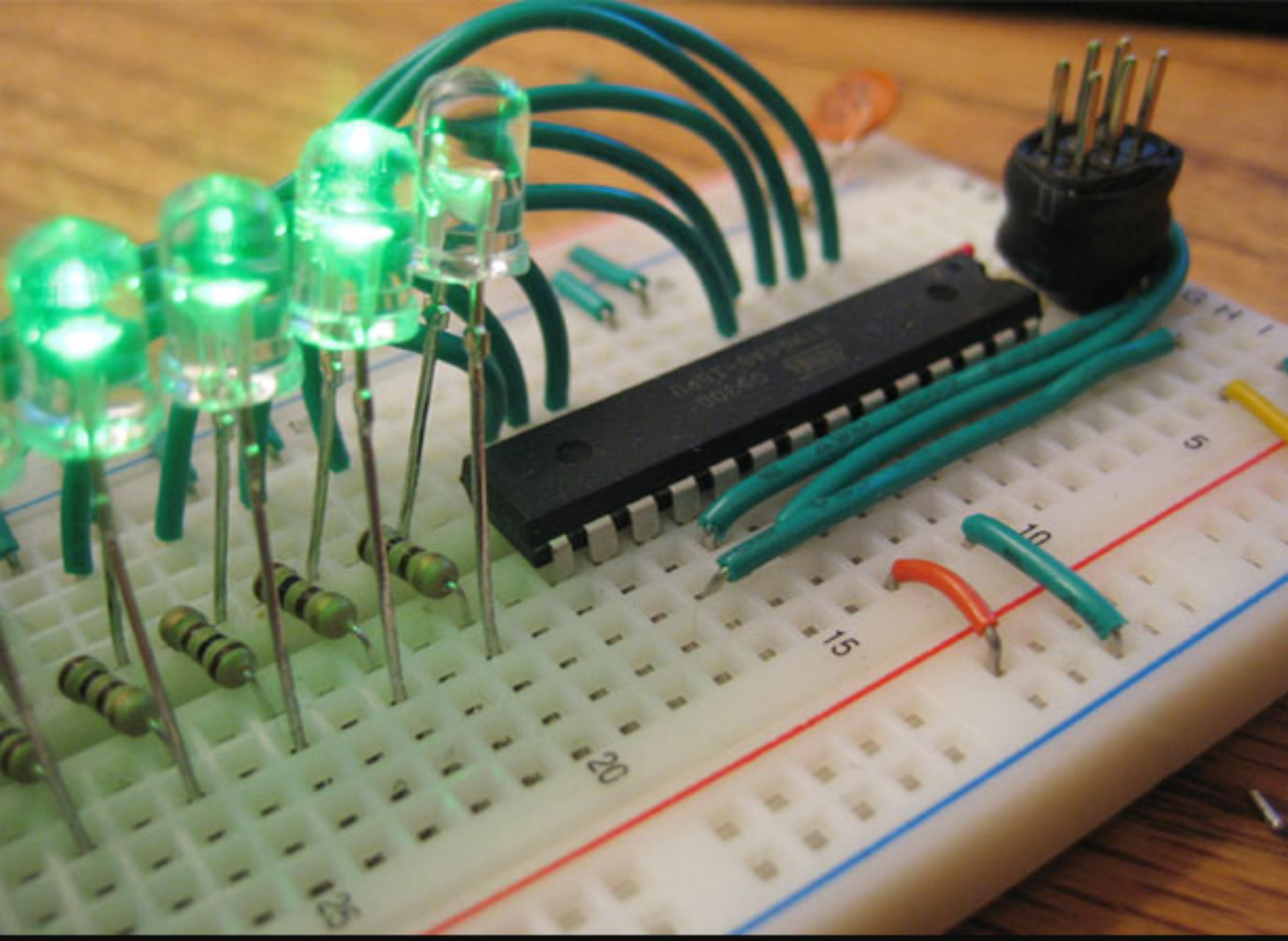


Analogue Circuits in Electronic Engineering

Marlena Cromwell

First Edition, 2012

ISBN 978-81-323-2829-2

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

Table of Contents

Chapter 1 - Passive Analogue Filter Development

Chapter 2 - Audio Crossover

Chapter 3 - Composite Image Filter

Chapter 4 - Current Source

Chapter 5 - Current-to-voltage Converter

Chapter 6 - Equivalent Impedance Transforms

Chapter 7 - Integrating ADC

Chapter 8 - LED Circuit

Chapter 9 - Mechanical Filter

Chapter- 1

Passive Analogue Filter Development

Analogue filters are a basic building block of signal processing much used in electronics. Amongst their many applications are the separation of an audio signal before application to bass, mid-range and tweeter loudspeakers; the combining and later separation of multiple telephone conversations onto a single channel; the selection of a chosen radio station in a radio receiver and rejection of others.

Passive linear electronic analogue filters are those filters which can be described with linear differential equations (linear); they are composed of capacitors, inductors and, sometimes, resistors (passive) and are designed to operate on continuously varying (analogue) signals. There are many linear filters which are not analogue in implementation (digital filter), and there are many electronic filters which may not have a passive topology – both of which may have the same transfer function of the filters described here. Analogue filters are most often used in wave filtering applications, that is, where it is required to pass particular frequency components and to reject others from analogue (continuous-time) signals.

Analogue filters have played an important part in the development of electronics. Especially in the field of telecommunications, filters have been of crucial importance in a number of technological breakthroughs and have been the source of enormous profits for telecommunications companies. It should come as no surprise, therefore, that the early development of filters was intimately connected with transmission lines. Transmission line theory gave rise to filter theory, which initially took a very similar form, and the main application of filters was for use on telecommunication transmission lines. However, the arrival of network synthesis techniques greatly enhanced the degree of control of the designer.

Today, it is often preferred to carry out filtering in the digital domain where complex algorithms are much easier to implement, but analogue filters do still find applications, especially for low-order simple filtering tasks and are often still the norm at higher frequencies where digital technology is still impractical, or at least, less cost effective. Wherever possible, and especially at low frequencies, analogue filters are now implemented in a filter topology which is active in order to avoid the wound components required by passive topology.

It is possible to design linear analogue mechanical filters using mechanical components which filter mechanical vibrations or acoustic waves. While there are few applications for such devices in mechanics per se, they can be used in electronics with the addition of transducers to convert to and from the electrical domain. Indeed some of the earliest ideas for filters were acoustic resonators because the electronics technology was poorly understood at the time. In principle, the design of such filters can be achieved entirely in terms of the electronic counterparts of mechanical quantities, with kinetic energy, potential energy and heat energy corresponding to the energy in inductors, capacitors and resistors respectively.

Historical overview

There are three main stages in the history of **passive analogue filter development**:

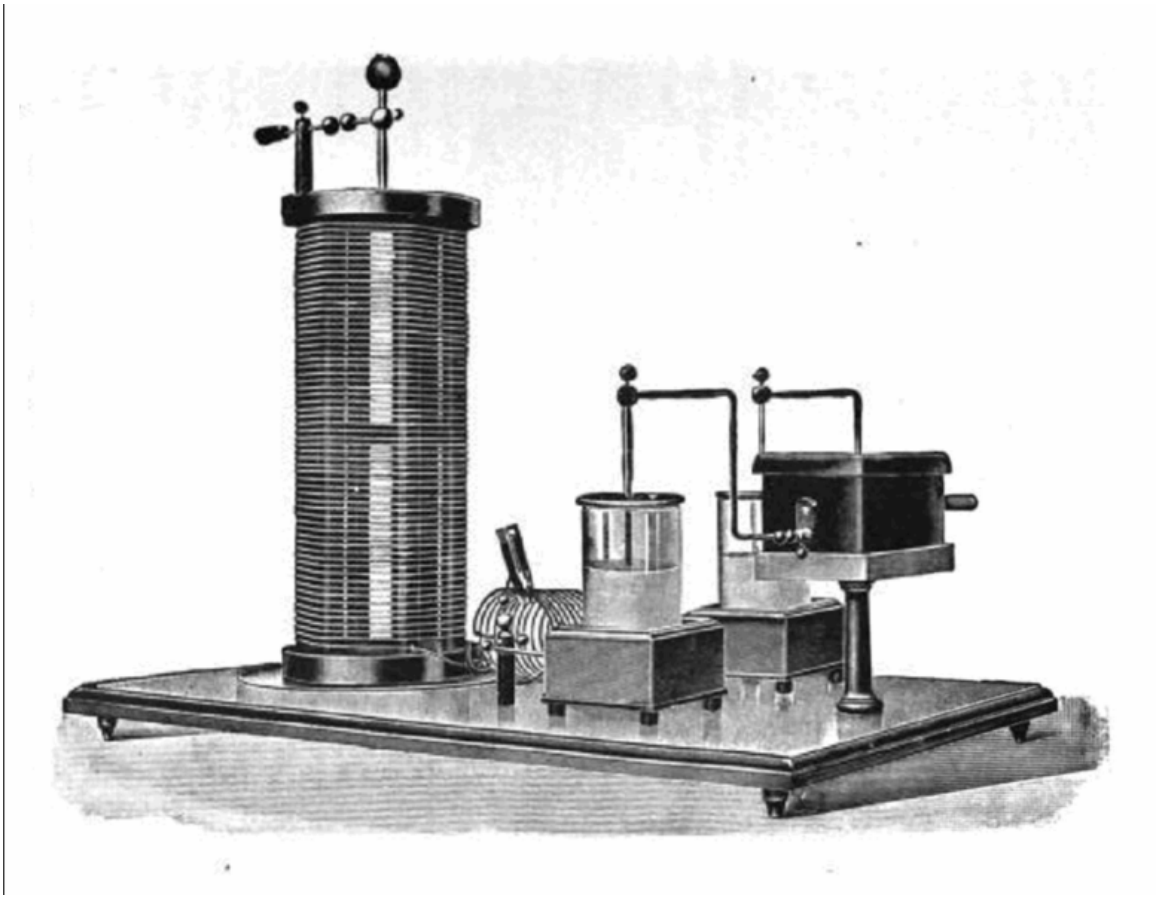
1. **Simple filters**. The frequency dependence of electrical response was known for capacitors and inductors from very early on. The resonance phenomenon was also familiar from an early date and it was possible to produce simple, single-branch filters with these components. Although attempts were made in the 1880s to apply them to telegraphy, these designs proved inadequate for successful frequency division multiplexing. Network analysis was not yet powerful enough to provide the theory for more complex filters and progress was further hampered by a general failure to understand the frequency domain nature of signals.
2. **Image filters**. Image filter theory grew out of transmission line theory and the design proceeded in a similar manner to transmission line analysis. For the first time filters could be produced that had precisely controllable passbands and other parameters. These developments took place in the 1920s and filters produced to these designs were still in widespread use in the 1980s, only declining as the use of analogue telecommunications has declined. Their immediate application was the economically important development of frequency division multiplexing for use on intercity and international telephony lines.
3. **Network synthesis filters**. The mathematical bases of network synthesis were laid in the 1930s and 1940s. After the end of World War II network synthesis became the primary tool of filter design. Network synthesis put filter design on a firm mathematical foundation, freeing it from the mathematically sloppy techniques of image design and severing the connection with physical lines. The essence of network synthesis is that it produces a design that will (at least if implemented with ideal components) accurately reproduce the response originally specified in black box terms.

Here throughout, the letters R,L and C are used with their usual meanings to represent resistance, inductance and capacitance, respectively. In particular they are used in combinations, such as LC, to mean, for instance, a network consisting only of inductors and capacitors. Z is used for electrical impedance, any 2-terminal combination of RLC elements and in some sections D is used for the rarely seen quantity elastance, which is the inverse of capacitance.

Resonance

Early filters utilised the phenomenon of resonance to filter signals. Although electrical resonance had been investigated by researchers from a very early stage, it was at first not widely understood by electrical engineers. Consequently, the much more familiar concept of acoustic resonance (which in turn, can be explained in terms of the even more familiar mechanical resonance) found its way into filter design ahead of electrical resonance. Resonance can be used to achieve a filtering effect because the resonant device will respond to frequencies at, or near, to the resonant frequency but will not respond to frequencies far from resonance. Hence frequencies far from resonance are filtered out from the output of the device.

Electrical resonance



A 1915 example of an early type of resonant circuit known as an Oudin coil which uses Leyden jars for the capacitance.

Resonance was noticed early on in experiments with the Leyden jar, invented in 1746. The Leyden jar stores electricity due to its capacitance, and is, in fact, an early form of capacitor. When a Leyden jar is discharged by allowing a spark to jump between the electrodes, the discharge is oscillatory. This was not suspected until 1826, when Felix Savary in France, and later (1842) Joseph Henry in the US noted that a steel needle

placed close to the discharge does not always magnetise in the same direction. They both independently drew the conclusion that there was a transient oscillation dying with time.

Hermann von Helmholtz in 1847 published his important work on conservation of energy in part of which he used those principles to explain why the oscillation dies away, that it is the resistance of the circuit which dissipates the energy of the oscillation on each successive cycle. Helmholtz also noted that there was evidence of oscillation from the electrolysis experiments of William Hyde Wollaston. Wollaston was attempting to decompose water by electric shock but found that both hydrogen and oxygen were present at both electrodes. In normal electrolysis they would separate, one to each electrode.

Helmholtz explained why the oscillation decayed but he had not explained why it occurred in the first place. This was left to Sir William Thomson (Lord Kelvin) who, in 1853, postulated that there was inductance present in the circuit as well as the capacitance of the jar and the resistance of the load. This established the physical basis for the phenomenon - the energy supplied by the jar was partly dissipated in the load but also partly stored in the magnetic field of the inductor.

So far, the investigation had been on the natural frequency of transient oscillation of a resonant circuit resulting from a sudden stimulus. More important from the point of view of filter theory is the behaviour of a resonant circuit when driven by an external AC signal: there is a sudden peak in the circuits response when the driving signal frequency is at the resonant frequency of the circuit. James Clerk Maxwell heard of the phenomenon from Sir William Grove in 1868 in connection with experiments on dynamos, and was also aware of the earlier work of Henry Wilde in 1866. Maxwell explained resonance mathematically, with a set of differential equations, in much the same terms that an RLC circuit is described today.

Heinrich Hertz (1887) experimentally demonstrated the resonance phenomena by building two resonant circuits, one of which was driven by a generator and the other was tunable and only coupled to the first electromagnetically (i.e., no circuit connection). Hertz showed that the response of the second circuit was at a maximum when it was in tune with the first. The diagrams produced by Hertz in this paper were the first published plots of an electrical resonant response.

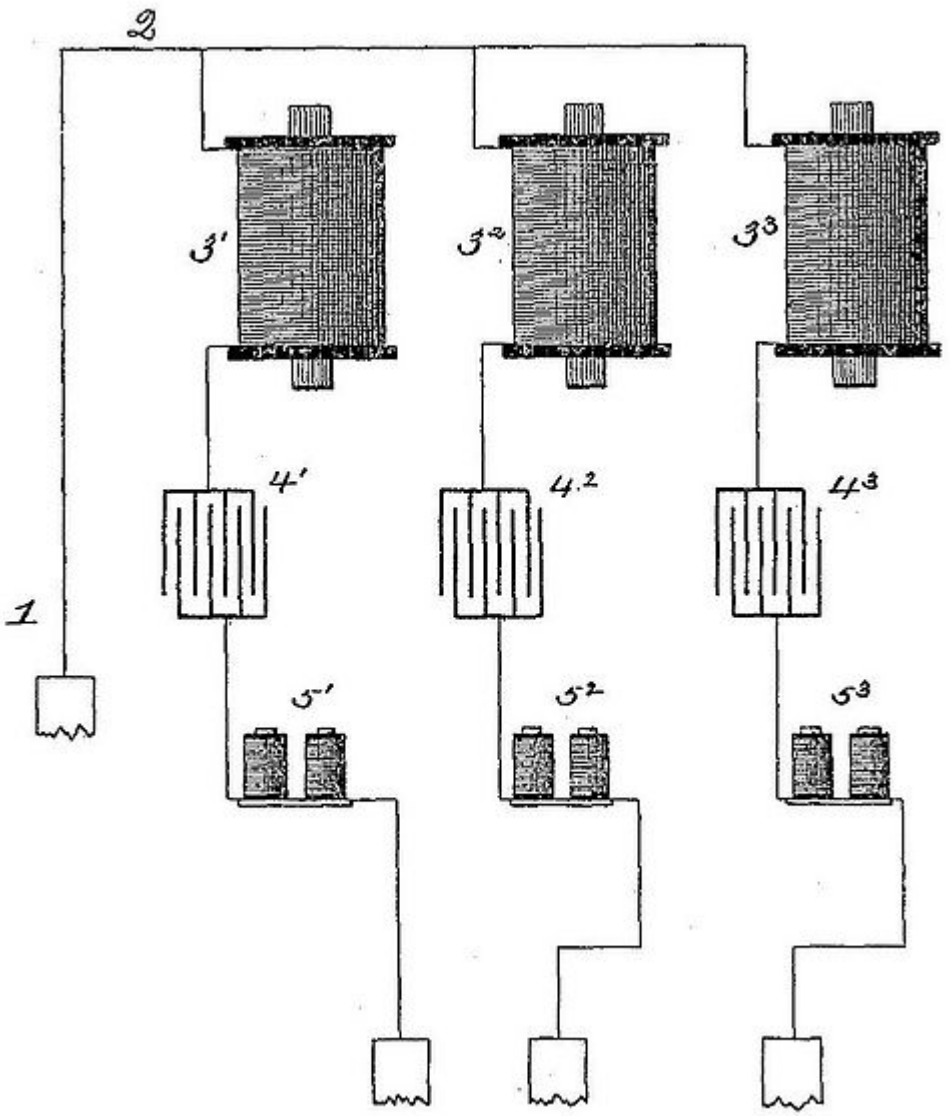
Acoustic resonance

As mentioned earlier, it was acoustic resonance that inspired filtering applications, the first of these being a telegraph system known as the "harmonic telegraph". Versions are due to Elisha Gray, Alexander Graham Bell (1870s), Ernest Mercadier and others. Its purpose was to simultaneously transmit a number of telegraph messages over the same line and represents an early form of frequency division multiplexing (FDM). FDM requires the sending end to be transmitting at different frequencies for each individual communication channel. This demands individual tuned resonators, as well as filters to separate out the signals at the receiving end. The harmonic telegraph achieved this with

electromagnetically driven tuned reeds at the transmitting end which would vibrate similar reeds at the receiving end. Only the reed with the same resonant frequency as the transmitter would vibrate to any appreciable extent at the receiving end.

Incidentally, the harmonic telegraph directly suggested to Bell the idea of the telephone. The reeds can be viewed as transducers converting sound to and from an electrical signal. It is no great leap from this view of the harmonic telegraph to the idea that speech can be converted to and from an electrical signal.

Early multiplexing



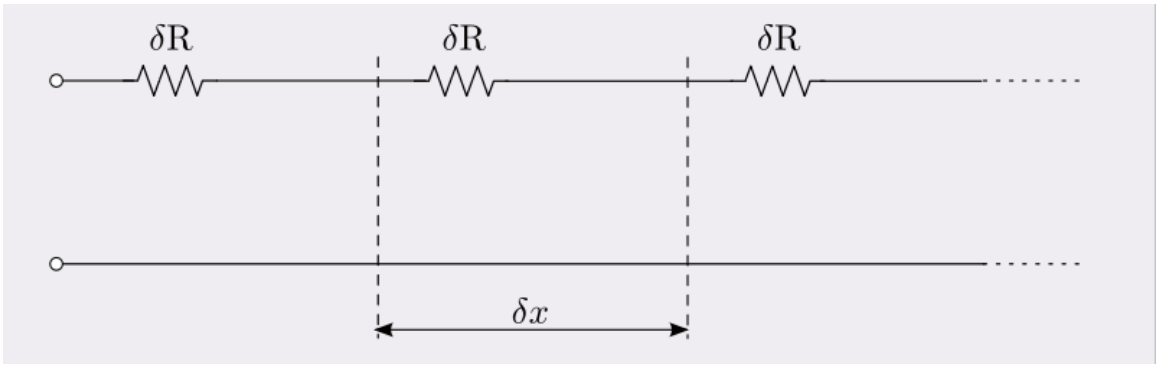
Hutin and Leblanc's multiple telegraph filter of 1891 showing the use of resonant circuits in filtering.

By the 1890s electrical resonance was much more widely understood and had become a normal part of the engineer's toolkit. In 1891 Hutin and Leblanc patented an FDM scheme for telephone circuits using resonant circuit filters. Rival patents were filed in 1892 by Michael Pupin and John Stone Stone with similar ideas, priority eventually being awarded to Pupin. However, no scheme using just simple resonant circuit filters can successfully multiplex (i.e. combine) the wider bandwidth of telephone channels (as opposed to telegraph) without either an unacceptable restriction of speech bandwidth or a channel spacing so wide as to make the benefits of multiplexing uneconomic.

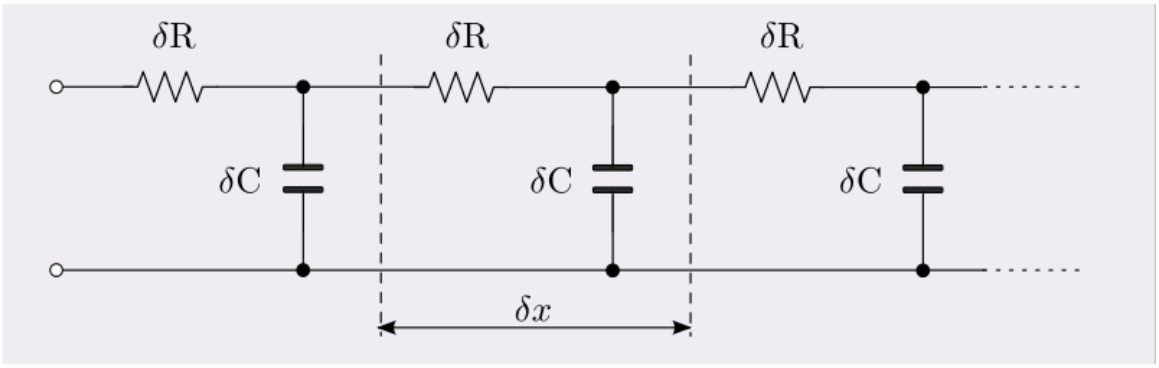
The basic technical reason for this difficulty is that the frequency response of a simple filter approaches a fall of 6 dB/octave far from the point of resonance. This means that if telephone channels are squeezed in side-by-side into the frequency spectrum, there will be crosstalk from adjacent channels in any given channel. What is required is a much more sophisticated filter that has a flat frequency response in the required passband like a low-Q resonant circuit, but that rapidly falls in response (much faster than 6 dB/octave) at the transition from passband to stopband like a high-Q resonant circuit. Obviously, these are contradictory requirements to be met with a single resonant circuit. The solution to these needs was founded in the theory of transmission lines and consequently the necessary filters did not become available until this theory was fully developed. At this early stage the idea of signal bandwidth, and hence the need for filters to match to it, was not fully understood; indeed, it was as late as 1920 before the concept of bandwidth was fully established. For early radio, the concepts of Q-factor, selectivity and tuning sufficed. This was all to change with the developing theory of transmission lines on which image filters are based, as explained in the next section.

At the turn of the century as telephone lines became available, it became popular to add telegraph on to telephone lines with an earth return phantom circuit. An LC filter was required to prevent telegraph clicks being heard on the telephone line. From the 1920s onwards, telephone lines, or balanced lines dedicated to the purpose, were used for FDM telegraph at audio frequencies. The first of these systems in the UK was a Siemens and Halske installation between London and Manchester. GEC and AT&T also had FDM systems. Separate pairs were used for the send and receive signals. The Siemens and GEC systems had six channels of telegraph in each direction, the AT&T system had twelve. All of these systems used electronic oscillators to generate a different carrier for each telegraph signal and required a bank of band-pass filters to separate out the multiplexed signal at the receiving end.

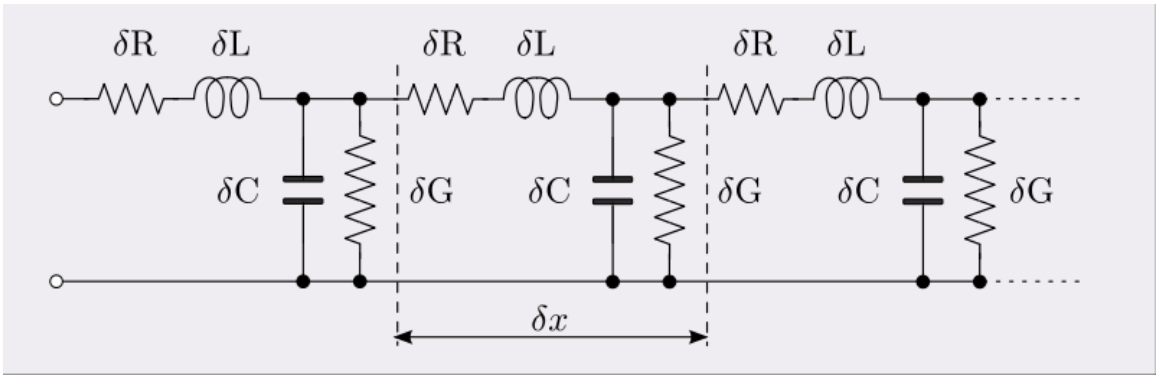
Transmission line theory



Ohm's model of the transmission line was simply resistance.



Lord Kelvin's model of the transmission line accounted for capacitance and the dispersion it caused. The diagram represents Kelvin's model translated into modern terms using infinitesimal elements, but this was not the actual approach used by Kelvin.



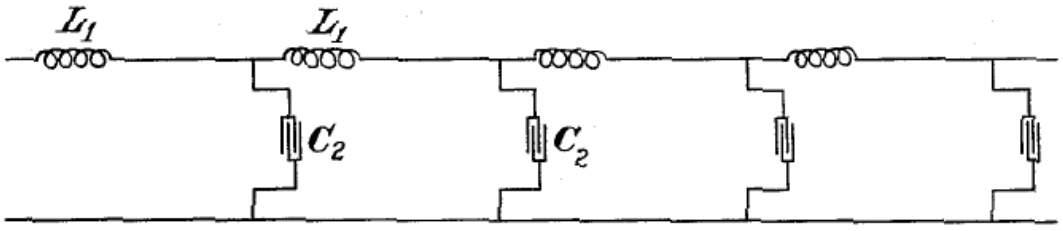
Heaviside's model of the transmission line. L, R, C and G in all three diagrams are the primary line constants. The infinitesimals δL, δR, δC and δG are to be understood as Lδx, Rδx, Cδx and Gδx respectively.

The earliest model of the transmission line was probably described by Georg Ohm (1827) who established that resistance in a wire is proportional to its length. The Ohm model thus included only resistance. Latimer Clark noted that signals were delayed and

elongated along a cable, an undesirable form of distortion now called dispersion but then called retardation, and Michael Faraday (1853) established that this was due to the capacitance present in the transmission line. Lord Kelvin (1854) found the correct mathematical description needed in his work on early transatlantic cables; he arrived at an equation identical to the conduction of a heat pulse along a metal bar. This model incorporates only resistance and capacitance, but that is all that was needed in undersea cables dominated by capacitance effects. Kelvin's model predicts a limit on the telegraph signalling speed of a cable but Kelvin still did not use the concept of bandwidth, the limit was entirely explained in terms of the dispersion of the telegraph symbols. The mathematical model of the transmission line reached its fullest development with Oliver Heaviside. Heaviside (1881) introduced series inductance and shunt conductance into the model making four distributed elements in all. This model is now known as the telegrapher's equation and the distributed elements are called the primary line constants.

From the work of Heaviside (1887) it had become clear that the performance of telegraph lines, and most especially telephone lines, could be improved by the addition of inductance to the line. George Campbell at AT&T implemented this idea (1899) by inserting loading coils at intervals along the line. Campbell found that as well as the desired improvements to the line's characteristics in the passband there was also a definite frequency beyond which signals could not be passed without great attenuation. This was a result of the loading coils and the line capacitance forming a low-pass filter, an effect that is only apparent on lines incorporating lumped components such as the loading coils. This naturally led Campbell (1910) to produce a filter with ladder topology, a glance at the circuit diagram of this filter is enough to see its relationship to a loaded transmission line. The cut-off phenomenon is an undesirable side-effect as far as loaded lines are concerned but for telephone FDM filters it is precisely what is required. For this application, Campbell produced band-pass filters to the same ladder topology by replacing the inductors and capacitors with resonators and anti-resonators respectively. Both the loaded line and FDM were of great benefit economically to AT&T and this led to fast development of filtering from this point onwards.

Image filters



Campbell's sketch of the low-pass version of his filter from his 1915 patent showing the now ubiquitous ladder topology with capacitors for the ladder rungs and inductors for the stiles. Filters of more modern design also often adopt the same ladder topology as used by Campbell. It should be understood that although superficially similar, they are really quite different. The ladder construction is essential to the Campbell filter and all the

sections have identical element values. Modern designs can be realised in any number of topologies, choosing the ladder topology is merely a matter of convenience. Their response is quite different (better) than Campbell's and the element values, in general, will all be different.

The filters designed by Campbell were named wave filters because of their property of passing some waves and strongly rejecting others. The method by which they were designed was called the image parameter method and filters designed to this method are called image filters. The image method essentially consists of developing the transmission constants of an infinite chain of identical filter sections and then terminating the desired finite number of filter sections in the image impedance. This exactly corresponds to the way the properties of a finite length of transmission line are derived from the theoretical properties of an infinite line, the image impedance corresponding to the characteristic impedance of the line.

From 1920 John Carson, also working for AT&T, began to develop a new way of looking at signals using the operational calculus of Heaviside which in essence is working in the frequency domain. This gave the AT&T engineers a new insight into the way their filters were working and led Otto Zobel to invent many improved forms. Carson and Zobel steadily demolished many of the old ideas. For instance the old telegraph engineers thought of the signal as being a single frequency and this idea persisted into the age of radio with some still believing that frequency modulation (FM) transmission could be achieved with a smaller bandwidth than the baseband signal right up until the publication of Carson's 1922 paper. Another advance concerned the nature of noise, Carson and Zobel (1923) treated noise as a random process with a continuous bandwidth, an idea that was well ahead of its time, and thus limited the amount of noise that it was possible to remove by filtering to that part of the noise spectrum which fell outside the passband. This too, was not generally accepted at first, notably being opposed by Edwin Armstrong (who ironically, actually succeeded in reducing noise with wide-band FM) and was only finally settled with the work of Harry Nyquist whose thermal noise power formula is well known today.

Several improvements were made to image filters and their theory of operation by Otto Zobel. Zobel coined the term constant k filter (or k-type filter) to distinguish Campbell's filter from later types, notably Zobel's m-derived filter (or m-type filter). The particular problems Zobel was trying to address with these new forms were impedance matching into the end terminations and improved steepness of roll-off. These were achieved at the cost of an increase in filter circuit complexity.

A more systematic method of producing image filters was introduced by Hendrik Bode (1930), and further developed by several other investigators including Piloty (1937-1939) and Wilhelm Cauer (1934-1937). Rather than enumerate the behaviour (transfer function, attenuation function, delay function and so on) of a specific circuit, instead a requirement for the image impedance itself was developed. The image impedance can be expressed in terms of the open-circuit and short-circuit impedances of the filter as $Z_i = \sqrt{Z_o Z_s}$. Since the image impedance must be real in the passbands and imaginary in the stopbands

according to image theory, there is a requirement that the poles and zeroes of Z_o and Z_s cancel in the passband and correspond in the stopband. The behaviour of the filter can be entirely defined in terms of the positions in the complex plane of these pairs of poles and zeroes. Any circuit which has the requisite poles and zeroes will also have the requisite response. Cauer pursued two related questions arising from this technique: what specification of poles and zeroes are realisable as passive filters; and what realisations are equivalent to each other. The results of this work led Cauer to develop a new approach, now called network synthesis.

This "poles and zeroes" view of filter design was particularly useful where a bank of filters, each operating at different frequencies, are all connected across the same transmission line. The earlier approach was unable to deal properly with this situation, but the poles and zeroes approach could embrace it by specifying a constant impedance for the combined filter. This problem was originally related to FDM telephony but frequently now arises in loudspeaker crossover filters.

Network synthesis filters

The essence of network synthesis is to start with a required filter response and produce a network that delivers that response, or approximates to it within a specified boundary. This is the inverse of network analysis which starts with a given network and by applying the various electric circuit theorems predicts the response of the network. The term was first used with this meaning in the doctoral thesis of Yuk-Wing Lee (1930) and apparently arose out of a conversation with Vannevar Bush. The advantage of network synthesis over previous methods is that it provides a solution which precisely meets the design specification. This is not the case with image filters, a degree of experience is required in their design since the image filter only meets the design specification in the unrealistic case of being terminated in its own image impedance, to produce which would require the exact circuit being sought. Network synthesis on the other hand, takes care of the termination impedances simply by incorporating them into the network being designed.

The development of network analysis needed to take place before network synthesis was possible. The theorems of Gustav Kirchhoff and others and the ideas of Charles Steinmetz (phasors) and Arthur Kennelly (complex impedance) laid the groundwork. The concept of a port also played a part in the development of the theory, and proved to be a more useful idea than network terminals. The first milestone on the way to network synthesis was an important paper by Ronald Foster (1924), *A Reactance Theorem*, in which Foster introduces the idea of a driving point impedance, that is, the impedance that is connected to the generator. The expression for this impedance determines the response of the filter and vice versa, and a realisation of the filter can be obtained by expansion of this expression. It is not possible to realise any arbitrary impedance expression as a network. Foster's reactance theorem stipulates necessary and sufficient conditions for realisability: that the reactance must be algebraically increasing with frequency and the poles and zeroes must alternate.

Wilhelm Cauer expanded on the work of Foster (1926) and was the first to talk of realisation of a one-port impedance with a prescribed frequency function. Foster's work considered only reactances (i.e., only LC-kind circuits). Cauer generalised this to any 2-element kind one-port network, finding there was an isomorphism between them. He also found ladder realisations of the network using Thomas Stieltjes' continued fraction expansion. This work was the basis on which network synthesis was built, although Cauer's work was not at first used much by engineers, partly because of the intervention of World War II, partly for reasons explained in the next section and partly because Cauer presented his results using topologies that required mutually coupled inductors and ideal transformers. Although on this last point, it has to be said that transformer coupled double tuned amplifiers are a common enough way of widening bandwidth without sacrificing selectivity.

Image method versus synthesis

Image filters continued to be used by designers long after the superior network synthesis techniques were available. Part of the reason for this may have been simply inertia, but it was largely due to the greater computation required for network synthesis filters, often needing a mathematical iterative process. Image filters, in their simplest form, consist of a chain of repeated, identical sections. The design can be improved simply by adding more sections and the computation required to produce the initial section is on the level of "back of an envelope" designing. In the case of network synthesis filters, on the other hand, the filter is designed as a whole, single entity and to add more sections (i.e., increase the order) the designer would have no option but to go back to the beginning and start over. The advantages of synthesised designs are real, but they are not overwhelming compared to what a skilled image designer could achieve, and in many cases it was more cost effective to dispense with time-consuming calculations. This is simply not an issue with the modern availability of computing power, but in the 1950s it was non-existent, in the 1960s and 1970s available only at cost, and not finally becoming widely available to all designers until the 1980s with the advent of the desktop personal computer. Image filters continued to be designed up to that point and many remained in service into the 21st century.

The computational difficulty of the network synthesis method was addressed by tabulating the component values of a prototype filter and then scaling the frequency and impedance and transforming the bandform to those actually required. This kind of approach, or similar, was already in use with image filters, for instance by Zobel, but the concept of a "reference filter" is due to Sidney Darlington. Darlington (1939), was also the first to tabulate values for network synthesis prototype filters, nevertheless it had to wait until the 1950s before the Cauer-Darlington elliptic filter first came into use.

Once computational power was readily available, it became possible to easily design filters to minimise any arbitrary parameter, for example time delay or tolerance to component variation. The difficulties of the image method were firmly put in the past, and even the need for prototypes became largely superfluous. Furthermore, the advent of

active filters eased the computation difficulty because sections could be isolated and iterative processes were not then generally necessary.

Realisability and equivalence

Realisability (that is, which functions are realisable as real impedance networks) and equivalence (which networks equivalently have the same function) are two important questions in network synthesis. Following an analogy with Lagrangian mechanics, Cauer formed the matrix equation,

$$[\mathbf{A}] = s^2[\mathbf{L}] + s[\mathbf{R}] + [\mathbf{D}] = s[\mathbf{Z}]$$

where [**Z**],[**R**],[**L**] and [**D**] are the *n*x*n* matrices of, respectively, impedance, resistance, inductance and elastance of an *n*-mesh network and *s* is the complex frequency operator $s=\sigma+i\omega$. Here [**R**],[**L**] and [**D**] have associated energies corresponding to the kinetic, potential and dissipative heat energies, respectively, in a mechanical system and the already known results from mechanics could be applied here. Cauer determined the driving point impedance by the method of Lagrange multipliers;

$$Z_{\mathrm{p}}(s) = \frac{\det[\mathbf{A}]}{s\, a_{11}}$$

where a_{11} is the complement of the element A_{11} to which the one-port is to be connected. From stability theory Cauer found that [**R**], [**L**] and [**D**] must all be positive-definite matrices for $Z_{\mathrm{p}}(s)$ to be realisable if ideal transformers are not excluded. Realisability is only otherwise restricted by practical limitations on topology. This work is also partly due to Otto Brune (1931), who worked with Cauer in the US prior to Cauer returning to Germany. A well known condition for realisability of a one-port rational impedance due to Cauer (1929) is that it must be a function of *s* that is analytic in the right halfplane ($\sigma>0$), have a positive real part in the right halfplane and take on real values on the real axis. This follows from the Poisson integral representation of these functions. Brune coined the term positive-real for this class of function and proved that it was a necessary and sufficient condition (Cauer had only proved it to be necessary) and they extended the work to LC multiports. A theorem due to Sidney Darlington states that any positive-real function $Z(s)$ can be realised as a lossless two-port terminated in a positive resistor R. No resistors within the network are necessary to realise the specified response.

As for equivalence, Cauer found that the group of real affine transformations,

$$[\mathbf{T}]^T[\mathbf{A}][\mathbf{T}]$$

where,

$$[\mathbf{T}] = \begin{bmatrix} 1 & 0 \cdots 0 \\ T_{21} & T_{22} \cdots T_{2n} \\ \cdot & \cdots \\ T_{n1} & T_{n2} \cdots T_{nn} \end{bmatrix}$$

is invariant in $Z_p(s)$, that is, all the transformed networks are equivalents of the original.

Approximation

The approximation problem in network synthesis is to find functions which will produce realisable networks approximating to a prescribed function of frequency within limits arbitrarily set. The approximation problem is an important issue since the ideal function of frequency required will commonly be unachievable with rational networks. For instance, the ideal prescribed function is often taken to be the unachievable lossless transmission in the passband, infinite attenuation in the stopband and a vertical transition between the two. However, the ideal function can be approximated with a rational function, becoming ever closer to the ideal the higher the order of the polynomial. The first to address this problem was Stephen Butterworth (1930) using his Butterworth polynomials. Independently, Cauer (1931) used Chebyshev polynomials, initially applied to image filters, and not to the now well-known ladder realisation of this filter.

Butterworth filter

Butterworth filters are an important class of filters due to Stephen Butterworth (1930) which are now recognised as being a special case of Cauer's elliptic filters. Butterworth discovered this filter independently of Cauer's work and implemented it in his version with each section isolated from the next with a valve amplifier which made calculation of component values easy since the filter sections could not interact with each other and each section represented one term in the Butterworth polynomials. This gives Butterworth the credit for being both the first to deviate from image parameter theory and the first to design active filters. It was later shown that Butterworth filters could be implemented in ladder topology without the need for amplifiers, possibly the first to do so was William Bennett (1932) in a patent which presents formulae for component values identical to the modern ones. Bennett, at this stage though, is still discussing the design as an artificial transmission line and so is adopting an image parameter approach despite having produced what would now be considered a network synthesis design. He also does not appear to be aware of the work of Butterworth or the connection between them.

Insertion-loss method

The insertion-loss method of designing filters is, in essence, to prescribe a desired function of frequency for the filter as an attenuation of the signal when the filter is inserted between the terminations relative to the level that would have been received were the terminations connected to each other via an ideal transformer perfectly matching them. Versions of this theory are due to Sidney Darlington, Wilhelm Cauer and others all

working more or less independently and is often taken as synonymous with network synthesis. Butterworth's filter implementation is, in those terms, an insertion-loss filter, but it is a relatively trivial one mathematically since the active amplifiers used by Butterworth ensured that each stage individually worked into a resistive load. Butterworth's filter becomes a non-trivial example when it is implemented entirely with passive components. An even earlier filter which influenced the insertion-loss method was Norton's dual-band filter where the input of two filters are connected in parallel and designed so that the combined input presents a constant resistance. Norton's design method, together with Cauer's canonical LC networks and Darlington's theorem that only LC components were required in the body of the filter resulted in the insertion-loss method. However, ladder topology proved to be more practical than Cauer's canonical forms.

Darlington's insertion-loss method is a generalisation of the procedure used by Norton. In Norton's filter it can be shown that each filter is equivalent to a separate filter unterminated at the common end. Darlington's method applies to the more straightforward and general case of a 2-port LC network terminated at both ends. The procedure consists of the following steps:

1. determine the poles of the prescribed insertion-loss function,
2. from that find the complex transmission function,
3. from that find the complex reflection coefficients at the terminating resistors,
4. find the driving point impedance from the short-circuit and open-circuit impedances,
5. expand the driving point impedance into an LC (usually ladder) network.

Darlington additionally used a transformation found by Hendrik Bode that predicted the response of a filter using non-ideal components but all with the same Q. Darlington used this transformation in reverse to produce filters with a prescribed insertion-loss with non-ideal components. Such filters have the ideal insertion-loss response plus a flat attenuation across all frequencies.

Elliptic filters

Elliptic filters are filters produced by the insertion-loss method which use elliptic rational functions in their transfer function as an approximation to the ideal filter response and the result is called a Chebyshev approximation. This is the same Chebyshev approximation technique used by Cauer on image filters but follows the Darlington insertion-loss design method and uses slightly different elliptic functions. Cauer had some contact with Darlington and Bell Labs before WWII (for a time he worked in the US) but during the war they worked independently, in some cases making the same discoveries. Cauer had disclosed the Chebyshev approximation to Bell Labs but had not left them with the proof. Sergei Schelkunoff provided this and a generalisation to all equal ripple problems. Elliptic filters are a general class of filter which incorporate several other important classes as special cases: Cauer filter (equal ripple in passband and stopband), Chebyshev

filter (ripple only in passband), reverse Chebyshev filter (ripple only in stopband) and Butterworth filter (no ripple in either band).

