AC electrical theory

An introduction to phasors, impedance and admittance, with emphasis on radio frequencies.

By David W. Knight*

Version 0.12. 18th Dec. 2015.

* Ottery St Mary, Devon, England.

Please check the author's website to ensure that you have the most recent versions of this article and its associated documents: **http://www.g3ynh.info/**

Recent changes (0.11→0.12): Grammatical corrections. More whitespace. Minimum linewidth in diagrams increased to 2px to improve rendering in online pdf viewers. Adoption of SI notation guidelines.

Table of Contents

Preface

This document provides an introduction to the subject of AC circuit analysis, with particular emphasis on radio-frequency applications. It was developed as part of a collection of writings on the subject of radio-frequency impedance matching and measurement; which were first made available via the Internet in 2005, and were grouped under the working title 'From Transmitter to Antenna'. Its purpose was (and still is) to widen the audience for the other articles by providing essential background material, but it can just as well be read by those who have a more general interest.

The approach adopted is that of starting with the basic laws of DC electricity and expanding them to deal with AC. The modified laws are then used to derive and explore results that are normally accepted without proof, thereby explaining the origins of various standard formulae and demonstrating the general method by which linear circuit design equations are obtained. The level of treatment is one that does not demand a high level of mathematical skill at the outset; because the required techniques are introduced as the narrative progresses. Hence the discussion should be accessible to anyone who has some knowledge of basic algebra and is reasonably familiar with circuit diagrams and electrical terminology.

Apart from providing a conventional introduction to AC theory however, there is a subtext. This relates to the author's concerns as a scientist: one being that there appears to be an almost universal public misconception regarding the nature of electricity; and the other being a lack of mathematical rigour in the way in which phasor techniques are commonly used. Both of these issues can (and should) be addressed at this stage in the development of working knowledge, and so the accompanying discussion attempts to do that.

As all experienced engineers and physicists know, our understanding of electricity comes from Maxwell's equations. The problem for those who wish to teach electrical subjects however, is that the electromagnetic field approach requires advanced mathematics and does not lead directly to the practicalities of circuit design. Hence it is sensible to hold back on the more abstract ideas until they become unavoidable; but that leaves the problem of how to dispel the notion that electricity is synonymous with electrons flowing through wires. The author's solution is to provide an extended preamble; which gives a purely qualitative explanation of electricity in terms of fields; and is intended to leave the reader with the same mental picture as will be held by those who are familiar with Maxwell's theory.

On the matter of mathematical rigour; it is not the intention to wrap the subject in formalism, but merely to eliminate certain bad practices. To this end, we pay particular attention to the definitions and properties of the mathematical objects involved, and develop a way of working that identifies and preserves the algebraic signs of the circuit parameters. The outcome is an internally-consistent theory of circuits, which produces results that are correct in both magnitude and phase. In this way, we eliminate the need for the so-called 'physical considerations' traditionally used to resolve ambiguities; and we discover which of the two commonly used definitions of admittance is actually the correct one.

Some, of course, will ask: 'why should we bother to learn phasor analysis when we can model circuits using SPICE?' In fact, the use of SPICE is highly recommended; but simulation is essentially a way of checking existing design work. It does not offer a systematic approach to the business of optimising circuits or inventing new ones. The techniques discussed here, on the other hand, allow the equations describing the behaviour of a circuit to be written down explicitly. We might then, for example, separate out the terms describing unwanted behaviour with a view to making alterations that will eliminate them. Such is the basis for the development of precision measuring instruments and all manner of other high-performance circuitry.

David Knight. June 2012.

1. Field electricity

When AC theory is introduced, and especially when there is a bias towards radio frequencies, the very first new idea required (by many people at least) is a correct understanding of the word 'electricity'. The teaching of basic science often involves what are known as 'lies to children', and the one about electricity being "electrons flowing through wires" is an intellectual dead-end. Electricity is actually an invisible form of light. Specifically, it is electromagnetic energy of very long wavelength (in comparison to visible light); which is why we can build devices called 'radio transmitters' that cause electricity to propagate off into space. Hence we will never understand AC electricity by counting electrons; and we must first refine our ideas of voltage and current by thinking about certain mysterious entities known as 'fields'.

The term 'field' has a vernacular meaning: "region of influence", and this is the route by which it came into the language of physics. In scientific parlance however, a field is more rigorously defined as 'a quantity that can take on different values, and possibly also different directions of maximum action, at different points in space and time'. The geometric field idea can be used (say) to describe the 3D temperature gradient around a hot object, or the average velocities of molecules in a flowing liquid; but the fields that are the most perplexing, and that ultimately reveal the deepest secrets of the Universe, are those that appear to produce action at a distance. Of these so-called 'force fields'; the gravitational, electric and magnetic are the most familiar; and of course, it is the latter two that concern us here.

The beginning of what is loosely called 'modern physics' can be traced to a single deduction made by James Clerk Maxwell in the latter part of the 19th Century. Maxwell collected the details of every known scientific result concerning electricity and magnetism, and lent his phenomenal mathematical skill to the problem of finding a single theory. This led him to discover inconsistencies in the laws of induction (i.e., those laws that govern the effects of time-varying electric and magnetic fields), which would nowadays be interpreted as violations of the principle of conservation of charge. Electrons had not been discovered at the time, and electric charge was thought to be some kind of fluid; but whatever it was, the physical ideas of the day did not permit it to disappear from one place and reappear in another. He fixed the problem by making a bold and unprecedented step; which was to postulate the existence of a new kind of electric current, not associated with flowing charges, which he called 'displacement current'.

Speculation is one thing; but Maxwell had a test for his theory. With the inclusion of displacement current, the modified laws still allow the existence of electricity when all of the terms relating to physical matter are deleted. This 'free electricity' has to be in the form an oscillating electric field combined with an oscillating magnetic field, with the directions of action of the fields disposed at right-angles. It had turned out that the fields do not represent mysterious action at a distance after all (although it took some years before that point was fully accepted). They are instead stores of and agents for the transfer of pure energy. The liberation of electricity is however subject to a strict condition; which is that the energy exists by virtue of continuous transfer between the two fields according to the laws of induction; i.e., a decaying electric field gives rise to a magnetic field, and a decaying magnetic field gives rise to an electric field; and so the energy constantly swaps from one to the other and back again. The mechanism only works if the energy is propagating through space in a direction at right-angles to the crossed fields with a velocity given by the expression:

$$v = 1/\sqrt{(\mu\varepsilon)}$$

where μ (Greek lower case "mu") is the *magnetic permeability* and ε ("epsilon") is the *electric permittivity* of the surrounding medium.

Permeability μ is a constant of proportionality obtained from the force of magnetic attraction or repulsion that occurs between wires carrying an electric current. Permittivity ε is a constant obtained from the relationship between the physical dimensions of a capacitor and its capacitance. Maxwell found that the best available measurements for the permeability and permittivity of vacuum, μ_0 and ε_0 ("mu nought" and "epsilon nought") gave a propagation velocity for free energy, $c = 1/\sqrt{(\mu_0\, \varepsilon_0)}$, which turned out to be the same as the speed of light. Thus he was able to confirm a suggestion put forward by Michael Faraday some years before, which is that light is composed of electromagnetic waves. Maxwell had also shown, of course, that electrical energy is a form of light; and that older ideas derived from DC experiments were no longer tenable.

Maxwell died in 1879 at the age of 46, only six years after the publication of his great treatise on electricity and magnetism. Thus it was left to others to explore the ramifications of his work. In the latter part of the 19th Century, there were two great interpreters of Maxwell's electromagnetic theory: Oliver Heaviside and Heinrich Hertz; both of whom were brilliant mathematicians in their own right. These two scientists independently cleared-up Maxwell's notation and reduced a nest of algebraic clutter to a set of four equations that describe the fields. The four 'Maxwell's equations' that we know today are actually a variant of the form preferred by Heaviside (extra terms, which are zero for the Universe in its present state, are nowadays usually deleted).

The climax of Hertz's work was the creation and detection of Maxwellian waves under laboratory conditions[1]; which means that Hertz is the father of radio telecommunications, and also the inventor of the first radio antennas. His clarification of Maxwell's theory was also the basis of the work of one Albert Einstein, a Zurich patent examiner with a habit of daydreaming about objects in relative motion.

Einstein realised that Maxwell's separation of light and matter implies that the speed of light is constant regardless of any motion on the part of the observer. This led to the Special and General Theories of Relativity, which overturned all 19th Century notions of space and time. He also gave us the explicit unification of electricity and magnetism, by showing that electromagnetic induction is a relativistic phenomenon. Most readers will be aware that an electro-mechanical generator works by moving a coil of wire relative to a strong magnetic field. The changing magnetic field (as seen from the coil's viewpoint) gives rise to an electric field, which manifests itself as a voltage across the ends of the coil. Einstein tells us that the magnetic field does not so much create an electric field; it *is* an electric field when seen from a moving frame of reference. Likewise, an electric field is a magnetic field when viewed by a moving observer. This means that generators (and by a converse principle, electric motors) make use of relativistic effects when they convert energy between its electrical and mechanical forms.

Heaviside's extended version of Maxwell's equations was background to the work of Paul Dirac, who later went on to predict the existence of anti-matter. Heaviside's most important work however was carried out before the advent of radio as a technology, and was primarily related to the problems of long-distance electrical communication (telegraphs and telephones). It is Heaviside who gave us the correct picture of electricity, by way of another corollary of Maxwell's theory called 'the principle of continuity of energy' (not to be confused with the principle of *conservation* of energy).

The principle of continuity dictates that energy cannot simply disappear from one location and reappear in another, it must, in some sense, be transferred. This, incidentally, is not the same as imagining that energy follows a specific route; because we can only explain phenomena such as optical diffraction (and remain consistent with quantum theory) if we allow that even the very smallest quantity of energy can follow a multiplicity of paths during flight. Nevertheless, it retains a form of integrity (it is conserved [2]), which means that electric and magnetic fields from different energy sources cannot combine to make electromagnetic radiation. It is intriguing to note that, were such combination possible, every stray field would interact and the Universe would explode,

[1] **Hertz, the Discoverer of Electric Waves**. Julian Blanchard, Bell System Technical Journal, July 1938, Vol. 17, No. 3, p326 - 337. [Available from http://bstj.bell-labs.com/]

[2] **Is the Universe leaking energy?** Tamara M Davis. Scientific American, July 2010, p21-27.

perhaps to expand to a state in which such interaction can no longer occur. The continuity principle allows us to break reality down into separate energy transfer processes; and so without it, we would not be able to understand the Universe. On a more immediate level however, it tells us exactly how the energy flows in an electrical circuit, and indeed, how it flows into space from a radio antenna.

The explanation for the principle of continuity comes, once again, from the work of Einstein, this time from his investigation of the photoelectric effect. It also follows logically from Maxwell's equations; the issue being that there must be a reason for the rate (frequency) at which energy swaps between the fields in a propagating electromagnetic wave. On the assumption that energy is the simplest thing in the Universe, the only possible governing factor is the *amount* of energy being transported. Hence it was bound to be discovered that light is made up of 'particles' (i.e., discrete units, not solid objects), each oscillating at a frequency dictated by the amount of energy it contains. The particles, of course, are nowadays called 'photons', and the relationship between energy and frequency is known to be a direct proportionality:

$$E = h\, f$$

where h is 'Plank's constant', and has a value of $6.62606896 \times 10^{-34}$ Joule seconds. This introduces another strange concept, known as 'wave-particle duality'; which is sometimes claimed to be a paradox but is actually nothing of the sort. For a glib explanation, we can say that it would only be paradoxical if a batch of propagating energy was not made up of discrete units, because then the energy would have no way of knowing its own frequency and so would not be able to form a wave. For a more formal way of thinking about this issue however; note that an electromagnetic wave is defined in relation to a route through a field. The wave-like nature of the energy flow is detected by inserting probes (measuring devices) into the field and building up a picture by intercepting photons. From this we infer that photons travel as waves, even though we can only discern that by adding together the small packets of energy delivered by them.

This, incidentally, raises a general point in relation to scientific observation; which is that there are no paradoxes in nature. Paradoxes exist only in the mind of the observer, and result from attempts to interpret information using faulty starting assumptions. When we detect waves, we do so on the assumption that our probes measure field strength. This is very convenient, because it allows us solve problems using field theory; but when the meaning of an observation is in doubt, or when discrepancies begin to accrue, it is important to remember that all measurements are ultimately purely dependent on what can be inferred from the absorption and emission of energy.

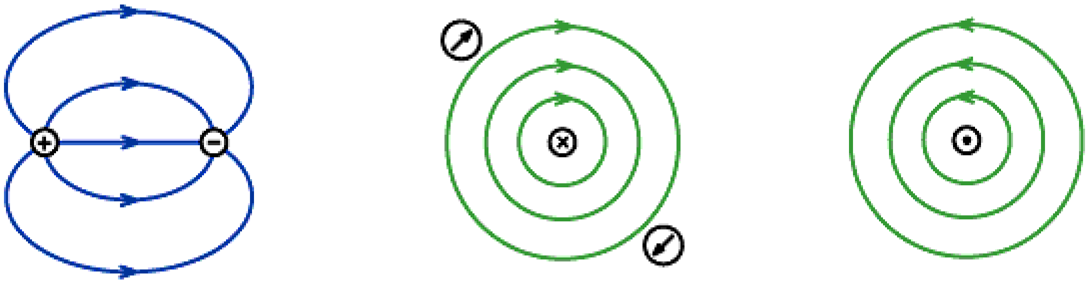
We rarely need to think about the granularity of light when working at electrical frequencies, because the amount of energy in each photon is exceedingly small. The wave property is instead the most dominant feature, and is often reinforced by a behaviour called 'coherence'; which is the ability of identical photons traversing the same set of paths to synchronise their fields. In this way, we see smooth waves in the collective behaviour of many particles. Heaviside knew nothing of this, but the photon theory explains his continuity principle by identifying the energy carrier. The fields extend throughout space, because they represent the propensity to exchange energy, and the photons can turn up anywhere that the fields have finite intensity; but the photons themselves undergo no additional exchange processes during transit.

Our ideas on the flow of electromagnetic energy are nowadays associated with John Henry Poynting, who formalised the continuity principle in a theorem that bears his name. Heaviside however, had already been using the principle for some time, and had a far more elegant derivation lodged with his publisher at the time of Poynting's first public presentation. Poynting also, did not interpret his findings correctly, whereas Heaviside had no such trouble; and so it was the latter who first described the underlying mechanism[3]. What follows may come as a shock to those who have

[3] **Oliver Heaviside**, Paul J Nahin. 2nd edition (paperback). John Hopkins University Press 2002. ISBN 0-8018-6909-9. Ch. 7, Tech note 3, p129-131.

been taught the 'lies-to-children' version. It turns out that the inside of a good conductor is the one part of a circuit where the transmission of electrical energy does not take place.

We can understand Heaviside's explanation by using Faraday's "lines of force", which provide a way of visualising the electric and magnetic fields. Some field patterns relevant to to the workings of circuits are shown below:



The left-hand diagram represents an electric field as it might exist between two charged spheres, or between two electrical conductors seen in cross-section. Recall that like charges repel, and opposite charges attract; and so a positively charged particle will be repelled by the (+) electrode and attracted to the (-) electrode. Hence, depending on the starting point, the arrows show the direction in which a positive charge will be accelerated, and the lines show the path that will be followed. There are, of course, an infinite number of possible starting positions, and so the field has an infinite number of lines; but a sparse representation is sufficient to give the general idea. The curvature of the lines arises because the mutual force between pairs of charged bodies is governed by an inverse-square law, i.e., the attraction or repulsion is strong when the bodies are close, but falls off rapidly with distance. Hence a particle close to one electrode will have a trajectory almost perpendicular to the surface, but the field line becomes curved further away because the particle is then influenced by both electrodes.

The middle diagram shows the lines of the magnetic field surrounding a wire when a current of positive charges is flowing away from the observer. The wire is shown in cross section, and the cross within its boundary represents the flow direction as the tail fins of a receding dart. The right-hand diagram shows the field when the current is flowing towards the observer; the dot being interpreted as the point of an approaching dart. Notice that we use a convention established before the discovery of the electron (by J J Thomson in 1897); which is that current flows from (+) to (-). Electrons flow the other way; but the continued use of conventional current makes no difference to the theory and serves to preserve the intelligibility of past scientific literature.

In the case of a magnetic field, the arrows drawn on the lines of force show the direction in which a compass will point when placed in proximity to the wire (presuming that the current is large enough to overcome the Earth's magnetic field). The clockwise 'rotation' of the field lines when the (conventional) current is flowing away is known as 'Maxwell's corkscrew rule'. This rule derives from the convention that the field lines around a bar magnet (or compass needle) emerge from the North-seeking pole and return to the South-seeking pole. Magnetic bodies repel when their force lines are in opposition (North pole to North pole), and attract when their force-lines point in the same direction (North pole to South pole). Hence a compass needle is repelled by the field lines coming towards it and attracted to the field lines going away from it.

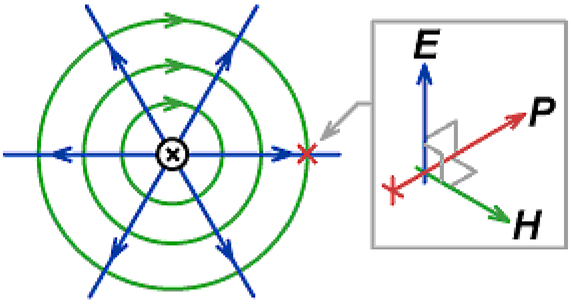
The Maxwellian fields have the same geometric properties as Faraday lines; and so we can now forget about forces on charged bodies and magnets and think about electromagnetic energy. Recall that Maxwell discovered that light travels with its electric field oscillating at right angles to its magnetic field, the direction of propagation being at right angles to both fields. Heaviside and Poynting now tell us that the transport of energy in electrical circuits occurs in exactly the same way.

An arbitrarily chosen point in an electric or magnetic field has both intensity (i.e., magnitude, or strength) and a direction of action. Such points are known as 'vectors', and a region of space filled with vectors is called a 'vector field'. There is also a mathematical operation called 'vector multiplication', which can be applied at points where two fields cross to produce a new vector at right-angles to the original two. Notional directions were assigned to the electric and magnetic vectors in the discussion above; and the adopted convention is, of course, one that gives the correct direction of energy transport when the fields are combined using the vector product (or "cross product", as it is also known). The electric field is usually given the symbol ***E***, and the magnetic field the symbol ***H***. The cross product is then written:

$$\boldsymbol{P} = \boldsymbol{E} \times \boldsymbol{H}$$

where ***P*** is known as the 'Poynting vector', and gives the intensity and average direction of the energy flow at some point in the combined electric and magnetic (i.e., electromagnetic) field.

Now consider the electromagnetic field around a wire in a circuit (as shown on the right). Electric field lines emerge perpendicular to the surface, magnetic field lines encircle the conduction current; and there are an infinite number of points at which they cross at right angles. Thus, assuming that the ***E*** and ***H*** fields are related in accordance with the continuity principle, we can work out the direction in which energy is travelling (and also the rate of flow) at any location.



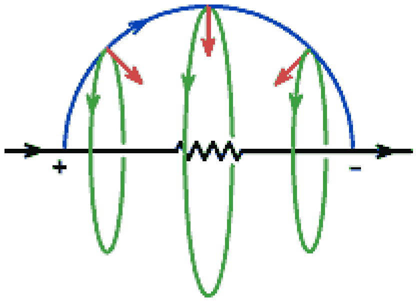
The key to the right of the diagram gives the direction of the Poynting vector in relation to the ***E*** and ***H*** fields. It follows that if the electric field is strictly perpendicular to the wire, then the Poynting vector lies parallel to the wire; and the fields as they are depicted have it running away from the observer. Notice also, that all of the propagating energy is on the outside of the wire (albeit in greatest concentration close to the surface where the magnetic field is strongest); and it transpires that if the wire is a perfect conductor, there is none on the inside at all.

It requires both an electric field and a magnetic field for the transportation of energy. A perfect conductor however is a material that, by definition, cannot sustain an electric field. This can be understood by noting that the electrical resistance between any two points within the body of a perfectly conducting object is zero, in which case there can be no voltage difference and so no electric field. Hence *electrical energy* cannot flow inside a good conductor.

This understanding, incidentally, gives rise to a semantic difficulty regarding whether there is a difference between 'electricity' and 'electrical energy'. It is hard to justify the preservation of different meanings for the two terms, and yet people will persist in saying that electricity "flows through" conductors. We can sidestep the issue by saying that electricity *flows along* the wires, but that does little to rectify the basic misconception. The general consensus now seems to be that unqualified use of the word 'electricity' should be avoided altogether in any rigorous scientific context. The electricity for which the utility company demands payment however, is definitely of the Heaviside, rather than the electrons in wires, variety.

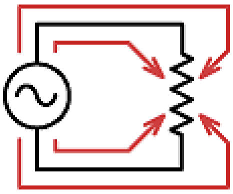
Now that we have established the location of the electrical energy; it must be added that a small amount does flow into (but not through) practical conductors. This is because metal (presuming that the temperature is too high for it to be superconducting) always has some resistance. The inflowing energy is, of course, lost from the fields and converted into heat.

The mechanism of *energy delivery* can be understood, once again, using the Poynting vector; and it explains not only unwanted losses, but also what happens in relation to devices that are deliberately made resistive so that they can absorb large amounts of energy. We start by imagining a small particle of resistance, such as an infinitesimal resistive region in an otherwise perfectly conducting wire. We know of course, that resistance is distributed throughout conducting materials; but the continuity principle allows us to break energy transfer processes down into separate components, which can later be combined to give the overall picture.



When a current flows along a conductor, a resistive region gives rise to a voltage drop or 'potential difference'. Hence an electric field exists between a point just upstream and and a point just downstream of the obstacle. The diagram on the right shows the interaction between the magnetic field encircling the wire and a single electric field line (other lines are left to the imagination). Using the sense of the Poynting vector given earlier, we see that energy flows into the resistance from both the upstream and downstream sides. Now, mentally, rotate the diagram about the wire axis and it becomes a spherical wave-front converging onto the point resistance.

A further important observation arises when we consider what happens when the direction of the current is reversed. In that case, the directions of the electric and the magnetic vectors are both reversed, and so energy continues to flow into the resistance. The fields explain the strange fact that the direction of energy flow cannot be reversed by swapping the polarity of the power supply. It can be reversed however, by replacing the resistance with a device that is a source of energy (a battery or generator) or by devices that can store energy (inductors and capacitors); this being a matter that we will explore in detail later.



When we think of the Poynting vector in relation to a complete electrical system, we are really thinking of the average of a large number of microscopic energy transfer processes. In the case of a simple circuit consisting only of a generator and a resistive load; the Poynting vector is directed along the wires, from generator to load; and the direction is the same on both sides of the generator. Should we examine the average energy flow close to a wire however, we will see that the direction is tilted very slightly towards the wire, on account of the distributed resistive losses.

So now we have the basic field picture of electricity, but there remain a few issues that need to be explained. Particularly, we need to look again at electric current, and the matter of why it is defined in terms of moving charges. As it says in every school physics textbook, an Ampere is a current of one Coulomb per second; and since the charge of an electron is -1.6021892×10^{-19} Coulombs, an Amp flowing from (+) to (-) corresponds to $1/1.6021892\times10^{-19} = 6.241460122\times10^{18}$ electrons per second flowing from (-) to (+). This is correct (assuming that the current is due to electrons), but only in the special limiting case when the frequency of the electromagnetic energy being transferred tends to zero. In other words; it is only strictly true for DC electricity. As the frequency increases, the correspondence between conduction current and effective current becomes progressively less accurate; which is why Maxwell invented displacement current.

The role of the electrons in conduction is actually an optical one. They interact with the electromagnetic field in such a way as to increase the amount of energy that can be stored in the region of space immediately surrounding the conductor. This creates a duct through which the energy prefers to flow; in a manner analogous to the way in which mirages are sometimes seen in hot deserts, and over-the-horizon VHF radio communication becomes possible on hot days. Ducting occurs when the refractive index of the medium increases with distance from the surface, i.e., a light ray that tries to move away is bent towards the parallel direction.

The existence of electrons was hypothesised, and indeed the name was coined, some time before

J J Thomson identified them as the current-carriers in cathode-ray tubes (1897). It must then have seemed to many that the electrical fluid theory was confirmed; but the discovery was actually its nemesis. Upon estimating the number of free electrons in a given volume (say) of copper, it turns out that the average velocity of propagation of an electron current through a solid medium is of the order of a few millimetres per second. The mass of the electron is also so small that the amount of energy transferred by collisions with the atoms of the conductor comes nowhere near to the amount transferred by the electromagnetic field. In fact, the collision energy is merely the resistive loss that occurs in imperfect conductors. The electrons do not carry the electrical energy, but they do extract a small tax for the service they provide in guiding it along the outside surfaces.

So how are we now to think of electric current? The current that conveys energy would seem to be best described as a displacement current, at least in the sense that it is not carried by electrons. It merely becomes correlated with the number of electrons passing a given point in a conductor per second when the generator frequency is low. For historical reasons however (i.e., because Maxwell modified the definition of current instead of replacing it) we think of current as the sum of conduction and displacement currents. This turns out to be a reasonable approach, because there are many situations in which conduction current is important. Electronics, for example, is the art of controlling electricity by controlling the conduction current. Charged particles are also involved in the workings of chemical energy sources (batteries and fuel cells) and electrochemical processes in general (e.g., electro-plating). For the vast range of AC electrical problems however (including the design of electronic circuits) conduction current is a source of misconceptions, and needs to be distinguished from current in the general electromagnetic sense.

We shall proceed by thinking in terms of a type of current called 'current'; which may or may not be correlated with the movements of charged particles, but for most of the time we don't care whether it is or not. This might seem to imply that we have turned current into an abstract idea, but actually we have simply unloaded some unnecessary baggage. Many readers will be aware that voltage is sometimes referred to as 'electromotive force' (EMF). Its unit of measurement is not a pure force in the Newtonian sense, but it is proportional to the force exerted on a charged particle; and so the habit of referring to it as a force is loosely (and widely) accepted. Likewise, current is proportional to the force exerted on a magnet in the field, and its unit, the Ampere, is the measure of *magnetomotive* force. The electric field in a circuit is everywhere proportional to the driving voltage. Likewise, the magnetic field is everywhere proportional to the effective current; and the two fields between them set the intensity of the Poynting vector. Also, it has to be said that AC ammeters are calibrated according to the amount of energy delivered, and so read the true, or magnetomotive, quantity.

Before we move on, it is perhaps worth making a few observations on the use of the term 'displacement current'. There are occasional (non peer-reviewed) publications that agonise about the supposedly deep philosophical implications of this strange quantity. That is unfortunate, because it doesn't actually exist. Maxwell coined the term because he initially imagined it as distortions of the Æther, the latter being an elastic medium supposed to permeate all space and thereby 'explain' the paradoxical phenomenon of action at a distance. The 19th Century luminiferous Æther[4] has now gone the way of the Earth-centred Universe (and good riddance); and Science has come to explain all electromagnetic phenomena in terms of fields and particles. It will do no harm to think of displacement current as a convenient fudge; which allows us to extend the laws of DC electricity to higher frequencies and thereby avoid having to subject every problem to the full electromagnetic treatment. Hence, 'displacement current' is that which has to be evoked because electromagnetic energy doesn't always follow the wires. It is not a physical current. It is just a quantity that corrects for the difference between the magnetomotive force and the conduction current.

[4] The term 'Ether' has however come back into favour in the discussion of the properties of the quantum vacuum. see: **The Lightness of Being**, Frank Wilczek. 2008, (Penguin edn. ISBN 9780141043142), especially ch. 8.

Magnetomotive force (or 'MMF') is incidentally, not completely synonymous with current. Were it so, we would gladly drop the misleading concept 'current' altogether; but unfortunately, we are stuck with it. The reason is that MMF and current are only identical in circuits formed of a single conducting loop. When the circuit is composed of overlapping loops, disposed in such a way that adjacent conductors carry current in the same direction, the MMF is increased due to a phenomenon called 'magnetic flux linkage'. Such overlapping structures are, of course, known as coils or inductors, and have the property that they allow the amount of magnetic energy that can be stored in a given volume of space to be magnified. Still, for the greater (non-overlapping) part of an electrical circuit, current and MMF are practically the same; and to a good approximation, we can dispense with the details and treat coils as separate objects having a single magnetic concentrating property called 'inductance'. Certainly, it is very useful to know how to calculate inductance from the number of turns and the physical dimensions of a coil; but it is a matter that can separated from the general business of designing electrical systems.

2. Circuit analysis overview

The basic theory of electrical circuits is known as 'lumped component analysis'. The verb 'to analyse', incidentally, means 'to break down into simpler or more-fundamental parts'; and in this case, analysis is the art of describing and predicting the behaviour of circuits by treating them as networks of interconnected resistances, capacitances, inductances and generators. This turns out to be an extraordinarily accurate technique when correctly applied; but it has limitations and idiosyncrasies of which the practitioner needs to be aware.

A peculiarity, which is often introduced without comment, is that AC generators (of the analytical variety) are considered to produce sinusoidal outputs. Many practical generators (e.g., mechanical alternators, radio transmitters) do indeed produce something approximating a voltage or current sine-wave; but the reason goes somewhat deeper than that. If we take, for example, a moving-coil microphone (which is a type of generator that produces electricity from air-pressure variations), we will find that its output in response to (say) the sound of the human voice, is extremely complicated. A technique known as 'Fourier analysis' however, shows that all waveforms can be built-up by adding-together sinusoidal waves of different frequencies; and physical investigation shows that these separate frequency components actually exist. A set of one or more frequency components is known as a signal. It transpires that no new frequencies will be added to a signal when it is processed (e.g., passed from the input to the output of an electrical network), provided that the materials encountered in the transmission path behave in a linear manner. In general, a material is said to be 'linear' when its change in response to some force is proportional to the intensity of that force (i.e., the graph of change versus force is a straight line), and the change is reversible. Resistance, of course, obeys a straight-line law called 'Ohm's law'. The AC voltage versus current laws laws governing inductance and capacitance are also linear. Thus, the basic circuit elements combine to make linear networks; which lack the ability to produce new signal components. This means, overall, that a linear network treats each frequency component as if it exists in isolation. We can therefore quantify its behaviour one frequency at a time, which is why the basic generator of circuit analysis produces only a single frequency. The response of a circuit to more-complicated waveforms can always be built-up (when required) by adding the results of analyses carried out at the component frequencies. For many purposes however, the focus of interest when several frequencies are involved is the *frequency response*, which is just a stepwise application of the one-frequency-at-a-time approach.

Despite the simplicity offered by single-frequency analysis, there are, of course, numerous electrical and electronic components that behave in a non-linear fashion. Semiconductor devices (diodes, transistors, etc.) are an obvious example; but there are also materials that change their characteristics according to field strength. This might seem to place a limitation on linear network theory; but actually, there is a straightforward solution. A non-linear device is one that accepts energy at one or more excitation frequencies and converts it into energy at one or more new frequencies. From a circuit analysis point-of-view, anything that absorbs energy is a resistance, and anything that creates a new frequency component is a generator. Hence we can put the behaviour of an alien device into a metaphorical 'black box'; with one or more two-terminal connections called 'ports', which look to the outside world like networks of basic circuit elements. The rule by which energy disappears into a resistance inside the box, and reappears from one or more generators inside the box is called the 'transfer function'. The fact that anything with electrical connections and a known transfer function can be incorporated into circuit theory confers enormous power upon the method.

As mentioned previously however; circuit analysis does have its limitations. It is after all, not a general theory, but a projection or 'degenerate form' of electromagnetic theory. Naturally, a price is paid for the simplification, and it is instructive to consider what that is. So, look at a circuit diagram and try to find where the lengths of the wires and the physical dimensions of the components are written. That information is conspicuous by its absence, because a circuit diagram is a purely topological representation (like the famous London Underground map). It was a curious and usually unremarked discovery of the early circuit experimenters, that it doesn't matter how the equipment is laid out, or whether a component is large or small, or how long the wires are (provided that they are a lot more conductive than any of the designated resistances). This, of course, ceases to be true as the frequency is increased; and this breakdown of DC theory is partly due to the finite speed of light.

Consider a sine-wave generator connected by means of relatively long wires to a resistive load. If we measure the voltage difference between any two points in the circuit, we obtain a quantity that is proportional to the total electric field existing between those points. In this case, the field is associated with electromagnetic energy propagating from the generator to the load. Since it takes a finite time for the energy to make the journey, this means that the voltage measured across the generator will not be identical to the voltage measured across the load. If we presume that the resistive loss in the wires is negligible, the main difference will be in the relative phases of the two sine waves; i.e., if we take some reference point on the waveform, such as the zero-crossing-point on going from negative to positive, we will find that the load waveform is delayed relative to the generator waveform. This will not be noticeable if the measurements are made using an ordinary AC voltmeter, but the time difference can certainly be demonstrated using a dual-channel oscilloscope; and the same effect will give rise to performance deviations in more complicated (i.e., phase critical) circuits. There are ways of dealing with such problems (and in the example case, it is to represent the wires as inductances with some capacitance between them), but it is important to understand that the 'truth' of circuit diagrams is contingent upon unspecified factors.

Knowing the difference between representation and reality is the art (as opposed to the science) of circuit analysis. No one would want to apply the full electromagnetic theory to routine circuit problems; and indeed, success in the solving of Maxwell's equations for some particular class of problems is often regarded as a scientific event. Hence electricity is primarily associated with circuits, rather than fields and waves, and being 'good at it' requires a level of understanding that is difficult to formalise. Experience comes with time, but we can at least invite entry to the Guild by offering a straightforward rule of thumb. A light wave in vacuum completes one cycle of variation of its electric and magnetic fields upon travelling a distance given by the expression:

$$\lambda = c / f$$

where f is the frequency;
c = 299 792 458 metres/second is the speed of light;
and λ (Greek "lambda") is the wavelength.

