceedingly small. This gives extended frequency response and low distortion. The distortion is further minimised by the fact that there is more inherent linearity in the induction principle than there is in the piezo-electric one. Since the lighter stylus requires very low tracking forces, it requires a more sophisticated counterweighted arm, but reduces record wear.

The output from the magnetic cartridge is only a few millivolts compared to several tenths of a volt from a crystal or ceramic pickup. This requires an additional preamplifier stage. Careful design and shielding in the signal cables and amplifier is needed to prevent unwanted noise (shot noise or EMI). The magnetic induction principle also naturally leads to a linearly rising response with increasing frequency, and this needs to be compensated for to give correct (flat) frequency reproduction. Conversely, the very low bass frequencies are not efficiently picked up, so a strong bass boost is needed. This can amplify unwanted low frequency noise such as that from the turntable motor and drive mechanism itself (rumble). Crystal pickups do not suffer from these drawbacks and give

a much better bass performance, so they may be preferred in some applications (such as DJ-ing), where robustness and good bass are favoured over highest fidelity reproduction. The moving coil pickups tend to have an even lower level of output, and so usually require a step-up transformer or very special preamplifiers to bring these signals up to the level of input that a standard amplifier requires. These special preamplifiers must have very low noise indeed (some have even been cryogenically cooled), and hence tend to be expensive. Many audiophiles claim that the benefits are worthwhile. There are higher output moving coil pickup cartridges that can be used directly into a normal moving magnet input, but these are in the minority.

The bass boost and high frequency rolloff can be conveniently incorporated into the preamplifier which implements the RIAA equalization curve, a noise reduction technique used on all modern records.

The magnetic nature of the modern pickup means it must be shielded from external fields, especially from the loudspeakers of the same system, or unpleasant and possibly damaging feedback can occur. For this reason, the cartridge itself has a shield, usually of mu-metal, to help screen out unwanted fields.

The stylus

The stylus, or "needle", is a crucial part of the record player (or 'phonograph' or 'gramophone', both terms now archaic), as it is the one part of the system that actually contacts the recorded disc and transfers its vibrations to the rest of the system. It is the part which also suffers the greatest wear. There are two desired qualities in a stylus: first, that it faithfully follows the contours of the recorded groove and transfers the vibration to the system, and second, that it does not damage the recorded disc.

Electric motor

Electric motors

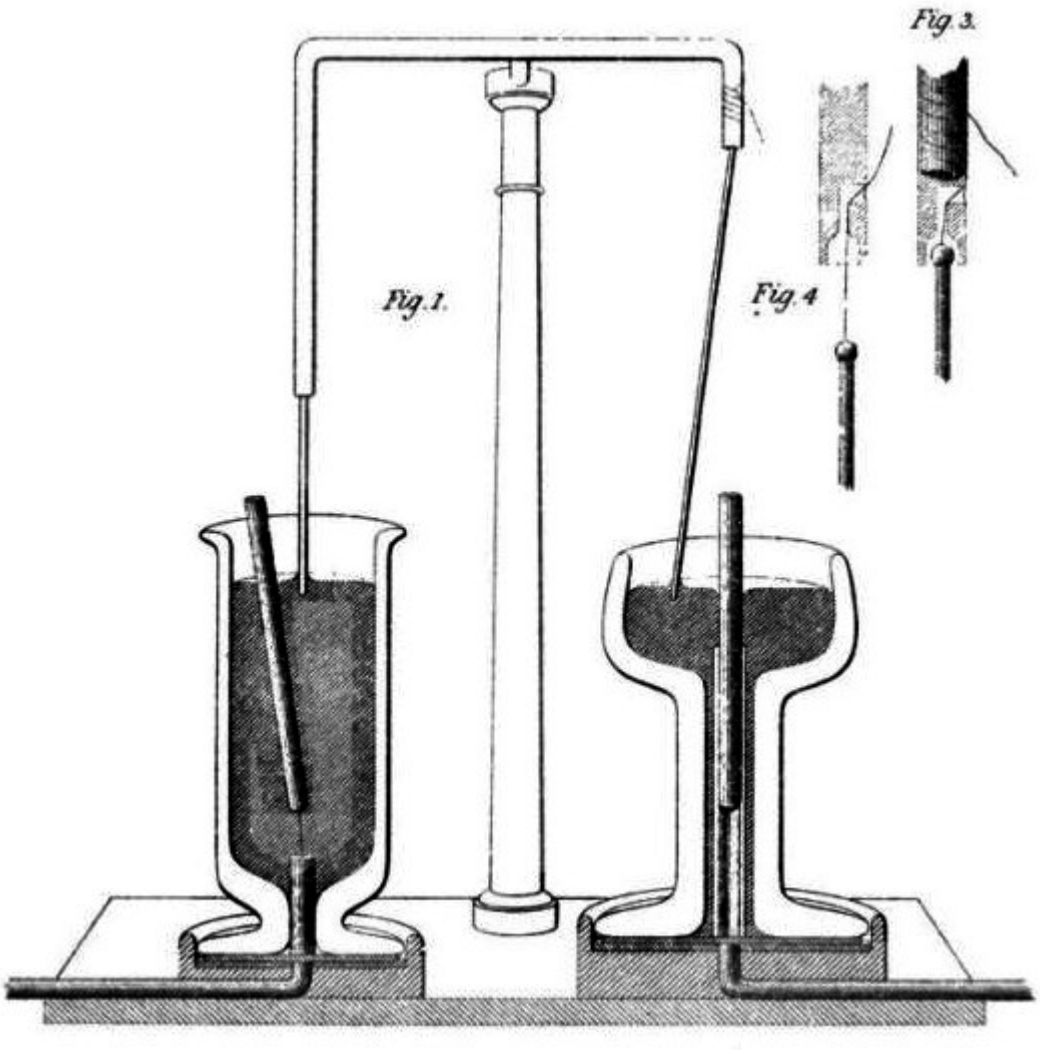
An **electric motor** is a type of machine that converts electrical energy into mechanical energy. Electric motors operate through interacting magnetic fields and current-carrying conductors to generate force, although a few use electrostatic forces. The reverse process, producing electrical energy from mechanical energy, is accomplished by an alternator, generator or dynamo. Many types of electric motors can be run as generators, and vice versa. For example a starter/generator for a gas turbine or Traction motors used on vehicles often perform both tasks.

Electric motors are found in applications as diverse as industrial fans, blowers and pumps, machine tools, household appliances, power tools, and disk drives. They may be powered by direct current (e.g., a battery powered portable device or motor vehicle), or by alternating current from a central electrical distribution grid. The smallest motors may be found in electric wristwatches. Medium-size motors of highly standardized dimensions and characteristics provide convenient mechanical power for industrial uses. The very largest electric motors are used for propulsion of large ships, and for such purposes as pipeline compressors, with ratings in the millions of watts. Electric motors may be classified by the source of electric power, by their internal construction, by their application, or by the type of motion they give.

The physical principle of production of mechanical force by the interactions of an electric current and a magnetic field was known as early as 1821. Electric motors of increasing efficiency were constructed throughout the 19th century, but commercial exploitation of electric motors on a large scale required efficient electrical generators and electrical distribution networks.

Some devices, such as magnetic solenoids and loudspeakers, although they generate some mechanical power, are not generally referred to as electric motors, and are usually termed actuators and transducers, respectively.

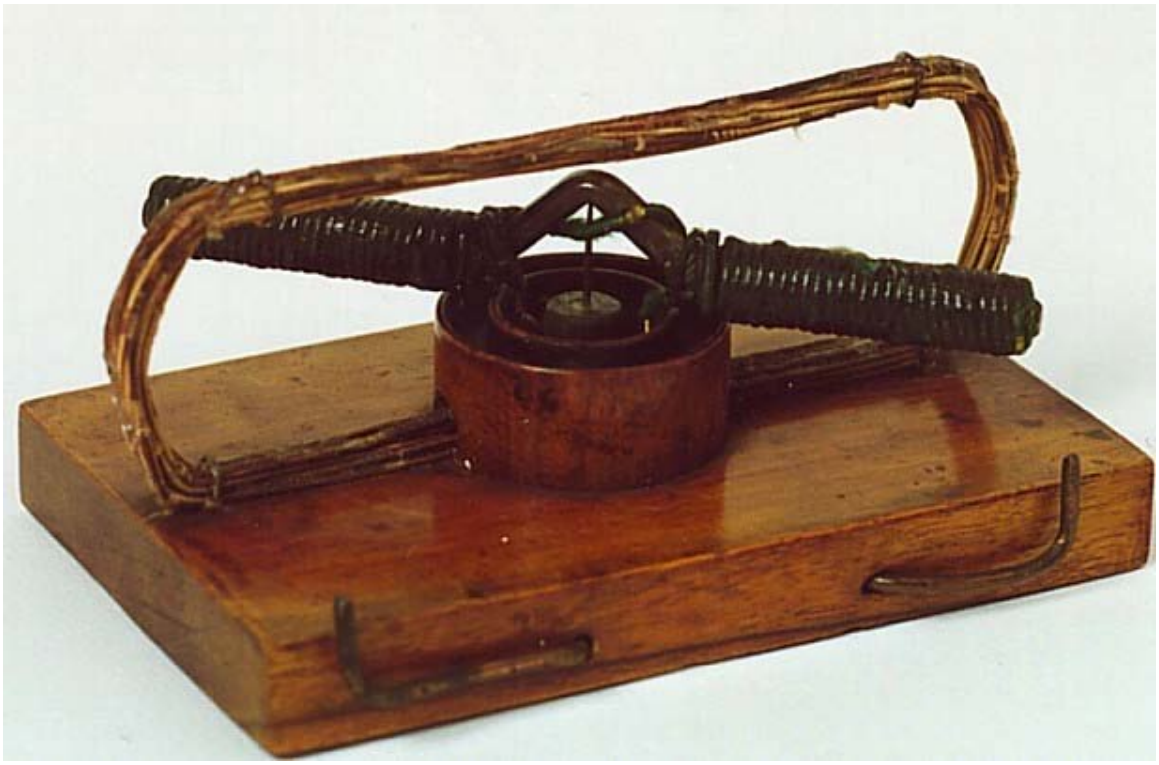
History and development



Faraday's Electromagnetic experiment, 1821.

The principle

The conversion of electrical energy into mechanical energy by a electromagnetic means was demonstrated by the British scientist Michael Faraday in 1821. A free-hanging wire was dipped into a pool of mercury, on which a permanent magnet was placed. When a current was passed through the wire, the wire rotated around the magnet, showing that the current gave rise to a close circular magnetic field around the wire. This motor is often demonstrated in school physics classes, but brine (salt water) is sometimes used in place of the toxic mercury. This is the simplest form of a class of devices called homopolar motors. A later refinement is the Barlow's Wheel. These were demonstration devices only, unsuited to practical applications due to their primitive construction.



Jedlik's "electromagnetic self-rotor", 1827. (Museum of Applied Arts, Budapest. The historic motor still works perfectly today.)

In 1827, Hungarian Ányos Jedlik started experimenting with electromagnetic rotating devices he called "electromagnetic self-rotors". He used them for instructive purposes in universities, and in 1828 demonstrated the first device which contained the three main components of practical direct current motors: the stator, rotor and commutator. Both the stationary and the revolving parts were electromagnetic, employing no permanent magnets. Again, the devices had no practical application.

The first electric motors

The first commutator-type direct current electric motor capable of turning machinery was invented by the British scientist William Sturgeon in 1832. Following Sturgeon's work, a commutator-type direct-current electric motor made with the intention of commercial use was built by Americans Emily and Thomas Davenport and patented in 1837. Their motors ran at up to 600 revolutions per minute, and powered machine tools and a printing press. Due to the high cost of the zinc electrodes required by primary battery power, the motors were commercially unsuccessful and the Davenports went bankrupt. Several inventors followed Sturgeon in the development of DC motors but all encountered the same cost issues with primary battery power. No electricity distribution had been developed at the time. Like Sturgeon's motor, there was no practical commercial market for these motors.

In 1855 Jedlik built a device using similar principles to those used in his electromagnetic self-rotors that was capable of useful work. He built a model electric motor-propelled vehicle that same year. There is no evidence that this experimentation was communicated to the wider scientific world at that time, or that it influenced the development of electric motors in the following decades.

The modern DC motor was invented by accident in 1873, when Zénobe Gramme connected the dynamo he had invented to a second similar unit, driving it as a motor. The Gramme machine was the first electric motor that was successful in the industry.

In 1886 Frank Julian Sprague invented the first practical DC motor, a non-sparking motor capable of constant speed under variable loads. Other Sprague electric inventions about this time greatly improved grid electric distribution (prior work done while employed by Thomas Edison), allowed power from electric motors to be returned to the electric grid, provided for electric distribution to trolleys via overhead wires and the trolley pole, and provided controls systems for electric operations. This allowed Sprague to use electric motors to invent the first electric trolley system in 1887-88 in Richmond VA, the electric elevator and control system in 1892, and the electric subway with independently powered centrally controlled cars, which was first installed in 1892 in Chicago by the South Side Elevated Railway where it became popularly known as the "L". Sprague's motor and related inventions led to an explosion of interest and use in electric motors for industry, while almost simultaneously another great inventor was developing its primary competitor, which would become much more widespread.

In 1888 Nikola Tesla invented the first practicable AC motor and with it the polyphase power transmission system. Tesla continued his work on the AC motor in the years to follow at the Westinghouse company.

The development of electric motors of acceptable efficiency was delayed for several decades by failure to recognize the extreme importance of a relatively small air gap between rotor and stator. Early motors, for some rotor positions, had comparatively huge air gaps which constituted a very high reluctance magnetic circuit. They produced far-lower torque than an equivalent amount of power would produce with efficient designs. The cause of the lack of understanding seems to be that early designs were based on familiarity of distant attraction between a magnet and a piece of ferromagnetic material, or between two electromagnets. Efficient designs, as described here, are based on a rotor with a comparatively small air gap, and flux patterns that create torque.

Note that the armature bars are at some distance (unknown) from the field pole pieces when power is fed to one of the field magnets; the air gap is likely to be considerable. The text tells of the inefficiency of the design. (Electricity was created, as a practical matter, by consuming zinc in wet primary cells!)

In his workshops Froment had an electromotive engine of one-horse power. But, though an interesting application of the transformation of energy, these machines will never be practically applied on the large scale in manufactures, for the expense of the acids and the

zinc which they use very far exceeds that of the coal in steam-engines of the same force. [...] motors worked by electricity, independently of any question as to the cost of construction, or of the cost of the acids, are at least sixty times as dear to work as steam-engines.

Although Gramme's design was comparatively much more efficient, apparently the Froment motor was still considered illustrative, years later. It is of some interest that the St. Louis motor, long used in classrooms to illustrate motor principles, is extremely inefficient for the same reason, as well as appearing nothing like a modern motor. Photo of a traditional form of the motor: Note the prominent bar magnets, and the huge air gap at the ends opposite the rotor. Even modern versions still have big air gaps if the rotor poles are not aligned.

Application of electric motors revolutionized industry. Industrial processes were no longer limited by power transmission using shaft, belts, compressed air or hydraulic pressure. Instead every machine could be equipped with its own electric motor, providing easy control at the point of use, and improving power transmission efficiency. Electric motors applied in agriculture eliminated human and animal muscle power from such tasks as handling grain or pumping water. Household uses of electric motors reduced heavy labor in the home and made higher standards of convenience, comfort and safety possible. Today, electric motors consume more than half of all electric energy produced.

Categorization of electric motors

The classic division of electric motors has been that of Alternating Current (AC) types vs Direct Current (DC) types. This is more a *de facto* convention, rather than a rigid distinction. For example, many classic DC motors run on AC power, these motors being referred to as universal motors.

Rated output power is also used to categorise motors, those of less than 746 Watts, for example, are often referred to as fractional horsepower motors (FHP) in reference to the old imperial measurement.

The ongoing trend toward electronic control further muddles the distinction, as modern drivers have moved the commutator out of the motor shell. For this new breed of motor, driver circuits are relied upon to generate sinusoidal AC drive currents, or some approximation thereof. The two best examples are: the brushless DC motor and the stepping motor, both being poly-phase AC motors requiring external electronic control, although historically, stepping motors (such as for maritime and naval gyrocompass repeaters) were driven from DC switched by contacts.

Considering all rotating (or linear) electric motors require synchronism between a moving magnetic field and a moving current sheet for average torque production, there is a clearer distinction between an asynchronous motor and synchronous types. An asynchronous motor requires slip between the moving magnetic field and a winding set to induce current in the winding set by mutual inductance; the most ubiquitous example

being the common AC induction motor which must slip to generate torque. In the synchronous types, induction (or slip) is not a requisite for magnetic field or current production (e.g. permanent magnet motors, synchronous brush-less wound-rotor doubly-fed electric machine).