Generally, for insertion-loss filters where the transmission zeroes and infinite losses are all on the real axis of the complex frequency plane (which they usually are for minimum component count), the insertion-loss function can be written as;

$$\frac{1}{1 + JF^2}$$

where F is either an even (resulting in an antimetric filter) or an odd (resulting in an symmetric filter) function of frequency. Zeroes of F correspond to zero loss and the poles of F correspond to transmission zeroes. J sets the passband ripple height and the stopband loss and these two design requirements can be interchanged. The zeroes and poles of F and J can be set arbitrarily. The nature of F determines the class of the filter;

- if F is a Chebyshev approximation the result is a Chebyshev filter,
- if F is a maximally flat approximation the result is a passband maximally flat filter,
- if $1/F$ is a Chebyshev approximation the result is a reverse Chebyshev filter,
- if $1/F$ is a maximally flat approximation the result is a stopband maximally flat filter,

A Chebyshev response simultaneously in the passband and stopband is possible, such as Cauer's equal ripple elliptic filter.

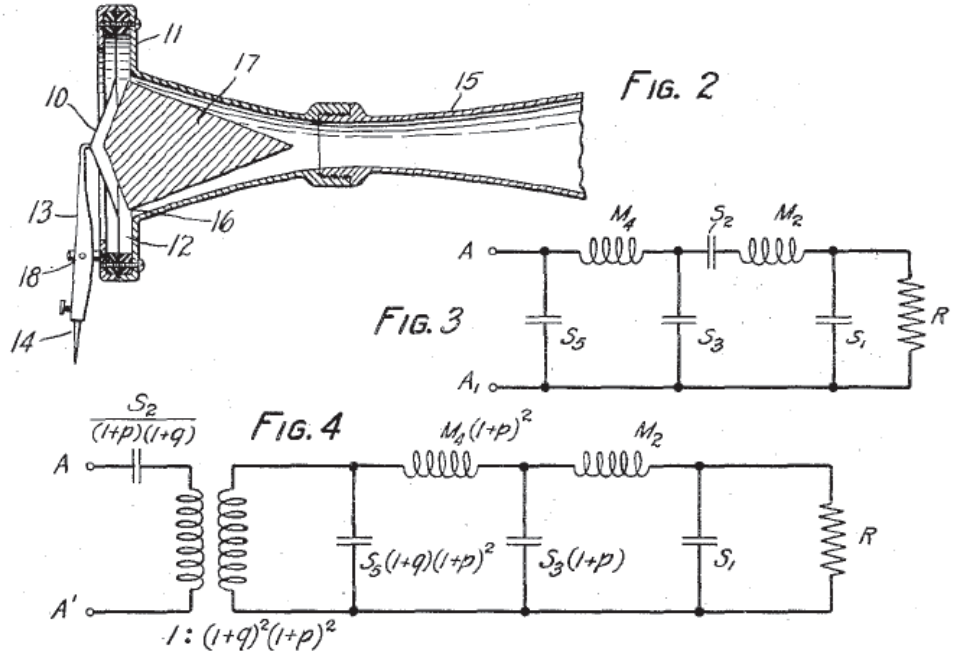
Darlington relates that he found in the New York City library Carl Jacobi's original paper on elliptic functions, published in Latin in 1829. In this paper Darlington was surprised to find foldout tables of the exact elliptic function transformations needed for Chebyshev approximations of both Cauer's image parameter, and Darlington's insertion-loss filters.

Other methods

Darlington considers the topology of coupled tuned circuits to involve a separate approximation technique to the insertion-loss method, but also producing nominally flat passbands and high attenuation stopbands. The most common topology for these is shunt anti-resonators coupled by series capacitors, less commonly, by inductors, or in the case of a two-section filter, by mutual inductance. These are most useful where the design requirement is not too stringent, that is, moderate bandwidth, roll-off and passband ripple.

Other notable developments and applications

Mechanical filters



Norton's mechanical filter together with its electrical equivalent circuit. Two equivalents are shown, "Fig.3" directly corresponds to the physical relationship of the mechanical components; "Fig.4" is an equivalent transformed circuit arrived at by repeated application of a well known transform, the purpose being to remove the series resonant circuit from the body of the filter leaving a simple *LC* ladder network.

Edward Norton, around 1930, designed a mechanical filter for use on phonograph recorders and players. Norton designed the filter in the electrical domain and then used the correspondence of mechanical quantities to electrical quantities to realise the filter using mechanical components. Mass corresponds to inductance, stiffness to elastance and damping to resistance. The filter was designed to have a maximally flat frequency response.

In modern designs it is common to use quartz crystal filters, especially for narrowband filtering applications. The signal exists as a mechanical acoustic wave while it is in the crystal and is converted by transducers between the electrical and mechanical domains at the terminals of the crystal.

Transversal filters

Transversal filters are not usually associated with passive implementations but the concept can be found in a Wiener and Lee patent from 1935 which describes a filter

consisting of a cascade of all-pass sections. The outputs of the various sections are summed in the proportions needed to result in the required frequency function. This works by the principle that certain frequencies will be in, or close to antiphase, at different sections and will tend to cancel when added. These are the frequencies rejected by the filter and can produce filters with very sharp cut-offs. This approach did not find any immediate applications, and is not common in passive filters. However, the principle finds many applications as an active delay line implementation for wide band discrete-time filter applications such as television, radar and high-speed data transmission.

Matched filter

The purpose of matched filters is to maximise the signal-to-noise ratio (S/N) at the expense of pulse shape. Pulse shape, unlike many other applications, is unimportant in radar while S/N is the primary limitation on performance. The filters were introduced during WWII (described 1943) by Dwight North and are often eponymously referred to as "North filters".

Filters for control systems

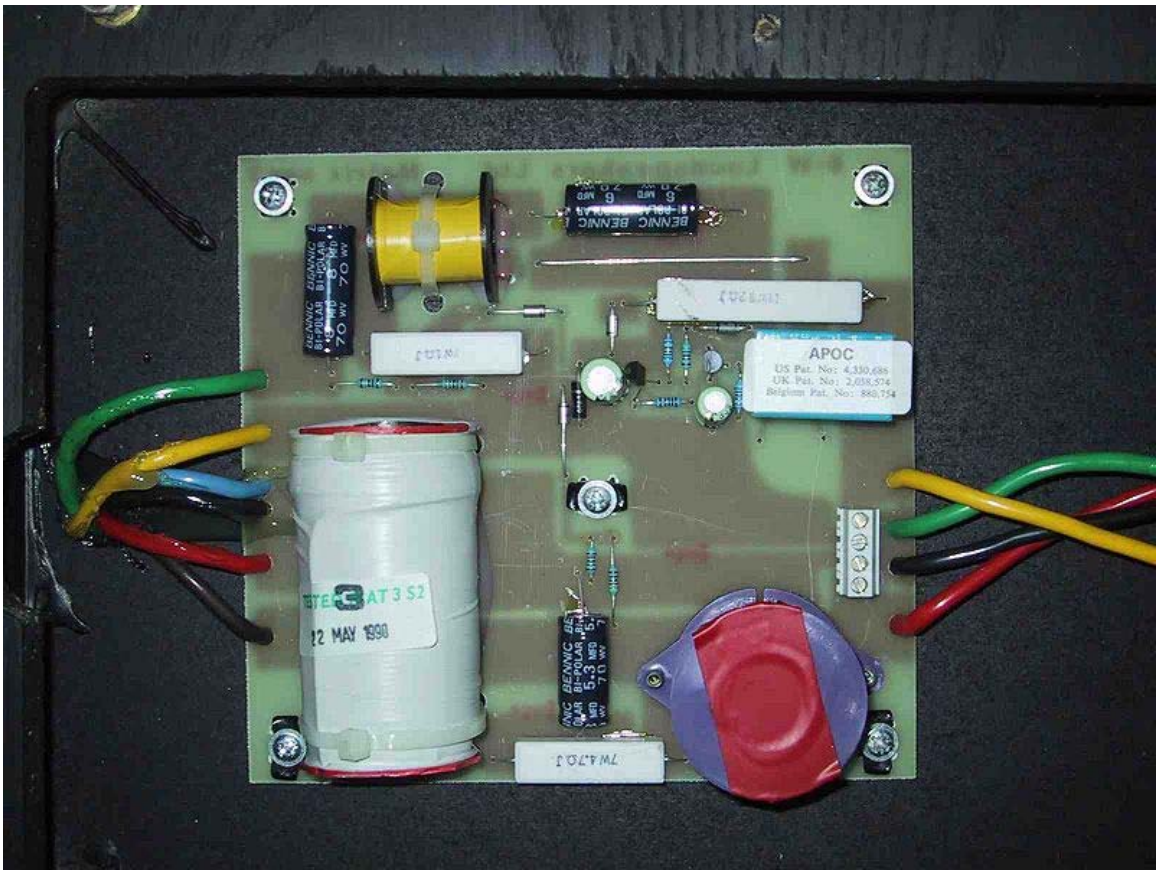
Control systems have a need for smoothing filters in their feedback loops with criteria to maximise the speed of movement of a mechanical system to the prescribed mark and at the same time minimise overshoot and noise induced motions. A key problem here is the extraction of Gaussian signals from a noisy background. An early paper on this was published during WWII by Norbert Wiener with the specific application to anti-aircraft fire control analogue computers. Rudy Kalman (Kalman filter) later reformulated this in terms of state-space smoothing and prediction where it is known as the linear-quadratic-Gaussian control problem. Kalman started an interest in state-space solutions, but according to Darlington this approach can also be found in the work of Heaviside and earlier.

Modern practice

LC passive filters gradually became less popular as active amplifying elements, particularly operational amplifiers, became cheaply available. The reason for the change is that wound components (the usual method of manufacture for inductors) are far from ideal, the wire adding resistance as well as inductance to the component. Inductors are also relatively expensive and are not "off-the-shelf" components. On the other hand, the function of LC ladder sections, LC resonators and RL sections can be replaced by RC components in an amplifier feedback loop (active filters). These components will usually be much more cost effective, and smaller as well. Cheap digital technology, in its turn, has largely supplanted analogue implementations of filters. However, there is still an occasional place for them in the simpler applications such as coupling where sophisticated functions of frequency are not needed.

Chapter- 2

Audio Crossover



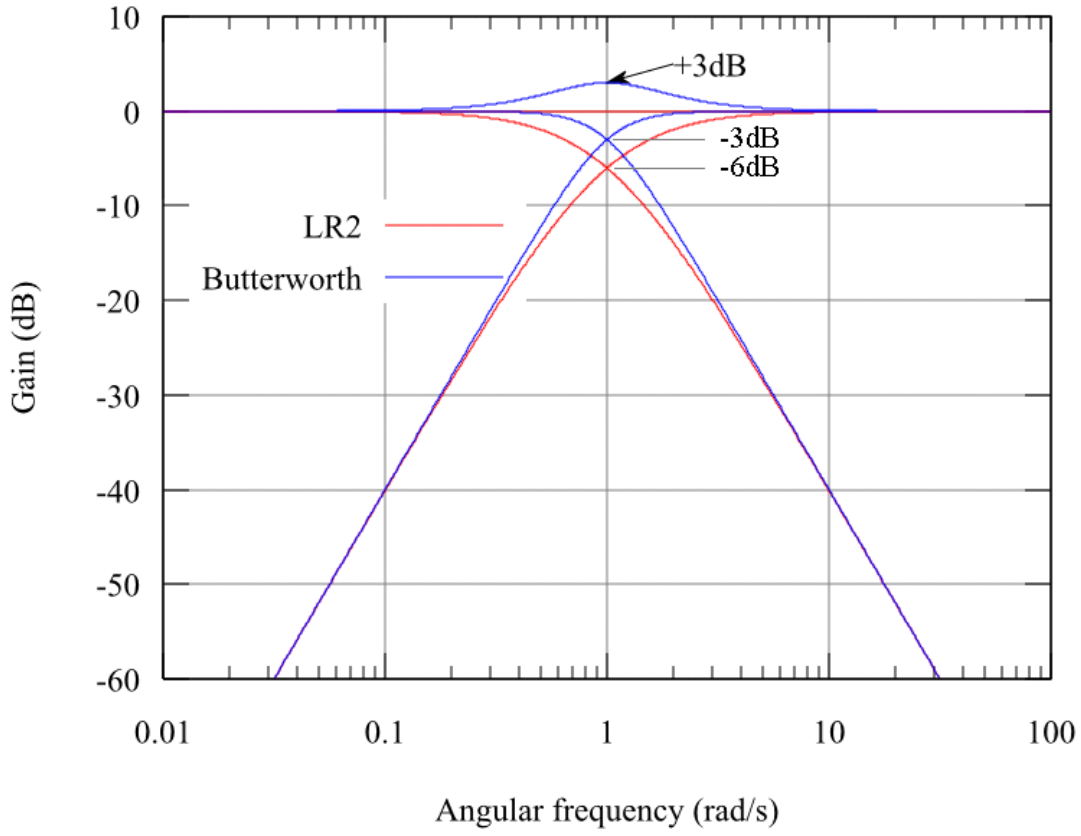
A passive 2-way crossover designed to operate at loudspeaker voltages

Audio crossovers are a class of electronic filter used in audio applications. Most individual loudspeaker drivers are incapable of covering the entire audio spectrum from low frequencies to high frequencies with acceptable relative volume and lack of distortion so most hi-fi speaker systems use a combination of multiple loudspeakers or drivers, each catering to a different frequency band. Crossovers split the audio signal into

separate frequency bands that can be separately routed to loudspeakers optimized for those bands.

Crossovers also enable multiband processing and multiple amplification where the audio signal is split into bands that are adjusted (equalized, compressed, echoed, etc.) separately before they are mixed together again. Some examples are: multiband dynamics (compression, limiting, de-essing), multiband distortion, bass enhancement, high frequency exciters, and noise reduction (for example: Dolby A noise reduction).

Overview



Comparison of the magnitude response of 2 pole Butterworth and Linkwitz-Riley crossover filters. The summed output of the Butterworth filters has a +3dB peak at the crossover frequency.

The definition of an ideal audio crossover changes relative to the task at hand. If the separate bands are to be mixed back together again (as in multiband processing), then the ideal audio crossover would split the incoming audio signal into separate bands that do not overlap or interact and which result in an output signal unchanged in frequency, relative levels, and phase response. This ideal performance can only be approximated. How to implement the best approximation is a matter of lively debate. On the other hand, if the audio crossover separates the audio bands in a loudspeaker, there is no requirement for mathematically ideal characteristics within the crossover itself, as the frequency and

phase response of the loudspeaker drivers within their mountings will eclipse the results. Satisfactory output of the complete system comprising the audio crossover *and* the loudspeaker drivers in their enclosure(s) is the design goal. Such a goal is often achieved using non-ideal, asymmetric crossover filter characteristics.

Many different crossover types are used in audio, but they generally belong to one of the following classes.

Classification

Classification based on the number of filter sections

In loudspeaker specifications, one often sees a speaker classified as an "N-way" speaker. N is a positive whole number greater than 1, and it indicates the number of filter sections. A 2-way crossover consists of a low-pass and a high-pass filter. A 3-way crossover is constructed as a combination of low-pass, band-pass and high-pass filters (LPF, BPF and HPF respectively). The BPF section is in turn a combination of HPF and LPF sections. 4 (or more) way crossovers are not very common in speaker design, primarily due to the complexity involved, which is not generally justified by better acoustic performance.

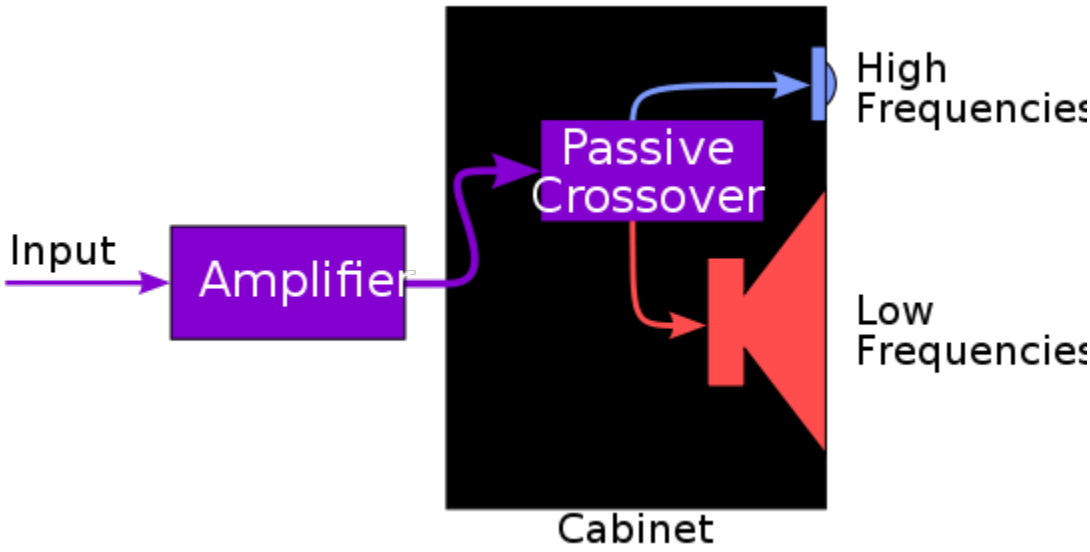
An extra HPF section may be present in an "N-way" loudspeaker crossover to protect the lowest-frequency driver from frequencies lower than it can safely handle. Such a crossover would then have a bandpass filter for the lowest-frequency driver. Similarly, the highest-frequency driver may have a protective LPF section to prevent high frequency damage, though this is far less common.

Recently, a number of manufacturers have begun using what is often called "N.5-way" crossover techniques for stereo loudspeaker crossovers. This usually indicates the addition of a second woofer that plays the same bass range as the main woofer but rolls off far before the main woofer does.

Classification based on components

Crossovers can also be classified based on the design approach; by the type of components used.

Passive



A passive crossover

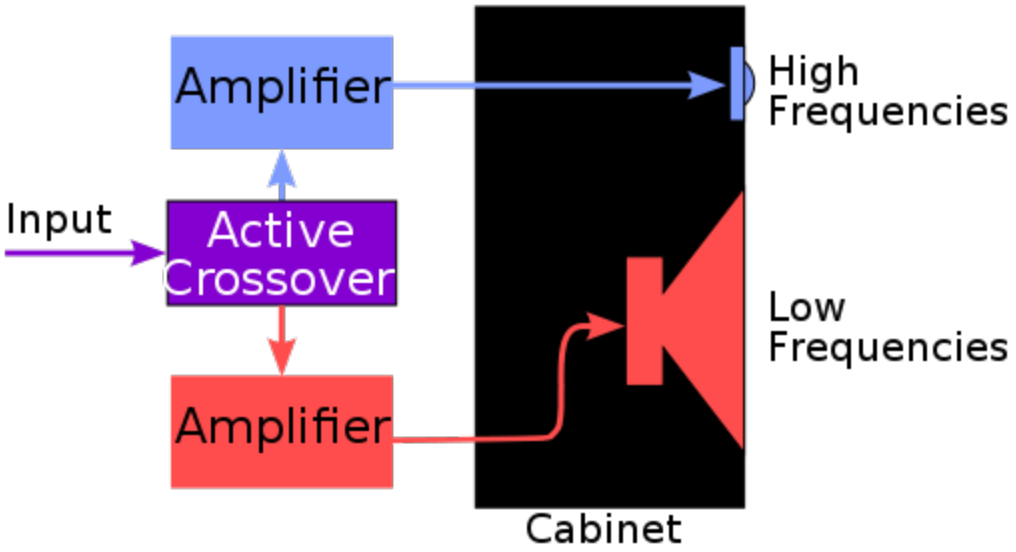
A passive crossover is made entirely of passive components, arranged most commonly in a Cauer topology to achieve a Butterworth filter. Passive filters use non-reactive resistors combined with reactive components such as capacitors and inductors. Very high performance passive crossovers are likely to be more expensive than active crossovers since individual components capable of good performance at the high currents and voltages at which speaker systems are driven are hard to make, and expensive. Polypropylene, metalized polyester foil, and paper-electrolytic capacitors are common. Inductors may have air cores, powdered metal cores, ferrite cores, or laminated silicon steel cores, and most are wound with enamelled copper wire. Some passive networks include devices such as fuses, PTC devices, bulbs or circuit breakers to protect the loudspeaker drivers from accidental overpowering. Modern passive crossovers increasingly incorporate equalization networks (e.g., Zobel networks) that compensate for the changes in impedance with frequency inherent in virtually all loudspeakers. The issue is complex, as part of the change in impedance is due to acoustic loading changes across a driver's passband.

On the negative side, passive networks may be bulky and cause power loss. They are not only frequency specific, but also impedance specific. This prevents interchangeability with speaker systems of different impedances. Ideal crossover filters, including impedance compensation and equalization networks, can be very difficult to design, as the components interact in complex ways. Crossover design expert Siegfried Linkwitz said of them that "the only excuse for passive crossovers is their low cost. Their behavior changes with the signal level dependent dynamics of the drivers. They block the power amplifier from taking maximum control over the voice coil motion. They are a waste of time, if accuracy of reproduction is the goal."

Alternatively, passive components can be utilised to construct filter circuits before the amplifier. This is called passive line-level crossover.

Active

An active crossover contains active components (i.e., those with gain) in its filters. In recent years, the most commonly used active device is an op-amp; active crossovers are operated at levels suited to power amplifier inputs in contrast to passive crossovers which operate after the power amplifier's output, at high current and in some cases high voltage. On the other hand, all circuits with gain introduce noise, and such noise has a more deleterious effect when introduced prior to the signal being amplified by the power amplifiers.



Typical usage of an active crossover, though a passive crossover can be positioned similarly before the amplifiers

Active crossovers always require the use of power amplifiers for each output band. Thus a 2-way active crossover needs two amplifiers—one each for the woofer and tweeter. This means that an active crossover based system will often cost more than a passive crossover based system, although none of the amplifiers needs to provide output as high as for an equivalent sound level full-frequency, power amplifier, which reduces cost. The cost and complication disadvantages of active crossovers are offset by the following gains:

- a frequency response independent of the dynamic changes in a driver's electrical characteristics.
- typically, the possibility of an easy way to vary or fine tune each frequency band to the specific drivers used. Examples would be crossover slope, filter type (e.g., Bessel, Butterworth, etc.), relative levels, ...

- isolation of each driver from signals handled by drivers, thus reducing intermodulation distortion and overdriving
- The power amplifiers are directly connected to the speaker drivers, thereby maximizing amplifier damping control of the speaker voice coil, reducing consequences of dynamic changes in driver electrical characteristics, all of which are likely to improve the transient response of the system
- reduction in power amplifier output requirement. With no energy being lost in passive components, amplifier requirements are reduced considerably (up to 1/2 in some cases), reducing costs, and potentially increasing quality.

Digital

Active crossovers can be implemented digitally using a DSP chip or other microprocessor. They either use digital approximations to traditional analog circuits, known as IIR filters (Bessel, Butterworth, Linkwitz-Riley etc.), or they use Finite impulse response (FIR) filters. IIR filters have many similarities with analog filters and are relatively undemanding of CPU resources; FIR filters on the other hand usually have a higher order and therefore require more resources for similar characteristics. They can be designed and built so that they have a linear phase response, which is thought desirable by many involved in sound reproduction. There are drawbacks though—in order to achieve linear phase response, a longer delay time is incurred than would be necessary with an IIR or minimum phase FIR filters. IIR filters, which are by nature recursive have the drawback that if not carefully designed they may enter limit cycles resulting in non-linear distortion.

Mechanical

This crossover type is mechanical and uses the properties of the materials in a driver diaphragm to achieve the necessary filtering. Such crossovers are commonly found in full-range speakers which are designed to cover as much of the audio band as possible. One such is constructed by coupling the cone of the speaker to the voice coil bobbin through a compliant section and directly attaching a small lightweight *whizzer* cone to the bobbin. This compliant section serves as a compliant filter, so the main cone is not vibrated at higher frequencies. The whizzer cone responds to all frequencies, but due to its smaller size only gives a useful output at higher frequencies, thereby implementing a mechanical crossover function. Careful selection of materials used for the cone, whizzer and suspension elements determines the crossover frequency and the effectiveness of the crossover. Such mechanical crossovers are complex to design, especially if high fidelity is desired. Computer aided design has largely replaced the laborious trial and error approach that was historically used. Over several years, the compliance of the materials may change, negatively affecting the frequency response of the speaker.

A more common approach is to employ the dust cap as a high frequency radiator. The dust cap radiates low frequencies, moving as part of the main assembly, but due to low-mass and reduced damping, radites increased energy at higher frequencies. As with whizzer cones, careful selection of material, shape and position are required to provide

smooth, extended output. High frequency dispersion is somewhat different for this approach than for whizzer cones. A related approach is to shape the main cone with such profile, and of such materials, that the neck area remains more rigid, radiating all frequencies, while the outer areas of the cone are selectively decoupled, radiating only at lower frequencies. Cone profiles and materials can be modeled in FEA software and the results predicted to excellent tolerances.

Speakers which use these mechanical crossovers have some advantages in sound quality despite the difficulties of designing and manufacturing them, and despite the inevitable output limitations. Full-range drivers have a single acoustic center, and can have relatively modest phase change across the audio spectrum. For best performance at low frequencies, these drivers require careful enclosure design. Their small size (typically 165 to 200 mm) requires considerable cone excursion to reproduce bass effectively, but the short voice coils required for reasonable high frequency performance can only move over a limited range. Nevertheless, within these constraints, cost and complications are reduced, as no crossovers are required.

Classification based on filter order or slope

Just as filters have different orders, so do crossovers, depending on the filter slope they implement. The final acoustic slope may be completely determined by the electrical filter or may be achieved by combining the electrical filter's slope with the natural characteristics of the driver. In the former case, the only requirement is that each driver has a flat response at least to the point where its signal is approximately −10dB down from the passband. In the latter case, the final acoustic slope is usually steeper than that of the electrical filters used. A third- or fourth-order acoustic crossover often has just a second order electrical filter. This requires that speaker drivers be well behaved a considerable way from the nominal crossover frequency, and further that the high frequency driver be able to survive a considerable input in a frequency range below its crossover point. This is difficult in actual practice. In the discussion below, the characteristics of the electrical filter order is discussed, followed by a discussion of crossovers having that acoustic slope and their advantages or disadvantages.

Most audio crossovers use first to fourth order electrical filters. Higher orders are not generally implemented in passive crossovers for loudspeakers, but are sometimes found in electronic equipment under circumstances for which their considerable cost and complexity can be justified.

First order

First-order filters have a 20 dB/decade (or 6 dB/octave) slope. All first-order filters have a Butterworth filter characteristic. First-order filters are considered by many audiophiles to be ideal for crossovers. This is because this filter type is 'transient perfect', meaning it passes both amplitude and phase unchanged across the range of interest. It also uses the fewest parts and has the lowest insertion loss (if passive). A first-order crossover allows more signals of unwanted frequencies to get through in the LPF and HPF sections than do

higher order configurations. While woofers can easily take this (aside from generating distortion at frequencies above those they can properly handle), smaller high frequency drivers (especially tweeters) are more likely to be damaged since they are not capable of handling large power inputs at frequencies below their crossovers.

In practice, speaker systems with true first order acoustic slopes are difficult to design because they require large overlapping driver bandwidth, and the shallow slopes mean that non-coincident drivers interfere over a wide frequency range and cause large response shifts off-axis.

Second order

Second-order filters have a 40 dB/decade (or 12 dB/octave) slope. Second-order filters can have a Bessel, Linkwitz-Riley or Butterworth characteristic depending on design choices and the components used. This order is commonly used in passive crossovers as it offers a reasonable balance between complexity, response, and higher frequency driver protection. When designed with time aligned physical placement, these crossovers have a symmetrical polar response, as do all even order crossovers.

It is commonly thought that there will always be a phase difference of 180° between the outputs of a (second order) low-pass filter and a high-pass filter having the same crossover frequency. And so, in a 2-way system, the high-pass section's output is usually connected to the high frequency driver 'inverted', to correct for this phase problem. For passive systems, the tweeter is wired with opposite polarity to the woofer; for active crossovers the high-pass filter's output is inverted. In 3-way systems the mid-range driver or filter is inverted. However, this is generally only true when the speakers have a wide response overlap and the acoustic centers are physically aligned.

Third order

Third-order filters have a 60 dB/decade (or 18 dB/octave) slope. These crossovers usually have Butterworth filter characteristics; phase response is very good, the level sum being flat and in phase quadrature, similar to a first order crossover. The polar response is asymmetric. In the original D'Appolito MTM arrangement, a symmetrical arrangement of drivers is used to create a symmetrical off-axis response when using third-order crossovers.

Third-order acoustic crossovers are often built from first- or second-order filter circuits.

Fourth order

Fourth-order crossover slopes shown on a Smaart transfer function

Fourth-order filters have an 80 dB/decade (or 24 dB/octave) slope. These filters are complex to design in passive form, as the components interact with each other. Steep-slope passive networks are less tolerant of parts value deviations or tolerances, and more sensitive to mis-termination with reactive driver loads. A 4th order crossover with −6 dB crossover point and flat summing is also known as a Linkwitz-Riley crossover (named after its inventors), and can be constructed in active form by cascading two 2nd order Butterworth filter sections. The output signals of this crossover order are in phase, thus avoiding partial phase inversion if the crossover bandpasses are electrically summed, as they would be within the output stage of a multiband compressor. Crossovers used in loudspeaker design do not require the filter sections to be in phase: smooth output characteristics are often achieved using non-ideal, asymmetric crossover filter characteristics. Bessel, Butterworth and Chebyshev are among the possible crossover topologies.

Such steep-slope filters have greater problems with overshoot and ringing but there are several key advantages, even in their passive form, such as the potential for a lower crossover point and increased power handling for tweeters, together with less overlap between drivers, dramatically reducing lobing, or other unwelcome off-axis effects. With less overlap between adjacent drivers, their location relative to each other becomes less critical and allows more latitude in speaker system cosmetics or (in car audio) practical installation constraints.

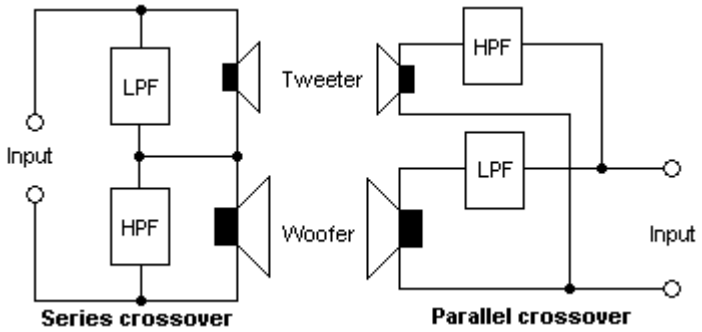
Higher order

Passive crossovers giving acoustic slopes higher than fourth-order are not common because of cost and complexity. Filters of up to 96 dB per octave are available in active crossovers and loudspeaker management systems.

Mixed order

Crossovers can also be constructed with mixed order filters. For example, a second order lowpass combined with a third order highpass. These are generally passive and are used for several reasons, often when the component values are found by computer program optimization. A higher order tweeter crossover can sometimes help compensate for the time offset between the woofer and tweeter, caused by non aligned acoustic centers.

Classification based on circuit topology



Series and parallel crossover topologies. The HPF and LPF sections for the series crossover are interchanged with respect to the parallel crossover since they appear in shunt with the low & high frequency drivers.

Parallel

Parallel crossovers are by far the most common. Electrically the filters are in parallel and thus the various filter sections do not interact. This makes two-way crossovers easier to design because the sections can be considered separately, and because component tolerance variations will be isolated. In the years before computer modeling, three-way crossovers were designed using the same value, but the advent of iterative design software has taught that this old technique creates excess gain and a 'haystack' response in the midrange output, together with a lower than anticipated input impedance.

Series

In this topology, the individual filters are connected in series, and a driver or driver combination is connected in parallel with each filter. To understand the signal path in this type of crossover, refer to the "Series Crossover" figure, and consider a high frequency signal that, during a certain moment, has a positive voltage on the upper Input terminal

compared to the lower Input terminal. The low pass filter (LPF) presents a high impedance to the signal, and the tweeter presents a low impedance; so the signal passes through the tweeter. The signal continues to the connection point between the woofer and the high pass filter (HPF). There, the HPF presents a low impedance to the signal, so the signal passes through the HPF, and appears at the lower Input terminal. A low frequency signal with a similar instantaneous voltage characteristic first passes through the LPF, then the woofer, and appears at the lower Input terminal.

Derived

Derived crossovers include active crossovers in which one of the crossover responses is derived from the other through the use of a differential amplifier. For example, the difference between the input signal and the output of the high pass section is a low pass response. Thus, when a differential amplifier is used to extract this difference, its output constitutes the low pass filter section. The main advantage of derived filters is that they produce no phase difference between the high pass and low pass sections at any frequency. The disadvantages are either

- (a) that the high pass and low pass sections often have different levels of attenuation in their stop bands, *i.e.* their slopes are asymmetrical, or
- (b) that the response of one or both sections peaks near the crossover frequency,

or both. In case (a), above, the usual situation is that the derived low pass response attenuates at a much slower rate than the fixed response. This requires the speaker to which it is directed to continue to respond to signals deep into the stopband where its physical characteristics may not be ideal. In the case of (b), above, both speakers are required to operate at higher volume levels as the signal nears the crossover points. This uses more amplifier power and may drive the speaker cones into non-linearity.

Chapter- 3

Composite Image Filter

A **composite image filter** is an electronic filter consisting of multiple image filter sections of two or more different types.

The image method of filter design determines the properties of filter sections by calculating the properties they have in an infinite chain of such sections. In this, the analysis parallels transmission line theory on which it is based. Filters designed by this method are called *image parameter filters*, or just *image filters*. An important parameter of image filters is their image impedance, the impedance of an infinite chain of identical sections.

The basic sections are arranged into a ladder network of several sections, the number of sections required is mostly determined by the amount of stopband rejection required. In its simplest form, the filter can consist entirely of identical sections. However, it is more usual to use a composite filter of two or three different types of section to improve different parameters best addressed by a particular type. The most frequent parameters considered are stopband rejection, steepness of the filter skirt (transition band) and impedance matching to the filter terminations.

Image filters are linear filters and are invariably also passive in implementation.

History

The image method of designing filters originated at AT&T, who were interested in developing filtering that could be used with the multiplexing of many telephone channels on to a single cable. The researchers involved in this work and their contributions are briefly listed below;

- John Carson provided the mathematical underpinning to the theory. He invented single sideband modulation for the purpose of multiplexing telephone channels. It was the need to recover these signals that gave rise to the need for advanced filtering techniques. He also pioneered the use of operational calculus (what has

now become Laplace transforms in its more formal mathematical guise) to analyse these signals.
- George Campbell worked on filtering from 1910 onwards and invented the constant k filter. This can be seen as a continuation of his work on loading coils on transmission lines, a concept invented by Oliver Heaviside. Heaviside, incidentally, also invented the operational calculus used by Carson.
- Otto Zobel provided a theoretical basis (and the name) for Campbell's filters. In 1920 he invented the m-derived filter. Zobel also published composite designs incorporating both constant k and m-derived sections.
- R S Hoyt also contributed.

The image method

The image analysis starts with a calculation of the input and output impedances (the image impedances) and the transfer function of a section in an infinite chain of identical sections. This can be shown to be equivalent to the performance of a section terminated in its image impedances. The image method, therefore, relies on each filter section being terminated with the correct image impedance. This is easy enough to do with the internal sections of a multiple section filter, because it is only necessary to ensure that the sections facing the one in question have identical image impedances. However, the end sections are a problem. They will usually be terminated with fixed resistances that the filter cannot match perfectly except at one specific frequency. This mismatch leads to multiple reflections at the filter terminations and at the junctions between sections. These reflections result in the filter response deviating quite sharply from the theoretical, especially near the cut-off frequency.

The requirement for better matching to the end impedances is one of the main motivations for using composite filters. A section designed to give good matching is used at the ends but something else (for instance stopband rejection or passband to stopband transition) is designed for the body of the filter.

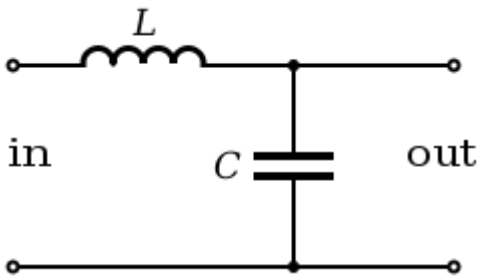
Filter section types

Each filter section type has particular advantages and disadvantages and each has the capability to improve particular filter parameters. The sections described below are the prototype filters for low-pass sections. These prototypes may be scaled and transformed to the desired frequency bandform (low-pass, high-pass, band-pass or band-stop).

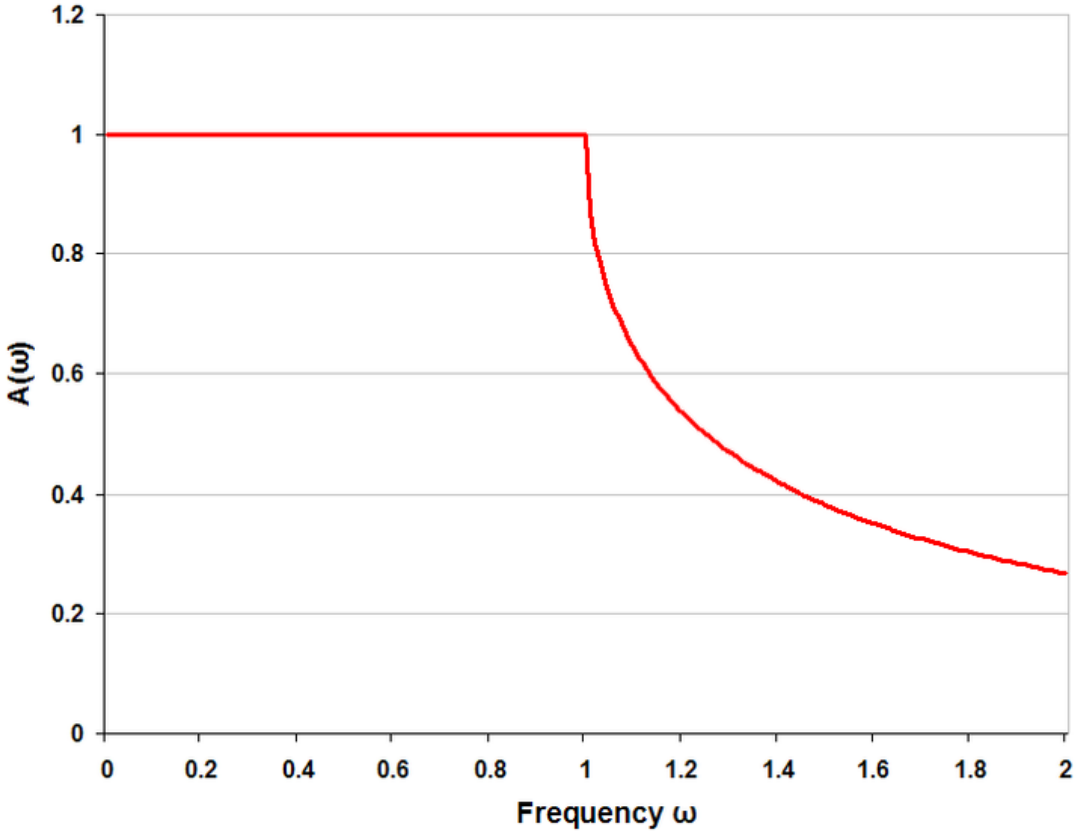
The smallest unit of an image filter is an L half-section. Because the L section is not symmetrical, it has different image impedances (Z_{i}) on each side. These are denoted $Z_{\mathrm{iT}}$and $Z_{\mathrm{i\Pi}}$. The T and the Π in the suffix refer to the shape of the filter section that would be formed if two half sections were to be connected back-to-back. T and Π are the smallest symmetrical sections that can be constructed, as shown in diagrams in the topology chart (below). Where the section in question has an image impedance different from the general case a further suffix is added identifying the section type, for instance $Z_{\mathrm{iT}m}$.

Constant k section

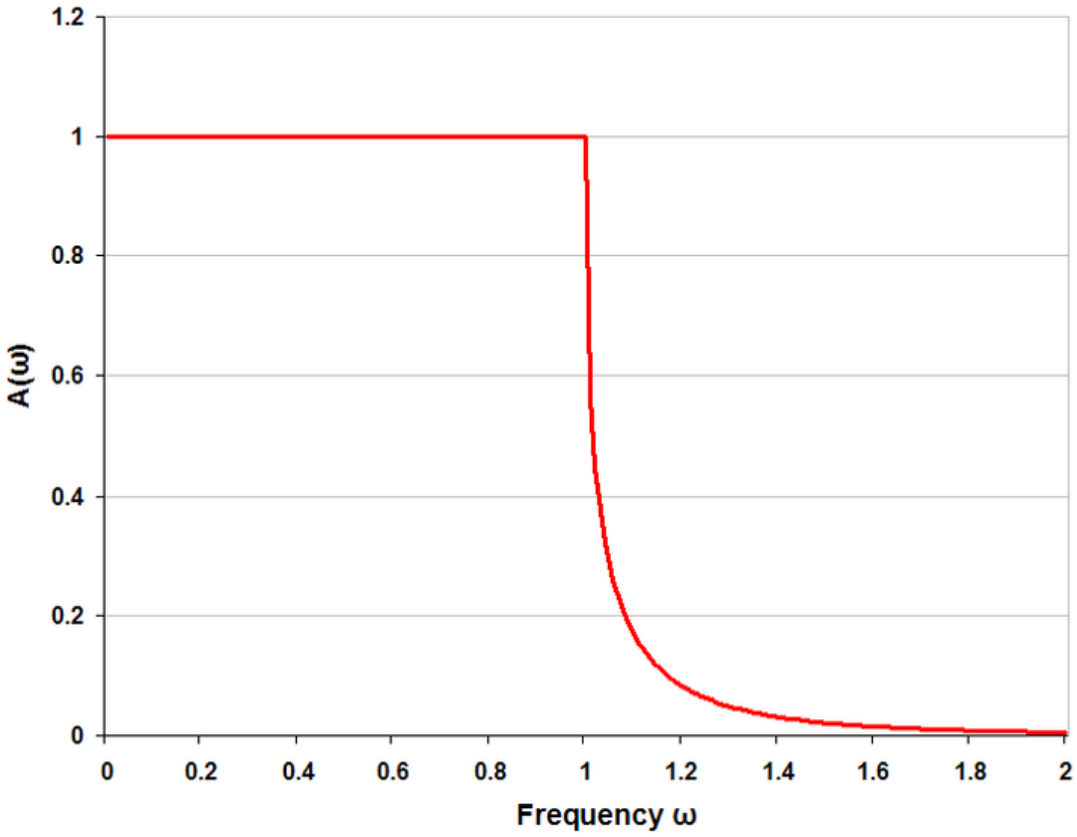
The **constant k** or **k-type** filter section is the basic image filter section. It is also the simplest circuit topology. The k-type has moderately fast transition from the passband to the stopband and moderately good stopband rejection.



k-type low-pass filter half section



k-type low-pass response, single half-section



k-type low-pass response with four (half) sections

m-derived section

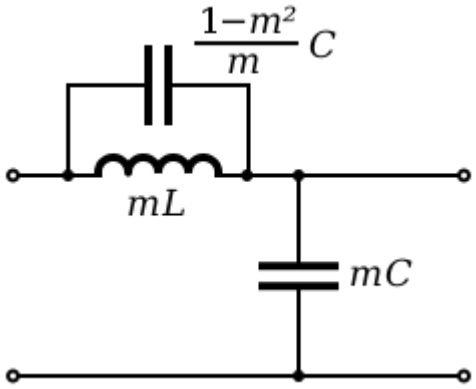
The **m-derived** or **m-type** filter section is a development of the k-type section. The most prominent feature of the m-type is a pole of attenuation just past the cut-off frequency inside the stopband. The parameter m ($0<m<1$) adjusts the position of this pole of attenuation. Smaller values of m put the pole closer to the cut-off frequency. Larger values of m put it further away. In the limit, as m approaches unity, the pole approaches ω of infinity and the section approaches a k-type section.