Due to the essentially refractive nature of circuits, the apparent velocity of signal propagation through an electrical network is never exactly c, but it rarely deviates from c by more than a few %. Hence we can easily obtain an idea of the phase errors that will accumulate as a result of constructing a circuit on a particular physical scale. A given amount of phase error does not necessarily translate into the same error in some other quantity, but if we confine ourselves to thinking of the order of the error (i.e., its magnitude thereabouts), it is a fairly good guide. On that basis, we can answer the question: "If I build the circuit as drawn, how well will its performance agree with my analysis?" The scale on which attention to physical detail (layout, wire lengths, component size, etc.) will be required, for a given level of agreement, is shown for various frequencies in the table below (using the mental-arithmetic approximation $c = 3\times10^8$ m/s).

Frequency f	**Wavelength** λ = c / f	**Construction scale for a given accuracy**		
		10%	**1%**	**0.1%**
300 kHz	1 km	100 m	10 m	1 m
3 MHz	100 m	10 m	1 m	10 cm
30 MHz	10 m	1 m	10 cm	1 cm
300 MHz	1 m	10 cm	1 cm	1 mm
3 GHz	10 cm	1 cm	1 mm	0.1 mm

Assuming that most signal-processing circuits are constructed on a scale of about 10 cm, we can see that there will be no serious discrepancies between analysis and practical results for frequencies up to about 3 MHz (ignoring displacement currents for the time being). Beyond that, we are definitely into the realm of radio-frequency engineering, where layout is important; but note that this does not mean that analysis will fail. Rather, as was alluded-to earlier, it requires a modified approach, where some physical variables have to be turned into theoretical circuit components. As we approach ultra-high frequencies (UHF) however, the struggle to adapt the lumped component representation will become increasingly difficult; and the need to resort to Maxwell's equations (or at least, to standard solutions obtained from the scientific literature) will become more and more apparent.

While on the subject of scale incidentally; note that the wavelength range of visible light runs from 0.7 to 0.4 microns (where 1 micron = 1 μm = 10^{-6} m). If we were to represent the interaction of visible light and matter using circuit diagrams, the circuits would have to be built on the molecular scale (around 1 nm = 10^{-9} m). Hence visible light has no measurable tendency to be guided by ordinary electrical circuitry; but it does have the convenient habit of reflecting from the components, so that we can see them.

We can now draw together two threads from the preceding discussion. Representing a circuit diagrammatically begins with a basic assumption; that either the circuit is infinitesimally small, or that the speed of light is infinite. Either way, it means that every part of the circuit is assumed (initially) to be in instantaneous communication with every other part. It is also assumed that the electrical energy always follows the wires, whereas it is actually distributed in the fields surrounding the circuit. In both cases, in the absence of corrective measures, this can result in disagreement between the behaviour calculated from circuit theory and the measured performance of the actual circuit.

In resolving the potential for discrepancy, we must first recognise that circuit diagrams fall into two categories: those that are used for production engineering and end up in service manuals; and theoretical diagrams used by designers. As we will see in this and subsequent articles; a theoretical diagram is actually a type of mathematical statement; which can be extended to describe the behaviour of a physical circuit to an almost arbitrary degree of precision. A production diagram, on the other hand, is just a record of the interconnections in a set of manufactured sub-assemblies (resistors, transistors, coils, etc.). As has already been implied: for equipment operating at audio frequencies and below, there may be a great deal of similarity between the diagram used by the design engineer and the diagram in the service manual; but for well-designed radio equipment, this will not necessarily be the case.

On the subject of circuit diagrams; it will be noted that the North American or Japanese preferred (zig-zag line) symbol for resistance is used here. This convention is adopted (or, in the author's case, was never un-adopted) because the rectangular box symbol was already in use by theoreticians long before European standards (apparently intended solely for the convenience of draughtsmen) were put forward. In this, and all of the other documents produced by this author, the box symbol is used strictly to represent a generalised electrical network. It is also, in particular (and in keeping with long-standing practice), used to represent a generalised two-terminal linear network called an 'impedance' (a mathematical construct with which we are about to become very familiar).

In the following sections of this article, we will derive the basic AC theory, which deals with notionally discrete resistances, inductances and capacitances, these being known as *ideal components*. As will be shown, the behaviour of networks of these components can be determined by starting with a handful of empirical electrical formulae (i.e., formulae determined by experiment) and then using the properties of the Poynting vector to determine the rules of combination. We will not, incidentally, need to know how to carry out 3D vector multiplication explicitly; we simply need to know that, for multiplication operations of any type, when the two quantities being multiplied have the same algebraic sign, the answer is positive, and when they have opposite signs, the answer is negative. This will result in an extendible theory of linear networks; extension being a matter of incorporating circuit modules (black boxes) called 'component models' or 'equivalent circuits', which have ports defined as networks of ideal components. These modules were mentioned previously as a way of dealing with non-linear devices; but the approach can be applied to any electrical device that requires the use of a more sophisticated theory (such as electromagnetics), or that is not well described by treating it as a discrete ideal component.

It will be noticed that we have talked of 'resistances, capacitances and inductances', rather than 'resistors, capacitors, and inductors'. There may be no great difference between the two conceptions from the point-of-view the audio engineer; but for the radio engineer, ideal components and practical components are not the same. Whether a practical component can be regarded as an ideal component is a matter of internal dimensions and wavelength. Some capacitors, for example, are made by rolling-up long lengths of metallised plastic film, giving an assembly that behaves quite like a pure capacitance at low frequencies, but turns into an inductance at radio frequencies (thus radio engineers have a certain fondness for capacitors that have small electrodes). Similarly, an inductor is made by coiling-up a length of wire, and its property of pure inductance is modified by the resistance of the wire and by the time it takes for an electromagnetic wave to propagate along it. Even the humble resistor is not perfect; and in general: every practical two-terminal device is replaced by an equivalent circuit, which may (sometimes) correspond to a discrete ideal linear component at low frequencies, but at high frequencies will always mutate into a network of resistances, capacitances and inductances. The subject of component models is developed in detail in later articles in this series; but for now, the important point is that circuits designed on the assumption of ideal behaviour may require extra work if they are to be realised in practice.

Neglecting possible (but usually minor) non-linearities; the difference between ideal and practical components can always be attributed to a combination of three factors: time delays; displacement currents; and the finite resistance of conductors at ordinary temperatures. The same can also be said of wiring and layout. Hence, an accurate theoretical model for a practical electrical system may also require components to represent spurious effects in the circuit at large. In the case of extreme time delays (i.e., long connecting wires), we can introduce a two-port module called a 'transmission line', which is a solution of Maxwell's equations in a box. Minor effects can be accounted for by including 'stray' or 'parasitic' resistances, capacitances and inductances here and there. In the absence constructional details and related corrections however; it is implicit in any theoretical circuit: that the components will be grouped on a scale that is small in comparison to wavelength; that interconnection resistance is negligible in comparison to specified resistance; and that there may be a need for shielding to prevent energy from turning-up in places where it is not wanted.

Having cautioned against the pitfalls however; we should also caution against over-modelling. Even at radio frequencies, there need not always be a great deal of difference between the idealised and the practical representations. This is because, firstly: circuits will often work adequately when built according to the assumption that practical components are nearly ideal; and secondly, nominal component values are subject to manufacturing tolerances, which means that accurate performance can only be achieved by making some components adjustable. Adjustment not only serves to correct the nominal component value to the required value, it can also absorb deficiencies in the original analysis. Hence, it is important to understand that a simple approach will often do the trick; and extreme attention to subtleties is usually only needed when attempting to achieve the most exacting standards of radio frequency (RF) performance.

3. Basic electrical formulae

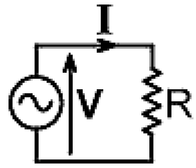
The theory that will be developed in the sections to follow is of a type classed as 'steady-state analysis'. This does not mean that nothing changes, but that everything that does change is assumed to do so sinusoidally; i.e., there are no sudden or 'transient' events. We can always represent more complicated phenomena (if necessary) by adding-together sine waves of different frequencies, or we can look to other bodies of theory better suited to systems that undergo non-periodic (i.e., non cyclical) change; but for a large range of problems, the steady-state approach is all that is needed. The translation from DC to the AC steady-state is a matter of replacing the battery with a sine-wave generator, and then adopting definitions for voltage and current such that the established DC laws continue to be true (insofar as that is possible). The necessary choice is to use RMS values of voltage and current; where the RMS (which stands for: the square-Root of the Mean of the Squares[5]) is an average chosen so that AC and DC electricity both have the same heating (or long-term energy-delivery) effect[6].

By using RMS values, we conscript various formulae and diagrammatic conventions into AC service; except that, due to the alternation of the power supply, it no longer makes any sense to write the symbols (+) and (-) against the generator terminals, and it no longer makes any sense to equate current to the flow of electrons. The solution is to ensure that the polarities of voltages and currents are defined in a way that gives the correct sense to the Poynting vector (which is what the DC conventions achieve by rote instead of reason in any case).

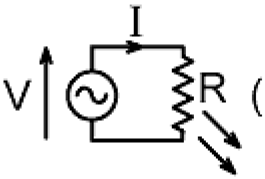
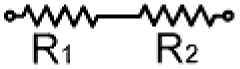
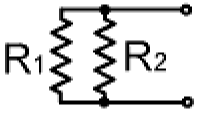
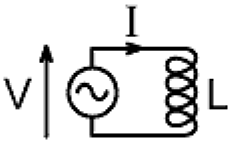
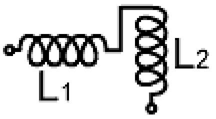
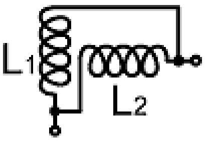
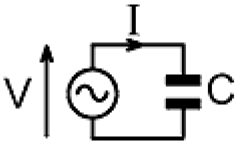
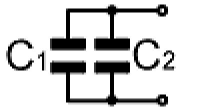
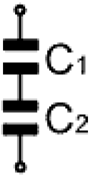
[5] Not "root mean squared" or some similar garbled version that throws away the actual concept.

[6] "**The RMS Average**" D. W. Knight. [Available from g3ynh.info].

An accepted usage, which is adopted here, is to draw an arrow across the generator terminals to indicate the direction in which electric potential is assumed to increase. Current then 'flows away' from the high-potential terminal. If both current and voltage are then taken to be positive (or both negative if you like, it amounts to the same thing), the chosen convention is that energy (or power) flowing away from a generator is positive.



As was mentioned earlier, AC analysis is the art of representing circuits as networks of interconnected resistances, inductances, capacitances and generators. Hence we have the task of developing the general rules of combination for those elements. What we have to combine is summarised in the table below; which gives the basic electrical formulae much as they appear in numerous textbooks.

 $V=IR$ (Ohm's Law) Power $P=IV=V^2/R=I^2R$	 $R=R_1+R_2$	 $\frac{1}{R}=\frac{1}{R_1}+\frac{1}{R_2}$ $R=\frac{R_1 R_2}{(R_1+R_2)}$
 $X_L=2\pi f L$ $V=IX_L$ $P=0$	 $L=L_1+L_2$	 $\frac{1}{L}=\frac{1}{L_1}+\frac{1}{L_2}$ $L=\frac{L_1 L_2}{(L_1+L_2)}$
 $X_C=-1/(2\pi f C)$ $V=IX_C$ $P=0$	 $C=C_1+C_2$	 $\frac{1}{C}=\frac{1}{C_1}+\frac{1}{C_2}$ $C=\frac{C_1 C_2}{(C_1+C_2)}$

Observe that only the entries in the left-hand column contain fundamental scientific information. The uppermost entry is a statement of Ohm's law; which is that the electrical current I is proportional to the voltage applied across the ends of a conductor, the constant of proportionality being known as the resistance. The entry below it is the power law; which represents the observation that a conductor (resistor) heats up as a consequence of an electric current (i.e., it dissipates energy) and the power consumed (i.e., the energy delivered per unit-of-time) is the product of the applied voltage and the current, i.e., $P = V I$ [Watts]. Also, by using the substitutions $I = V/R$ and $V = I R$, we obtain two alternative power laws: $P = V^2/R$ and $P = I^2 R$. The expression $P = I^2 R$ is known as *Joule's law*, and is a statement of the fundamental relationship between electricity and thermodynamics. One important point to note about the standard power and resistance formulae however, is that they are all derived from experiments with DC electricity. They represent *incomplete statements* of Ohm's law and Joule's law, because they can only be applied to AC circuits when the load on the generator is a pure resistance. Later-on we will show

how to state these laws in a completely general way, but some groundwork will be required before that can be done.

Notice incidentally, the correspondence between the power law P = V I and the earlier-given definition of the Poynting vector $\boldsymbol{P} = \boldsymbol{E} \times \boldsymbol{H}$. The former is a dimensionally-reduced version of the latter, as can be seen by noting that the unit of electrical field strength is 'Volts per metre', and the unit of magnetic field strength is 'Amperes per metre'. Also, even though we do not need to know how to perform 3-dimensional vector multiplication; it is not difficult to understand that any type of multiplication also multiplies the units of measurement. Hence the unit of the Poynting vector is 'Watts per metre-squared', which is a measure of illumination; i.e., the delivery of electrical power is a matter of illuminating the receiving object with electromagnetic energy.

In the case of inductors and capacitors, the entries in the left-hand column tell us that they also obey Ohm's law when connected to a generator of alternating voltage but, insofar as we can construct them without inadvertently including resistance, they consume no power. The reason why the ideal versions of these components cannot dissipate energy is that they have no resistance by definition, i.e., they cannot convert energy into heat or work. Instead, over the course of a generator cycle, the amount of energy that flows into the component is exactly equal to the amount that flows out. This property forces a ±90° phase difference between the voltage and current waveforms; a ±¼ cycle or 'quadrature' offset being that which causes the Poynting vector to reverse its direction four times per cycle. Hence, although the average or steady-state power consumption is zero, the instantaneous power-flow is alternating at twice the generator frequency.

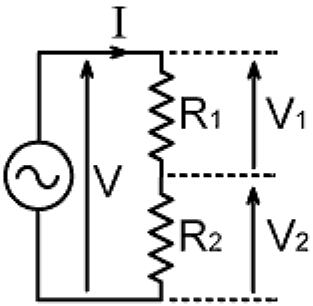
In the case of an inductor, the AC resistance or *reactance* X_L (measured in Ohms) is *directly* proportional to the inductance (in Henrys) and to the frequency f (in Hertz, i.e., cycles per second) of the applied voltage. The quantity $2\pi f$ is known as the angular frequency (i.e., the frequency in radians per second, where 2π radians corresponds to 360°) and is often given the symbol ω (Greek lower-case "omega"). In the case of a capacitor, the reactance X_C is *inversely* proportional to the angular frequency, and also inversely proportional to the capacitance (in Farads). Note also that capacitive reactance is shown as being negative; because it transpires that when capacitors and inductors are connected to form resonant circuits, the reactance of the inductor, in some sense, cancels the reactance of the capacitor. This means that one of the types of reactance has to be considered to be negative and, as will be explained later, we choose it to be the capacitive variety in order to be consistent with the conventions of trigonometry.

The other entries in the table are derived from the formulae in the left-hand column, using only Ohm's law and a basic electrical rule known as *Kirchhoff's first law* (pronounced: "kir-khov"). Kirchhoff's law tells us that the sum of all the currents flowing into a given point in a circuit is equal to the sum of all the currents flowing out. This law was originally regarded as proof of the principle of conservation of charge in DC circuits (what goes in must come out); but it also turns out to be true of current in the general (magnetomotive) sense, provided that we use the correct rules of addition (to be determined shortly) in circuits involving both resistance and reactance.

The entries in the middle and right-hand columns are, of course, the well-known series and parallel combination formulae for passive electrical components. These expressions can all be regarded as examples of simple mathematical models (in this case, in the sense that a single component can serve to represent a combination of several components). Of these, the formula for resistances in series is the simplest of all, and tells us that whenever we encounter two resistors in series, we can treat them as a single resistor with a value equal to the sum of the two resistances. That this statement is derived from existing physical laws can be seen by applying some basic techniques of circuit analysis to the circuit shown below:

Resistors in series

To analyse this circuit, we first observe that, as a requirement of Kirchhoff's first law, the current in the two resistors must be the same. Ohm's law then tells us that $V_1 = I R_1$ and $V_2 = I R_2$. Now, since voltage is analogous to pressure, common sense (otherwise known as Kirchhoff's second law) tells us that the total pressure-drop is equal to the sum of the pressure-drops across the two resistors, i.e.,



$V = V_1 + V_2$

Putting these ideas together we have:

$V = I R_1 + I R_2$

$\quad = I (R_1 + R_2)$

Now, if we postulate a hypothetical resistance R that represents the series combination of R_1 and R_2, it must be possible to replace R_1 and R_2 with this resistance and obtain the same current for a given voltage, i.e.,

$V = I R = I (R_1 + R_2)$

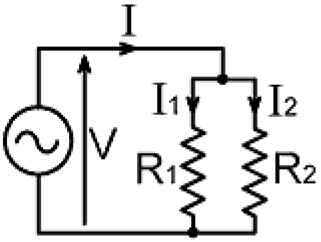
Hence:

$\boxed{R = R_1 + R_2}$

Resistors in parallel

In the case of two resistors in parallel, the *voltage across* the two resistors is the same. Hence Ohm's law tells us that:

$V = I_1 R_1 = I_2 R_2$ .... (**3.1**)



and Kirchhoff's first law tells us that:

$I = I_1 + I_2$

Now, if we postulate a hypothetical resistance R that represents the parallel combination of R_1 and R_2, we have:

$V = I R = (I_1 + I_2) R$

We can eliminate I_1 and I_2 by using equation (**3.1**) above, i.e., $I_1 = V/R_1$ and $I_2 = V/R_2$, hence:

$V = [(V/R_1) + (V/R_2)] R$

The voltage can then be factored out and cancelled to give:

$1 = [(1/R_1) + (1/R_2)] R$

and dividing each side of the equation by R gives:

$$\boxed{1/R = (1/R_1) + (1/R_2)}$$

This is one form of the standard expression for resistors in parallel, and a little rearrangement will give us the other. Inverting the expression above gives:

$R = 1 / [\ (1/R_1) + (1/R_2)\ ]$

We then arrange the terms inside the square brackets to have a common denominator (multiply the $1/R_1$ term by R_2/R_2 and multiply the $1/R_2$ term by R_1/R_1), i.e.,

$R = 1 / [\ \{(R_2/(R_1R_2)\} + \{R_1/(R_1R_2)\}\ ]$

hence:

$R = 1 / [\ (R_1 + R_2\ ) / R_1\ R_2\ ]$

which, upon inversion, gives:

$$\boxed{R = R_1\ R_2 / (R_1 + R_2\ )}$$

The formulae for inductors and capacitors in series and parallel may also be derived by using exactly the same approach as was used above; the only difference being that inductive reactance $X_L = 2\pi fL$, or capacitive reactance $X_C = -1/(2\pi fC)$, is substituted in place of resistance. The 2πf factors and any minus signs disappear by cancellation, leaving formulae involving only inductance or capacitance. Note incidentally, that the inductors in the illustrations in the previous table are shown orientated at right-angles to each other, this being done as a reminder that the formulae are only true when there is no magnetic coupling between the coils. Note also, that the capacitor formulae take on the opposite forms of their resistance counterparts; this being due to the reciprocal (inverse) relationship between capacitance and capacitive reactance.

4. Resonance

The combination rules discussed above allow us to deal with resistors, or capacitors, or inductors in series and parallel, but for reasons that will become clear in the following sections, they do not provide a method for dealing with combinations of resistance and reactance (if we try to add resistance to reactance directly, our calculations will not agree with our measurements). We can however deal with combinations of inductive and capacitive reactance, provided that we observe the convention that capacitive reactance is negative. We can therefore add to our repertoire of standard formulae by writing general expressions for pure reactances in series and parallel, i.e.:

Reactances in series
$X = X_1 + X_2$

Reactances in parallel
$X = X_1 X_2 / (X_1 + X_2)$

Now, since inductive reactance is positive and increases with frequency, and capacitive reactance is negative and decreases with frequency; if an inductance is placed in series or parallel with a capacitance, there will occur a frequency at which the two reactances cancel. That frequency, of course, is the *resonant frequency* of the combined reactances. A resonant frequency is usually denoted by the symbol ' f_0' ("f nought")

In the case of an inductor and a capacitor in series, the reactance goes to *zero*, i.e., the combination behaves like a short-circuit (neglecting resistance), when $X_L+X_C = 0$. In the case of an inductor and a capacitor in parallel, since the term X_L+X_C is on the bottom of a fraction, it would appear that the reactance goes to *infinity*, i.e., the combination behaves like an open-circuit when $X_L+X_C = 0$. A complete open-circuit does not appear in practice however, because in the parallel case, it transpires that we are not at liberty to neglect the resistances of the coil and the capacitor. We therefore cannot calculate the *exact* resonant frequency of a practical parallel tuned circuit, nor the resistance that remains when the reactance has been cancelled, until we have developed a more comprehensive theory; and so that is another matter that we must leave until later. We can say however, that the exact resonant frequency of a series tuned circuit, and the approximate resonant frequency of a typical parallel tuned circuit occurs when:

$$X_L = -X_C$$

i.e.,

$$2\pi\ f_0\ L = 1/(2\pi\ f_0\ C\ )$$

Now, if we rearrange this equation to put both instances of f_0 on one side we have:

$$f_0^2 = 1/(\ 4\pi^2\ L\ C\ )$$

and taking the square-root gives:

$f_0 = 1/[\ 2\pi \sqrt{(L\ C)}\ ]$	**4.1**

(pronounced: "f nought equals one over two pi root LC", or in rhyme: "One over two pi root LC gives the resonant frequency").

Equation (**4.1**) is, of course, is the standard resonance formula; but before accepting it we should note that, because it contains a square root, every combination of L and C has two resonant frequencies associated with it. Every equation involving a square root has two solutions because the square root of a number is, by definition: 'a quantity that, when multiplied by itself, gives the number in question'. When two negative numbers are multiplied, the result is a positive number. Hence, if x is a positive number, we must note not only that:

$x^2 = x \times x$

but also that:

$x^2 = (-x) \times (-x)$

hence:

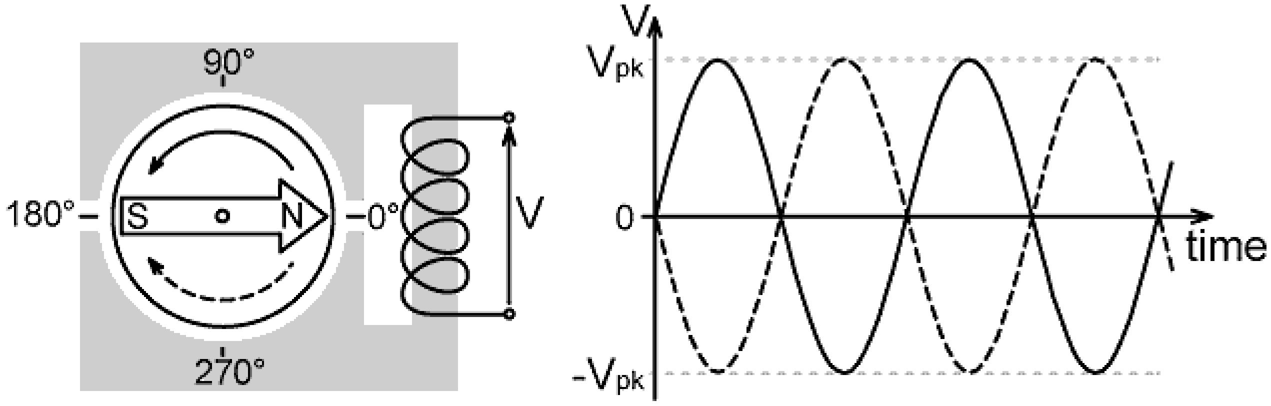
$\sqrt{(x^2)} = \pm x$

So, the resonance formula stated explicitly becomes:

$f_0 = \pm 1/[\ 2\pi\sqrt{(L\ C)}\ ]$

and there are two solutions, numerically identical but of opposite sign. By convention, we usually assume the positive result; but since there were no restrictions on the validity of the arguments used in deriving the formula, the negative frequency solution must exist and must mean something.

AC electrical theory, as we delve more deeply into it, will present us with various little conundrums (often involving square roots); and although negative frequency is one of the most trivial, we will be ill prepared for the others if we simply let it pass. The negative frequency solution arises because a sinusoidal waveform is derived from circular motion, and there are two possibilities for this: clockwise and anti-clockwise. This does not mean that the positive and negative frequencies are identical however, as the following argument will illustrate: Consider an alternator (mechanical AC generator) stopped at a position where it would give zero output voltage (or current) if it were spinning, and where it will give an initially positive output if its shaft is turned anti-clockwise (see illustration below). Now, if the shaft is turned *clockwise*, the output will initially go negative. It follows that the difference between the positive and negative frequency outputs is that while the voltage (or current) associated with one is positive the voltage associated with the other is always negative, and vice versa. Hence changing the sign of a frequency has the effect of shifting the phase of the associated waveform by 180°. Incidentally, for anyone who might insist on taking the direction of rotation analogy too seriously; it is of course obvious that if the generator is an electronic oscillator, the concept of rotation is meaningless. In that case, the negative frequency solution can be obtained by swapping the connections (as it can with any generator or resonator).



5. Impedance, resistance, reactance

The basic electrical laws discussed earlier tell us that resistors consume power when connected across a generator but that perfect inductors and capacitors do not. The combination formulae then tell us how to deal with resistances or reactances in series and parallel; but they do not tell us how to deal with combinations of resistance and reactance. This is a serious limitation, which can only be overcome by introducing the generalised concept of *impedance*, i.e., the theory of two-terminal devices that obey Ohm's law but do not necessarily consume all of the electrical power delivered to them.

A concept that needs to be formalised at this point is that of a linear, passive, two terminal network. An electrical device is *linear* if its graph of voltage versus current is a straight line, i.e., if it obeys Ohm's law; and it is *passive* if it contains no sources of energy. The general term 'network' is used because, although the theory we are about to develop covers 'simple' devices like capacitors and resistors, it also covers any combination (actual or hypothetical) of resistances and reactances in series and parallel connected to a single pair of terminals. A network can be hypothetical in the sense that it behaves in the same way as (i.e., serves as a model for) an actual two terminal device. For example, when an antenna system is connected as a load to a radio transmitter, we can treat it as a hypothetical network of resistances and reactances. An antenna, incidentally, is not completely passive, because it also receives radio signals, but we can model the receiving case by considering it to be exactly the same network as in the transmitting case, but with one or more generators connected in series with it.

Any linear passive two-terminal network can be regarded as an impedance. This means that its electrical behaviour at a particular frequency can be explained by invoking two (and only two) mutually independent properties; namely resistance and reactance. Resistance R is that property of the network that enables it to dissipate (i.e., consume or dispose of) energy, and reactance X is that property that enables it to store energy. It is also a special property of our Universe that; while energy dissipation is, on average, a one-way process; the electromagnetic energy storage mechanisms come in the form of a complementary pair. This means that true resistance is always positive (a statement that we will qualify later), but there are two opposing types of reactance, which of course we know as inductive reactance, $X_L = 2\pi fL$, and capacitive reactance, $X_C = -1/(2\pi fC)$. Inductive reactance arises through the storage of energy in a magnetic field, and capacitive reactance through the storage of energy in an electric field. When inductance, capacitance, and resistance are combined within the same two-terminal 'black box', the opposing reactances will always tend to cancel-out to some degree, and so the two types of reactance make only a single contribution to the impedance at any particular frequency. There is however, no way in which resistance and reactance can be combined to form a single numerical quantity, because the physical processes they represent turn out to be mutually exclusive.

A natural distinction arises between resistance and reactance because perfect energy dissipation implies that the Poynting vector never reverses; whereas perfect storage and return implies alternation, with the Poynting vector spending equal amounts of time in the two possible flow directions. Hence, for a resistance; when the instantaneous voltage is positive, the current is positive; and when the voltage is negative, the current is negative; i.e., the voltage and current waveforms are perfectly in phase. For the Poynting vector to alternate and give zero average power delivery however, there must be a ±¼-cycle difference between the voltage and current waveforms. It is the 0° difference in the resistive case, and the ±90° difference in the reactive case, that gives rise to a condition of mathematical independence, or othogonality, which we can exploit to obtain a generalised form of Ohm's law. Once we have that generalisation, the rules of combination follow and give rise to a complete and internally consistent AC theory.

For DC circuits, we can write Ohm's law as $V = I R$. For AC circuits therefore, we must suspect that we can write something along the lines of $V = I Z$, as long as we recognise that impedance, Z, the generalised attribute of objects that obey Ohm's law, must be represented by some composite quantity containing two distinct elements R and X. In circumstances such as this, it is traditional to see if anyone has developed a branch of mathematics that suits the problem, and the clue regarding where to look lies in the independence of R and X. If two quantities are completely independent, they must in some sense exist in different dimensions (i.e., they always move at right-angles to each other). This means that impedance cannot be represented by an ordinary number, i.e., a one-dimensional quantity lying on a line between $-\infty$ and $+\infty$, it must be represented by a point on a two-dimensional plane, which is another way of saying that Z can be plotted as a point on a graph of R against X. With regard then to solving problems involving impedance, it so happens that we are spoilt for choice, because there are no less than two appropriate branches of mathematics, namely *vectors* and *complex numbers*. The vector approach traditionally preferred by engineers is that of making sketches or graphs, and using trigonometry to work out the actual numbers; whereas the complex number approach is algebraic, in that it allows equations involving two-dimensional objects to be written-down and re-arranged. Both approaches are equivalent however, and sometimes one can clarify the other, and so we will adopt a notation and a way of thinking that enables us to switch freely between them.

6. Vectors & scalars

A ***vector*** is, by definition, a mathematical object that must be described by two or more independently variable numbers. Impedances, as we have noted, fall into this category; and so vectors can be used to describe them. One very useful property of vectors is that they can be mixed with ordinary numbers and manipulated using the normal rules of arithmetic, provided that the rules are generalised to accommodate them. Because vectors *are* different from ordinary numbers however, it is helpful to note each one as a letter in a **bold typeface** (or in handwriting by putting a little arrow above the symbol), and an optional comma-separated list in brackets may be included to denote its extent in its various dimensions. Thus we can represent an impedance as $\mathbf{Z}(R , X)$, by which we mean that **Z** is characterised by an amount R in the resistance dimension and an amount X in the reactance dimension. In the context of vectors, ordinary numbers are known as ***scalars***, because the effect of multiplying a vector by a scalar is to scale it (i.e., magnify or shrink it) without otherwise changing it. Thus, if s is a scalar, we can write:

$$s\mathbf{Z}(R , X) = \mathbf{Z'}(sR , sX)$$

Note also, a widely used mathematical notation, which is to use an apostrophe or "prime" (') to indicate that an object has been modified.

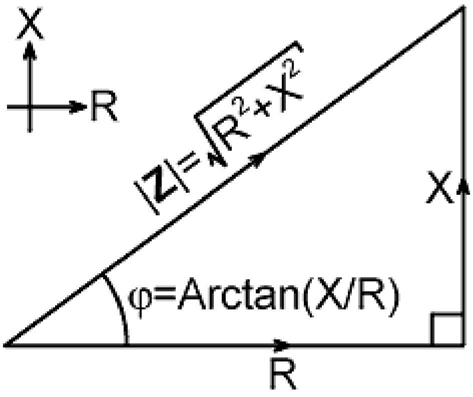
We can immediately deduce a rule for adding vectors by observing that two quantities will only add together if they exist in the same dimension (you can't increase the length of an object by adding to its width). Thus, if we want to add two impedance vectors $\mathbf{Z}_1(R_1 , X_1)$ and $\mathbf{Z}_2(R_2 , X_2)$, i.e., find out what happens when the impedances are placed in series, all we have to do is add the R parts and the X parts separately to find the new impedance $\mathbf{Z}(R_1{+}R_2 , X_1{+}X_2)$. This operation is indicated by the '+' symbol, just as in ordinary arithmetic, i.e. if

$$\mathbf{Z}(R , X) = \mathbf{Z}_1(R_1 , X_1) + \mathbf{Z}_2(R_2 , X_2)$$

then $R = R_1 + R_2$ and $X = X_1 + X_2$

One point in treating impedances as vectors, is that it enables us to draw diagrams in order to visualise what is going on. We can do this by representing an impedance as a line in a plane, with a particular length and orientation. In this sense, a vector diagram is like a navigation chart, with the distances, in this case, measured in Ohms. Mathematicians calls such maps 'spaces', by analogy with ordinary space; and a space in which distance is measured in Ohms is called ***impedance space***. Now observe that although the R and X parts of an impedance exist in different dimensions, they both exist in the same space because they are connected by the fact that they are measured using the same units (i.e., Ohms). We may therefore deduce that the difference between a space and a graph is that all of the axes in a space must be labelled in the same units; whereas the axes of a graph can have different units (e.g., temperature vs. time). You may, of course, have heard of four-dimensional space-time, which appears to disobey the rule just stated, but in fact the unit of the fourth physical dimension is not time but *the speed of light multiplied by time*, i.e., ct. The units of ct are metres per second × seconds, i.e., metres, and so Einsteinian space has four dimensions with units of length.