Comparison of motor types

Comparison of motor types

Type	Advantages	Disadvantages	Typical Application	Typical Drive
AC Induction (Shaded Pole)	Least expensive Long life high power	Rotation slips from frequency Low starting torque	Fans	Uni/Poly-phase AC
AC Induction (split-phase capacitor)	High power high starting torque	Rotation slips from frequency	Appliances Stationary Power Tools	Uni/Poly-phase AC
Universal motor	High starting torque, compact, high speed	Maintenance (brushes) Medium lifespan	Drill, blender, vacuum cleaner, insulation blowers	Uni-phase AC or Direct DC
AC Synchronous	Rotation in-sync with freq - hence no slip long-life (alternator)	More expensive	Industrial motors Clocks Audio turntables tape drives	Uni/Poly-phase AC
Stepper DC	Precision positioning High holding torque	High initial cost Requires a controller	Positioning in printers and floppy drives	DC
Brushless DC	Long lifespan low maintenance High efficiency	High initial cost Requires a controller	Hard drives CD/DVD players electric vehicles	DC
Brushed DC	Low initial cost Simple speed control	Maintenance (brushes) Medium lifespan	Treadmill exercisers automotive motors (seats, blowers, windows)	Direct DC or PWM
Pancake DC	Compact design Simple speed control	Medium cost Medium lifespan	Office Equip Fans/Pumps	Direct DC or PWM

Servo motor

A servomechanism, or servo is an automatic device that uses error-sensing feedback to correct the performance of a mechanism. The term correctly applies only to systems where the feedback or error-correction signals help control mechanical position or other parameters. For example, an automotive power window control is not a servomechanism, as there is no automatic feedback which controls position—the operator does this by observation. By contrast the car's cruise control uses closed loop feedback, which classifies it as a servomechanism.

Synchronous electric motor

A synchronous electric motor is an AC motor distinguished by a rotor spinning with coils passing magnets at the same rate as the alternating current and resulting magnetic field which drives it. Another way of saying this is that it has zero slip under usual operating conditions. Contrast this with an induction motor, which must slip to produce torque. A synchronous motor is like an induction motor except the rotor is excited by a DC field. Slip rings and brushes are used to conduct current to rotor. The rotor poles connect to each other and move at the same speed hence the name synchronous motor.

Induction motor

An induction motor (IM) is a type of asynchronous AC motor where power is supplied to the rotating device by means of electromagnetic induction. Another commonly used name is squirrel cage motor because the rotor bars with short circuit rings resemble a squirrel cage (hamster wheel). An electric motor converts electrical power to mechanical power in its rotor (rotating part). There are several ways to supply power to the rotor. In a DC motor this power is supplied to the armature directly from a DC source, while in an induction motor this power is induced in the rotating device. An induction motor is sometimes called a rotating transformer because the stator (stationary part) is essentially the primary side of the transformer and the rotor (rotating part) is the secondary side. Induction motors are widely used, especially polyphase induction motors, which are often used in industrial drives.

Electrostatic motor (capacitor motor)

An electrostatic motor or capacitor motor is a type of electric motor based on the attraction and repulsion of electric charge. Usually, electrostatic motors are the dual of conventional coil-based motors. They typically require a high voltage power supply, although very small motors employ lower voltages. Conventional electric motors instead employ magnetic attraction and repulsion, and require high current at low voltages. In the 1750s, the first electrostatic motors were developed by Benjamin Franklin and Andrew Gordon. Today the electrostatic motor finds frequent use in micro-mechanical (MEMS) systems where their drive voltages are below 100 volts, and where moving, charged plates are far easier to fabricate than coils and iron cores. Also, the molecular machinery which runs living cells is often based on linear and rotary electrostatic motors.

DC Motors

A DC motor is designed to run on DC electric power. Two examples of pure DC designs are Michael Faraday's homopolar motor (which is uncommon), and the ball bearing motor, which is (so far) a novelty. By far the most common DC motor types are the brushed and brushless types, which use internal and external commutation respectively to create an oscillating AC current from the DC source—so they are not purely DC machines in a strict sense.

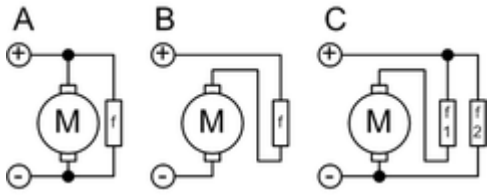
Brushed DC motors

DC motor design generates an oscillating current in a wound rotor, or armature, with a split ring commutator, and either a wound or permanent magnet stator. A rotor consists of one or more coils of wire wound around a core on a shaft; an electrical power source is connected to the rotor coil through the commutator and its brushes, causing current to flow in it, producing electromagnetism. The commutator causes the current in the coils to be switched as the rotor turns, keeping the magnetic poles of the rotor from ever fully aligning with the magnetic poles of the stator field, so that the rotor never stops (like a compass needle does) but rather keeps rotating indefinitely (as long as power is applied and is sufficient for the motor to overcome the shaft torque load and internal losses due to friction, etc.)

Many of the limitations of the classic commutator DC motor are due to the need for brushes to press against the commutator. This creates friction. Sparks are created by the brushes making and breaking circuits through the rotor coils as the brushes cross the insulating gaps between commutator sections. Depending on the commutator design, this may include the brushes shorting together adjacent sections—and hence coil ends—momentarily while crossing the gaps. Furthermore, the inductance of the rotor coils causes the voltage across each to rise when its circuit is opened, increasing the sparking of the brushes. This sparking limits the maximum speed of the machine, as too-rapid sparking will overheat, erode, or even melt the commutator. The current density per unit area of the brushes, in combination with their resistivity, limits the output of the motor. The making and breaking of electric contact also causes electrical noise, and the sparks additionally cause RFI. Brushes eventually wear out and require replacement, and the commutator itself is subject to wear and maintenance (on larger motors) or replacement (on small motors). The commutator assembly on a large motor is a costly element, requiring precision assembly of many parts. On small motors, the commutator is usually permanently integrated into the rotor, so replacing it usually requires replacing the whole rotor.

Large brushes are desired for a larger brush contact area to maximize motor output, but small brushes are desired for low mass to maximize the speed at which the motor can run without the brushes excessively bouncing and sparking (comparable to the problem of "valve float" in internal combustion engines). (Small brushes are also desirable for lower cost.) Stiffer brush springs can also be used to make brushes of a given mass work at a higher speed, but at the cost of greater friction losses (lower efficiency) and accelerated

brush and commutator wear. Therefore, DC motor brush design entails a trade-off between output power, speed, and efficiency/wear.



A: shunt
B: series
C: compound
f = field coil

There are five types of brushed DC motor:

A. DC shunt wound motor

B. DC series wound motor

C. DC compound motor (two configurations):

- Cumulative compound
- Differentially compounded

D. Permanent Magnet DC Motor (not shown)

E. Separately excited (sepex) (not shown).

Brushless DC motors

Some of the problems of the brushed DC motor are eliminated in the brushless design. In this motor, the mechanical "rotating switch" or commutator/brushgear assembly is replaced by an external electronic switch synchronised to the rotor's position. Brushless motors are typically 85-90% efficient or more (higher efficiency for a brushless electric motor of up to 96.5% were reported by researchers at the Tokai University in Japan in 2009), whereas DC motors with brushgear are typically 75-80% efficient.

Midway between ordinary DC motors and stepper motors lies the realm of the brushless DC motor. Built in a fashion very similar to stepper motors, these often use a permanent magnet external rotor, three phases of driving coils, one or more Hall effect sensors to sense the position of the rotor, and the associated drive electronics. The coils are activated, one phase after the other, by the drive electronics as cued by the signals from either Hall effect sensors or from the back EMF (electromotive force) of the undriven coils. In effect, they act as three-phase synchronous motors containing their own variable-frequency drive electronics. A specialized class of brushless DC motor controllers utilize

EMF feedback through the main phase connections instead of Hall effect sensors to determine position and velocity. These motors are used extensively in electric radio-controlled vehicles. When configured with the magnets on the outside, these are referred to by modellers as outrunner motors.

Brushless DC motors are commonly used where precise speed control is necessary, as in computer disk drives or in video cassette recorders, the spindles within CD, CD-ROM (etc.) drives, and mechanisms within office products such as fans, laser printers and photocopiers. They have several advantages over conventional motors:

- Compared to AC fans using shaded-pole motors, they are very efficient, running much cooler than the equivalent AC motors. This cool operation leads to much-improved life of the fan's bearings.
- Without a commutator to wear out, the life of a DC brushless motor can be significantly longer compared to a DC motor using brushes and a commutator. Commutation also tends to cause a great deal of electrical and RF noise; without a commutator or brushes, a brushless motor may be used in electrically sensitive devices like audio equipment or computers.
- The same Hall effect sensors that provide the commutation can also provide a convenient tachometer signal for closed-loop control (servo-controlled) applications. In fans, the tachometer signal can be used to derive a "fan OK" signal.
- The motor can be easily synchronized to an internal or external clock, leading to precise speed control.
- Brushless motors have no chance of sparking, unlike brushed motors, making them better suited to environments with volatile chemicals and fuels. Also, sparking generates ozone which can accumulate in poorly ventilated buildings risking harm to occupants' health.
- Brushless motors are usually used in small equipment such as computers and are generally used to get rid of unwanted heat.
- They are also very quiet motors which is an advantage if being used in equipment that is affected by vibrations.

Modern DC brushless motors range in power from a fraction of a watt to many kilowatts. Larger brushless motors up to about 100 kW rating are used in electric vehicles. They also find significant use in high-performance electric model aircraft.

Coreless or ironless DC motors

Nothing in the design of any of the motors described above requires that the iron (steel) portions of the rotor actually rotate; torque is exerted only on the windings of the electromagnets. Taking advantage of this fact is the **coreless or ironless DC motor**, a specialized form of a brush or brushless DC motor. Optimized for rapid acceleration, these motors have a rotor that is constructed without any iron core. The rotor can take the form of a winding-filled cylinder, or a self-supporting structure comprising only the magnet wire and the bonding material. The rotor can fit inside the stator magnets; a

magnetically soft stationary cylinder inside the rotor provides a return path for the stator magnetic flux. A second arrangement has the rotor winding basket surrounding the stator magnets. In that design, the rotor fits inside a magnetically soft cylinder that can serve as the housing for the motor, and likewise provides a return path for the flux.

Because the rotor is much lighter in weight (mass) than a conventional rotor formed from copper windings on steel laminations, the rotor can accelerate much more rapidly, often achieving a mechanical time constant under 1 ms. This is especially true if the windings use aluminum rather than the heavier copper. But because there is no metal mass in the rotor to act as a heat sink, even small coreless motors must often be cooled by forced air.

Related limited-travel actuators have no core and a bonded coil placed between the poles of high-flux thin permanent magnets. These are the fast head positioners for rigid-disk ("hard disk") drives.

Printed Armature or Pancake DC Motors

A rather unusual motor design the pancake/printed armature motor has the windings shaped as a disc running between arrays of high-flux magnets, arranged in a circle, facing the rotor and forming an axial air gap. This design is commonly known the pancake motor because of its extremely flat profile, although the technology has had many brand names since its inception, such as ServoDisc.

The printed armature (originally formed on a printed circuit board) in a printed armature motor is made from punched copper sheets that are laminated together using advanced composites to form a thin rigid disc. The printed armature has a unique construction, in the brushed motor world, in that it does not have a separate ring commutator. The brushes run directly on the armature surface making the whole design very compact.

An alternative manufacturing method is to use wound copper wire laid flat with a central conventional commutator, in a flower and petal shape. The windings are typically stabilized by being impregnated with electrical epoxy potting systems. These are filled epoxies that have moderate mixed viscosity and a long gel time. They are highlighted by low shrinkage and low exotherm, and are typically UL 1446 recognized as a potting compound for use up to 180°C (Class H) (UL File No. E 210549).

The unique advantage of ironless DC motors is that there is no cogging (vibration caused by attraction between the iron and the magnets) and parasitic eddy currents cannot form in the rotor as it is totally ironless. This can greatly improve efficiency, but variable-speed controllers must use a higher switching rate (>40 kHz) or direct current because of the decreased electromagnetic induction.

These motors were originally invented to drive the capstan(s) of magnetic tape drives, in the burgeoning computer industry. Pancake motors are still widely used in high-performance servo-controlled systems, humanoid robotic systems, industrial automation and medical devices. Due to the variety of constructions now available the technology is

used in applications from high temperature military to low cost pump and basic servo applications.

Universal motors

A series-wound motor is referred to as a **universal motor** when it has been designed to operate on either AC or DC power. The ability to operate on AC is because the current in both the field and the armature (and hence the resultant magnetic fields) will alternate (reverse polarity) in synchronism, and hence the resulting mechanical force will occur in a constant direction.

Operating at normal power line frequencies, universal motors are often found in a range rarely larger than one kilowatt (about 1.3 horsepower). Universal motors also form the basis of the traditional railway traction motor in electric railways. In this application, the use of AC to power a motor originally designed to run on DC would lead to efficiency losses due to eddy current heating of their magnetic components, particularly the motor field pole-pieces that, for DC, would have used solid (un-laminated) iron. Although the heating effects are reduced by using laminated pole-pieces, as used for the cores of transformers and by the use of laminations of high permeability electrical steel, one solution available at start of the 20th Century was for the motors to be operated from very low frequency AC supplies, with 25 and 16.7 hertz (Hz) operation being common. Because they used universal motors, locomotives using this design were also commonly capable of operating from a third rail powered by DC.

An advantage of the universal motor is that AC supplies may be used on motors which have some characteristics more common in DC motors, specifically high starting torque and very compact design if high running speeds are used. The negative aspect is the maintenance and short life problems caused by the commutator. As a result, such motors are usually used in AC devices such as food mixers and power tools which are used only intermittently, and often have high starting-torque demands. Continuous speed control of a universal motor running on AC is easily obtained by use of a thyristor circuit, while (imprecise) stepped speed control can be accomplished using multiple taps on the field coil. Household blenders that advertise many speeds frequently combine a field coil with several taps and a diode that can be inserted in series with the motor (causing the motor to run on half-wave rectified AC).

Universal motors generally run at high speeds, making them useful for appliances such as blenders, vacuum cleaners, and hair dryers where high RPM operation is desirable. They are also commonly used in portable power tools, such as drills, sanders (both disc and orbital), circular and jig saws, where the motor's characteristics work well. Many vacuum cleaner and weed trimmer motors exceed 10,000 RPM, while Dremel and other similar miniature grinders will often exceed 30,000 RPM.

Universal motors also lend themselves to electronic speed control and, as such, are an ideal choice for domestic washing machines. The motor can be used to agitate the drum

(both forwards and in reverse) by switching the field winding with respect to the armature. The motor can also be run up to the high speeds required for the spin cycle.

Motor damage may occur due to overspeeding (running at an RPM in excess of design limits) if the unit is operated with no significant load. On larger motors, sudden loss of load is to be avoided, and the possibility of such an occurrence is incorporated into the motor's protection and control schemes. In some smaller applications, a fan blade attached to the shaft often acts as an artificial load to limit the motor speed to a safe value, as well as a means to circulate cooling airflow over the armature and field windings.

AC motors

In 1882, Nikola Tesla discovered the rotating magnetic field, and pioneered the use of a rotary field of force to operate machines. He exploited the principle to design a unique two-phase induction motor in 1883. In 1885, Galileo Ferraris independently researched the concept. In 1888, Ferraris published his research in a paper to the Royal Academy of Sciences in Turin.

Tesla had suggested that the commutators from a machine could be removed and the device could operate on a rotary field of force. Professor Poeschel, his teacher, stated that would be akin to building a perpetual motion machine. Tesla would later attain U.S. Patent 0,416,194, *Electric Motor* (December 1889), which resembles the motor seen in many of Tesla's photos. This classic alternating current electro-magnetic motor was an induction motor.

Michail Osipovich Dolivo-Dobrovolsky later invented a three-phase "cage-rotor" in 1890. This type of motor is now used for the vast majority of commercial applications.

Components

A typical AC motor consists of two parts:

- An outside stationary stator having coils supplied with AC current to produce a rotating magnetic field, and;
- An inside rotor attached to the output shaft that is given a torque by the rotating field.

Torque motors

A torque motor (also known as a limited torque motor) is a specialized form of induction motor which is capable of operating indefinitely while stalled, that is, with the rotor blocked from turning, without incurring damage. In this mode of operation, the motor will apply a steady torque to the load (hence the name).