The m-type has a particularly fast cut-off, going from fully pass at the cut-off frequency to fully stop at the pole frequency. The cut-off can be made faster by moving the pole nearer to the cut-off frequency. This filter has the fastest cut-off of any filter design; note that the fast transition is achieved with just a single section, there is no need for multiple sections. The drawback with m-type sections is that they have poor stopband rejection past the pole of attenuation.

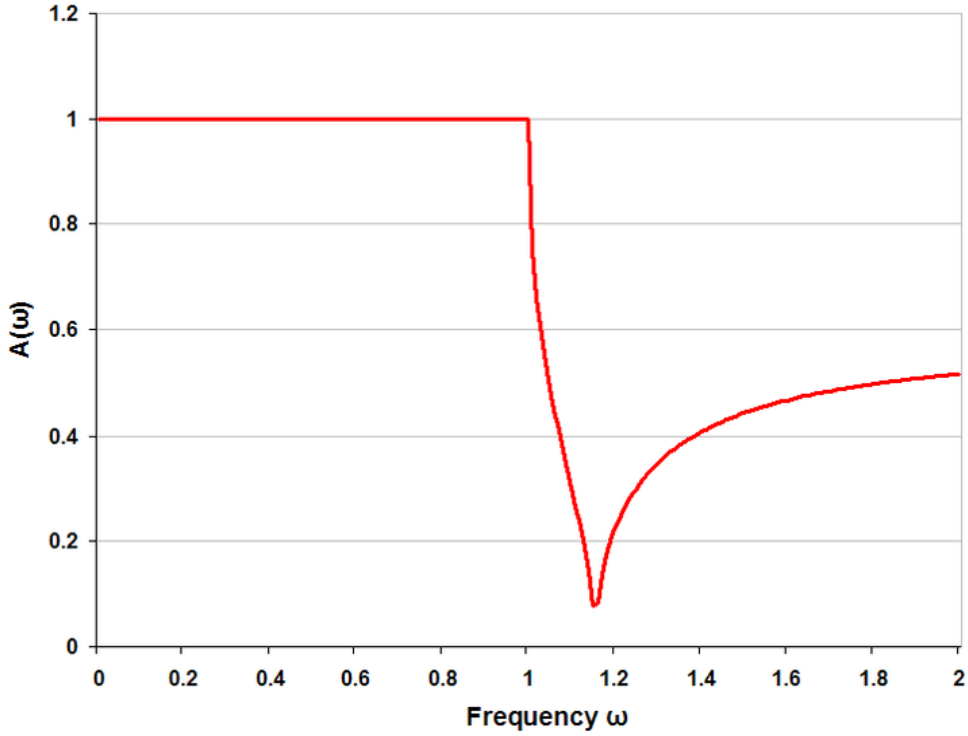
There is a particularly useful property of m-type filters with m=0.6. These have maximally flat image impedance Z_{im} in the passband. They are therefore good for

matching in to the filter terminations, in the passband at least, the stopband is another story.

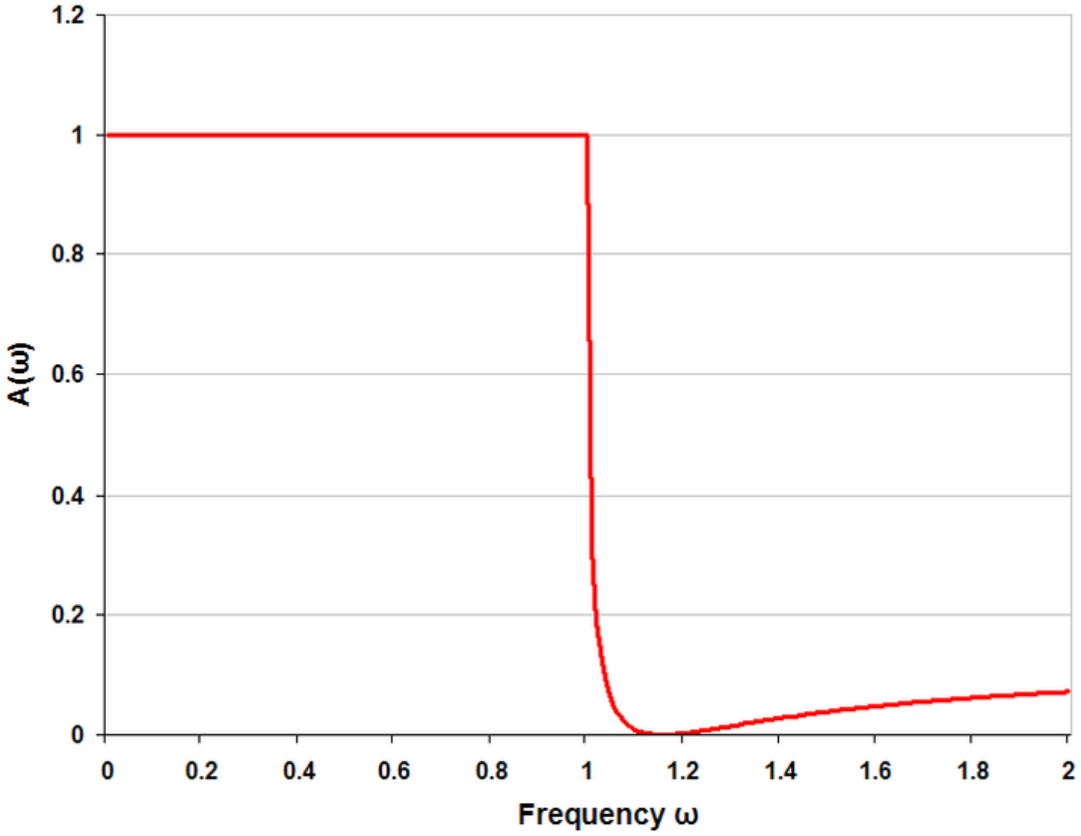
There are two variants of the m-type section, *series* and *shunt*. They have identical transfer functions but their image impedances are different. The shunt half-section has an image impedance which matches $Z_{\mathrm{i}\Pi}$on one side but has a different impedance, $Z_{\mathrm{i}\mathrm{T}m}$on the other. The series half-section matches $Z_{\mathrm{i}\mathrm{T}}$on one side and has $Z_{\mathrm{i}\Pi m}$on the other.



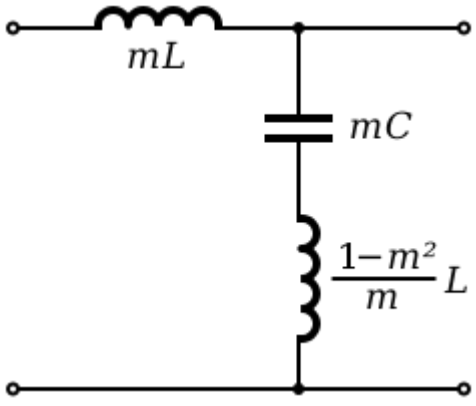
m-type low-pass filter shunt half section



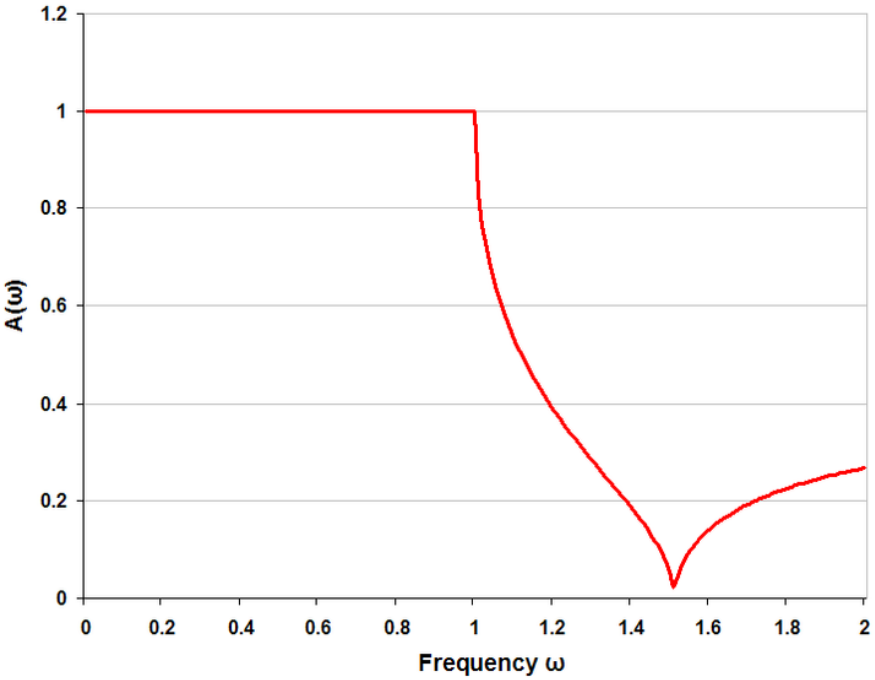
m-type low-pass response single half-section m=0.5



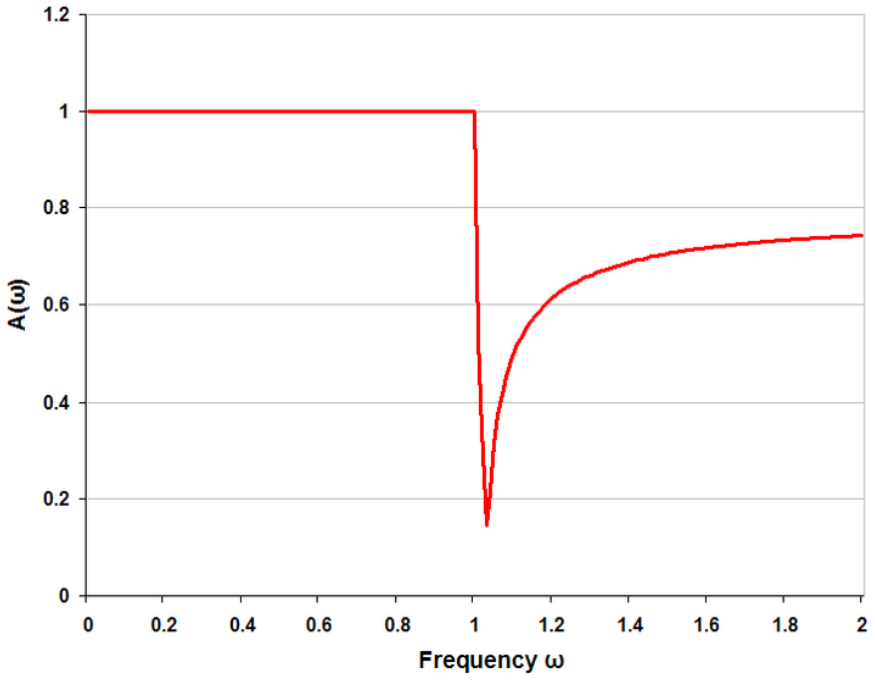
m-type low-pass response with four (half) sections m=0.5



m-type low-pass filter series half section



m-type low-pass response single half-section *m*=0.75



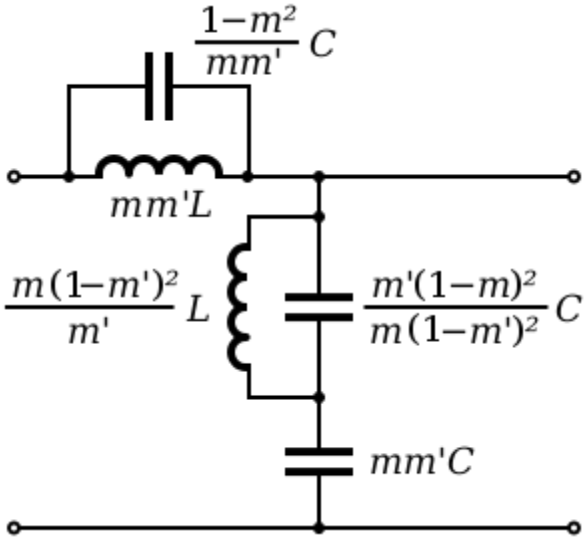
m-type low-pass response single half-section *m*=0.25

mm'-type section

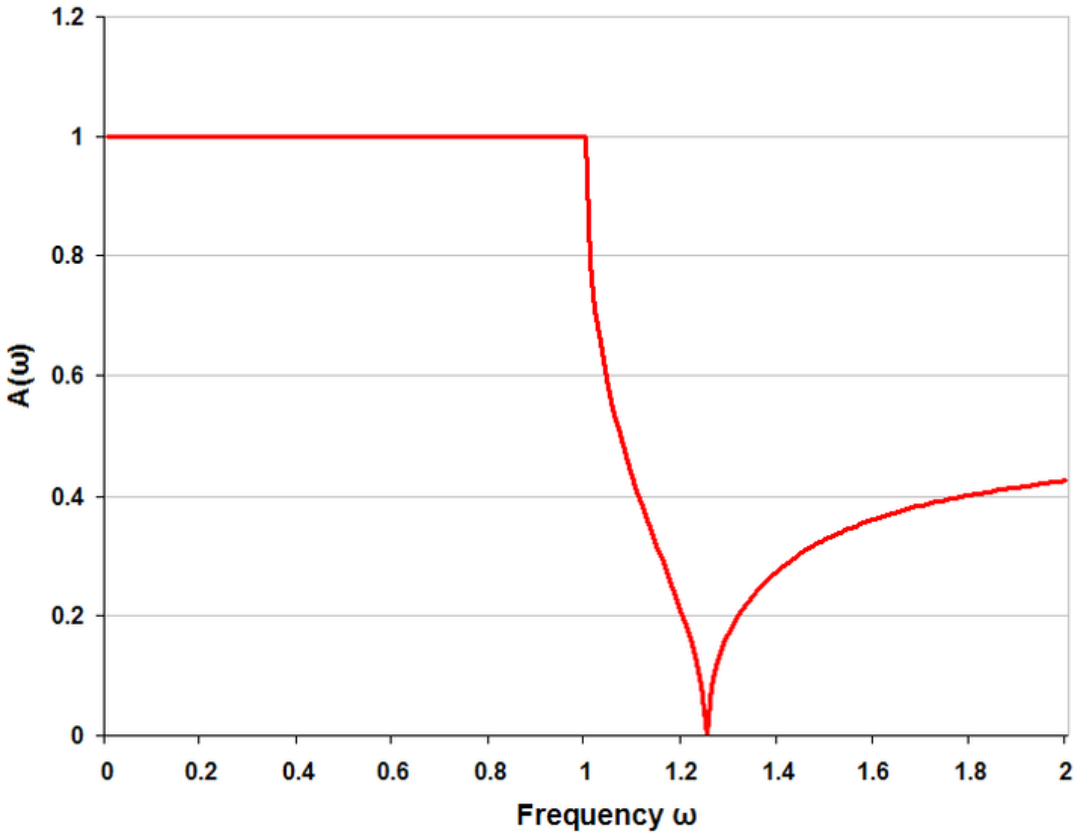
The **mm'-type** section has two independent parameters (*m* and *m*') that the designer can adjust. It is arrived at by double application of the *m*-derivation process. Its chief advantage is that it rather better at matching in to resistive end terminations than the k-type or m-type. The image impedance of a half-section is $Z_{\mathrm{i}m}$on one side and a different impedance, $Z_{\mathrm{i}mm'}$on the other. Like the m-type, this section can be constructed as a series or shunt section and the image impedances will come in T and Π variants. Either a series construction is applied to a shunt m-type or a shunt construction is applied to a series m-type. The advantages of the *mm*'-type filter are achieved at the expense of greater circuit complexity so it would normally only be used where it is needed for impedance matching purposes and not in the body of the filter.

The transfer function of an mm'-type is the same as an m-type with *m* set to the product *mm*'. To choose values of *m* and *m*' for best impedance match requires the designer to choose two frequencies at which the match is to be exact, at other frequencies there will be some deviation. There is thus some leeway in the choice but Zobel suggests the values *m*=0.7230 and *m*'=0.4134 which give a deviation of the impedance of less than 2% over the useful part of the band. Since *mm*'=0.3, this section will also have a much faster cut-off than an m-type of *m*=0.6 which is an alternative for impedance matching.

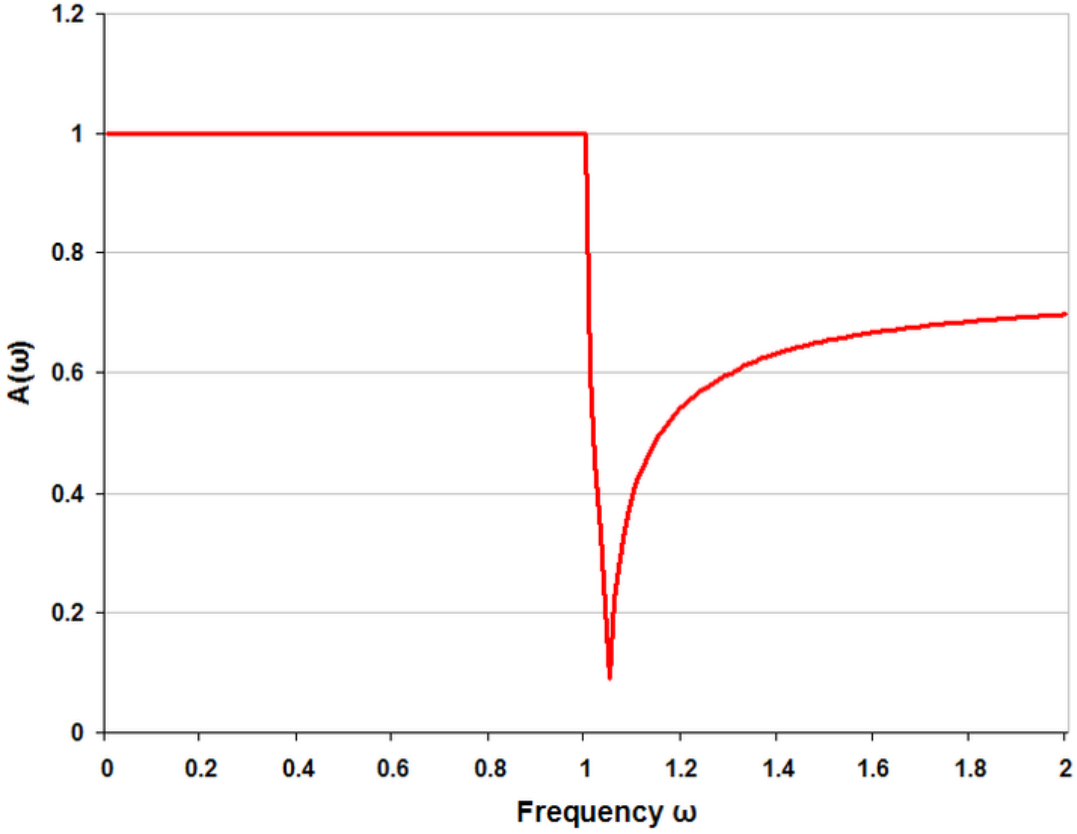
It is possible to continue the m-derivation process repeatedly and produce mm'm"-types and so on. However, the improvements obtained diminish at each iteration and are not usually worth the increase in complexity.



mm'-type low-pass filter series half section

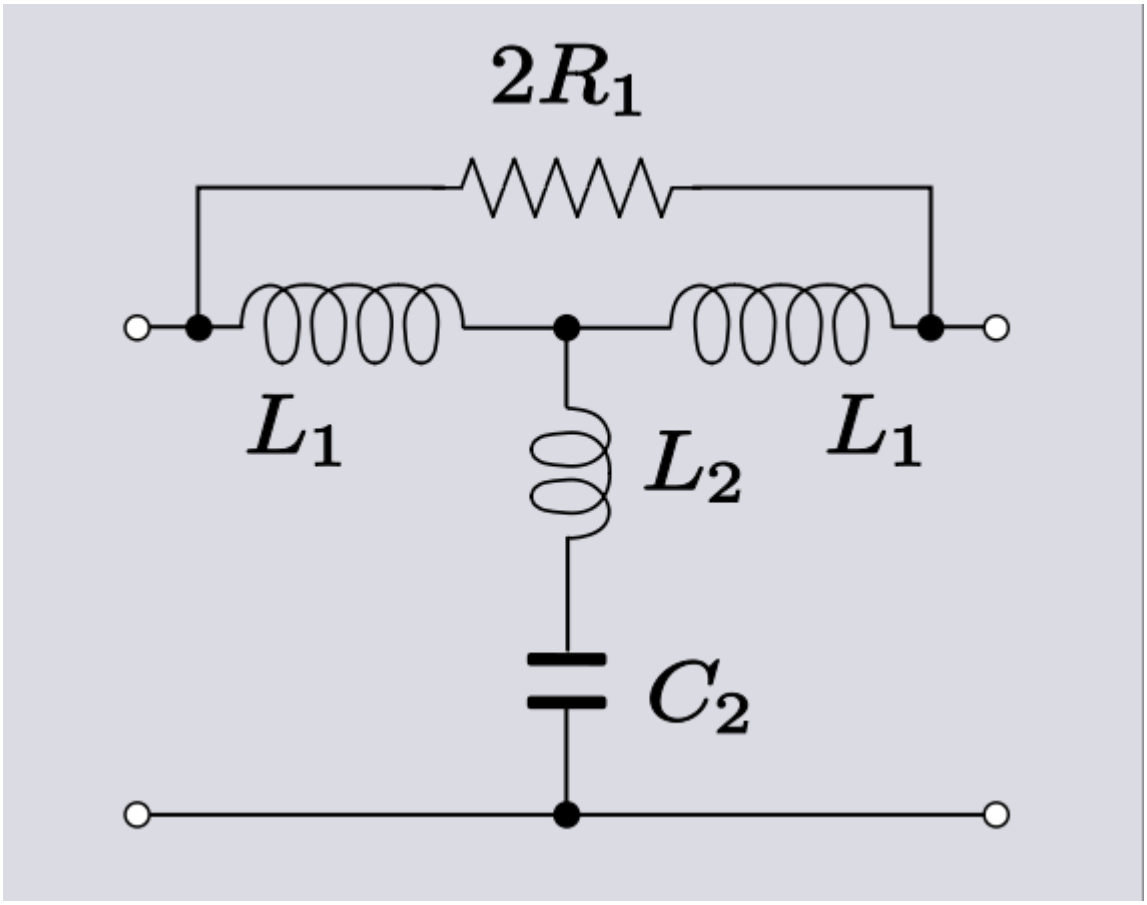


m-type low-pass response single half-section *m*=0.6



mm'-type low-pass response single half-section *mm*'=0.3

Bode's filter

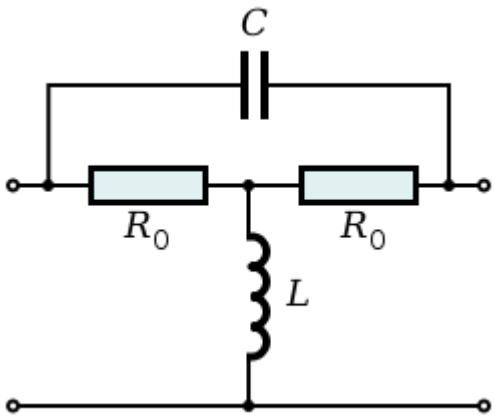


One incarnation of Bode's filter as a low-pass filter.

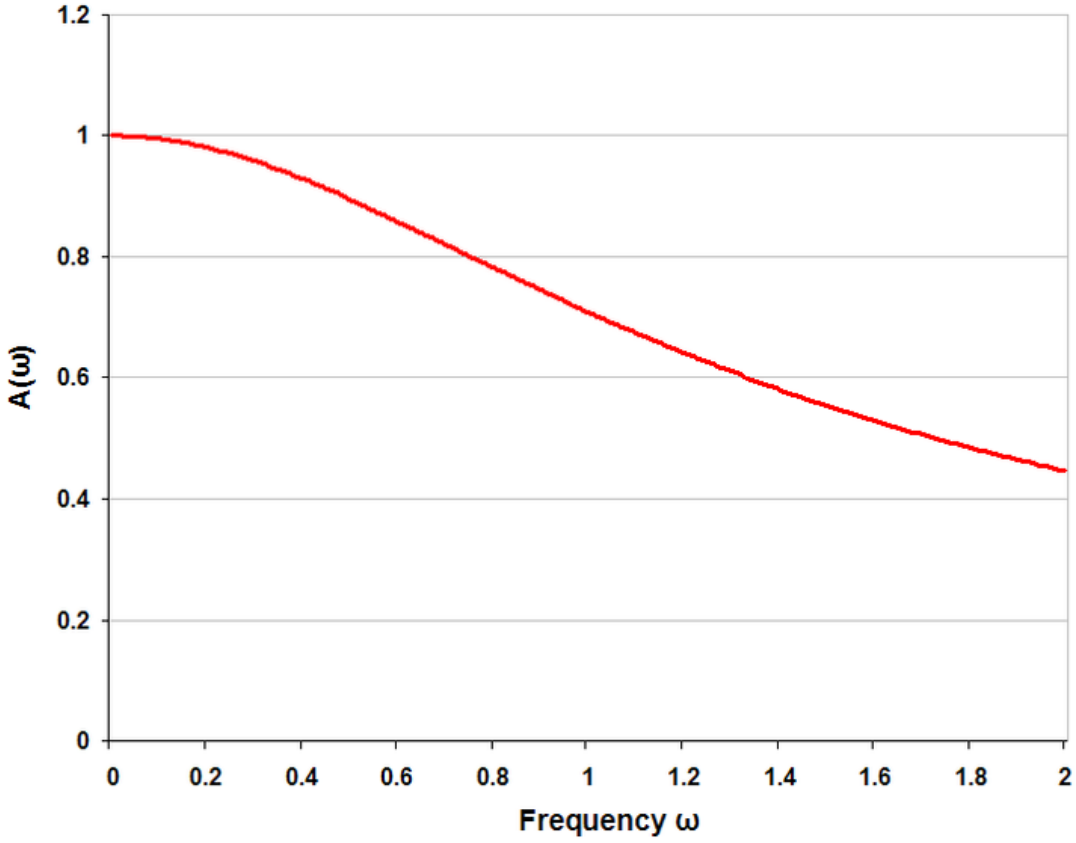
Another variation on the m-type filter was described by Hendrik Bode. This filter uses as a prototype a mid-series m-derived filter and transforms this into a bridged-T topology with the addition of a bridging resistor. This section has the advantage of being able to place the pole of attenuation much closer to the cut-off frequency than the Zobel filter, which starts to fail to work properly with very small values of *m* because of inductor resistance.

Zobel network

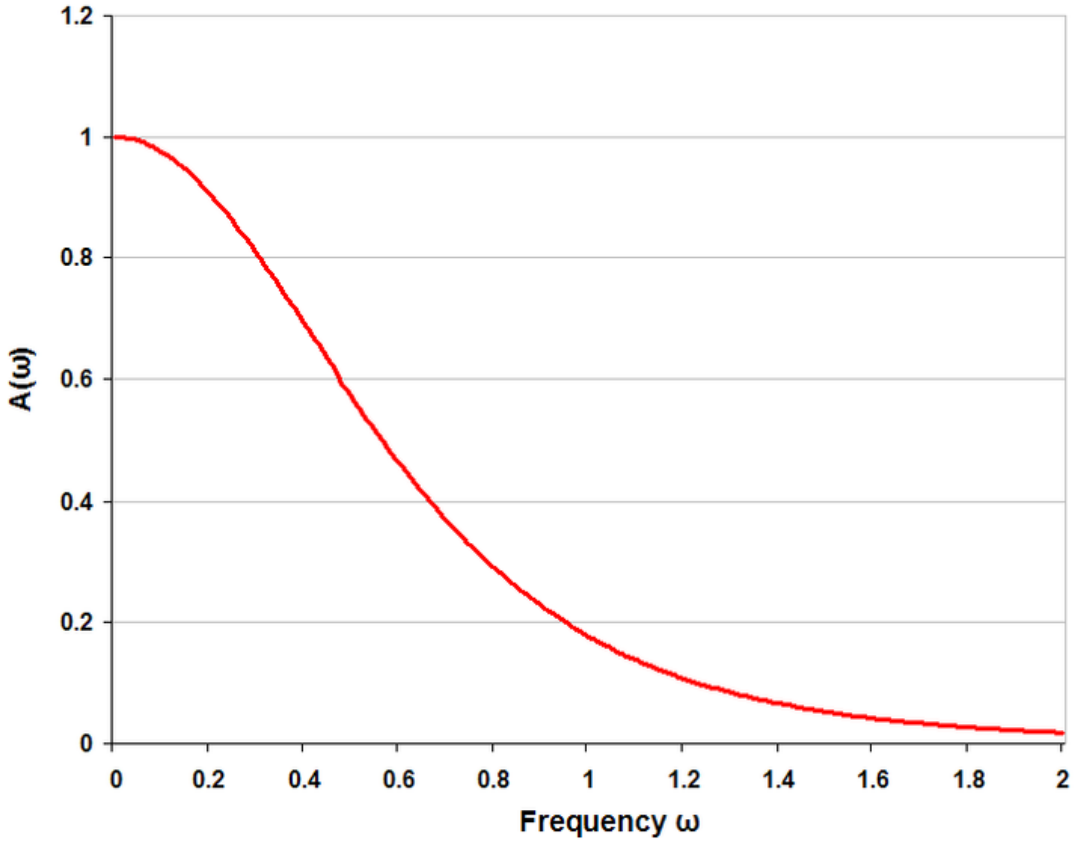
The distinguishing feature of **Zobel network** filters is that they have a constant resistance image impedance and for this reason are also known as **constant resistance networks**. Clearly, the Zobel network filter does not have a problem matching to its terminations and this is its main advantage. However, other filter types have steeper transfer functions and sharper cut-offs. In filtering applications, the main role of Zobel networks is as equalisation filters. Zobel networks are in a different group from other image filters. The constant resistance means that when used in combination with other image filter sections the same problem of matching arises as with end terminations. Zobel networks also suffer the disadvantage of using far more components than other equivalent image sections.



Zobel network bridge T high-pass filter section



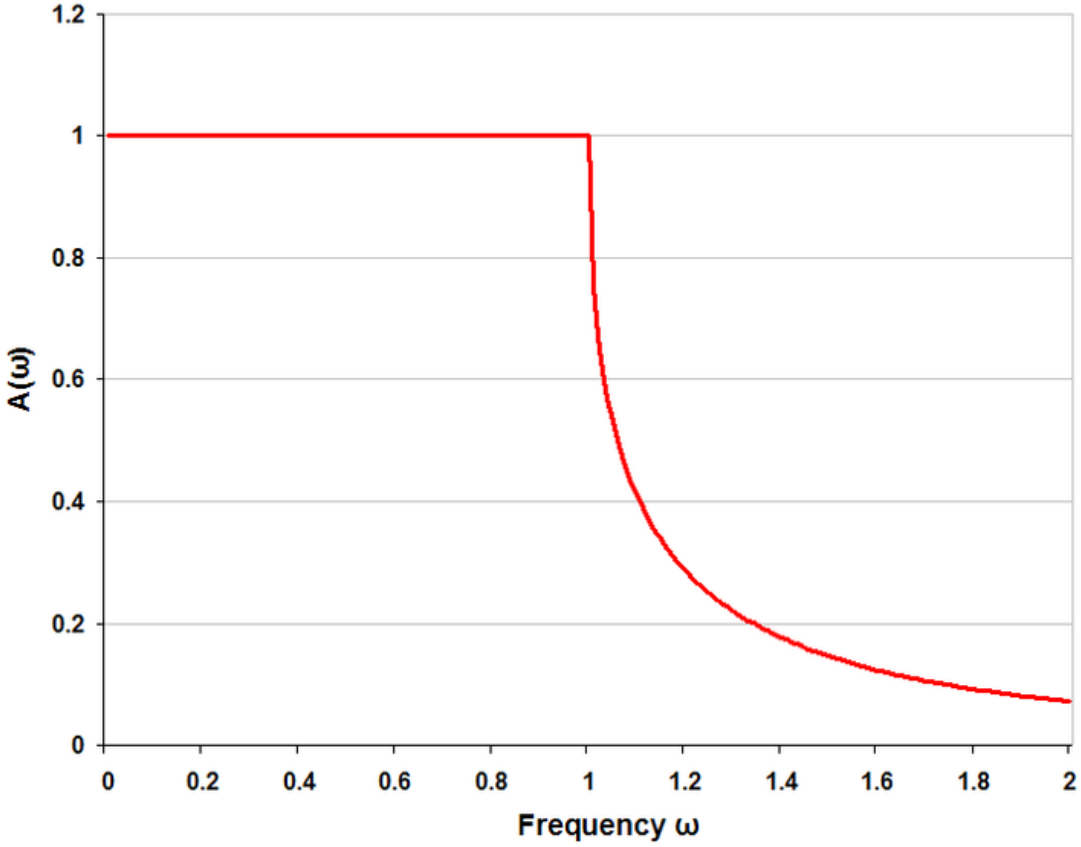
Zobel network low-pass response single section



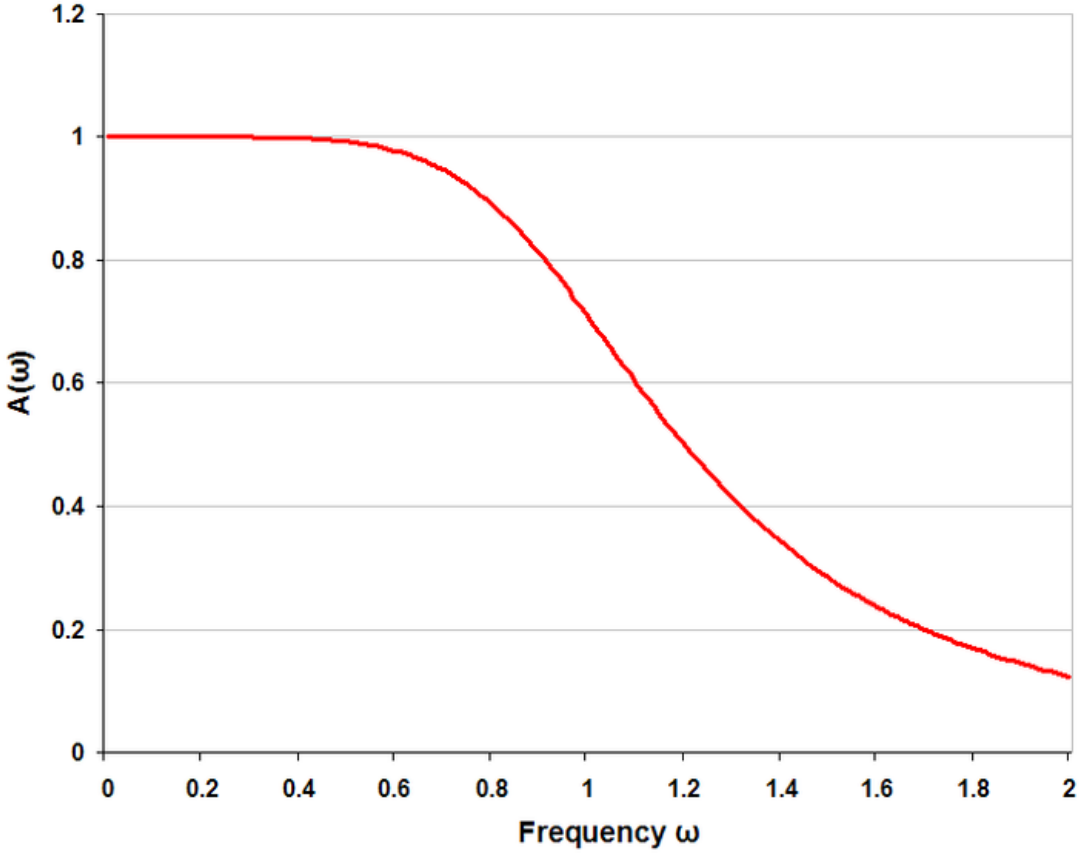
Zobel network low-pass response five sections

Effect of end terminations

A consequence of the image method of filter design is that the effect of the end terminations has to be calculated separately if its effects on response are to be taken into account. The most severe deviation of the response from that predicted occurs in the passband close to cut-off. The reason for this is twofold. Further into the passband the impedance match progressively improves, thus limiting the error. On the other hand, waves in the stopband are reflected from the end termination due to mismatch but are attenuated twice by the filter stopband rejection as they pass through it. So while stopband impedance mismatch may be severe, it has only limited effect on the filter response.



Theoretical k-type low-pass T-filter (two half-sections) response when correctly terminated in image impedance



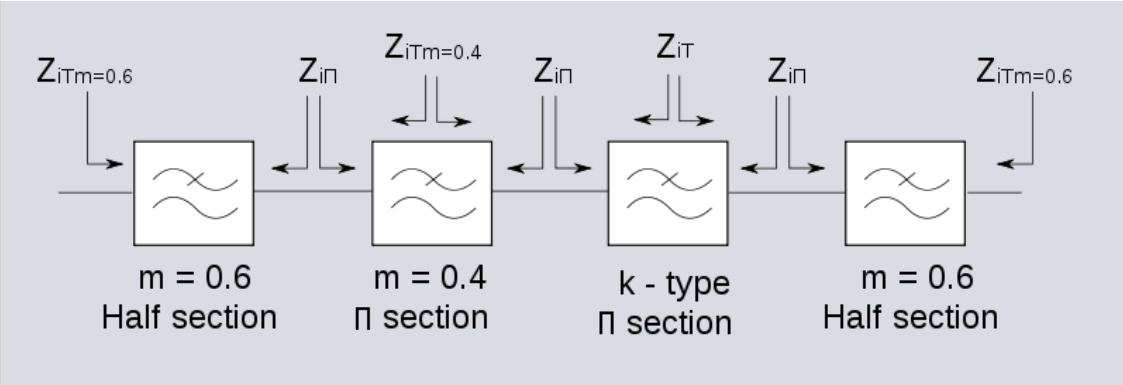
Practical k-type low-pass T-filter (two half-sections) response when terminated with fixed resistors

Cascading sections

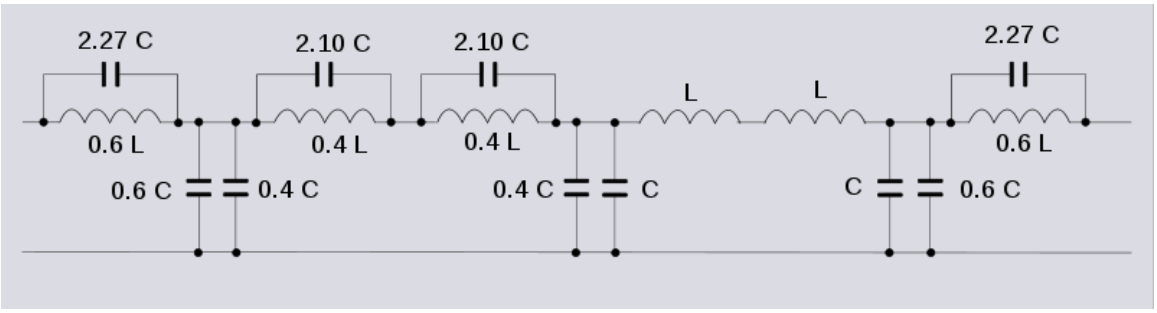
Several L half-sections may be cascaded to form a composite filter. The most important rule when constructing a composite image filter is that the image impedances must always face an identical impedance; like must always face like. T sections must always face T sections, Π sections must always face Π sections, k-type must always face k-type (or the side of an m-type which has the k-type impedance) and m-type must always face m-type. Furthermore, m-type impedances of different values of m cannot face each other. Nor can sections of any type which have different values of cut-off frequency.

Sections at the beginning and end of the filter are often chosen for their impedance match in to the terminations rather than the shape of their frequency response. For this purpose, m-type sections of m = 0.6 are the most common choice. An alternative is mm'-type sections of m=0.7230 and m'=0.4134 although this type of section is rarely used. While it has several advantages noted below, it has the disadvantages of being more complex and also, if constant k sections are required in the body of the filter, it is then necessary to include m-type sections to interface the mm'-type to the k-types.