Working in impedance space; if we adopt the standard convention that resistance increases to the right and reactance increases upwards, we can obtain the line representing an impedance by plotting a point, then moving right by a distance R and upwards by a distance X (or downwards if X is negative), and plotting another point. The length of the line that joins the two points is called the *magnitude* or 'modulus' of **Z**, and is written |**Z**| (and pronounced "mod Z"). The magnitude is always positive by definition, and is obtained by using Pythagoras' theorem (the square on the hypotenuse of a right-angled triangle is equal to the sum of the squares of the other two sides). Hence:



$\lvert \mathbf{Z} \rvert = +\sqrt{(R^2 + X^2)}$	**6.1**

$\tan\varphi = X / R$ $\cos\varphi = R / \lvert \mathbf{Z} \rvert$ $\sin\varphi = X / \lvert \mathbf{Z} \rvert$

Notice also that the definition of magnitude has a meaning for ordinary numbers because they can be regarded as a one-dimensional vectors. Hence, if s is a scalar:

$$|s| = +\sqrt{(s^2)}$$

i.e., the effect of taking the magnitude of an ordinary number is simply to remove the sign (+ or -).

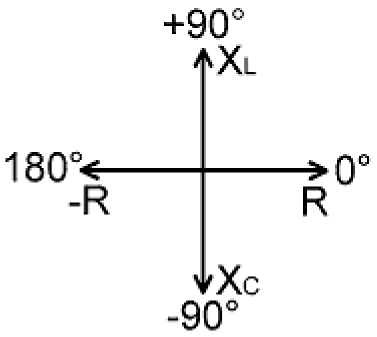
The direction of **Z** is given by the angle φ (lower-case "phi") it makes with the horizontal (resistance) axis, which is the angle whose tangent is X/R, i.e.,

$$X/R=\tan\varphi$$

Hence:

$\varphi = \arctan(X/R)$	**6.2**

('Arctan' is sometimes written as $\tan^{-1}$, but the unambiguous Arc- notation is to be preferred).



Note that φ can be positive or negative; and in particular, if we adopt the standard trigonometric convention that a positive angle is obtained by going anti-clockwise from zero (see diagram right), φ will be positive for an impedance with an inductive reactance and negative for an impedance with a capacitive reactance.

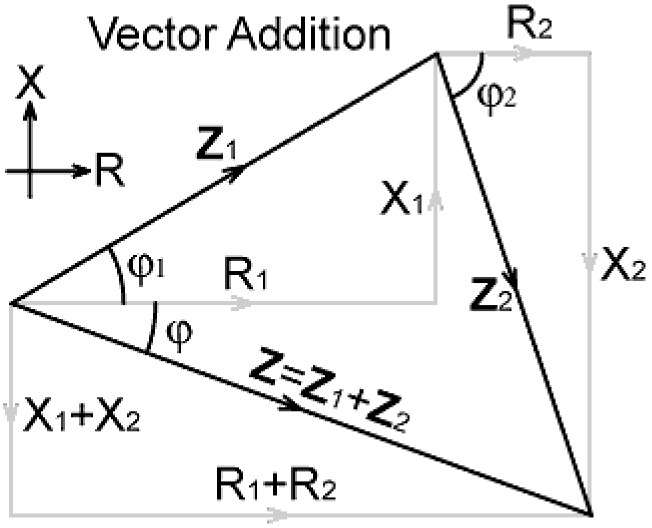
Notice also that $|\mathbf{Z}|$ and φ, taken together, provide a complete characterisation of a two-dimensional vector and so give us an alternative way of recording its properties. The form introduced earlier: $\mathbf{Z}(R , X)$ is known as the ***rectangular form*** because it contains a list of values in dimensions chosen to be at right-angles to each other. The alternative: $\mathbf{Z}(|\mathbf{Z}| , \varphi)$ is known as the ***polar form***, because it uses polar co-ordinates (distance and bearing). The polar form uses different units in its two dimensions (Ohms, degrees or radians); whereas the rectangular form has the same units in both dimensions (Ohms, Ohms). There is no ambiguity between the rectangular and polar forms because the list in brackets is optional, and a vector has the same properties regardless of how it is defined. Also, if a specific vector quantity is to be noted by putting actual numbers into the brackets, a degrees (°) symbol next to the angle will indicate that the polar form is intended (unless the angle is given in radians of course).

We can now regard equations (**6.1**) and (**6.2**) as the transformations that take a two-dimensional vector from the rectangular to the polar form. The reverse transformations are obtained from the standard trigonometric relations: $\text{Cos}\varphi = R / |\mathbf{Z}|$ and $\text{Sin}\varphi = X / |\mathbf{Z}|$, i.e.,

$R = |\mathbf{Z}|\ \text{Cos}\varphi$ and $X = |\mathbf{Z}|\ \text{Sin}\varphi$

The full set of transformations is summarised in the following table:

Rectangular form		Polar form	
$\mathbf{Z}(R , X)$	→	$\mathbf{Z}(\sqrt{[R^2 + X^2]} , \text{Arctan}[X/R])$	(**6.3**)
$\mathbf{Z}(\lvert\mathbf{Z}\rvert\text{Cos}\varphi , \lvert\mathbf{Z}\rvert\text{Sin}\varphi)$	←	$\mathbf{Z}(\lvert\mathbf{Z}\rvert , \varphi)$	



If we want to add two impedances graphically, we simply place the beginning of the second against the end of the first, and draw a new line from the beginning of the first to the end of the second. Thus we get a new impedance, with a new magnitude and a new direction. This might all seem rather unnecessary in view of the simple addition rule given earlier, but the meaning of vector addition is (hopefully) obvious when it is visualised in this way. Whatever the method used in performing the arithmetic however, the point in doing it, as we shall see, is that it allows us to keep track of the relationship between the voltage applied across an impedance and the corresponding current.

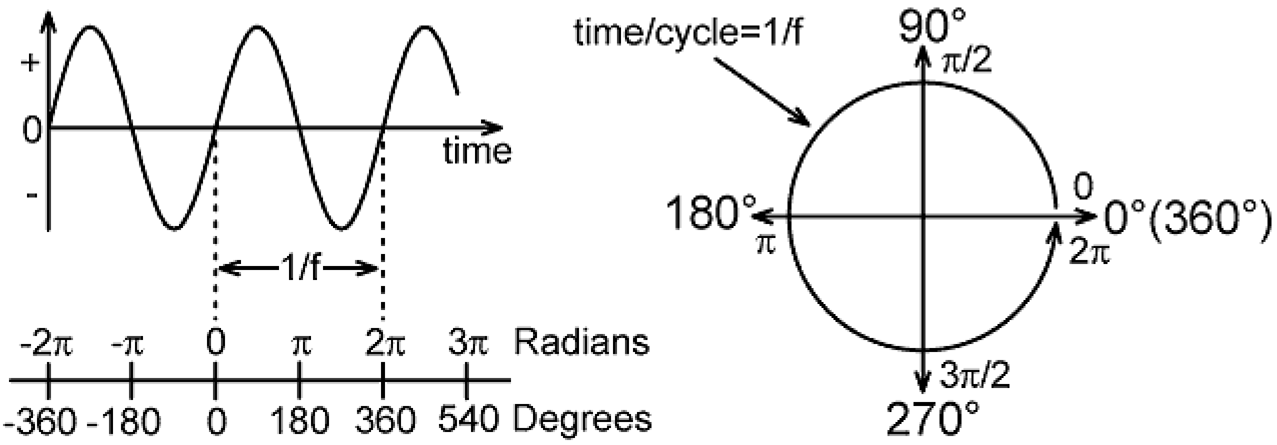
7. Balanced vector equations

It is here that we must observe the principal property of the 'equals' symbol, which is that if a given type of mathematical object lies on one side of it, then exactly the same type of object must lie on the other. Thus, now that we know that impedances are vectors, we must re-write Ohm's law in such a way that equality is never violated. There are numerous ways in which that can be done, but for the moment we will examine three possibilities:

V = **I Z**	**V** = I **Z**	**I** = V / **Z**

It is by no means obvious that all of these expressions must be true; but, as we shall see, they all are when interpreted correctly.

Since impedance is a vector, then either voltage is a vector, or current is a vector; or (presuming that the product of two vectors is also a vector) both current and voltage are vectors. In fact both voltages and currents are vectors because they each have associated with them a magnitude, a frequency, and a *phase*; the phase being defined as: *the time at which a chosen event in the wave cycle occurs* (e.g., the time of zero-crossing from negative to positive in the illustration below). The generator frequency is not an independent variable in the definition of impedance, because it already appears in the reactance ($X_L=2\pi fL$, $X_C= -1/[2\pi fC]$), and so we may deduce that the direction of the impedance vector (φ) constitutes phase information, i.e., it gives us the time difference (in degrees or radians) between corresponding events in the voltage and current waveforms. Hence φ is known as the *phase angle*, and can be converted into a time difference in seconds by dividing a complete cycle of the waveform into 360° or 2π radians and noting that the time-per-cycle or *period* of the waveform is 1/f.



Now note that since **I** and **V** are vectors, we can write them in polar or rectangular forms using the transformations (6.3) given earlier, i.e.,

I(|**I**| , φ) = **I**(|**I**|Cosφ , |**I**|Sinφ)

V(|**V**| , φ) = **V**(|**V**|Cosφ , |**V**|Sinφ)

In general, it is natural to think of currents and voltages in their polar forms; but the rectangular form is important for understanding what happens when the phase angle is either 0° or 180°. Taking a current vector as an example:

$\mathbf{I}(\,|\mathbf{I}|\,,0°\,) = \mathbf{I}(\,|\mathbf{I}|\mathrm{Cos}0°,\,|\mathbf{I}|\mathrm{Sin}0°\,)$

$= \mathbf{I}(\,+|\mathbf{I}|\,,0\,)$

and

$\mathbf{I}(\,|\mathbf{I}|\,,180°\,) = \mathbf{I}(\,|\mathbf{I}|\mathrm{Cos}180°,\,|\mathbf{I}|\mathrm{Sin}180°\,)$

$= \mathbf{I}(\,-|\mathbf{I}|\,,0\,)$

When a two dimensional vector lies along the 0° direction, either pointing with it or in opposition, its extent in one of its spatial (i.e., rectangular form) dimensions is zero; and, as in our interpretation of negative frequency given in section **4**, the minus (-) symbol is associated with a 180° phase shift (or *phase reversal*).

So, now that we know that both current and voltage are vectors, we must conclude that **V**=**IZ** is the general statement of Ohm's law. It transpires however, that we may admit the validity of the other possibilities **I**=V/**Z** and **V**=I**Z** under certain circumstances. The point is that, in AC theory, we are usually interested not in the absolute phases of the voltages and currents (i.e., the phases relative to some external reference), but in their phases relative to each other. This means that we are often at liberty to choose the direction of one of the vectors in order to learn the directions of the others relative to it. The direction chosen for this special ***reference vector*** is in principle arbitrary; but a simplification occurs if we choose it to be either 0° or 180° because Sinφ goes to zero in either case, and a vector that is zero in one of its spatial dimensions behaves, *in this context*, as though it has one less dimension. A two-dimensional vector that drops a dimension in this way, of course, becomes a one-dimensional vector, i.e., a *scalar*. Hence, whenever a voltage or current appearing in a mathematical expression is written as a scalar, the symbol can be (and, as we shall see later, *must be*) interpreted to mean that the corresponding vector is lying along the 0° axis.

A vector that transforms as a scalar in some specific context is called a ***pseudoscalar***. A pseudoscalar has the property that when its space co-ordinates are reflected with respect to the origin (0,0) it changes sign, whereas a true scalar remains unchanged[7]. Hence voltages and currents become pseudoscalars when we choose their directions to be 0° or 180°. Another electrical pseudoscalar is resistance; a special kind of impedance that can be treated as a scalar, but which becomes negative if the co-ordinates of impedance space are reversed.

If we choose the current in Ohm's law to be our reference vector, and set its phase angle to 0°, it becomes a pseudoscalar of value equal to its extent in the 0° direction; i.e., it is identifiable as the quantity |**I**|Cos0° or +|**I**|. Thus, in the relationship **V**=I**Z**, we can recognise I as the reference vector against which the phase of **V** will be determined:

$\mathrm{I} = \mathbf{I}(\,|\mathbf{I}|\,,0°\,) = +|\mathbf{I}|$

and

$\mathbf{V} = \mathrm{I}\,\mathbf{Z} = (+|\mathbf{I}|)\,\mathbf{Z}$

The pseudoscalar current I is therefore *equal* to the current magnitude |**I**| , the latter being the quantity registered by an ordinary AC ammeter (an device ignorant of phase). It is however not

[7] **Elementary Particles**, Enrico Fermi, Silliman memorial lecture series, Yale University Press, 1951. Definition of pseudoscalar: p9.

identical to the magnitude because it can be negative in principle, even if not usually in practice; i.e., if, for some reason, the reference phase is chosen to be 180°, then:

$\mathrm{I} = \mathbf{I}(\, |\mathbf{I}|\, ,\, 180°\,) = -|\mathbf{I}|$

An AC ammeter must be considered to register magnitude |**I**| rather than pseudoscalar current I because swapping the connections makes no difference to the reading, i.e., the instrument can *never* give a negative indication (and putting a minus sign in front of each of the numbers on the scale won't help, because then it will never be able to give a positive indication). We can however equate the meter reading |**I**| with the reference vector I if we want to know the phase of the voltage relative to 0°.

A similar logic applies in the case of the relationship **I** = V/**Z** , where, on the (correct) assumption that the reciprocal of a vector (i.e., 1/**Z**) is also a vector, we can identify the pseudoscalar voltage V as the reference vector against which the phase of **I** will be determined:

$\mathrm{V} = \mathbf{V}(\, |\mathbf{V}|\, ,\, 0°\,) = +|\mathbf{V}|$

|**V**| is, of course, the quantity registered by an ordinary AC voltmeter, and we can equate it to V if we want to know the phase of the current relative to 0°.

So, what we have seen here is that if one of a set of voltage or current vectors is replaced by its magnitude, it becomes a reference vector pointing at 0°. We may also deduce the converse, which is that if a vector should happen to be pointing at 0° by virtue of a choice made elsewhere, then it too can be replaced by its magnitude. It is however important to understand that there is a difference between a vector that has dropped a dimension and a magnitude, because there will be many circumstances in which we will want to use the magnitudes of vectors that are free to point in any direction. In particular, we will need this distinction later in order to generalise Joule's law. It will however become apparent that adoption of the convention that vectors written as scalars (i.e., un-bold) are pointing at 0° (or 180°) preserves the meaning of most of the DC and pure-resistance-only formulae that appear in standard textbooks. The correspondence arises because, whenever a vector is written as a scalar, a statement is made to the effect that the phase of that vector (except for the sign) can be ignored. A DC formula works for AC when the circuit contains only pure resistance because, in that case, rotating one vector to point at 0° rotates all of the others to point at 0°, and so they can all drop a dimension. Hence **V**=**IZ** becomes V=IR (for example). One consequence of all of this is that, in formulae, we should *avoid* writing voltages and currents as scalars unless we really mean them to be pointing at 0° or 180°. We must however permit a common convention, without which the notation will appear very cumbersome; which is that whenever we refer to a current or a voltage without mentioning phase we mean *magnitude*, i.e., the observable quantity that can be measured with a two-terminal meter. In other words, a measurement taken from a voltmeter may be written in isolation (say) $\mathrm{V}_{\mathbf{out}}$=27 V , but as soon as it is inserted into a formula with other vectors it acquires a phase, even if we don't need to know what that is, and must then be identified as $|\mathbf{V}_{\mathbf{out}}|$.

So, mindful of the warning that reference vectors and magnitudes are not quite the same thing, the expression **V**=I**Z**, now tells us that if we multiply an impedance by the magnitude of the current passing through it, we will obtain a vector representing the magnitude of the applied voltage and its phase relative to the phase of the current. This is an extremely useful result, and stems from the fact that the vector representation has captured the physics of the situation exactly. In effect, having observed that resistance and reactance act independently on the current, and that inductive and capacitive reactances act in opposition; we have elected to represent pure inductive reactance as a vector pointing at +90°, pure resistance as a vector pointing at 0°, and pure capacitive reactance as a vector pointing at -90°. Thus we have satisfied the requirement that the Poynting vector must alternate for reactance, but not for resistance, and we have incorporated it into the definition of impedance itself. Now however, it follows, that when some *mixture* of resistance and reactance is connected across a generator, the angle for the voltage-current phase difference will lie at some intermediate value, and the use of vectors allows this angle, the phase angle φ, to be determined from simple geometry. More to the point, a phase angle of 0° implies that an impedance will absorb all of the power delivered to it, and a phase angle of ±90° implies that an impedance will not accept any power. Thus we can observe that the phase angle represents not only the relationship between voltage and current for an impedance, but also the effectiveness with which power can be delivered to it.

8. Phasors

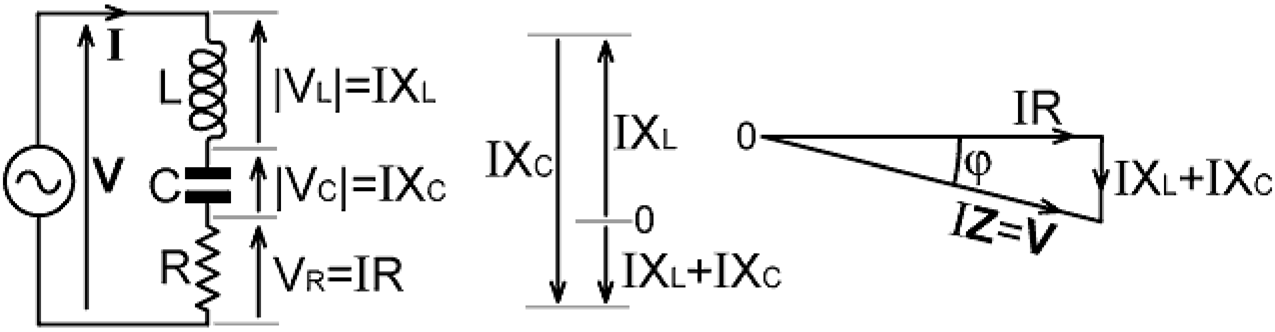
As we have just shown; one of the interpretations of Ohm's law is that, if an impedance vector is scaled by a current magnitude, it is transformed into a voltage vector. Since the act of scaling a vector does not change its direction, it transpires that both the impedance vector and the voltage vector contain the same phase information, and that this information is conserved after multiplication by a scalar. Put in plain language, this means that, although the current through an impedance will change according to Ohm's law as the applied voltage is changed, the **V-I** phase relationship will not change provided that the frequency is held constant. It is for this reason that vectors used in impedance related applications are known as '***Phasors***' (i.e., phase-vectors, or 'carriers of phase'), and diagrams involving them as '***Phasor Diagrams***'. The special properties of phasors (as distinct from vectors in general) are as follows:

- The phase co-ordinate is defined in relation to other phasors rather than to an absolute time reference.
- Time is measured in degrees or radians relative to one cycle of the frequency at which the analysis is being carried out.
- A phasor deemed to be pointing at 0° may be replaced by its magnitude, and a phasor deemed to be pointing at 180° may be replaced by the negative of its magnitude.
- Phasors are strictly two-dimensional; i.e., the vector cross product (which produces a new vector at right angles to the original two) has no meaning for phasors.

Shown below is a phasor diagram illustrating what happens when an impedance consisting of a resistance, an inductance, and a capacitance in series is connected across a generator. We can easily deduce the total impedance by inspection in this case; but notionally, it is obtained by regarding the individual series elements as phasors: $\mathbf{Z}_R(R , 0)$, $\mathbf{Z}_L(0 , X_L)$, and $\mathbf{Z}_C(0 , X_C)$; and adding them together. Thus:

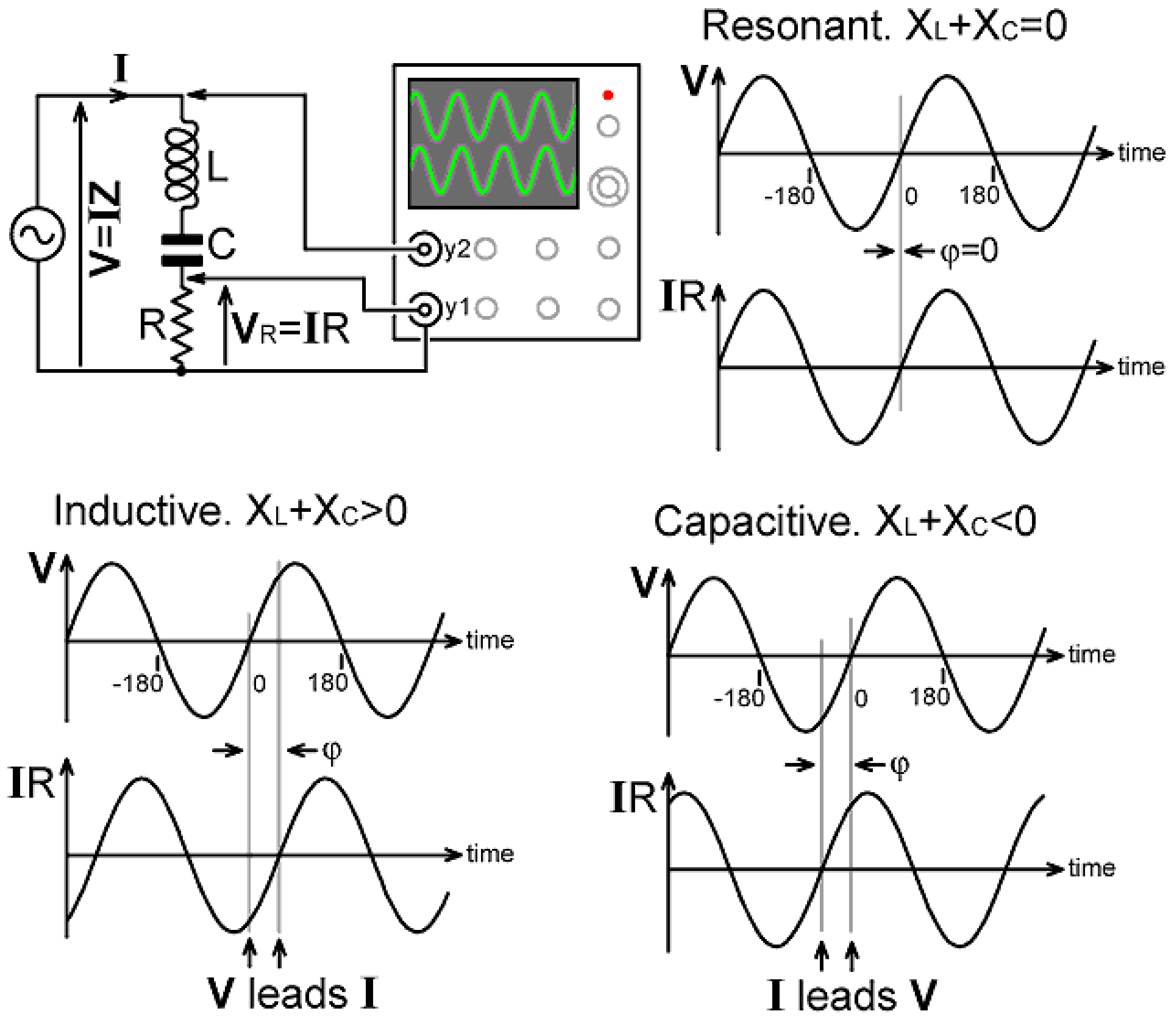
$\mathbf{Z}_R(R , 0) + \mathbf{Z}_L(0 , X_L) + \mathbf{Z}_C(0 , X_C) = \mathbf{Z}(R , X_L{+}X_C)$

We can draw the resultant phasor **Z** by moving along by a distance R and moving up by a distance X_L+X_C (or down if X_L+X_C is negative), but notice that in the diagram, the resistances and reactances have all been scaled by a reference phasor I, which is equal to the magnitude of the current. By so doing, all of the quantities have been turned into voltages, and so the diagram has become a *voltage phasor* diagram.



With regard to the physical phenomena represented here; observe that, since R, C, and L are in series, they must all carry the same current. We can deduce the magnitudes of the voltages across across the three components using Ohm's law, i.e., $|\mathbf{V}_L| = IX_L$, $|\mathbf{V}_C| = IX_C$, and $V_R = IR$ (the latter being written as a scalar because it is in phase with I and therefore pointing at 0°). We also know the relative phases of these voltages because they are all linked to the phase of the common current; i.e., the voltage IR across the resistance is in phase with the current, the voltage IX_L across the inductance is at +90° relative to the current, and the voltage IX_C across the capacitance is at -90° relative to the current. We can therefore add these three voltages as vectors to obtain the magnitude of the generator voltage and its phase relative to the phase of the current; although in the diagram the voltages across the two reactances have been added first to produce the more diagrammatically convenient quantity IX_L+IX_C (this being the voltage across the total reactance in the system). Note that the voltages across the two reactances *always* tend to cancel because there is a fixed 180° phase difference between them, and so the magnitude of the voltage across the total reactance is always smaller than the magnitude of either IX_L or IX_C .

The relationship between the phase-angle obtained from a phasor diagram and the waveforms that can be observed using a two-channel oscilloscope is shown below
(where > means "greater than" and < means "less than") :



Here we have obtained a waveform that is exactly in phase with the current by measuring the voltage across the resistive component (bottom trace). When this is compared against the waveform of the total voltage **V** (using the upward zero-crossing as an arbitrary reference point), we find that **V** is advanced in time (i.e., leading) relative to **I** when the impedance is inductive ($X_L+X_C > 0$), and **V** is retarded (lagging) relative to **I** when the impedance is capacitive ($X_L+X_C <0$). If we call the time difference observed on the oscilloscope Δt (where 'Δ' is upper-case "Delta", a symbol normally used to mean "the difference in"), then the ratio of Δt to the time of a complete cycle is the same as the ratio of the phase angle φ to a complete circle. The time-per-cycle (also known as the *period* of the waveform) is of course the reciprocal of the frequency (1/f), hence, if φ is measured in radians:

$\Delta t / (1/f) = \varphi / (2\pi)$

i.e.,

$\boxed{\Delta t = \varphi / (2\pi f)}$

Note incidentally, that it is impossible (neglecting the use of superconductors) to make a series LCR network from which all of the resistance can be isolated; because practical inductors and capacitors always have some internal resistance. A measurement made across *any part* of the total series resistance will however always produce a voltage that is in phase with **I**. A device that measures current by sampling the voltage across a resistance is, of course, an *ammeter.*

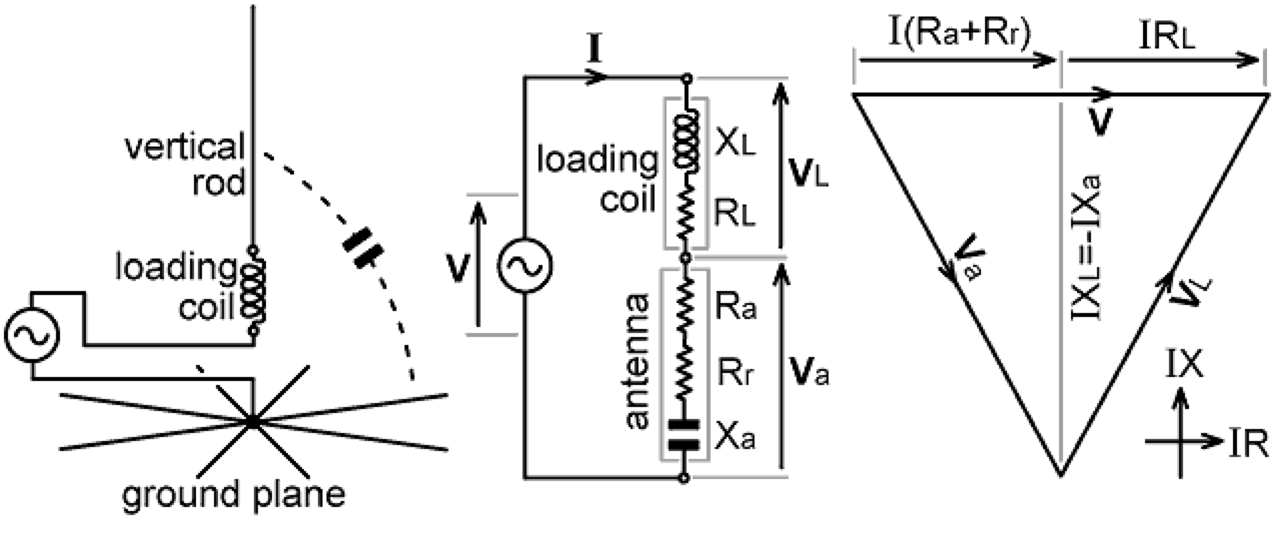
It was stated earlier that capacitive reactance is defined as a negative quantity in order to make AC theory consistent with trigonometry. The convention we follow is, of course, that which says that the phase angle of a vector increases as it rotates in the anti-clockwise direction. Hence the choice of X_C as the negative reactance stems from the fact that voltage lags (i.e., peaks 90° later than) current for a capacitor, whereas voltage leads (peaks 90° ahead of) current for an inductor. This can be remembered by considering what happens when a capacitor in series with a resistor is connected to a battery: a large inrush of current precedes the build-up of the voltage across the capacitor terminals. If the capacitor is replaced by a coil, the opposite happens; the build-up of current is delayed by a back-voltage produced by the growing magnetic field. There is no need for convoluted reasoning in AC theory however; just remembering the sign of X_C takes care of everything. Note however, that many technical articles follow the hallowed tradition of treating X_C as negative in some statements and positive in others. This is done as an aid to comprehension, because it encourages the reader to re-derive all of the mathematics in order to find out what the writer was trying to say.

9. Voltage magnification & Q

One of the curious properties of the series LCR network discussed above is that the voltages across the reactances can be much larger than the applied voltage. Take for example, an impedance consisting of a 1 Ω resistance in series with an inductance having X_L=100 Ω and a capacitance having X_C= -100 Ω, all connected across a generator giving 1 V output. In this case, the system is resonant because $X_L+X_C= 0$; and so the voltage across the total reactance, $|\mathbf{V}_X|= 0$ and the phase angle, φ=0°. Because the two reactances have cancelled each other out, the impedance looks like a pure 1 Ω resistance, but there is a current of 1 A flowing through each reactance, and so each has a voltage of 100 V across it. This voltage magnification (100:1) is referred to as the Q of the tuned circuit formed by L C and R; i.e., in this case, $Q = X_L/R$ and also $Q = -X_C/R$ (or $|X_C|/R$). Q is often taken to stand for 'quality', although that was not the basis on which the symbol was originally chosen[8]. Still, it is useful to remember that the smaller the energy loss (in this case due to a series resistance), the better the quality.

In the case of an impedance such as an antenna, of course, we cannot get inside it and measure the voltages across the individual components (and a simple series LCR combination is nowhere near complicated enough to account for the way in which antenna impedance varies with frequency). When maximising the power delivered to an antenna however, we frequently need to place the antenna impedance in series with another impedance in such a way as to create a pure resistance into which the transmitter can deliver all of its power. We would of course, like to cancel the reactance of the antenna by placing a pure opposite *reactance* in series with it, but pure reactance is unattainable, and so our compensating (or *conjugate*) reactance always brings some extra (loss) resistance with it. In this case, although the voltage appearing across the terminals of the combined impedance may be very low, the voltage across the antenna terminals can be enormous, and we must choose the voltage ratings of our matching network components accordingly.

For an illustration of the voltage magnification effect, consider the short vertical antenna system depicted in the diagram below:



In order to avoid misconceptions it is important to be aware that the vertical rod itself is not the antenna. To use the rod as a radiator, we must apply a voltage to the pair of terminals formed by it and the ground-plane, and so the antenna is the combination of the rod *and* the ground-plane. The

[8] "**Q**", Ken Smith, Electronics and Wireless World, July 1986, p51-53.

input impedance of an electrically short (less than a quarter-wavelength long, i.e., $< \lambda/4$) vertical antenna looks predominantly like a very small capacitor, which is essentially the capacitance that exists between the vertical section and the ground. A small capacitor has a large negative reactance (recall $X_C = -1/2\pi fC$), and so we need to place a large inductive reactance ($X_L=2\pi fL$) in series with the antenna to make the whole thing look like a pure resistance. If we now fill-in some of the details about the antenna and the loading coil, we are in a position to calculate the voltages across the antenna terminals and the loading coil for a given generator power, and also the overall efficiency of the complete antenna system (i.e., the proportion of the applied power that is actually radiated).

Any mechanism that dissipates energy, i.e., consumes power, must look electrically like a resistance. The resistive part of the antenna input impedance is shown to contain two components $\mathbf{R_a}$ and $\mathbf{R_r}$. $\mathbf{R_a}$ is the electrical losses of the antenna, due mainly to the RF resistance of the metal conductors used to make the rod and the ground plane and the dielectric losses (RF heating) of any insulating materials used. $\mathbf{R_r}$ is the *radiation resistance* of the antenna, i.e., a resistive component associated with energy radiated into space. Both $\mathbf{R_a}$ and $\mathbf{R_r}$ are in some sense distributed over the whole antenna, but they appear as a single resistive component ($\mathbf{R_a}+\mathbf{R_r}$) in the antenna input impedance. Take for example an antenna with a radiation resistance of 2 Ω and an input reactance of -3000 Ω. These are the approximate values to be expected for vertical section and ground-plane radial lengths of about 7% of the wavelength at the frequency of operation, i.e., 0.07λ (graphs for estimation of radiation resistance for short antennas are given in the references listed below[9] [10]). If we are very careful about the materials used in the antenna system, we might keep the loss component $\mathbf{R_a}$ down to about 0.5 Ω, so that the input impedance of the antenna will look like 2.5 Ω of resistance and -3000 Ω of reactance. To cancel the antenna reactance ($\mathbf{X_a}$=-3000 Ω), we need to place a coil having X_L=+3000 Ω in series with it. Such a 'loading coil' is normally placed out with the antenna, mainly because coils inside metal boxes have more losses than coils mounted in wide-open spaces, but even so, the coil will not be perfect and will have a distributed RF resistance that looks like another resistive component in the antenna input impedance. The amount of coil resistance is given by the Q of the coil, which is the ratio of reactance to loss resistance, i.e., $Q_L=X_L/R_L$. A well-made loading coil might have a Q of about 400, and so $R_L=X_L/Q_L$=7.5 Ω. With the reactance of the antenna now cancelled by the coil, the input impedance of the whole antenna system now looks like a pure resistance of 2.5 + 7.5 = 10 Ω .