A common application of a torque motor would be the supply- and take-up reel motors in a tape drive. In this application, driven from a low voltage, the characteristics of these motors allow a relatively constant light tension to be applied to the tape whether or not the capstan is feeding tape past the tape heads. Driven from a higher voltage, (and so delivering a higher torque), the torque motors can also achieve fast-forward and rewind operation without requiring any additional mechanics such as gears or clutches. In the computer gaming world, torque motors are used in force feedback steering wheels.

Another common application is the control of the throttle of an internal combustion engine in conjunction with an electronic governor. In this usage, the motor works against a return spring to move the throttle in accordance with the output of the governor. The latter monitors engine speed by counting electrical pulses from the ignition system or from a magnetic pickup and depending on the speed, makes small adjustments to the amount of current applied to the motor. If the engine starts to slow down relative to the desired speed, the current will be increased, the motor will develop more torque, pulling against the return spring and opening the throttle. Should the engine run too fast, the governor will reduce the current being applied to the motor, causing the return spring to pull back and close the throttle.

Slip ring

The slip ring is a component of the wound rotor motor as an induction machine (best evidenced by the construction of the common automotive alternator), where the rotor comprises a set of coils that are electrically terminated in slip rings. These are metal rings rigidly mounted on the rotor, and combined with brushes (as used with commutators), provide continuous unswitched connection to the rotor windings.

In the case of the wound-rotor induction motor, external impedances can be connected to the brushes. The stator is excited similarly to the standard squirrel cage motor. By changing the impedance connected to the rotor circuit, the speed/current and speed/torque curves can be altered.

(Slip rings are most-commonly used in automotive alternators as well as in synchro angular data-transmission devices, among other applications.)

The slip ring motor is used primarily to start a high inertia load or a load that requires a very high starting torque across the full speed range. By correctly selecting the resistors used in the secondary resistance or slip ring starter, the motor is able to produce maximum torque at a relatively low supply current from zero speed to full speed. This type of motor also offers controllable speed.

Motor speed can be changed because the torque curve of the motor is effectively modified by the amount of resistance connected to the rotor circuit. Increasing the value of resistance will move the speed of maximum torque down. If the resistance connected to the rotor is increased beyond the point where the maximum torque occurs at zero speed, the torque will be further reduced.

When used with a load that has a torque curve that increases with speed, the motor will operate at the speed where the torque developed by the motor is equal to the load torque. Reducing the load will cause the motor to speed up, and increasing the load will cause the motor to slow down until the load and motor torque are equal. Operated in this manner, the slip losses are dissipated in the secondary resistors and can be very significant. The speed regulation and net efficiency is also very poor.

Stepper motors

Closely related in design to three-phase AC synchronous motors are stepper motors, where an internal rotor containing permanent magnets or a magnetically soft rotor with salient poles is controlled by a set of external magnets that are switched electronically. A stepper motor may also be thought of as a cross between a DC electric motor and a rotary solenoid. As each coil is energized in turn, the rotor aligns itself with the magnetic field produced by the energized field winding. Unlike a synchronous motor, in its application, the stepper motor may not rotate continuously; instead, it "steps" — starts and then quickly stops again — from one position to the next as field windings are energized and de-energized in sequence. Depending on the sequence, the rotor may turn forwards or backwards, and it may change direction, stop, speed up or slow down arbitrarily at any time.

Simple stepper motor drivers entirely energize or entirely de-energize the field windings, leading the rotor to "cog" to a limited number of positions; more sophisticated drivers can proportionally control the power to the field windings, allowing the rotors to position between the cog points and thereby rotate extremely smoothly. This mode of operation is often called microstepping. Computer controlled stepper motors are one of the most versatile forms of positioning systems, particularly when part of a digital servo-controlled system.

Stepper motors can be rotated to a specific angle in discrete steps with ease, and hence stepper motors are used for read/write head positioning in computer floppy diskette drives. They were used for the same purpose in pre-gigabyte era computer disk drives, where the precision and speed they offered was adequate for the correct positioning of the read/write head of a hard disk drive. As drive density increased, the precision and speed limitations of stepper motors made them obsolete for hard drives—the precision limitation made them unusable, and the speed limitation made them uncompetitive—thus newer hard disk drives use voice coil-based head actuator systems. (The term "voice coil" in this connection is historic; it refers to the structure in a typical (cone type) loudspeaker. This structure was used for a while to position the heads. Modern drives have a pivoted coil mount; the coil swings back and forth, something like a blade of a rotating fan. Nevertheless, like a voice coil, modern actuator coil conductors (the magnet wire) move perpendicular to the magnetic lines of force.)

Stepper motors were and still are often used in computer printers, optical scanners, and digital photocopiers to move the optical scanning element, the print head carriage (of dot matrix and inkjet printers), and the platen. Likewise, many computer plotters (which

since the early 1990s have been replaced with large-format inkjet and laser printers) used rotary stepper motors for pen and platen movement; the typical alternatives here were either linear stepper motors or servomotors with complex closed-loop control systems.

So-called quartz analog wristwatches contain the smallest commonplace stepping motors; they have one coil, draw very little power, and have a permanent-magnet rotor. The same kind of motor drives battery-powered quartz clocks. Some of these watches, such as chronographs, contain more than one stepping motor.

Stepper motors were upscaled to be used in electric vehicles under the term SRM (Switched Reluctance Motor).

Linear motors

A linear motor is essentially an electric motor that has been "unrolled" so that, instead of producing a torque (rotation), it produces a straight-line force along its length by setting up a traveling electromagnetic field.

Linear motors are most commonly induction motors or stepper motors. You can find a linear motor in a maglev (Transrapid) train, where the train "flies" over the ground, and in many roller-coasters where the rapid motion of the motorless railcar is controlled by the rail. On a smaller scale, at least one letter-size (8.5" x 11") computer graphics X-Y pen plotter made by Hewlett-Packard (in the late 1970s to mid 1980's) used two linear stepper motors to move the pen along the two orthogonal axes.

Feeding and windings

Doubly-fed electric motor

Doubly-fed electric motors have two independent multiphase windings that actively participate in the energy conversion process with at least one of the winding sets electronically controlled for variable speed operation. Two is the most active multiphase winding sets possible without duplicating singly-fed or doubly-fed categories in the same package. As a result, doubly-fed electric motors are machines with an effective constant torque speed range that is twice synchronous speed for a given frequency of excitation. This is twice the constant torque speed range as singly-fed electric machines, which have only one active winding set.

A doubly-fed motor allows for a smaller electronic converter but the cost of the rotor winding and slip rings may offset the saving in the power electronics components. Difficulties with controlling speed near synchronous speed limit applications.

Singly-fed electric motor

Singly-fed electric motors incorporate a single multiphase winding set that is connected to a power supply. Singly-fed electric machines may be either induction or synchronous. The active winding set can be electronically controlled. Induction machines develop starting torque at zero speed and can operate as standalone machines. Synchronous machines must have auxiliary means for startup, such as a starting induction squirrel-cage winding or an electronic controller. Singly-fed electric machines have an effective constant torque speed range up to synchronous speed for a given excitation frequency.

The induction (asynchronous) motors (i.e., squirrel cage rotor or wound rotor), synchronous motors (i.e., field-excited, permanent magnet or brushless DC motors, reluctance motors, etc.), which are discussed on this page, are examples of singly-fed motors. By far, singly-fed motors are the predominantly installed type of motors.

Nanotube nanomotor

Researchers at University of California, Berkeley, recently developed rotational bearings based upon multiwall carbon nanotubes. By attaching a gold plate (with dimensions of the order of 100 nm) to the outer shell of a suspended multiwall carbon nanotube (like nested carbon cylinders), they are able to electrostatically rotate the outer shell relative to the inner core. These bearings are very robust; devices have been oscillated thousands of times with no indication of wear. These nanoelectromechanical systems (NEMS) are the next step in miniaturization and may find their way into commercial applications in the future.

Efficiency

To calculate a motor's efficiency, the mechanical output power is divided by the electrical input power:

$$\eta = \frac{P_m}{P_e},$$

where η is energy conversion efficiency, P_e is electrical input power, and P_m is mechanical output power.

In simplest case $P_e = VI$, and $P_m = T\omega$, where V is input voltage, I is input current, T is output torque, and ω is output angular velocity. It is possible to derive analytically the point of maximum efficiency. It is typically at less than 1/2 the stall torque.

Implications

Because a DC motor operates most efficiently at less than 1/2 its stall torque, an "oversized" motor runs with the highest efficiency: using a bigger motor than necessary enables the motor to operate closest to no load, or peak operating conditions.

Torque capability of motor types

When optimally designed for a given active current (i.e., torque current), voltage, pole-pair number, excitation frequency (i.e., synchronous speed), and core flux density, all categories of electric motors or generators will exhibit virtually the same maximum continuous shaft torque (i.e., operating torque) within a given physical size of electromagnetic core. Some applications require bursts of torque beyond the maximum operating torque, such as short bursts of torque to accelerate an electric vehicle from standstill. Always limited by magnetic core saturation or safe operating temperature rise and voltage, the capacity for torque bursts beyond the maximum operating torque differs significantly between categories of electric motors or generators.

Note: Capacity for bursts of torque should not be confused with Field Weakening capability inherent in fully electromagnetic electric machines (Permanent Magnet (PM) electric machine are excluded). Field Weakening, which is not readily available with PM electric machines, allows an electric machine to operate beyond the designed frequency of excitation without electrical damage.

Electric machines without a transformer circuit topology, such as Field-Wound (i.e., electromagnet) or Permanent Magnet (PM) Synchronous electric machines cannot realize bursts of torque higher than the maximum designed torque without saturating the magnetic core and rendering any increase in current as useless. Furthermore, the permanent magnet assembly of PM synchronous electric machines can be irreparably damaged, if bursts of torque exceeding the maximum operating torque rating are attempted.

Electric machines with a transformer circuit topology, such as Induction (i.e., asynchronous) electric machines, Induction Doubly-Fed electric machines, and Induction or Synchronous Wound-Rotor Doubly-Fed (WRDF) electric machines, exhibit very high bursts of torque because the active current (i.e., Magneto-Motive-Force or the product of current and winding-turns) induced on either side of the transformer oppose each other and as a result, the active current contributes nothing to the transformer coupled magnetic core flux density, which would otherwise lead to core saturation.

Electric machines that rely on Induction or Asynchronous principles short-circuit one port of the transformer circuit and as a result, the reactive impedance of the transformer circuit becomes dominant as slip increases, which limits the magnitude of active (i.e., real) current. Still, bursts of torque that are two to three times higher than the maximum design torque are realizable.

The Synchronous WRDF electric machine is the only electric machine with a truly dual ported transformer circuit topology (i.e., both ports independently excited with no short-circuited port). The dual ported transformer circuit topology is known to be unstable and requires a multiphase slip-ring-brush assembly to propagate limited power to the rotor winding set. If a precision means were available to instantaneously control torque angle and slip for synchronous operation during motoring or generating while simultaneously providing brushless power to the rotor winding set, the active current of the Synchronous WRDF electric machine would be independent of the reactive impedance of the transformer circuit and bursts of torque significantly higher than the maximum operating torque and far beyond the practical capability of any other type of electric machine would be realizable. Torque bursts greater than eight times operating torque have been calculated.

Materials

There is an impending shortage of many rare raw materials used in the manufacture of hybrid and electric cars (Nishiyama 2007) (Cox 2008). For example, the rare earth element dysprosium is required to fabricate many of the advanced electric motors used in hybrid cars (Cox 2008). However, over 95% of the world's rare earth elements are mined in China (Haxel et al. 2005), and domestic Chinese consumption is expected to consume China's entire supply by 2012 (Cox 2008).

While permanent magnet motors, favored in hybrids such as those made by Toyota, often use rare earth materials in their magnets, AC traction motors used in production electric vehicles such as the GM EV1, Toyota RAV4 EV and Tesla Roadster do not use permanent magnets or the associated rare earth materials. AC motors typically use conventional copper wire for their stator coils and copper or aluminum rods or bars for their rotor. AC motors do not significantly use rare earth materials.

Motor standards

The following are major design and manufacturing standards covering electric motors:

- International Electrotechnical Commission: IEC 60034 Rotating Electrical Machines
- National Electrical Manufacturers Association (USA): NEMA MG 1 Motors and Generators
- Underwriters Laboratories (USA): UL 1004 - Standard for Electric Motors

Uses

Electric motors are used in many, if not most, modern machines. Obvious uses would be in rotating machines such as fans, turbines, drills, the wheels on electric cars, locomotives and conveyor belts. Also, in many vibrating or oscillating machines, an electric motor

spins an irregular figure with more area on one side of the axle than the other, causing it to appear to be moving up and down.

Electric motors are also popular in robotics. They are used to turn the wheels of vehicular robots, and servo motors are used to turn arms and legs in humanoid robots. In flying robots, along with helicopters, a motor causes a propeller or wide, flat blades to spin and create lift force, allowing vertical motion.

Electric motors are replacing hydraulic cylinders in airplanes and military equipment.

In industrial and manufacturing businesses, electric motors are used to turn saws and blades in cutting and slicing processes, and to spin gears and mixers (the latter very common in food manufacturing). Linear motors are often used to push products into containers horizontally.

Many kitchen appliances also use electric motors to accomplish various jobs. Food processors and grinders spin blades to chop and break up foods. Blenders use electric motors to mix liquids, and microwave ovens use motors to turn the tray food sits on. Toaster ovens also use electric motors to turn a conveyor to move food over heating elements.

Chapter 4

Loudspeaker

An inexpensive, low fidelity 3½-inch **speaker**, typically found in small radios

A four-way, high fidelity **loudspeaker system**.

A **loudspeaker** (or "speaker") is an electroacoustic transducer that converts an electrical signal into sound. The speaker moves in accordance with the variations of an electrical signal and causes sound waves to propagate through a medium such as air or water.

After the acoustics of the listening space, loudspeakers (and other electroacoustic transducers) are the most variable elements in a modern audio system and are usually responsible for most distortion and audible differences when comparing sound systems.

Terminology

The term "loudspeaker" may refer to individual transducers (known as "drivers") or to complete speaker systems consisting of an enclosure including one or more drivers. To adequately reproduce a wide range of frequencies, most loudspeaker systems employ more than one driver, particularly for higher sound pressure level or maximum accuracy. Individual drivers are used to reproduce different frequency ranges. The drivers are named subwoofers (for very low frequencies); woofers (low frequencies); mid-range speakers (middle frequencies); tweeters (high frequencies); and sometimes *supertweeters*, optimized for the highest audible frequencies. The terms for different speaker drivers differ, depending on the application. In two-way systems there is no mid-range driver, so the task of reproducing the mid-range sounds falls upon the woofer and tweeter. Home stereos use the designation "tweeter" for the high frequency driver, while professional concert systems may designate them as "HF" or "highs". When multiple drivers are used in a system, a "filter network", called a crossover, separates the incoming signal into different frequency ranges and routes them to the appropriate driver. A loudspeaker system with *n* separate frequency bands is described as "*n*-way speakers": a two-way system will have a woofer and a tweeter; a three-way system employs a woofer, a mid-range, and a tweeter.

History

Johann Philipp Reis installed an electric loudspeaker in his telephone in 1861; it was capable of reproducing pure tones, but also could reproduce speech. Alexander Graham Bell patented his first electric loudspeaker (capable of reproducing intelligible speech) as part of his telephone in 1876, which was followed in 1877 by an improved version from Ernst Siemens. Nikola Tesla reportedly made a similar device in 1881, but he was not issued a patent. During this time, Thomas Edison was issued a British patent for a system using compressed air as an amplifying mechanism for his early cylinder phonographs, but he ultimately settled for the familiar metal horn driven by a membrane attached to the stylus. In 1898, Horace Short patented a design for a loudspeaker driven by compressed air; he then sold the rights to Charles Parsons, who was issued several additional British patents before 1910. A few companies, including the Victor Talking Machine Company and Pathé, produced record players using compressed-air loudspeakers. However, these designs were significantly limited by their poor sound quality and their inability to reproduce sound at low volume. Variants of the system were used for public address applications, and more recently, other variations have been used to test space-equipment resistance to the very loud sound and vibration levels that the launching of rockets produces.

The modern design of moving-coil (also called *dynamic*) drivers was established by Oliver Lodge in 1898. The first practical application of moving-coil loudspeakers was established by Peter L. Jensen and Edwin Pridham, in Napa, California. Jensen was denied patents. Being unsuccessful in selling their product to telephone companies, in 1915 they changed strategy to public address, and named their product Magnavox. Jensen

was, for years after the invention of the loudspeaker, a part owner of The Magnavox Company.