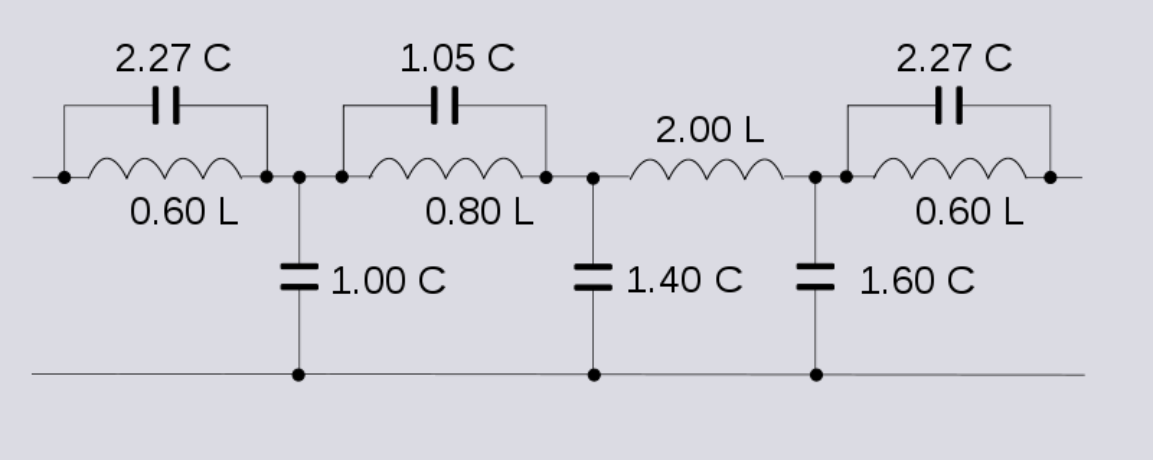
The inner sections of the filter are most commonly chosen to be constant k since these produce the greatest stopband attenuation. However, one or two m-type sections might also be included to improve the rate of fall from pass to stopband. A low value of m is chosen for m-types used for this purpose. The lower the value of *m*, the faster the transition, while at the same time, the stopband attenuation becomes less, increasing the need to use extra k-type sections as well. An advantage of using mm'-types for impedance matching is that these type of end sections will have a fast transition anyway (much more so than *m*=0.6 m-type) because *mm*'=0.3 for impedance matching. So the need for sections in the body of the filter to do this may be dispensed with.



Typical example of a composite image filter in block diagram form. The image impedances and how they match are shown.



The above filter realised as a ladder low-pass filter. Component values are given in terms of L and C, the component values of a constant k half-section.



The same filter minimised by combining components in series or parallel where appropriate.

Another reason for using m-types in the body of the filter is to place an additional pole of attenuation in the stopband. The frequency of the pole directly depends on the value of *m*. The smaller the value of *m*, the closer the pole is to the cut-off frequency. Conversely, a large value of *m* places the pole further away from cut-off until in the limit when *m*=1 the pole is at infinity and the response is the same as the k-type section. If a value of *m* is chosen for this pole which is different from the pole of the end sections it will have the effect of broadening the band of good stopband rejection near to the cut-off frequency. In this way the m-type sections serve to give good stopband rejection near to cut-off and the k-type sections give good stopband rejection far from cut-off. Alternatively, m-type sections can be used in the body of the filter with different values of m if the value found in the end sections is unsuitable. Here again, the mm'-type would have some advantages if used for impedance matching. The mm'-type used for impedance matching places the pole at *m*=0.3. However, the other half of the impedance matching section needs to be an m-type of *m*=0.723. This automatically gives a good spread of stopband rejection and as with the steepness of transition issue, use of mm'-type sections may remove the need for additional m-type sections in the body.

Constant resistance sections may also be required, if the filter is being used on a transmission line, to improve the flatness of the passband response. This is necessary because the transmission line response is not usually anywhere near perfectly flat. These sections would normally be placed closest to the line since they present a predictable impedance to the line and also tend to mask the indeterminate impedance of the line from the rest of the filter. There is no issue with matching constant resistance sections to each other even when the sections are operating on totally different frequency bands. All sections can be made to have precisely the same image impedance of a fixed resistance.

Chapter- 4

Current Source

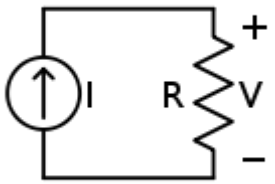


Figure 1: An ideal current source, *I*, driving a resistor, *R*, and creating a voltage *V*

A **current source** is an electrical or electronic device that delivers or absorbs electric current. A current source is the dual of a voltage source. The term constant-current **sink** is sometimes used for sources fed from a negative voltage supply. Figure 1 shows a schematic for an ideal current source driving a resistor load.

Ideal current sources

Voltage source

Current Source

Controlled Voltage Source Controlled Current Source

Battery of cells Single cell

Figure 2: Source symbols

In circuit theory, an **ideal current source** is a circuit element where the current through it is independent of the voltage across it. It is a mathematical model, which real devices can only approach in performance. If the current through an ideal current source can be specified independently of any other variable in a circuit, it is called an *independent* current source. Conversely, if the current through an ideal current source is determined by some other voltage or current in a circuit, it is called a **dependent** or **controlled current source**. Symbols for these sources are shown in Figure 2.

An independent current source with zero current is identical to an ideal open circuit. For this reason, the internal resistance of an ideal current source is infinite. The voltage across an ideal current source is completely determined by the circuit it is connected to. When connected to a short circuit, there is zero voltage and thus zero power delivered. When connected to a load resistance, the voltage across the source approaches infinity as the load resistance approaches infinity (an open circuit). Thus, an ideal current source, if such a thing existed in reality, could supply unlimited power and so would represent an unlimited source of energy.

No real current source is ideal (no unlimited energy sources exist) and all have a finite internal resistance (none can supply unlimited voltage). However, the internal resistance of a physical current source is effectively modeled in circuit analysis by combining a non-zero resistance in parallel with an ideal current source (the Norton equivalent circuit). The connection of an ideal open circuit to an ideal non-zero current source does not represent any physically realizable system.

Physical current sources

Resistor current source

The simplest non-ideal current source consists of a voltage source in series with a resistor. The current available from such a source is given by the ratio of the voltage across the voltage source to the resistance of the resistor. This value of current will only be delivered to a load with zero voltage drop across its terminals (a short circuit, an uncharged capacitor, a charged inductor, a virtual ground circuit, etc.) The current delivered to a load with nonzero voltage (drop) across its terminals (a linear or nonlinear

resistor with a finite resistance, a charged capacitor, an uncharged inductor, a voltage source, etc.) will always be different. It is given by the ratio of the voltage drop across the resistor (the difference between the exciting voltage and the voltage across the load) to its resistance. For a nearly ideal current source, the value of the resistor should be very large but this implies that, for a specified current, the voltage source must be very large (in the limit as the resistance and the voltage go to infinity, the current source will become ideal and the current will not depend at all on the voltage across the load). Thus, efficiency is low (due to power loss in the resistor) and it is usually impractical to construct a 'good' current source this way. Nonetheless, it is often the case that such a circuit will provide adequate performance when the specified current and load resistance are small. For example, a 5 V voltage source in series with a 4.7 kilohm resistor will provide an *approximately* constant current of 1 mA (±5%) to a load resistance in the range of 50 to 450 ohm.

A Van de Graaff generator is an example of such a high voltage current source. It behaves as an almost constant current source because of its very high output voltage coupled with its very high output resistance and so it supplies the same few microamperes at any output voltage up to hundreds of thousands of volts (or even tens of megavolts) for large laboratory versions.

Active current sources without negative feedback

Active current sources have many important applications in electronic circuits. Current sources (current-stable resistors) are often used in place of ohmic resistors in analog integrated circuits to generate a current that depends slightly on the voltage across the load.

Transistor current sources with constant input voltage

The common collector, common drain and common cathode configurations driven by a constant input voltage naturally behave as current sources (or sinks) because the output impedance of these devices is naturally high. The simple current mirror is an example of such a current source widely used in integrated circuits.

FET current sources with zero input voltage

A JFET can be made to act as a current source by tying its gate to its source. The current then flowing is the I_{DSS} of the FET. These can be purchased with this connection already made and in this case the devices are called current regulator diodes or constant current diodes or current limiting diodes (CLD). An enhancement mode N channel MOSFET can be used in the circuits listed below.

Current sources with series negative feedback

Simple transistor current source

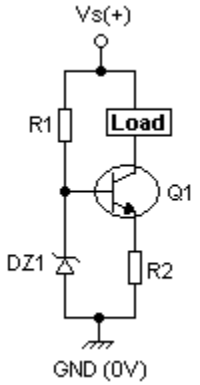


Figure 3: Typical constant current source (CCS)

Figure 3 shows a typical constant current source (CCS). DZ1 is a zener diode which, when reverse biased (as shown in the circuit) has a constant voltage drop across it irrespective of the current flowing through it. Thus, as long as the zener current (I_Z) is above a certain level (called holding current), the voltage across the zener diode (V_Z) will be constant. Resistor R1 supplies the zener current and the base current (I_B) of NPN transistor (Q1). The constant zener voltage is applied across the base of Q1 and emitter resistor R2. The operation of the circuit is as follows:

Voltage across R2 (V_{R2}) is given by V_Z - V_{BE}, where V_{BE} is the base-emitter drop of Q1. The emitter current of Q1 which is also the current through R2 is given by

$$I_{R2}(= I_E) = \frac{V_{R2}}{R2} = \frac{V_Z - V_{BE}}{R2}$$

Since V_Z is constant and V_{BE} is also (approximately) constant for a given temperature, it follows that V_{R2} is constant and hence I_E is also constant. Due to transistor action, emitter current I_E is very nearly equal to the collector current I_C of the transistor (which in turn, is the current through the load). Thus, the load current is constant (neglecting the output resistance of the transistor due to the Early effect) and the circuit operates as a constant current source. As long as the temperature remains constant (or doesn't vary much), the load current will be independent of the supply voltage, R1 and the transistor's gain. R2 allows the load current to be set at any desirable value and is calculated by

$$R2 = \frac{V_Z - V_{BE}}{I_{R2}}$$

or

$$R2 = \frac{V_Z - 0.65}{I_{R2}},$$

since V_{BE} is typically 0.65 V for a silicon device.

(I_{R2} is also the emitter current and is assumed to be the same as the collector or required load current, provided h_{FE} is sufficiently large). Resistance R_1 at resistor R1 is calculated as

$$R_1 = \frac{V_S - V_Z}{I_Z + K \cdot I_B}$$

where K = 1.2 to 2 (so that R_1 is low enough to ensure adequate I_B),

$$I_B = \frac{I_C (= I_E = I_{R2})}{h_{FE(min)}}$$

and $h_{FE(min)}$ is the lowest acceptable current gain for the particular transistor type being used.

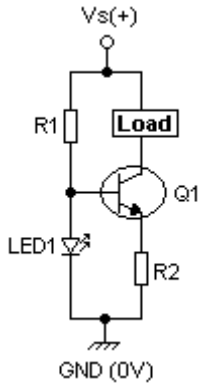


Figure 5: Typical constant current source (CCS) using LED instead of zener

The Zener diode can be replaced by any other diode, e.g. a light-emitting diode LED1 as shown in Figure 5. The LED voltage drop (V_D) is now used to derive the constant voltage and also has the additional advantage of tracking (compensating) V_{BE} changes due to temperature. R_2 is calculated as

$$R_2 = \frac{V_D - V_{BE}}{I_{R2}}$$

and R_1 as

$$R_1 = \frac{V_S - V_D}{I_D + K \cdot I_B}$$, where I_D is the LED current.

Simple transistor current source with diode compensation

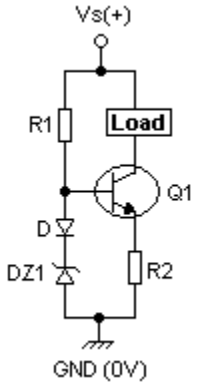


Figure 4: Typical constant current source (CCS) with diode compensation

Temperature changes will change the output current delivered by the circuit of Figure 3 because V_{BE} is sensitive to temperature. Temperature dependence can be compensated using the circuit of Figure 4 that includes a standard diode D (of the same semiconductor material as the transistor) in series with the Zener diode as shown in the image on the left. The diode drop (V_D) tracks the V_{BE} changes due to temperature and thus significantly counteracts temperature dependence of the CCS.

Resistance R_2 is now calculated as

$$R_2 = \frac{V_Z + V_D - V_{BE}}{I_{R2}}$$

Since $V_D = V_{BE} = 0.65$ V,

$$R_2 = \frac{V_Z}{I_{R2}}$$

(In practice V_D is never exactly equal to V_{BE} and hence it only suppresses the change in V_{BE} rather than nulling it out.)

R_1 is calculated as

$$R_1 = \frac{V_S - V_Z - V_D}{I_Z + K \cdot I_B}$$ (the compensating diode's forward voltage drop V_D appears in the equation and is typically 0.65 V for silicon devices.)

This method is most effective for Zener diodes rated at 5.6 V or more. For breakdown diodes of less than 5.6 V, the compensating diode is usually not required because the breakdown mechanism is not as temperature dependent as it is in breakdown diodes above this voltage.

Current mirror with emitter degeneration

Series negative feedback is also used in the two-transistor current mirror with emitter degeneration. Negative feedback is a basic feature in some current mirrors using multiple transistors, such as the Widlar current source and the Wilson current source.

Op-amp current sources...

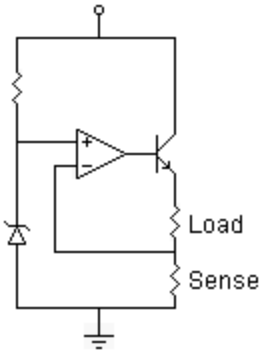


Figure 6: Typical op-amp current source. The transistor is not needed if the required current doesn't exceed the sourcing ability of the op-amp. The current will be the zener voltage divided by the sense resistor.

...with series negative feedback...

Another common method is to use a negative feedback to set the current and remove the dependence on the Vbe of the transistor. Figure 6 shows a very common approach using an op amp with the non-inverting input connected to a voltage source (such as the Zener in an above example) and the inverting input connected to the same node as the resistor and emitter of the transistor. The circuit is actually a non-inverting amplifier driven by a constant input voltage. It keeps up this constant voltage across the constant sense resistor; as a result, the current flowing through the load is constant as well.

...with parallel negative feedback

In the case of op-amp circuits (e.g., an op-amp voltage-to-current converter) sometimes it is desired to inject a precisely known current through a resistor into the inverting input (as an offset of signal input for instance). The combination of the input voltage source and the resistor will approximate an ideal current source with value V/R. The op-amp inverting input will be at virtual ground.

Current source made by a voltage regulator

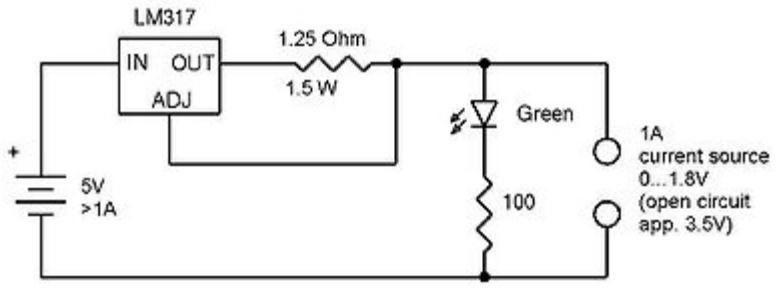


Figure 7: Constant current source using the LM317 voltage regulator

The circuit of Figure 7 using the LM317 voltage regulator is used to present a constant current source. The voltage regulator keeps up a constant voltage drop (1.25 V) across a constant resistor (1.25 Ω); so, a constant current (1 A) flows through the resistor and the load. The LED is on when the voltage across the load exceeds 1.8 V (the indicator circuit introduces some error).

Current and voltage source comparison

Most sources of electrical energy (mains electricity, a battery, ...) are best modeled as voltage sources. Such sources provide constant voltage, which means that as long as the amount of current drawn from the source is within the source's capabilities, its output voltage stays constant. An ideal voltage source provides no energy when it is loaded by an open circuit (i.e. an infinite impedance), but approaches infinite power and current when the load resistance approaches zero (a short circuit). Such a theoretical device would have a zero ohm output impedance in series with the source. A real-world voltage source has a very low, but non-zero output impedance: often much less than 1 ohm.

Conversely, a current source provides a constant current, as long as the load connected to the source terminals has sufficiently low impedance. An ideal current source would provide no energy to a short circuit and approach infinite energy and voltage as the load resistance approaches infinity (an open circuit). An *ideal* current source has an infinite output impedance in parallel with the source. A *real-world* current source has a very high, but finite output impedance. In the case of transistor current sources, impedances of a few megohms (at DC) are typical.

An *ideal* current source cannot be connected to an *ideal* open circuit because this would create the paradox of running a constant, non-zero current (from the current source) through an element with a defined zero current (the open circuit). Nor can an *ideal* voltage source be connected to an *ideal* short circuit (R=0), since this would result a similar paradox of finite non zero voltage across an element with defined zero voltage (the short circuit).

Because no ideal sources of either variety exist (all real-world examples have finite and non-zero source impedance), any current source can be considered as a voltage source with the *same* source impedance and vice versa. These concepts are dealt with by Norton's and Thévenin's theorems.

Chapter- 5

Current-to-voltage Converter

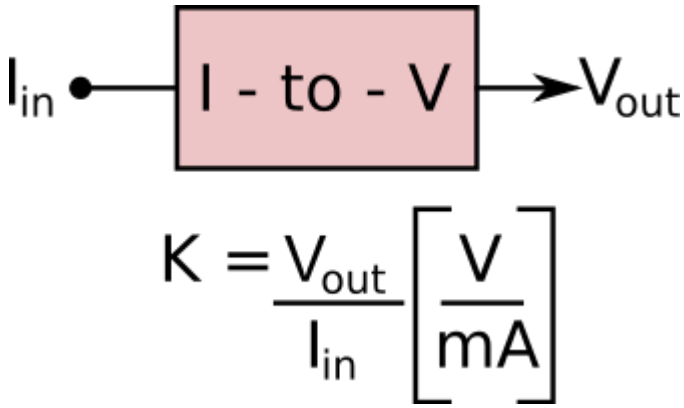


Fig. 1. Current-to-voltage converter (a block diagram)

A current-to-voltage converter (or transimpedance amplifier) is an electrical device that takes an electric current as an input signal and produces a corresponding voltage as an output signal. Three kinds of devices are used in electronics: generators (having only outputs), converters (having inputs and outputs) and loads (having only inputs). Most frequently, electronic devices use voltage as input/output quantity, as it generally requires less power consumption than using current.

In some cases, there is a need for converters having current as the input and voltage as the output. A typical situation is the measuring of a current using instruments having voltage inputs. A **current-to-voltage converter** is a circuit that performs current to voltage transformation. In electronic circuitry operating at signal voltages, it usually changes the electric attribute carrying information from current to voltage. The converter acts as a linear circuit with transfer ratio $k = V_{OUT}/I_{IN}$, called the transimpedance, which has dimensions of [V/A] (also known as resistance). That is why the active version of the circuit is also referred to as a **transresistance** or **transimpedance amplifier**.

Typical applications of current-to-voltage converter are measuring currents by using instruments having voltage inputs, creating current-controlled voltage sources, building various passive and active voltage-to-voltage converters, etc. In some cases, the simple

passive current-to-voltage converter works well; in other cases, there is a need of using *active current-to-voltage converters*. There is a close interrelation between the two versions - the active version has come from the passive one.

Ideal current-to-voltage converters have zero input resistance (impedance), so that they actually short the input source. Therefore, in this case, the input source has to have some resistance; ideally, it has to behave as a constant current source. Otherwise, the input source and the current-to-voltage converter can saturate.

The basic idea behind the passive version

Non-electrical domain: *Flow causes pressure*

In physical terms, there are many situations where a pressure-like quantity induces flow of a substance through an impediment. However, there are also corresponding situations where a flow-like quantity induces pressure at an impediment: *mechanical* (if one tries to stop a moving car with his body, the "flowing" car exerts pressure on him, the impediment), *pneumatic* (pinch a hose in the middle and you will see that a pressure appears at the pinch point).

In this arrangement, the flow-, pressure-, and impediment-like attributes are interrelated. Usually, the output pressure-like variable is proportional to the input flow-like one; in this way, the flow-like quantity creates (is converted to) a pressure-like one.

To induce pressure, an impediment must be put in the way of a flowing quantity.

Electrical domain: Current causes voltage

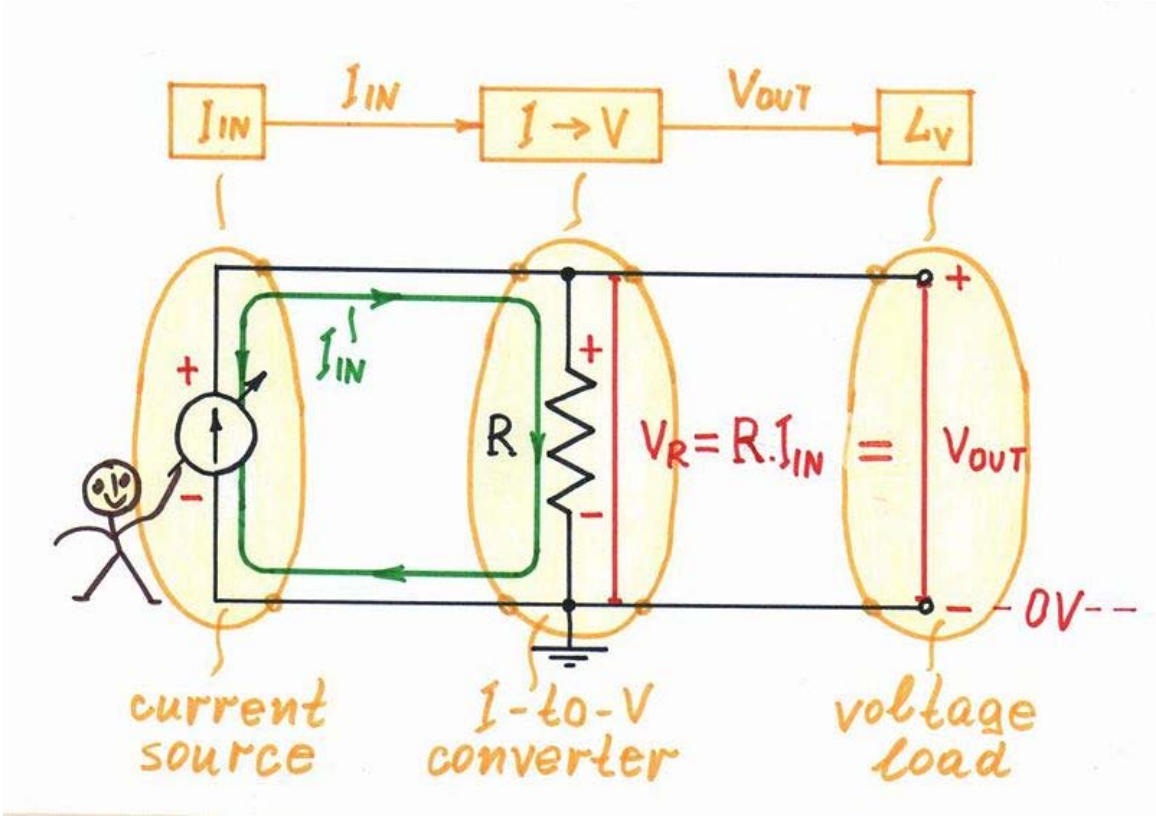


Fig. 2. The passive current-to-voltage converter is based on the *current-causes-voltage* phenomenon.

Building the circuit. Similarly, in electricity, if a current I_{IN} flows through a resistor R (Fig. 2), the latter impedes (resists) the current; as a result, a proportional voltage drop $V_R = R.I_{IN}$ appears across the resistor according to *current-causes-voltage* formulation of Ohm's law (V = R.I). In this *current-supplied circuit*, the voltage drop V_R acts as an output voltage V_{OUT} (the voltage drop V_R is created not by the resistor; it is created by the excitation voltage source inside the input current source). In this way, the current I_{IN} is converted to a proportional voltage V_{OUT}; the resistor R serves as a *current-to-voltage converter* - a linear circuit with transfer ratio $k = V_{OUT}/I_{IN}$ [V/mA] having dimensions of resistance.

Circuit operation. Fig. 2 represents graphically the circuit operation by using a current loop and voltage bars. The thickness of the current loop is proportional to the magnitude of the current and the height of the voltage bars is proportional to the corresponding voltages.

A graphoanalytical interpretation of the circuit (and of the Ohm's law) is shown on Fig. 3. As the current through and the voltage across the two components (the current source and the resistor) are the same, their IV curves are superimposed on a common coordinate system. The intersection of the two lines is the *operating point* A; it represents the present

magnitudes of the current I_A and the voltage V_A. When the current I_{IN} of the input current source varies, its IV curve moves vertically. As a result, the working point A slides over the IV curve of the resistor R; its slope represents the converter's ratio.

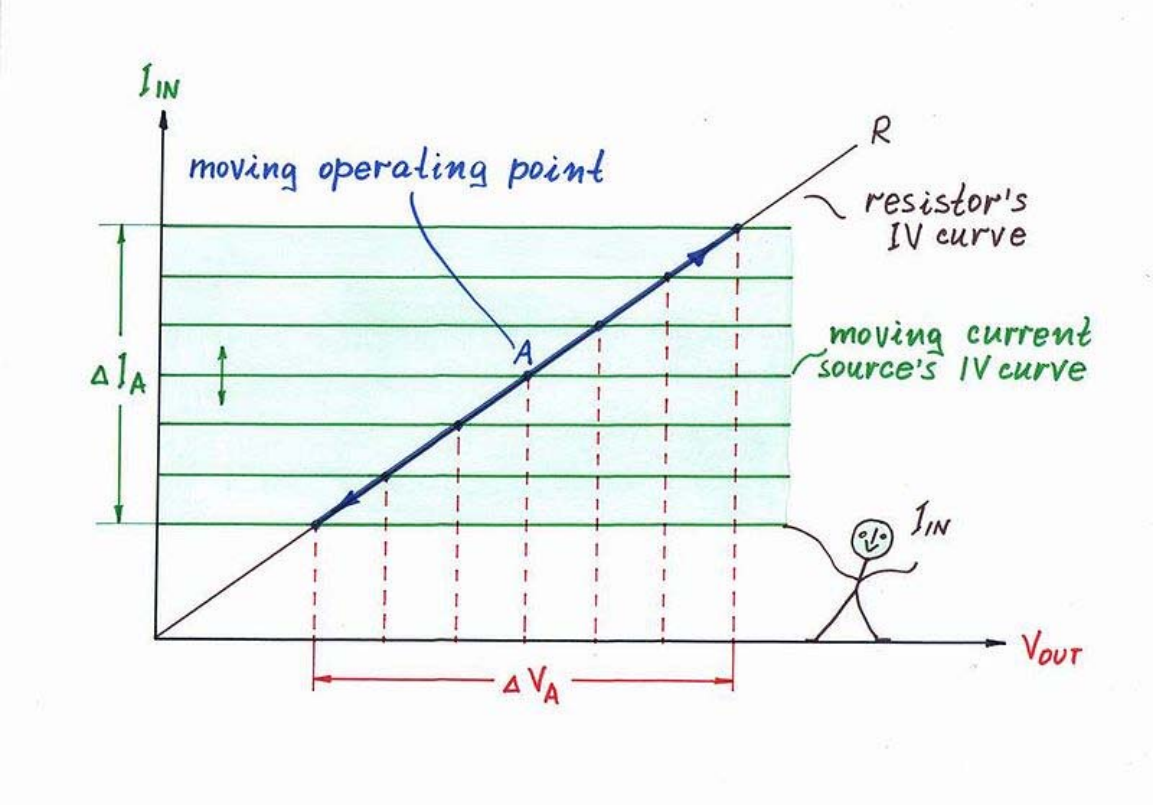


Fig. 3. A graphoanalytical presentation of the circuit operation

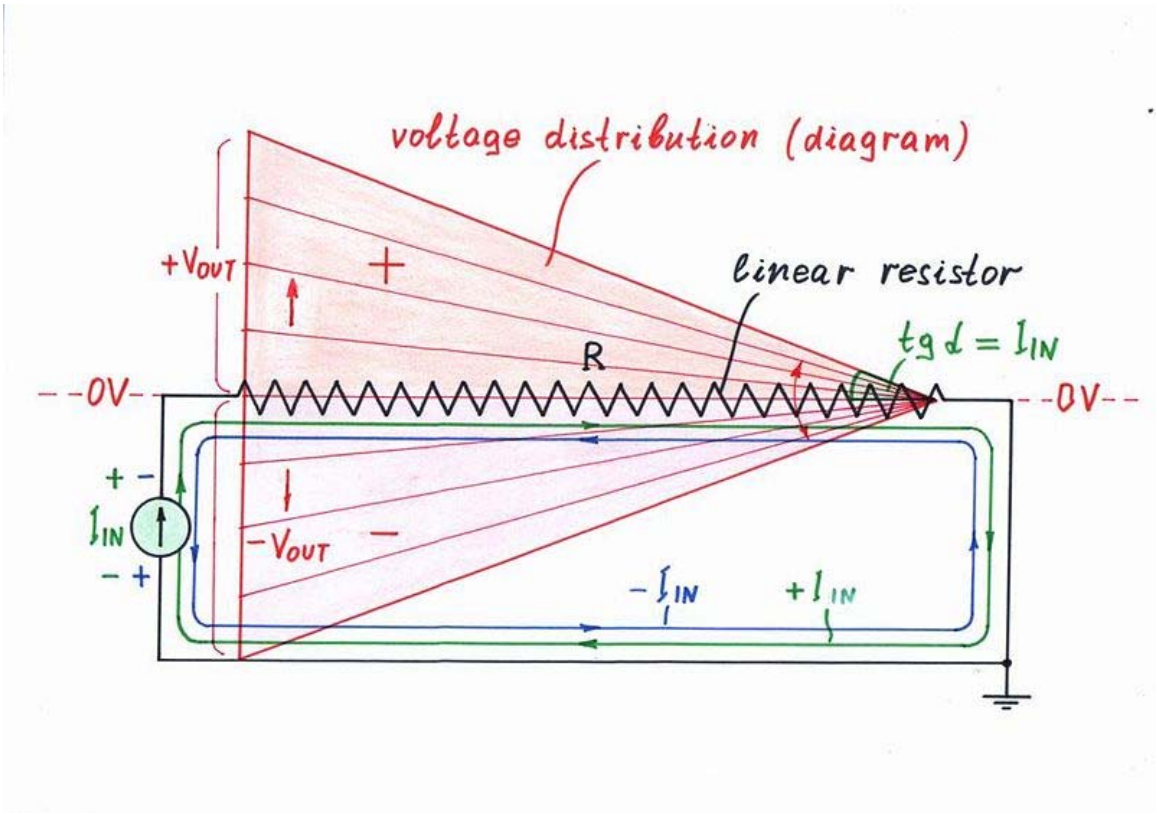


Fig. 4. Voltage distribution over the resistor R

Fig. 4 shows another attractive graphical interpretation of Ohm's law - the voltage diagram (the voltage distribution along the resistive film inside a linear resistor). When the input current varies, the local voltages along the resistive film vary decreasing gradually from left to right. In this arrangement, the angle α represents the input current I_{IN}.

Passive version applications

I-to-V converter acting as an output device

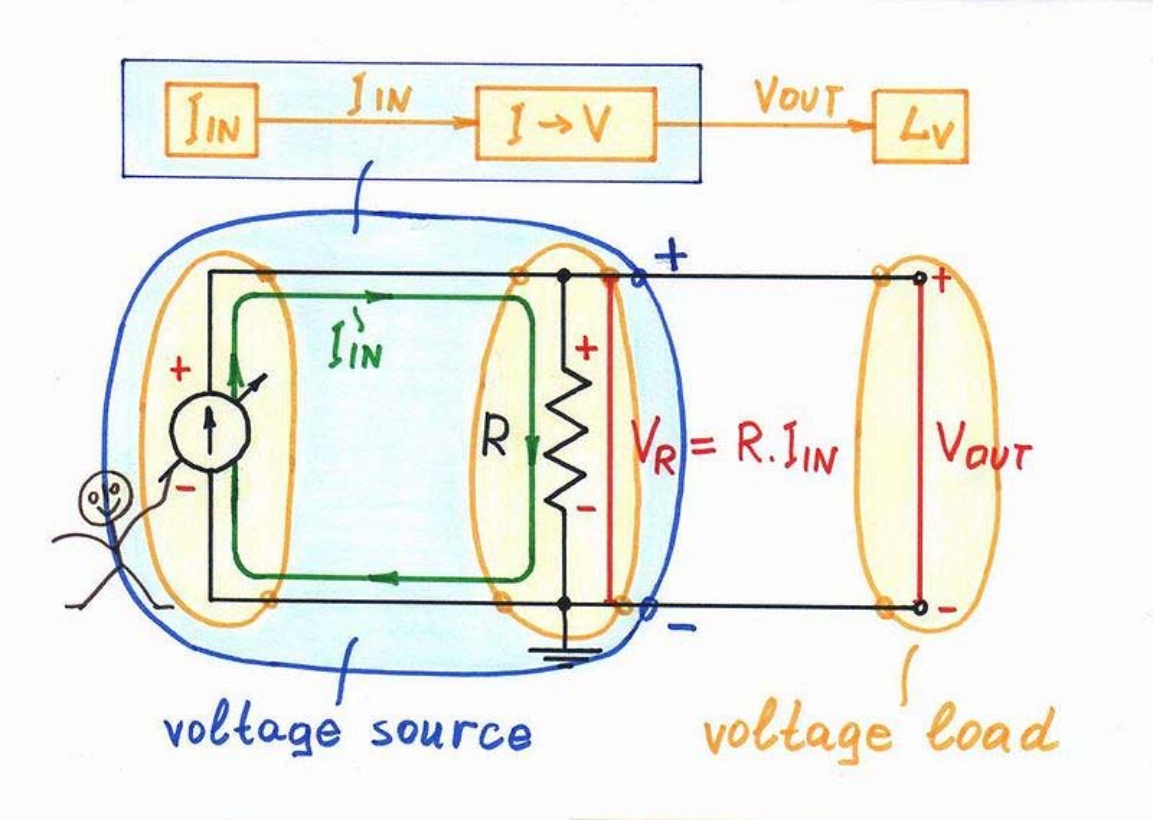


Fig. 5. Current-controlled voltage source

Current-controlled voltage source. Although there are enough constant voltage sources in nature (primary and secondary batteries), if a current source is available but there is a need of a voltage source, it may be built. For this purpose, a current-to-voltage converter has to be connected after the current source, according to the building formula below:

Voltage source = Current source + Current-to-voltage converter

The simplest implementation of this idea is shown in Fig. 5 where a resistor R is connected in parallel to the input current source I_{IN} (the Norton's idea in electricity).

If the load is ideal (that is, it has an infinite resistance), a constant voltage $V_{OUT} = R.I_{IN}$ will be generated. This voltage will affect the current, if the input current source is imperfect.

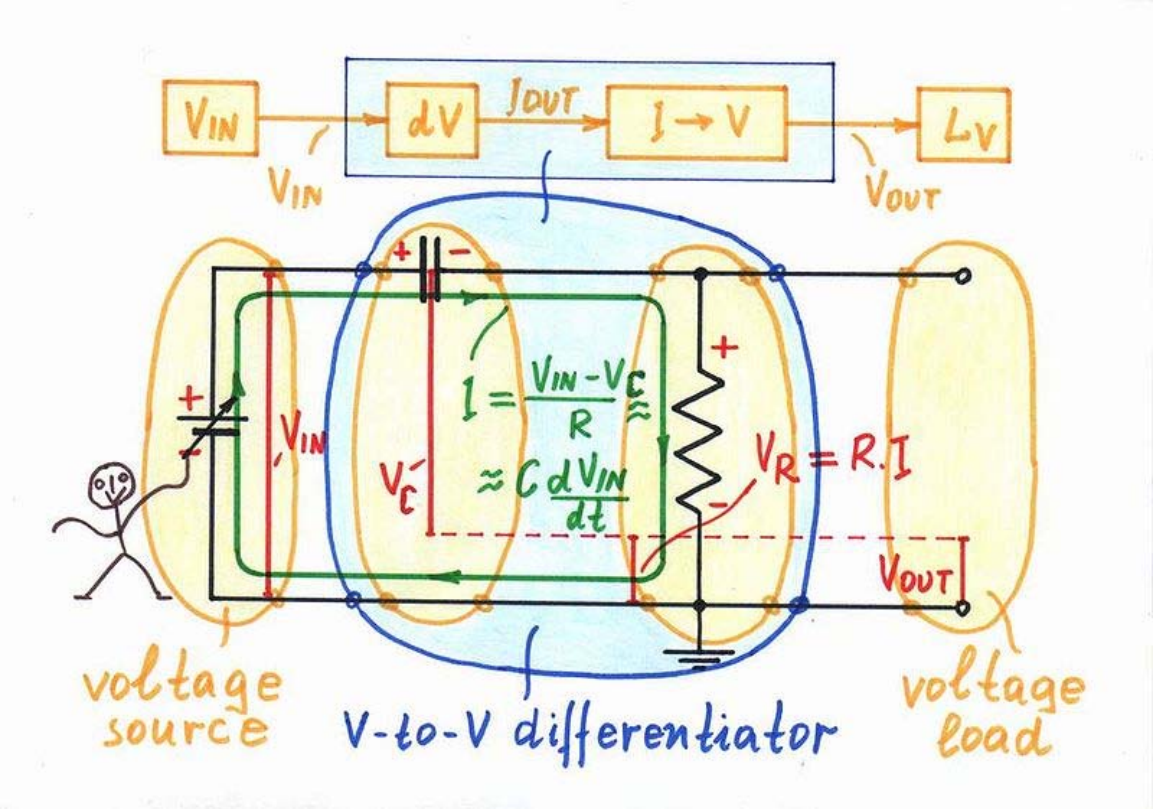


Fig. 6. V-to-V RC differentiator = V-to-I C differentiator + I-to-V converter

Compound passive converters: Similarly, in the popular passive circuits of capacitive differentiator, inductive integrator, antilogarithmic converter, etc., the resistor acts as a current-to-voltage converter:

V-to-V CR differentiator = V-to-I C differentiator + I-to-V converter

V-to-V LR integrator = V-to-I L integrator + I-to-V converter

V-to-V DR antilog converter = V-to-I D antilog converter + I-to-V

For example, a classic capacitive-resistive differentiator is built on Fig. 6 by using the simpler voltage-to-current capacitive differentiator (a bare capacitor) and a current-to-voltage converter.

In these circuits, the resistor R acting as a current-to-voltage converter introduces some voltage drop V_R, which affects the excitation voltage V_{IN}. As a result, the current decreases and an error appears.

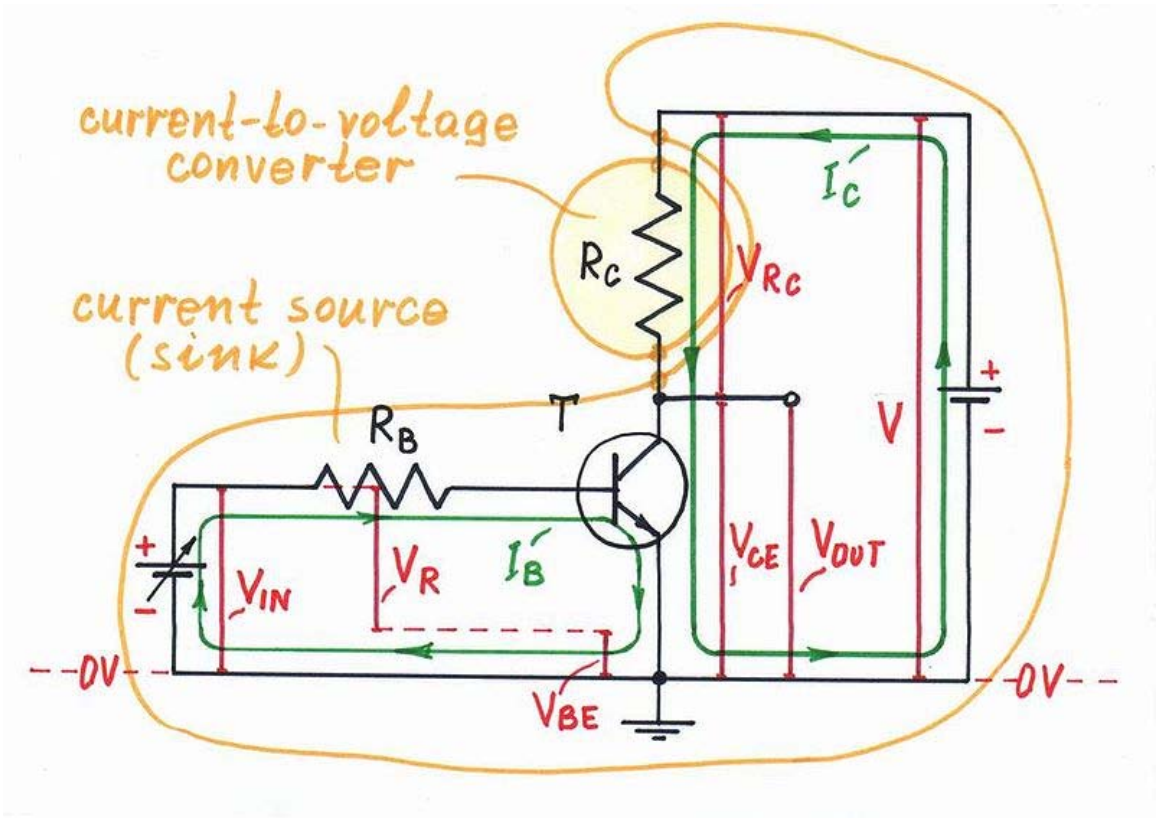


Fig. 7. The collector resistor Rc acts as a current-to-voltage converter

Transistor collector resistor. A transistor is a current-controlling device. Therefore, to obtain a voltage as an output, a collector resistor is connected in the output circuit of the transistor stage (Fig. 7). Examples of this technique are the common-emitter, common-base and differential amplifier, a transistor switch, etc.

Voltage-output transistor = Current-output transistor + I-to-V converter

The transistor's collector resistor acts as a current-to-voltage converter.

Since the voltage drop V_{Rc} is floating, usually the complementary (to the power supply) voltage drop V_{CE} is used as an output. As a result, these transistor circuits are inverting (when the input voltage rises, the output voltage drops and v.v.)

A similar technique is used to obtain a voltage in the transistor emitter. Examples of this technique are all the transistor circuits using series negative feedback.

The transistor's emitter resistor acts as a current-to-voltage converter.