Suppose we now decide to deliver 10 Watts (10 W) from a generator (transmitter) into this 10 Ω resistance. Knowledge of the power level enables us to calculate the antenna current and hence the voltages that appear between the various terminals, but a word of caution is in order before using the standard power formulae for this purpose. The expressions: "P=IV", "$P=I^2R$", and "$P=V^2/R$" are all deeply suspect because, if we try to convert them into vector expressions simply by changing the voltages and currents into vectors, then the equations that result will be nonsense because power (energy per unit-of-time) is scalar (strictly, pseudoscalar, as we will see later). We need to develop some additional ideas on the subject of vectors before this matter can be fully resolved, but the standard expressions *will* balance if all of the vectors involved can drop a dimension. Hence we can concede that the expressions are true for voltage and current *magnitudes*, provided that the generator is driving a purely resistive load (a somewhat restrictive condition, but we happen to satisfy it here). Thus, using the standard formulae, we can work out that the current in the antenna will be

$$|\mathbf{I}| = I = \sqrt{(P/R)} = \sqrt{(10/10)} = 1 \text{ A}$$

[9] "**Efficiency of Short Antennas**", Stan Gibilisco W1GV, Ham Radio, Sept 1982, p18-21. Graphs of radiation resistance vs. electrical length for short verticals and dipoles. Efficiency calculations.

[10] "**How long is a piece of wire?**" J J Wiseman, Electronics and Wireless World, April 1985, p24-25. Discussion of the efficiency (or lack thereof) of electrically short verticals. The effect of top loading.

and the voltage at the generator will be

$|\mathbf{V}| = V = \surd(PR) = \surd(10\times10) = 10$ V.

Note however that the current will result in a voltage of 3000 V (IX) across both reactances, and although these voltages are cancelled at the generator, we can certainly experience them as real by touching the junction between the rod and the loading coil (not recommended). In fact, the electric field-strength at the top of the loading coil is so great that a neon lamp or a fluorescent tube held there will light without any wires connected to it (see photograph below). The actual voltage appearing across the coil, $|\mathbf{V}_L|$, can be obtained by using Pythagoras' Theorem, i.e.,

$|\mathbf{V}_L| = I\surd(X_L^2+R_L^2) = I\surd(3000^2+7.5^2) = 3000.01$ V

and the voltage across the antenna terminals (i.e., the voltage between the bottom of the rod and the ground plane) is,

$|\mathbf{V_a}| = I\surd(\ X_\mathbf{a}^2 + [R_\mathbf{a}+R_\mathbf{r}]^2\ ) = I\surd(3000^2+2.5^2) = 3000.001$ V.

These voltages are barely different from the voltages across the (theoretical) pure reactances, and reflect the fact that reactance dominates the impedances of both the coil and the antenna; but despite the reactive input impedance of the antenna we have nevertheless turned it into an effective radiator. One way to look at this is to say that by resonating the antenna with a loading reactance, we exploit the voltage magnification of the resulting tuned circuit in order to force power into a reactive load. We could, of course, do the same by sheer brute force; but that would involve using a generator with an output of just over 3000 V to get a measly 1 A into the antenna.

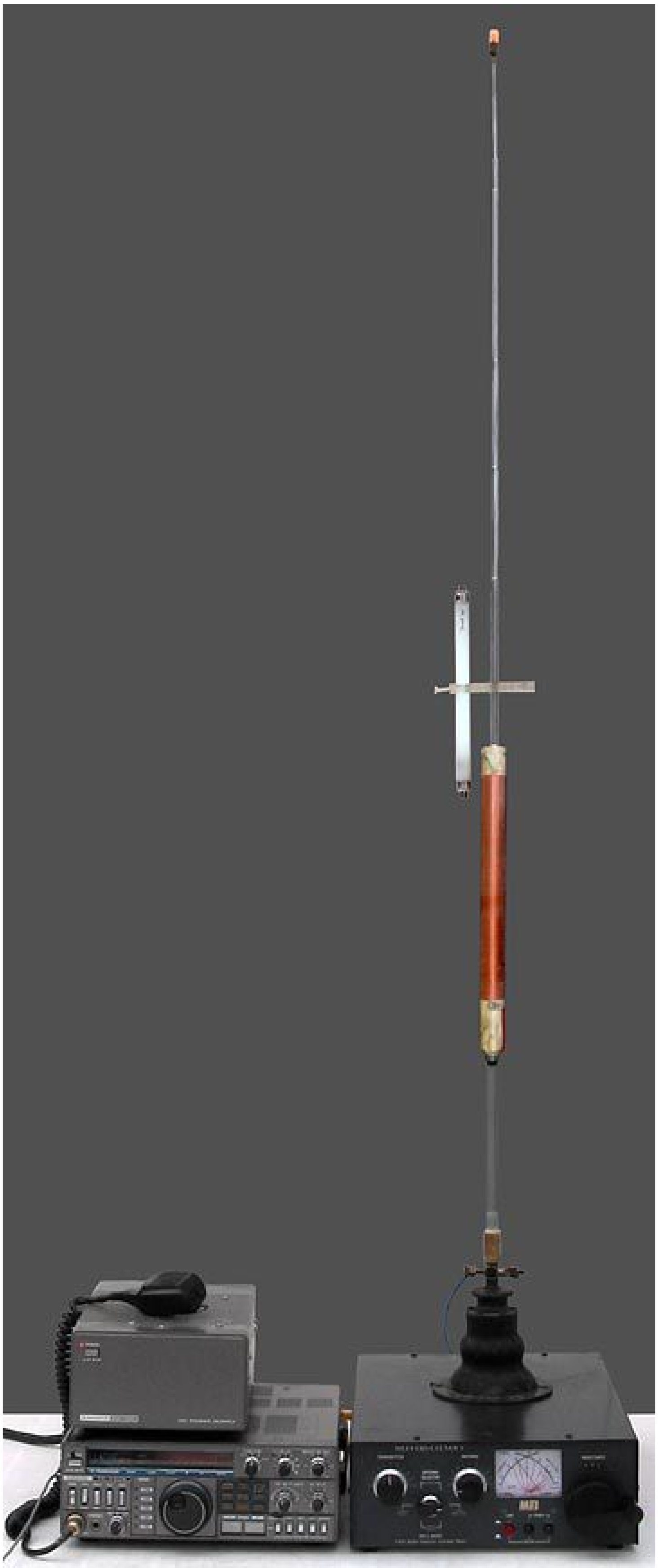
One further point to note here is that the voltages calculated by the above methods are all RMS voltages. RMS, as mentioned earlier, stands for 'square-**R**oot of the **M**ean of the **S**quares', this being a mathematical trick to find an equivalent constant (DC) voltage or current that gives the same heating effect as an alternating voltage or current. The voltages and currents referred to in the theory of impedance must be RMS values by definition, because that is the only way in which Ohm's law can be generalised to include both DC and AC (DC becomes a special case of AC with $f = 0$). The need for an RMS average arises because the ordinary average of a sinusoidal alternating voltage is zero (the voltage spends as much time being positive as it does negative). If however, we square the instantaneous voltage, we obtain a function that is proportional to the power it will deliver to a resistance. If we average that power function (i.e., take the mean of the squares) and then take the square root, we will obtain the equivalent direct voltage, i.e., the constant voltage that will deliver the same amount of power to a given resistance. The RMS average of a sine wave is the instantaneous peak value divided by the square-root of 2, i.e., $V_{RMS} = V_{Pk}/\surd2$, and $V_{Pk} = V_{RMS}\times\surd2$. Thus if we calculate a voltage of 3000 V RMS across the antenna terminals, the maximum instantaneous (peak) voltage will be $3000\times1.4142 = 4247$ V , and it is this higher figure that must be used in calculating the voltage ratings of the components used.

The final piece of information we can extract from the vertical antenna example under discussion is the efficiency of the system. In this case, all we have to do is note that the total input resistance was 10 Ω, whereas the radiation resistance was 2 Ω. This gives an efficiency of 2/10 or 20%, i.e., 2 W radiated for 10 W in, 20 W radiated for 100 W in. This incidentally, is not a disaster, and represents a good figure for a short loaded-vertical; around 10 W of SSB radiated from a reasonably high location being sufficient for worldwide short-wave communication in suitable atmospheric conditions.

Voltage magnification in action

For the antenna in the illustration on the right, the frequency of operation is 1.84 MHz and the physical length of the antenna assembly is 1.45 m from the bottom of the rubber mounting base to the top of the neon lamp soldered to the tip. The section above the loading coil is 0.76 m long (0.0047λ). The clamp holding the fluorescent tube is made from acrylic resin (Perspex), and there is no electrical connection between the tube and the whip. The glow from both lamps is visible at an input level of 1 W, but since the photograph was taken on a bright summer's day (albeit in the shade), the power input to the antenna-matching unit (AMU) was turned up to 100 W to overcome the daylight. The antenna is a 160 m band mobile whip used by the author in the 1970s. It gave a useful service (a range of several miles using about 1 W of AM) despite having an efficiency of considerably less than 1%. The long thin shape of the coil does not give maximum Q, but it does cause the coil to radiate to some extent (some of its 'loss' resistance is actually radiation resistance). The 6 W fluorescent tube was added for this demonstration, but the neon bulb at the tip was always used as a tuning aid. The generator in the photograph is a Kenwood TS430s HF transceiver with its mains power supply; and the AMU is an MFJ989C T-network. The input to the antenna is resistive when the length is adjusted correctly (about 25 Ω, mainly due to the coil), and the AMU was used to transform this resistance to 50 Ω, as required by the generator.

Those wishing to reproduce this demonstration should note that, apart from the mains lead, there is no proper ground-plane for the set-up, and the author had to tune-up wearing thick rubber gloves in order to avoid getting burnt fingers. Mounting the antenna on a car is safer.



10. Power factor & scalar product

In the preceding discussion, we observed that reactance acts as an impediment to the delivery of power to an impedance, and that the applied voltage must be increased in order to overcome it. This means that the DC power formula " P=IV ", if we interpret it to mean the product of the current and voltage *magnitudes*, is not generally true for impedances because, except in the special case that X=0 (i.e., when the impedance is a pure resistance) it will give a result that is larger than the actual power delivered. As mentioned earlier, we can deduce what is wrong with the equation in a purely abstract way by noting that **I** and **V** are phasors, whereas P is scalar. Now we will fix the problem by finding a method of vector multiplication that produces a scalar. The first step in doing so is to refer to the product of the magnitudes $|\mathbf{I}||\mathbf{V}|$ as the *apparent power*:

$P_{\mathbf{apparent}} = |\mathbf{I}|\ |\mathbf{V}|$

The true power, on the other hand, is the power dissipated in the resistive part of the impedance. It can be determined from the magnitude of the current; i.e., using a properly balanced version of the DC formula:

$P = |\mathbf{I}|^2\ R$

and if we choose **I** as a 0° reference vector:

$P = I^2\ R$

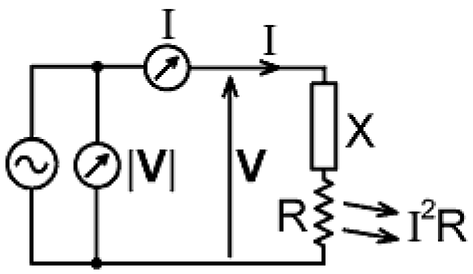
Earlier in this chapter, we showed how an impedance phasor diagram can be scaled by a reference phasor I to obtain a voltage phasor diagram (i.e., every resistance or reactance in the diagram is multiplied by I). The phasor diagram below has been scaled by I² to obtain a '*power phasor*' *diagram*. Here we should be aware that the phrase 'power phasor' is an oxymoron (i.e., a contradiction in terms like "encrypted broadcast") because average power is scalar; but *apparent power* is not power, and we *can* think of it as a vector. In particular, having set the phase of the current to be 0°, the 'phase' of the apparent power is given by the expression:

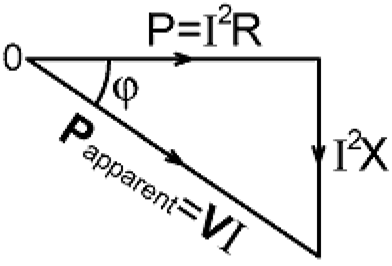
$\mathbf{P}_{\mathbf{apparent}} = I\ \mathbf{V}$

and since $\mathbf{V} = I\ \mathbf{Z}(R\ ,\ X)$, then

$\mathbf{P}_{\mathbf{apparent}} = I^2\ \mathbf{Z}(R\ ,\ X)$

which gives the definition of apparent power as: $\mathbf{P}_{\mathbf{apparent}}(I^2R\ ,\ I^2X)$
Thus the phase of **V** relative to **I** is the 'phase' of the apparent power relative to the true power ($P = I^2\ R$); and the magnitude of the apparent power is the diagonal of the 'phasor' diagram shown below:





From this, we can determine a correction factor for the $|\mathbf{I}||\mathbf{V}|$ power formula, particularly by observing that the cosine (adjacent / hypotenuse) of the phase angle is $P/(|\mathbf{V}I|)$, i.e.;

$$P = |\mathbf{V}\ I|\ \text{Cos}\varphi$$

or, after factoring out the pseudoscalar I:

$$P = |\mathbf{V}|\ I\ \text{Cos}\varphi$$

or, since $I=|\mathbf{I}|$:

$$\boxed{P = |\mathbf{V}|\ |\mathbf{I}|\ \text{Cos}\varphi}$$

Notice that Cosφ is zero for $\varphi = \pm 90°$ (no power is delivered to a pure reactance), and $\text{Cos}\varphi = 1$ for $\varphi = 0°$ (real and apparent power are the same for a pure resistance). Be aware also that the formula above appears in standard textbooks as:

" $P = V\ I\ \text{Cos}\varphi$ "

but unfortunately, there is nothing we can do to salvage this traditional version. We will prove later that the un-bold symbols V and I, when used in an equation, *must* be interpreted as phasors pointing at 0° or 180° because that is the only way in which we can incorporate DC and AC into the same theory. **V** and **I** however can only point in the same direction when $\varphi = 0°$. The standard formula is therefore internally inconsistent (a mathematical oxymoron). The best that can be said for it is that there is little choice but to assume V and I to be magnitudes in this instance, since the expression is nonsense otherwise.

The quantity $|\mathbf{V}||\mathbf{I}|\text{Cos}\varphi$ is known as the *scalar product* or *dot product* of the two vectors, and is defined in the same way for all vectors (regardless of the number of dimensions):

$$\mathbf{A}\bullet\mathbf{B} = |\mathbf{A}|\ |\mathbf{B}|\ \text{Cos}\varphi$$

It is the component (shadow length) of **B** when projected onto the direction of **A** multiplied by the length of **A** (*and* vice versa, i.e., **A** and **B** are interchangeable). Had we attacked the DC power formula " P=IV " with a foreknowledge of vector theory, we would have failed it on the grounds of dimensional inconsistency (P has too many dimensions) and deduced that the scalar product is required, i.e.,

$P = \mathbf{V}\bullet\ \mathbf{I} = \lvert\mathbf{V}\rvert\ \lvert\mathbf{I}\rvert\ \text{Cos}\varphi$	**10.1**

Instead, we attacked the problem backwards and discovered the definition of the scalar product instead. Note however, that there is a subtle difference between the general vector dot product and the phasor dot product, which will be discussed shortly.

In the context of impedance, Cosφ is known as the ***power-factor*** (PF), and is of particular interest to electricity generating companies, which prefer their customers to place pure resistances across the supply so that they do not have to run their generators into reactive loads. Thus if a load such as an electric motor is inductive as well as resistive, a suitable capacitor placed across it or in series with it can be used to cancel the reactance and bring the power-factor to unity, i.e., $\varphi = 0°$ and $\text{Cos}\varphi = 1$. This brings the apparent power into coincidence with the actual power consumed, and has the effect of minimising the consumer's electricity bill as well as minimising the stress on the generators and power transmission equipment. Thus *power-factor correction*, in relation to electricity distribution, is equivalent to the business of bringing an antenna system into resonance. The reactance-cancelling step in antenna matching, and the insertion of a loading coil into a vertical antenna, can both perfectly well be regarded as a forms of power-factor correction.

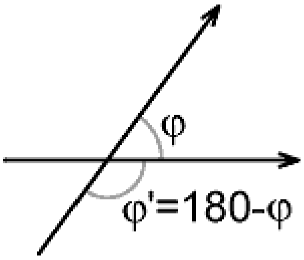
Now, since power can only be delivered to the resistive part of an impedance, only that part of voltage-multiplied-by-current that corresponds to true power (i.e., the $|\mathbf{I}|^2 R$ component) can be measured in Watts. The reason is that power (the amount of energy delivered or work done in unit time) establishes the relationship between electricity and thermodynamics, and the connection is through energy dissipation. It is therefore the convention in electrical engineering, to express apparent power in volt-amps (VA) and only true power in Watts. Many readers will already be aware that mains transformers and portable electric generators (for example) are rated in VA; the implication being that to get the full power output without over-stressing the device, it is necessary to make the apparent power in VA equal to the true power in Watts, i.e., to provide the transformer or generator with a resistive load. Since maximum power output will be associated with a particular value of load resistance; it transpires that *all* generators, not just radio transmitters, require impedance matching if the maximum allowable output is to be obtained.

One further point that should be noted on the subject of power, is that power flowing from a generator to a resistance is, by convention, positive. In DC circuits, this means that, if the voltage applied to a resistance is taken to be positive, then the current in the resistance must also be taken to be positive; and if the voltage is taken to be negative, then the current is negative.

Also note that, in AC circuits, power is calculated from RMS voltages and currents. This means that it is already an average or steady-state quantity, and there is therefore never a need to compute the RMS value (it can be done, out of mathematical curiosity, but it is numerically not the same as $|\mathbf{V}_{\mathrm{RMS}}|\,|\mathbf{I}_{\mathrm{RMS}}|\,\mathrm{Cos}\varphi$, see ref [11]). Thus the term "RMS Watts", commonly seen in the Hi-Fi literature, is nonsense and should be avoided.

[11] **"RMS watt, or not?"** Lawrence Woolf, Electronics World Dec 1998, p1043-1045.
Why $V_{\mathrm{RMS}} \times I_{\mathrm{RMS}}$ is not RMS power.

11. Phasor dot product



When the angle between two vectors is taken for the purpose of computing the dot product, there are actually two possible choices: the acute angle (< 90°); and the obtuse angle (> 90°). These are shown in the diagram on the right as φ and φ' . Now, using the trigonometric identity:

$\mathrm{Cos}(180- \varphi) = -\mathrm{Cos}\varphi$

we can see that there are two possible solutions to

$\mathbf{A}\bullet\mathbf{B} = |\mathbf{A}|\ |\mathbf{B}|\ \mathrm{Cos}\varphi$

which are numerically identical but of opposite sign. Recall from the earlier discussion that there are also two possible solutions to the taking of a magnitude (because it involves a square-root), but that by convention we take the positive answer. So it is with the dot product, in *general vector theory*; but in phasor theory it transpires that there is also a meaning to the negative solution.

Notice that when computing power from the dot product, we don't actually specify the acute angle. Strictly, the angle that must be used is the phase angle, and the power will be negative if the phase angle should happen to be obtuse. Noting that:

$P = |\mathbf{I}|^2 R$

it should be apparent that the magnitude of the phase angle will be greater than 90°, and the power will be negative, for an impedance that has a *negative* resistive component.

Negative power dissipation does not occur in nature, because it violates the principle of conservation of energy. It can occur in circuit analysis however; when a network that has been defined as passive turns out to be active. The point is that power can flow out of an 'impedance' if the network inside it should happen to include a generator.

One situation in which negative resistance can be encountered is when modelling antenna systems with multiple feed points. Due to the coupling between the different parts of the antenna, it sometimes occurs that more power flows out of one of the ports than flows in; and the computed input impedance then has a negative resistive component. The same can happen in any network with multiple ports. Normally the situation is avoided by defining the port as active in advance; but when modelling a complicated antenna, there is generally no analytical solution for the input impedance of a port (i.e., it is not possible to find a soluble algebraic expression) and the computation involves so-called 'numerical methods'. Hence the existence of an active element in a port cannot always be foreseen.

It follows that; when writing computer programs for network analysis, it is important to code for the possibility of negative resistance (rather than terminating with an error message when it occurs, or worse still, ignoring the sign). There is nothing intrinsically wrong with it; it is just an unconventional way of defining an active network, and it should not be taken to constitute a fault or inconsistency in the mathematics.

12. Complex numbers

Although the graphical 'phasor diagram' approach outlined in the previous sections is suitable for problems involving phasor addition and scaling (i.e., series networks); it is somewhat less tractable for solving problems involving phasor multiplication and division, one of particular importance being that of how to analyse networks involving *impedances in parallel*.

In section **3** we derived an expression for resistances in parallel, and also by inference an expression for reactances in parallel, i.e.,

$R = R_1 R_2 / (R_1 + R_2)$

$X = X_1 X_2 / (X_1 + X_2)$

It should come as no surprise that, if we repeat the exercise with impedances instead of reactances or resistances, applying ordinary arithmetical operations to the phasors without knowing what they mean, we end up with the expression:

$$\boxed{\mathbf{Z} = \mathbf{Z}_1 \mathbf{Z}_2 / (\mathbf{Z}_1 + \mathbf{Z}_2)}$$

The problem now is that of how to interpret this equation, a somewhat inconvenient matter if we continue to define phasors as comma-separated lists; but it transpires that there is a short-cut, due to the fact that we are only dealing with *two-dimensional* vectors, which is that such vectors can be treated as *complex numbers*.

Complex numbers were first discovered as a 'necessary evil' in solving quadratic equations, i.e., equations that can be written in the form: $ax^2+bx+c=0$. They were once described as the work of the Devil, but in fact, they merely indicate that ordinary numbers are not the whole story. Those who studied quadratic equations at school, but never got as far as complex numbers, may be surprised to learn that all of the examples they were given were deliberately chosen so as not to involve complex numbers; and that education systems in general expend more effort trying to protect students from the knowledge of complex numbers than they expend trying to teach the subject. A derivation of the general solution for all quadratic equations is shown in the box below:

General solution for quadratic equations:
Obtaining the general solution to all quadratic equations is a matter of re-arranging the general form $ax^2+bx+c=0$ so that x is all alone on one side of the equation. We can start by subtracting c from both sides, so that:

$ax^2 + bx = -c$

and then divide both sides by a , so that:

$x^2 + (bx/a) = -(c/a)$ . . . . (12.1).

We now need to find a substitution for the term $x^2+(bx/a)$ such that x is on its own. We can do that by observing that $x^2+(bx/a)$ looks similar to part of the expansion of a quantity in the form $(x+p)^2$, (where p is just an arbitrarily chosen symbol) i.e.,

$(x+p)^2 = x^2 + 2px + p^2$ . . . . (12.2)

To use this substitution, we equate the term $2px$ in equation (12.2) with the term bx/a in equation (12.1), i.e., we put $p=b/2a$ and rewrite equation (12.2) thus:

$$[\ x + (b/2a)\ ]^2 = x^2 + (bx/a) + b^2/4a^2$$

which can be rearranged by subtracting $b^2/4a^2$ from both sides to give:

$$x^2 + (bx/a) = [\ x + (b/2a)\ ]^2 - b^2/4a^2$$

Substituting this into expression (12.1) gives:

$$[\ x + (b/2a)\ ]^2 - (b^2/4a^2) = -c/a$$

and adding $b^2/4a^2$ to both sides gives:

$$[\ x + (b/2a)\ ]^2 = (b^2/4a^2) - c/a$$

We then put the terms of the right-hand side onto a common denominator, thus:

$$[\ x + (b/2a)\ ]^2 = (b^2 - 4ac)/4a^2$$

Now we can take the square root of both sides to get x on its own, but note that when a square-root is taken, there are two possibilities because $q \times q$ is the same as $(-q)\times(-q)$, i.e.,

$\sqrt{(q^2)} = \pm q$.

Hence:

$$x + (b/2a) = \pm\sqrt{[\ (b^2 - 4ac)/4a^2\ ]}$$

$$= [\ \pm\sqrt{(b^2 - 4ac)}\ ]/2a$$

finally, we subtract $b/2a$ from both sides to obtain:

$x = [\ -b \pm\sqrt{(b^2 - 4ac)}\ ] / 2a$	12.3

which is, of course, the standard school formula for solving quadratic equations.

The formula (12.3) looks innocuous enough, but what happens when 4ac is larger than b^2 ? In that case, the solution for x has a term containing the square-root of a negative number (i.e., a number that is negative when multiplied by itself) even though the basic rules of arithmetic demand that when a number is squared, the answer must always be positive.

Take, for example, the seemingly innocent quadratic equation $x^2 - x + 1 = 0$. In this case: a = 1 , b = -1 , and c = 1 , and the solution is:

$$x = (½) \pm(\sqrt{(-3)}\ )/2$$

The best simplification we can manage is to factor out the square root of -1 , i.e.,

$$x = 0.5 \pm 0.866\sqrt{(-1)}$$

Thus there are two solutions, $x = 0.5 + 0.866\sqrt{(-1)}$ and $x = 0.5 - 0.866\sqrt{(-1)}$, both of which contain a part that is a real number, and a part that is not a real number. That which is not real is *imaginary*, and so the oddball quantity $\sqrt{(-1)}$ ' was given the symbol ' i ', (by Leonhard Euler, 1707-1783) and this symbol is still used by mathematicians. When it became apparent to scientists researching into electricity that this branch of mathematics is useful however, the symbol ' i ' had already been allocated to represent current, and so the next letter in the alphabet, ' **j** ', was allocated for use in conjunction with electrical problems (here we will write the symbol in **bold**, to make it easier to spot). Thus we can write the unsimplifiable solution to the previous example as:

$$x = 0.5 \pm 0.866\mathbf{j}$$

That which is not simplifiable is *complex*, and so in this case, x is a *complex number*. **j** is called the *imaginary operator*, because it operates on a number in such a way as to make it impossible to add it to a *real* number.

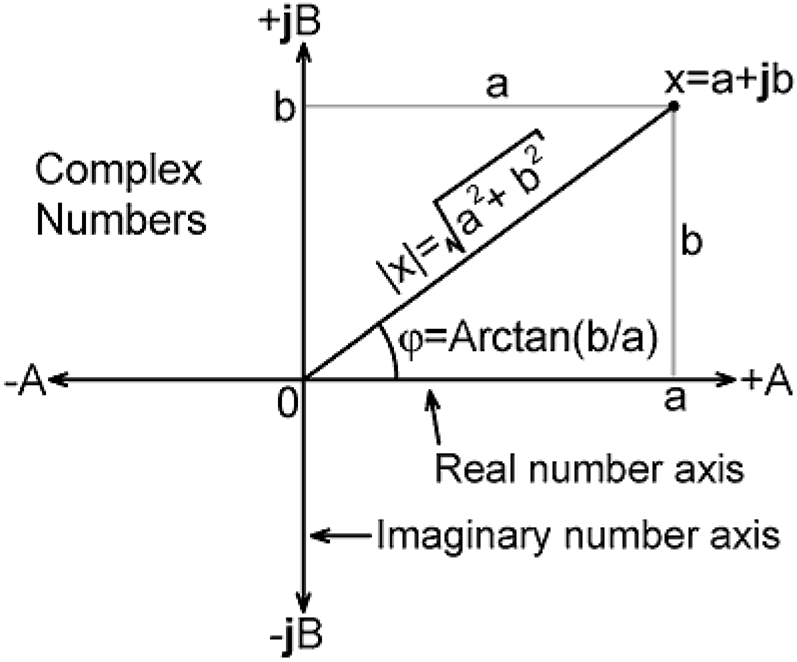
Once **j** (or i) was discovered, mathematicians went on to find general solutions for cubic equations, and quartic equations (i.e., equations involving x^3 and x^4), and it was proved that no other type of imaginary operator was required. This means that *all* numbers can be reduced to the sum of a real part and an imaginary part, and expressed in the general form:

$$x = a + \mathbf{j}b$$

with the proviso that sometimes b = 0 and the number is purely real, and sometimes a = 0 and the number is purely imaginary. Thus it is not so much that complex numbers are peculiar, but that real numbers are a special class of complex numbers that just happen to have the imaginary part equal to zero.

Once it was understood that numbers are in general complex, the next step was to work out what that meant. The clue comes from our earlier discussion of vectors. Firstly, we may observe that all real numbers must lie on a line stretching between $-\infty$ and $+\infty$. Secondly we may observe that **j** causes imaginary numbers to exist in a dimension separate from real numbers. Therefore the effect of **j** is to rotate the number-line through 90°. Thirdly, we may observe that the numbers 0 and $0+\mathbf{j}0$ are the same, so that the real and imaginary number-lines must cross at 0. The upshot is that complex numbers (i.e., all numbers) can be represented as points in a plane, which is the same as saying that the number $a+\mathbf{j}b$ can be plotted as a point on a graph of a vs. b . That graph is, of course, *number space*, and maps in this space are known as *Argand diagrams*.

We must observe, at this point, that complex numbers are so like impedances that had they been discovered by electrical engineers, they might well have been called impedances. Naturally, since complex numbers are the general class of numbers to which all numbers belong, they are essential for solving all kinds of mathematical problems, but nowhere is the association so direct and so profound that all we have to do to convert an impedance into a complex number is to write:



$$\boxed{\mathbf{Z} = \mathrm{R} + \mathbf{j}\mathrm{X}}$$

This says that impedance is a quantity with a real part R and an imaginary part X. The original terms 'real' and 'imaginary' are also perfectly appropriate, because the apparent power $P = IV_R$ dissipated in a resistance is indeed real, while the apparent power $P = I\,|\mathbf{V}_X|$ dissipated in a pure reactance is entirely *imaginary*. Thus it is hard to make a logical distinction between the two statements: "impedances can be represented by complex numbers" and "impedances *are* complex numbers".

It follows also, from the relationships implicit in Ohm's law, that if impedances can be treated as complex numbers, then so too can voltages and currents. This does not mean that these objects have somehow ceased to be vectors however, far from it. The complex number form is just another two-dimensional vector representation, which complements the rectangular and polar forms we have already met. In fact, it is merely a version of the rectangular form in which the 90° difference between the dimensions is imposed by the **j** operator; and a vector always behaves in the same way regardless of how it is defined. This minor change makes a huge difference however, because it allows a phasor to be written as an ordinary algebraic sum. An expression with **j** in it might not seem ordinary of course; but it is so in the sense that the existence of **j** is required by the rules of common arithmetic, and so **j** is by definition subject to those rules.

The complex form of a phasor makes the rectangular form effectively redundant. The transformations from the complex to the polar form are given below, and are very similar to the transformations given earlier in table 6.3.

Complex form		Polar form	
$\mathbf{Z} = \mathrm{R} + \mathbf{j}\mathrm{X}$	→	$\mathbf{Z}(\sqrt{[\mathrm{R}^2 + \mathrm{X}^2]}\,,\ \mathrm{Arctan}[\mathrm{X}/\mathrm{R}])$	(12.4)
$\mathbf{Z} = \lvert\mathbf{Z}\rvert(\mathrm{Cos}\varphi + \mathbf{j}\mathrm{Sin}\varphi)$	←	$\mathbf{Z}(\lvert\mathbf{Z}\rvert\,,\ \varphi)$	

Notice also that **j** can be regarded as a *phasor operator*, because its effect on an algebraic expression is to turn that expression into a phasor (another good reason for writing **j** in bold). Hence, in the matter of writing properly balanced vector equations, we may note that if a live phasor (i.e., one that has not been turned into a scalar by taking a magnitude or a scalar product) exists one one side of the '=' symbol, then there must be a live phasor or an expression with **j** in it (i.e., a live phasor) on the other side.

Euler's Formula:
For those familiar with exponents, note that:

$\text{Cos}\ \varphi + \mathbf{j}\text{Sin}\ \varphi = e^{\mathbf{j}\ \varphi}$

This equation is known as *Euler's formula*, and arises because imaginary exponential growth is equivalent to moving around in a circle. It defines the relationship between algebra and trigonometry; where 'e' is sometimes referred to as *Euler's number* and is, to more decimal places then you'll probably ever need: 2.718 281 828 459 045 235 360 287 471 352 662 497 757 247 093 699 959 574 966 967 627 724 076 630 353 547 594 571 382 178 525 166 427 427 466 391 932 003 059 921 817 413 596 629 043 572 900 334 295 260 595 630 738 132 328 627 943 490 763 . . . . etc., etc.