The moving-coil principle as commonly used today in direct radiators was patented in 1924 by Chester W. Rice and Edward W. Kellogg. The key difference between previous attempts and the patent by Rice and Kellogg was the adjustment of mechanical parameters so that the fundamental resonance of the moving system took place at a lower frequency than that at which the cone's radiation impedance had become uniform.

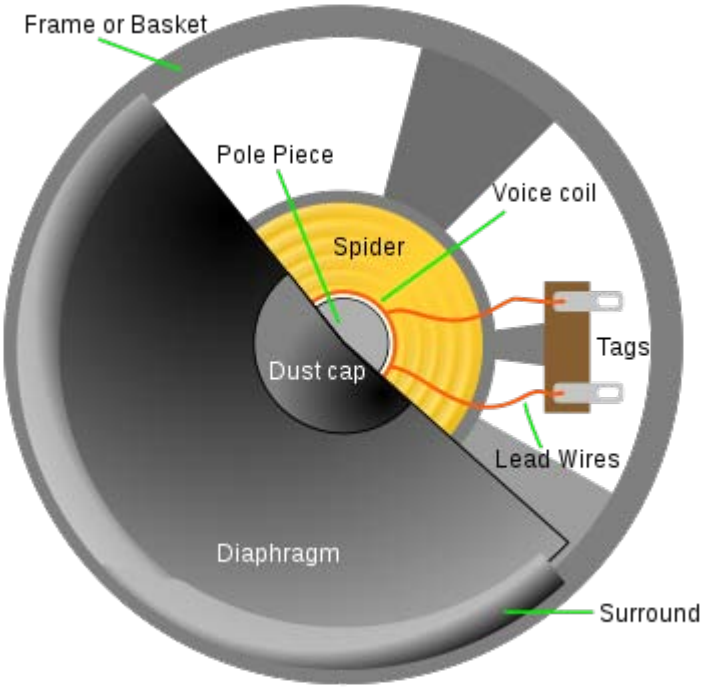
About this same period, Walter H. Schottky invented the first ribbon loudspeaker.

These first loudspeakers used electromagnets, because large, powerful permanent magnets were generally not available at a reasonable price. The coil of an electromagnet, called a field coil, was energized by current through a second pair of connections to the driver. This winding usually served a dual role, acting also as a choke coil, filtering the power supply of the amplifier to which the loudspeaker was connected. AC ripple in the current was attenuated by the action of passing through the choke coil; however, AC line frequencies tended to modulate the audio signal being sent to the voice coil and added to the audible hum of a powered-up sound reproduction device.

In the 1930s, loudspeaker manufacturers began to combine two and three bandpasses' worth of drivers in order to increase frequency response and sound pressure level. In 1937, the first film industry-standard loudspeaker system, "The Shearer Horn System for Theatres" (a two-way system), was introduced by Metro-Goldwyn-Mayer. It used four 15″ low-frequency drivers, a crossover network set for 375 Hz, and a single multi-cellular horn with two compression drivers providing the high frequencies. John Kenneth Hilliard, James Bullough Lansing, and Douglas Shearer all played roles in creating the system. At the 1939 New York World's Fair, a very large two-way public address system was mounted on a tower at Flushing Meadows. The eight 27″ low-frequency drivers were designed by Rudy Bozak in his role as chief engineer for Cinaudagraph. High-frequency drivers were likely made by Western Electric.

Altec Lansing introduced the '604', which was to become their most famous coaxial Duplex driver, in 1943, incorporating a high-frequency horn sending sound through the middle of a 15-inch woofer for near-point-source performance. Altec's "Voice of the Theatre" loudspeaker system arrived in the marketplace in 1945, offering better coherence and clarity at the high output levels necessary in movie theaters. The Academy of Motion Picture Arts and Sciences immediately began testing its sonic characteristics; they made it the film house industry standard in 1955. Subsequently, continuous developments in enclosure design and materials led to significant audible improvements. The most notable improvements in modern speakers are improvements in cone materials, the introduction of higher-temperature adhesives, improved permanent magnet materials, improved measurement techniques, computer-aided design, and finite element analysis.

Driver design



Cutaway view of a dynamic loudspeaker.

The most common type of driver uses a lightweight diaphragm, or *cone*, connected to a rigid *basket*, or *frame*, via a flexible suspension that constrains a coil of fine wire to move axially through a cylindrical magnetic gap. When an electrical signal is applied to the voice coil, a magnetic field is created by the electric current in the voice coil, making it a variable electromagnet. The coil and the driver's magnetic system interact, generating a mechanical force that causes the coil (and thus, the attached cone) to move back and forth, thereby reproducing sound under the control of the applied electrical signal coming from the amplifier. The following is a description of the individual components of this type of loudspeaker.

The diaphragm is usually manufactured with a cone- or dome-shaped profile. A variety of different materials may be used, but the most common are paper, plastic, and metal. The ideal material would be 1) rigid, to prevent uncontrolled cone motions; 2) have low mass, to minimize starting force requirements and energy storage issues; 3) be well damped, to reduce vibrations continuing after the signal has stopped with little or no audible ringing due to its resonance frequency as determined by its usage. In practice, all three of these criteria cannot be met simultaneously using existing materials; thus, driver design involves trade-offs. For example, paper is light and typically well damped, but is not stiff; metal may be stiff and light, but it usually has poor damping; plastic can be light, but typically, the stiffer it is made, the poorer the damping. As a result, many cones are made of some sort of composite material. For example, a cone might be made of cellulose paper, into which some carbon fiber, Kevlar, fiberglass, hemp or bamboo fibers have

been added; or it might use a honeycomb sandwich construction; or a coating might be applied to it so as to provide additional stiffening or damping.

A stamped steel loudspeaker basket frame is clearly visible (here, blue-grey).

The chassis, frame, or basket, is designed to be rigid, avoiding deformation which would change critical alignments with the magnet gap, perhaps causing the voice coil to rub against the sides of the gap. Chassis are typically cast from aluminum alloy, or stamped from thin steel sheet, although molded plastic and damped plastic compound baskets are becoming common, especially for inexpensive, low-mass drivers. Metallic chassis can play an important role in conducting heat away from the voice coil; heating during operation changes resistance, causing physical dimensional changes, and if extreme, may even demagnetize permanent magnets.

The suspension system keeps the coil centered in the gap and provides a restoring (centering) force that returns the cone to a neutral position after moving. A typical suspension system consists of two parts: the "spider", which connects the diaphragm or

voice coil to the frame and provides the majority of the restoring force, and the "surround", which helps center the coil/cone assembly and allows free pistonic motion aligned with the magnetic gap. The spider is usually made of a corrugated fabric disk, impregnated with a stiffening resin. The name comes from the shape of early suspensions, which were two concentric rings of Bakelite material, joined by six or eight curved "legs". Variations of this topology included the addition of a felt disc to provide a barrier to particles that might otherwise cause the voice coil to rub. The German firm Rulik still offers drivers with uncommon spiders made of wood.

The cone surround can be rubber or polyester foam, or a ring of corrugated, resin coated fabric; it is attached to both the outer diaphragm circumference and to the frame. These different surround materials, their shape and treatment can dramatically affect the acoustic output of a driver; each class and implementation having advantages and disadvantages. Polyester foam, for example, is lightweight and economical, but is degraded by exposure to ozone, UV light, humidity and elevated temperatures, limiting its useful life to about 15 years.

The wire in a voice coil is usually made of copper, though aluminum—and, rarely, silver—may be used. Voice-coil wire cross sections can be circular, rectangular, or hexagonal, giving varying amounts of wire volume coverage in the magnetic gap space. The coil is oriented co-axially inside the gap; it moves back and forth within a small circular volume (a hole, slot, or groove) in the magnetic structure. The gap establishes a concentrated magnetic field between the two poles of a permanent magnet; the outside of the gap being one pole, and the center post (called the pole piece) being the other. The pole piece and backplate are often a single piece, called the poleplate or yoke.

Modern driver magnets are almost always permanent and made of ceramic, ferrite, Alnico, or, more recently, rare earth such as neodymium and Samarium cobalt. A trend in design—due to increases in transportation costs and a desire for smaller, lighter devices (as in many home theater multi-speaker installations)—is the use of the last instead of heavier ferrite types. Very few manufacturers still use electrically powered field coils, as was common in the earliest designs (one such is French). When high field-strength permanent magnets became available, Alnico, an alloy of aluminum, nickel, and cobalt became popular, since it dispensed with the power supply issues of field-coil drivers. Alnico was used for almost exclusively until about 1980. Alnico magnets can be partially degaussed (i.e., demagnetized) by accidental 'pops' or 'clicks' caused by loose connections, especially if used with a high power amplifier. This damage can be reversed by "recharging" the magnet.

After 1980, most (but not quite all) driver manufacturers switched from Alnico to ferrite magnets, which are made from a mix of ceramic clay and fine particles of barium or strontium ferrite. Although the energy per kilogram of these ceramic magnets is lower than Alnico, it is substantially less expensive, allowing designers to use larger yet more economical magnets to achieve a given performance.

The size and type of magnet and details of the magnetic circuit differ, depending on design goals. For instance, the shape of the pole piece affects the magnetic interaction between the voice coil and the magnetic field, and is sometimes used to modify a driver's behavior. A "shorting ring", or Faraday loop, may be included as a thin copper cap fitted over the pole tip or as a heavy ring situated within the magnet-pole cavity. The benefits of this complication is reduced impedance at high frequencies, providing extended treble output, reduced harmonic distortion, and a reduction in the inductance modulation that typically accompanies large voice coil excursions. On the other hand, the copper cap requires a wider voice-coil gap, with increased magnetic reluctance; this reduces available flux, requiring a larger magnet for equivalent performance.

Driver design—including the particular way two or more drivers are combined in an enclosure to make a speaker system—is both an art and science. Adjusting a design to improve performance is done using some combination of magnetic, acoustic, mechanical, electrical, and material science theory; high precision measurements, generally with the observations of experienced listeners. A few of the issues speaker and driver designers must confront are distortion, lobing, phase effects, off-axis response, and crossover complications. Designers can use an anechoic chamber to ensure the speaker can be measured independently of room effects, or any of several electronic techniques which can, to some extent, replace such chambers. Some developers eschew anechoic chambers in favor of specific standardized room setups intended to simulate real-life listening conditions.

The fabrication of finished loudspeaker systems has become segmented, depending largely on price, shipping costs, and weight limitations. High-end speaker systems, which are typically heavier (and often larger) than economic shipping allows outside local regions, are usually made in their target market region and can cost $140,000 or more per pair. The lowest-priced speaker systems and most drivers are manufactured in China or other low-cost manufacturing locations.

Driver types

Individual electrodynamic drivers provide optimal performance within a limited pitch range. Multiple drivers (e.g., subwoofers, woofers, mid-range drivers, and tweeters) are generally combined into a complete loudspeaker system to provide performance beyond that constraint.

Full-range drivers

A full-range driver is designed to have the widest frequency response possible. These drivers are small, typically 3 to 8 inches (7.6 to 20 cm) in diameter to permit reasonable high frequency response, and carefully designed to give low-distortion output at low frequencies, though with reduced maximum output level. Full-range (or more accurately, wide-range) drivers are most commonly heard in public address systems, in televisions (although some models are suitable for hi-fi listening), small radios, intercoms, some computer speakers, etc. In hi-fi speaker systems, the use of wide-range drive units can

avoid undesirable interactions between multiple drivers caused by non-coincident driver location or crossover network issues. Fans of wide-range driver hi-fi speaker systems claim a coherence of sound, said to be due to the single source and a resulting lack of interference, and likely also to the lack of crossover components. Detractors typically cite wide-range drivers' limited frequency response and modest output abilities (most especially at low frequencies), together with their requirement for large, elaborate, expensive enclosures—such as transmission lines, or horns—to approach optimum performance.

Full-range drivers often employ an additional cone called a *whizzer*: a small, light cone attached to the joint between the voice coil and the primary cone. The whizzer cone extends the high-frequency response of the driver and broadens its high frequency directivity, which would otherwise be greatly narrowed due to the outer diameter cone material failing to keep up with the central voice coil at higher frequencies. The main cone in a whizzer design is manufactured so as to flex more in the outer diameter than in the center. The result is that the main cone delivers low frequencies and the whizzer cone contributes most of the higher frequencies. Since the whizzer cone is smaller than the main diaphragm, output dispersion at high frequencies is improved relative to an equivalent single larger diaphragm.

Limited-range drivers, also used alone, are typically found in computers, toys, and clock radios. These drivers are less elaborate and less expensive than wide-range drivers, and they may be severely compromised to fit into very small mounting locations. In these applications, sound quality is a low priority. The human ear is remarkably tolerant of poor sound quality, and the distortion inherent in limited-range drivers may enhance their output at high frequencies, increasing clarity when listening to spoken word material.

Subwoofer

A subwoofer is a woofer driver used only for the lowest part of the audio spectrum: typically below 200 Hz for consumer systems, below 100 Hz for professional live sound, and below 80 Hz in THX-approved systems. Because the intended range of frequencies is limited, subwoofer system design is usually simpler in many respects than for conventional loudspeakers, often consisting of a single driver enclosed in a suitable box or enclosure

To accurately reproduce very low bass notes without unwanted resonances (typically from cabinet panels), subwoofer systems must be solidly constructed and properly braced; good speakers are typically quite heavy. Many subwoofer systems include power amplifiers and electronic sub-filters, with additional controls relevant to low-frequency reproduction. These variants are known as "active subwoofers". In contrast, "passive" subwoofers require external amplification.

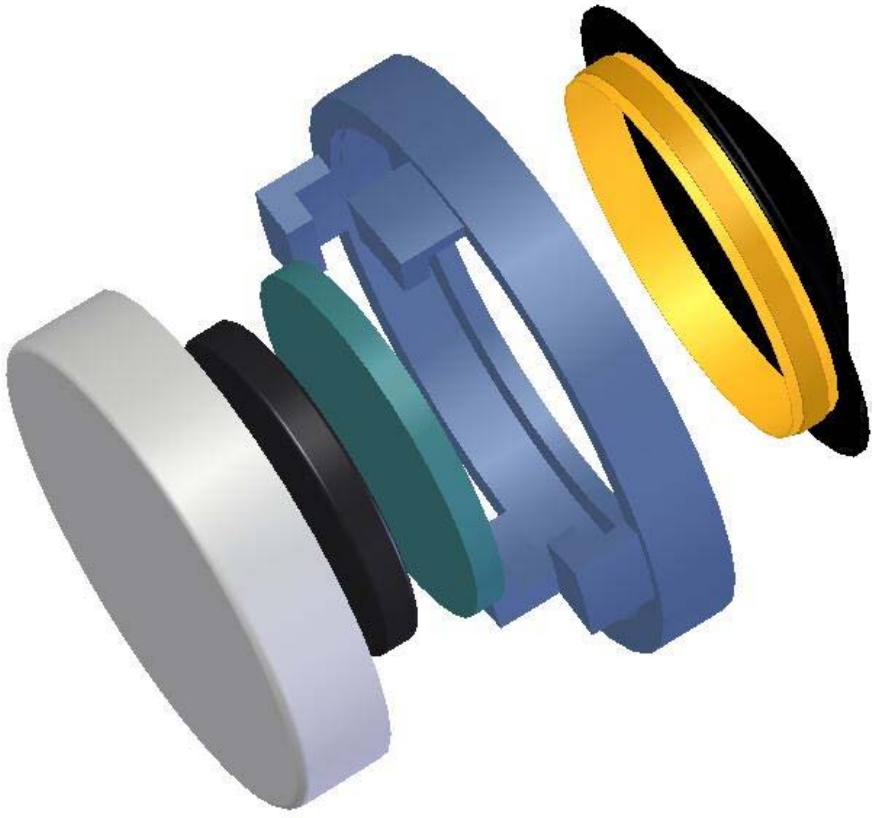
Woofer

A woofer is a driver that reproduces low frequencies. The driver combines with the enclosure design to produce suitable low frequencies. Some loudspeaker systems use a woofer for the lowest frequencies, sometimes well enough that a subwoofer is not needed. Additionally, some loudspeakers use the woofer to handle middle frequencies, eliminating the mid-range driver. This can be accomplished with the selection of a tweeter that can work low enough that, combined with a woofer that responds high enough, the two drivers add coherently in the middle frequencies.