I-to-V converter acting as an input device

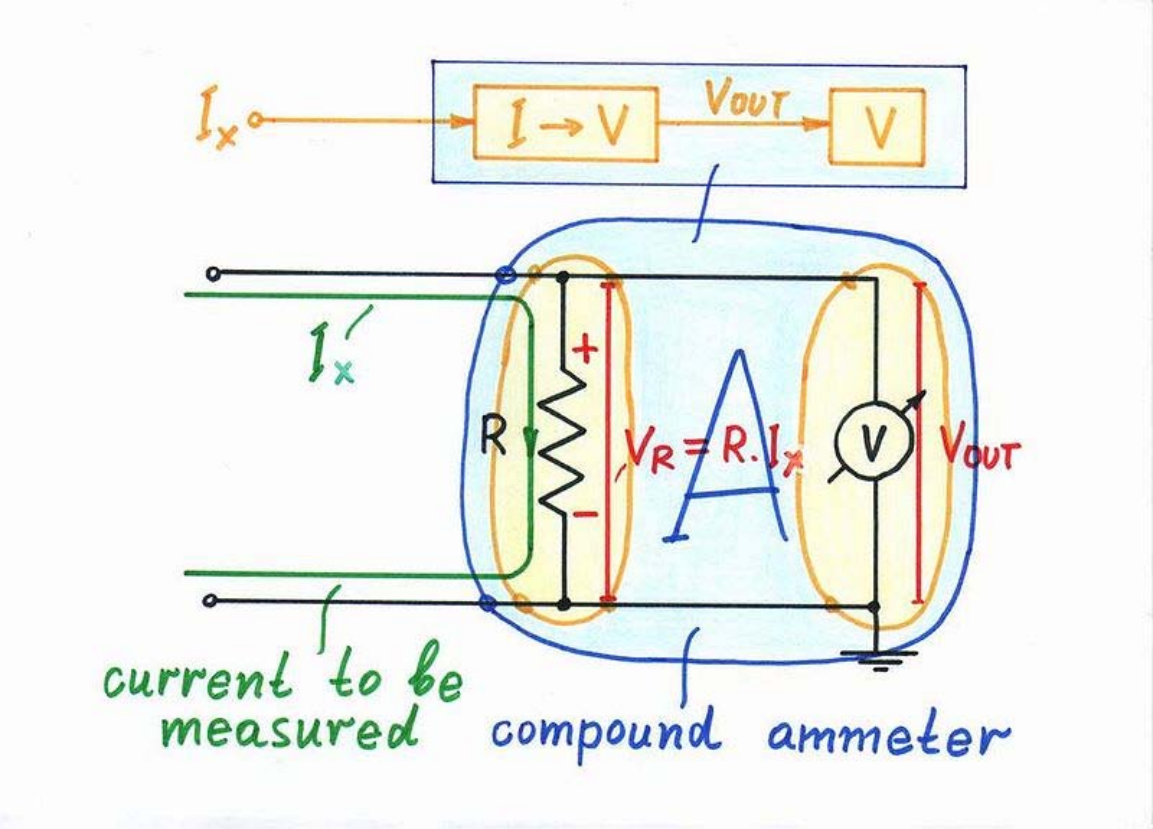


Fig. 8. Compound ammeter = I-to-V converter + voltmeter

Compound ammeter. Today's measuring instruments (DVM's, analog-to-digital converters, etc.) are mainly voltmeters. If there is a need to measure a current, a simple current-to-voltage converter (a shunt resistor) is connected before the voltmeter (Fig. 8). This ammeter is a composed device consisting of two components:

Compound ammeter = Current-to-voltage converter + voltmeter

The shunt resistor of a composed ammeter acts as a current-to-voltage converter.

Although the active version is the perfect current measurement solution, the popular multimeters use the passive version to measure big currents.

I-to-V converter as a part of negative feedback V-to-I converters

Negative feedback systems have the unique property to reverse the causality in the electronic converters connected in the feedback loop. Examples: an op-amp non-inverting amplifier is actually a reversed voltage divider, an op-amp integrator is a reversed differentiator and v.v., an op-amp logarithmic converter is a reversed antilogarithmic converter and v.v., etc.

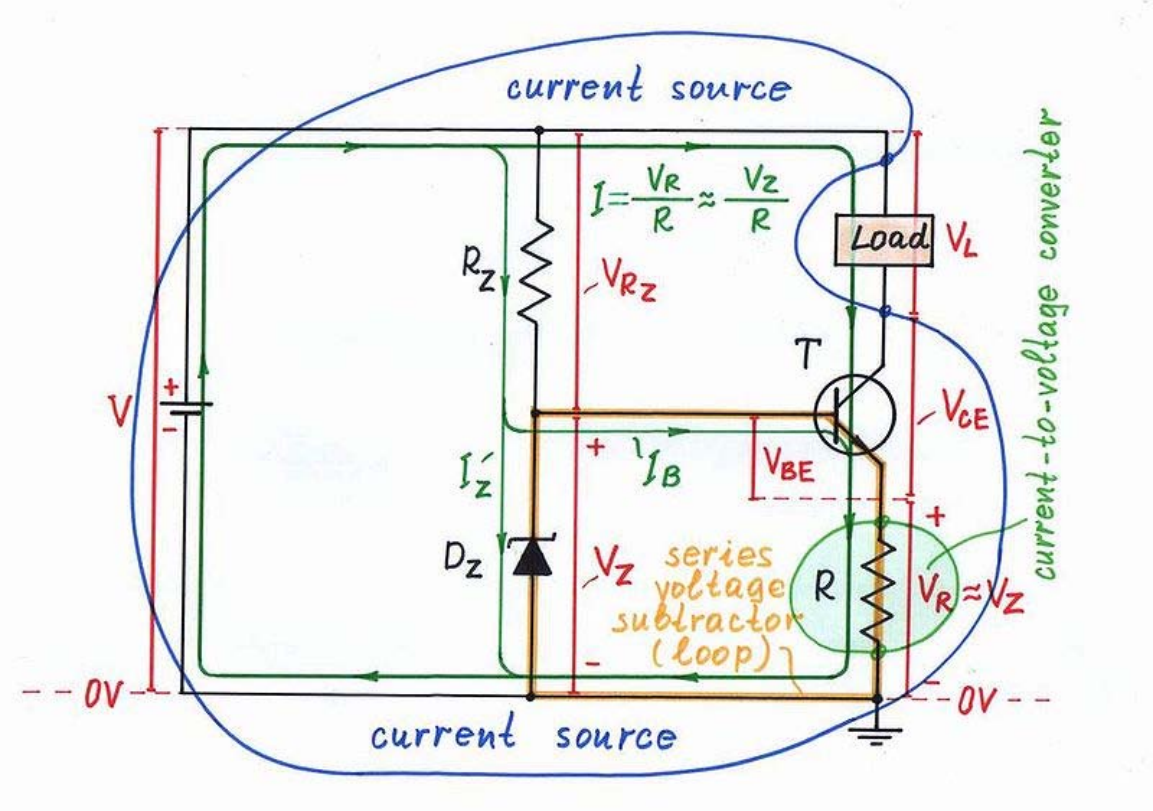


Fig. 9. A transistor current source using a current-to-voltage converter

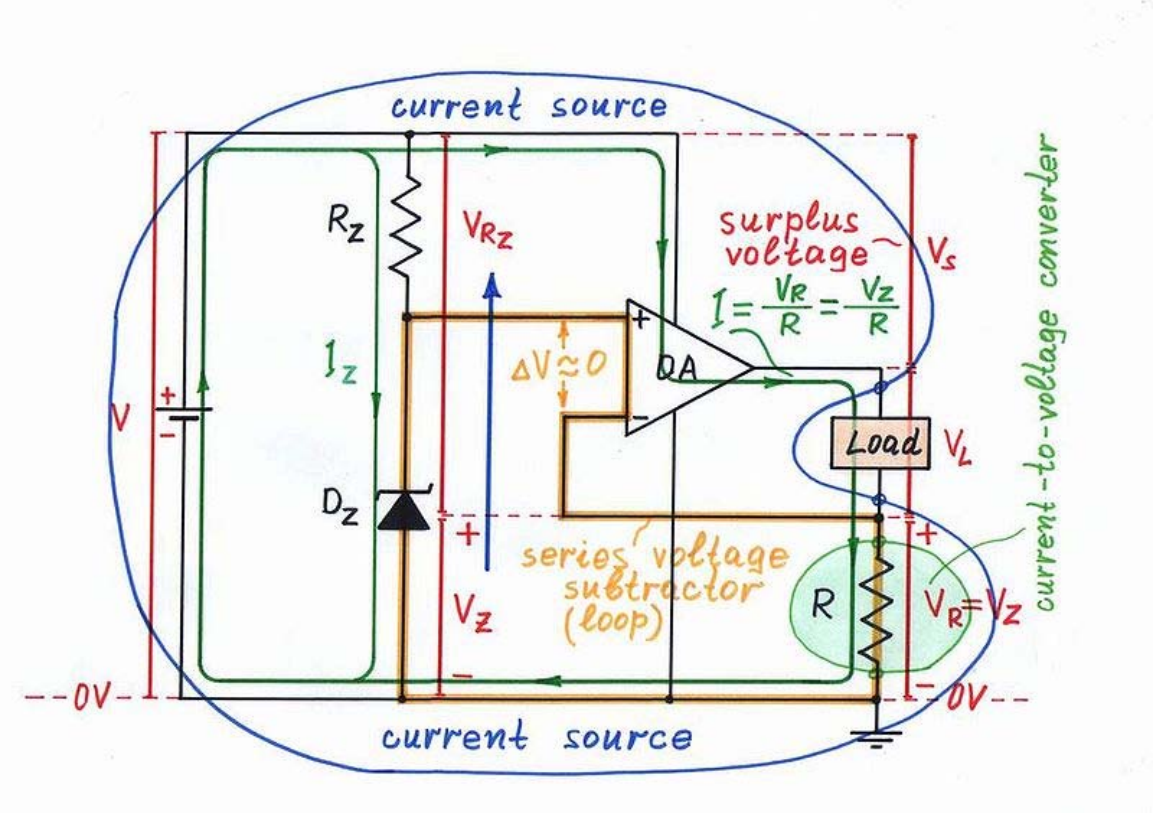


Fig. 10. An op-amp current source using a current-to-voltage converter

Similarly, an op-amp voltage-to-current converter (a voltage-controlled constant current source) built by using a negative feedback is actually a reversed current-to-voltage converter. This powerful idea is implemented on Fig. 9 (a transistor version of a current source) and on Fig. 10 (an op-amp version of a current source) where a current-to-voltage converter (the bare resistor R) is connected in the negative feedback loop. The voltage drop V_R proportional to the load current I is compared with the input voltage V_Z. For this purpose, the two voltages are connected in series and their difference $dV = V_Z - V_R$ is applied to the input part of the regulating element (the base-emitter junction of the transistor T or the differential input of the op-amp OA). As a result, the regulating element establishes the current $I = V_R/R \approx V_Z/R$ by changing its output resistance so that to zero the voltage difference dV. In this way, the output current is proportional to the input voltage; the whole circuit acts as a voltage-to-current converter.

Passive version imperfections

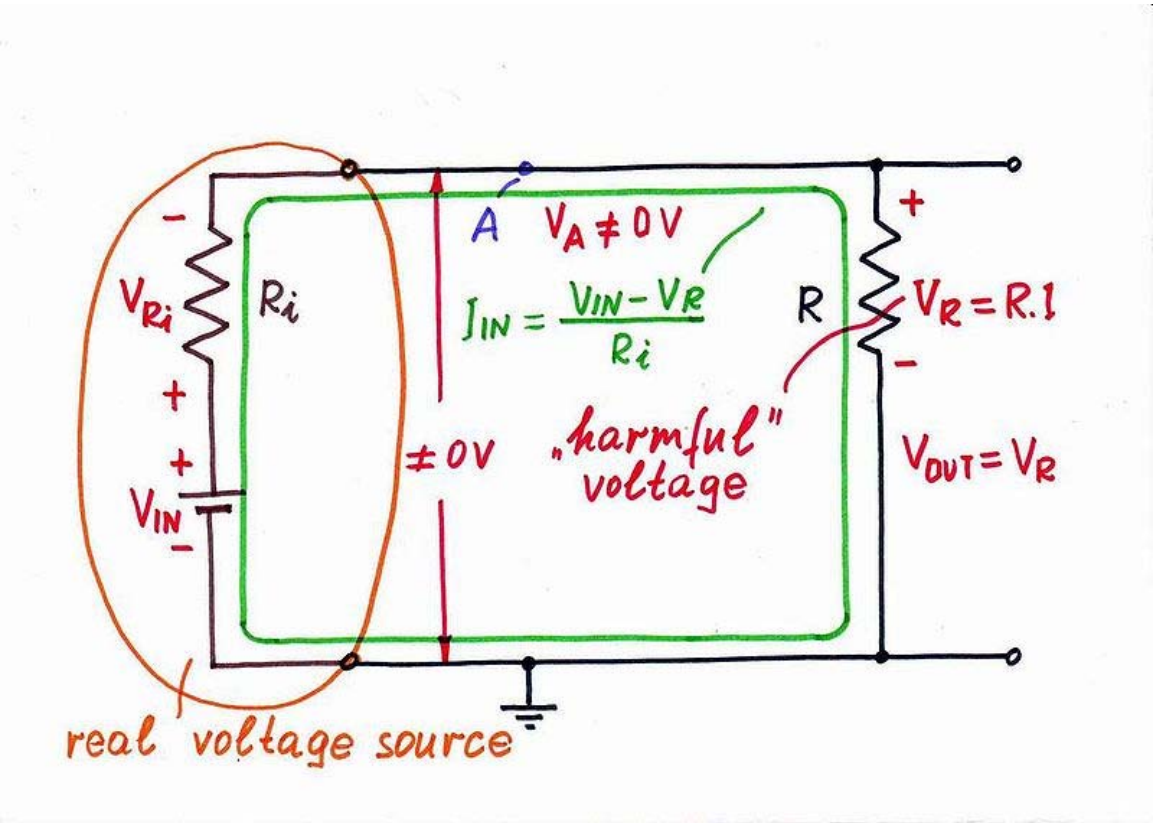


Fig. 11. The resistor R affects the current I_{in} when the input current source is imperfect

The passive current-to-voltage converter (as all the passive circuits) is imperfect because of two reasons:

Resistor R. The voltage drop V_R affects the input current I_{IN} as the resistor R consumes energy from the input source (Fig. 11). A contradiction exists in this circuit: from one side, the voltage drop V_R is useful as it serves as an output voltage; from the other side, this voltage drop is harmful as it effectively modifies the actual current-creating voltage V_{Ri}. In this arrangement, the voltage difference V_{IN} - V_R determines the current instead the voltage V_{IN} (the resistor Ri actually acts as the opposite voltage-to-current converter). As a result, the current decreases.

Load resistance. In addition, if the load has some finite resistance (instead of infinite resistance), a part of the current I_{IN} will diverts through it. As a result, both the current I_{IN} and the voltage V_{OUT} decrease. The problem is again that the load consumes energy from the passive circuit.

Improvement: Active current-to-voltage converter

The basic idea behind the active version

Non-electrical domain: *Removing disturbance by equivalent "antidisturbance"*

The active version of the current-to-voltage converter is based on a well-known technique from human routine, where we compensate the undesirable effects **caused by ourselves** using equivalent "anti-quantities". This idea is implemented by using an additional power source, which "helps" the main source by compensating the local losses caused by the **internal** undesired quantity (conversely, in the opposite active voltage-to-current converter, the additional power source compensates the losses caused by the **external** quantity). Example: if we have broken our window in winter, we turn on a heater that compensates the thermal losses; and v.v., in summer, we turn on an air-conditioner. More examples: if our car has come into collision with other car, the insurance company compensates the damages we caused to the other car. If we cause trouble to others, we apologize. If we spend money from an account, we deposit funds. In all these cases, we prepared "standby" resources to use if there is a need to compensate internal losses.

Electrical domain: *Removing voltage by equivalent "antivoltage"*

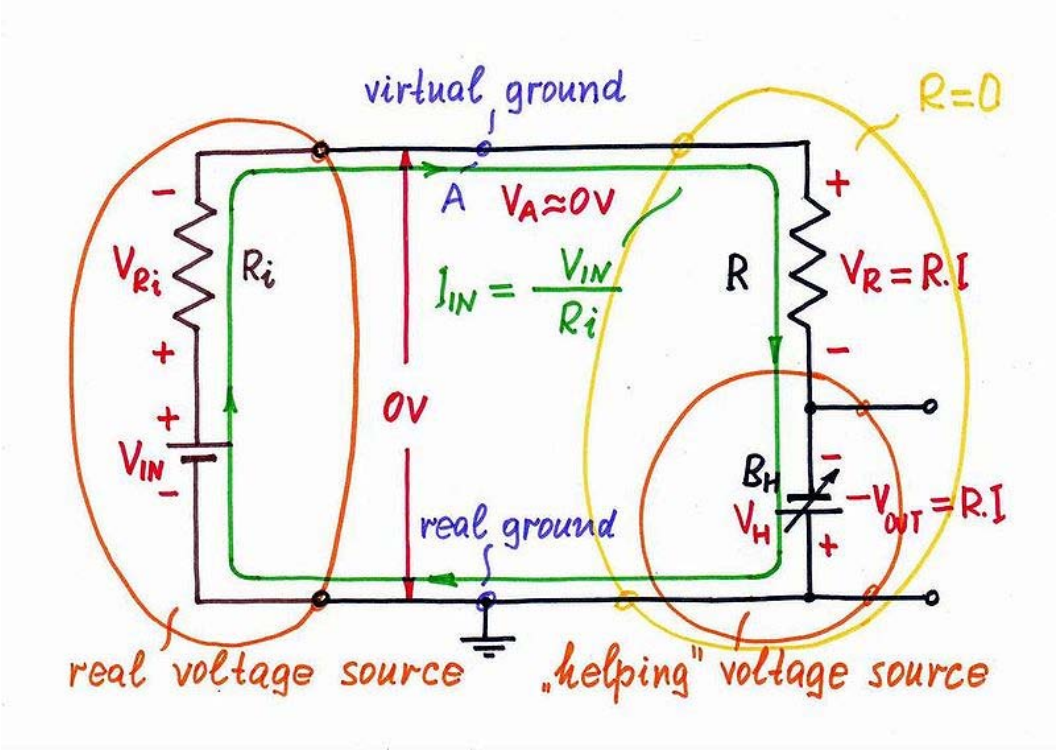


Fig. 12. Active current-to-voltage converter

Electrical implementation. To show how this powerful basic idea is applied to improve the passive current-to-voltage converter, first, an equivalent electrical circuit is used (Fig. 12). In this *active current-to-voltage converter*, the voltage drop V_R across the **internal** resistor R is compensated by adding the same voltage $V_H = V_R$ to the input voltage V_{IN} . For this purpose, an additional following voltage source B_H is connected in series with the resistor. It "helps" the input voltage source; as a result, the undesired voltage V_R and the resistance R disappear (the point A becomes a virtual ground).

Active I-to-V converter = passive I-to-V converter + "helping" voltage source

Where to take an output from? The magnitude of the compensating quantity is frequently used to measure **indirectly** the initial quantity (an example - weighing by using scales). This idea is applied in the circuit of active current-to-voltage converter by connecting the load to the compensating voltage source B_H instead to the resistor. There are two advantages of this arrangement: first, the load is connected to the common ground; second, it consumes energy from the additional source instead from the input source. Therefore, it might possess small resistance.

Op-amp implementation

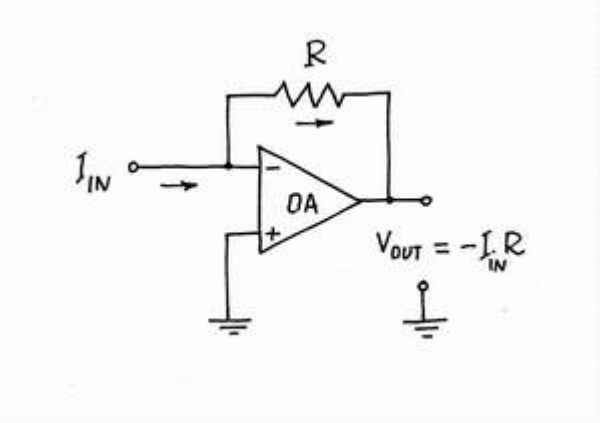


Fig. 13. Op-amp current-to-voltage converter

The basic idea above is implemented in the op-amp current-to-voltage converter (Fig. 13, 14) . In this circuit, the output of the operational amplifier is connected in series with the input voltage source; the op-amp's inverting input is connected to point A. As a result, the op-amp's output voltage and the input voltage are summed.

From other viewpoint, the output of the operational amplifier is connected in series with the resistor R in the place of the compensating voltage source B_H from Fig. 12. As a result, the op-amp's output voltage and the voltage drop V_R are subtracted; the potential of the point A represents the result of this subtraction (it behaves as a virtual ground).

Op-amp I-to-V converter = passive I-to-V converter + "helping" op-amp

Op-amp circuit operation

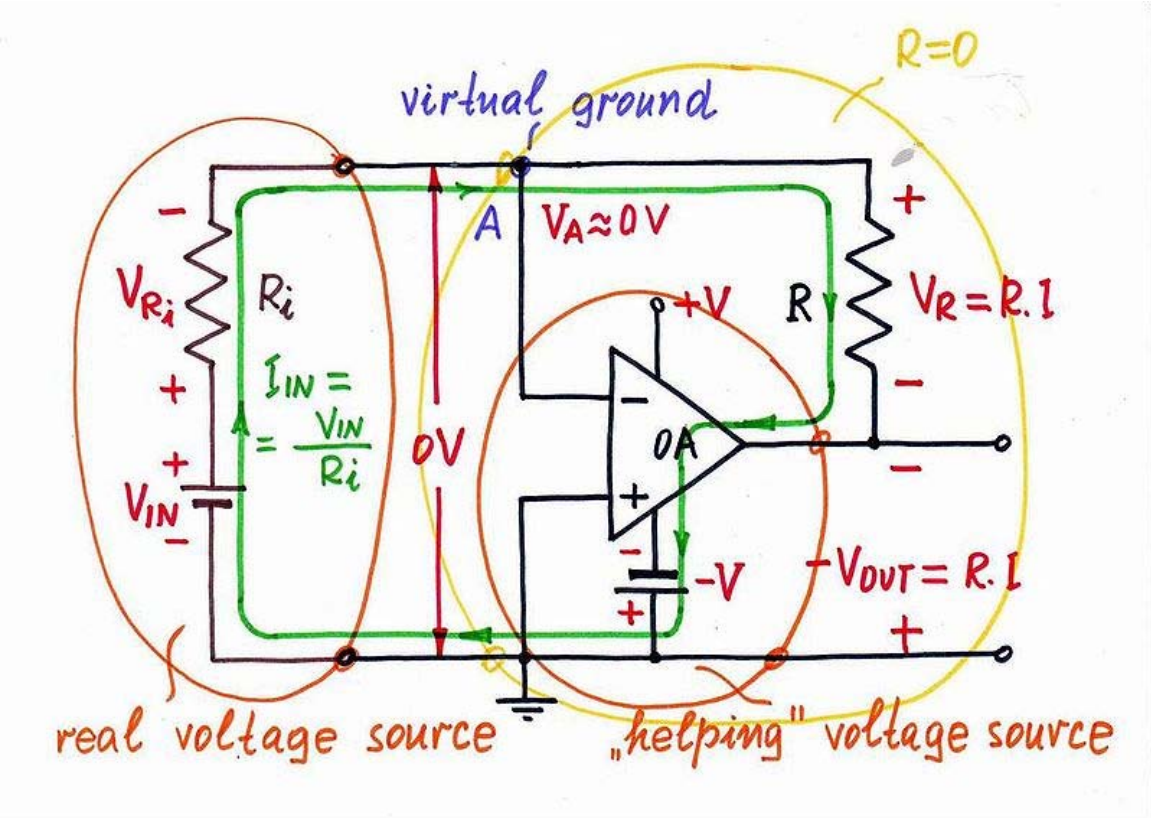


Fig. 14. Op-amp current-to-voltage converter ($+V_{IN}$)

Zero input voltage results in no voltage drops or currents in the circuit.

Positive input voltage. If the input voltage V_{IN} increases above the ground, an input current I_{IN} begins flowing through the resistor R. As a result, a voltage drop V_R appears across the resistor and the point A begins raising its potential (the input source "pulls" the point A up toward the positive voltage V_{IN}). Only, the op-amp "observes" that and immediately reacts: it decreases its output voltage under the ground sucking the current. Figuratively speaking, the op-amp "pulls" the point A down toward the negative voltage -V until it manages to zero its potential (to establish a virtual ground). It does this work by connecting a portion of the voltage produced by the negative power supply -V in series with the input voltage V_{IN}. The two voltage sources are connected in series, in the same direction (traversing the loop clockwise, the signs are - V_{IN} +, - V_{OA} +) so that their voltages are added. However, regarding to the ground, they have opposite polarities.

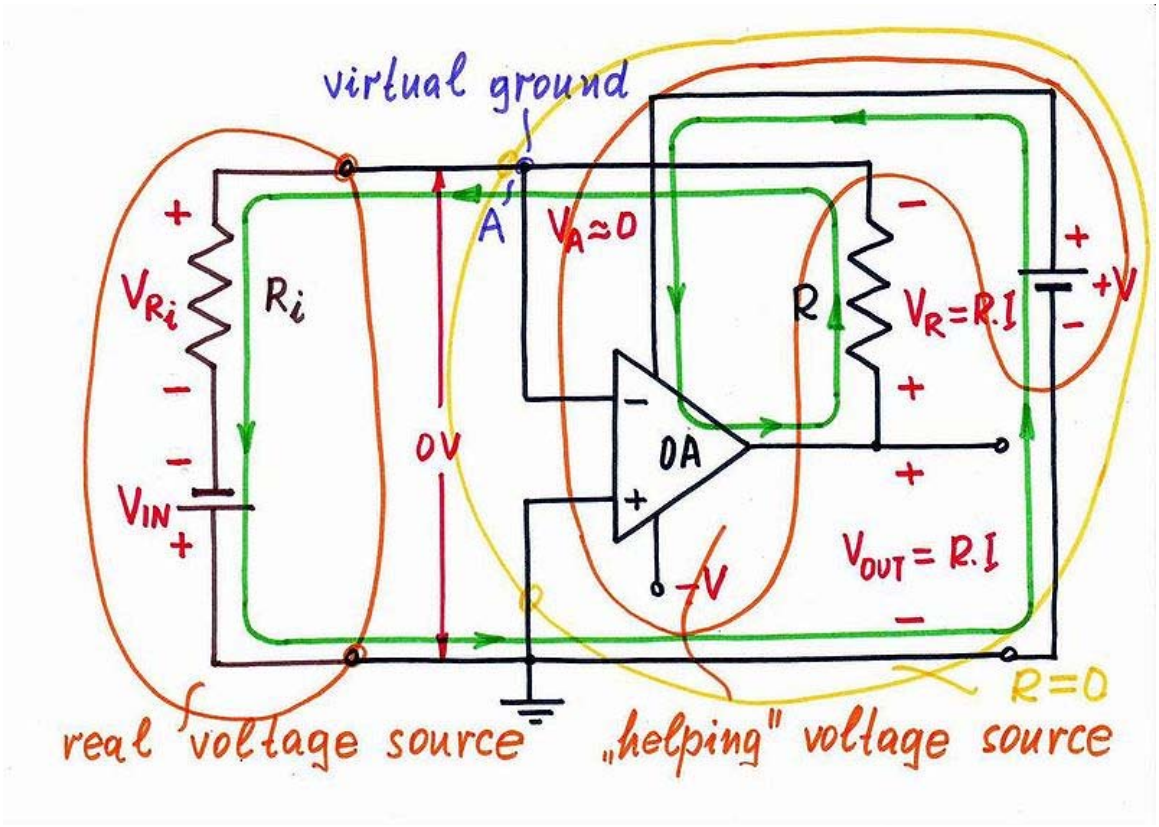


Fig. 15. Op-amp current-to-voltage converter ($-V_{IN}$)

Negative input voltage. If the input voltage V_{IN} decreases under the ground, the input current flows through the resistor R in opposite direction (Fig. 15). As a result, a voltage drop V_R appears across the resistor again and the point A begins dropping its potential (now, the input source "pulls" the point A down toward the negative voltage $-V_{IN}$). The op-amp "observes"' that and immediately reacts: it increases its output voltage above the ground "pushing out" the current. Now, the op-amp "pulls" the point A up toward the positive voltage +V until it manages to zero again the potential V_A (the virtual ground). For this purpose, the op-amp puts a portion of the voltage produced by the positive power supply +V in series with the input voltage V_{IN}. The two voltage sources are connected again, in the same direction (traversing the loop clockwise, + V_{IN} -, + V_{OA} -) so that their voltages are added. However, regarding to the ground, they have opposite polarities as above.

Conclusion. *In the circuit of an op-amp current-to-voltage converter, the op-amp adds as much voltage to the voltage of the input source as it loses across the resistor. The op-amp compensates the local losses caused by this* ***internal resistor*** *(conversely, in the opposite op-amp voltage-to-current converter, the op-amp compensates the losses caused by the* ***external load****).*

I-to-V converters versus transimpedance amplifiers

The active current-to-voltage converter is an amplifier with current input and voltage output. The gain of this amplifier is represented by the resistance R ($K = V_{OUT}/I_{IN} = R$); it is expressed in units of ohms. That is why this circuit is named *transresistance amplifier* or more generally, *transimpedance amplifier* . Both terms are used to designate the circuit considered.

Active version applications

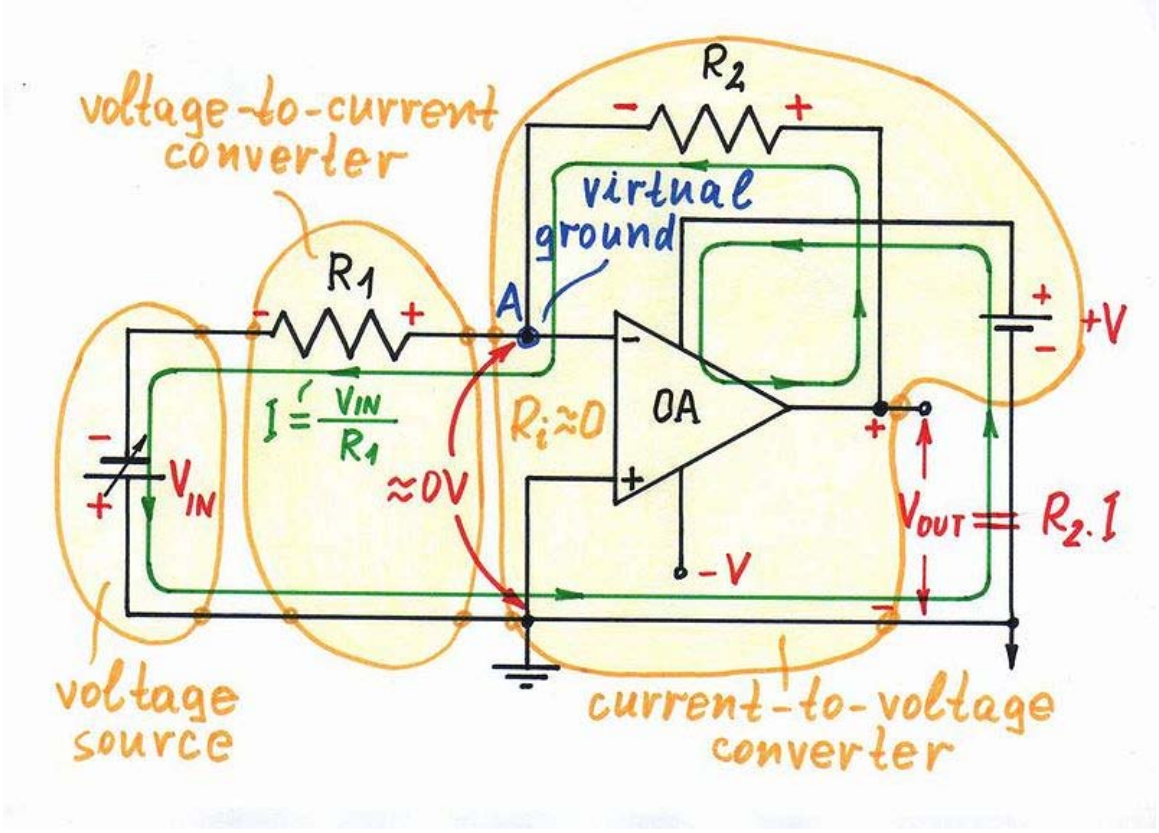


Fig. 16. *Op-amp inverting amplifier = V-to-I converter + op-amp I-to-V converter*

Transimpedance amplifiers are commonly used in receivers for optical communications. The current generated by a photodetector generates photo voltage, but in a nonlinear fashion. Therefore the amplifier has to prevent any large voltage by its low input impedance and generate either a 50 Ohm signal (considered low impedance by many) to drive a coaxial cable or a voltage signal for further amplification. But note that the most linear amplification is current amplification by a bipolar transistor, so you may want to amplify before the impedance conversion.

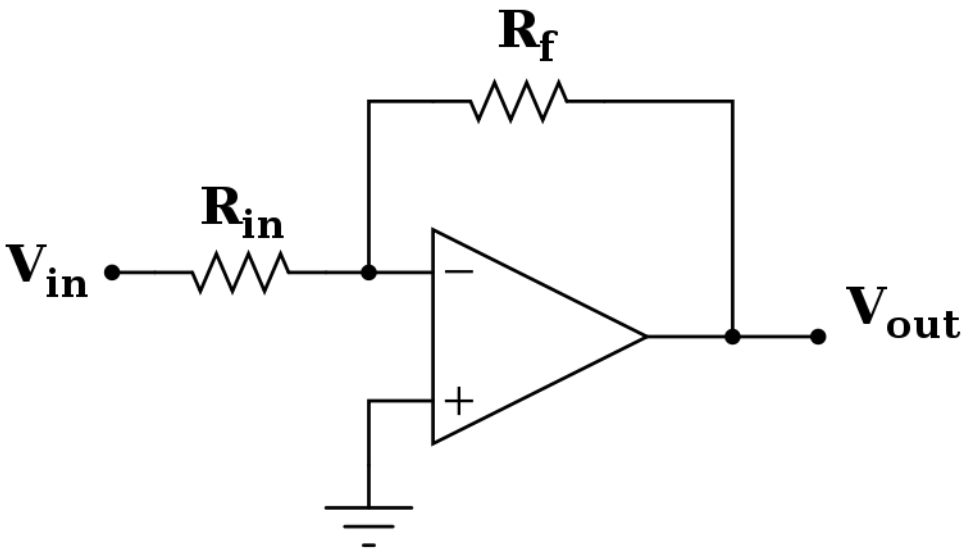


Fig. 17. Inverting amplifier configuration of an op-amp becomes a transimpedance amplifier when Rin is 0 ohms

The circuit considered is also used as a main part of more complex op-amp inverting circuits with (parallel) negative feedback: inverting amplifier (Fig. 16, 17), CR differentiator, LR integrator, inverting voltage summer etc. Here are the building formulas of these circuits:

Op-amp inverting amplifier = V-to-I converter + op-amp I-to-V converter

Op-amp V-to-V CR differentiator = V-to-I C differentiator + op-amp I-to-V converter

Op-amp V-to-V LR integrator = V-to-I L integrator + op-amp I-to-V converter

Op-amp V-to-V DR antilog converter = V-to-I D antilog converter + op-amp I-to-V converter

Active version imperfections (power considerations)

Although the active current-to-voltage converter is a perfect circuit, the popular multimeters do not work that way. To measure a current, they use the imperfect passive current-to-voltage converter instead of the almost ideal op-amp current-to-voltage converter. The reason for applying such an old-fashioned approach to current measurements is that all the input current I_{IN} flows through the "helping" voltage source B_H in the active version (Fig. 3). Therefore, the source has to be able to endure such a current. Accordingly, in the practical op-amp circuit (Fig. 4), both the power source and the op-amp have to endure the input current measured. For example, if they try to measure a current of 10A (a normal maximum current range in all purpose DVMs), they must use a car battery as a power supply and a power "op-amp" that can dissipate 100W.

Chapter- 6

Equivalent Impedance Transforms

An **equivalent impedance** is an equivalent circuit of an electrical network of impedance elements which presents the same impedance between all pairs of terminals as did the given network.

There are a number of very well known and often used equivalent circuits in linear network analysis. These include resistors in series, resistors in parallel and the extension to series and parallel circuits for capacitors, inductors and general impedances. Also well known are the Norton and Thévenin equivalent current generator and voltage generator circuits respectively, as is the Y-Δ transform. None of these are discussed in detail here; the individual linked articles should be consulted.

The number of equivalent circuits that a linear network can be transformed into is unbounded. Even in the most trivial cases this can be seen to be true, for instance, by asking how many different combinations of resistors in parallel are equivalent to a given combined resistor. Wilhelm Cauer found a transformation that could generate all possible equivalents of a given rational, passive, linear one-port, or in other words, any given two-terminal impedance. Transformations of 4-terminal, especially 2-port, networks are also commonly found and transformations of yet more complex networks are possible.

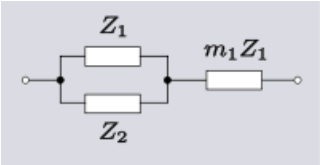
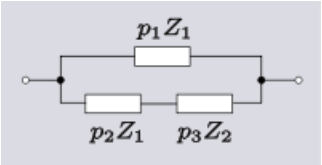
The vast scale of the topic of equivalent circuits is underscored in a story told by Sidney Darlington. According to Darlington, a large number of equivalent circuits were found by Ronald Foster, following his and George Campbell's 1920 paper on non-dissipative four-ports. In the course of this work they looked at the ways four ports could be interconnected with ideal transformers. They found a number of combinations which might have practical applications and asked the AT&T patent department to have them patented. The patent department replied that it was pointless just patenting some of the circuits if a competitor could use an equivalent circuit to get around the patent; they should patent all of them or not bother. Foster therefore set to work calculating every last one of them. He arrived at an enormous total of 83,539 equivalents. This was too many to patent, so instead the information was released into the public domain in order to prevent any of AT&T's competitors from patenting them in the future.

2-terminal, 2-element-kind networks

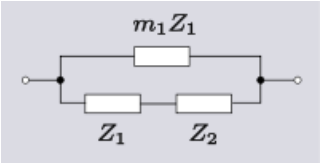
A single impedance has two terminals to connect to the outside world, hence can be described as a 2-terminal, or a one-port, network. Despite the simple description, there is no limit to the number of meshes, and hence complexity and number of elements, that the impedance network may have. 2-element-kind networks are common in circuit design; filters, for instance, are often LC-kind networks and printed circuit designers favour RC-kind networks because inductors are less easy to manufacture. Transformations are simpler and easier to find than for 3-element-kind networks. One-element-kind networks can be thought of as a special case of two-element-kind. It is possible to use the transformations in this section on a certain few 3-element-kind networks by substituting a network of elements for element Z_n. However, this is limited to a maximum of two impedances being substituted; the remainder will not be a free choice. All the transformation equations given in this section are due to Otto Zobel.

3-element networks

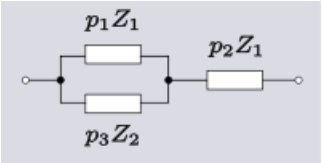
One-element networks are trivial and two-element, two-terminal networks are either two elements in series or two elements in parallel, also trivial. The smallest number of elements that is non-trivial is three, and there are two 2-element-kind non-trivial transformations possible, one being both the reverse transformation and the topological dual, of the other.

Description	Network	Transform equations	Transformed network
Transform 1.1 Transform 1.2 is the reverse of this transform.		$p_1 = 1 + m_1$, $p_2 = m_1(1 + m_1)$, $p_3 = (1 + m_1)^2$.	

Transform 1.2 The reverse transform, and topological dual, of Transform 1.1.



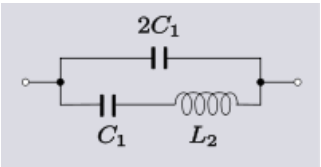
$$p_1 = \frac{{m_1}^2}{1+m_1} \ ,$$



$$p_2 = \frac{m_1}{1+m_1} \ ,$$

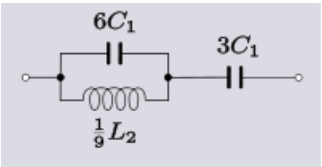
$$p_3 = \left(\frac{m_1}{1+m_1}\right)^2 \ .$$

Example 1. An example of Transform 1.2. The reduced size of the inductor has practical advantages.



$$m_1 = 0.5 \ ,$$

$$p_1 = \tfrac{1}{6} \ , \ p_2 = \tfrac{1}{3} \ ,$$



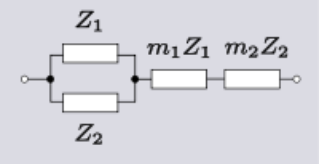
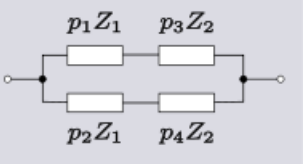
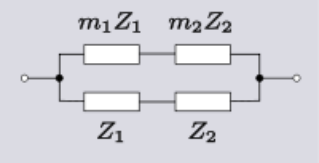
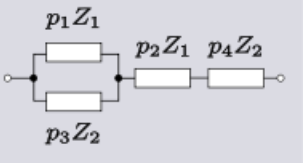
$$p_3 = \tfrac{1}{9} \ .$$

4-element networks

There are four non-trivial 4-element transformations for 2-element-kind networks. Two of these are the reverse transformations of the other two and two are the dual of a different two. Further transformations are possible in the special case of Z_2 being made the same element kind as Z_1, that is, when the network is reduced to one-element-kind. The number of possible networks continues to grow as the number of elements is increased. For all entries in the following table it is defined:

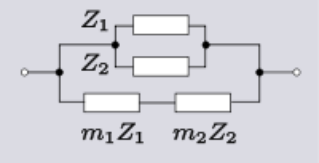
$$\begin{aligned} q_1 &:= 1 + m_1 + m_2 \ , \\ q_2 &:= \sqrt{{q_1}^2 - 4m_1m_2} \ , \\ q_3 &:= \frac{(1+m_1)(1+m_2)}{(m_1-m_2)^2} \ , \\ q_4 &:= \frac{q_2 - q_1 + 2m_2}{2q_2} \ , \end{aligned}$$

$$q_5 := \frac{q_2 + q_1 - 2m_2}{2q_2} .$$

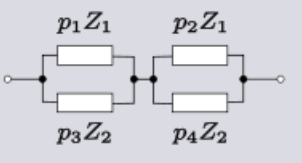
Description	Network	Transform equations	Transformed network
Transform 2.1 Transform 2.2 is the reverse of this transform. Transform 2.3 is the topological dual of this transform.		$p_1 = \frac{q_1 + q_2}{2q_5}$, $p_2 = \frac{q_1 - q_2}{2q_4}$, $p_3 = \frac{m_2}{q_5}$, $p_4 = \frac{m_2}{q_4}$.	
Transform 2.2 Transform 2.1 is the reverse of this transform. Transform 2.4 is the topological dual of this transform.		$p_1 = \frac{1}{q_3(1 + m_2)}$, $p_2 = \frac{m_1}{1 + m_1}$, $p_3 = \frac{1}{q_3(1 + m_1)}$,	

$$p_4 = \frac{m_2}{1 + m_2} \ .$$

Transform 2.3
Transform 2.4 is the reverse of this transform. Transform 2.1 is the topological dual of this transform.



$$p_1 = \frac{q_4(q_1 + q_2)}{2m_2} \ ,$$

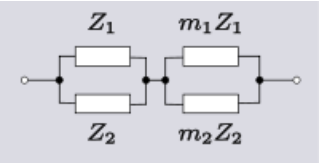


$$p_2 = \frac{q_5(q_1 - q_2)}{2m_2} \ ,$$

$$p_3 = q_4 \ ,$$

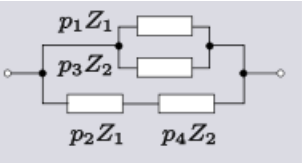
$$p_4 = q_5 \ .$$

Transform 2.4
Transform 2.3 is the reverse of this transform. Transform 2.2 is the topological dual of this transform.



$$p_1 = 1 + m_1 \ ,$$

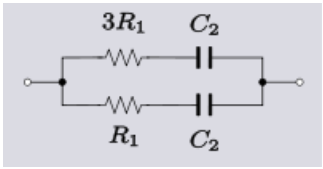
$$p_2 = m_1 q_3(1 + m_1) \ ,$$



$$p_3 = 1 + m_2 \ ,$$

$$p_4 = m_1 q_3(1 + m_2) \ .$$

Example 2. An example of Transform 2.2.