Euler's Identity:
The following expression was described by a young Richard Feynman[12] as "the most remarkable formula in math":

$e^{i\ \pi} + 1 = 0$

(where, for the purposes of a general mathematical discussion, we write i instead of **j**). This statement is known as *Euler's identity*, and manages to combine e , π , √-1 , 1 and 0, as well as the operations of addition, multiplication and exponentiation. Here we can see that it is true by noting that Cosπ = -1 and Sinπ = 0 , and putting those values into Euler's formula.

People who learned about logarithms at secondary (high) school will of course know that you can't take the logarithm of a negative number. Well, actually, you can; and Euler's identity tells us how to do it. Rearranging it gives:

$-1 = e^{i\ \pi}$

and taking the natural logarithm of both sides:

$\log_e(-1) = i\ \pi$

The subject of logarithms is discussed in more detail in **section 28.**

[12] **An Imaginary Tale. The story of √-1** , Paul J Nahin. 1998, Princeton Univ. Press. ISBN 978 0 691 12798 9. See p67. Also Ch 6

13. Complex arithmetic

Complex numbers can be added in the same way as vectors, i.e.,

$$(R_1 + \mathbf{j}X_1) + (R_2 + \mathbf{j}X_2) = (R_1 + R_2) + \mathbf{j}(X_1 + X_2)$$

and they can be scaled in the same way as vectors, i.e.,

$$s(R + \mathbf{j}X) = sR + \mathbf{j}sX$$

(it is traditional to move **j** to the beginning of the term it operates on, to make its presence more obvious).

The real power of the representation however, comes from the fact that we know immediately how to perform multiplication involving complex numbers because, although expressions having non-zero real and imaginary parts cannot be reduced to a single number, we can deal with the multiplication cross-terms by observing that $\mathbf{j}^2 = -1$. Hence:

$$(R_1 + \mathbf{j}X_1)(R_2 + \mathbf{j}X_2) = R_1R_2 + \mathbf{j}X_1R_2 + \mathbf{j}X_2R_1 + \mathbf{j}^2X_1X_2$$

$$= (R_1R_2 - X_1X_2) + \mathbf{j}(R_1X_2 + X_1R_2)$$

Thus we can multiply two complex numbers and always obtain a result that can be re-arranged into the form a+**j**b . This outcome demonstrates also that the ordinary algebraic product of two phasors, **AB** , is another phasor; and is not the same as the dot (scalar) product **A•B** . The ordinary product is known as the *complex product*, or the *phasor product*, (and is also *not* the same as the cross product used in general vector theory). The statement $\mathbf{j}^2 = -1$ incidentally, is the same as saying that rotation of a number through 90° followed by another rotation through 90° has the effect of reversing its original direction, i.e., multiplying it by -1 .

We now have part of the solution of how to interpret the expression: $\mathbf{Z}=\mathbf{Z}_1\mathbf{Z}_2/(\mathbf{Z}_1+\mathbf{Z}_2)$. One further trick is required in order to cope with the division part of the problem however, and this comes from noticing what happens when the complex number a+**j**b is multiplied by the complex number a-**j**b :

$$(a + \mathbf{j}b)(a - \mathbf{j}b) = a^2 + \mathbf{j}ab - \mathbf{j}ab - \mathbf{j}^2b^2$$

$$= a^2 + b^2$$

a-**j**b is called the *complex conjugate* of a+**j**b , and vice versa. An asterisk is normally used to denote the complex conjugate of a number, e.g., if **Z**=R +**j**X , then **Z*** = R -**j**X (**Z*** is pronounced "Z-star"). When a number is multiplied by its complex conjugate, the result is *always* real. Thus if **j** appears in the denominator (the bottom part) of a fraction, we can multiply both the numerator (the top part) and the denominator by the complex conjugate of the denominator. Multiplying both the top and bottom of a fraction by the same number makes no difference to the value, but the operation makes the denominator real, so that the fraction can then be rearranged into a form that looks, once again, like a+**j**b .

We now have a complete set of definitions for mathematical operations involving phasors, and thus armed, we are in a position to attack the parallel impedance problem.

14. Impedances in parallel

If $\mathbf{Z}_1 = R_1 + \mathbf{j}X_1$ and $\mathbf{Z}_2 = R_2 + \mathbf{j}X_2$, what is the impedance $\mathbf{Z} = R + \mathbf{j}X$ that results from placing $\mathbf{Z}_1$ in parallel with $\mathbf{Z}_2$?

$$\mathbf{Z} = \frac{\mathbf{Z}_1\,\mathbf{Z}_2}{(\mathbf{Z}_1 + \mathbf{Z}_2)} = \frac{(R_1 + \mathbf{j}X_1)(R_2 + \mathbf{j}X_2)}{(R_1 + R_2) + \mathbf{j}(X_1 + X_2)}$$

$$\mathbf{Z} = \frac{(R_1R_2 - X_1X_2) + \mathbf{j}(R_1X_2 + X_1R_2)}{(R_1 + R_2) + \mathbf{j}(X_1 + X_2)}$$

Now multiply numerator and denominator by the complex conjugate of the denominator:

$$\mathbf{Z} = \frac{[(R_1R_2 - X_1X_2) + \mathbf{j}(R_1X_2 + X_1R_2)]\,[(R_1 + R_2) - \mathbf{j}(X_1 + X_2)]}{[(R_1 + R_2) + \mathbf{j}(X_1 + X_2)]\,[(R_1 + R_2) - \mathbf{j}(X_1 + X_2)]}$$

and multiply out the terms in the denominator to show that it is now real:

$$\mathbf{Z} = \frac{[(R_1R_2 - X_1X_2) + \mathbf{j}(R_1X_2 + X_1R_2)]\,[(R_1 + R_2) - \mathbf{j}(X_1 + X_2)]}{(R_1 + R_2)^2 + (X_1 + X_2)^2}$$

The terms in the numerator are now multiplied out and rearranged so as to separate the real and imaginary parts, i.e., the numerator is put into the form a+jb as follows:

$$\mathbf{Z} = \frac{(R_1R_2-X_1X_2)(R_1+R_2)+(R_1X_2+X_1R_2)(X_1+X_2) + \mathbf{j}[(R_1X_2+X_1R_2)(R_1+R_2)-(R_1R_2-X_1X_2)(X_1+X_2)]}{(R_1 + R_2)^2 + (X_1 + X_2)^2}$$

Simplification of this expression involves multiplying out the brackets and crossing out any pairs of terms that are equal and opposite:

$$\mathbf{Z} = \frac{\begin{array}{l} R_1^2R_2 + R_1R_2^2 - X_1X_2R_1 - X_1X_2R_2 + X_2X_1R_1 + R_1X_2^2 + X_1^2R_2 + X_1X_2R_2 \\ \qquad + \mathbf{j}[R_1^2X_2 + R_1R_2X_2 + X_1R_1R_2 + X_1R_2^2 - X_1R_1R_2 - R_1R_2X_2 + X_1^2X_2 + X_1X_2^2] \end{array}}{(R_1 + R_2)^2 + (X_1 + X_2)^2}$$

Which leaves us with:

$$\mathbf{Z} = \frac{[R_1^2R_2 + R_2^2R_1 + R_1X_2^2 + R_2X_1^2] + \mathbf{j}[R_1^2X_2 + R_2^2X_1 + X_1^2X_2 + X_2^2X_1]}{(R_1 + R_2)^2 + (X_1 + X_2)^2}$$

This solution can be written in various ways, depending on preference; e.g.:

$$\boxed{\mathbf{Z} = \frac{[\, R_1R_2\,(R_1+R_2) + R_1X_2^2 + R_2X_1^2 \,] + \mathbf{j}[\, X_1X_2\,(X_1+X_2) + X_1R_2^2 + X_2R_1^2 \,]}{(\, R_1 + R_2\,)^2 + (\, X_1 + X_2\,)^2}} \tag{\textbf{14.1}}$$

or:

$$\boxed{\mathbf{Z} = \frac{[\, R_1(R_2^2+X_2^2) + R_2(R_1^2+X_1^2) \,] + \mathbf{j}[\, X_1(R_2^2+X_2^2) + X_2(R_1^2+X_1^2) \,]}{(\, R_1 + R_2\,)^2 + (\, X_1 + X_2\,)^2}} \tag{\textbf{14.1a}}$$

The real part of expression (**14.1**) is R, and the imaginary part is X, and so we can write:

$$\boxed{R = \frac{R_1R_2\,(R_1+R_2) + R_1X_2^2 + R_2X_1^2}{(R_1 + R_2)^2 + (X_1 + X_2)^2}} \quad \text{and} \quad \boxed{X = \frac{X_1X_2\,(X_1+X_2) + X_1R_2^2 + X_2R_1^2}{(R_1 + R_2)^2 + (X_1 + X_2)^2}}$$

Or, alternatively, using expression (**14.1a**):

$$\boxed{R = \frac{R_1(R_2^2+X_2^2) + R_2(R_1^2+X_1^2)}{(R_1 + R_2)^2 + (X_1 + X_2)^2}} \quad \text{and} \quad \boxed{X = \frac{X_1(R_2^2+X_2^2) + X_2(R_1^2+X_1^2)}{(R_1 + R_2)^2 + (X_1 + X_2)^2}}$$

The formula (and variants) given above for impedances in parallel, while not exactly memorable, has the advantage of being completely general. First note that if we put $X_1=0$ and $X_2=0$, then all of the reactive terms vanish and we are left with the formula for resistors in parallel, i.e., $R=R_1R_2/(R_1+R_2)$. Similarly, if we put $R_1=R_2=0$, we end up with the parallel reactance formula $X=X_1X_2/(X_1+X_2)$. More usefully however, we can put only $X_2=0$ and find out what happens when a resistance is placed in parallel with an impedance, and we can put $R_2=0$ and find out what happens when a pure reactance is placed in parallel with an impedance. The latter operation is of particular importance in the matter of devising and analysing antenna matching networks.

15. Dimensional consistency

The solution to the parallel impedance problem is our first example of what might be called a 'messy' mathematical derivation. As such, it is fairly typical of circuit analysis problems, which involve no difficult logical steps, but tend to expand into large numbers of terms, many pairs of which subsequently turn out to be equal and opposite and so cancel. Thus the problem expands alarmingly, and then contracts again into one or more relatively simple expressions.

It can be difficult to keep track of the various parts of the equation when carrying out such manipulations, which means that mistakes are likely to occur. There is however a simple reality-check, which identifies invalid terms and gives an immediate indication of the likely correctness of the result. This is the test of ***dimensional consistency*** that, with a certain amount of practice, can be carried out at a glance. The rules are as follows:

- If two quantities are to be added together (or subtracted) it must be possible to express them in the same units.

It would make no sense to add a distance in metres to a temperature in °C. It would also make no sense to add a distance in metres to a distance in centimetres, but in that case the distance in centimetres can be divided by 100 to convert it into metres, and then the addition can be performed. It follows, that if a '+' or a '-' symbol appears anywhere in an equation, the dimensions of the quantities on either side of that symbol must be the same.

- An equation, which supposedly represents a certain quantity, must have dimensions appropriate to that quantity.

Take, for example, the expression for the real part of two impedances in parallel, as derived above:

$$R = \frac{R_1R_2(R_1+R_2) + R_1X_2^2 + R_2X_1^2}{(R_1 + R_2)^2 + (X_1 + X_2)^2}$$

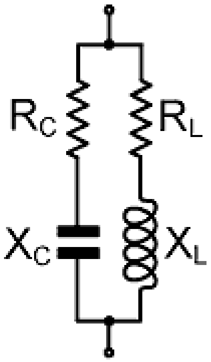
The truth of this statement is not immediately obvious, but a check of dimensional consistency can very quickly tell us if it is capable of being true. In this case, the denominator (the bottom part) of the fraction has two brackets each containing quantities having the units of resistance (Ohms). Hence the terms $(R_1+R_2)^2$ and $(X_1+X_2)^2$ have dimensions of $[\Omega^2]$, and the overall dimensions of the denominator are $[\Omega^2]$. In the case of the numerator; there are three terms to be added, each having the dimensions of $[\Omega^3]$, and the overall dimensions of the numerator are $[\Omega^3]$. Dividing the dimensions of the numerator by the dimensions of the denominator we obtain: $[\Omega^3]/[\Omega^2]=[\Omega]$; and so the equation is dimensionally consistent and represents a quantity that can be expressed in Ohms.

It is also possible to test the dimensional consistency of equations involving mixed units. The point here is that units have aliases that are composites of other units; and so we can check any equation, provided that we know the relationships between the units used. In the context of circuit analysis, these relationships are easily obtained, because they are embedded in the basic formulae from which the mathematical argument is constructed. Ohm's law, $\mathbf{V}/\mathbf{I}=\mathbf{Z}$, for example, tells us that Ohms are equivalent to Volts divided by Amperes, and so a quantity having the latter dimensions, i.e., a voltage divided by a current, may legitimately replace a quantity measured in Ohms. Similarly, the reactance laws $X_L=2\pi fL$ and $X_C= -1/2\pi fC$, tell us that Ohms can also be replaced with [Henrys × radians / second], or by [1/(Farads × radians / second)]. Thus we should not be confused by structures such as:

$$\mathbf{Z} = R + \mathbf{j}\{\, 2\pi fL - 1/(2\pi fC) \,\}$$

The bracket after the **j** *is* internally consistent, and represents a quantity measured in Ohms.

16. Parallel resonance



In an earlier section, we said that it is not possible to calculate the exact resonant frequency of a parallel tuned circuit, nor the impedance that it presents at resonance, without taking the resistances of the coil and capacitor into account. Now, of course, having derived a general equation for impedances in parallel, we are in a position to rectify that omission. The network we need to analyse is shown on the right; where R_C is the so-called equivalent series resistance (ESR) of the capacitor, and R_L is the loss resistance of the coil, which we previously defined as $R_L=X_L/Q_L$ (here Q is given the subscript 'L' to indicate that it is the Q of the coil, not the overall Q of the tuned circuit). For the purposes of this discussion, we will assume that both R_C and R_L are predominantly due to the RF resistance of the wires and other conducting materials used to make the components, and for reasons that are explained in subsequent articles[13], are considerably larger than the DC resistances. For the types of components used in HF antenna matching applications, R_C will be of the order of 0.1Ω, and R_L typically a few ohms.

In the general electronic literature, several different definitions are used for the resonant frequency of a parallel tuned circuit; the alternatives being the frequency at which the impedance of the circuit has its largest magnitude, and the frequency at which $X_L= -X_C$. Here however, we will adopt the most straightforward definition, which is the frequency at which the impedance is purely resistive (also known as the 'unity power-factor frequency'). We can find this frequency by setting the imaginary part equal to zero in equation (**14.1**) given earlier, i.e.:

$$X = [X_C X_L (X_C+X_L) +X_C R_L^2 +X_L R_C^2] / [(R_C + R_L)^2 +(X_C + X_L)^2] = 0$$

(Where the subscripts 1 and 2 have been changed to C and L as befits the current problem).
Now notice, that to make the reactance equal to zero, we only need to make the numerator of this expression equal to zero, i.e., we can ignore the denominator. Hence:

$$X_C X_L (X_C +X_L) +X_C R_L^2 +X_L R_C^2 = 0 \qquad ... \quad (\mathbf{16.1})$$

We now need to make the frequency dependence of this expression explicit by using the substitutions: $X_C = -1/2\pi f_0 C$, and $X_L=2\pi f_0 L$, i.e.:

$$-(2\pi f_0 L / 2\pi f_0 C)(2\pi f_0 L - 1/ 2\pi f_0 C) - (R_L^2 / 2\pi f_0 C) + (2\pi f_0 L R_C^2) = 0$$

The resonant frequency can now be found by re-arranging this expression to get f_0 on its own. Also, since we know that the series-resonance formula is an approximation for the expression we are about to derive, we expect the result to look like the series-resonance formula with an additional correction term or factor. We can begin by multiplying-out the first two brackets. Hence:

$$-(2\pi f_0 L^2 / C) + (L / 2\pi f_0 C^2) - (R_L^2 / 2\pi f_0 C) + (2\pi f_0 L R_C^2) = 0$$

Now we will put all of the terms containing $2\pi f_0$ on one side, and the terms containing $1/(2\pi f_0)$ on the other.

$$2\pi f_0 (LR_C^2 - L^2/C) = (1 / 2\pi f_0) [(RL^2/C) - (L/C^2)]$$

Then multiply both sides by $2\pi f_0$, and divide both sides by $LR_C^2 - L^2/C$:

[13] Components and Materials. www.g3ynh.info

$(2\pi f_0)^2 = [(R_L^2/C) - (L/C^2)] / (LR_C^2 - L^2/C)$

and factor-out 1/LC from the right-hand side:

$(2\pi f_0)^2 = (1 / LC) (R_L^2 - L/C) / (R_C^2 - L/C)$

Here we will also multiply top and bottom by -1 to put the L/C terms first, L/C generally being much larger than the resistance-squared terms, hence:

$(2\pi f_0)^2 = (1 / LC) ([L/C] - R_L^2) / ([L/C] - R_C^2)$

which rearranges to:

$$f_0 = \frac{1}{2\pi\sqrt{LC}}\sqrt{\left[\frac{(L/C)-R_L^2}{(L/C)-R_C^2}\right]} \tag{16.2}$$

Thus we find that the resonant frequency of a parallel tuned circuit is the same as that for a series tuned circuit except for a correction factor $\sqrt{[(L/C - R_L^2)/(L/C - R_C^2)]}$, which is usually close to unity. Notice that this factor is equal to 1 if R_L and R_C are zero; and also that the factor is 1 when $R_L = R_C$.

Example:
A 3μH coil is connected in parallel with a 42pF capacitor. The approximate resonant frequency is:

$1/(2\pi\sqrt{[LC]}) = 1/(2\pi\sqrt{[3\times10^{-6}\times42\times10^{-12}]})$

$= 14.178649$ MHz.

In the region of 14MHz, the coil has a loss resistance of 2Ώ and the capacitor has an equivalent series resistance (ESR) of 0.1Ώ. Thus L/C=71428.57, $R_L^2=4$, and $R_C^2=0.01$. Hence the correction factor is:

$\sqrt{[(71428.57-4)/(71428.57-0.01)]} = 0.99994414$.

The precise resonant frequency (to the nearest 1Hz) is therefore

0.999944 × 14.178649 = 14.178253MHz.

The quantity L/C is called the "'L C Ratio" of the tuned circuit (and it has units of 'Ohms squared'). Note that:

$L/C = - X_L\, X_C = |X_L\, X_C|$

It will turn out that the L/C ratio is an important parameter of resonant circuits. Also, there is some precedent for referring to the square root of the L/C ratio as the *characteristic resistance* of the tuned circuit, by analogy with the *characteristic impedance* of a lossless transmission line, which is

$R_0 = \sqrt{(L/C)}$

where *L* is inductance per unit of length and *C* is capacitance per unit of length; but the lengths cancel and so the characteristic resistance of an ideal transmission line is the square root of its L/C ratio.

In the example given above, the resonant frequency differed from the ideal case by only 0.0028% or 396Hz, the reason being that the L/C ratio was very large in comparison to the squares of the loss resistances. In HF radio applications, the L/C ratios of tuned circuits are generally in the order of several tens of thousands of Ω^2, whereas the value tolerances of radio components are seldom better than 1% and often considerably worse. In order to obtain an exact resonant frequency, it is necessary to make either the coil or the capacitor adjustable; and the required adjustment range will easily swallow any deviation caused by using the ideal-case formula $f_0=1/[2\pi\sqrt{(LC)}]$. We can therefore conclude that, in normal circumstances, the assumption of zero losses is perfectly acceptable when calculating the resonant frequency of a parallel-tuned circuit; but, as we will see in the next section, it is not acceptable when calculating the impedance at resonance.

17. Dynamic resistance

For an ideal parallel tuned circuit (i.e., $R_L = 0$ and $R_C = 0$), the impedance becomes infinite at resonance. This, of course, does not happen in practice; but provided that the loss resistances of the components are small, it does rise to a high value. Since we have defined resonance as the frequency at which the reactance is cancelled, this impedance is also purely resistive, and it is known as the *dynamic resistance* of the parallel tuned circuit. Here we will give it the symbol R_{p0} (effective parallel resistance when $f = f_0$). It is, of course, given by the real part of equation (**14.1**) (the parallel impedance formula given earlier); i.e.:

$$\boxed{R_{p0} = \frac{R_L R_C(R_L+R_C) + R_L X_C^2 + R_C X_L^2}{(R_L+R_C)^2 + (X_L+X_C)^2}} \qquad (\mathbf{17.1})$$

In the example from the previous section we had: $R_L = 2\ \Omega$, $R_C=0\ .1\ \Omega$, $L = 3\ \mu H$, $C=42$ pF , $f_0=14.178253$ MHz , $X_C = -267.2687112\ \Omega$ and $X_L=267.2537728\ \Omega$. If we apply the above formula to these data, we obtain:

$$R_{p0} = [\ 0.42 + 2(71432.56399) + 0.1(71424.57907)\ ] / [\ (2.1)^2 + (-0.0149384)^2\ ]$$

$$= 150008.0059 / 4.410223156$$

$$= 34.0137\ k\Omega$$

The only problem with equation (**17.1**) is that it is very cumbersome (and resistant to simplification). We might therefore be inclined to look for some simplifying assumptions; and the most obvious of these is to note that since X_L is very nearly equal to $-X_C$, we might as well assume the term X_L+X_C to be zero. This also implies that $X_C^2 = X_L^2 = -X_C X_L = L/C$, hence equation (**17.1**) becomes:

$$R_{p0} = [\ R_L R_C\,(R_L+R_C) + (L/C)(R_L+R_C)\ ] / (R_L+R_C)^2$$

i.e.

$$\boxed{R_{p0} = [\ R_L R_C + (L/C)\ ] / (R_L+R_C)} \tag{17.2}$$

Notice that this formula has lost all of its reactance terms, which is very convenient. If we apply it to our example data, where $L/C = 71428.57\ \Omega^2$, we obtain:

$$R_{p0} = (2\times0.1/2.1) + 71428.57/2.1$$

$$= 0.095 + 34013.61\ \Omega$$

$$R_{p0} = 34.0137\ \mathrm{k}\Omega$$

The approximation is almost exact for components of moderate Q. Also we may observe that the term $R_L R_C/(R_L+R_C)$ is much smaller than the term $(L/C)/(R_L+R_C)$, and given that we are unlikely to know the component resistances very accurately, we might as well drop the first term. Hence the appropriate formula for calculating the dynamic resistance is:

$$\boxed{R_{p0} = (L/C) / (R_L+R_C)} \tag{17.3}$$

This equation is an excellent approximation for the dynamic resistance, but strangely, it is not the one offered in most textbooks. The usual approximation is that, in addition to X_L+X_C being zero, the ESR of the capacitor is assumed to be zero. This causes all of the terms containing RC in equation (17.1) to disappear, and gives rise to a considerable simplification, viz.:

$$R_{p0} = R_L\, X_C^2 / R_L^2$$

i.e.,

$$R_{p0} = X_C^2 / R_L$$

If we apply this formula to our example data we obtain:

$$R_{p0} = 71432.56399 / 2$$

$$= 35.7163\ \mathrm{k}\Omega$$

In this case the deviation from the true value is 1702.6Ω, or 5%, which might be a reasonable approximation for many purposes, but needs to be treated with caution. Also, the failure to eliminate reactance from the formula makes computation more difficult.

18. Double-slash notation

In geometry, the expression: ' AB//CD ' means: "the line drawn from point A to point B lies in parallel with the line drawn from point C to point D". Hence, by existing convention, the symbol // means "in parallel with". In electrical engineering, of course, we are frequently interested in circuits in which components are connected in parallel, and so we can usefully adapt the double slash notation to have a non-geometric meaning. We can, for example, re-state our basic parallel component formulae as follows:

$R_1 // R_2 = R_1 R_2 / (R_1 + R_2)$

$X_1 // X_2 = X_1 X_2 / (X_1 + X_2)$

$\mathbf{Z}_1 // \mathbf{Z}_2 = \mathbf{Z}_1 \mathbf{Z}_2 / (\mathbf{Z}_1 + \mathbf{Z}_2)$

$L_1 // L_2 = L_1 L_2 / (L_1 + L_2)$

and possibly, but best avoided:

$C_1 // C_2 = C_1 + C_2$

This convention is often convenient, because it saves the bother of having to define a temporary variable to represent the parallel combination; i.e., instead of writing: "Let R represent the parallel combination of R_1 and R_2", and then having to remember what R is; we simply work with the quantity ($R_1//R_2$) , which can be expanded or calculated when necessary, but more to the point is just a resistance with an obvious definition.

While straightforward however, the use of the // notation involves a subtlety that lies in the distinction between physical and mathematical objects. In describing a test procedure, for example, we might put an entry in a table: "Test load: 68 Ω // 100 pF". The item "68 Ω // 100 pF" is a physical object, a capacitor in parallel with a resistor, but it is *not* a complete mathematical statement of impedance and cannot be treated as an impedance in any calculation. In order to turn the parallel combination into a mathematical object; we must ensure that the quantities on either side of the // symbol are of the same type and that they are expressed in the same units. In this case we can fix the problem by noting that, if the report is to have any useful meaning, a test frequency must be stated somewhere. If that frequency is, say, 14 MHz, then the reactance of the capacitor becomes

$-1/(2\pi fC) = -113.7$ Ω, and its impedance (assuming that losses are negligible) is 0-**j**113.7 Ω. Hence we can re-state the test load as (68 // -**j**114) Ω . This is the same as saying (68+**j**0 // 0-**j**114) Ω ; and is, of course, a complete statement of the load impedance in the form $\mathbf{Z}_1//\mathbf{Z}_2$ that can be converted into the R+**j**X form if so desired.

A particular logic emerges from these observations and is summarised in the box below:

18.1) **A resistance is an impedance**.

Resistances and impedances are the same type of object. A resistance in parallel with an impedance is an impedance. A resistance is simply an impedance that happens to have its imaginary part equal to zero. This means, incidentally, that the preferred pseudoscalar symbol for **Z** is usually R, rather than Z.

18.2) **A reactance is not an impedance**.
The statement:

$\mathbf{Z} = 68 \,//\, 114$

has a completely different meaning to the statement:

$\mathbf{Z} = 68 \,//\, -\mathbf{j}114$

(the first is a resistance in parallel with a resistance, the second is a resistance in parallel with a reactance). Mathematically, a *reactance* cannot be combined directly with an impedance, but a reactance can be converted into an impedance by multiplying it by **j**. Looking at this another way: impedance and reactance have reference directions that are 90° apart. To make them compatible, it is necessary to rotate one of them through 90°.

18.3) **Scalability is preserved**.
When the double slash notation is used to create a mathematical object, i.e., the same type of phasor exists on both sides of the // symbol, it has the useful property that a common factor can be multiplied-in or divided-out of the parallel object, i.e.:

$s\mathbf{Z}_1 \,//\, s\mathbf{Z}_2 = s\,(\mathbf{Z}_1 \,//\, \mathbf{Z}_2)$

Proof:

$s\mathbf{Z}_1 \,//\, s\mathbf{Z}_2 = s\mathbf{Z}_1\, s\mathbf{Z}_2 \,/\, (s\mathbf{Z}_1 + s\mathbf{Z}_2\,) = s\,\mathbf{Z}_1\,\mathbf{Z}_2 \,/\, (\mathbf{Z}_1 + \mathbf{Z}_2\,) = s\,(\mathbf{Z}_1 \,//\, \mathbf{Z}_2)$

18.4) **The associative rule**.
The double slash notation can be extended to represent any number of impedances in parallel:

$\mathbf{Z}_1 \,//\, \mathbf{Z}_2 \,//\, \mathbf{Z}_3 \,//....//\, \mathbf{Z}_n = 1 \,/\, [\,(1/\mathbf{Z}_1) + (1/\mathbf{Z}_2) + (1/\mathbf{Z}_3) + .\,.\,.\,.\,. + (1/\mathbf{Z}_n)\,]$

and the associative rule of arithmetic (and linear electrical devices in parallel) is obeyed, i.e.:

$(\mathbf{Z}_1 \,//\, \mathbf{Z}_2) \,//\, \mathbf{Z}_3 = \mathbf{Z}_1 \,//\, \mathbf{Z}_2 \,//\, \mathbf{Z}_3$

18.5) **Double-slash product definition**.
The // notation implies a specialised kind of phasor multiplication, which we might call the ***double-slash product*** or the ***parallel product*** of a pair of phasors. Since its use in conjunction with parallel capacitors is pointless, we will adopt the following strict mathematical definition:

$$\boxed{\mathbf{a} \,//\, \mathbf{b} = \mathbf{ab}/(\mathbf{a}+\mathbf{b})}$$

19a. Parallel-to-series transformation

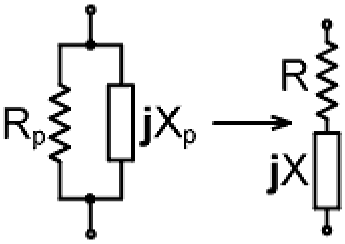
In the discussion so far, we have adopted the habit of representing every impedance as a resistance *in series* with a reactance. It makes good sense to do so in most circumstances, because it allows the impedance to be written directly in the form R+**j**X. There are many situations however, in which circuit analysis can be simplified by representing an impedance as a resistance in *parallel* with a reactance. The two possible representations are equally valid; but it should be obvious from the parallel impedance equation (**14.1**) derived earlier, that the parallel representation for a particular impedance requires a different combination of resistance and reactance to that of the series representation. In the next section we will explore the relationships between the two representations, beginning with the transformation of an impedance from its parallel to its series form:

To derive this transformation, we simply regard the parallel elements as two separate impedances R_p+j0 and $0+\mathbf{j}X_p$, and apply the formula for impedances in parallel (i.e., $[\mathbf{Z}_1 // \mathbf{Z}_2] = \mathbf{Z}_1\mathbf{Z}_2/[\mathbf{Z}_1+\mathbf{Z}_2]$). Hence:

$$R + \mathbf{j}X = (R_p // \mathbf{j}X_p)$$

i.e.:

$$R + \mathbf{j}X = \mathbf{j}X_p R_p / (R_p + \mathbf{j}X_p)$$



and R and X are simply the real and imaginary parts of the right hand side of this expression once it has been put into the form a+**j**b. We proceed as usual by multiplying the top (numerator) and bottom (denominator) by the complex conjugate of the denominator, thus:

$$R + \mathbf{j}X = \frac{\mathbf{j}\, R_p X_p (R_p - \mathbf{j}X_p)}{(R_p + \mathbf{j}X_p) (R_p - \mathbf{j}X_p)}$$

Which rearranges to:

$$\boxed{R + \mathbf{j}X = \frac{R_p X_p^2 + \mathbf{j}\, X_p R_p^2}{(R_p^2 + X_p^2)}} \quad (\mathbf{19.1})$$

Hence, for the series representation:

$$\boxed{R = \frac{R_p X_p^2}{(R_p^2 + X_p^2)}} \quad \text{and} \quad \boxed{X = \frac{X_p R_p^2}{(R_p^2 + X_p^2)}}$$

Further pieces of information that we can extract from the parallel-to-series transformation, and which will be useful later, are the phase-angle, magnitude and Q of an impedance in its parallel form, as follows.

Phase angle and Q of an impedance in parallel form:
The phase angle for an impedance in its series form was given earlier as expression (6.2):

$$\varphi = \text{Arctan}(X / R)$$

By using expression (**19.1**) above we can substitute for X and R to obtain:

$$\varphi = \text{Arctan}(X_p\ R_p^2 / R_p\ X_p^2)$$

i.e.,

$$\boxed{\varphi = \text{Arctan}(R_p / X_p)}$$

which also tells us that $X/R = R_p/X_p$, i.e. the ratio of resistance to reactance of an impedance in its series form is the inverse of the ratio for the impedance in its parallel form. Also, since we know that $|X|/R_{\mathbf{Loss}}$ is an expression for the Q of an electrical component, we may further note that component Q can be expressed as:

$$\boxed{Q_{\mathbf{comp}} = R_{pLoss} / |X_p|}$$

(the higher the parallel loss resistance, the higher the Q).

Magnitude of an impedance in parallel form:
The magnitude of an impedance in its series form is given by (**6.1**):

$$|\mathbf{Z}| = \sqrt{(R^2 + X^2)}$$

Substituting for R and X using expression (**19.1**) we obtain:

$$|\mathbf{Z}| = \sqrt{[\{ (R_p\ X_p^2)^2 + (X_p\ R_p^2)^2 \}/ (R_p^2 + X_p^2)^2]}$$

$$= \sqrt{[\{ R_p^2\ X_p^2\ (X_p^2 + R_p^2) \}/\{ (R_p^2 + X_p^2)^2 \}]}$$

We can take the square root of the $R_p^2X_p^2$ term and so factor it out of the square-root part of the expression, provided that we only use the positive result (magnitudes are always positive). Hence:

$$\boxed{|\mathbf{Z}| = |\ R_p\ X_p / \sqrt{\ (R_p^2 + X_p^2)}\ |} \qquad \textbf{19.2}$$

A convenient rearrangement of this expression can be obtained by forcibly factoring X_p from the denominator:

$$|\mathbf{Z}| = |\ R_p\ X_p / \{\ X_p \sqrt{\ [(R_p^2/X_p^2) + 1]}\ \}\ |$$

Now, since R_p and R_p^2/X_p^2 are always positive, we can drop the magnitude brackets to obtain:

$$\boxed{|\mathbf{Z}| = R_p / +\sqrt{\ [(R_p/X_p)^2 + 1]}} \qquad \textbf{19.3}$$

This form is particularly useful for frequency response calculations, because it allows the reactance contribution to be treated as a correction factor:

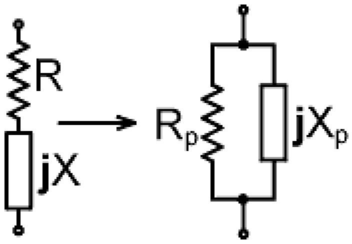
$1 / \sqrt{[(R_p/X_p)^2 + 1]}$

which goes to unity (→1) when the reactance is large in comparison to the resistance.