Mid-range driver

A mid-range speaker is a loudspeaker driver that reproduces middle frequencies. Mid-range driver diaphragms can be made of paper or composite materials, and can be direct radiation drivers (rather like smaller woofers) or they can be compression drivers (rather like some tweeter designs). If the mid-range driver is a direct radiator, it can be mounted on the front baffle of a loudspeaker enclosure, or, if a compression driver, mounted at the throat of a horn for added output level and control of radiation pattern.

Tweeter



Exploded view of a dome tweeter.

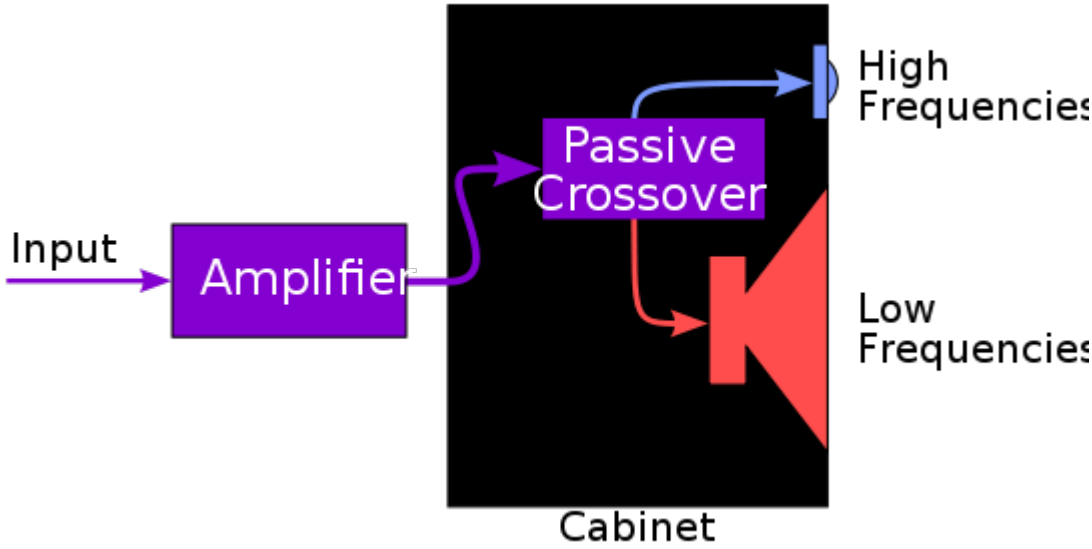
A tweeter is a high-frequency driver that reproduces the highest frequencies in a speaker system. Many varieties of tweeter design exist, each with differing abilities with regard to frequency response, output fidelity, power handling, maximum output level, etc. Soft-dome tweeters are widely found in home stereo systems, and horn-loaded compression drivers are common in professional sound reinforcement. Ribbon tweeters have gained popularity in recent years, as their output power has been increased to levels useful for professional sound reinforcement, and their output pattern is wide in the horizontal plane, a pattern that has convenient applications in concert sound.

Coaxial drivers

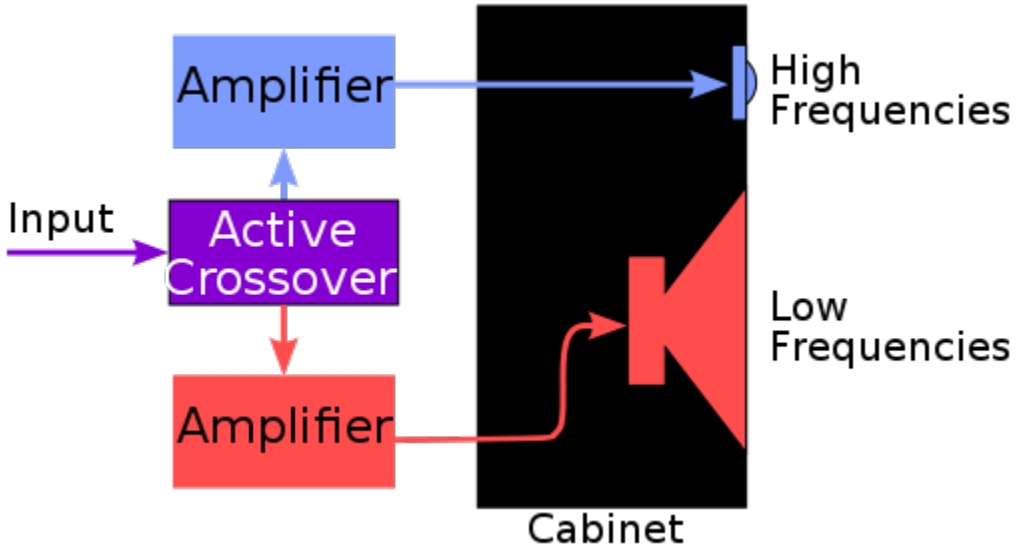
A coaxial driver is a loudspeaker driver with two or several combined concentric drivers. Coaxial drivers have been produced by many companies, such as Altec, Tannoy, Pioneer, KEF, BMS, Cabasse and Genelec.

Loudspeaker system design

Crossover



A passive crossover.



Bi-amped.

Used in multi-driver speaker systems, the crossover is a subsystem that separates the input signal into different frequency ranges suited to each driver. The drivers receive only the power in their usable frequency range (the range they were designed for), thereby reducing distortion in the drivers and interference between them.

Crossovers can be *passive* or *active*. A passive crossover is an electronic circuit that uses a combination of one or more resistors, inductors, or non-polar capacitors. These parts are formed into carefully designed networks and are most often placed between the power amplifier and the loudspeaker drivers to divide the amplifier's signal into the necessary frequency bands before being delivered to the individual drivers. Passive crossover circuits need no external power beyond the audio signal itself, but do cause overall signal loss and a significant reduction in damping factor between the voice coil and the crossover. An active crossover is an electronic filter circuit that divides the signal into individual frequency bands *before* power amplification, thus requiring at least one power amplifier for each bandpass. Passive filtering may also be used in this way before power amplification, but it is an uncommon solution, due to inflexibility compared to active filtering. Any technique that uses crossover filtering followed by amplification is commonly known as bi-amping, tri-amping, quad-amping, and so on, depending on the minimum number of amplifier channels. Some loudspeaker designs use a combination of passive and active crossover filtering, such as a passive crossover between the mid- and high-frequency drivers and an active crossover between the low-frequency driver and the combined mid- and high frequencies.

Passive crossovers are commonly installed inside speaker boxes and are by far the most usual type of crossover for home and low-power use. In car audio systems, passive crossovers may be in a separate box, necessary to accommodate the size of the components used. Passive crossovers may be simple for low-order filtering, or complex

to allow steep slopes such as 18 or 24 dB per octave. Passive crossovers can also be designed to compensate for undesired characteristics of driver, horn, or enclosure resonances, and can be tricky to implement, due to component interaction. Passive crossovers, like the driver units that they feed, have power handling limits, have insertion losses (10% is often claimed), and change the load seen by the amplifier. The changes are matters of concern for many in the hi-fi world. When high output levels are required, active crossovers may be preferable. Active crossovers may be simple circuits that emulate the response of a passive network, or may be more complex, allowing extensive audio adjustments. Some active crossovers, usually digital loudspeaker management systems, may include facilities for precise alignment of phase and time between frequency bands, equalization, and dynamics (compression and limiting) control.

Some hi-fi and professional loudspeaker systems now include an active crossover circuit as part of an onboard amplifier system. These speaker designs are identifiable by their need for AC power in addition to a signal cable from a pre-amplifier. This active topology may include driver protection circuits and other features of a digital loudspeaker management system. Powered speaker systems are common in computer sound (for a single listener) and, at the other end of the size spectrum, in modern concert sound systems, where their presence is significant and steadily increasing.

Enclosures

An unusual three-way speaker system. The cabinet is narrow in order to reduce a diffraction effect called the "baffle step".

Most loudspeaker systems consist of drivers mounted in an enclosure, or cabinet. The role of the enclosure is to provide a place to physically mount the drivers, and to prevent sound waves emanating from the back of a driver from interfering destructively with those from the front; these typically cause cancellations (e.g., comb filtering) and significantly alter the level and quality of sound at low frequencies.

The simplest driver mount is a flat panel (i.e., baffle) with the drivers mounted in holes in it. However, in this approach, sound frequencies with a wavelength longer than the baffle dimensions are canceled out, because the antiphase radiation from the rear of the cone

interferes with the radiation from the front. With an infinitely large panel, this interference could be entirely prevented. A sufficiently large sealed box can approach this behavior.

Since panels of infinite dimensions are impractical, most enclosures function by containing the rear radiation from the moving diaphragm. A sealed enclosure prevents transmission of the sound emitted from the rear of the loudspeaker by confining the sound in a rigid and airtight box. Techniques used to reduce transmission of sound through the walls of the cabinet include thicker cabinet walls, lossy wall material, internal bracing, curved cabinet walls—or more rarely, visco-elastic materials (e.g., mineral-loaded bitumen) or thin lead sheeting applied to the interior enclosure walls.

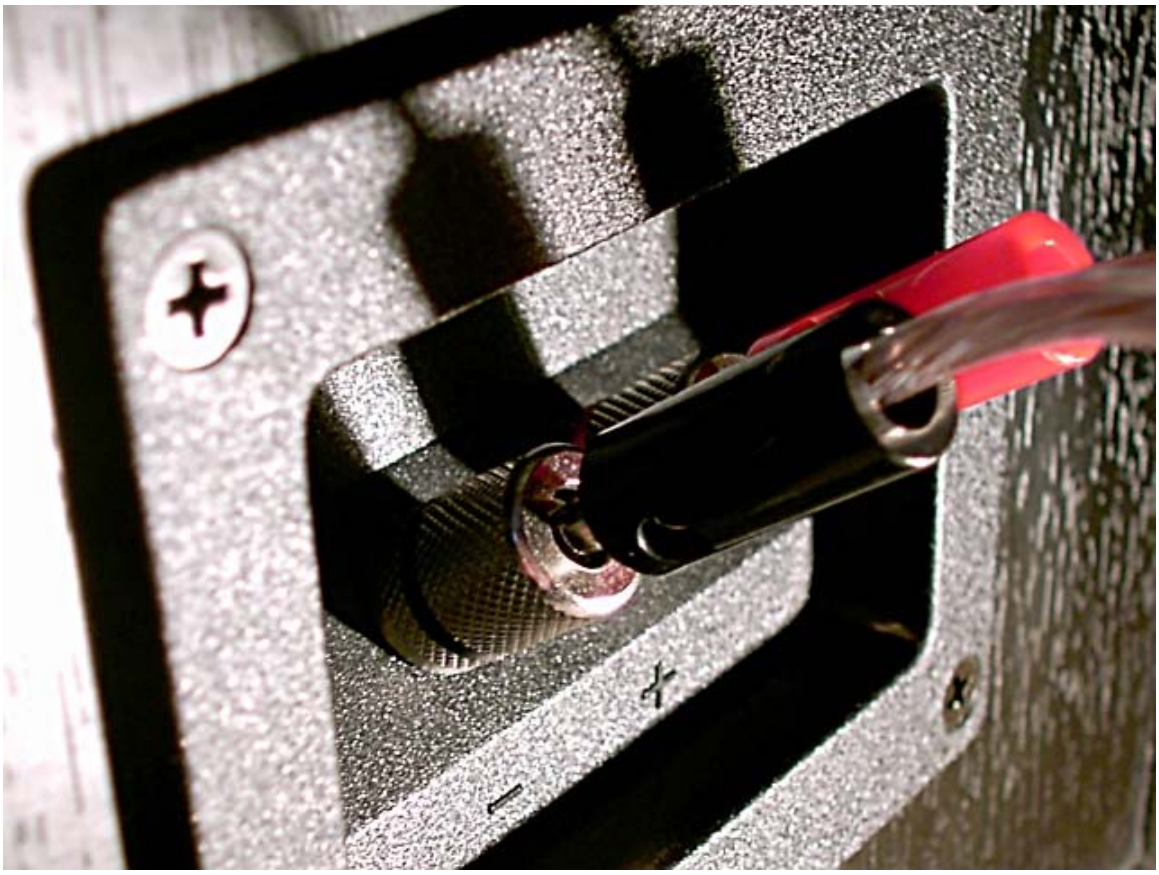
However, a rigid enclosure reflects sound internally, which can then be transmitted back through the loudspeaker diaphraghm—again resulting in degradation of sound quality. This can be reduced by internal absorption using absorptive materials (often called "damping"), such as fiberglass, wool, or synthetic fiber batting, within the enclosure. The internal shape of the enclosure can also be designed to reduce this by reflecting sounds away from the loudspeaker diaphragm, where they may then be absorbed.

Other enclosure types alter the rear sound radiation so it can add constructively to the output from the front of the cone. Designs that do this (including *bass reflex*, *passive radiator*, *transmission line*, etc.) are often used to extend the effective low-frequency response and increase low-frequency output of the driver.

To make the transition between drivers as seamless as possible, system designers have attempted to time-align (or phase adjust) the drivers by moving one or more driver mounting locations forward or back so that the acoustic center of each driver is in the same vertical plane. This may also involve tilting the face speaker back, providing a separate enclosure mounting for each driver, or (less commonly) using electronic techniques to achieve the same effect. These attempts have resulted in some unusual cabinet designs.

The speaker mounting scheme (including cabinets) can also cause diffraction, resulting in peaks and dips in the frequency response. The problem is usually greatest at higher frequencies, where wavelengths are similar to, or smaller than, cabinet dimensions. The effect can be minimized by rounding the front edges of the cabinet, curving the cabinet itself, using a smaller or narrower enclosure, choosing a strategic driver arrangement, using absorptive material around a driver, or some combination of these and other schemes.

Wiring connections



Two-way binding posts on a loudspeaker, connected using banana plugs.

A 4-ohm loudspeaker with two pairs of binding posts capable of accepting bi-wiring after the removal of two metal straps.

Most loudspeakers use two wiring points to connect to the source of the signal (for example, to the audio amplifier or receiver). This is usually done using binding posts or spring clips on the back of the enclosure. If the wires for the left and right speakers (in a stereo setup) are not connected "in phase" with each other (the + and − connections on the speaker and amplifier should be connected + to + and − to −), the loudspeakers will be out of polarity. Given identical signals, motion in one cone will be in the opposite direction of the other. This will typically cause monophonic material within a stereo recording to be canceled out, reduced in level, and made more difficult to localize, all due to destructive interference of the sound waves. The cancellation effect is most noticeable at frequencies where the speakers are separated by a quarter wavelength or less; low frequencies are affected the most. This type of wiring error doesn't damage speakers, but isn't optimal.

Specifications

Specifications label on a loudspeaker.

Speaker specifications generally include:

- **Speaker or driver type** (individual units only) – Full-range, woofer, tweeter, or mid-range.
- **Size** of individual drivers. For cone drivers, the quoted size is generally the outside diameter of the basket. However, it may less commonly also be the diameter of the cone surround, measured apex to apex, or the distance from the center of one mounting hole to its opposite. Voice-coil diameter may also be specified. If the loudspeaker has a compression horn driver, the diameter of the horn throat may be given.
- **Rated Power** – Nominal (or even continuous) power, and peak (or maximum short-term) power a loudspeaker can handle (i.e., maximum input power before destroying the loudspeaker; it is never the sound output the loudspeaker produces). A driver may be damaged at much less than its rated power if driven past its mechanical limits at lower frequencies. Tweeters can also be damaged by amplifier clipping (amplifier circuits produce large amounts of energy at high frequencies in such cases) or by music or sine wave input at high frequencies. Each of these situations passes more energy to a tweeter than it can survive without damage. In some jurisdictions, power handling has a legal meaning

allowing comparisons between loudspeakers under consideration. Elsewhere, the variety of meanings for power handling capacity can be quite confusing.
- **Impedance** – typically 4 Ω (ohms), 8 Ω, etc.
- **Baffle or enclosure type** (enclosed systems only) – Sealed, bass reflex, etc.
- **Number of drivers** (complete speaker systems only) – two-way, three-way, etc.

and optionally:

- **Crossover frequency(ies)** (multi-driver systems only) – The nominal frequency boundaries of the division between drivers.
- **Frequency response** – The measured, or specified, output over a specified range of frequencies for a constant input level varied across those frequencies. It sometimes includes a variance limit, such as within "± 2.5 dB".
- **Thiele/Small parameters** (individual drivers only) – these include the driver's F_s (resonance frequency), Q_{ts} (a driver's Q (more or less, its damping factor at resonant frequency), V_{as} (the equivalent air compliance volume of the driver), etc.
- **Sensitivity** – The sound pressure level produced by a loudspeaker in a non-reverberant environment, often specified in dB and measured at 1 meter with an input of 1 watt (2.83 rms volts into 8 Ω), typically at one or more specified frequencies. This rating is often specified by manufacturers to be impressive.
- **Maximum SPL** – The highest output the loudspeaker can manage, short of damage or not exceeding a particular distortion level. This rating is often specified by manufacturers to be impressive, and is commonly given without reference to frequency range or distortion level.