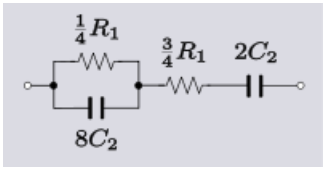


$$m_1 = 3 \ ,$$

$$m_2 = 1 \ , \ q_3 = 2 \ ,$$

$$p_1 = \tfrac{1}{4} \ , \ p_2 = \tfrac{3}{4} \ ,$$

$$p_3 = \tfrac{1}{8} \ , \ p_4 = \tfrac{1}{2} \ .$$



2-terminal, n*-element, 3-element-kind networks*

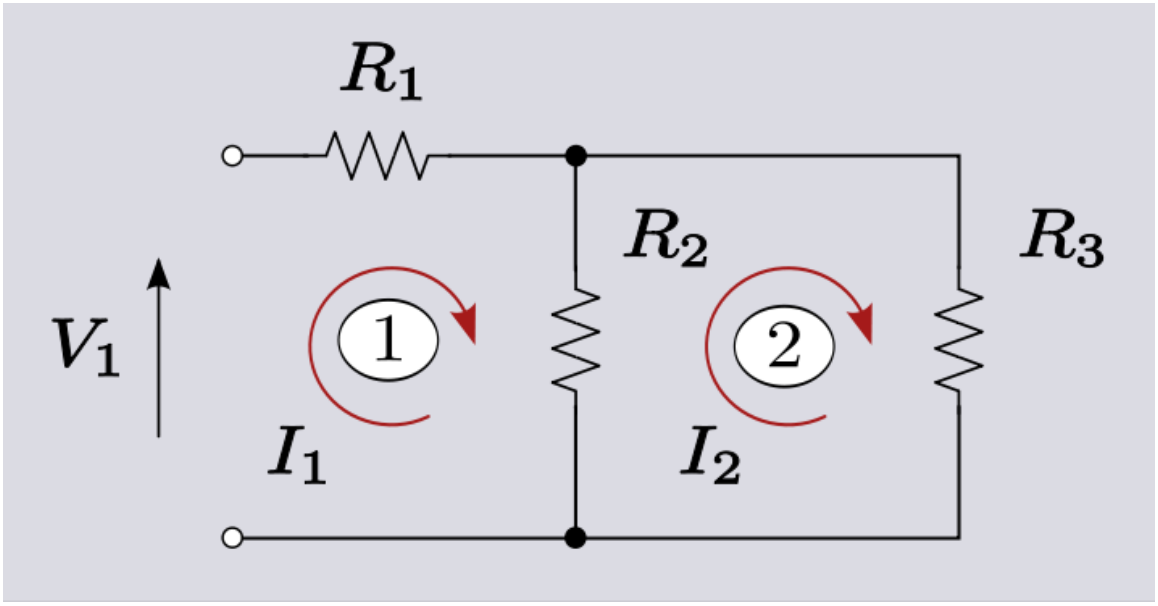


Fig. 1. Simple example of a network of impedances using resistors only for clarity. However, analysis of networks with other impedance elements proceed by the same principles. Two meshes are shown, with numbers in circles. The sum of impedances around each mesh, p, will form the diagonal of the entries of the matrix, Z_{pp}. The impedance of branches shared by two meshes, p and q, will form the entries $-Z_{pq}$. Z_{pq}, $p \neq q$, will always have a minus sign provided that the convention of loop currents are defined in the same (conventionally counter-clockwise) direction and the mesh contains no ideal transformers or mutual inductors.

Simple networks with just a few elements can be dealt with by formulating the network equations "by hand" with the application of simple network theorems such as Kirchhoff's laws. Equivalence is proved between two networks by directly comparing the two sets of equations and equating coefficients. For large networks more powerful techniques are

required. A common approach is to start by expressing the network of impedances as a matrix. This approach is only good for rational networks. Any network that includes distributed elements, such as a transmission line, cannot be represented by a finite matrix. Generally, an *n*-mesh network requires an *n*x*n* matrix to represent it. For instance the matrix for a 3-mesh network might look like;

$$[\mathbf{Z}] = \begin{bmatrix} Z_{11} & Z_{12} & Z_{13} \\ Z_{21} & Z_{22} & Z_{23} \\ Z_{31} & Z_{32} & Z_{33} \end{bmatrix}$$

The entries of the matrix are chosen so that the matrix forms a system of linear equations in the mesh voltages and currents (as defined for mesh analysis);

$$[\mathbf{V}] = [\mathbf{Z}][\mathbf{I}]$$

The example diagram in Figure 1, for instance, can be represented as an impedance matrix by;

$$[\mathbf{Z}] = \begin{bmatrix} R_1 + R_2 & -R_2 \\ -R_2 & R_2 + R_3 \end{bmatrix}$$

and the associated system of linear equations are,

$$\begin{bmatrix} V_1 \\ 0 \end{bmatrix} = \begin{bmatrix} R_1 + R_2 & -R_2 \\ -R_2 & R_2 + R_3 \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \end{bmatrix}$$

In the most general case, each branch, Z_p, of the network may be made up of three elements so that,

$$Z_\mathrm{p} = sL_\mathrm{p} + R_\mathrm{p} + \frac{1}{sC_\mathrm{p}}$$

where *L*, *R* and *C* represent inductance, resistance, and capacitance respectively and *s* is the complex frequency operator $s=\sigma+i\omega$.

This is the conventional way of representing a general impedance but for the purposes here it is mathematically more convenient to deal with elastance, *D*, the inverse of capacitance, *C*. In those terms the general branch impedance can be represented by,

$$sZ_\mathrm{p} = s^2 L_\mathrm{p} + sR_\mathrm{p} + D_\mathrm{p}$$

Likewise, each entry of the impedance matrix can consist of the sum of three elements. Consequently, the matrix can be decomposed into three *n*x*n* matrices, one for each of the three element kinds;

$$s[\mathbf{Z}] = s^2[\mathbf{L}] + s[\mathbf{R}] + [\mathbf{D}]$$

It is desired that the matrix [**Z**] represent an impedance, $Z(s)$. For this purpose, the loop of one of the meshes is cut and $Z(s)$ is the impedance measured between the points so cut. It is conventional to assume the external connection port is in mesh 1, and is therefore connected across matrix entry Z_{11}, although it would be perfectly possible to formulate this with connections to any desired nodes. In the following discussion $Z(s)$ taken across Z_{11} is assumed. $Z(s)$ may be calculated from [**Z**] by;

$$Z(s) = \frac{|\mathbf{Z}|}{z_{11}}$$

where z_{11} is the complement of Z_{11} and |**Z**| is the determinant of [**Z**].

For the example network above;

$$|\mathbf{Z}| = (R_1 + R_2)(R_2 + R_3) - {R_2}^2 = R_1R_2 + R_1R_3 + R_2R_3 \ ,$$
$$z_{11} = Z_{22} = R_2 + R_3 \ , \text{and,}$$
$$Z(s) = R_1 + \frac{R_2R_3}{R_2 + R_3} \ .$$

This result is easily verified to be correct by the more direct method of resistors in series and parallel. However, such methods rapidly become tedious and cumbersome with the growth of the size and complexity of the network under analysis.

The entries of [**R**], [**L**] and [**D**] cannot be set arbitrarily. For [**Z**] to be able to realise the impedance $Z(s)$ then [**R**],[**L**] and [**D**] must all be positive-definite matrices. Even then, the realisation of $Z(s)$ will, in general, contain ideal transformers within the network. Finding only those transforms that do not require mutual inductances or ideal transformers is a more difficult task. Similarly, if starting from the "other end" and specifying an expression for $Z(s)$, this again cannot be done arbitrarily. To be realisable as a rational impedance, $Z(s)$ must be positive-real. The positive-real (PR) condition is both necessary and sufficient but there may be practical reasons for rejecting some topologies.

A general impedance transform for finding equivalent rational one-ports from a given instance of [**Z**] is due to Wilhelm Cauer. The group of real affine transformations,

$$[\mathbf{Z}'] = [\mathbf{T}]^T[\mathbf{Z}][\mathbf{T}]$$

where,

$$[\mathbf{T}] = \begin{bmatrix} 1 & 0 \cdots 0 \\ T_{21} & T_{22} \cdots T_{2n} \\ \cdot & \cdots \\ T_{n1} & T_{n2} \cdots T_{nn} \end{bmatrix}$$

is invariant in $Z(s)$. That is, all the transformed networks are equivalents according to the definition given here. If the $Z(s)$ for the initial given matrix is realisable, that is, it meets the PR condition, then all the transformed networks produced by this transformation will also meet the PR condition.

3 and 4-terminal networks

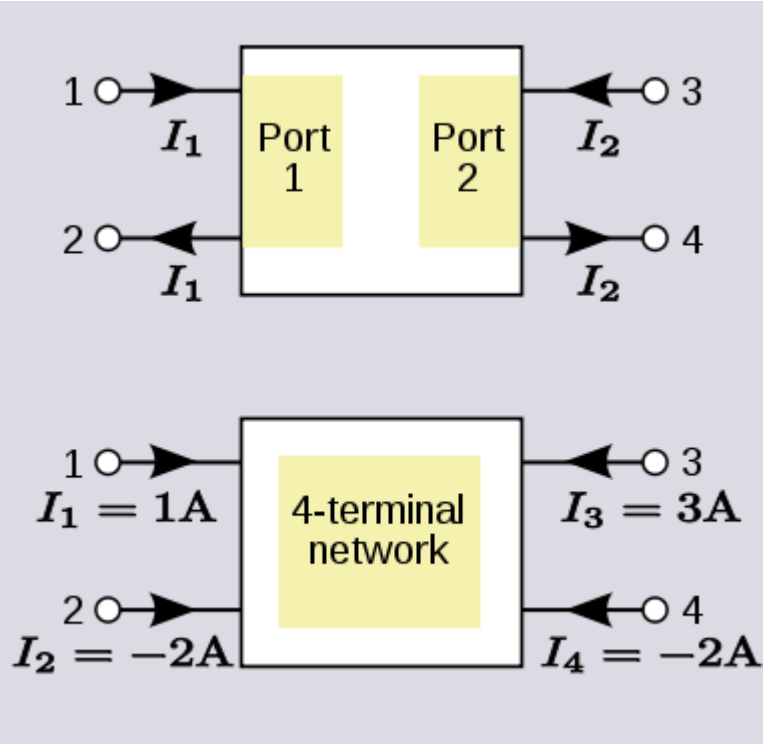
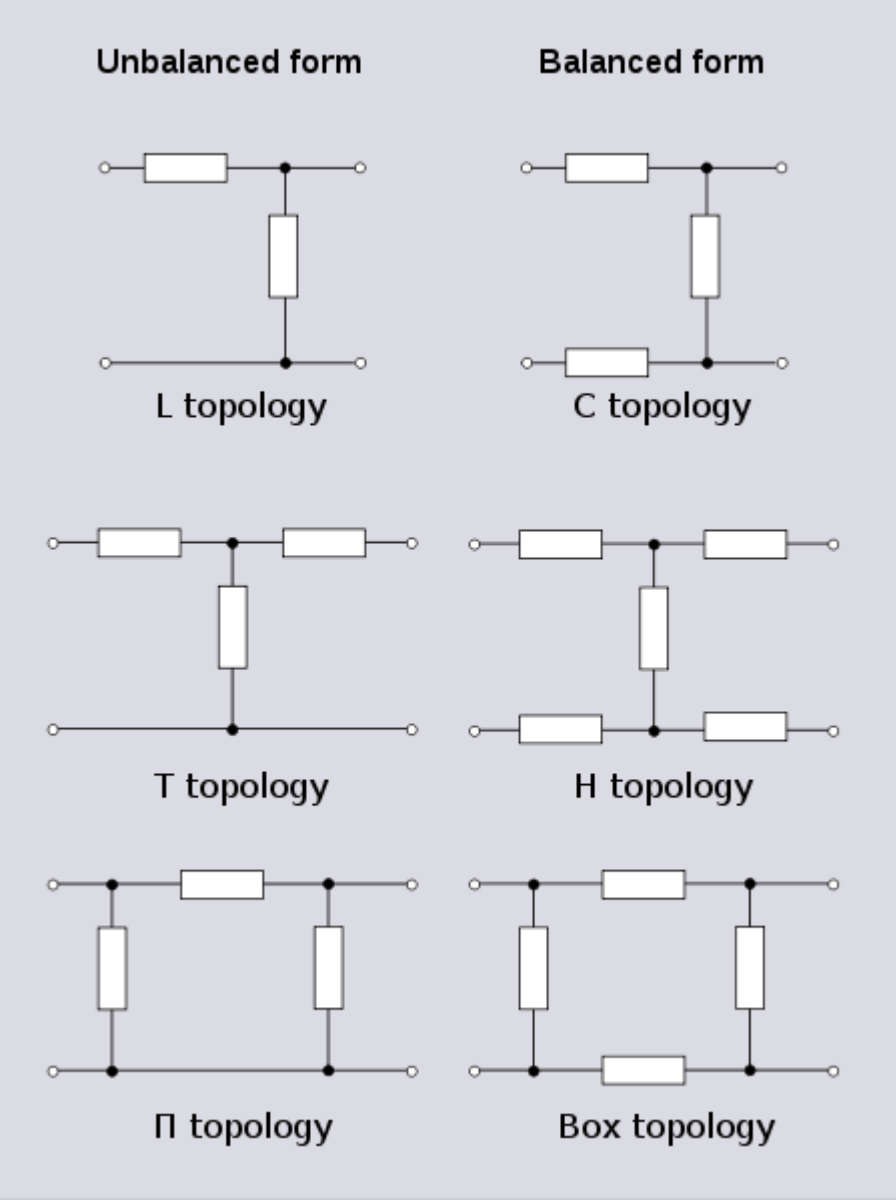


Fig. 2. A 4-terminal network connected by ports (top) has equal and opposite currents in each pair of terminals. The bottom network does not meet the port condition and cannot be treated as a 2-port. It could, however, be treated as an unbalanced 3-port by splitting one of the terminals into three common terminals shared between the ports.

When discussing 4-terminal networks, network analysis often proceeds in terms of 2-port networks, which covers a vast array of practically useful circuits. "2-port", in essence, refers to the way the network has been connected to the outside world: that the terminals have been connected in pairs to a source or load. It is possible to take exactly the same network and connect it to external circuitry in such a way that it is no longer behaving as a 2-port. This idea is demonstrated in Figure 2.



Equivalent unbalanced and balanced networks. The impedance of the series elements in the balanced version is half the corresponding impedance of the unbalanced version.

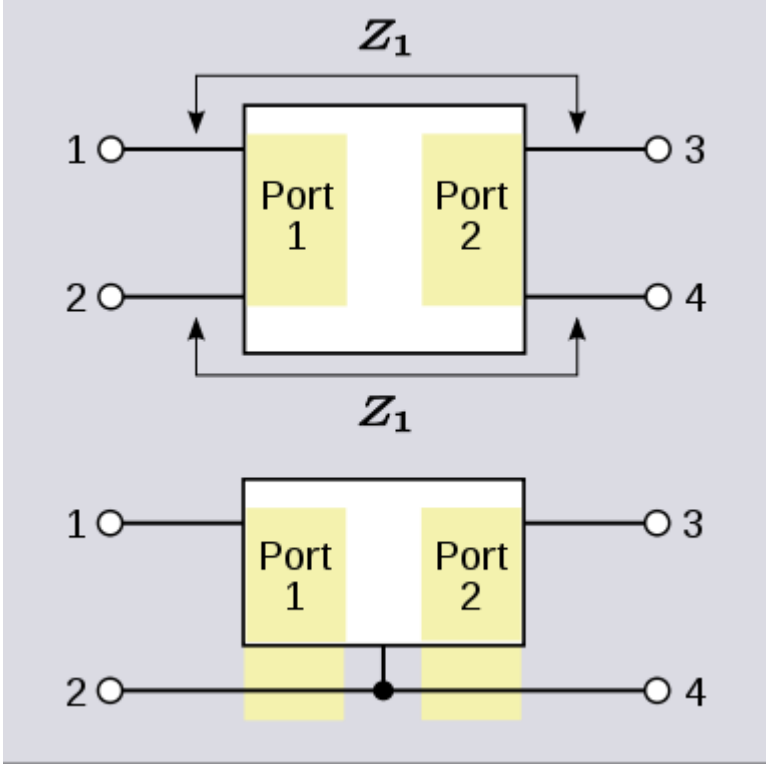


Fig. 3. To be balanced, a network must have the same impedance in each "leg" of the circuit.

A 3-terminal network can also be used as a 2-port. To achieve this, one of the terminals is connected in common to one terminal of both ports. In other words, one terminal has been split into two terminals and the network has effectively been converted to a 4-terminal network. This topology is known as unbalanced topology and is opposed to balanced topology. Balanced topology requires, referring to Figure 3, that the impedance measured between terminals 1 and 3 is equal to the impedance measured between 2 and 4. This is the pairs of terminals *not* forming ports: the case where the pairs of terminals forming ports have equal impedance is referred to as symmetrical. Strictly speaking, any network that does not meet the balance condition is unbalanced, but the term is most often referring to the 3-terminal topology described above and in Figure 3. Transforming an unbalanced 2-port network into a balanced network is usually quite straightforward: all series connected elements are divided in half with one half being relocated in what was the common branch. Transforming from balanced to unbalanced topology will often be possible with the reverse transformation but there are certain cases of certain topologies which cannot be transformed in this way.

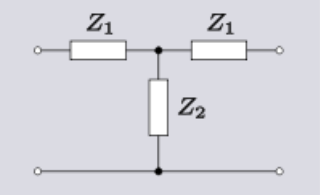
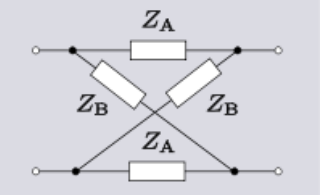
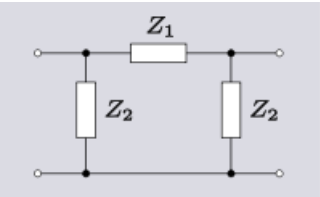
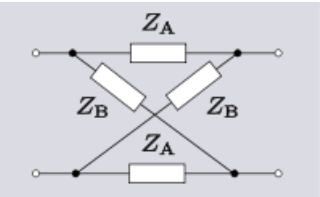
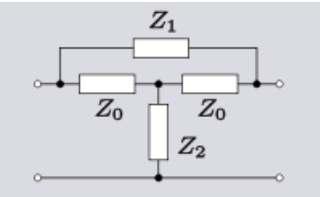
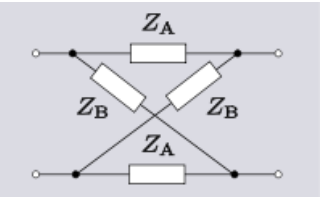
An example of a 3-terminal network transform that is not restricted to 2-ports is the Y-Δ transform. This is a particularly important transform for finding equivalent impedances. Its importance arises from the fact that the total impedance between two terminals cannot be determined solely by calculating series and parallel combinations except for a certain restricted class of network. In the general case additional transformations are required. The Y-Δ transform, its inverse the Δ-Y transform, and the *n*-terminal analogues of these

two transforms (star-polygon transforms) represent the minimal additional transforms required to solve the general case. Series and parallel are, in fact, the 2-terminal versions of star and polygon topology. A common simple topology that cannot be solved by series and parallel combinations is the input impedance to a bridge network (except in the special case when the bridge is in balance). The rest of the transforms in this section are all restricted to use with 2-ports only.

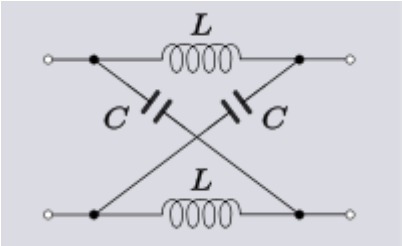
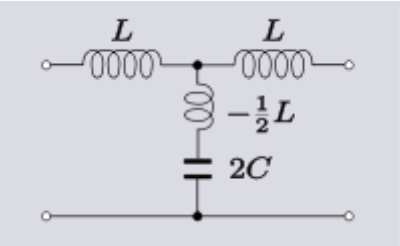
Lattice transforms

Symmetric 2-port networks can be transformed into lattice networks using Bartlett's bisection theorem. The method is limited to symmetric networks but this includes many topologies commonly found in filters, attenuators and equalisers. The lattice topology is intrinsically balanced, there is no unbalanced counterpart to the lattice and it will usually require more components than the transformed network.

Some common networks transformed to lattices (X-networks)

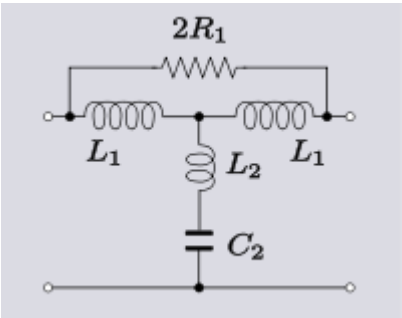
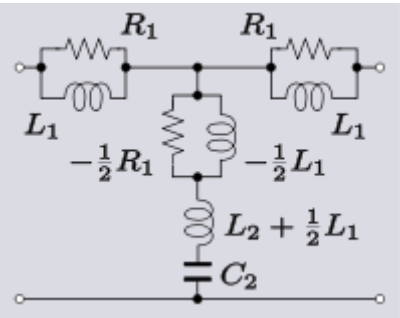
Description	Network	Transform equations	Transformed network
Transform 3.1 Transform of T network to lattice network.		$Z_\mathrm{A} = Z_1$, $Z_\mathrm{B} = Z_1 + 2Z_2$.	
Transform 3.2 Transform of Π network to lattice network.		$Z_\mathrm{A} = \frac{Z_1 Z_2}{Z_1 + 2Z_2}$, $Z_\mathrm{B} = Z_2$.	
Transform 3.3 Transform of Bridged-T network to lattice network.		$Z_\mathrm{A} = \frac{Z_1 Z_0}{Z_1 + 2Z_0}$, $Z_\mathrm{B} = Z_0 + 2Z_2$.	

Reverse transformations from a lattice to an unbalanced topology are not always possible in terms of passive components. For instance, this transform,

Description	Network	Transformed network
Transform 3.4 Transform of a lattice phase equaliser to a T network.		

cannot be realised with passive components because of the negative values arising in the transformed circuit. It can however be realised if mutual inductances and ideal transformers are permitted, for instance, in this circuit. Another possibility is to permit the use of active components which would enable negative impedances to be directly realised as circuit components.

It can sometimes be useful to make such a transformation, not for the purposes of actually building the transformed circuit, but rather, for the purposes of aiding understanding of how the original circuit is working. The following circuit in bridged-T topology is a modification of a mid-series m-derived filter T-section. The circuit is due to Hendrik Bode who claims that the addition of the bridging resistor of a suitable value will cancel the parasitic resistance of the shunt inductor. The action of this circuit is clear if it is transformed into T topology - in this form there is a negative resistance in the shunt branch which can be made to be exactly equal to the positive parasitic resistance of the inductor.

Description	Network	Transformed network
Transform 3.5 Transform of a bridged-T low-pass filter section to a T-section.		

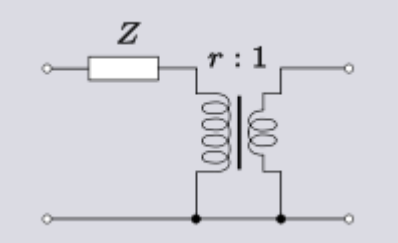
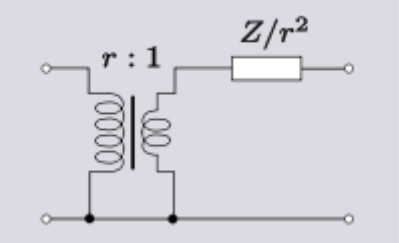
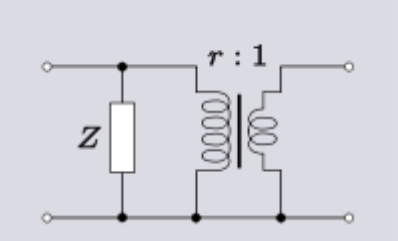
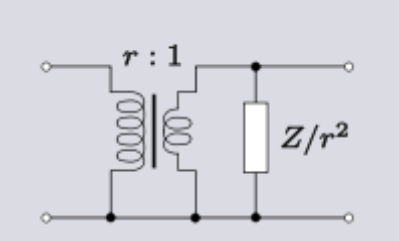
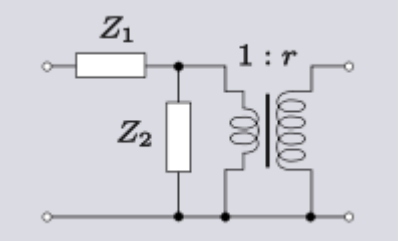
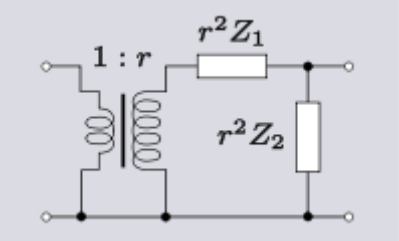
Any symmetrical network can be transformed into any other symmetrical network by the same method, that is, by first transforming into the intermediate lattice form (omitted for clarity from the above example transform) and from the lattice form into the required target form. As with the example, this will generally result in negative elements except in special cases.

Eliminating resistors

A theorem due to Sidney Darlington states that any PR function $Z(s)$ can be realised as a lossless two-port terminated in a positive resistor R. That is, regardless of how many resistors feature in the matrix [**Z**] representing the impedance network, a transform can be found that will realise the network entirely as an LC-kind network with just one resistor across the output port (which would normally represent the load). No resistors within the network are necessary in order to realise the specified response. Consequently, it is always possible to reduce 3-element-kind 2-port networks to 2-element-kind (LC) 2-port networks provided the output port is terminated in a resistance of the required value.

Eliminating ideal transformers

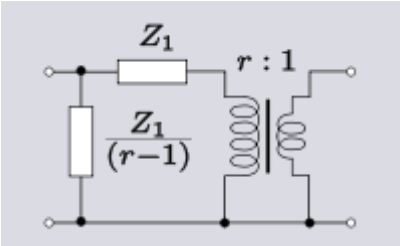
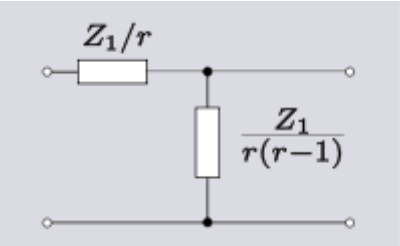
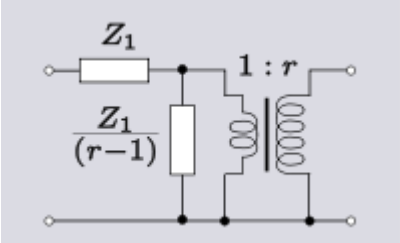
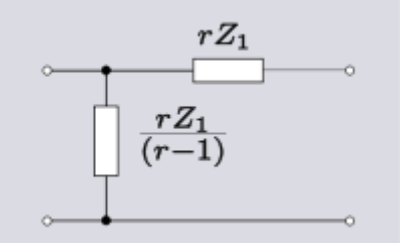
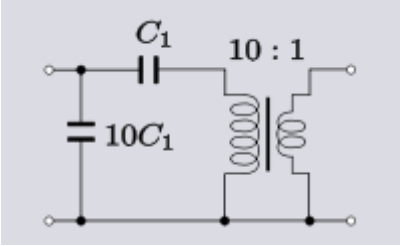
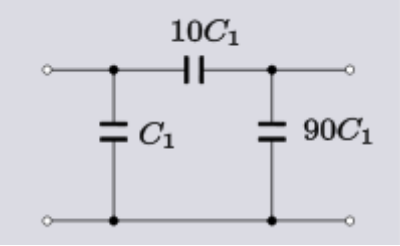
An elementary transformation that can be done with ideal transformers and some other impedance element is to shift the impedance to the other side of the transformer. In all the following transforms, r is the turns ratio of the transformer.

Description	Network	Transformed network
Transform 4.1 Series impedance through a step-down transformer.		
Transform 4.2 Shunt impedance through a step-down transformer.		
Transform 4.3 Shunt and series impedance network through a step-up transformer.		

These transforms do not just apply to single elements; entire networks can be passed through the transformer. In this manner, the transformer can be shifted around the network to a more convenient location.

Darlington gives an equivalent transform that can eliminate an ideal transformer altogether. This technique requires that the transformer is next to (or capable of being

moved next to) an "L" network of same-kind impedances. The transform in all variants results in the "L" network facing the opposite way, that is, topologically mirrored.

Description	Network	Transformed network
Transform 5.1 Elimination of a step-down transformer.		
Transform 5.2 Elimination of a step-up transformer.		
Example 3. Example of transform 5.1.		

Example 3 shows the result is a Π-network rather than an L-network. The reason for this is that the shunt element has more capacitance than is required by the transform so some is still left over after applying the transform. If the excess were instead, in the element nearest the transformer, this could be dealt with by first shifting the excess to the other side of the transformer before carrying out the transform.

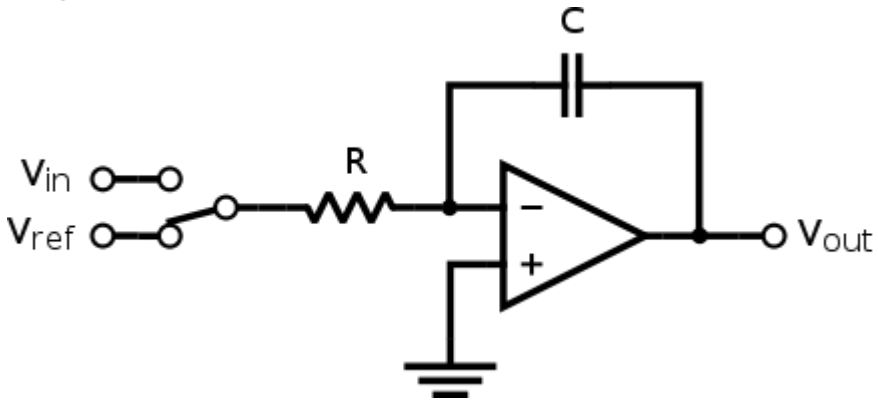
Chapter- 7

Integrating ADC

An **integrating ADC** is a type of analog-to-digital converter that converts an unknown input voltage into a digital representation through the use of an integrator. In its most basic implementation, the unknown input voltage is applied to the input of the integrator and allowed to ramp for a fixed time period (the run-up period). Then a known reference voltage of opposite polarity is applied to the integrator and is allowed to ramp until the integrator output returns to zero (the run-down period). The input voltage is computed as a function of the reference voltage, the constant run-up time period, and the measured run-down time period. The run-down time measurement is usually made in units of the converter's clock, so longer integration times allow for higher resolutions. Likewise, the speed of the converter can be improved by sacrificing resolution.

Converters of this type (or variations on the concept) are able to achieve high resolutions (8.5 digits, or 28 bits, in the case of the Agilent 3458A digital multimeter), but often do so at the expense of speed. The Agilent 3458A, for example, only achieves its highest resolution at a rate of six samples per second. For this reason, these converters are not found in audio or signal processing applications. Their use is typically limited to digital voltmeters and other instruments requiring highly accurate measurements.

Basic Design

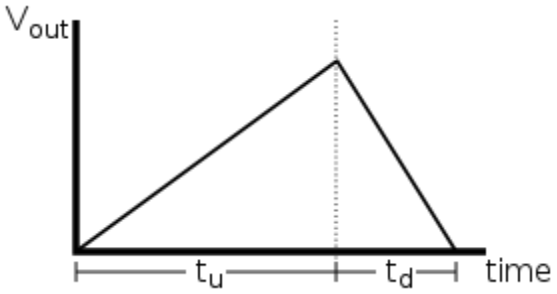


Basic integrator of a Dual-slope Integrating ADC. The comparator, the timer, and the controller are not shown.

The basic integrating ADC circuit consists of an integrator, a switch to select between the voltage to be measured and the reference voltage, a timer that determines how long to integrate the unknown and measures how long the reference integration took, a comparator to detect zero crossing, and a controller. Depending on the implementation, a switch may also be present in parallel with the integrator capacitor to allow the integrator to be reset (by discharging the integrator capacitor). The switches will be controlled electrically by means of the converter's controller (a microprocessor or dedicated control logic). Inputs to the controller include a clock (used to measure time) and the output of a comparator used to detect when the integrator's output reaches zero.

The conversion takes place in two phases: the run-up phase, where the input to the integrator is the voltage to be measured, and the run-down phase, where the input to the integrator is a known reference voltage. During the run-up phase, the switch selects the measured voltage as the input to the integrator. The integrator is allowed to ramp for a fixed period of time to allow a charge to build on the integrator capacitor. During the run-down phase, the switch selects the reference voltage as the input to the integrator. The time that it takes for the integrator's output to return to zero is measured during this phase.

In order for the reference voltage to ramp the integrator voltage down, the reference voltage needs to have a polarity opposite to that of the input voltage. In most cases, for positive input voltages, this means that the reference voltage will be negative. To handle both positive and negative input voltages, a positive and negative reference voltage is required. The selection of which reference to use during the run-down phase would be based on the polarity of the integrator output at the end of the run-up phase. That is, if the integrator's output were negative at the end of the run-up phase, a negative reference voltage would be required. If the integrator's output were positive, a positive reference voltage would be required.



Integrator output voltage in a basic dual-slope integrating ADC

The basic equation for the output of the integrator (assuming a constant input) is:

$$V_{out} = -\frac{V_{in}}{RC} t_{int} + V_{initial}$$

Assuming that the initial integrator voltage at the start of each conversion is zero and that the integrator voltage at the end of the run down period will be zero, we have the following two equations that cover the integrator's output during the two phases of the conversion:

$$V_{out-up} = -\frac{V_{in}}{RC}t_u$$
$$V_{out-down} = -\frac{V_{ref}}{RC}t_d + V_{out-up} = 0$$

The two equations can be combined and solved for V_{in}, the unknown input voltage:

$$V_{in} = -V_{ref}\frac{t_d}{t_u}$$

From the equation, one of the benefits of the dual-slope integrating ADC becomes apparent: the measurement is independent of the values of the circuit elements (R and C). This does not mean, however, that the values of R and C are unimportant in the design of a dual-slope integrating ADC (as will be explained below).

Note that in the graph to the right, the voltage is shown as going up during the run-up phase and down during the run-down phase. In reality, because the integrator uses the op-amp in a negative feedback configuration, applying a positive V_{in} will cause the output of the integrator to go *down*. The *up* and *down* more accurately refer to the process of adding charge to the integrator capacitor during the run-up phase and removing charge during the run-down phase.

The resolution of the dual-slope integrating ADC is determined primarily by the length of the run-down period and by the time measurement resolution (i.e., the frequency of the controller's clock). The required resolution (in number of bits) dictates the minimum length of the run-down period for a full-scale input ($V_{in} = -V_{ref}$):

$$t_d = \frac{2^r}{f_{clk}}$$

During the measurement of a full-scale input, the slope of the integrator's output will be the same during the run-up and run-down phases. This also implies that the time of the run-up period and run-down period will be equal ($t_u = t_d$) and that the total measurement time will be $2t_d$. Therefore, the total measurement time for a full-scale input will be based on the desired resolution and the frequency of the controller's clock:

$$t_m = 2\frac{2^r}{f_{clk}}$$

If a resolution of 16 bits is required with a controller clock of 10 MHz, the measurement time will be 13.1 milliseconds (or a sampling rate of just 76 samples per second). However, the sampling time can be improved by sacrificing resolution. If the resolution requirement is reduced to 10 bits, the measurement time is also reduced to only 0.2 milliseconds (almost 4900 samples per second).

Limitations

There are limits to the maximum resolution of the dual-slope integrating ADC. It is not possible to increase the resolution of the basic dual-slope ADC to arbitrarily high values by using longer measurement times or faster clocks. Resolution is limited by:

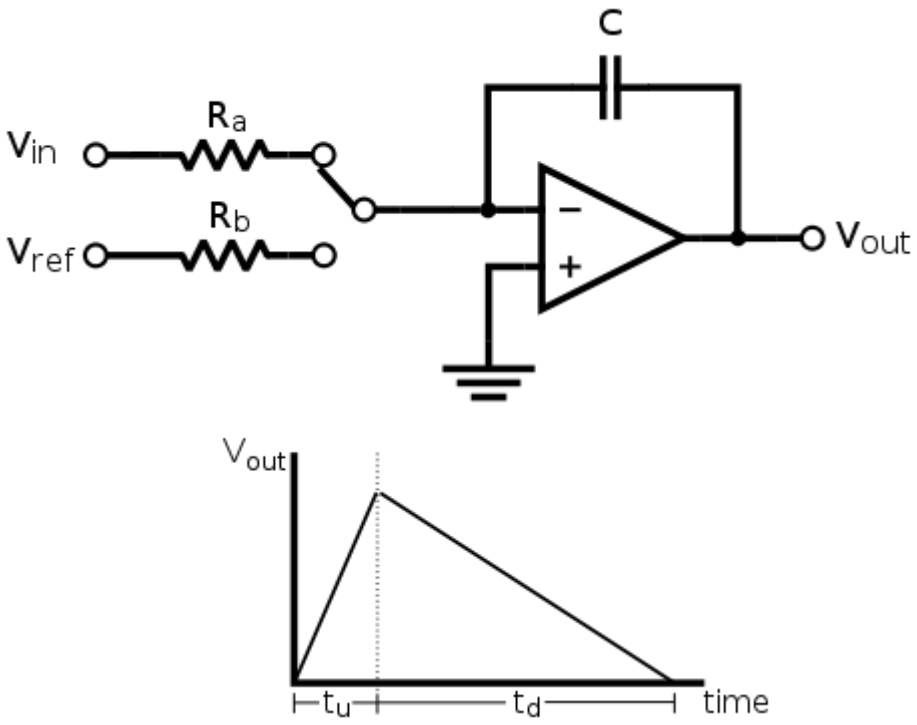
- The range of the integrating amplifier. The voltage rails on an op-amp limit the output voltage of the integrator. An input left connected to the integrator for too long will eventually cause the op amp to limit its output to some maximum value, making any calculation based on the run-down time meaningless. The integrator's resistor and capacitor are therefore chosen carefully based on the voltage rails of the op-amp, the reference voltage and expected full-scale input, and the longest run-up time needed to achieve the desired resolution.
- The accuracy of the comparator used as the null detector. Wideband circuit noise limits the ability of the comparator to identify exactly when the output of the integrator has reached zero. Goerke suggests a typical limit is a comparator resolution of 1 millivolt.
- The quality of the integrator's capacitor. Although the integrating capacitor need not be perfectly linear, it does need to be time-invariant. Dielectric absorption causes errors.

Enhancements

The basic design of the dual-slope integrating ADC has a limitations in both conversion speed and resolution. A number of modifications to the basic design have been made to overcome both of these to some degree.

Run-up improvements

Enhanced dual-slope



Enhanced run-up dual-slope integrating ADC

The run-up phase of the basic dual-slope design integrates the input voltage for a fixed period of time. That is, it allows an unknown amount of charge to build up on the integrator's capacitor. The run-down phase is then used to measure this unknown charge to determine the unknown voltage. For a full-scale input, half of the measurement time is spent in the run-up phase. For smaller inputs, an even larger percentage of the total measurement time is spent in the run-up phase. Reducing the amount of time spent in the run-up phase can significantly reduce the total measurement time.

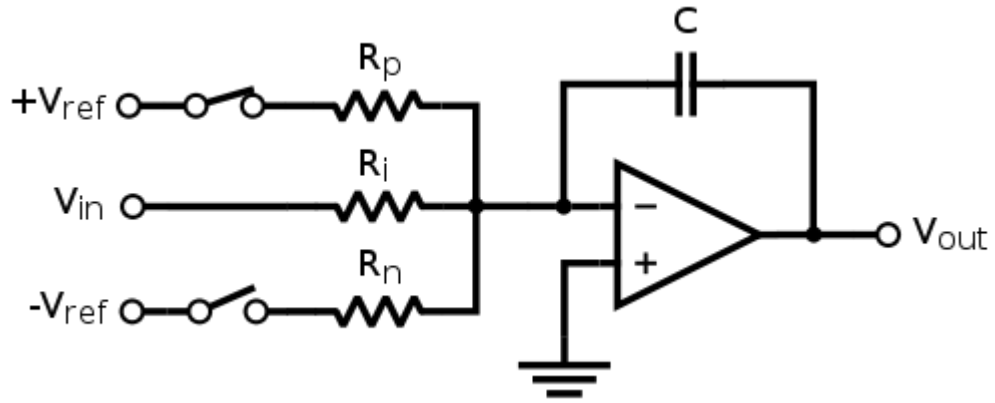
A simple way to reduce the run-up time is to increase the rate that charge accumulates on the integrator capacitor by reducing the size of the resistor used on the input, a method referred to as enhanced dual-slope. This still allows the same total amount of charge accumulation, but it does so over a smaller period of time. Using the same algorithm for the run-down phase results in the following equation for the calculation of the unknown input voltage (V_{in}):

$$V_{in} = -V_{ref}\frac{R_a}{R_b}\frac{t_d}{t_u}$$

Note that this equation, unlike the equation for the basic dual-slope converter, has a dependence on the values of the integrator resistors. Or, more importantly, it has a

dependence on the *ratio* of the two resistance values. This modification does nothing to improve the resolution of the converter (since it doesn't address either of the resolution limitations noted above).