19b. Series-to-parallel transformation

From expression (**19.1**), we have:

$R = R_p X_p^2 / (R_p^2 + X_p^2)$ . . . . (**19.4**)



and

$X = X_p R_p^2 / (R_p^2 + X_p^2)$ . . . . (**19.5**)

Obtaining the series-to-parallel transformation is a matter of using these two equations to obtain equations for R_p and X_p. This would prove to be a somewhat tricky problem, had we not noticed from the preceding derivation of the magnitude (equation **19.2**) that:

$|\mathbf{Z}|^2 = R^2 + X^2 = R_p^2 X_p^2 / (R_p^2 + X_p^2)$

The right hand side of this equation can also be obtained by multiplying expression (**19.4**) by R_p, or by multiplying expression (**19.5**) by X_p. Hence:

$R R_p = R^2 + X^2$

and

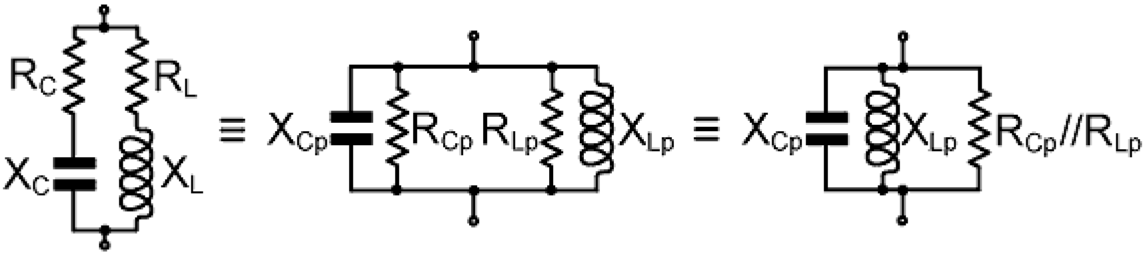
$X X_p = R^2 + X^2$

i.e.:

$R_p = (R^2 + X^2) / R$	**19.6a**
and	
$X_p = (R^2 + X^2) / X$	**19.6b**

20. Parallel resonator in parallel form

Having derived the series to parallel transformation, we are now in a position to analyse the parallel resonator in a different way. The outcome should be mathematically unsurprising, because we are bound to obtain the same results as before, but the technique will give us a new way of thinking about the circuit.



(The symbol " ≡ " means: "is by definition equal to". The symbol " // " means "in parallel with").

As the diagram above illustrates; the parallel impedance representation allows us to visualise the circuit as an ideal parallel resonator with a resistance connected across it. This separates the reactive and the resistive parts of the problem and tells us immediately that unity power-factor resonance occurs when $X_{Lp}= -X_{Cp}$, and that the dynamic resistance is given by the value of $R_{Lp}//R_{Cp}$ at f_0.

We can, of course, relate the parallel impedance form of the resonator to the series impedance form by using the transformations given in the previous section (equations **19.6**), i.e.,

$R_{Cp} = (R_C^2 + X_C^2) / R_C$ (**20.1**)

$X_{Cp} = (R_C^2 + X_C^2) / X_C$ (**20.2**)

$R_{Lp} = (R_L^2 + X_L^2) / R_L$ (**20.3**)

$X_{Lp} = (R_L^2 + X_L^2) / X_L$ (**20.4**)

Using the appropriate transformations (**20.2** and **20.4**), the resonance condition $X_{Lp}= -X_{Cp}$ becomes:

$$(R_L^2 + X_L^2) / X_L = -(R_C^2 + X_C^2) / X_C$$

which can be rearranged to:

$$X_C (R_L^2 + X_L^2) + X_L (R_C^2 + X_C^2) = 0$$

and then to:

$$X_C X_L (X_C + X_L) + X_C R_L^2 + X_L R_C^2 = 0$$

We have seen this expression before as equation (**16.1**), and so the derivation may continue as in section **16** to give the parallel resonance formula (**16.2**).

$f_0 = \{1/[\ 2\pi\sqrt{(LC)}\ ]\ \}\{\ \sqrt{[(L/C - R_L^2)/(L/C - R_C^2)]}\ \}$

The dynamic resistance R_{p0} is given by:

$R_{p0} = R_{Lp}//R_{Cp}$

Hence using the transformations (**20.1**) and (**20.3**) we have:

$$R_{p0} = \frac{[\ (R_L^2 + X_L^2)\ /\ R_L\ ]\ [\ (R_C^2 + X_C^2)\ /\ R_C\ ]}{[\ (R_L^2 + X_L^2)\ /\ R_L\ ] + [\ (R_C^2 + X_C^2)\ /\ R_C\ ]}$$

which simplifies to:

$$\boxed{R_{p0} = \frac{(R_L^2 + X_L^2)\ (R_C^2 + X_C^2)}{R_C\,(R_L^2 + X_L^2) + R_L\,(R_C^2 + X_C^2)}} \qquad (\mathbf{20.5})$$

Thus we obtain another formula for the dynamic resistance of a parallel resonator, and it is interesting to compare it with equation (**16.1**), which was our original derivation (here we show it rearranged slightly):

$$\boxed{R_{p0} = \frac{R_C\,(R_L^2 + X_L^2) + R_L\,(R_C^2 + X_C^2)}{(R_L + R_C)^2 + (X_L + X_C)^2}} \qquad (\mathbf{20.6})$$

The two formulae are radically different in appearance; but it is easy to verify, by plugging in the numbers from the example in section **17**, that they both give exactly the same answer. This leaves the issue of which one of them is the best simplification; and the answer in this case is that it is equation (**20.6**). We can tell by looking at the *power* or *degree* of the numerator and denominator of each equation. Observe first that all of the quantities involved in the expressions are measured on Ohms. Hence the numerator of (**20.5**) has dimensions of Ω^4 and the denominator has dimensions of Ω^3. In equation (**20.6**) however, the numerator has dimensions of Ω^3 and the denominator Ω^2. Hence the numerator of (**20.6**) is of lower degree than that of (**20.5**), and the denominators likewise. This means that (**20.5**) can be simplified further and ultimately transformed into (**20.6**); although for anyone who cares to try it, the manipulations required are laborious, and require the use of equation (**15.1**) as a substitution.

Something more tractable happens however when we multiply equations (**20.5**) and (**20.6**) and take the square root to obtain a new expression for R_{p0}, i.e., we take the *geometric mean* of the two formulae. In this case the denominator of (**20.5**) cancels the numerator of (**20.6**) and we obtain:

$$\boxed{R_{p0} = \sqrt{\left[\frac{(R_L^2 + X_L^2)(R_C^2 + X_C^2)}{(R_L + R_C)^2 + (X_L + X_C)^2}\right]}} \qquad (\mathbf{20.7})$$

Here we can make the following simplifying assumptions:

1) Since X_L^2 is normally much greater than R_L^2 in radio circuits, R_L^2 can be deleted from the numerator without making much difference.

2) Since X_C^2 is also usually much greater than R_C^2, R_C^2 can be deleted from the numerator without making much difference.

3) If the Qs of the resonator components are reasonably high, (X_L+X_C) is very nearly zero at resonance and can therefore be deleted from the denominator without making much difference.

The result is:

$$R_{p0} = \sqrt{\left[\frac{X_L^2\ X_C^2}{\left(R_L+R_C\right)^2}\right]}$$

This expression can be simplified by observing that everything inside the square root bracket is squared, but in doing so we must be mindful of a common fallacy. *The square root of the square of a number is not the number itself.* A square root always has two solutions, one positive, one negative; and if only one of the solutions can be true, additional information is required for selection of the correct one. In this case, we know that R_{p0} must be positive if the network is passive, and so we accept the positive square roots; but note that in section **6** we defined the positive square root of a square as a *magnitude*, i.e.;

$+\sqrt{(X^2)} = |X|$

This rule must be strictly applied, because simply deleting the superscripts and the square root symbol would have given us a negative value for R_{p0} because X_C is negative. Hence:

$R_{p0} = |X_L||X_C| / (R_L + R_C)$

We have noted before that:

$|X_L|\ |X_C| = L/C$

hence:

$$\boxed{R_{p0} = (L/C) / (R_L+R_C)}$$

which we have seen before as equation (**17.3**).

While it is instructive to attack a derivation from several directions and verify that all approaches lead to the same conclusion, the point of the parallel impedance representation is that it often makes problems easier to solve. The parallel resonator is a good example because the parallel representation gives a direct separation of the resistive and reactive parts of the problem. A further and very important point however, is that we do not use the parallel representation with a view to converting it into the series form at the earliest opportunity. It is simply another way of expressing impedance; and it is *no less authoritative* than the series form. Hence if we have data for an inductor or capacitor in series form, we can transform it into the parallel form and use it like that.

The parallel form may seem less authoritative than the series form because the expression for R_p (equation **19.6a**) has reactance in it, and so explicitly varies with frequency. In reality however, the resistive component in the series form also varies with frequency, due to a variety of frequency dependent losses such as, skin effect and dielectric absorption [see "Components and Materials"**]**, capacitive and inductive coupling to resistive materials in the vicinity of the component, and of course our old friend radiation. Thus, when solving problems using simple circuit models, we need to be aware that resistances inserted to represent losses are expected to vary with frequency, regardless of representation. Thus the practical problem of finding the dynamic resistance of a parallel resonator becomes that of measuring the impedances of the components at a frequency reasonably close to the desired resonance, transforming the losses into parallel resistances, and taking the parallel combination of those.

21. Imaginary resonance

It was observed in section 4 that the series-resonance formula:

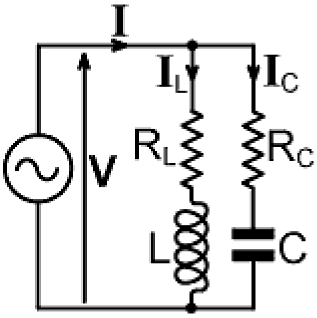
$f_0 = 1/[\ 2\pi\surd(L\ C)\ ]$

always gives two solutions for the resonant frequency, one positive and one negative. The parallel resonance formula (16.2) does the same, but presents us with a further conceptual challenge, in that it also allows *imaginary* solutions. If we inspect the formula:

$$f_0 = \frac{1}{2\pi\sqrt{LC}}\sqrt{\left[\frac{(L/C)-R_L^2}{(L/C)-R_C^2}\right]}$$

we can observe that if either R_L^2 or R_C^2 should become larger than L/C (but not both at the same time), then the quantity inside the right-most square-root bracket will become negative. Once again, there were no restrictions on the validity of the arguments that went into deriving the formula, and so imaginary resonance is possible and must have a physical meaning.

The answer to this conundrum can be obtained by considering the parallel resonator as two separate impedances connected across a generator (see diagram right). Real resonance implies a condition such that the current from the generator is in phase with the voltage it produces, i.e., it occurs at a frequency at which the resonator constitutes a resistive load. The output current **I** is the vector sum of the currents in the two branches of the resonator, i.e.;



$\mathbf{I} = \mathbf{I}_L + \mathbf{I}_C$

and so real resonance occurs when $\mathbf{I}_L + \mathbf{I}_C$ is real. Real resonance can only occur however, if the current in one branch can become large enough for its imaginary component to cancel the imaginary component of the current in the other branch. Notice that if we allow R_L to become extremely large, then practically no current will flow in the inductive branch and the circuit will not resonate. The same argument applies, of course, to the capacitive branch. The parallel resonance formula can therefore be seen to tell us that true (i.e., real) resonance cannot occur if the resistance in either branch rises above a certain critical value, that value being the square-root of the L/C ratio (the *characteristic resistance*), i.e.,

$R_0 = \surd(L/C)$

If the resistance in one branch rises above √(L/C), then the current in that branch will always be too feeble to bring the system into resonance. What happens instead is that the phase of the total current **I** can approach and move away from the phase of the generator voltage as the frequency is varied, but it is never able to reach it. The 'resonant frequency' is simply the imaginary frequency of closest approach (and it does not exist on the real frequency line). It is imaginary because the combined impedance of the two branches can never become real (i.e., resistive) by cancellation. Note however, by inspecting the circuit, that the combined impedance does become resistive at zero and infinite frequencies, but that this is not due to cancellation: At 0 Hz, $X_L = 0$ and $X_C \rightarrow \infty$ ('→' means "approaches" or "tends towards"), so the impedance is simply R_L; and at infinite frequency, $X_L \rightarrow \infty$ and $X_C = 0$, so the impedance is R_C. If a real resonant frequency does not exist therefore, what will be obtained is a network that has a voltage-current phase relationship always on one side or the other of zero degrees, only approaching 0° at zero or infinite frequency.

It was mentioned earlier that resonant circuits used in HF radio applications tend to have large L/C ratios, often greater than 10000 Ω^2. In the case of parallel resonance, one reason for this policy should now be apparent; i.e., we need to obtain a high characteristic resistance ($R_0 = +\sqrt{[L/C]}$) in order to ensure that the circuit will function properly with practically realisable inductors. A parallel resonator with an L/C ratio of 100 Ω^2, for example, will not work if the RF resistance of the inductive branch is greater than 10 Ω at the expected resonant frequency, and it is by no means impossible for a practical inductor to exceed such a limit.

We may conclude, from this discussion, that a parallel tuned circuit will only resonate usefully if $\sqrt{(L/C)}$ is made larger than the resistance in either of the branches. The qualification 'usefully' must be applied however, because if the resistance in *both* branches is allowed to become larger than the critical value, then both the numerator and the denominator of the term inside the square-root bracket will become negative, and so the term itself will be positive. Thus there will be a real resonance, but the current in both of the branches will be feeble, and so the resonance will also be feeble and of little practical use.

One final significance of the characteristic resistance that is worth remembering is that it is equal to the magnitudes of the reactances in the circuit at the 'ideal case' resonant frequency, i.e., the resonant frequency when the resistance in both branches is equal. This frequency, as was mentioned earlier, is given by the series resonance formula, i.e.,

$$f_{0s} = 1/[\ 2\pi\sqrt{(L\ C)}\ ]$$

or in radians / sec:

$$2\pi f_{0s} = 1/[\sqrt{(L\ C)}\ ]$$

Now, if we call the inductive reactance at this frequency X_{L0s}, then:

$$X_{L0s} = 2\pi f_{0s}\ L$$

$$= L\ /\ [\sqrt{(L\ C)}\ ]$$

and, since any number is the square of its own square root:

$$\boxed{X_{L0s} = +\sqrt{(L/C)}}$$

Similarly, for the capacitive reactance:

$$X_{C0s} = -1/[\ 2\pi f_{0s}\ C\ ]$$

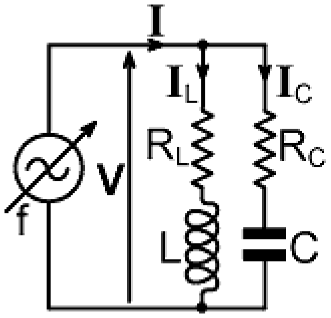
$$= -[\ \sqrt{(L\ C)}\ ]\ /\ C$$

$$\boxed{X_{C0s} = -\sqrt{(L/C)}}$$

22. Phase analysis

We can visualise the phase relationship between voltage and current in a parallel resonant circuit by deriving an expression for the **I-V** phase angle and plotting it as a graph against frequency for various values of included resistance. This is only one of the many situations in which graphs of phase vs. frequency are instructive, and so this section will serve as a general introduction to the technique of phase analysis as well as a specific investigation of the parallel resonator.

The circuit to be analysed is shown on the right, and we can use Ohm's law straight away to write an expression for the current:



$\mathbf{I} = V / \mathbf{Z}$

where **Z** is the parallel combination of the impedances in the two branches of the resonator, and we choose the phase of **V** to be 0° and treat it as a scalar. If we also define: $\mathbf{Z}_L = R_L + \mathbf{j}X_L$, and $\mathbf{Z}_C = R_C + \mathbf{j}X_C$, then:

$\mathbf{Z} = \mathbf{Z}_L\, \mathbf{Z}_C / (\mathbf{Z}_L + \mathbf{Z}_C)$

hence:

$\mathbf{I} = V\, (\mathbf{Z}_L + \mathbf{Z}_C) / (\mathbf{Z}_L\, \mathbf{Z}_C)$

Now, we noted earlier that the phase angle of a complex expression a+**j**b is given by:

$\varphi = \text{Arctan}(b/a)$

so in order to obtain the **I-V** phase difference we first write:

$\mathbf{I} / V = (\mathbf{Z}_L + \mathbf{Z}_C) / (\mathbf{Z}_L\, \mathbf{Z}_C)$

then split the right hand side of the equation into its real and imaginary parts, divide the imaginary by the real, and take the inverse tangent. Expanding the expression above we get:

$$\frac{\mathbf{I}}{V} = \frac{R_L + \mathbf{j}X_L + R_C + \mathbf{j}X_C}{(\,R_L + \mathbf{j}X_L\,)(\,R_C + \mathbf{j}X_C\,)}$$

and multiplying out the terms in the denominator gives:

$$\frac{\mathbf{I}}{V} = \frac{R_L + R_C + \mathbf{j}(X_L + X_C)}{R_L R_C - X_L X_C + \mathbf{j}(R_L X_C + X_L R_C)}$$

Now we multiply numerator and denominator by the complex conjugate of the denominator:

$$\frac{\mathbf{I}}{V} = \frac{[\,R_L + R_C + \mathbf{j}(X_L + X_C)\,][\,R_L R_C - X_L X_C - \mathbf{j}(R_L X_C + X_L R_C)\,]}{(R_L R_C - X_L X_C)^2 + (R_L X_C + X_L R_C)^2}$$

then multiply out the numerator, crossing out equal and opposite terms, to get:

$$\frac{\mathbf{I}}{V} = \frac{R_L(R_C^2 + X_C^2) + R_C(R_L^2 + X_L^2) - \mathbf{j}[\, X_C(R_L^2 + X_L^2) + X_L(R_C^2 + X_C^2)\,]}{(R_L R_C - X_L X_C)^2 + (R_L X_C + X_L R_C)^2}$$

This is in the form a+**j**b, so the phase angle is given by:

$$\boxed{\mathrm{Tan}\varphi = -\frac{X_L(R_C^2 + X_C^2) + X_C(R_L^2 + X_L^2)}{R_L(R_C^2 + X_C^2) + R_C(R_L^2 + X_L^2)}}$$

Now, there is no need to rearrange this formula any further in order to use it, but since we are analysing the phenomenon of parallel resonance, it is interesting to recall that $X_L X_C = -L/C$. If we multiply out the numerator, we will obtain two terms that contain $X_L X_C$, and this leads to an alternative expression; i.e.:

$$\mathrm{Tan}\varphi = -\frac{X_L R_C^2 + X_L X_C^2 + X_C R_L^2 + X_C X_L^2}{R_L(R_C^2 + X_C^2) + R_C(R_L^2 + X_L^2)}$$

becomes:

$$\mathrm{Tan}\varphi = -\frac{X_L R_C^2 - (L/C)X_C + X_C R_L^2 - (L/C)X_L}{R_L(R_C^2 + X_C^2) + R_C(R_L^2 + X_L^2)}$$

which rearranges to:

$$\boxed{\mathrm{Tan}\varphi = -\frac{X_L(R_C^2 - L/C) + X_C(R_L^2 - L/C)}{R_L(R_C^2 + X_C^2) + R_C(R_L^2 + X_L^2)}} \qquad (\mathbf{22.1})$$

Which, since the L/C ratio is a fixed parameter for the resonant circuit, somewhat simplifies calculation.

We will now use the expression above to evaluate the effect of resistance in a fairly representative parallel resonator. For this example we will use an inductance of 1 µH and a capacitance of 100 pF. This combination gives an L/C ratio of 10000 Ω^2 and hence a critical resistance $R_0=\sqrt{(L/C)}=100\ \Omega$. The 'ideal' resonant frequency, i.e., the resonant frequency when $R_L=R_C$ is:

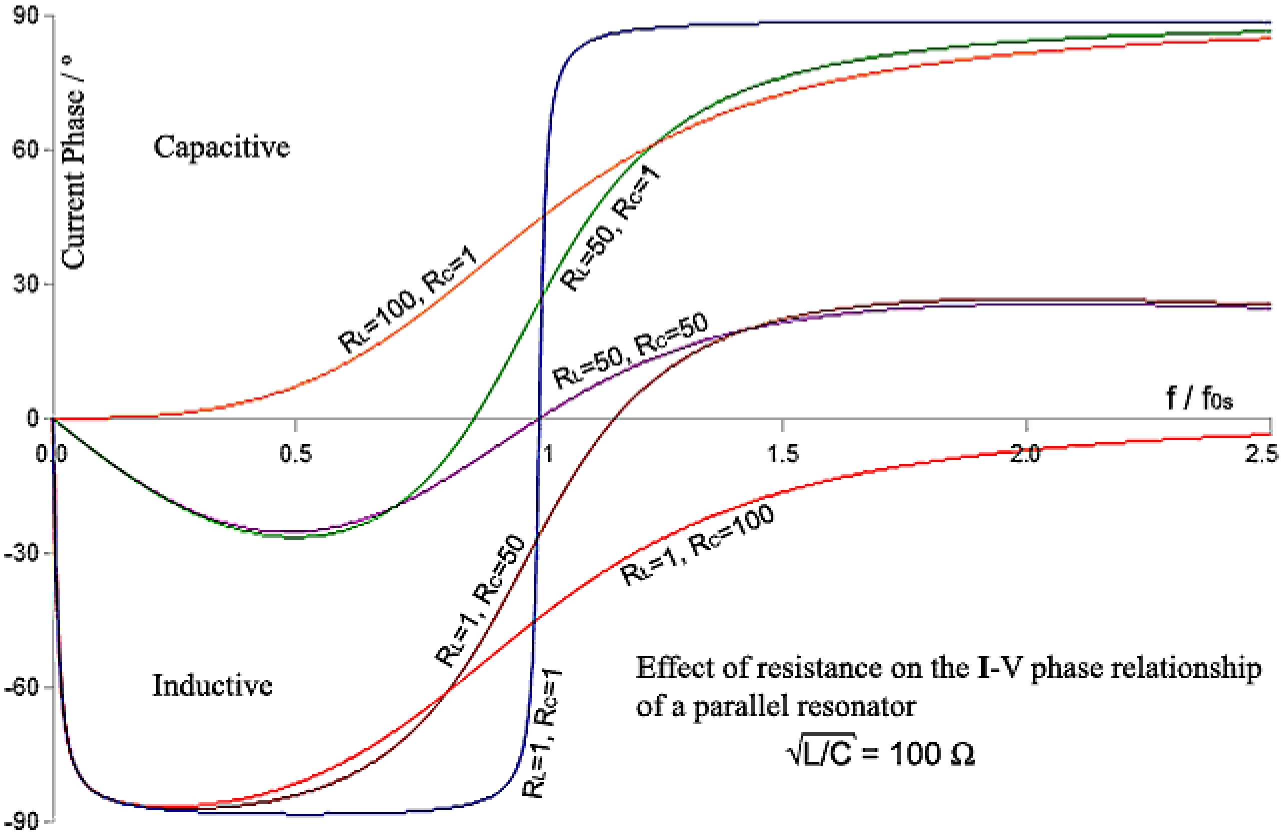
$f_{0s} = 1/[\, 2\pi\sqrt{(L\, C)}\,] = 15.91549431$ MHz

(i.e., $2\pi f_{0s} = 100$ M radians/sec), and at this frequency, $X_L = -X_C = \sqrt{(L/C)} = 100\ \Omega$.

Shown below is a set of graphs of the **I**-**V** phase relationship for our example resonator with various values of R_C and R_L between 1 Ω and √(L/C). These graphs were produced using the *Open Office Calc* spreadsheet program (available free from **OpenOffice.org**), the procedure being to create columns for frequency, X_L, and X_C, and use the calculated reactance values in the Arctangent (inverse tangent) of equation (**22.1**) given above. Note that spreadsheets often give the results of inverse trigonometric functions in radians, and so it is necessary to multiply the expression by 180/π=57.29577951 to get the result in degrees (there are 2π radians in 360°), i.e.:

$$\varphi = -57.29577951\mathrm{Arctan}\{ [X_L(R_C^2-L/C)+X_C(R_L^2-L/C)]/[R_L(R_C^2+X_C^2)+R_C(R_L^2+X_L^2)] \}$$

The plotted curves below were created using the spreadsheet "chart" tool [see accompanying spreadsheet file: **par_res_ph.ods**].

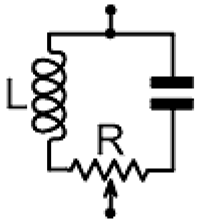


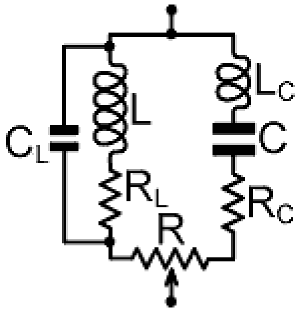
Of the curves shown, only the example with R_L=1 and R_C=1 constitutes a good healthy resonance. The choice of 1 Ω in each of the branches incidentally was made simply so that the resonant frequency would coincide with f_{0s}. Any curve with at total resistance R_L+R_C = 2 Ω will have an almost identical appearance. The Q of the resonance (as will be explained later) is 50 in this case (i.e., $Q_0=X_L/[R_L+R_C]$), which is fairly high; and so the phase of the current lags the voltage by nearly 90° at frequencies a few percent below the resonant frequency, and leads it by nearly 90° at frequencies a few percent above. Hence the circuit provides the generator with a nearly pure inductive load below resonance, and a nearly pure capacitive load above.

In the case where R_L=50 Ω and R_C=50 Ω, the Q of the resonance is 1. A large resistive component is present in the impedance at all frequencies, and so the **I**-V phase difference never approaches 90° in either direction.

The curves for R_L=50 Ω and R_C=1 Ω, and R_L=1 Ω and R_C=50 Ω, are included to show that the resonant frequency (the point where the curve crosses the zero phase-difference axis) moves to low frequency when R_L exceeds R_C, and vice versa. The curves for R_L=100 Ω and R_C=1 Ω, and R_L=1 Ω and R_C=100 Ω show that the 'resonant frequency' goes to zero when $R_L=\surd(L/C)$, and goes to infinity when $R_C=\surd(L/C)$. These results seem to indicate that the parallel resonator is infinitely tunable by means of a variable resistor, a proposition that warrants careful examination.

23. Resistance tuned LC resonator?





The parallel resonator shown in the upper diagram on the the right was offered as a "circuit idea" in Electronics World[14]; it being pointed out in the article that the tuning range is 0 to ∞ if R= √(L/C) , and the Q of the circuit is stable because the total resistance is constant. Both of these claims are true, *within the scope of the model*; but there are a couple of fatal flaws in the concept and we will address them lest people should start to believe that the circuit will work.

The author of the article was perhaps a little unsure of the 0 to ∞ claim, and so concluded that a variable resistor can give a "much wider" frequency range than a variable capacitor or inductor. We however, can straight-away dispense with the infinite upper limit by drawing the circuit model shown in the lower diagram. We might describe the first circuit as "what you try to build", whereas this one makes some attempt at simulating "what you actually get".

A capacitor is simply two pieces of electrically conducting material in proximity. The conductors do not have to be plates. Capacitance appears whenever two conductors have the ability to be at different relative voltages (i.e., capacitance is made by not shorting things together), and so there will always be some 'stray capacitance' across the coil. This precludes resonance at infinite frequency, but in fact, a coil behaves as though it has considerably more parallel capacitance than simple consideration of strays would predict. The reason is that it takes a finite amount of time for an electromagnetic wave to make its helical journey along the wire in the coil, and the resulting phase shift has to be represented by placing a hypothetical capacitance, the coil's *self-capacitance* C_L in parallel with the the idealised pure inductance. Self-capacitance is largely dependent on the length of the winding wire and the effective velocity for a wave travelling along it. This propagation velocity (the so-called *phase velocity*) is frequency dependent, but most radio coils are operated in a regime where the velocity is changing in such a way that the self-capacitance appears to take on a definite value. In this regime, the apparent self-capacitance turns out to depend only on the external dimensions of the coil (the turn-to-turn spacing and the number of turns are practically irrelevant). The inclusion of self-capacitance in the model allows for the fact that the coil has a *self-resonant frequency* (SRF) even when there is nothing whatsoever connected to it, and it is part of the HF resonator design procedure to ensure that the SRF is outside the frequency-range of interest.

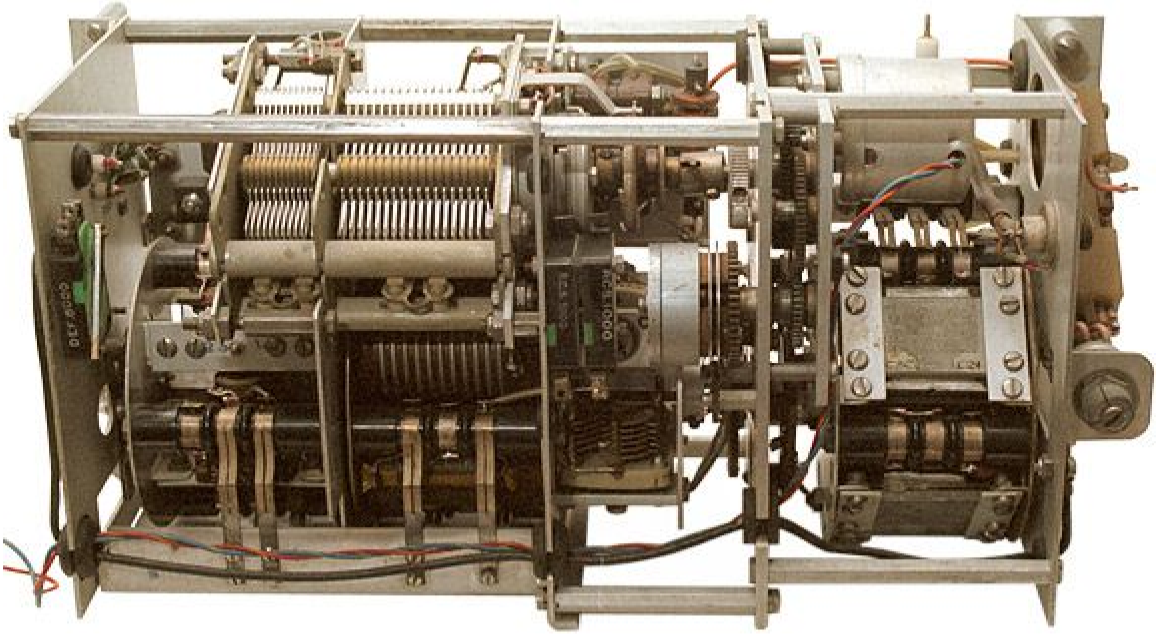
A physically small resonator coil suitable for radio receiver applications might have a self-capacitance of about 1 pF. Let us suppose therefore that this applies to the 1 μH coil from the previous example. This amount of unavoidable capacitance places an upper limit on the maximum attainable resonant frequency somewhere very roughly around $1/[2\pi\sqrt{(LC_L)}]$=160 MHz. Stray capacitance between the connecting wires will reduce this frequency, so if we construct the circuit carefully we should expect the inductive branch to self-resonate somewhere in a range from about 40 to 160 MHz.

All electrical conductors have inductance (a coil is simply a structure designed to enhance inductance by causing the magnetic fields developed by adjacent turns to add together). Hence the wires and plates involved in making up the capacitive branch of the resonator will constitute an additional series inductance, which we can model to a good approximation by imagining a small inductor L_C in series with the capacitor. For the 100 pF capacitor of our previous example, it will be very difficult to get this 'self-inductance' to be less than about 10 nH, so we might place the upper limit for the series self-resonance of the capacitive branch somewhere very roughly around $1/[2\pi\sqrt{(L_C C)}]$=160 MHz. Inductance of the connecting wires will reduce this frequency, so we should expect the capacitive branch also to self-resonate somewhere in a range from about 40 to

[14] **"An unusual tuned circuit"**, S Chekcheyev, Electronics World, Jan 2004, p41.

160 MHz.

In the days before synthesisers, simple (single-conversion) short-wave radio receivers used wide-range VFOs and variable capacitance tuning. For full short-wave coverage however, it was necessary to provide the receiver with a band-switch, one reason being that it was very difficult to obtain a tuning range of much greater than about an octave without changing coils. The larger the coil, the larger the self-capacitance, and so the band-switch selects progressively smaller coils with progressively shorter connecting wires as the frequency is increased. Winding a set of candidate coils and checking them for self-resonance will quickly indicate to the designer that a set of frequency ranges like 1-2, 2-4, 4-8, 8-16, and 16-32 MHz is easily achievable, but trying to reduce the number of bands to four (e.g., 1-2.35, 2.35-5.52, 5.52-13, 13-30.6) requires careful construction, and reducing the number to three (1-3.16, 3.16-10, 10-31.6) is very difficult using a conventional rotary switch. This does not mean that a three-band solution cannot be obtained, but it falls close to the borderline at which it becomes preferable to use an elaborate low-capacitance technique such a 'turret bandchanger', i.e., a rotating turret that carries the coils and enables them to be connected to the active circuit via very short leads (see below).



Motorised bandchanger turret from 1958 vintage Marconi AD307 aviation transmitter.