Electrical characteristics of a dynamic loudspeaker

The load that a driver presents to an amplifier consists of a complex electrical impedance—a combination of resistance and both capacitive and inductive reactance, which combines properties of the driver, its mechanical motion, the effects of crossover components (if any are in the signal path between amplifier and driver), and the effects of air loading on the driver as modified by the enclosure and its environment. Most amplifiers' output specifications are given at a specific power into an ideal resistive load; however, a loudspeaker does not have a constant resistance across its frequency range. Instead, the voice coil is inductive, the driver has mechanical resonances, the enclosure changes the driver's electrical and mechanical characteristics, and a passive crossover between the drivers and the amplifier contributes its own variations. The result is a load resistance that varies fairly widely with frequency, and usually a varying phase relationship between voltage and current as well, also changing with frequency. Some amplifiers can cope with the variation better than others can.

To make sound, a loudspeaker is driven by modulated electrical current (produced by an amplifier) that pass through a "speaker coil" (a coil of copper wire), which then (through resistance and other forces) magnetizes the coil, creating a magnetic field. The electrical current variations that pass through the speaker are thus converted to varying magnetic

forces, which move the speaker diaphraghm, which thus forces the driver to produce air motion that is similar to the original signal from the amplifier.

Electromechanical measurements

Fully characterizing the sound output quality of a loudspeaker driver or system in words is essentially impossible. Objective measurements provide information about several aspects of performance so that informed comparisons and improvements can be made, but no combination of measurements summarizes the performance of a loudspeaker system in use, if only because the test signals used are neither music nor speech.. Examples of typical measurements are: amplitude and phase characteristics vs. frequency; impulse response under one or more conditions (e.g., square waves, sine wave bursts, etc.); directivity vs. frequency (e.g., horizontally, vertically, spherically, etc.); harmonic and intermodulation distortion vs. SPL output, using any of several test signals; stored energy (i.e., ringing) at various frequencies; impedance vs. frequency; and small-signal vs. large-signal performance. Most of these measurements require sophisticated and often expensive equipment to perform, and also good judgment by the operator, but the raw sound pressure level output is rather easier to report and so is often the only specified value—sometimes in misleadingly exact terms. The sound pressure level (SPL) a loudspeaker produces is measured in decibels (dB_{spl}).

Efficiency vs. sensitivity

Loudspeaker efficiency is defined as the sound power output divided by the electrical power input. Most loudspeakers are actually very inefficient transducers; only about 1% of the electrical energy sent by an amplifier to a typical home loudspeaker is converted to acoustic energy. The remainder is converted to heat, mostly in the voice coil and magnet assembly. The main reason for this is the difficulty of achieving proper impedance matching between the acoustic impedance of the drive unit and that of the air into which it is radiating (at low frequencies improving this match is the main purpose of speaker enclosure designs). The efficiency of loudspeaker drivers varies with frequency as well. For instance, the output of a woofer driver decreases as the input frequency decreases because of the increasingly poor match between air and the driver.

Driver ratings based on the SPL for a given input are called sensitivity ratings and are notionally similar to efficiency. Sensitivity is usually defined as so many decibels at 1 W electrical input, measured at 1 meter (except for headphones), often at a single frequency. The voltage used is often 2.83 V_{RMS}, which is 1 watt into an 8 Ω (nominal) speaker impedance (approximately true for many speaker systems). Measurements taken with this reference are quoted as dB with 2.83 V @ 1 m.

The sound pressure output is measured at (or mathematically scaled to be equivalent to a measurement taken at) one meter from the loudspeaker and on-axis (directly in front of it), under the condition that the loudspeaker is radiating into an infinitely large space and mounted on an infinite baffle. Clearly then, sensitivity does not correlate precisely with efficiency, as it also depends on the directivity of the driver being tested and the acoustic

environment in front of the actual loudspeaker. For example, a cheerleader's horn produces more sound output in the direction it is pointed by concentrating sound waves from the cheerleader in one direction, thus "focusing" them. The horn also improves impedance matching between the voice and the air, which produces more acoustic power for a given speaker power. In some cases, improved impedance matching (via careful enclosure design) will allow the speaker to produce more acoustic power.

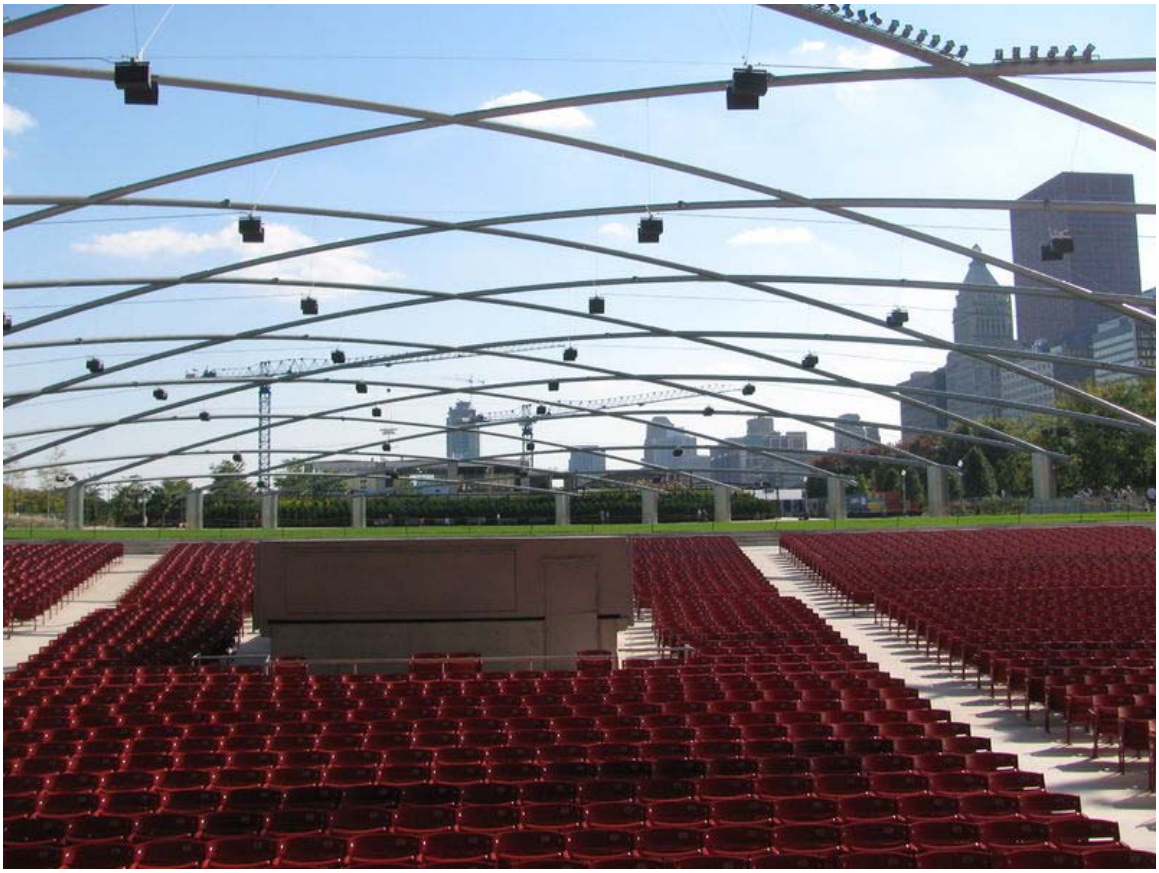
- Typical home loudspeakers have sensitivities of about 85 to 95 dB for 1 W @ 1 m—an efficiency of 0.5–4%.
- Sound reinforcement and public address loudspeakers have sensitivities of perhaps 95 to 102 dB for 1 W @ 1 m—an efficiency of 4–10%.
- Rock concert, stadium PA, marine hailing, etc. speakers generally have higher sensitivities of 103 to 110 dB for 1 W @ 1 m—an efficiency of 10–20%.

A driver with a higher maximum power rating cannot necessarily be driven to louder levels than a lower-rated one, since sensitivity and power handling are largely independent properties. In the examples that follow, assume (for simplicity) that the drivers being compared have the same electrical impedance; are operated at the same frequency within both driver's respective pass bands; and that power compression and distortion are low. For the first example, a speaker 3 dB more sensitive than another will produce double the sound power (or be 3 dB louder) for the same power input; thus, a 100 W driver ("A") rated at 92 dB for 1 W @ 1 m sensitivity will put out twice as much acoustic power as a 200 W driver ("B") rated at 89 dB for 1 W @ 1 m when both are driven with 100 W of input power. In this particular example, when driven at 100 W, speaker A will produce the same SPL, or loudness, as speaker B would produce with 200 W input. Thus, a 3 dB increase in sensitivity of the speaker means that it will need half the amplifier power to achieve a given SPL. This translates into a smaller, less complex power amplifier—and often, to reduced overall system cost.

It is not possible to combine high efficiency (especially at low frequencies) with compact enclosure size and adequate low frequency response. One can, more or less, choose only two of the three parameters when designing a speaker system. So, for example, if extended low-frequency performance and small box size are important, one must accept low efficiency. This rule of thumb is sometimes called Hoffman's Iron Law (after J.A. Hoffman, the "H" in KLH).

Listening environment

Jay Pritzker Pavilion



At Jay Pritzker Pavilion, a LARES system is combined with a zoned sound reinforcement system, both suspended on an overhead steel trellis, to synthesize an indoor acoustic environment outdoors.

The interaction of a loudspeaker system with its environment is complex and is largely out of the loudspeaker designer's control. Most listening rooms present a more or less reflective environment, depending on size, shape, volume, and furnishings. This means the sound reaching a listener's ears consists not only of sound directly from the speaker system, but also the same sound delayed by traveling to and from (and being modified by) one or more surfaces. These reflected sound waves, when added to the direct sound, cause cancellation and addition at assorted frequencies (e.g., from resonant room modes), thus changing the timbre and character of the sound at the listener's ears. The human brain is very sensitive to small variations, including some of these, and this is part of the

reason why a loudspeaker system sounds different at different listening positions or in different rooms.

A significant factor in the sound of a loudspeaker system is the amount of absorption and diffusion present in the environment. Clapping one's hands in a typical empty room, without draperies or carpet, will produce a zippy, fluttery echo which is due both to a lack of absorption and to reverberation (that is, repeated echoes) from flat reflective walls, floor, and ceiling. The addition of hard surfaced furniture, wall hangings, shelving and even baroque plaster ceiling decoration, will change the echoes, due primarily to the diffusion caused by somewhat reflective objects with shapes and surfaces having sizes on the order of the sound wavelengths being diffused. This somewhat breaks up the simple reflections otherwise caused by bare flat surfaces, and spreads the reflected energy of an incident wave over a larger angle on reflection.

Placement

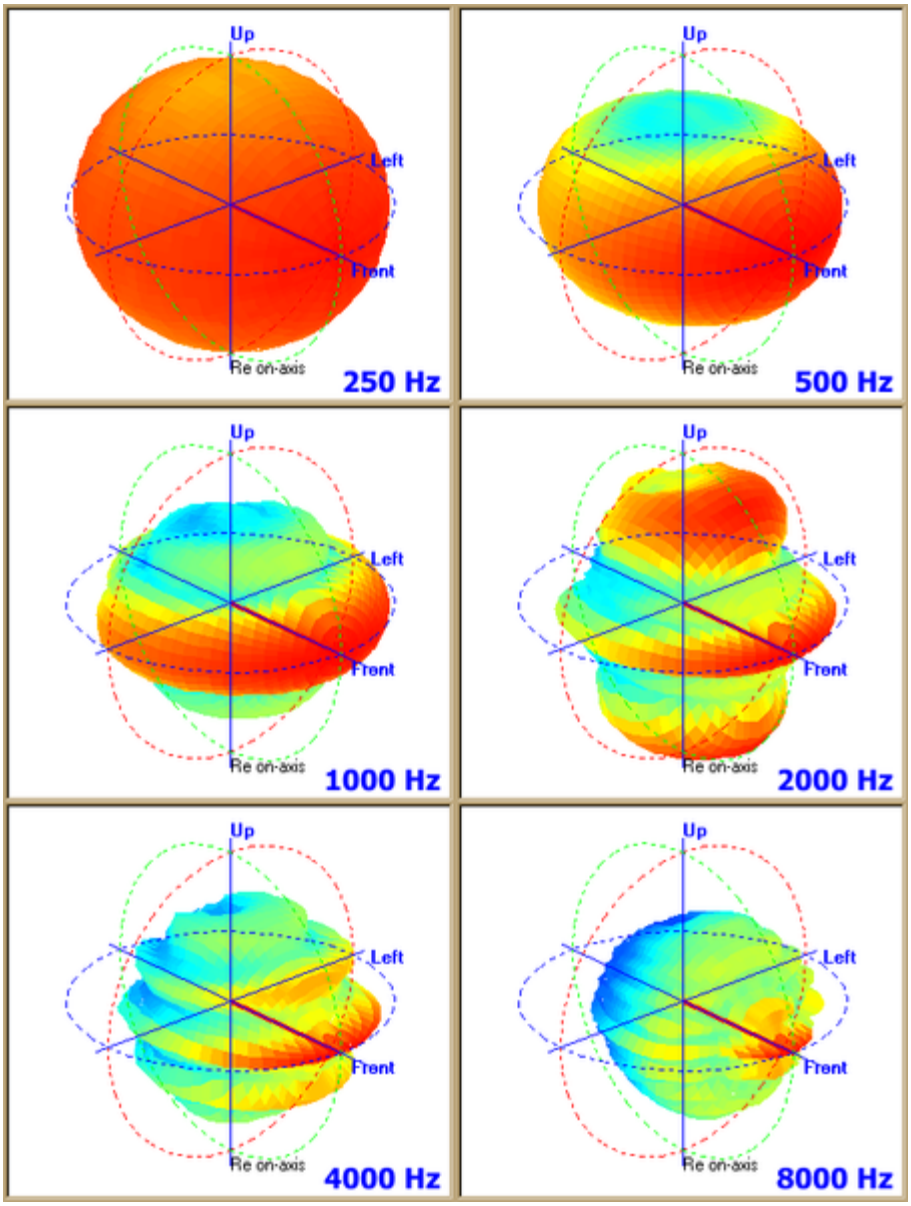
In a typical rectangular listening room, the hard, parallel surfaces of the walls, floor and ceiling cause primary acoustic resonance nodes in each of the three dimensions: left-right, up-down and forward-backward. Furthermore, there are more complex resonance modes involving three, four, five and even all six boundary surfaces combining to create standing waves. Low frequencies excite these modes the most, since long wavelengths are not much affected by furniture compositions or placement. The mode spacing is critical, especially in small and medium size rooms like recording studios, home theaters and broadcast studios. The proximity of the loudspeakers to room boundaries affects how strongly the resonances are excited as well as affecting the relative strength at each frequency. The location of the listener is critical, too, as a position near a boundary can have a great effect on the perceived balance of frequencies. This is because standing wave patterns are most easily heard in these locations and at lower frequencies, below the Schroeder frequency – typically around 200–300 Hz, depending on room size.

Directivity

Acousticians, in studying the radiation of sound sources have developed some concepts important to understanding how loudspeakers are perceived. The simplest possible radiating source is a point source, sometimes called a simple source. An ideal point source is an infinitesimally small point radiating sound. It may be easier to imagine a tiny pulsating sphere, uniformly increasing and decreasing in diameter, sending out sound waves in all directions equally, independent of frequency.

Any object radiating sound, including a loudspeaker system, can be thought of as being composed of combinations of such simple point sources. The radiation pattern of a combination of point sources will not be the same as for a single source, but rather will depend on the distance and orientation between the sources, the position relative to them from which the listener hears the combination, and the frequency of the sound involved. Using geometry and calculus, some simple combinations of sources are easily solved; others are not.

One simple combination is two simple sources separated by a distance and vibrating out of phase, one miniature sphere expanding while the other is contracting. The pair is known as a doublet, or dipole, and the radiation of this combination is similar to that of a very small dynamic loudspeaker operating without a baffle. The directivity of a dipole is a figure 8 shape with maximum output along a vector which connects the two sources and minimums to the sides when the observing point is equidistant from the two sources, where the sum of the positive and negative waves cancel each other. While most drivers are dipoles, depending on the enclosure to which they are attached, they may radiate as monopoles, dipoles (or bipoles). If mounted on a finite baffle, and these out of phase waves allowed to interact, dipole peaks and nulls in the frequency response result. When the rear radiation is absorbed or trapped in a box, the diaphragm becomes a monopole radiator. Bipolar speakers, made by mounting in-phase monopoles (both moving out of or into the box in unison) on opposite sides of a box, are a method of approaching omnidirectional radiation patterns.