Multi-slope run-up



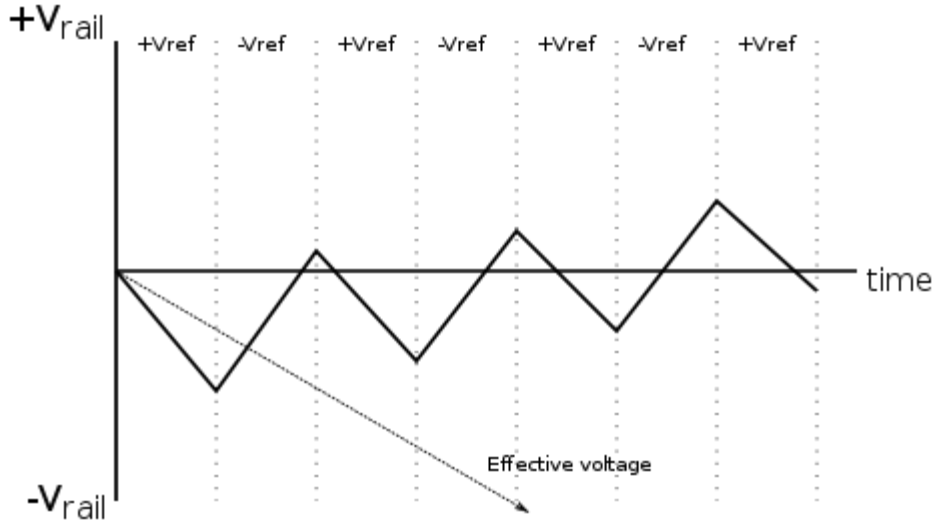
Circuit diagram for a multi-slope run-up converter

One method to improve the resolution of the converter is to artificially increase the range of the integrating amplifier during the run-up phase. As mentioned above, the purpose of the run-up phase is to add an unknown amount of charge to the integrator to be later measured during the run-down phase. Having the ability to add larger quantities of charge allows for more higher-resolution measurements. For example, assume that we are capable of measuring the charge on the integrator during the run-down phase to a granularity of 1 coulomb. If our integrator amplifier limits us to being able to add only up to 16 coulombs of charge to the integrator during the run-up phase, our total measurement will be limited to 4 bits (16 possible values). If we can increase the range of the integrator to allow us to add up to 32 coulombs, our measurement resolution is increased to 5 bits.

One method to increase the integrator capacity is by periodically adding or subtracting known quantities of charge during the run-up phase in order to keep the integrator's output within the range of the integrator amplifier. Then, the total amount of artificially-accumulated charge is the charge introduced by the unknown input voltage plus the sum of the known charges that were added or subtracted.

The circuit diagram shown to the right is an example of how multi-slope run-up could be implemented. The concept is that the unknown input voltage, V_{in}, is always applied to the integrator. Positive and negative reference voltages controlled by the two independent switches add and subtract charge as needed to keep the output of the integrator within its limits. The reference resistors, R_p and R_n are necessarily smaller than R_i to ensure that the references can overcome the charge introduced by the input. A comparator is connected to the output to compare the integrator's voltage with a threshold voltage. The output of the comparator is used by the converter's controller to decide which reference voltage should be applied. This can be a relatively simple algorithm: if the integrator's output

above the threshold, enable the positive reference (to cause the output to go down); if the integrator's output is below the threshold, enable the negative reference (to cause the output to go up). The controller keeps track of how often each switch is turned on in order to estimate how much additional charge was placed onto (or removed from) the integrator capacitor as a result of the reference voltages.



Output from multi-slope run-up

To the right is a graph of sample output from the integrator during a multi-slope run-up. Each dashed vertical line represents a decision point by the controller where it samples the polarity of the output and chooses to apply either the positive or negative reference voltage to the input. Ideally, the output voltage of the integrator at the end of the run-up period can be represented by the following equation:

$$V_{out} = -\frac{\frac{NV_{in}t_\Delta}{R_i} + \frac{N_pV_{ref}t_\Delta}{R_p} - \frac{N_nV_{ref}t_\Delta}{R_n}}{C}$$

where t_Δ is the sampling period, N_p is the number of periods in which the positive reference is switched in, N_n is the number of periods in which the negative reference is switched in, and N is the total number of periods in the run-up phase.

The resolution obtained during the run-up period can be determined by making the assumption that the integrator output at the end of the run-up phase is zero. This allows us to relate the unknown input, V_{in}, to just the references and the N values:

$$\frac{NV_{in}}{R_i} = -\left(\frac{N_pV_{ref}}{R_p} - \frac{N_nV_{ref}}{R_n}\right)$$

The resolution can be expressed in terms of the difference between single steps of the converter's output. In this case, if we solve the above equation for V_{in} using $N_p = 0, N_n = N$ and $N_p = 1, N_n = N - 1$ (the sum of N_p and N_n must always equal N), the difference will equal the smallest resolvable quantity. This results in an equation for the resolution of the multi-slope run-up phase (in bits) of:

$$r = log_2 \frac{R_i(R_p + R_n)}{N R_n R_p}$$

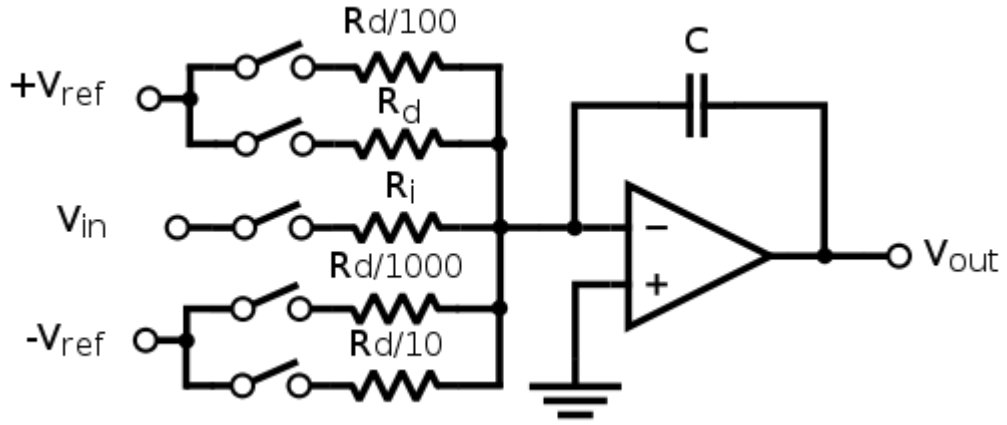
Using typical values of the reference resistors R_p and R_n of 10k ohms and an input resistor of 50k ohms, we can achieve a 16 bit resolution during the run-up phase with 655360 periods (65.5 milliseconds with a 10 MHz clock).

While it is possible to continue the multi-slope run-up indefinitely, it is not possible to increase the resolution of the converter to arbitrarily high levels just by using a longer run-up time. Error is introduced into the multi-slope run-up through the action of the switches controlling the references, cross-coupling between the switches, unintended switch charge injection, mismatches in the references, and timing errors.

Some of this error can be reduced by careful operation of the switches. In particular, during the run-up period, each switch should be activated a constant number of times. The algorithm explained above does not do this and just toggles switches as needed to keep the integrator output within the limits. Activating each switch a constant number of times makes the error related to switching approximately constant. Any output offset that is a result of the switching error can be measured and then subtracted from the result.

Run-down improvements

Multi-slope run-down

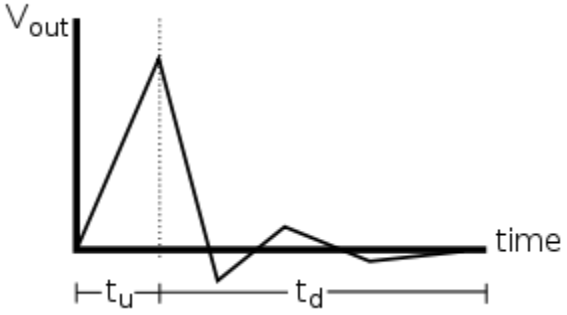


Multi-slope run-down integrating ADC

The simple, single-slope run-down is slow. Typically, the run down time is measured in clock ticks, so to get four digit resolution, the rundown time may take as long as 10,000 clock cycles. A multi-slope run-down can speed the measurement up without sacrificing accuracy. By using 4 slope rates that are each a power of ten more gradual than the previous, four digit resolution can be achieved in roughly 40 or fewer clock ticks—a huge speed improvement.

The circuit shown to the right is an example of a multi-slope run-down circuit with four run-down slopes with each being ten times more gradual than the previous. The switches control which slope is selected. The switch containing R_d / 1000 selects the steepest slope (i.e., will cause the integrator output to move toward zero the fastest). At the start of the run-down interval, the unknown input is removed from the circuit by opening the switch connected to V_{in} and closing the R_d / 1000 switch. Once the integrator's output reaches zero (and the run-down time measured), the R_d / 1000 switch is opened and the next slope is selected by closing the R_d / 100 switch. This repeats until the final slope of R_d has reached zero. The combination of the run-down times for each of the slopes determines the value of the unknown input. In essence, each slope adds one digit of resolution to the result.

In the example circuit, the slope resistors differ by a factor of 10. This value, known as the *base* (*B*), can be any value. As explained below, the choice of the base affects the speed of the converter and determines the number of slopes needed to achieve the desired resolution.



Output of the multi-slope run-down integrating ADC

The basis of this design is the assumption that there will always be overshoot when trying to find the zero crossing at the end of a run-down interval. This will necessarily be true given any hysteresis in the output of the comparator measuring the zero crossing and due to the periodic sampling of the comparator based on the converter's clock. If we assume that the converter switches from one slope to the next in a single clock cycle (which may or may not be possible), the maximum amount of overshoot for a given slope would be the largest integrator output change in one clock period:

$$V_\Delta = \frac{V_{ref}}{RC}\frac{1}{f_{clk}}$$

To overcome this overshoot, the next slope would require no more than B clock cycles, which helps to place a bound on the total time of the run-down. The time for the first-run down (using the steepest slope) is dependent on the unknown input (i.e., the amount of charge placed on the integrator capacitor during the run-up phase). At most, this will be:

$$T_{first} = \left\lceil \frac{V_{max} C R_{s1} f_{clk}}{V_{ref}} \right\rceil$$

where T_{first} is the maximum number of clock periods for the first slope, V_{max} is the maximum integrator voltage at the start of the run-down phase, and R_{s1} is the resistor used for the first slope.

The remainder of the slopes have a limited duration based on the selected base, so the remaining time of the conversion (in converter clock periods) is:

$$T_d \leq B(N-1)$$

where N is the number of slopes.

Converting the measured time intervals during the multi-slope run-down into a measured voltage is similar to the charge-balancing method used in the multi-slope run-up enhancement. Each slope adds or subtracts known amounts of charge to/from the integrator capacitor. The run-up will have added some unknown amount of charge to the integrator. Then, during the run-down, the first slope subtracts a large amount of charge, the second slope adds a smaller amount of charge, etc. with each subsequent slope moving a smaller amount in the opposite direction of the previous slope with the goal of reaching closer and closer to zero. Each slope adds or subtracts a quantity of charge proportional to the slope's resistor and the duration of the slope:

$$C_{slope} = \pm \frac{V_{ref} T_{slope}}{R_{slope} f_{clk}}$$

T_{clock} is necessarily an integer and will be less than or equal to B for the second and subsequent slopes. Using the circuit above as an example, the second slope, R_d / 100, can contribute the following charge, C_{slope2}, to the integrator:

$$\frac{100 V_{ref}}{R_d f_{clk}} \leq C_{slope2} \leq \frac{1000 V_{ref}}{R_d f_{clk}} \text{ in steps of } \frac{100 V_{ref}}{R_d f_{clk}}$$

That is, B possible values with the largest equal to the first slope's smallest step, or one (base 10) digit of resolution per slope. Generalizing this, we can represent the number of slopes, N, in terms of the base and the required resolution, M:

$N = \log_B M$

Substituting this back into the equation representing the run-down time required for the second and subsequent slopes gives us this:

$$T_d \leq B(log_B(M) - 1)$$

Which, when evaluated, shows that the minimum run-down time can be achieved using a base of e. This base may be difficult to use both in terms of complexity in the calculation of the result and of finding an appropriate resistor network, so a base of 2 or 4 would be more common.

Residue ADC

When using run-up enhancements like the multi-slope run-up, where a portion of the converter's resolution is resolved during the run-up phase, it is possible to eliminate the run-down phase altogether by using a second type of analog-to-digital converter. At the end of the run-up phase of a multi-slope run-up conversion, there will still be an unknown amount of charge remaining on the integrator's capacitor. Instead of using a traditional run-down phase to determine this unknown charge, the unknown voltage can be converted directly by a second converter and combined with the result from the run-up phase to determine the unknown input voltage.

Assuming that multi-slope run-up as described above is being used, the unknown input voltage can be related to the multi-slope run-up counters, N_p and N_n, and the measured integrator output voltage, V_{out} using the following equation (derived from the multi-slope run-up output equation):

$$V_{in} = \frac{R_i \left(-\frac{N_p V_{ref} t_\Delta}{R_p} + \frac{N_n V_{ref} t_\Delta}{R_n} - C V_{out} \right)}{N t_\Delta}$$

This equation represents the theoretical calculation of the input voltage assuming ideal components. Since the equation depends on nearly all of the circuit's parameters, any variances in reference currents, the integrator capacitor, or other values will introduce errors in the result. A calibration factor is typically included in the CV_{out} term to account for measured errors (or, as described in the referenced patent, to convert the residue ADC's output into the units of the run-up counters).

Instead of being used to eliminate the run-down phase completely, the residue ADC can also be used to make the run-down phase more accurate than would otherwise be possible. With a traditional run-down phase, the run-down time measurement period ends with the integrator output crossing through zero volts. There is a certain amount of error involved in detecting the zero crossing using a comparator (one of the short-comings of the basic dual-slope design as explained above). By using the residue ADC to rapidly sample the integrator output (synchronized with the converter controller's clock, for example), a voltage reading can be taken both immediately before and immediately after

the zero crossing (as measured with a comparator). As the slope of the integrator voltage is constant during the run-down phase, the two voltage measurements can be used as inputs to an interpolation function that more accurately determines the time of the zero-crossing (i.e., with a much higher resolution than the controller's clock alone would allow).

Other improvements

Continuously-integrating Converter

By combining some of these enhancements to the basic dual-slope design (namely multi-slope run-up and the residue ADC), it is possible to construct an integrating analog-to-digital converter that is capable of operating continuously without the need for a run-down interval. Conceptually, the multi-slope run-up algorithm is allowed to operate continuously. To start a conversion, two things happen simultaneously: the residue ADC is used to measure the approximate charge currently on the integrator capacitor and the counters monitoring the multi-slope run-up are reset. At the end of a conversion period, another residue ADC reading is taken and the values of the multi-slope run-up counters are noted.

The unknown input is calculated using a similar equation as used for the residue ADC, except that two output voltages are included (V_{out1} representing the measured integrator voltage at the start of the conversion, and V_{out2} representing the measured integrator voltage at the end of the conversion.

$$V_{in} = \frac{R_i \left(-\frac{N_p V_{ref} t_\Delta}{R_p} + \frac{N_n V_{ref} t_\Delta}{R_n} - C(V_{out2} - V_{out1}) \right)}{N t_\Delta}$$

Such a continuously-integrating converter is very similar to a delta-sigma analog-to-digital converter.

Calibration

In most variants of the dual-slope integrating converter, the converter's performance is dependent on one or more of the circuit parameters. In the case of the basic design, the output of the converter is in terms of the reference voltage. In more advanced designs, there are also dependencies on one or more resistors used in the circuit or on the integrator capacitor being used. In all cases, even using expensive precision components there may be other effects that are not accounted for in the general dual-slope equations (dielectric effect on the capacitor or frequency or temperature dependencies on any of the components). Any of these variations result in error in the output of the converter. In the best case, this is simply gain and/or offset error. In the worst case, nonlinearity or nonmonotonicity could result.

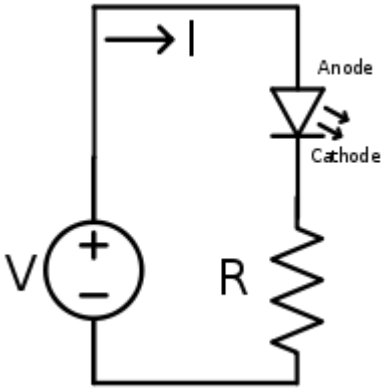
Some calibration can be performed internal to the converter (i.e., not requiring any special external input). This type of calibration would be performed every time the converter is turned on, periodically while the converter is running, or only when a special calibration mode is entered. Another type of calibration requires external inputs of known quantities (e.g., voltage standards or precision resistance references) and would typically be performed infrequently (every year for equipment used in normal conditions, more often when being used in metrology applications).

Of these types of error, offset error is the simplest to correct (assuming that there is a constant offset over the entire range of the converter). This is often done internal to the converter itself by periodically taking measurements of the ground potential. Ideally, measuring the ground should always result in a zero output. Any non-zero output indicates the offset error in the converter. That is, if the measurement of ground resulted in an output of 0.001 volts, one can assume that all measurements will be offset by the same amount and can subtract 0.001 from all subsequent results.

Gain error can similarly be measured and corrected internally (again assuming that there is a constant gain error over the entire output range). The voltage reference (or some voltage derived directly from the reference) can be used as the input to the converter. If the assumption is made that the voltage reference is accurate (to within the tolerances of the converter) or that the voltage reference has been externally calibrated against a voltage standard, any error in the measurement would be a gain error in the converter. If, for example, the measurement of a converter's 5 volt reference resulted in an output of 5.3 volts (after accounting for any offset error), a gain multiplier of 0.94 (5 / 5.3) can be applied to any subsequent measurement results.

Chapter- 8

LED Circuit



Simple LED circuit diagram

In electronics, the basic **LED circuit** is an electric power circuit used to power a light-emitting diode or LED. The simplest such circuit consists of a voltage source powering two components connected in series: A current-limiting resistor (sometimes called the ballast resistor), and an LED. Optionally, a switch may be introduced to open and close the circuit. The switch may be replaced with another component or circuit to form a continuity tester.

(Although simple, this circuit is not necessarily the most energy efficient circuit to drive an LED, since energy is lost in the resistor. More complicated circuits may be used to improve energy efficiency).

The LED used will have a voltage drop, specified at the intended operating current. Ohm's law and Kirchhoff's circuit laws are used to calculate the resistor that is used to attain the correct current. The resistor value is computed by subtracting the LED voltage drop from the supply voltage, and then dividing by the desired LED operating current. If the supply voltage is equal to the LED's voltage drop, no resistor is needed.

This basic circuit is used in a wide range of applications, including many consumer appliances.

Simple resistance formula for optimum brightness of the LED

The formula to calculate the correct resistance to use is:

$$\text{resistance}(R) = \frac{\text{power supply voltage}(V_s) - \text{LED voltage drop}(V_f)}{\text{LED current rating}(I_f)}$$

where:

- **Power supply voltage** (V_s) is the voltage of the power supply: e.g. a 9 volt battery.
- **LED voltage drop** (V_f) is the voltage drop across the LED. Typically, this is about 1.8 – 3.3 volts; it varies by the color of the LED. A *red* LED typically drops 1.8 volts, but voltage drop normally rises as the light frequency increases, so a *blue* LED may drop around 3.3 volts.
- **LED current rating** (I_f) is the manufacturer rating of the LED (Although the above formula requires the current in amperes, this value is usually given by the manufacturer in milliamperes, such as 20 mA).

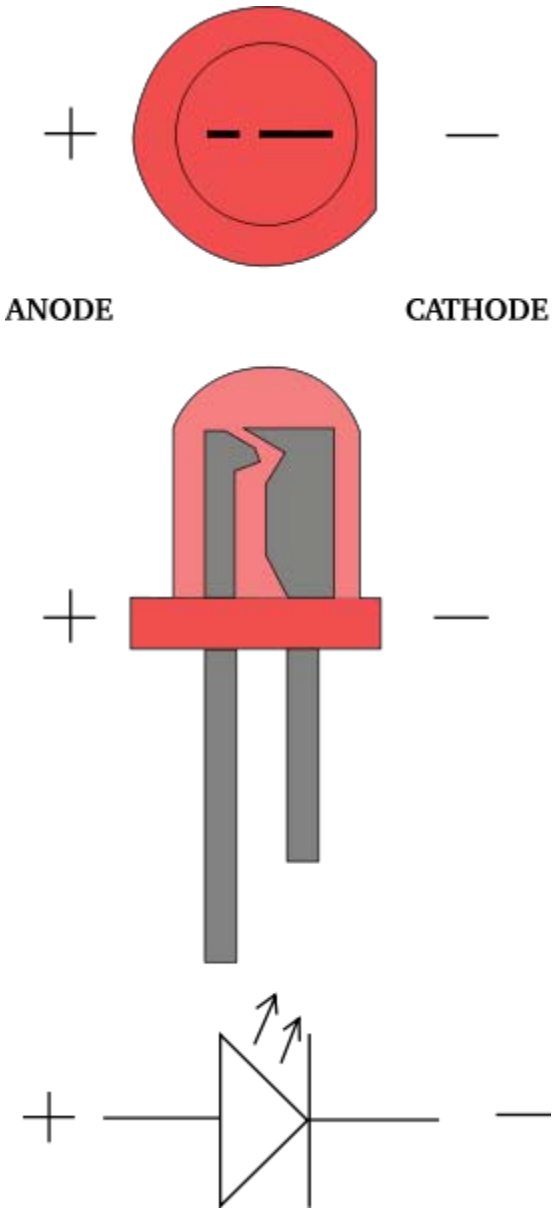
Analysis using Kirchhoff's Laws

The formula can be explained considering the LED as a $\frac{V_f}{I_f}\ \Omega$ resistance, and applying the KVL (R is the unknown quantity):

$$V_s = V_r + V_f = RI_f + \frac{V_f}{I_f} I_f$$

$$RI_f = V_s - V_f$$

$$R = \frac{V_s - V_f}{I_f}$$



LED orientation

Polarity

Unlike incandescent light bulbs, which illuminate regardless of the electrical polarity, LEDs will only light with correct electrical polarity. When the voltage across the *p-n junction* is in the correct direction, a significant current flows and the device is said to be *forward-biased*. If the voltage is of the wrong polarity, the device is said to be *reverse biased*, very little current flows, and no light is emitted. LEDs can be operated on an alternating current voltage, but they will only light with positive voltage, causing the LED to turn on and off at the frequency of the AC supply.

Most LEDs have low reverse breakdown voltage ratings, so they will also be damaged by an applied reverse voltage above this threshold. The cause of damage is overcurrent resulting from the diode breakdown, not the voltage itself. LEDs driven directly from an AC supply of more than the reverse breakdown voltage may be protected by placing a diode (or another LED) in inverse parallel.

The manufacturer will normally advise how to determine the polarity of the LED in the product datasheet. However, these methods may also be used:

sign:	+	-
terminal:	anode (A)	cathode (K)
leads:	long	short
exterior:	round	flat
wiring:	red	black
marking:*	none	stripe
pin:*	1	2
PCB:*	round	square

(*)Less reliable methods of determining polarity

Power sources

The voltage versus current characteristics of an **LED** are much like any diode. Current is approximately an exponential function of voltage, so a small voltage change results in a large change in current. It is therefore important that the power source gives the right voltage.

If the voltage is below the threshold or on-voltage no current will flow and the result is an unlit LED. If the voltage is too high the current will go above the maximum rating, heating and potentially destroying the LED. As the LED heats, its voltage drop decreases (band gap decrease), further increasing current. Consequently, LEDs should only be connected directly to constant-voltage sources if special care is taken. Series resistors are a simple way to stabilize the LED current, but wastes energy in the resistor. A constant current regulator is commonly used for high power LEDs. Low drop-out (LDO) constant current regulators also allow the total LED string voltage to be a higher percentage of the power supply voltage, resulting in improved efficiency and reduced power use. Switched-mode power supplies are used in some LED flashlights, stabilizing light output over a wide range of battery voltages and increasing the useful life of the batteries.

Miniature indicator LEDs are normally driven from low voltage DC via a current limiting resistor. Currents of 2 mA, 10 mA and 20 mA are common. Sub-mA indicators may be made by driving ultrabright LEDs at very low current. Efficiency tends to reduce at low currents, but indicators running on 100 μA are still practical. The cost of ultrabright LEDs is higher than that of 2 mA indicator LEDs.

Strings of LEDs are normally operated in series LEDs, with the total LED voltage typically adding up to around two-thirds of the supply voltage, with resistor current control for each string. In disposable coin cell powered keyring type LED lights, the resistance of the cell itself is usually the only current limiting device. The cell should not therefore be replaced with a lower resistance type.

LEDs can be purchased with built in series resistors. These can save printed circuit board space and are especially useful when building prototypes or populating a PCB in a way other than its designers intended. However, the resistor value is set at the time of manufacture, removing one of the key methods of setting the LED's intensity. Alphanumeric LEDs use the same drive strategy as indicator LEDs, the only difference being the larger number of channels, each with its own resistor. Seven-segment and starburst LED arrays are available in both common-anode and common-cathode form.

Lighting LEDs on mains

LEDs, by nature, require direct current (DC) with low voltage, as opposed to the mains electricity from the electrical grid which supplies a high voltage with an alternating current (AC).

A CR dropper (capacitor and resistor) followed by full-wave rectification is the usual electrical ballast with series-parallel LED clusters. A single series string minimizes dropper losses, while paralleled strings increase reliability. In practice usually three strings or more are used. An advantage of a capacitor is that it can reduce the high line voltage to an appropriate low voltage, without wasting power, with a very simple circuit; a disadvantage is that there may be a high surge of current for a short time when it is first turned on.

Operation on square wave and modified sine wave (MSW) sources, such as many inverters, causes heavily-increased resistor dissipation in CR dropper, and LED ballasts designed for sine wave use tend to burn on non-sine waveforms. The non-sine waveform also causes high peak LED currents, heavily shortening LED life. An inductor and rectifier make a more suitable ballast for such use, and other options are also possible. Dedicated integrated circuits are available that provide optimal drive for LEDs and maximum overall efficiency.

Multiple LEDs can be connected in series with a single current limiting resistor provided the source voltage is greater than the sum of the individual LED threshold voltages. Parallel operation is also possible but can be more problematic. Parallel LEDs must have closely matched forward voltages (Vf) in order to have equal branch currents and, therefore, equal light output. Variations in the manufacturing process can make it difficult to obtain satisfactory operation when connecting some types of LEDs in parallel.

To increase efficiency (or to allow intensity control without the complexity of a DAC), the power may be applied periodically or intermittently; so long as the flicker rate is

greater than the human flicker fusion threshold, the LED will appear to be continuously lit.

Resonant asymmetric inductive supply (RAIS)

RAIS is an off-line LED driver topology with TRIAC dimmer compatibility and near unity power factor with no loss in efficiency.

This LED driver technology is especially suited to use with domestic TRIAC dimmers. Resonant asymmetric inductive supply (RAIS) sits between the mains and the LED. Report L10270 from the Lighting Association Laboratories found that the technology would work across different dimmer types while still maintaining a power factor of 0.96, and with an input to output system efficiency of 91 per cent. As the light is dimmed, the system is able to hold a significantly higher efficiency than a Buck converter as RAIS draws a continuous current without using bleed resistors. RAIS is a single stage supply that also delivers a constant current output to the LED with no sense resistor. The report also found that RAIS technology fitted into the dimensions of a standard GU10 lampholder (bayonet mount).

LED lighting typically involves LEDs connected in series. The resultant forward voltage may be in the region of 10 to 20V. In such cases, the ratio between the mains voltage and the voltage required to drive the load is between 10 and 20. With such a large ratio, conventional circuits used to drive LEDs become very inefficient because the switching will be operating at extremes of duty ratio with very short conduction times and high peak currents. This inevitably means that extra components, such as a common mode choke, need to be used. It is therefore common to include a magnetic or piezoelectric (ceramic) transformer with an input-to-output ratio suitable to create a step down in voltage and a corresponding step up in current, introducing further efficiency loss, cost and bulk.

In contrast, the RAIS technology can drive high current, low voltage LED strings from a 240V AC mains supply without high peak currents, a transformer, or common mode choke, and still achieve the turns ratio.

Typical Use: Between the mains and the LED within retro fit lamps, such as a GU10, where small size and use with conventional TRIAC dimmers is required.

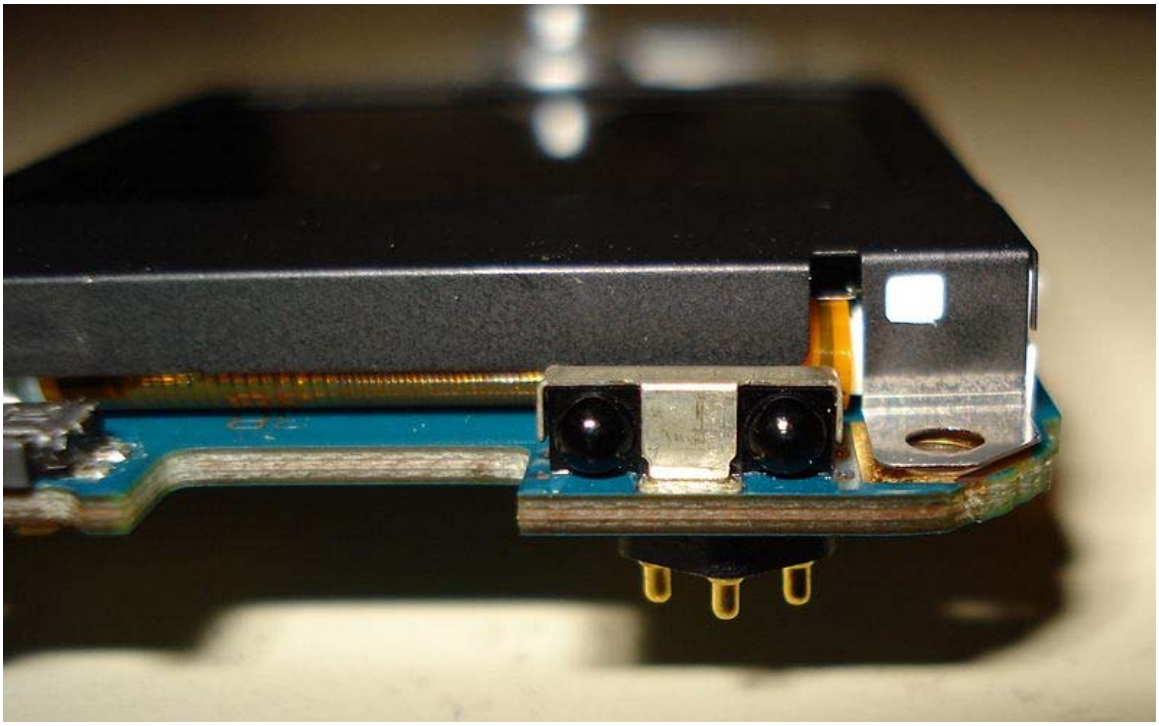
Special Features : Inherent compatibility with a TRIAC dimmers through continuous current draw in the same way as a conventional lamp, in other words it looks like a resistor to the mains. This also has an impact on the efficiency as it does not require a bleeder circuit or holding current resistor to ensure proper TRIAC operation that can cause efficiency loss during dimming.

The topology is inherently constant current. There is no need for a second stage, current sensing, feedback or short circuit protection.

Typical performance of a TRIAC dimmable LED retrofit lamp, Power Factor 0.96, efficiency > 90%.

The RAIS technology has now been granted UK patent # GB2449616 dated 17th February 09 and is applied for in most other world-wide territories.

LED as light sensor



Mobile phone IrDA

An LED can be used as a photodiode used for light detection as well as emission. This capability has been demonstrated and used in a variety of applications including ambient light detection and bidirectional communications. This implementation of LEDs is important because functionality can be added to designs with only minor modifications, usually at little or no cost.

An LED is simply a diode that has been doped specifically for efficient light emission and has been packaged in a transparent case. Therefore, if inserted into a circuit in the same way as a photodiode, which is essentially the same thing, the LED will perform the same function. As a photodiode, it is sensitive to wavelengths equal to or shorter than the predominant wavelength it emits. For example, a green LED will be sensitive to blue light and to some green light, but not to yellow or red light. Additionally, the LED can be multiplexed in such a circuit, such that it can be used for both light emission and sensing at different times.

Several applications for this technology have been suggested and/or implemented, ranging from use as simple ambient light sensors to full bidirectional communications

using a single LED. Most of these applications benefit from this technology because of the cost reduction of using the same component for multiple functions.

Ambient light sensors

LEDs have been used as ambient light sensors. For example, a remote control keypad backlight would be turned on by capacitive proximity sensors only in the absence of ambient light. The LED used for the backlight was also used as the ambient light sensor. This resulted in increased functionality for no increase in manufacturing costs.

Bidirectional communications

LEDs can be used as both emitters and detectors of light, which means that a device having only a single LED can be used to achieve bidirectional communications with another device meeting these requirements. Using this technology, any of the ubiquitous LEDs connected to household appliances, computers and other electronic devices can be used as a bidirectional communications port.

One application for bidirectional communication with a single LED is fiber optic communications. In typical plastic optical fiber communications, a single optical fiber is used only for communication in one direction. This is because a single LED transmitter is placed at one end of the fiber, and a photodiode receiver is placed at the other end. Thus, two fibers are needed for bidirectional communication. However, if a single LED is placed at each end of a fiber, then the optical fiber can carry information in both directions using half the number of components as a typical system. This reduces system weight, cost and complexity.

Another application of this use of LEDs is a proposed alternative to RFID tags called the iDropper, developed by Mitsubishi Electric Research Laboratories in 2003. The iDropper is a small device that consists of a microcontroller, a battery, an LED, and a single push-button. The device records or transmits a small amount of data upon command from the user. Compared to RFID tags, the iDropper is more secure because the user must press a button to reveal personal information, and is similar in cost.

One major limitation of this scheme is the fact that a single LED can only operate as a half-duplex transceiver. A single LED can either transmit or receive information at one time, not both simultaneously. A simple way to put this is that an LED transceiver behaves like a walkie-talkie, in contrast to a telephone. This means that it takes a considerable time for two devices to "talk" to each other.

Chapter- 9

Mechanical Filter

Figure 1. A mechanical filter made by the Kokusai Electric Company intended for selecting the narrow 2 kHz bandwidth signals in SSB radio receivers. It operates at 455 kHz, a common IF for these receivers, and is dimensioned 45×15×15 mm ($1\,{}^{3}\!/_{4}\times{}^{7}\!/_{12}\times{}^{7}\!/_{12}$ in).

A **mechanical filter** is a signal processing filter usually used in place of an electronic filter at radio frequencies. Its purpose is the same as that of a normal electronic filter: to pass a range of signal frequencies, but to block others. The filter acts on mechanical vibrations which are the analogue of the electrical signal. At the input and output of the filter there are transducers which convert the electrical signal into, and then back from, these mechanical vibrations.

The components of a mechanical filter are all directly analogous to the various elements found in electrical circuits. The mechanical elements obey mathematical functions which are identical to their corresponding electrical elements. This makes it possible to apply electrical network analysis and filter design methods to mechanical filters. Electrical theory has developed a large library of mathematical forms that produce useful filter

frequency responses and the mechanical filter designer is able to make direct use of these. It is only necessary to set the mechanical components to appropriate values to produce a filter with an identical response to the electrical counterpart.

Steel and nickel–iron alloys are common materials for mechanical filter components; nickel is sometimes used for the input and output couplings. Resonators in the filter made from these materials need to be machined to precisely adjust their resonance frequency before final assembly.

While the meaning of *mechanical filter* here is one that is used in an electromechanical role, it is fully possible to use a mechanical design to filter mechanical vibrations or sound waves (which are also essentially mechanical) directly. For example, filtering of audio frequency response in the design of loudspeaker cabinets can be achieved with mechanical components. In the electrical application, in addition to mechanical components which correspond to their electrical counterparts, transducers are needed to convert between the mechanical and electrical domains. There are a wide variety of component forms and topologies for mechanical filters, a representative selection of which are presented here.

The theory of mechanical filters was first applied to improving the mechanical parts of phonographs in the 1920s. By the 1950s mechanical filters were being manufactured as self-contained components for applications in radio transmitters and high-end receivers. The high "quality factor", Q, that mechanical resonators can attain, far higher than that of an all-electrical LC circuit, made possible the construction of mechanical filters with excellent selectivity. Good selectivity, being important in radio receivers, made such filters highly attractive. Contemporary researchers are working on microelectromechanical filters, the mechanical devices corresponding to electronic integrated circuits.

Elements

The elements of a passive linear electrical network consist of inductors, capacitors and resistors which have the properties of inductance, elastance (inverse capacitance) and resistance, respectively. The mechanical counterparts of these properties are, respectively, mass, stiffness and damping. In most electronic filter designs, only inductor and capacitor elements are used in the body of the filter (although the filter may be terminated with resistors at the input and output). Resistances are not present in a theoretical filter composed of ideal components and only arise in practical designs as unwanted parasitic elements. Likewise, a mechanical filter would ideally consist only of components with the properties of mass and stiffness, but in reality some damping is present as well.

The mechanical counterparts of voltage and electric current in this type of analysis are, respectively, force (F) and velocity (v) and represent the signal waveforms. From this, a mechanical impedance can be defined in terms of the imaginary angular frequency, $j\omega$, which entirely follows the electrical analogy.

Mechanical element	Formula (in one dimension)	Mechanical impedance	Electrical counterpart
Stiffness, S	$S = \frac{F}{x}$	$Z = \frac{S}{j\omega}$	Elastance, $1/C$, the inverse of capacitance
Mass, M	$M = \frac{F}{dv/dt} = \frac{F}{a}$	$Z = j\omega M$	Inductance, L
Damping, D	$D = \frac{F}{v}$	$Z = D$	Resistance, R

Notes:

- The symbols x, t, and a represent their usual quantities; distance, time, and acceleration respectively.
- The mechanical quantity *compliance*, which is the inverse of stiffness, can be used instead of stiffness to give a more direct correspondence to capacitance, but stiffness is used in the table as the more familiar quantity.

The scheme presented in the table is known as the impedance analogy. Circuit diagrams produced using this analogy match the electrical impedance of the mechanical system seen by the electrical circuit, making it intuitive from an electrical engineering standpoint. There is also the mobility analogy, in which force corresponds to current and velocity corresponds to voltage. This has equally valid results but requires using the reciprocals of the electrical counterparts listed above. Hence, $M \rightarrow C$, $S \rightarrow 1/L$, $D \rightarrow G$ where G is electrical conductance, the inverse of resistance. Equivalent circuits produced by this scheme are similar, but are the dual impedance forms whereby series elements become parallel, capacitors become inductors, and so on. Circuit diagrams using the mobility analogy more closely match the mechanical arrangement of the circuit, making it more intuitive from a mechanical engineering standpoint. In addition to their application to electromechanical systems, these analogies are widely used to aid analysis in acoustics.

Any mechanical component will unavoidably possess both mass and stiffness. This translates in electrical terms to an LC circuit, that is, a circuit consisting of an inductor and a capacitor, hence mechanical components are resonators and are often used as such. It is still possible to represent inductors and capacitors as individual lumped elements in a mechanical implementation by minimising (but never quite eliminating) the unwanted property. Capacitors may be made of thin, long rods, that is, the mass is minimised and the compliance is maximised. Inductors, on the other hand, may be made of short, wide pieces which maximise the mass in comparison to the compliance of the piece.

Mechanical parts act as a transmission line for mechanical vibrations. If the wavelength is short in comparison to the part then a lumped element model as described above is no longer adequate and a distributed element model must be used instead. The mechanical distributed elements are entirely analogous to electrical distributed elements and the

mechanical filter designer can use the methods of electrical distributed element filter design.

History

Harmonic telegraph

Mechanical filter design was developed by applying the discoveries made in electrical filter theory to mechanics. However, a very early example (1870s) of acoustic filtering was the "harmonic telegraph", which arose precisely because electrical resonance was poorly understood but mechanical resonance (in particular, acoustic resonance) was very familiar to engineers. This situation was not to last for long; electrical resonance had been known to science for some time before this, and it was not long before engineers started to produce all-electric designs for filters. In its time, though, the harmonic telegraph was of some importance. The idea was to combine several telegraph signals on one telegraph line by what would now be called frequency division multiplexing thus saving enormously on line installation costs. The key of each operator activated a vibrating electromechanical reed which converted this vibration into an electrical signal. Filtering at the receiving operator was achieved by a similar reed tuned to precisely the same frequency, which would only vibrate and produce a sound from transmissions by the operator with the identical tuning.

Versions of the harmonic telegraph were developed by Elisha Gray, Alexander Graham Bell, Ernest Mercadier and others. Its ability to act as a sound transducer to and from the electrical domain was to inspire the invention of the telephone.

Mechanical equivalent circuits

Once the basics of electrical network analysis began to be established, it was not long before the ideas of complex impedance and filter design theories were carried over into mechanics by analogy. Kennelly, who was also responsible for introducing complex impedance, and Webster were the first to extend the concept of impedance into mechanical systems in 1920. Mechanical admittance and the associated mobility analogy came much later and are due to Firestone in 1932.

It was not enough to just develop a mechanical analogy. This could be applied to problems that were entirely in the mechanical domain, but for mechanical filters with an electrical application it is necessary to include the transducer in the analogy as well. Poincaré in 1907 was the first to describe a transducer as a pair of linear algebraic equations relating electrical variables (voltage and current) to mechanical variables (force and velocity). These equations can be expressed as a matrix relationship in much the same way as the z-parameters of a two-port network in electrical theory, to which this is entirely analogous:

$$\begin{bmatrix} V \\ F \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \end{bmatrix} \begin{bmatrix} I \\ v \end{bmatrix}$$

where V and I represent the voltage and current respectively on the electrical side of the transducer.