In the matter of making a resistance tuned parallel resonator therefore; our rough calculations, and observations of what others have been able to achieve in practice, seem to indicate an approximately 2:1 rule-of-thumb for the upper limit of the frequency range. What this means in this instance however, is that we should not try to push the resonator to much more than about twice its ideal-case resonant frequency; so if we consider our example 1 μH in parallel with 100 pF resonator, which resonates at about 16 MHz when the resistance in both branches is equal, we might reasonably expect to be able to tune it from 0-32 MHz. This, although not infinite, is nevertheless a phenomenal tuning range; but unfortunately, there is a catch.

The problem is that if we use a variable resistor of value equal to √(L/C), the Q of the circuit will be approximately 1. We will investigate the relationship between Q and bandwidth shortly, but we can pre-empt those findings by stating that such a circuit will be completely useless as a band-bass filter such as might be used to provide a radio receiver with selectivity. We might therefore consider raising the circuit Q to 10, by reducing the value of the variable resistor in our example √(L/C)=100 Ω resonator so that the total resistance in both branches adds up to 10 Ω. In this way, as we shall see, we sacrifice 'some' of the tuning range in order to obtain a poor but possibly useful Q. To find the tuning range that results, we can use the full parallel resonance formula:

$f_0 = \{1/[\ 2\pi\sqrt{(L\ C)}\ ]\ \}\{\ \sqrt{[(L/C - R_L^2)/(L/C - R_C^2)]}\ \}$

i.e.,

$f_0 = f_{0s}\ \sqrt{[(L/C - R_L^2)/(L/C - R_C^2)]}$

Now if, for the sake of simplicity, we assume that all of the resistance is in the inductive branch at the low frequency limit, and in the capacitive branch at the high frequency limit; the correction factor $\sqrt{[(L/C - R_L^2)/(L/C - R_C^2)]}$ becomes 0.99499 when R_L=10 Ω, and 1.00504 when R_C=10 Ω. So, for our 1 μH in parallel with 100 pF resonator, with its ideal-case resonance of 15.915 MHz, we obtain a tuning range of 15.836 to 15.996 MHz, a spectacular ±0.5% with a Q of 10. Of course, if we dispense with the variable resistor and use a variable capacitor or inductor instead, we can easily obtain a tuning range of more than 2:1, while sustaining a Q of around 50.

So much for the resistance-tuned resonator as a variable band-pass filter, but perhaps we can use it as the frequency-determining device in an oscillator? That was certainly the suggestion in the "circuit idea" article from whence it came. An oscillator is effectively an amplifier with some of its output fed back into its input via a frequency-selective network, and we can easily design an amplifier with sufficient gain to overcome the losses of the candidate circuit. The first problem however, is that the spurious self-resonances of the network are of higher Q than the desired resonance, and if nothing is done about them the circuit will probably oscillate at somewhere around the self-resonance frequency of the inductive branch. We might however decide to use an amplifier that is too slow to oscillate at VHF, or implement some kind of low-pass filter in order to ensure that the system has no gain in the troublesome self-resonance region. Thus by some artifice we might actually get the oscillator to submit to our will, at which point it will deliver its final insult by producing a signal that appears to be modulated by a hissing noise. This ill-mannered behaviour will occur because an oscillator is effectively a generator of filtered white noise. No oscillator produces a pure sine-wave. A practical RF oscillator always produces a band of noise centred on a selected resonance of the frequency-determining network, with a width determined by the network Q and the amplifier gain. The usual objective in oscillator design for radio applications is therefore to obtain as high a Q as possible, so that the desired output is a spike in the amplitude vs frequency domain sufficiently narrow to be *regarded* as a sine wave.

Thus the resistively tuned parallel LC resonator is of little practical appeal in situations demanding spectral purity. Its significance to this discussion lies instead in the fact that the "circuit idea" is a plausible fallacy; and that its appearance in an electronics publication did not generate a flurry of letters pointing out its flaws. We might comment, at this point, that it is necessary to build a circuit and try it before recommending it to others; but that is no help in finding out what went wrong if the circuit should fail to work as expected. There is a great deal of difference between a resonance and a useful resonance; and a practical circuit component operated at radio frequencies does not bear description as a pure inductance, capacitance, or resistance.

Wide-range resistance-tuned LC oscillators (operating at low frequencies) have nevertheless been built[15], but the theory of operation depends on more than the simple observation that resistance appears in the parallel resonance formula. Such circuits were used many years ago for resistance-to-frequency converters in scientific instrumentation applications, but are nowadays rendered obsolete by simple (and spectrally noisy) RC oscillators such as can be implemented using CMOS logic inverters or the 555 timer IC.

[15] "Theory and Application of Resistance Tuning", C. Brunetti and E. Weiss. Proc. IRE, June 1941, p333-344.

24. Phasor theorems

Early in this chapter we observed that the standard electrical formulae represent incomplete statements of Ohm's Law and Joule's Law. We then went on to generalise Ohm's law, but have yet to state all of its implications; and we repaired the VI power law by introducing the scalar product, but have yet to analyse Joule's law. We also introduced the the idea that if a phasor is pointing at 0° or 180° it can be treated as a scalar, a trick that obviously works, but for which we offered no convincing mathematical proof. All of these discomforts arise because of a narrative expediency, which is that of delaying the introduction of complex numbers until after that of vectors. We will now resolve all of the residual issues, with the aid of a handful of simple theorems that require complicated geometrical arguments if they are to be proved using vectors, but are easy to prove using complex numbers. These theorems incidentally are also true for real numbers, which are effectively one-dimensional vectors.

24.1 Magnitude ratio theorem:

The magnitude of the ratio of two (complex) numbers is equal to the ratio of their magnitudes.

$\lvert \mathbf{N}_1/\mathbf{N}_2 \rvert = \lvert \mathbf{N}_1 \rvert / \lvert \mathbf{N}_2 \rvert$	**24.1**

Proof:

Let $\mathbf{N}_1 = a_1 + \mathbf{j}b_1$ and $\mathbf{N}_2 = a_2 + \mathbf{j}b_2$

Then:

$$\mathbf{N}_1 / \mathbf{N}_2 = (a_1 + \mathbf{j}b_1) / (a_2 + \mathbf{j}b_2)$$

Now multiply numerator and denominator by the complex conjugate of the denominator:

$$\mathbf{N}_1 / \mathbf{N}_2 = (a_1 + \mathbf{j}b_1)(a_2 - \mathbf{j}b_2) / (a_2^2 + b_2^2)$$

$$= [a_1 a_2 + b_1 b_2 + \mathbf{j}(b_1 a_2 - a_1 b_2)] / (a_2^2 + b_2^2)$$

Now $|a+\mathbf{j}b| = \sqrt{}(a^2 + b^2)$, therefore:

$$\lvert \mathbf{N}_1/\mathbf{N}_2 \rvert = \sqrt{}\, [\{ (a_1a_2 + b_1b_2)^2 + (b_1a_2 - a_1b_2)^2 \} / (a_2^2 + b_2^2)^2]$$

$$= \sqrt{}[\{(a_1a_2)^2 +(b_1b_2)^2 +2a_1a_2b_1b_2 +(b_1a_2)^2 +(b_2a_1)^2 -2a_1a_2b_1b_2\}/(a_2^2+b_2^2)^2]$$

$$= \sqrt{}[\{ (a_1a_2)^2 +(b_1b_2)^2 +(b_1a_2)^2 +(b_2a_1)^2 \}/(a_2^2+b_2^2)^2]$$

$$= \sqrt{}[\{ a_2^2(a_1^2 + b_1^2) + b_2^2(a_1^2 + b_1^2) \}/ (a_2^2 + b_2^2)^2]$$

$$= \sqrt{}[(a_1^2 + b_1^2) (a_2^2 + b_2^2) / (a_2^2 + b_2^2)^2]$$

$$= \sqrt{}[(a_1^2 + b_1^2) / (a_2^2 + b_2^2)]$$

$$= [\sqrt{}(a_1^2 + b_1^2)] / [\sqrt{}(a_2^2 + b_2^2)]$$

$$= \lvert \mathbf{N}_1 \rvert / \lvert \mathbf{N}_2 \rvert$$

24.2 Magnitude reciprocal theorem:
The magnitude of the reciprocal of a (complex) number is equal to the reciprocal of its magnitude.

$\lvert 1/\mathbf{N}\rvert = 1 / \lvert\mathbf{N}\rvert$	**24.2**

Proof:
Let $\mathbf{N}_1 = 1+\mathbf{j}0 = 1$

Now

$\lvert\mathbf{N}_1/\mathbf{N}_2\rvert = \lvert\mathbf{N}_1\rvert / \lvert\mathbf{N}_2\rvert$

Therefore:

$\lvert 1/\mathbf{N}_2\rvert = \lvert 1\rvert / \lvert\mathbf{N}_2\rvert$

$\lvert 1/\mathbf{N}_2\rvert = 1 / \lvert\mathbf{N}_2\rvert$

24.3 Magnitude product theorem:
The magnitude of the product of two (complex) numbers is equal to the product of their magnitudes.

$\lvert\mathbf{N}_1\ \mathbf{N}_2\rvert = \lvert\mathbf{N}_1\rvert\ \lvert\mathbf{N}_2\rvert$	**24.3**

Proof:
Let $\mathbf{N}_1 = a_1 + \mathbf{j}b_1$ and $\mathbf{N}_2 = a_2 + \mathbf{j}b_2$

Then:

$\mathbf{N}_1\ \mathbf{N}_2 = (\ a_1 + \mathbf{j}b_1\ )(\ a_2 + \mathbf{j}b_2\ )$

$\quad = a_1a_2 - b_1b_2 + \mathbf{j}(a_1b_2 + a_2b_1)$

$\lvert\mathbf{N}_1\ \mathbf{N}_2\rvert = \sqrt{}[(a_1a_2 - b_1b_2)^2 + (a_1b_2 + a_2b_1)^2]$

$\quad = \sqrt{}[(a_1a_2)^2 + (b_1b_2)^2 - 2a_1a_2b_1b_2 + (a_1b_2)^2 + (a_2b_1)^2 + 2a_1a_2b_1b_2\ ]$

$\quad = \sqrt{}[a_1^2(a_2^2 + b_2^2) + b_1^2(a_2^2 + b_2^2)]$

$\quad = \sqrt{}[(a_1^2 + b_1^2)(a_2^2 + b_2^2)]$

$\quad = [\sqrt{}(a_1^2 + b_1^2)][\sqrt{}(a_2^2 + b_2^2)]$

$\quad = \lvert\mathbf{N}_1\rvert\ \lvert\mathbf{N}_2\rvert$

24.4 Scaling theorem:
The magnitude of the product of a scalar and a complex number is equal to the product of the scalar and the magnitude.

$\lvert s\mathbf{N}\rvert = s\lvert\mathbf{N}\rvert$	**24.4**

Proof:
Let s be a scalar, and $\mathbf{N} = a + \mathbf{j}b$

$s\mathbf{N} = sa + \mathbf{j}sb$

$\lvert s\mathbf{N}\rvert = \sqrt{[(sa)^2 + (sb)^2]}$

$= \sqrt{[s^2(a^2 + b^2)]}$

$= s\sqrt{(a^2 + b^2)}$

$= s\lvert\mathbf{N}\rvert$

i.e., a scalar can be factored out of or multiplied into a magnitude bracket in the same way that it can be done with any other type of bracket.

24.5 Drop-dimension theorem:
A phasor with a phase angle of 0° or 180° transforms as a scalar:

$\mathbf{N}(\,\lvert\mathbf{N}\rvert\,,\,0°) = +\lvert\mathbf{N}\rvert$ $\mathbf{N}(\,\lvert\mathbf{N}\rvert\,,\,180°) = -\lvert\mathbf{N}\rvert$	**24.5**

Proof:
A phasor pointing at 0° can be represented as a complex number with a *positive* real part and a zero imaginary part. A phasor pointing at 180° can be represented as a complex number with a *negative* real part and a zero imaginary part. Hence if:

$\mathbf{N} = a + \mathbf{j}0$

then

$\mathbf{N} = a$

and

$\lvert\mathbf{N}\rvert = +\sqrt{(a^2 + 0^2)} = \lvert a\rvert$

Hence:

$\mathbf{N} = N = \pm\lvert\mathbf{N}\rvert$

where N is a pseudoscalar equal in value and sign to the real part of **N**.

This may appear trivial, but it shows that our assumption that a phasor that has dropped a dimension can be treated as a scalar is universal, rather than a special interpretation of a particular phasor expression. A further implication however is that N is not *identical* to the magnitude of **N**, because magnitudes are always positive whereas N can be positive or negative. We can force N to become equal to $|\mathbf{N}|$ by stipulating that $\varphi = 0°$. We can also drop a dimension, i.e., set the imaginary part to zero, by choosing $\varphi = 180°$, but in that case we get $N = -|\mathbf{N}|$.

Thus the alleged scalar that results from dropping a dimension is not a magnitude, but it is a quantity that is equal in magnitude to a magnitude, and if $\varphi = 0°$ it is positive. This may seem a pedantic distinction, but the point in making it is that if we restrict the scope of our phasor algebra through erroneous interpretation, we lose the ability to include DC electricity in our theory, and we lose the ability to explore exotic ideas such as negative resistance. The pseudoscalar we obtain by dropping a dimension *can be negative*, even if usually it isn't.

24.6 Square magnitude theorem:

The product of a complex number and its complex conjugate is the square of the complex number's magnitude.

$$\mathbf{N}\,\mathbf{N}^* = |\mathbf{N}|^2 \tag{24.6}$$

Proof:

Let $\mathbf{N} = a + \mathbf{j}b$ and $\mathbf{N}^* = a - \mathbf{j}b$

$$\mathbf{N}\,\mathbf{N}^* = a^2 + b^2$$

but

$$|\mathbf{N}| = \sqrt{}(a^2 + b^2)$$

therefore

$$\mathbf{N}\,\mathbf{N}^* = |\mathbf{N}|^2$$

Hence the product of a complex number and its complex conjugate is a true scalar. It is also literally a *scalar product*. Recall that the definition of a scalar product is:

$$\mathbf{a}\bullet\mathbf{b} = |\mathbf{a}|\,|\mathbf{b}|\,\mathrm{Cos}\varphi$$

but if **a** and **b** are identical, then $\varphi=0°$ and $\mathrm{Cos}\varphi=1$. Hence:

$$\mathbf{N}\bullet\mathbf{N} = |\mathbf{N}|^2 = \mathbf{N}\,\mathbf{N}^*$$

24.7 Conjugate product theorem:
The complex conjugate of the product of two complex numbers is the product of the complex conjugates.

$(\mathbf{N}_1\ \mathbf{N}_2)^* = \mathbf{N}_1{}^*\ \mathbf{N}_2{}^*$	**24.7**

Proof:
Let $\mathbf{N}_1 = a_1 + \mathbf{j}b_1$ and $\mathbf{N}_2 = a_2 + \mathbf{j}b_2$

Then:

$$\mathbf{N}_1\ \mathbf{N}_2 = (a_1 + \mathbf{j}b_1)(a_2 + \mathbf{j}b_2)$$

$$= a_1a_2 - b_1b_2 + \mathbf{j}(a_1b_2 + a_2b_1)$$

Therefore:

$$(\mathbf{N}_1\ \mathbf{N}_2)^* = a_1a_2 - b_1b_2 - \mathbf{j}(a_1b_2 + a_2b_1)$$

$$= a_1(a_2 - \mathbf{j}b_2) - \mathbf{j}b_1(a_2 - \mathbf{j}b_2)$$

$$= (a_1 - \mathbf{j}b_1)(a_2 - \mathbf{j}b_2)$$

$$= \mathbf{N}_1{}^*\ \mathbf{N}_2{}^*$$

24.8 In-phase quotient theorem:
If two phasors are in phase, their ratio can be treated as a scalar.

$\mathbf{N}_1(\lvert\mathbf{N}_1\rvert, \varphi) / \mathbf{N}_2(\lvert\mathbf{N}_2\rvert, \varphi) = \lvert\mathbf{N}_1\rvert / \lvert\mathbf{N}_2\rvert$	**24.8**

proof:
Using the polar to complex transformation (**12.4**):

$$\mathbf{N}_1(\lvert\mathbf{N}_1\rvert, \varphi) = \lvert\mathbf{N}_1\rvert(\mathrm{Cos}\varphi + \mathbf{j}\mathrm{Sin}\varphi)$$

$$\mathbf{N}_2(\lvert\mathbf{N}_2\rvert, \varphi) = \lvert\mathbf{N}_2\rvert(\mathrm{Cos}\varphi + \mathbf{j}\mathrm{Sin}\varphi)$$

where the phase angle φ is the same in both cases. Therefore:

$$\mathbf{N}_1 / \mathbf{N}_2 = \lvert\mathbf{N}_1\rvert(\mathrm{Cos}\varphi + \mathbf{j}\mathrm{Sin}\varphi) / [\ \lvert\mathbf{N}_2\rvert(\mathrm{Cos}\varphi + \mathbf{j}\mathrm{Sin}\varphi)\ ]$$

$$= \lvert\mathbf{N}_1\rvert / \lvert\mathbf{N}_2\rvert$$

24.9 Magnitude caveat:

The mathematical operation of 'taking a magnitude' destroys information. Specifically, it is important to be aware that if $|\mathbf{a}| = |\mathbf{b}|$, then it is *not necessarily true* that $\mathbf{a} = \mathbf{b}$. The magnitude operation discards the directional information of a vector, and the sign information of a scalar. Note for example that although:

$$|\mathbf{a}| = |\mathbf{a}^*| = |-\mathbf{a}| = |-\mathbf{a}^*|$$

any one of the quantities inside magnitude brackets is definitely not identical to any of the others. The magnitude retains only the length of the object. In so doing however, it does retain the unit of measurement; i.e., a magnitude is a length in impedance space, or voltage space, or current space, etc., and so has the units of the space in which it exists.

24.10 Magnitude equivalence:

As noted above, for any complex number $\mathbf{Z}$:

$$|\mathbf{Z}| = |\mathbf{Z}^*| = |-\mathbf{Z}| = |-\mathbf{Z}^*|$$

When designing electrical circuits, it is not unusual to meet situations in which the magnitude of a voltage or current needs to be determined, but the phase is unimportant. As the theorems above show, when only the magnitude is needed, all of the impedances involved in the calculation can be replaced by their magnitudes (provided that the impedances are factors, i.e., multipliers or divisors, not terms in a summation). What is less obvious however, is that an impedance $\mathbf{Z}$ enclosed between magnitude brackets can then be replaced by one of the alternatives having the same magnitude, namely $\mathbf{Z}^*$, $-\mathbf{Z}$ and $-\mathbf{Z}^*$. This principle of *Magnitude Equivalence* allows us to deduce alternative networks that will produce the same outcome. In particular, it allows us to identify situations in which inductance can be replaced by capacitance and vice versa.

24.11 About these theorems:

The theorems given above do not appear in standard engineering textbooks. Therefore it is legitimate to ask: 'Why have they been stated here when everyone else manages without them?' The answer to the question is this: By sticking to the mathematical rules: particularly by ensuring that we always use properly balanced vector equations, and by using any simplifications that can be proved in a general way; *we eliminate the need for phasor diagrams*. Essentially, we can let the algebra do all of the reasoning. We can still use phasor diagrams for the purpose of explaining what is going on, but they become merely illustrative and make no difference whatsoever to the outcome of a problem-solving exercise. The traditional role of the phasor diagram has been to help in resolving the ambiguities caused by unrigorous mathematical definitions. But the mathematics is self-consistent. If the problem is defined correctly, the hand-waving becomes unnecessary.

Also, it should be said, that the simple rules of substitution given above are the sort of thing that engineers work out for themselves after years of experience. Thus, by stating them initially, we stand to make the learning process a little less arduous.

25. Generalisation of Ohm's law

We have already arrived at a general statement of Ohm's law in section **7** by observing that it can be written as a phasor equation: **V**= **I Z** . We have also observed that we are at liberty to treat either **I** or **V** as a scalar equal in value to its own magnitude in order to learn the phase of the other relative to it. The drop-dimension theorem (**24.5**) gives our justification for doing so, but also allows us to generalise our phasor Ohm's law to include DC. We do this by noting that if a circuit has any series capacitive reactance, then as the frequency goes to zero, X_C goes to infinity; hence the magnitude of the impedance goes to infinity and the impedance becomes an open circuit. The only type of reactance which gives DC continuity is, of course, inductive reactance, and at f=0, X_L=0. Thus **Z**(R,X) drops a dimension and becomes **Z**(R,0)=R. Hence we can write **V**=**I**R for DC (or for pure resistance and AC), but since both **V** and **I** are then in phase, they can drop dimensions also. Thus we obtain V=IR if we drop dimensions at φ=0°; but more to the point, we are also at liberty to drop dimensions at φ=180° and obtain the perfectly valid alternative:

(-V) = (-I) R

i.e., we have a theory which covers all aspects of AC electricity and also allows us to have the negative voltages and currents required for the analysis of DC circuits. This is why we must insist that the un-bold symbols V and I are not magnitudes, they are pseudoscalars (or, if you prefer, complex numbers in the form a+**j**0) which can point in either a positive or a negative direction. It is only resistance which can never be negative in a passive network, and that is for physical rather than for mathematical reasons.

An additional interpretation of Ohm's law is also given to us by the magnitude ratio and product theorems (**24.1-24.3**). These allow that if **V**=**IZ**, then:

|**V**| = |**I Z**| = |**I**| |**Z**|

(and all possible rearrangements). This says that the need for complex arithmetic is removed if all you want to know is a magnitude, i.e., if the left hand side of an equation is a magnitude, then all of the phasors on the right can be replaced by their magnitudes. This observation simplifies some problems enormously, since failure to apply the magnitude ratio and product theorems when the situation allows results in unwitting repetition of the working used in the proofs in sections **24.1** to **24.3**.

Shown below are some of the possible interpretations of Ohm's law that stem from the discussion in this chapter. There is no need to memorise these formulae because they are all derived from the statement **V** = **I Z** . What they show is that all manner of complicated arguments involving phasor diagrams are in fact trivial and can be deduced by inspection of the master equation.

$\mathbf{V} = \mathbf{I}\,\mathbf{Z}$	$\mathbf{Z} = \mathbf{V} / \mathbf{I} = \mathbf{V}\,\mathbf{I}^* / \lvert\mathbf{I}\rvert^2$	$\mathbf{I} = \mathbf{V} / \mathbf{Z} = \mathbf{V}\,\mathbf{Z}^* / \lvert\mathbf{Z}\rvert^2$
$\mathbf{V} = \mathbf{I}\,(R+\mathbf{j}X)$	$R + \mathbf{j}X = \mathbf{V} / \mathbf{I}$	$\mathbf{I} = \mathbf{V} / (R+\mathbf{j}X) = \mathbf{V}\,(R-\mathbf{j}X) / (R^2 + X^2)$
$V = \mathbf{I}\,\mathbf{Z}$	$\mathbf{Z} = V / \mathbf{I} = V\mathbf{I}^* / \lvert\mathbf{I}\rvert^2$	$\mathbf{I} = V / \mathbf{Z} = V\,\mathbf{Z}^* / \lvert\mathbf{Z}\rvert^2$
$\mathbf{V} = I\,\mathbf{Z}$	$\mathbf{Z} = \mathbf{V} / I$	$I = \mathbf{V} / \mathbf{Z} = \mathbf{V}\,\mathbf{Z}^* / \lvert\mathbf{Z}\rvert^2$
$\lvert\mathbf{V}\rvert = \lvert\mathbf{I}\rvert\,\lvert\mathbf{Z}\rvert$	$\lvert\mathbf{Z}\rvert = \lvert\mathbf{V}\rvert / \lvert\mathbf{I}\rvert$	$\lvert\mathbf{I}\rvert = \lvert\mathbf{V}\rvert / \lvert\mathbf{Z}\rvert$
$\lvert\mathbf{V}\rvert = I\,\lvert\mathbf{Z}\rvert$	$\lvert\mathbf{Z}\rvert = \lvert\mathbf{V}\rvert / I$	$I = \lvert\mathbf{V}\rvert / \lvert\mathbf{Z}\rvert$
$V = \lvert\mathbf{I}\rvert\,\lvert\mathbf{Z}\rvert$	$\lvert\mathbf{Z}\rvert = V / \lvert\mathbf{I}\rvert$	$\lvert\mathbf{I}\rvert = V / \lvert\mathbf{Z}\rvert$
$V = I\,R$	$R = V / I$	$I = V / R$
$(-V) = (-I)\,R$	$R = (-V) / (-I)$	$(-I) = (-V) / R$

Where:
$\mathbf{V}$ $\mathbf{I}$ and $\mathbf{Z}$ are phasors,
$\mathbf{V}^*$ $\mathbf{I}^*$ and $\mathbf{Z}^*$ are complex conjugates,
$|\mathbf{V}|$ $|\mathbf{I}|$ and $|\mathbf{Z}|$ are magnitudes,
V and I are phasors pointing at 0°,
$(-V)$ and $(-I)$ are negative values of V and I (and thus are phasors pointing at 180°),
and un-bold Z is not normally used[16], because an impedance pointing at 0° already has the symbol R.

[16] The impedance of free space, Z_0 (which is real for a centrosymmetcic Universe), is an exception.

26. General statement of Joule's law

In section **3**, we gave Joule's law in its standard form:

$$P = I^2 R$$

This, now that we know that I should be interpreted as a phasor pointing at 0° or 180°, proves to be a correctly balanced vector equation; but it is only so by an accident of notation and, as we shall see shortly, the restriction on the phase of the current is unnecessary and limits the scope of the formula.

Joule's law, even in its standard form, is a more fundamental statement than P=IV; because the squaring of the current prevents the direction of the current from having any effect on the direction of the power. It therefore tells us that power is positive when resistance is positive, i.e., the dissipation of energy is a uni-directional process. The direction in question is that of entropy, the general spreading out and cooling down of the Universe, which is associated with the irreversibility of time. Thus, assuming that an impedance is by definition a strictly passive network; our relationship with impedance space is skewed, in that we are not allowed to venture into the regions where resistance is negative. One far-reaching consequence is that we cannot devise electrical networks that will give an output before receiving an input, i.e., we cannot build circuits that violate causality.

There are however non-linear passive electronic devices that have a *negative resistance characteristic* (such as the Esaki diode or tunnel diode[17] [18]), but this is only in the sense that there is a region in the graph of I vs. V where the current goes down as the voltage is increased. Thus negative resistance devices can go from a particular level of power dissipation to a lower level as the applied voltage is increased, but they can never achieve a state of negative power dissipation.

In order to generalise Joule's law completely, we must write it in a way that allows the current phasor **I** to adopt an arbitrary phase but which gives an explicitly scalar result. The obvious candidate expression is:

$$\boxed{P = |\mathbf{I}|^2 R \qquad \mathbf{26.1}}$$

It transpires that this *is* the definitive statement of Joule's law, as we will demonstrate; but first it is interesting to note a parallel between it and energy laws in the wider context. The square magnitude theorem (**24.6**) tells us that $|\mathbf{I}|^2 = \mathbf{I}\,\mathbf{I}^*$, and so we can write expression (**26.1**) as

$$P = \mathbf{I}\,\mathbf{I}^*\, R$$

or more to the point:

$$\boxed{P = \mathbf{I}^*\, R\, \mathbf{I} \qquad \mathbf{26.2}}$$

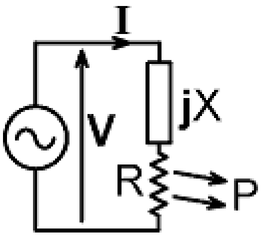
Mathematical structures of this type occur everywhere in physics. Quantum-mechanical energy equations, for example, are of the same form; and in that context the vector that occupies the position if **I** is known as the *eigenvector* or 'state-vector' (and also the wavefunction). If **I** describes the state of the system, the implication is that all we have to do is define **I** in order to determine the energy. If the analogy holds, then the expression above is universally true upon provision of a

[17] **The Art of Electronics**, Paul Horowitz [W1HFA] and Winfield Hill, 2nd edition 1989, Cambridge University Press. ISBN 0-521-37095-7. Tunnel (Esaki) diode p14-15, & p1060. Back diode p891, 893.

[18] **Physical Electronics**, C L Hemenway, R W Henry, M Caulton, Wiley & Sons, New York, 2nd edn. 1967. Library of Congress cat. card no. 67-23327. Section 14.6: The tunnel diode, p290-294.

definition for **I**. This is not difficult to demonstrate, because all electrical power transmission problems boil down to the matter of delivering power to an impedance.

Consider the system shown on the right. In this case, **V** and **I** are not necessarily in phase, but we can easily obtain an expression for **I** in terms of **V**, and since we now appear to have a version of Joule's law that allows **I** to point in any direction, there is no need to impose a restriction on any of the phasors involved. Thus we can write:



$\mathbf{I} = \mathbf{V} / (R+\mathbf{j}X)$

Now, putting the reciprocal impedance into the a+**j**b form by multiplying numerator and denominator by the complex conjugate of the denominator we obtain:

$\mathbf{I} = \mathbf{V}\ (R - \mathbf{j}X) / (R^2 + X^2)$

and, using the conjugate product theorem (**24.7**):

$\mathbf{I}^* = \mathbf{V}^*\ (R + \mathbf{j}X) / (R^2 + X^2)$

Now we can insert these definitions into equation (**26.2**): P=**I***R**I**, thus:

$P = \mathbf{V}\ \mathbf{V}^*\ R\ (R - \mathbf{j}X)(R + \mathbf{j}X)/[(R^2 + X^2)^2]$

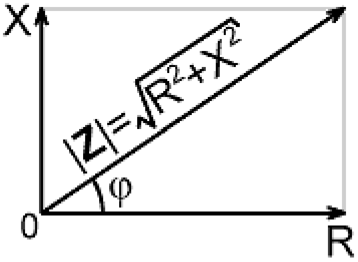
and using the square magnitude theorem (**24.6**) we obtain:

$P = \lvert\mathbf{V}\rvert^2 R / (R^2 + X^2)$	**26.3**

Thus without any convoluted discussion about phases or reference phasors, we have obtained in a few lines of algebra a general expression for the power dissipated in an impedance in terms of the applied voltage. Now let us check that this is consistent with the conventional approach:
Here we use the power factor (**V I** scalar product) rule (**10.1**).

$P = \mathbf{V}\bullet\mathbf{I} = \lvert\mathbf{V}\rvert\ \lvert\mathbf{I}\rvert \cos\varphi$

Now, using the diagram on the right, we can see that Cosφ (adjacent / hypotenuse) is $R/\sqrt{(R^2 + X^2)}$. Hence:



$P = \lvert\mathbf{V}\rvert\ \lvert\mathbf{I}\rvert\ R / \sqrt{(R^2 + X^2)}$ . . . **26.4**

From Ohm's law we know that:

$\mathbf{I} = \mathbf{V} / (R+\mathbf{j}X)$

and using the magnitude ratio theorem (**24.1**) we obtain:

$\lvert\mathbf{I}\rvert = \lvert\mathbf{V}\rvert / \lvert(R+\mathbf{j}X)\rvert$

i.e.:

$|\mathbf{I}| = |\mathbf{V}| / \sqrt{(R^2 + X^2)}$

Now, substituting this into expression (**26.4**) we have:

$P = |\mathbf{V}|^2 R / (R^2 + X^2)$

Which is the same as equation (**26.3**) and so demonstrates that the power-factor rule is already embedded in Joule's law when we write the latter as a properly balanced and *un-restricted* vector equation. Notice also, that however we manipulate the power law, the average direction of power flow is always dictated by the sign of the resistance.

We can, incidentally, also obtain equation (**26.3**) by using the series to parallel transformation discussed in section **19b**. If we write the impedance in parallel form, then the power is simply given by the square of the voltage magnitude divided by the equivalent parallel resistance. Thus, if:

$\mathbf{Z} = R + \mathbf{j}X = R'' // \mathbf{j}X''$

Then:

$P = |\mathbf{V}|^2 / R''$

where:

$R'' = (R^2 + X^2) / R$

i.e.:

$P = |\mathbf{V}|^2 R / (R^2 + X^2)$

The universal steady-state electrical power laws can be summarised as follows:

$P = \lvert\mathbf{I}\rvert^2 R$	$P = \lvert\mathbf{V}\rvert^2 R / (R^2 + X^2)$	$P = \mathbf{V} \bullet \mathbf{I} = \lvert\mathbf{V}\rvert \, \lvert\mathbf{I}\rvert \operatorname{Cos}\varphi$

where $|\mathbf{I}|$ is the reading obtained from an ammeter, and $|\mathbf{V}|$ is the reading obtained from a voltmeter. If the impedance has no reactive component, or if the frequency approaches 0Hz (DC), the general formulae above revert to their standard textbook forms:

$P = I^2 R$	$P = V^2 / R$	$P = V I$

where V and I are the readings from AC or DC instruments and can be positive or negative; but if V is negative then I must be negative (presuming that the resistance to which the power is being delivered is positive).

27. Bandwidth

We are often interested the way in which the gain or loss of a network or circuit varies over a particular band of frequencies. We will introduce this type of analysis shortly in connection with resonant networks, but before doing so it is necessary to define the term 'bandwidth'.