Polar plots of a four-driver industrial columnar public address loudspeaker taken at six frequencies. Note how the pattern is nearly omnidirectional at low frequencies, converging to a wide fan-shaped pattern at 1 kHz, then separating into lobes and getting weaker at higher frequencies

In real life, individual drivers are actually complex 3D shapes such as cones and domes, and they are placed on a baffle for various reasons. A mathematical expression for the directivity of a complex shape, based on modeling combinations of point sources, is usually not possible, but in the farfield, the directivity of a loudspeaker with a circular diaphragm will be close to that of a flat circular piston, so it can be used as an illustrative simplification for discussion. As a simple example of the mathematical physics involved, consider the following: the formula for farfield directivity of a flat circular piston in an

infinite baffle is $p(\theta) = \frac{p_0 J_1(k_a \sin\theta)}{k_a \sin\theta}$ where $k_a = \frac{2\pi a}{\lambda}$, p_0 is the pressure on axis, a is the piston radius, λ is the wavelength (i.e. $\lambda = \frac{c}{f} = \frac{\text{speed of sound}}{\text{frequency}}$) θ is the angle off axis and J_1 is the Bessel function of the first kind.

A planar source will radiate sound uniformly for low frequencies whose wavelength is longer than the dimensions of the planar source, and as frequency increases, the sound from such a source will be focused into an increasingly narrower angle. The smaller the driver, the higher the frequency where this narrowing of directivity occurs. Even if the diaphragm is not perfectly circular, this effect occurs such that larger sources are more directive. Several loudspeaker designs have been built which have approximately this behavior. Most are electrostatic or planar magnetic designs.

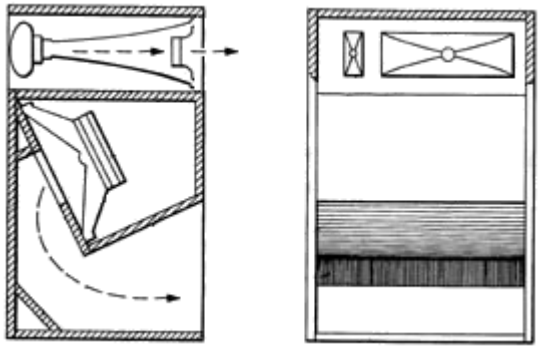
Various manufacturers use different driver mounting arrangements to create a specific type of sound field in the space for which they are designed. The resulting radiation patterns may be intended to more closely simulate the way sound is produced by real instruments, or simply create a controlled energy distribution from the input signal (some using this approach are called monitors, as they are useful in checking the signal just recorded in a studio). An example of the first is a room corner system with many small drivers on the surface of a 1/8 sphere. A system design of this type was patented by, and actually produced commercially, by Professor Amar Bose—the 2201. Later Bose models have deliberately emphasized production of both direct and reflected sound by the loudspeaker itself, regardless of its environment. The designs are controversial in high fidelity circles, but have proven commercially successful. Several other manufacturers' designs follow similar principles.

Directivity is an important issue because it affects the frequency balance of sound a listener hears, and also the interaction of the speaker system with the room and its contents. A speaker which is very directive (i.e., on an axis perpendicular to the speaker face) may result in a reverberant field lacking in high frequencies, giving the impression the speaker is deficient in treble even though it measures well on axis (e.g., "flat" across the entire frequency range). Speakers with very wide, or rapidly increasing directivity at high frequencies, can give the impression that there is too much treble (if the listener is on axis) or too little (if the listener is off axis). This is part of the reason why on-axis frequency response measurement is not a complete characterization of the sound of a given loudspeaker.

Other driver designs

Other types of drivers which depart from the most commonly used direct radiating electro-dynamic driver mounted in an enclosure include:

Horn loudspeakers



A three-way loudspeaker that uses horns in front of each of the three drivers: a shallow horn for the tweeter, a long, straight horn for mid frequencies and a folded horn for the woofer

Horn loudspeakers are the oldest form of loudspeaker system. The use of horns as voice-amplifying megaphones dates at least to the 17th century, and horns were used in mechanical gramophones as early as 1857. Horn loudspeakers use a shaped waveguide in front of or behind the driver to increase the directivity of the loudspeaker and to transform a small diameter, high pressure condition at the driver cone surface to a large diameter, low pressure condition at the mouth of the horn. This increases the sensitivity of the loudspeaker and focuses the sound over a narrower area. The size of the throat, mouth, the length of the horn, as well as the area expansion rate along it must be carefully chosen to match the drive to properly provide this transforming function over a range of frequencies (every horn performs poorly outside its acoustic limits, at both high and low frequencies). The length and cross-sectional mouth area required to create a bass or sub-bass horn require a horn many feet long. 'Folded' horns can reduce the total size, but compel designers to make compromises and accept increased complication such as cost and construction. Some horn designs not only fold the low frequency horn, but use the walls in a room corner as an extension of the horn mouth. In the late 1940s, horns whose mouths took up much of a room wall were not unknown amongst hi-fi fans. Room sized installations became much less acceptable when two or more were required.

A horn loaded speaker can have a sensitivity as high as 110 dB at 2.83 volts (1 watt at 8 ohms) at 1 meter. This is a hundredfold increase in output compared to a speaker rated at 90 dB sensitivity, and is invaluable in applications where high sound levels are required or amplifier power is limited.

Piezoelectric speakers

Piezoelectric speakers are frequently used as beepers in watches and other electronic devices, and are sometimes used as tweeters in less-expensive speaker systems, such as computer speakers and portable radios. Piezoelectric speakers have several advantages over conventional loudspeakers: they are resistant to overloads which would normally destroy most high frequency drivers, and they can be used without a crossover due to

their electrical properties. There are also disadvantages: some amplifiers can oscillate when driving capacitive loads like most piezoelectrics, which results in distortion or damage to the amplifier. Additionally, their frequency response, in most cases, is inferior to that of other technologies. This is why they are generally used in single frequency (beeper) or non-critical applications.

Piezoelectric speakers can have extended high frequency output, and this is useful in some specialized circumstances; for instance, sonar applications in which piezoelectric variants are used as both output devices (generating underwater sound) and as input devices (acting as the sensing components of underwater microphones). They have advantages in these applications, not the least of which is simple and solid state construction which resists the effects of seawater better than, say, a ribbon based device would.

Magnetostrictive speakers

Magnetostrictive transducers, based on magnetostriction, have been predominantly used as sonar ultrasonic sound wave radiators, but their usage has spread also to audio speaker systems. Magnetostrictive speaker drivers have some special advantages: they can provide greater force (with smaller excursions) than other technologies; low excursion can avoid distortions from large excursion as in other designs; the magnetizing coil is stationary and therefore more easily cooled; they are robust because delicate suspensions and voice coils are not required. Magnetostrictive speaker modules have been produced by Fostex and FeONIC and subwoofer drivers have also been produced.

Electrostatic loudspeakers

Electrostatic loudspeakers use a high voltage electric field (rather than a magnetic field) to drive a thin statically charged membrane. Because they are driven over the entire membrane surface rather than from a small voice coil, they ordinarily provide a more linear and lower distortion motion than dynamic drivers. They have the disadvantage that the diaphragm excursion is severely limited because of practical construction limitations; the further apart the stators are positioned, the higher the voltage must be to achieve acceptable efficiency, which increases the tendency for electrical arcs as well as the increasing the speaker's attraction of dust particles. For many years electrostatic loudspeakers had a reputation as a generally unreliable and occasionally dangerous product. Arcing remains a potential problem with current technologies, especially when the panels are allowed to collect dust or dirt, or when driven with high signal levels.

Electrostatics are inherently dipole radiators and due to the thin flexible membrane are less suited for use in enclosures to reduce low frequency cancellation as with common cone drivers. Due to this and the low excursion capability, full range electrostatic loudspeakers are large by nature, and the bass will roll off at a frequency corresponding to a quarter wavelength of the narrowest panel dimension. To reduce the size of commercial products, they are sometimes used as a high frequency driver in combination with a conventional dynamic driver which handles the bass frequencies.

Ribbon and planar magnetic loudspeakers

A **ribbon speaker** consists of a thin metal-film ribbon suspended in a magnetic field. The electrical signal is applied to the ribbon which moves with it, thus creating the sound. The advantage of a ribbon driver is that the ribbon has very little mass; thus, it can accelerate very quickly, yielding very good high-frequency response. Ribbon loudspeakers are often very fragile—some can be torn by a strong gust of air. Most ribbon tweeters emit sound in a dipole pattern; a very few have backings which limit the dipole radiation pattern. Above and below the ends of the more or less rectangular ribbon, there is less audible output due to phase cancellation, but the precise amount of directivity depends on ribbon length. Ribbon designs generally require exceptionally powerful magnets which make them costly to manufacture. Ribbons have a very low resistance that most amplifiers cannot drive directly. As a result, a step down transformer is typically used to increase the current through the ribbon. The amplifier "sees" a load that is the ribbon's resistance times the transformer turns ratio squared. The transformer must be carefully designed so that its frequency response and parasitic losses do not degrade the sound, further increasing cost and complication relative to conventional designs.

Planar magnetic speakers (having printed or embedded conductors on a flat diaphragm) are sometimes described as "ribbons", but are not truly ribbon speakers. The term planar is generally reserved for speakers which have roughly rectangular shaped flat surfaces that radiate in a bipolar (i.e., front and back) manner. Planar magnetic speakers consist of a flexible membrane with a voice coil printed or mounted on it. The current flowing through the coil interacts with the magnetic field of carefully placed magnets on either side of the diaphragm, causing the membrane to vibrate more or less uniformly and without much bending or wrinkling. The driving force covers a large percentage of the membrane surface and reduces resonance problems inherent in coil-driven flat diaphragms.

Bending wave loudspeakers

Bending wave transducers use a diaphragm that is intentionally flexible. The rigidity of the material increases from the center to the outside. Short wavelengths radiate primarily from the inner area, while longer waves reach the edge of the speaker. To prevent reflections from the outside back into the center, long waves are absorbed by a surrounding damper. Such transducers can cover a wide frequency range (80 Hz to 35,000 Hz) and have been promoted as being close to an ideal point sound source. This uncommon approach is being taken by only a very few manufacturers, in very different arrangements. The line of Ohm Walsh speakers use a unique driver designed by Lincoln Walsh. Lincoln Walsh was a brilliant engineer who was part of the engineering team that developed radar during World War II. He later designed audio amplifiers, and his final project was a unique, one-way speaker with one driver. It was a large cone that faced down into a sealed, airtight enclosure. Rather than move back-and-forth as conventional speakers do, the cone rippled and created sound using a principle known as “transmission line”. The new speaker created a single, perfectly rendered sound wave of remarkable clarity. A new company, Ohm Acoustics, was formed to develop and market Walsh’s

new speaker design. Lincoln Walsh died before his speaker was released to the public. After developing the Ohm A prototype, in 1973 Ohm introduced the Ohm F speaker to critical acclaim.

Flat panel loudspeakers

There have been many attempts to reduce the size of speaker systems, or alternatively to make them less obvious. One such attempt was the development of voice coils mounted to flat panels to act as sound sources. These can then be made in a neutral color and hung on walls where they will be less noticeable than many speakers, or can be deliberately painted with patterns in which case they can function decoratively. There are two related problems with flat panel techniques: first, a flat panel is necessarily more flexible than a cone shape in the same material, and therefore will move as a single unit even less, and second, resonances in the panel are difficult to control, leading to considerable distortions. Some progress has been made using such lightweight, rigid, materials such as Styrofoam, and there have been several flat panel systems commercially produced in recent years.

Distributed mode loudspeakers

A newer implementation of the flat panel speaker system involves an intentionally flexible panel and an "exciter", mounted off-center and located so as to excite the panel to vibrate with minimal resonances. Speakers using such techniques can reproduce sound with a wide directivity pattern (paradoxically somewhat like a point source) and have been used in some computer speaker designs and bookshelf loudspeakers.

Heil air motion transducers

Oskar Heil invented the air motion transducer in the 1960s. In this approach, a pleated diaphragm is mounted in a magnetic field and forced to close and open under control of a music signal. Air is forced from between the pleats in accordance with the imposed signal, generating sound. The drivers are less fragile than ribbons and considerably more efficient (and able to produce higher absolute output levels) than ribbon, electrostatic, or planar magnetic tweeter designs.

ESS, a California manufacturer, licensed the design, employed Heil, and produced a range of speaker systems using his tweeters during the 1970s and 1980s. Radio Shack, a large US retail store chain, also sold speaker systems using such tweeters for a time. At present, there are two manufacturers of these drivers, both in Germany, one of which produces a range of high end professional speakers using tweeters and mid-range drivers based on the technology.

Plasma arc speakers

Plasma arc loudspeakers use electrical plasma as a radiating element. Since plasma has minimal mass, but is charged and therefore can be manipulated by an electric field, the

result is a very linear output at frequencies far higher than the audible range. Problems of maintenance and reliability for this approach tend to make it unsuitable for mass market use. In 1978 Alan E. Hill of the Air Force Weapons Laboratory in Albuquerque, NM, designed the Plasmatronics Hill Type I, a tweeter whose plasma was generated from helium gas. This avoided the ozone and nitrous oxide produced by RF decomposition of air in an earlier generation of plasma tweeters made by the pioneering DuKane Corporation, who produced the Ionovac (marketed as the Ionofane in the UK) during the 1950s. Currently, there remain a few manufacturers in Germany who use this design, and a do-it-yourself design has been published and has been available on the Internet.

A less expensive variation on this theme is the use of a flame for the driver, as flames contain ionized (electrically charged) gases.

Digital speakers

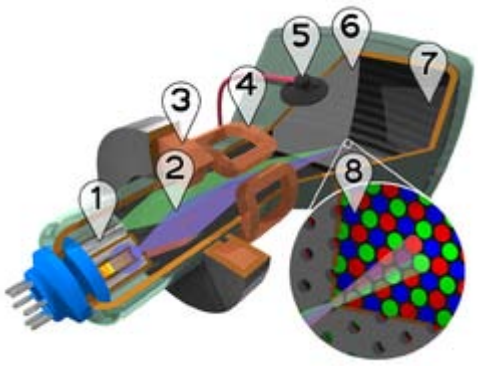
Digital speakers have been the subject of experiments performed by Bell Labs as far back as the 1920s. The design is simple; each bit controls a driver, which is either fully 'on' or 'off'.

There are problems with this design which have led to it being abandoned as impractical for the present. First, for a reasonable number of bits (required for adequate sound reproduction quality), the physical size of a speaker system becomes very large. Secondly, due to inherent analog digital conversion problems, the effect of aliasing is unavoidable, so that the audio output is "reflected" at equal amplitude in the frequency domain, on the other side of the sampling frequency, causing an unacceptably high level of ultrasonics to accompany the desired output. No workable scheme has been found to adequately deal with this.

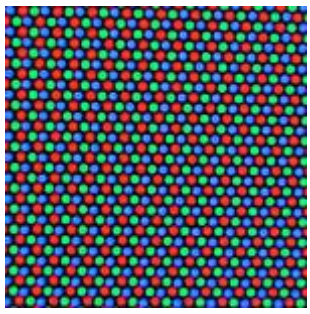
The term "digital" or "digital-ready" is often used for marketing purposes on speakers or headphones, but these systems are not digital in the sense described above. Rather, they are conventional speakers which can be used with digital sound sources, as can any conventional speaker, (e.g., optical media, MP3 players, etc.).

Chapter 5

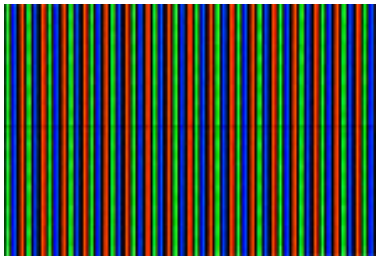
Cathode Ray Tube



Cutaway rendering of a color CRT:
1. Three Electron guns (for red, green, and blue phosphor dots)
2. Electron beams
3. Focusing coils
4. Deflection coils
5. Anode connection
6. Mask for separating beams for red, green, and blue part of displayed image
7. Phosphor layer with red, green, and blue zones
8. Close-up of the phosphor-coated inner side of the screen



Magnified view of a shadow mask color CRT



Magnified view of an aperture grille color CRT

The **Cathode Ray Tube** (**CRT**) is a vacuum tube containing an electron gun (a source of electrons) and a fluorescent screen, with internal or external means to accelerate and deflect the electron beam, used to create images in the form of light emitted from the fluorescent screen. The image may represent electrical waveforms (oscilloscope), pictures (television, computer monitor), radar targets and others.