Wegel, in 1921, was the first to express these equations in terms of mechanical impedance as well as electrical impedance. The element $z_{22}$is the open circuit mechanical impedance, that is, the impedance presented by the mechanical side of the transducer when no current is entering the electrical side. The element z_{11}, conversely, is the clamped electrical impedance, that is, the impedance presented to the electrical side when the mechanical side is clamped and prevented from moving (velocity is zero). The remaining two elements, $z_{21}$and z_{12}, describe the transducer forward and reverse transfer functions respectively. Once these ideas were in place, engineers were able to extend electrical theory into the mechanical domain and analyse an electromechanical system as a unified whole.

Sound reproduction

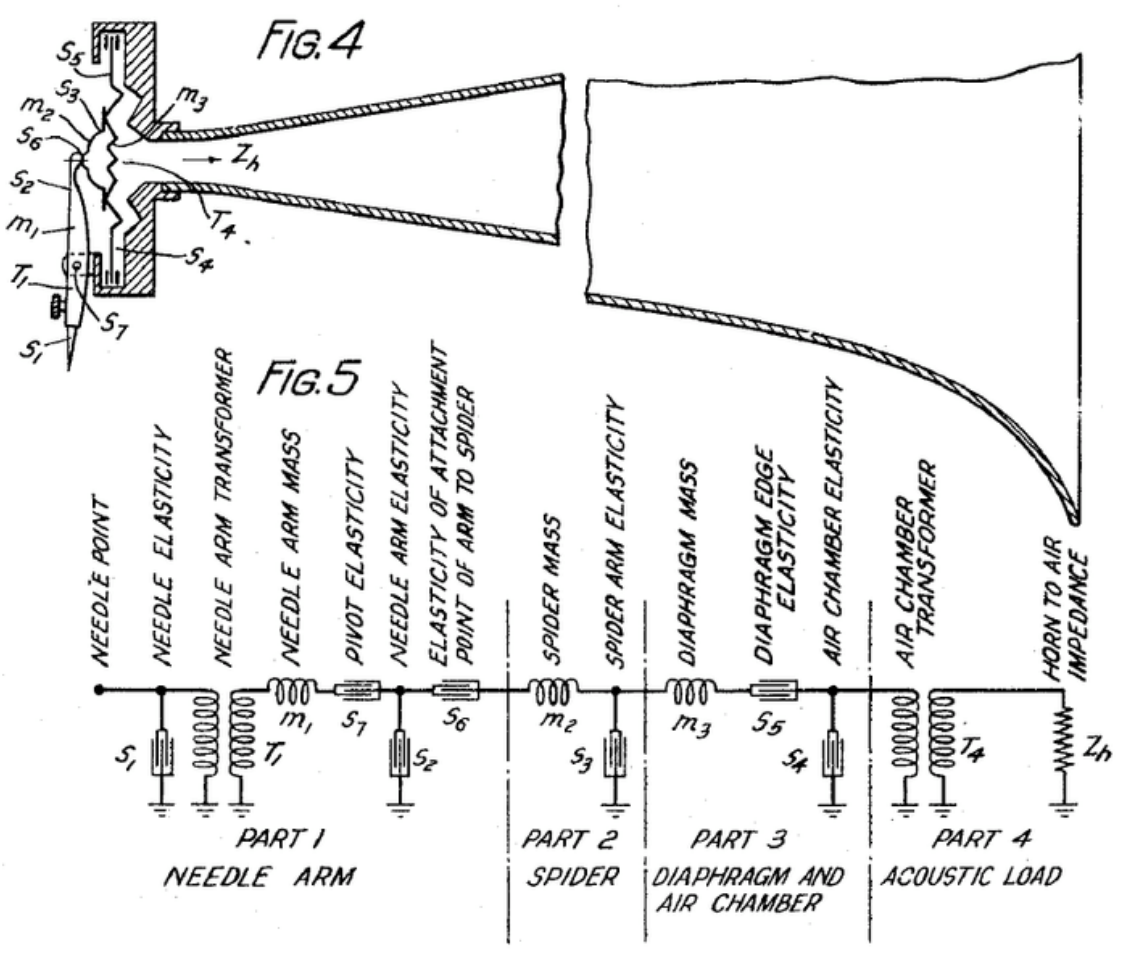


Figure 2. Harrison's phonograph mechanism and its electrical equivalent circuit.

An early application of these new theoretical tools was in phonographic sound reproduction. A recurring problem with early phonograph designs was that mechanical resonances in the pickup and sound transmission mechanism caused excessively large peaks and troughs in the frequency response, resulting in poor sound quality. In 1923, Harrison of the Western Electric Company filed a patent for a phonograph in which the mechanical design was entirely represented as an electrical circuit. The horn of the phonograph is represented as a transmission line, and is a resistive load for the rest of the circuit, while all the mechanical and acoustic parts—from the pickup needle through to the horn—are translated into lumped components according to the impedance analogy. The circuit arrived at is a ladder topology of series resonant circuits coupled by shunt capacitors. This can be viewed as a bandpass filter circuit. Harrison designed the component values of this filter to have a specific passband corresponding to the desired audio passband (in this case 100 Hz to 6 kHz) and a flat response. Translating these electrical element values back into mechanical quantities provided specifications for the mechanical components in terms of mass and stiffness, which in turn could be translated into physical dimensions for their manufacture. The resulting phonograph has a flat frequency response in its passband and is free of the resonances previously experienced. Shortly after this, Harrison filed another patent using the same methodology on telephone transmit and receive transducers.

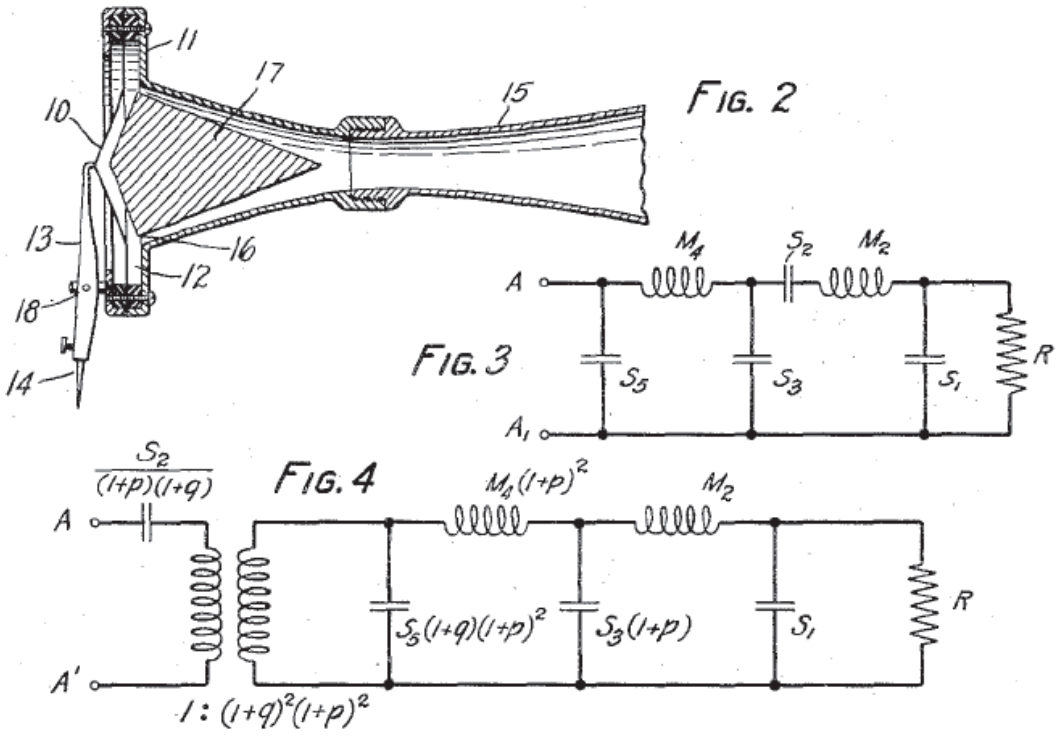


Figure 3. Norton's mechanical filter together with its electrical equivalent circuit.

Harrison used Campbell's image filter theory, which was the most advanced filter theory available at the time. In this theory, filter design is viewed essentially as an impedance

matching problem. More advanced filter theory was brought to bear on this problem by Norton in 1929 at Bell Labs. Norton followed the same general approach though he later described to Darlington the filter he designed as being "maximally flat". Norton's mechanical design predates the paper by Butterworth who is usually credited as the first to describe the electronic maximally flat filter. The equations Norton gives for his filter correspond to a singly terminated Butterworth filter, that is, one driven by an ideal voltage source with no impedance, whereas the form more usually given in texts is for the doubly terminated filter with resistors at both ends, making it hard to recognise the design for what it is. Another unusual feature of Norton's filter design arises from the series capacitor, which represents the stiffness of the diaphragm. This is the only series capacitor in Norton's representation, and without it, the filter could be analysed as a low-pass prototype. Norton moves the capacitor out of the body of the filter to the input at the expense of introducing a transformer into the equivalent circuit (Norton's figure 4). Norton has used here the "turning round the L" impedance transform to achieve this.

The definitive description of the subject from this period is Maxfield and Harrison's 1926 paper. There, they describe not only how mechanical bandpass filters can be applied to sound reproduction systems, but also apply the same principles to recording systems and describe a much improved disc cutting head.

Volume production

The first volume production of mechanical filters was undertaken by Collins Radio Company starting in the 1950s. These were originally designed for telephone frequency-division multiplex applications where there is commercial advantage in using high quality filters. Precision and steepness of the transition band leads to a reduced width of guard band, which in turn leads to the ability to squeeze more telephone channels into the same cable. This same feature is useful in radio transmitters for much the same reason. Mechanical filters quickly also found popularity in VHF/UHF radio intermediate frequency (IF) stages of the high end radio sets (military, marine, amateur radio and the like) manufactured by Collins. They were favoured in the radio application because they could achieve much higher Q-factors than the equivalent *LC* filter. High *Q* allows filters to be designed which have high selectivity, important for distinguishing adjacent radio channels in receivers. They also had an advantage in stability over both *LC* filters and monolithic crystal filters. The most popular design for radio applications was torsional resonators because radio IF typically lies in the 100 to 500 kHz band.

Transducers

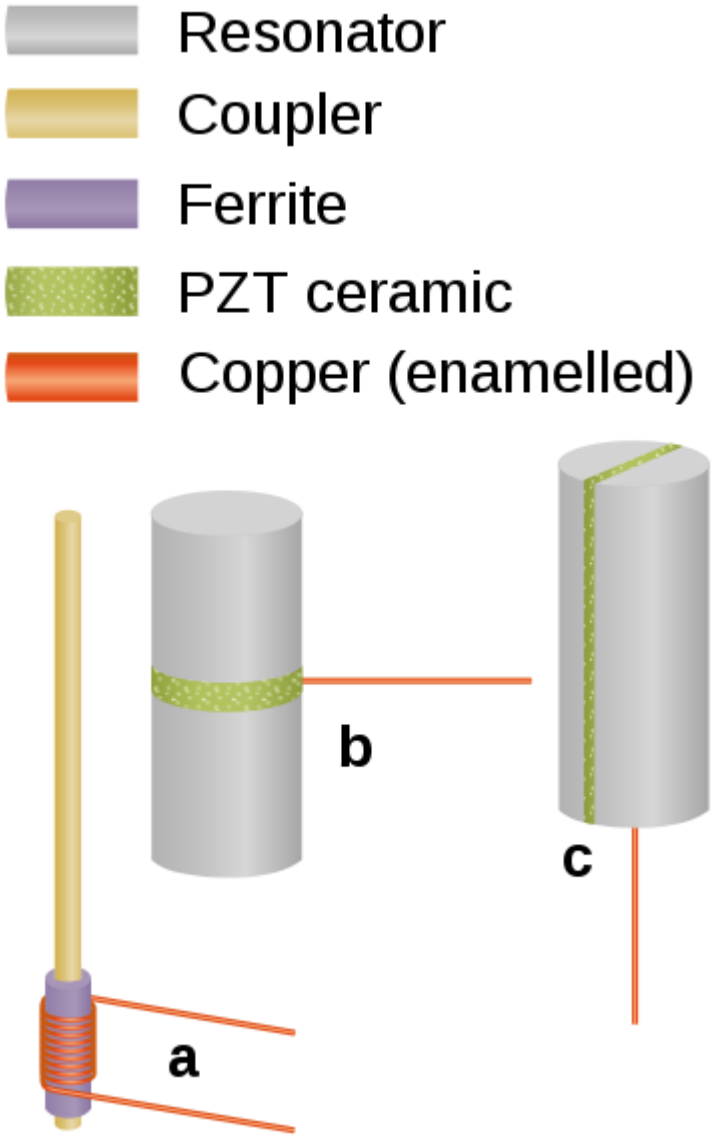


Figure 4. Mechanical filter transducers. **a** magnetostrictive transducer. **b** Langevin type piezoelectric transducer. **c** torsional piezoelectric transducer.

There are two general types of transducers used with mechanical filters: magnetostrictive and piezoelectric. Piezoelectric is favoured in more recent designs since the piezoelectric material can also be used as one of the resonators of the filter, thus reducing the number of components and thereby saving space. They also avoid the susceptibility to extraneous magnetic fields from which the magnetostrictive type suffers.

Magnetostrictive

A magnetostrictive material is one which changes shape when a magnetic field is applied. In reverse, it produces a magnetic field when distorted. The magnetostrictive transducer requires a coil of conducting wire around the magnetostrictive material. The coil either induces a magnetic field in the transducer and sets it in motion or else picks up an induced current from the motion of the transducer at the filter output. It is also usually necessary to have a small magnet to bias the magnetostrictive material into its operating range. It is possible to dispense with the magnets if the biasing is taken care of on the electronic side by providing a d.c. current superimposed on the signal, but this approach would detract from the generality of the filter design.

The usual magnetostrictive materials used for the transducer are either ferrite or compressed powdered iron. Mechanical filter designs often have the resonators coupled with steel or nickel-iron wires, but on some designs, especially older ones, nickel wire may be used for the input and output rods. This is because it is possible to wind the transducer coil directly on to a nickel coupling wire since nickel is slightly magnetostrictive. However, it is not strongly so and coupling to the electrical circuit is weak. This scheme also has the disadvantage that there are no measures taken to prevent eddy currents, a problem that is avoided if ferrites are used instead of nickel.

The coil of the transducer adds some inductance on the electrical side of the filter. It is common practice to add a capacitor in parallel with the coil so that an additional resonator is formed which can be incorporated into the filter design. While this will not improve performance to the extent that an additional mechanical resonator would, there is some benefit and the coil has to be there in any case.

Piezoelectric

A piezoelectric material is one which changes shape when an electric field is applied. In reverse, it produces an electric field when it is distorted. A piezoelectric transducer, in essence, is made simply by plating electrodes on to the piezoelectric material. Early piezoelectric materials used in transducers such as barium titanate had poor temperature stability. This precluded the transducer from functioning as one of the resonators; it had to be a separate component. This problem was solved with the introduction of lead zirconate titanate (abbreviated PZT) which is stable enough to be used as a resonator. Another common piezoelectric material is quartz, which has also been used in mechanical filters. However, ceramic materials such as PZT are preferred for their greater electromechanical coupling coefficient.

One type of piezoelectric transducer is the Langevin type, named after a transducer used by Paul Langevin in early sonar research. This is good for longitudinal modes of vibration. It can also be used on resonators with other modes of vibration if the motion can be mechanically converted into a longitudinal motion. The transducer consists of a layer of piezoelectric material sandwiched transversally into a coupling rod or resonator.

Another kind of piezoelectric transducer has the piezoelectric material sandwiched in longitudinally, usually into the resonator itself. This kind is good for torsional vibration modes and is called a torsional transducer.

Resonators

Material	Q-factor
Nickel	several 100
Steel	several 1000
Aluminium	~10,000
Nickel-iron alloy	10,000 to 25,000 depending on composition

It is possible to achieve an extremely high *Q* with mechanical resonators. Mechanical resonators typically have a *Q* of 10,000 or so, and 25,000 can be achieved in torsional resonators using a particular nickel-iron alloy. This is an unreasonably high figure to achieve with LC circuits, whose *Q* is limited by the resistance of the inductor coils.

Early designs in the 1940s and 1950s started by using steel as a resonator material. This has given way to nickel-iron alloys, primarily to maximise the *Q* since this is often the primary appeal of mechanical filters rather than price. Some of the metals that have been used for mechanical filter resonators and their *Q* are shown in the table.

Piezoelectric crystals are also sometimes used in mechanical filter designs. This is especially true for resonators that are also acting as transducers for inputs and outputs.

One advantage that mechanical filters have over LC electrical filters is that they can be made very stable. The resonance frequency can be made so stable that it varies only 1.5 parts per billion (ppb) from the specified value over the operating temperature range (−25 to 85 °C), and its average drift with time can be as low as 4 ppb per day. This stability with temperature is another reason for using nickel-iron as the resonator material. Variations with temperature in the resonance frequency (and other features of the frequency function) are directly related to variations in the Young's modulus, which is a measure of stiffness of the material. Materials are therefore sought that have a small temperature coefficient of Young's modulus. In general, Young's modulus has a negative temperature coefficient (materials become less stiff with increasing temperature) but additions of small amounts of certain other elements in the alloy can produce a material with a temperature coefficient that changes sign from negative through zero to positive with temperature. Such a material will have a zero coefficient of temperature with resonance frequency around a particular temperature. It is possible to adjust the point of zero temperature coefficient to a desired position by heat treatment of the alloy.

Resonator modes

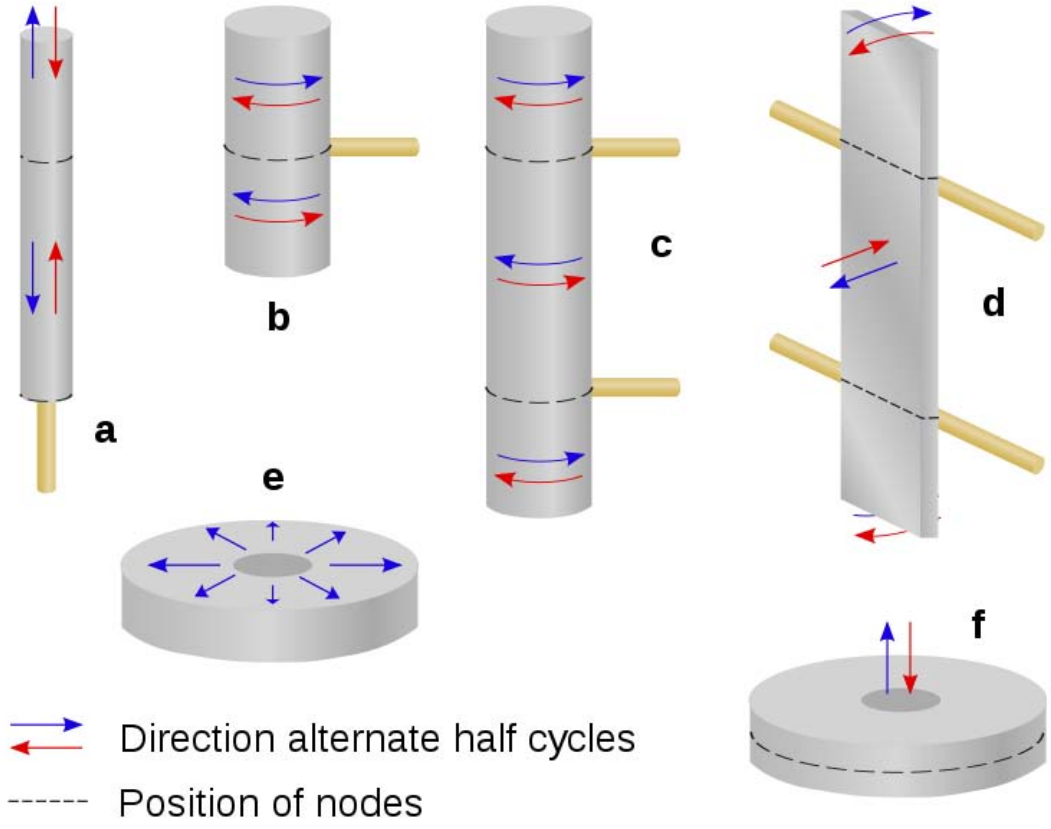


Figure 5. Some possible vibrational modes of resonators

It is usually possible for a mechanical part to vibrate in a number of different modes, however the design will be based on a particular vibrational mode and the designer will take steps to try to restrict the resonance to this mode. As well as the straightforward longitudinal mode some others which are used include flexural mode, torsional mode, radial mode and drumhead mode.

Modes are numbered according to the number of half-wavelengths in the vibration. Some modes exhibit vibrations in more than one direction (such as drumhead mode which has two) and consequently the mode number consists of more than one number. When the vibration is in one of the higher modes, there will be multiple nodes on the resonator where there is no motion. For some types of resonator, this can provide a convenient place to make a mechanical attachment for structural support. Wires attached at nodes will have no effect on the vibration of the resonator or the overall filter response. In figure 5, some possible anchor points are shown as wires attached at the nodes. The modes shown are (5a) the second longitudinal mode fixed at one end, (5b) the first torsional mode, (5c) the second torsional mode, (5d) the second flexural mode, (5e) first radial expansion mode and (5f) first radially symmetric drumhead mode.

Circuit designs

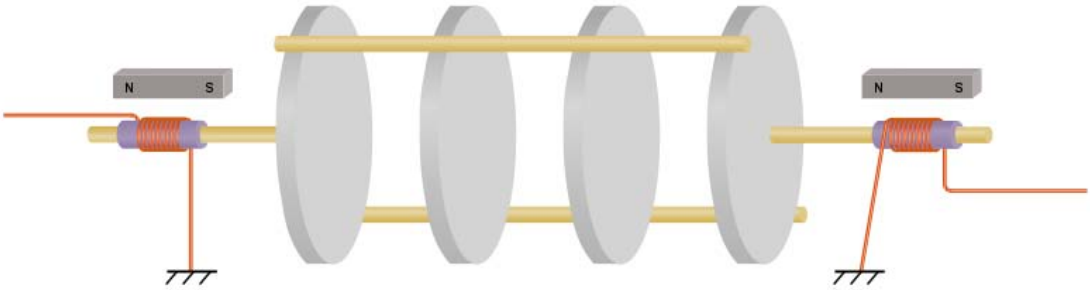


Figure 6. A mechanical filter using disc flexural resonators and magnetostrictive transducers

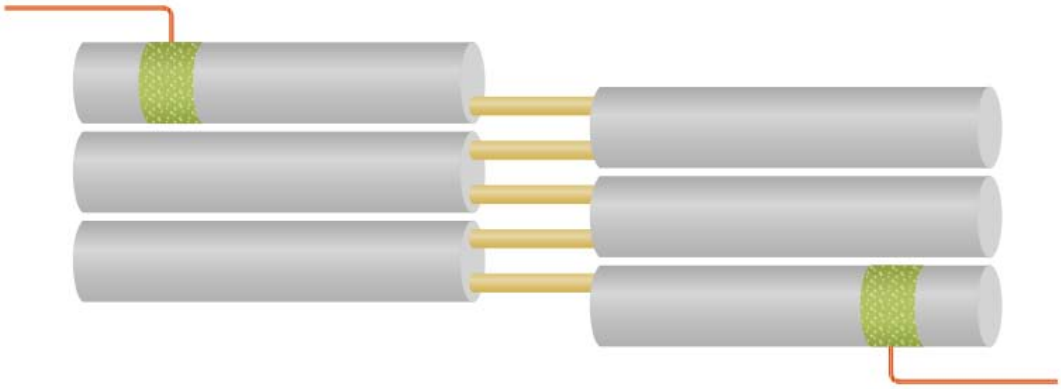


Figure 7. A filter using longitudinal resonators and Langevin type transducers

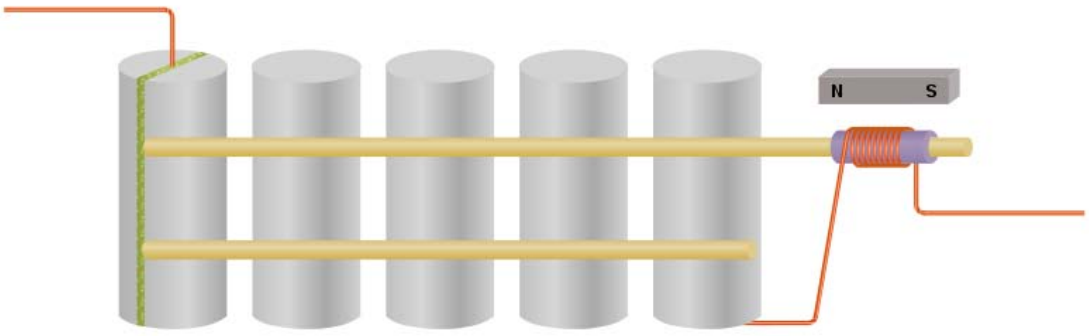


Figure 8a. A filter using torsional resonators. The input is shown with a torsional piezoelectric transducer and the output has a magnetostrictive transducer.

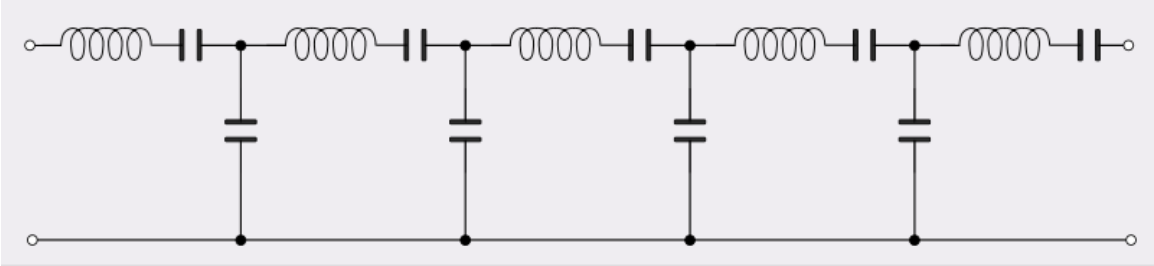


Figure 8b. Equivalent circuit of the torsional resonator circuit above

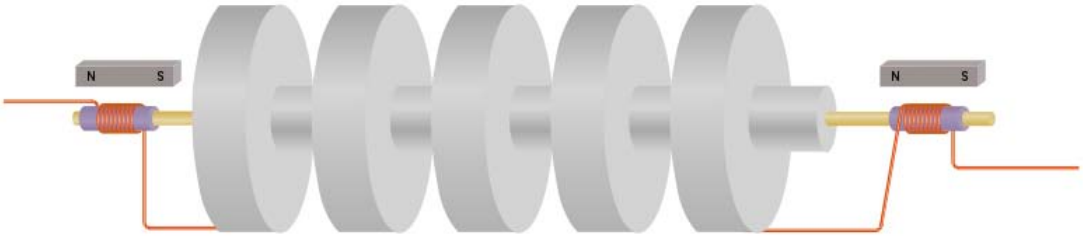


Figure 9. A filter using disc drumhead resonators

There are a great many combinations of resonators and transducers that can be used to construct a mechanical filter. A selection of some of these is shown in the diagrams. Figure 6 shows a filter using disc flexural resonators and magnetostrictive transducers. The transducer drives the centre of the first resonator, causing it to vibrate. The edges of the disc move in antiphase to the centre when the driving signal is at, or close to, resonance, and the signal is transmitted through the connecting rods to the next resonator. When the driving signal is not close to resonance, there is little movement at the edges, and the filter rejects (does not pass) the signal. Figure 7 shows a similar idea involving longitudinal resonators connected together in a chain by connecting rods. In this diagram, the filter is driven by piezoelectric transducers. It could equally well have used magnetostrictive transducers. Figure 8 shows a filter using torsional resonators. In this diagram, the input has a torsional piezoelectric transducer and the output has a magnetostrictive transducer. This would be quite unusual in a real design, as both input and output usually have the same type of transducer. The magnetostrictive transducer is only shown here to demonstrate how longitudinal vibrations may be converted to torsional vibrations and vice versa. Figure 9 shows a filter using drumhead mode resonators. The edges of the discs are fixed to the casing of the filter (not shown in the diagram) so the vibration of the disc is in the same modes as the membrane of a drum. Collins calls this type of filter a disc wire filter.

The various types of resonator are all particularly suited to different frequency bands. Overall, mechanical filters with lumped elements of all kinds can cover frequencies from about 5 to 700 kHz although mechanical filters down as low as a few kilohertz (kHz) are rare. The lower part of this range, below 100 kHz, is best covered with bar flexural

resonators. The upper part is better done with torsional resonators. Drumhead disc resonators are in the middle, covering the range from around 100 to 300 kHz.

The frequency response behaviour of all mechanical filters can be expressed as an equivalent electrical circuit using the impedance analogy described above. An example of this is shown in figure 8b which is the equivalent circuit of the mechanical filter of figure 8a. Elements on the electrical side, such as the inductance of the magnetostrictive transducer, are omitted but would be taken into account in a complete design. The series resonant circuits on the circuit diagram represent the torsional resonators, and the shunt capacitors represent the coupling wires. The component values of the electrical equivalent circuit can be adjusted, more or less at will, by modifying the dimensions of the mechanical components. In this way, all the theoretical tools of electrical analysis and filter design can be brought to bear on the mechanical design. Any filter realisable in electrical theory can, in principle, also be realised as a mechanical filter. In particular, the popular finite element approximations to an ideal filter response of the Butterworth and Chebyshev filters can both readily be realised. As with the electrical counterpart, the more elements that are used, the closer the approximation approaches the ideal, however, for practical reasons the number of resonators does not normally exceed eight.

Semi-lumped designs

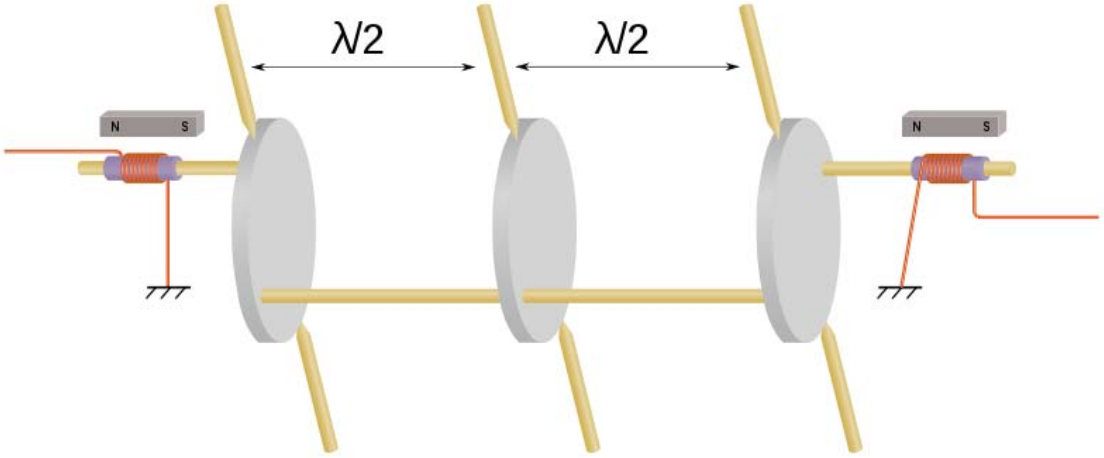


Figure 10a. A semi-lumped design using disc flexural resonators and λ/2 coupling wires

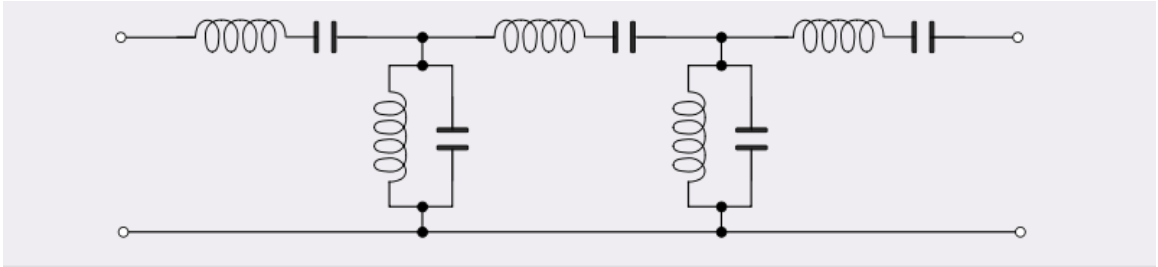


Figure 10b. Equivalent circuit of the semi-lumped circuit above

Frequencies of the order of megahertz (MHz) are above the usual range for mechanical filters. The components start to become very small, or alternatively the components are large compared to the signal wavelength. The lumped element model described above starts to break down and the components must be considered as distributed elements. The frequency at which the transition from lumped to distributed models takes place is much lower for mechanical filters than it is for their electrical counterparts. This is because mechanical vibrations travel at the speed of sound for the material the component is composed of. For solid components, this is many times (x15 for nickel-iron) the speed of sound in air (343 m/s) but still considerably less than the speed of electromagnetic waves (approx. $3x10^8$ m/s in vacuum). Consequently, mechanical wavelengths are much shorter than electrical wavelengths for the same frequency. Advantage can be taken of these effects by deliberately designing components to be distributed elements, and the components and methods used in electrical distributed element filters can be brought to bear. The equivalents of stubs and impedance transformers are both achievable. Designs which use a mixture of lumped and distributed elements are referred to as semi-lumped.

An example of such a design is shown in figure 10a. The resonators are disc flexural resonators similar to those shown in figure 6, except that these are energised from an edge, leading to vibration in the fundamental flexural mode with a node in the centre, whereas the top diagram design is energised in the centre leading to vibration in the second flexural mode at resonance. The resonators are mechanically attached to the housing by pivots at right angles to the coupling wires. The pivots are to ensure free turning of the resonator and minimise losses. The resonators are treated as lumped elements; however, the coupling wires are made exactly one half-wavelength ($\lambda/2$) long and are equivalent to a $\lambda/2$ open circuit stub in the electrical equivalent circuit. For a narrow-band filter, a stub of this sort has the approximate equivalent circuit of a parallel shunt tuned circuit as shown in figure 10b. Consequently, the connecting wires are being used in this design to add additional resonators into the circuit and will have a better response than one with just the lumped resonators and short couplings. For even higher frequencies, microelectromechanical methods can be used as described below.

Bridging wires

Bridging wires are rods that couple together resonators that are not adjacent. They can be used to produce poles of attenuation in the stopband. This has the benefit of increasing the stopband rejection. When the pole is placed near the passband edge, it also has the benefit of increasing roll-off and narrowing the transition band. The typical effects of some of these on filter frequency response are shown in figure 11. Bridging across a single resonator (figure 11b) can produce a pole of attenuation in the high stopband. Bridging across two resonators (figure 11c) can produce a pole of attenuation in both the high and the low stopband. Using multiple bridges (figure 11d) will result in multiple poles of attenuation. In this way, the attenuation of the stopbands can be deepened over a broad frequency range.

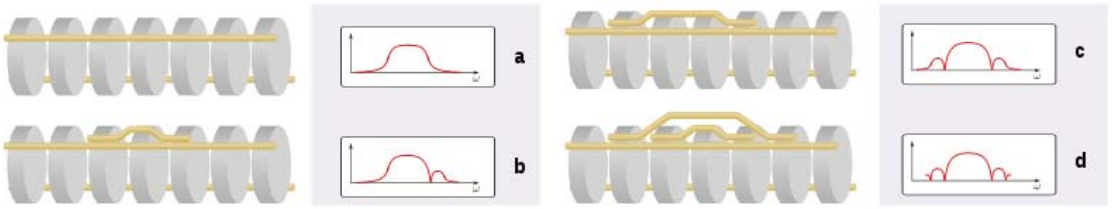


Figure 11. Schematic bridging arrangements and their effect on frequency response.

The method of coupling between non-adjacent resonators is not limited to mechanical filters. It can be applied to other filter formats. For instance, channels can be cut between cavity resonators, mutual inductance can be used with discrete component filters, and feedback paths can be used with active analogue or digital filters. Nor was the method first discovered in the field of mechanical filters; the earliest description is in a 1948 patent for filters using microwave cavity resonators. However, mechanical filter designers were the first (1960s) to develop practical filters of this kind and the method became a particular feature of mechanical filters.

Microelectromechanical filters

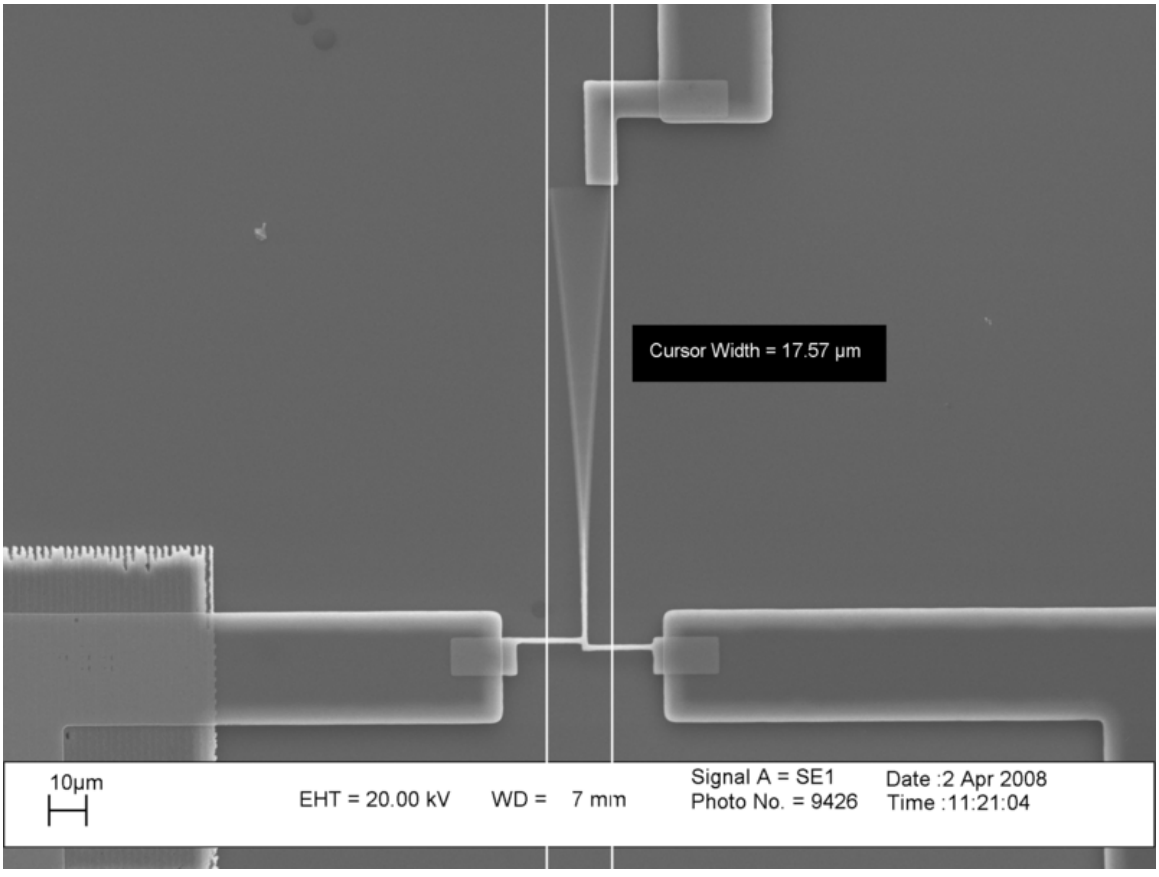


Figure 12. MEMS cantilever resonator. The device can be seen to be vibrating in this picture.

A new technology emerging in mechanical filtering is microelectromechanical systems (MEMS). MEMS are very small micromachines with component sizes measured in micrometres (μm), but not as small as nanomachines. These systems are mostly fabricated from silicon (Si), silicon nitride (Si_3N_4), or polymers. A common component used for radio frequency filtering (and MEMS applications generally), is the cantilever resonator. Cantilevers are simple mechanical components to manufacture by much the same methods used by the semiconductor industry; masking, photolithography and etching, with a final undercutting etch to separate the cantilever from the substrate. The technology has great promise since cantilevers can be produced in large numbers on a single substrate—much as large numbers of transistors are currently contained on a single silicon chip.

The resonator shown in figure 12 is around 120 μm in length. Experimental complete filters with an operating frequency of 30 GHz have been produced using cantilever varactors as the resonator elements. The size of this filter is around 4×3.5 mm. Cantilever resonators are typically applied at frequencies below 200 MHz, but other structures, such as micro-machined cavities, can be used in the microwave bands. Extremely high *Q* resonators can be made with this technology; flexural mode resonators with a *Q* in excess of 80,000 at 8 MHz are reported.

Adjustment

The precision applications in which mechanical filters are used require that the resonators are accurately adjusted to the specified resonance frequency. This is known as *trimming* and usually involves a mechanical machining process. In most filter designs, this can be difficult to do once the resonators have been assembled into the complete filter so the resonators are trimmed before assembly. Trimming is done in at least two stages; coarse and fine, with each stage bringing the resonance frequency closer to the specified value. Most trimming methods involve removing material from the resonator which will increase the resonance frequency. The target frequency for a coarse trimming stage consequently needs to be set below the final frequency since the tolerances of the process could otherwise result in a frequency higher than the following fine trimming stage could adjust for.

The coarsest method of trimming is grinding of the main resonating surface of the resonator; this process has an accuracy of around ±800 ppm. Better control can be achieved by grinding the edge of the resonator instead of the main surface. This has a less dramatic effect and consequently better accuracy. Processes that can be used for fine trimming, in order of increasing accuracy, are sandblasting, drilling, and laser ablation. Laser trimming is capable of achieving an accuracy of ±40 ppm.

Trimming by hand, rather than machine, was used on some early production components but would now normally only be encountered during product development. Methods available include sanding and filing. It is also possible to add material to the resonator by hand, thus reducing the resonance frequency. One such method is to add solder, but this

is not suitable for production use since the solder will tend to reduce the high Q of the resonator.

In the case of MEMS filters, it is not possible to trim the resonators outside of the filter because of the integrated nature of the device construction. However, trimming is still a requirement in many MEMS applications. Laser ablation can be used for this but material deposition methods are available as well as material removal. These methods include laser or ion-beam induced deposition.