The CRT uses an evacuated glass envelope which is large, deep, heavy, and relatively fragile.

History

A common CRT used in computer monitors and television sets

The experimentation of cathode rays is largely accredited to J.J. Thomson, an English physicist who, in his three famous experiments, was able to deflect cathode rays, a fundamental function of the modern CRT. The earliest version of the CRT was invented by the German physicist Ferdinand Braun in 1897 and is also known as the *Braun tube*. It was a cold-cathode diode, a modification of the Crookes tube with a phosphor-coated screen.

In 1907, Russian scientist Boris Rosing used a CRT in the receiving end of an experimental video signal to form a picture. He managed to display simple geometric shapes onto the screen, which marked the first time that CRT technology was used for what is now known as television.

The first cathode ray tube to use a hot cathode was developed by John B. Johnson (who gave his name to the term Johnson noise) and Harry Weiner Weinhart of Western Electric, and became a commercial product in 1922.

Overview

A cathode ray tube is a vacuum tube which consists of one or more electron guns, possibly internal electrostatic deflection plates, and a phosphor target. In television sets and computer monitors, the entire front area of the tube is scanned repetitively and systematically in a fixed pattern called a raster. An image is produced by controlling the intensity of each of the three electron beams, one for each additive primary color (red, green, and blue) with a video signal as a reference. In all modern CRT monitors and televisions, the beams are bent by *magnetic deflection*, a varying magnetic field generated by coils and driven by electronic circuits around the neck of the tube, although electrostatic deflection is commonly used in oscilloscopes, a type of diagnostic instrument.

Electron gun

Oscilloscope CRTs

In oscilloscope CRTs, electrostatic deflection is used, rather than the magnetic deflection commonly used with television and other large CRTs. The beam is deflected horizontally by applying an electric field between a pair of plates to its left and right, and vertically by applying an electric field to plates above and below. Oscilloscopes use electrostatic rather than magnetic deflection because the inductive reactance of the magnetic coils would limit the frequency response of the instrument.

Phosphor persistence

Various phosphors are available depending upon the needs of the measurement or display application. The brightness, color, and persistence of the illumination depends upon the type of phosphor used on the CRT screen. Phosphors are available with persistences ranging from less than one microsecond to several seconds. For visual observation of brief transient events, a long persistence phosphor may be desirable. For events which are fast and repetitive, or high frequency, a short-persistence phosphor is generally preferable.

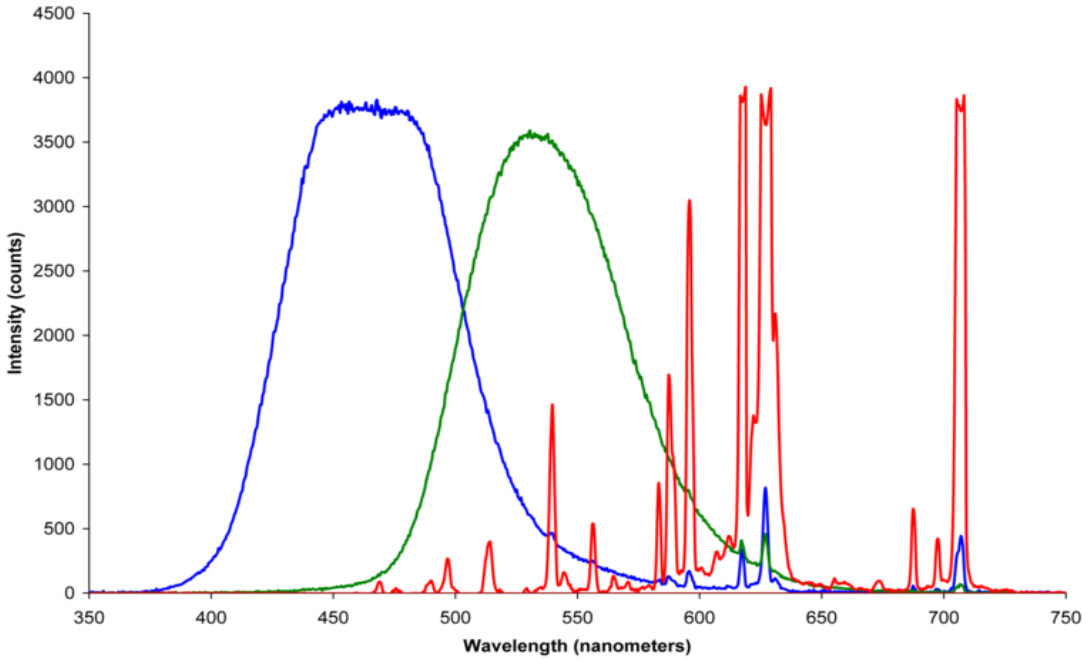
Microchannel plate

When displaying fast one-shot events the electron beam must deflect very quickly, with few electrons impinging on the screen; leading to a faint or invisible image on the display. Oscilloscope CRTs designed for very fast signals can give a brighter display by passing the electron beam through a micro-channel plate just before it reaches the screen. Through the phenomenon of secondary emission this plate multiplies the number of electrons reaching the phosphor screen, giving a significant improvement in writing rate (brightness), and improved sensitivity and spot size as well.

Graticules

Most oscilloscopes have a graticule as part of the visual display, to facilitate measurements. The graticule may be permanently marked inside the face of the CRT, or it may be a transparent external plate. External graticules are typically made of glass or acrylic plastic. An internal graticule provides an advantage in that it eliminates parallax error. Unlike an external graticule, an internal graticule can not be changed to accommodate different types of measurements. Oscilloscopes commonly provide a means for the graticule to be side-illuminated, which improves its visibility when used in a darkened room or when shaded by a camera hood.

Color CRTs



Spectra of constituent blue, green and red phosphors in a common CRT

Color tubes use three different phosphors which emit red, green, and blue light respectively. They are packed together in stripes (as in aperture grille designs) or clusters called "triads" (as in shadow mask CRTs). Color CRTs have three electron guns, one for each primary color, arranged either in a straight line or in a triangular configuration (the guns are usually constructed as a single unit). A grille or mask absorbs the electrons that would otherwise hit the wrong phosphor. A shadow mask tube uses a metal plate with tiny holes, placed so that the electron beam only illuminates the correct phosphors on the face of the tube. Another type of color CRT uses an aperture grille to achieve the same result.

Convergence in color CRTs

The three beams in color CRTs would not strike the screen at the same point without convergence calibration. Instead, the set would need to be manually adjusted to converge the three color beams together to maintain color accuracy.

Degaussing

Most CRT television sets and computer monitors have a built-in degaussing (demagnetizing) coil, which upon power-up creates a brief, alternating magnetic field which decays in strength over the course of a few seconds. This degaussing field is strong enough to remove most cases of shadow mask magnetization.

Vector monitors

Vector monitors were used in early computer aided design systems and in some late-1970s to mid-1980s arcade games such as *Asteroids*. They draw graphics point-to-point, rather than scanning a raster.

CRT resolution

Dot pitch defines the maximum resolution of the display, assuming delta-gun CRTs. In these, as the scanned resolution approaches the dot pitch resolution, moiré appears, as the detail being displayed is finer than what the shadow mask can render. Aperture grille monitors do not suffer from vertical moiré, however, because their phosphor stripes have no vertical detail. In smaller CRTs, these strips maintain position by themselves, but larger aperture grille CRTs require one or two crosswise (horizontal) support strips.

Gamma

CRTs have a pronounced triode characteristic, which results in significant gamma (a nonlinear relationship in an electron gun between applied video voltage and light intensity).

Other types of CRTs

Cat's eye

In better quality tube radio sets a tuning guide consisting of a phosphor tube was used to aid the tuning adjustment. This was also known as a "Magic Eye" or "Tuning Eye". Tuning would be adjusted until the width of a radial shadow was minimized. This was used instead of a more expensive electromechanical meter, which later came to be used on higher-end tuners when transistor sets lacked the high voltage required to drive the device. The same type of device was used with tape recorders as a recording level meter.

Charactrons

Some displays for early computers (those that needed to display more text than was practical using vectors, or that required high speed for photographic output) used Charactron CRTs. These incorporate a perforated metal character mask (stencil), which shapes a wide electron beam to form a character on the screen. The system selects a character on the mask using one set of deflection circuits, but that causes the extruded beam to be aimed off-axis, so a second set of deflection plates has to re-aim the beam so it is headed toward the center of the screen. A third set of plates places the character wherever required. The beam is unblanked (turned on) briefly to draw the character at that position. Graphics could be drawn by selecting the position on the mask corresponding to the code for a space (in practice, they were simply not drawn), which had a small round hole in the center; this effectively disabled the character mask, and the system reverted to regular vector behavior. Charactrons had exceptionally-long necks, because of the need for three deflection systems.

Nimo

Nimo tube BA0000-P31

Nimo was the trademark of a family of small specialised CRTs manufactured by Industrial Electronics Engineers. These had 10 electron guns which produced electron beams in the form of digits in a manner similar to that of the charactron. The tubes were either simple single-digit displays or more complex 4- or 6- digit displays produced by means of a suitable magnetic deflection system. Having little of the complexities of a

standard CRT, the tube required a relatively simple driving circuit, and as the image was projected on the glass face, it provided a much wider viewing angle than competitive types (e.g. nixie tubes).

Zeus Thin CRT Displays

In the late 1990s and early 2000s Philips Research Laboratories experimented with a type of thin CRT known as the *Zeus* display which contained CRT-like functionality in a flat panel. The devices were demonstrated but never marketed.

The future of CRT technology

Demise

Although a mainstay of display technology for decades, the demand for CRT screens has dropped precipitously since 2000, and this falloff has been accelerating in the latter half of that decade. The rapid advances and falling prices of LCD flat panel technology, first for computer monitors and then for televisions, has been the key factor in the demise of competing display technologies such as CRT, rear-projection, and plasma display.

The end of most high-end CRT production by around 2010 (including high-end Sony and Mitsubishi product lines) means an erosion of the CRT's capability. In Canada and the United States, the sale and production of high-end CRT TVs (30-inch screens) in these markets has all but ended by 2007; just a couple years later inexpensive combo CRT TVs (20-inch screens with an integrated VHS or DVD player) have disappeared from discount stores. It has been common to replace CRT-based televisions and monitors in as little as 5–6 years, although they generally are capable of satisfactory performance for a much longer time.

Companies are responding to this trend. Electronics retailers such as Best Buy have been steadily reducing store spaces for CRTs. In 2005, Sony announced that they would stop the production of CRT computer displays. Samsung did not introduce any CRT models for the 2008 model year at the 2008 Consumer Electronics Show and on February 4, 2008 Samsung removed their 30" wide screen CRTs from their North American website and has not replaced them with new models.

The demise of CRT, however, has been happening more slowly in the developing world. According to iSupply, production in units of CRTs was not surpassed by LCDs production until 4Q 2007, owing largely to CRT production at factories in China.

In the United Kingdom, DSG (Dixons), the largest retailer of domestic electronic equipment, reported that CRT models made up 80–90% of the volume of televisions sold at Christmas 2004 and 15–20% a year later, and that they were expected to be less than 5% at the end of 2006. Dixons ceased selling CRT televisions in 2007.

Causes

CRTs, despite recent advances, have remained relatively heavy and bulky and take up a lot of space in comparison to other display technologies. CRT screens have much deeper cabinets compared to flat panels and rear-projection displays for a given screen size, and so it becomes impractical to have CRTs larger than 40 inches (102 cm). The CRT disadvantages became especially significant in light of rapid technological advancements in LCD and plasma flat-panels which allow them to easily surpass 40 inches (102 cm) as well as being thin and wall-mountable, two key features that were increasingly being demanded by consumers.

By 2006, although the price points of CRTs are generally much lower than LCD and plasma flat panels, large screen CRTs (30-inches or more) are as expensive as a similar-sized LCD. While LCDs are generally the most expensive TV display technology, major innovations have caused prices to drop significantly.

Monochrome CRTs are even more efficient than color CRTs. This is because up to 2/3 of the backlight power of LCD and rear-projection displays are lost to the RGB stripe filter. Most LCDs also have poorer color rendition and can change color with viewing angle, though modern PVA and IPS LCDs have greatly attenuated these problems.

Resurgence in specialized markets

In the first quarter of 2008, CRTs retook the #2 technology position in North America from plasma, due to the decline and consolidation of plasma display manufacturers. DisplaySearch has reported that although in the 4Q of 2007 LCDs surpassed CRTs in worldwide sales, CRTs then outsold LCDs in the 1Q of 2008.

CRTs are useful for displaying photos with high pixels per unit area and correct color balance. LCDs, as currently the most common flatscreen technology, have generally inferior color rendition (despite having greater overall brightness) due to the fluorescent lights commonly used as a backlight.

CRTs are still popular in the printing and broadcasting industries as well as in the professional video, photography, and graphics fields due to their greater color fidelity, contrast, and better viewing from off-axis (wider viewing angle). CRTs also still find adherents in video gaming because of their higher resolution per initial cost, lowest possible input lag, fast response time, and multiple native resolutions.

Health concerns

Ionizing radiation

CRTs can emit a small amount of X-ray radiation as a result of the electron beam's bombardment of the shadow mask/aperture grille and phosphors. The amount of radiation escaping the front of the monitor is widely considered unharmful. The Food and Drug

Administration regulations in 21 C.F.R. 1020.10 are used to strictly limit, for instance, television receivers to 0.5 milliroentgens per hour (mR/h) (0.13 μC/(kg·h) or 36 pA/kg) at a distance of 5 cm (2 in) from any external surface; since 2007, most CRTs have emissions that fall well below this limit.

Toxicity

Color and monochrome CRTs may contain toxic substances, such as cadmium, in the phosphors. The rear glass tube of modern CRTs may be made from leaded glass, which represent an environmental hazard if disposed of improperly. By the time personal computers were produced, glass in the front panel (the viewable portion of the CRT) used barium rather than lead, though the rear of the CRT was still produced from leaded glass. Monochrome CRTs typically do not contain enough leaded glass to fail EPA tests.

In October 2001, the United States Environmental Protection Agency created rules stating that CRTs must be brought to special recycling facilities. In November 2002, the EPA began fining companies that disposed of CRTs through landfills or incineration. Regulatory agencies, local and statewide, monitor the disposal of CRTs and other computer equipment.

In Europe, disposal of CRT televisions and monitors is covered by the WEEE Directive.

Flicker

At low refresh rates (below 50 Hz), the periodic scanning of the display may produce an irritating flicker that some people perceive more easily than others, especially when viewed with peripheral vision. A high refresh rate (above 72 Hz) reduces the effect. Computer displays and televisions with CRTs driven by digital electronics often use refresh rates of 100 Hz or more to largely eliminate any perception of flicker. Non-computer CRTs or CRT for sonar or radar may have long persistence phosphor and are thus flicker free. If the persistence is too long on a video display, moving images will be blurred.

High-frequency noise

CRTs used for television operate with horizontal scanning frequencies of 15,734 Hz (for NTSC systems) or 15,625 Hz (for PAL systems). These frequencies are at the upper range of human hearing and are inaudible to many people; some people will perceive a high-pitched tone near an operating television CRT. The sound is due to magnetostriction in the magnetic core of the flyback transformer.

Implosion

A high vacuum exists within all cathode ray tubes, putting the envelope under relatively high stress. If the outer glass envelope is damaged, the glass will break and pieces will fly out at high speed. While modern Cathode Ray Tubes used in televisions and computer

displays have epoxy-bonded face-plates or other measures to prevent shattering of the envelope, CRTs removed from equipment must be handled carefully to avoid personal injury.

Security concerns

Under some circumstances, the signal radiated from the electron guns, scanning circuitry, and associated wiring of a CRT can be captured and used to remotely reconstruct what is shown on the CRT, using a process called Van Eck phreaking. Special TEMPEST shielding can mitigate this effect. Such radiation of a potentially exploitable signal however occurs also with LCDs and with all electronics in general.

Recycling

As electronic waste, CRTs are considered one of the hardest types to recycle. CRTs have relatively high concentration of lead and phosphorus, both of which are necessary for the display. There are several companies in the United States that charge a small fee to collect CRTs, then subsidize their labor by selling the harvested copper, wire, and printed circuit boards. Leaded CRT glass is sold to get remelted into other CRTs, or even broken down and used in road construction.