Most readers will be aware that an amplifier will generally show a fall-off of gain at low frequencies, this often being due to the increasing magnitudes of the reactances of coupling capacitors in series with the signal path; and it will also show a fall-off of gain at high-frequencies, this being due to a variety of factors including the falling magnitudes of the reactances of any stray capacitances in parallel with the signal path. Consequently amplifiers, and indeed many other types of circuit, usually show a hump-like frequency response; and will only pass signals usefully over a particular frequency range. The problem in defining bandwidth therefore lies in the definition of what we mean by 'useful' and, since this will vary according to the particular application, there is really no resolution to the issue. We must therefore eschew vague concepts like 'usefulness' in favour of a definition on which everyone can agree; and so it is universally accepted that bandwidth, unless stated otherwise, is defined in terms of what are known as the 'half-power points', i.e., the upper and lower frequency points at which the power delivered by the system (for a constant input) has fallen to half of that which is delivered at the frequency at which the maximum response occurs. The half-power points are chosen, as we shall see, because they have special mathematical significance; and for simple networks at least, knowledge of where they lie provides a complete definition of the frequency-response function of the system.

Now, if we call the power delivered at the frequency of maximum response P_{max}, then the power delivered at the half-power points is:

$$P = P_{max} / 2$$

and

$$P / P_{max} = ½$$

We can express this ratio in decibels using the general definition:

$$\text{Ratio in dB} = 10\text{Log}_{10}(P / P_{ref})$$

(where P_{ref} is the reference power level against which power P is being compared).
i.e.

$$10 \text{ Log}_{10}(½) = -3.010299957$$

Hence the half-power points are also known as the ' -3 dB ' points, and the frequency interval between the lower point and the upper point is also often called the ' -3 dB bandwidth '. It is a good idea to be specific in this way, because when the term 'bandwidth' is used without qualification, there is always the fear that it may involve some non-standard definition.

28. decibels & logarithms

Having just found cause to mention the dB notation, we should perhaps clarify some of the aspects of its use that can be a source of confusion. Firstly, a decibel is a tenth of a Bel, hence the small d and the capital B in 'dB'. The Bel, named after Alexander Graham Bell, is an internationally accepted unit equal to 10 "transmission units" (TU); the TU (now the dB) being a logarithmic relative signal measurement method introduced by AT&T in 1923 [19]. A ratio in Bels is the logarithm of a power ratio defined in the simplest possible way using base 10 (i.e. common) logarithms, hence:

$N/B = \mathrm{Log}_{10}(P / P_{\mathbf{ref}})$

Thus a ratio in decibels is:

$N/dB = 10\ \mathrm{Log}_{10}(P / P_{\mathbf{ref}})$

The practice of expressing audio power ratios on a logarithmic scale was developed because human hearing has an automatic gain-control system. This makes our hearing response logarithmic with respect to loudness, and thus enables us to extract information from sound over a vast range of sound pressures (it also makes our hearing asymmetric: if a loudspeaker produces a loud sound that is asymmetric about the ambient air-pressure axis, there will be a change in perceived quality if the amplifier connections are reversed). The use of logarithmic scales is also favoured because it allows us to represent the vast gains of electronic amplifiers, and the vast ranges of power encountered in communication systems, using convenient numbers. It carries over seamlessly into the field of radio, firstly because modulated radio signals are converted into sound, and secondly, radio receivers also use automatic gain control systems and so also have a roughly logarithmic response. Common logarithms are used because the notation hails from the time when log tables were used for multiplication. The decibel is preferred over the Bel because it transpires that 1 dB (25.9%) is close to the minimum change in audio power that can be detected by the human ear (this being 10 - 20%, i.e., 0.4 - 0.8 dB, depending on the waveform and the listener). Hence there is rarely a need to express power ratios in dB to a greater accuracy than to the nearest whole number.

There is also a log-ratio notation based on natural (Naperian) logarithms, with an alternative unit, the Neper (symbol Np), i.e.:

$N/Np = (½)\mathrm{Log}_e\ (P / P_{\mathbf{ref}})$

giving 1 Np = 8.686 dB

The Neper is preferred for some specialist applications, but is not normally used in a general context.

The use of log-tables for multiplication has of course died-out from the school curriculum, and this leaves modern students unprepared for the introduction of decibels and other logarithmic functions. A little revision of the subject will therefore not go amiss: The technique of multiplication using logarithms was introduced in 1614 by the Scottish mathematician and astronomer Jhone Neper (spelt variously but nowadays usually written as 'John Napier', this being a version that Neper himself would not have recognised). It arises from the observation that if two numbers are each

[19] **Admiralty Handbook of Wireless Telegraphy**, B.R.230 (Vol. II). HMSO London, **1938**., Appendix A: The Decibel and the Neper (Appendix available from http://www.g3ynh.info)

expressed as a base-number raised to a power, then the numbers can be multiplied simply by adding the powers, i.e.,

$B^{a} \times B^{b} = B^{a+b}$

This is obvious when the powers (also known as exponents) are whole numbers (e.g. $10^{3} \times 10^{5} = 10^{8}$), but it works just the same when they are not. Also, we can perform division just as easily by noting that:

$B^{a} \div B^{b} = B^{a-b}$

Consequently, in the 17th and 18th Centuries, long before the advent of affordable calculating machines, great effort was made to produce tables allowing difficult multiplications and divisions to be performed by looking up the exponent that, when used to raise a common base, represents a particular number. These exponents are known as *logarithms*, and are defined as follows:

If $n = B^{a}$ then $a = \mathrm{Log}_{B}(n)$

("if n equals B to the power of a, then a is the log to the base B of n")

If a logarithm is written without the base subscript, then base 10 is usually implied, i.e., 'Log' means 'Log_{10}' (although some older documents deviate from this convention). Naperian (natural) logarithms, which crop-up frequently in physics, use Euler's number e as the base (e≈2.71828), and can be written either "Log_{e}" or "ln", the latter pronounced "line" and being short for 'log-Naperian'.

Hence, working in base 10:

if $m=10^{a}$ and $n=10^{b}$ then $m\times n=10^{a+b}$

All we have to do to perform the multiplication is look-up the logarithms a and b, add them together, then look up the quantity a+b in a table of anti-logarithms to find the required m×n. The anti-log of a number x is simply 10^{x}. As an alternative to using tables, the same operations can be achieved by using two identical engraved logarithmic scales and sliding one relative to the other, the device for so doing (invented by William Oughtred in 1622) being known as a *slide rule*[20].

We no longer need to use logarithms for everyday multiplication, but we do need to memorise some of their basic properties in order to use logarithmic units with confidence. The first and most fundamental point is that any number raised to the power of zero is one, i.e.,

$B^{0} = 1$, always, regardless of B.

Hence:

$\mathrm{Log}(1) = 0$, always, regardless of base.

Now, if we represent power gain in decibels, i.e.,

$N/\mathrm{dB} = 10\ \mathrm{Log}_{10}(P / P_{\mathbf{ref}})$

then if $P = P_{\mathbf{ref}}$;

[20] **"When Slide Rules Ruled"** Cliff Stoll, Scientific American, May 2006, p68-75.

$N/dB = 10\ Log_{10}(1) = 0$

i.e., a system that neither amplifies nor attenuates a signal has a gain of 0dB.

Also we find that if P is greater than $P_{\mathbf{ref}}$, then N/dB is greater than zero, and vice versa. Hence a positive quantity in dB represents gain, and a negative quantity represents loss (i.e., negative gain). Finally, the additive property of logarithms allows that if we subject a signal to a number of processes, and note the gains in dB (positive or negative) for each of those processes, we can find the overall gain of the system simply by adding all of the individual stage gains together.

So much for the basics, but now we arrive at the point that causes greatest difficulty: A quantity in dB implies a logarithmic *power ratio*. It can however also be taken to represent a voltage ratio, or a current ratio, but the definition *must be modified* in that case. The reason why the definition can be extended to embrace current and voltage ratios is that power is a function of the voltage across, and also the current through, an impedance. Hence we can substitute for power using the general power laws derived earlier, i.e., $P = |\mathbf{V}|^2\ R/(R^2+X^2)$ and $P = |\mathbf{I}|^2\ R$. To obtain the voltage ratio formula we write:

$N/dB = 10\ Log_{10}\{\ [\ |\mathbf{V}|^2\ R/(R^2+X^2)\ ]\ /\ [\ |\mathbf{V}_{\mathbf{ref}}|^2\ R/(R^2+X^2)\ ]\ \}$

which reduces to:

$N/dB = 10\ Log_{10}[\ (\ |\mathbf{V}|\ /\ |\mathbf{V}_{\mathbf{ref}}|\ )^2\ ]$

and if we adopt the convention that the impedance against which the two voltages are compared is a resistance:

$N/dB = 10\ Log_{10}[\ (V\ /\ V_{\mathbf{ref}})^2\ ]$

A similar argument applies for the current ratio formula:

$N/dB = 10\ Log_{10}[\ (I\ /\ I_{\mathbf{ref}})^2\ ]$

Now everything would be fine of we left the quantity inside the logarithm brackets as the square of a voltage or current ratio, but everyone who teaches the subject will insist on performing a 'simplification', which is to note that a number can be squared by doubling its logarithm. Hence we get rid of the power of 2 by writing:

$N/dB = 20\ Log_{10}(V\ /\ V_{\mathbf{ref}})$

and

$N/dB = 20Log_{10}(I\ /\ I_{\mathbf{ref}})$

Which is all very clever, but leaves people struggling to decide whether they should use 10Log() or 20Log() , and so leads to lots of mistakes. So remember: a ratio in Bels is the Log of a *power ratio*, and a decibel is a tenth of a Bel (that's where the 10 comes from). By Joule's law, the *square* of a voltage or current magnitude ratio is also analogous to a power ratio, and the squaring can be obtained by doubling the logarithm (that's where the 20 comes from).

In using decibels, the basic approach is to consider the power levels at two points in a circuit or power transmission system and thereby define the gain. It is also useful however, to express power in relation to some external reference or standard, and this leads to an extension of the notation, some commonly encountered variants being as follows:

Unit	Definition	Reference voltage = √(PR)	Reference current = √(P/R)
dBm	dB relative to 1mW in 50Ω*	223.6mV	4.472mA
dBu	dB relative to 1mW in 600Ω	774.6mV	1.291mA
dBW	dB relative to 1W	-	-
dBV	dB relative to 1V	1V	

* in old audio publications and service manuals, ' dBm ' may be used to mean ' dB relative to 1mW in 600Ω '.

By extending the definition in this way, the dB notation may be used to express an absolute power (rather than a relative power); and if a reference resistance is specified, an absolute voltage or current as well. For example, if the line output level from an audio recorder is specified as -10 dBu, then the output voltage is obtained by rearranging the expression:

$-10 = 20\mathrm{Log}(V_{out}/V_{ref})$

where $V_{ref} = \sqrt{(0.001 \times 600)} = 774.6$ mV

Hence:

$V_{out} = V_{ref} \times 10^{-½} = V_{ref} / (\sqrt{10}) = 244.9$ mV RMS.

The dBW notation was brought into European Amateur Radio documents some years ago, this being the preference in the field of broadcast and professional radio engineering. Thus a 400 W transmitter (for example) becomes a 10 Log(400) = 26 dBW transmitter; and it is possible to determine the effective radiated power (ERP) of a radio installation by adding the transmitter power in dBW to the (negative) gain in dB of the antenna feeder and the gain in dB of the antenna. This is all very well of course, but it does beg the question: 'why, for a group of spectrum users generally only equipped to measure voltage and resistance to a reasonable accuracy, is it necessary to state power restrictions in a way that requires a knowledge of exponential functions in order to work out what they mean?' It would seem equally logical to state road speed restrictions in dBmph or dBkm/h, and so for the sake of any bureaucrats who might read this, we will also address the question: 'do speed ratios require the 10 Log() or the 20Log() formula?' This question, perhaps surprisingly, is not meaningless, and can be answered by noting that power is equivalent to energy delivered per unit-of-time. A power ratio is thus an energy per [unit-of-time] compared to a reference energy per [unit-of-time], and since the ' [unit-of-time]s ' will cancel (provided that they are the same - seconds are very popular), a power ratio is also an *energy ratio*. Hence:

$N/\mathrm{dB} = 10\ \mathrm{Log}_{10}(E / E_{ref})$

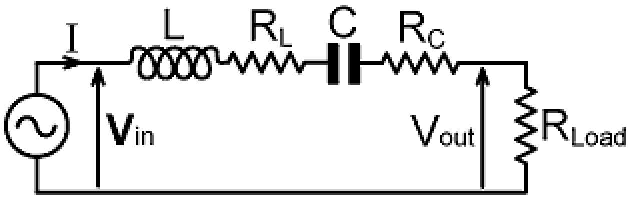
Newton's laws of motion tell us that the kinetic energy of a moving body is given by $E = mv^2/2$ (where m is the mass and v is the velocity), so energy is proportional to velocity squared as well as to voltage squared and current squared. Hence, a speed in dBmph is given by

20Log(v), so 30 mph becomes 29.5 dBmph and 70 mph becomes 36.9 dBmph.

Now, having upgraded all of our road signs to be in keeping with the preferred notation for Government standards documents, we are only left with the problem of how to measure money in decibels. Here we may note that currency names are often derived from weights (of silver, but there has been some devaluation since Roman times), and that Newton's and Einstein's laws tell us that mass is proportional to energy. Thus we can deduce that the 10 Log() formula is the correct one in this case. We might have solved this conundrum without recourse to physics however, by recalling the famous old saying: "money is power".

29. Bandwidth of a series resonator

When used to make filters for the RF and IF amplifiers of radio receivers, high-Q resonant circuits provide better selectivity than low-Q resonant circuits. Since good selectivity is synonymous with narrow bandwidth, there is evidently a relationship between bandwidth and Q, and as we shall see, this relationship is a particularly simple one if we define the bandwidth as the interval between the half-power (-3.01 dB) points. The question we must address next therefore is: "what is this power to which we must relate the bandwidth?" The answer in the case of an amplifier driving a resistive load is obvious, it is the power dissipated in the load. In the case of a resonant circuit however, there may or may not be a load in the normal sense, and we are left with the uncomfortable conclusion that bandwidth must be defined in terms of the power that would be dissipated in the load, if there were a load. We can crack this riddle for the *series* resonant case by considering the circuit shown below.



This is the circuit of a simple band-pass filter. It has an input voltage, an output voltage, and a load resistance; and the bandwidth is very clearly the frequency interval between the points where the power in the load is half of that which occurs at the frequency of maximum response (for constant $|\mathbf{V}_{in}|$). It is also however, a series resonant circuit, and the Q at resonance can be defined as:

$$Q_0 = X_{0L} / R = -X_{0C} / R$$

where R is the total resistance, i.e., $R = R_L+R_C+R_{Load}$; and the subscript ' 0 ' has been added to the reactances as a reminder that Q_0 is defined in terms of their values at the resonant frequency f_0.

Now, we can easily write an expression for the current that flows from the generator because the impedance connected across the generator is simply:

$$\mathbf{Z} = R + \mathbf{j}(X_L+X_C)$$

and so, taking the current as a reference phasor, and using Ohm's law and the magnitude ratio theorem (24.1), we obtain:

$$I = |\mathbf{I}| = |\mathbf{V}_{in}| / |\mathbf{Z}|$$

We can also state that the power delivered to the load is:

$P_{Load} = I^2 R_{Load}$

and that maximum power will occur at the resonant frequency because the total reactance will then be zero and the magnitude of the total impedance |**Z**| will be at a minimum. Hence we will call the maximum load power P_{0Load} , and the power at the bandwidth limits will be:

$P_{Load} = P_{0Load} / 2$

i.e.,

$I^2 R_{Load} = I_0^2 R_{Load} / 2$

where I_0 is the current at resonance. Hence the bandwidth limits lie at the points where
$I^2 = I_0^2 / 2$

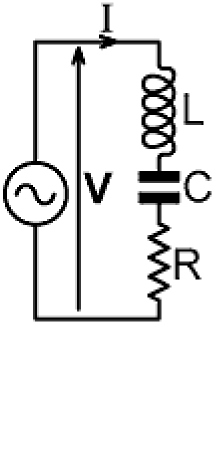
i.e., where

$I = I_0 / \surd2$

So the load resistance, having served to allow us to define the bandwidth, has promptly vanished; and the bandwidth becomes the interval between the points where the current has fallen to 1/√2 of its value at resonance. Furthermore, we can observe that we will always obtain this result regardless of which resistance we define as the load. R_L , R_C , and R_{Load} are only symbols, and since the corresponding resistances are connected in series, we can swap their designations at will. We can also consider any combination of these resistances to be the load, including the total resistance R, and this will always cancel and tell us that the half-power points occur when $I=I_0/\surd2$. Thus to define the bandwidth of a series resonant circuit, we do not need to designate any resistance as a load, we need only to consider the current.

So it transpires that we can choose any resistance in a series network and analyse the power dissipated in it to determine the bandwidth; and since we are interested here in the relationship between bandwidth and Q, the obvious resistance to choose is R, the total resistance. We can always isolate a portion of R to determine the power delivered to it or the voltage across it if we so wish, that is a trivial matter of proportions; but for a general analysis, the problem simplifies to that of understanding the behaviour of the simple series LCR network shown below.

The first part of the analysis is to determine the frequency response function for this circuit and plot it as a graph to see what it looks like. A good function to plot for this purpose is the ratio P/P_0 vs. frequency, because this ratio has a value of 1 at f_0 and is also in the correct form for conversion into decibels. The power ratio is equal to the square of the current ratio: $P/P_0 = I^2/I_0^2 = (I/I_0)^2$, because $P=I^2R$ and $P_0=I_0^2R$. Hence we will start by obtaining an expression for the current ratio.



The general expression for the current is:

$I = |\mathbf{I}| = |\mathbf{V}| / |\mathbf{Z}|$

where $\mathbf{Z}=R+\mathbf{j}(X_L+X_C)$

At the resonant frequency however, the impedance is purely resistive, so:

$I_0 = |\mathbf{V}| / R$

Hence:

$$I / I_0 = (|V| / |\mathbf{Z}|) / (|V| / R)$$

$$= R / |\mathbf{Z}|$$

$$= R / \{ \sqrt{}(R^2 + [X_L{+}X_C]^2) \}$$

which, by writing the reactances explicitly, gives:

$$I / I_0 = R / \{ \sqrt{}(R^2 + [2\pi fL - 1/(2\pi fC)]^2) \} \tag{29.1}$$

and since $P/P_0 = (I/I_0)^2$:

$$P / P_0 = R^2 / (R^2 + [2\pi fL - 1/(2\pi fC)]^2) \tag{29.2}$$

Graphs of both of these functions are shown below, the procedure used for generating them being to choose a value for f_0 and an L/C ratio, and then calculate (using the Open Office Calc spreadsheet program) a set of points at closely spaced intervals for various different values of Q_0 (see accompanying file **ser_res.ods**). The initial choices are arbitrary, since (as we are about to show) the shape of the curve obtained depends entirely on Q_0. In this case, the author chose f_0=10 MHz and $L/C = 10^4$, i.e., $C = L/10^4$. Values for L and C were then obtained by solving the resonance formula for L, i.e.,

$$10^7 = 1 / [2\pi \sqrt{}(LC)]$$

$$= 1/[2\pi \sqrt{}(L^2/10^4)]$$

$$= 100/[2\pi L]$$

$$L = 100 / [2\pi{\times}10^7]$$

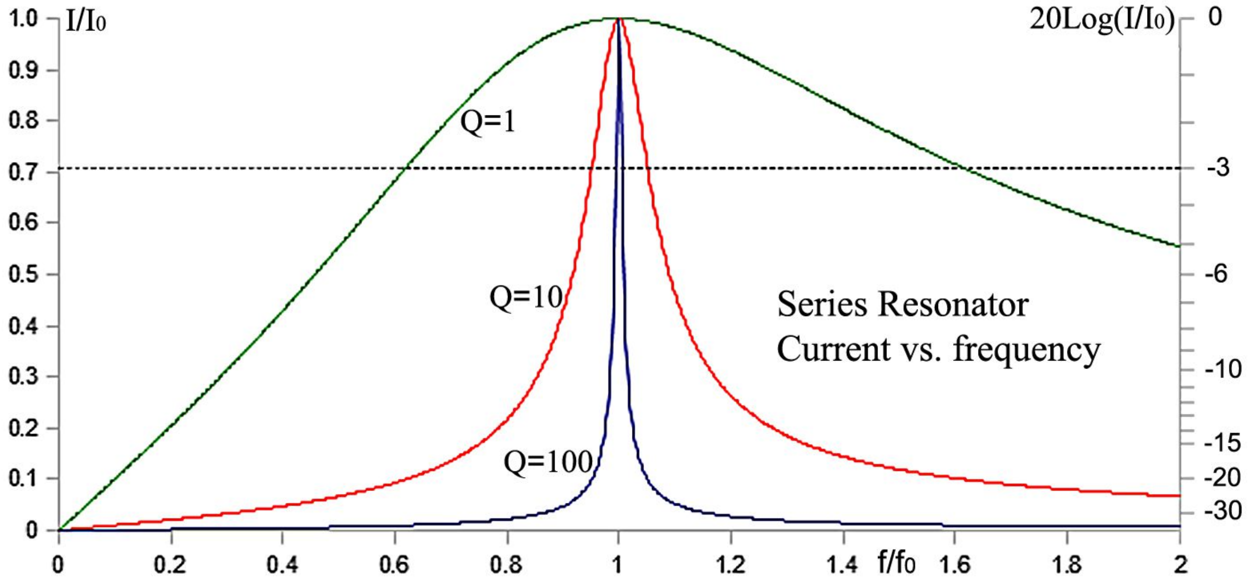
$$= 1.59154943\ \mu\text{H}$$

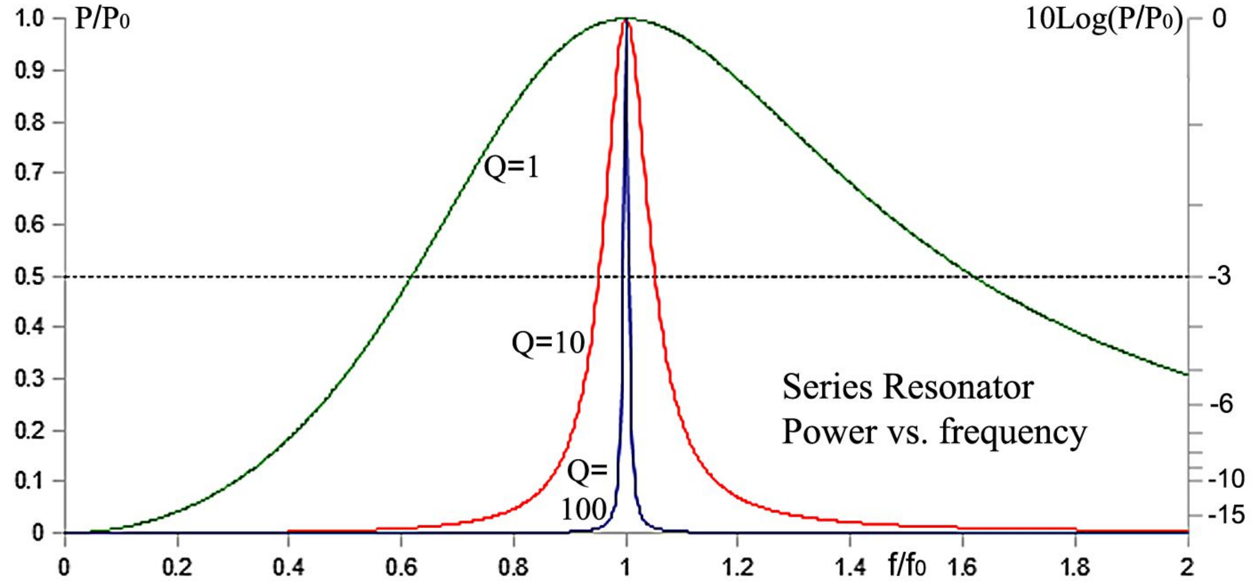
$$C = L / 10^4$$

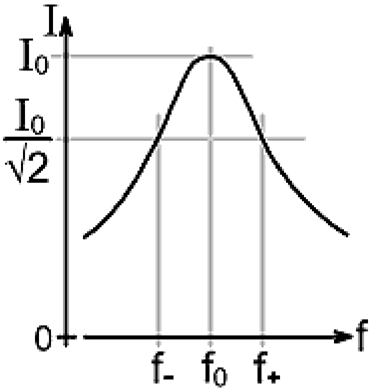
$$= 159.154943\ \text{pF}$$

Since $\sqrt{}(L/C) = 100\ \Omega$ is also the value of X_L and $-X_C$ at resonance, resonant Q values of 100, 10 and 1 correspond to total resistances (R) of 1 Ω, 10 Ω and 100 Ω respectively.

Notice in the graphs below how the squaring pushes the curve of P/P_0 downwards in comparison to I/I_0. Notice also that the half power level is $1/\sqrt{2} = 0.7071$ for I/I_0 and ½ for P/P_0, and that the decibel scales on the right differ accordingly.







So, having shown how the shape of the frequency response function varies with resonant Q, we will now derive an expression for the relationship between Q and bandwidth, the bandwidth being defined as the interval between the upper and lower half-power points. The procedure is to write a general expression for the current and solve it for the frequencies at which $I = I_0/\sqrt{2}$. Notice that the word "frequencies" is plural: the expression will be a quadratic equation.

As we have already determined, $I = |\mathbf{V}| / |\mathbf{Z}|$, and $I_0 = |\mathbf{V}| / R$. Hence, at the -3dB bandwidth limits:

$I = |\mathbf{V}| / |\mathbf{Z}| = I_0/\sqrt{2} = |\mathbf{V}| / (R\sqrt{2})$

Thus the bandwidth limits occur at the frequencies where:

$$|\mathbf{Z}| = R\sqrt{2}$$

i.e.,

$$\sqrt{(R^2 + X^2)} = R\sqrt{2}$$

$$R^2 + X^2 = 2R^2$$

$$X^2 = R^2$$

which, taking the square root of both sides and noting that there are two possibilities from so doing, gives:

$$X = \pm R$$

Now, writing X explicitly we obtain the expression:

$$\pm R = 2\pi fL - 1/(2\pi fC)$$

which we must solve for f. We may proceed by putting the right hand side onto a common denominator (i.e., by multiplying top and bottom of the $2\pi fL$ term by $2\pi fC$):

$$\pm R = [(2\pi f)^2 LC - 1] / (2\pi fC)$$

i.e.,

$$\pm 2\pi fCR = (2\pi f)^2 LC - 1$$

This rearranges to:

$$(2\pi f)^2 LC \pm 2\pi fCR - 1 = 0$$

Which is a quadratic equation in the form $af^2+bf+c=0$, with $a=4\pi^2LC$, $b=\pm 2\pi CR$ and $c=-1$. Notice however, that this particular equation will have four solutions, rather than the usual two, because the b term has a '±' symbol attached to it. The reason for that is that there are both positive and negative frequency solutions for each of the band-edges. To obtain all four of these frequencies we apply the general solution for quadratic equations (12.3):

$$f = [-b \pm\sqrt{(b^2 - 4ac)}\,] / 2a$$

$$f = \{ \pm 2\pi CR \pm\sqrt{[(2\pi CR)^2 + 4\times 4\pi^2 LC]}\, \}/(2\times 4\pi^2 LC)$$

and using the substitution $C=C^2/C$ to obtain a cancellation of C from all but one term:

$$f = \{\pm CR \pm\sqrt{[(CR)^2 + 4LC^2/C]}\, \}/(4\pi LC)$$

$$\boxed{f = [\pm R \pm\sqrt{(R^2 + 4L/C)}\,]/(4\pi L)}$$

In order to determine which are the positive frequency solutions among these four possibilities, observe that $+\sqrt{(R^2+4L/C)}$ is always larger than R. Hence the upper (positive) bandwidth limit is:

$f_+ = \{[+\sqrt{(R^2 +4L/C)}] + R\}/(4\pi L)$

and the lower (positive) bandwidth limit is:

$f_- = \{[+\sqrt{(R^2 +4L/C)}] - R\}/(4\pi L)$

and the bandwidth is:

$f_w = f_+ - f_- = \{[\sqrt{(R^2 +4L/C)}] + R - [\sqrt{(R^2 +4L/C)}] + R\}/(4\pi L)$

i.e.,

$f_w = R/(2\pi L)$ . . . . . (29.3)

Now recall that the resonant Q can be defined as $Q_0=X_L/R=2\pi f_0L/R$. Hence:

$Q_0/f_0 = 2\pi L/R$

Hence:

$f_w = f_0/Q_0$	**29.4**

This is the classic expression for the bandwidth of an LC resonator, and is exact when the resistance in the circuit remains constant with frequency. We have, of course, already observed that the loss resistances of inductors and capacitors vary with frequency, but it transpires that this will make practically no difference to the accuracy of the expression under most circumstances. Resistance makes only a small contribution to the overall shape of the bandwidth function because it only makes a significant contribution to the magnitude of the impedance (and hence to the current) when the frequency is close to resonance. Far from resonance, the impedance magnitude is dominated by the reactive component unless the Q of the resonator is very low. The physical laws governing the various processes that contribute to the loss resistance moreover are smoothly varying functions of frequency (unless some additional system resonance is encountered in the region of interest), and so the loss resistance component will not normally vary significantly over a small frequency interval. Consequently, for a reasonably high Q, the relationship: " Bandwidth = f_0/Q " is sufficiently accurate to be presumed exact for all normal engineering purposes.

30. Logarithmic frequency

One additional matter that might be of interest is that, although some authors refer to f_0 as the "centre frequency", the frequency interval between the half-power points is *not* symmetrical about f_0. We can find the mid-point or *median* frequency by taking the average of the upper and lower band limits, i.e.,

$$f_m = (f_+ + f_-)/2 = [\sqrt{(R^2 + 4L/C)} + R + \sqrt{(R^2 + 4L/C)} - R]/(2\times 4\pi L)$$

$$= [2\sqrt{(R^2 + 4L/C)}]\,/(2\times 4\pi L)$$

$$f_m = [\sqrt{(R^2 + 4L/C)}]\,/(4\pi L)\ .\ .\ .\ .\ .\ .\ (30.1)$$

This quantity is only equal to f_0 when $R\rightarrow 0$, i.e., (noting that $(\sqrt{L})/L = 1/\sqrt{L}$) :

$$f_m \rightarrow [\sqrt{(4L/C)}]\,/(4\pi L) = 1/(2\pi\sqrt{LC}) = f_0$$

This limiting condition arises because the bandwidth of the resonant circuit is infinitely narrow when $R\rightarrow 0$, i.e., the upper and lower band-limits become coincident with f_0 in the limit that $R\rightarrow 0$. It is therefore an essential property of a correct expression for the mid-band frequency, but it does give us a simplification for equation (**30.1**). Squaring (**30.1**) gives:

$$f_m^2 = (R^2 + 4L/C)]/(4\pi L)^2$$

$$= (R/4\pi L)^2 + f_0^2$$

but from expressions (**29.3**) and (**29.4**) given earlier, $f_0/Q = R/(2\pi L)$, hence:

$$f_m^2 = (f_0/2Q)^2 + f_0^2$$

$$= f_0^2(1 + [1/2Q]^2)$$

$f_m = \pm\, f_0\, \sqrt{(1 + [1/2Q]^2)}$	**30.2**

The (positive) mid point frequency is always very slightly above the resonant frequency for a practical resonator, but becomes coincident with f_0 when $Q\rightarrow\infty$ (i.e., $R\rightarrow 0$). This skewing of the bandwidth function can be said to arise because f_0 is always closer to zero than it is to infinity (the function spreads out on the high-frequency side because there is more room). The difference between the mid-band frequency and the resonant frequency is however, very small, being 0.125% (12.5 kHz at 10 MHz) for a Q of 10, and 0.00125% (125 Hz at 10 MHz) for a Q of 100. Hence the bandwidth function can be considered to be approximately symmetric for moderate Q.

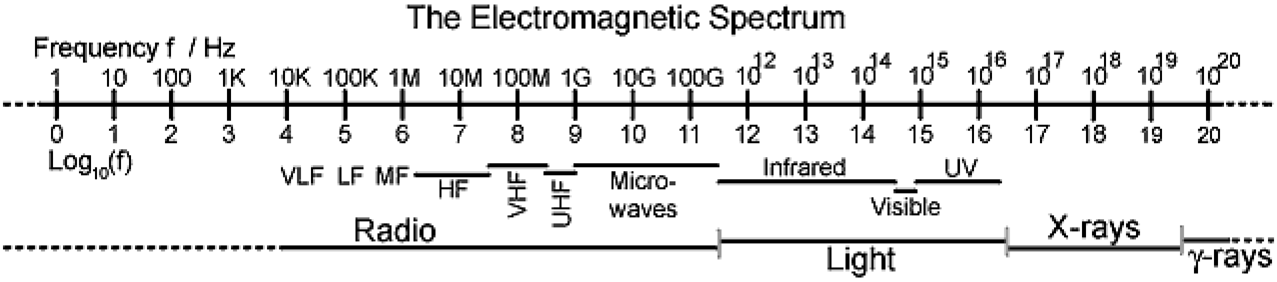
Although the bandwidth function is not symmetric about its peak if we choose frequency as the horizontal (x) axis, it can be made symmetrical if we instead plot it against an appropriately chosen *function of frequency*. In particular, we need a frequency function such that any resonance peak is always just as far from zero-frequency as it is from infinite frequency, i.e., we need an infinity to the left of the resonance, and an infinity to the right, and both infinities must be of the same type. Such a requirement is satisfied by, and indeed is one of the principal properties of, the logarithmic function:

$$x = \mathrm{Log}(f)$$

the choice of logarithm base being arbitrary. Hence the peak can be said to be symmetric about its *logarithmic centre frequency*

$x_0 = \mathrm{Log}(f_0)$

Since frequency can be scaled arbitrarily without affecting the shape of the bandwidth function (units of Hz are not mandatory); this matter can be 'proved' numerically by plotting P/P_0 against Log(f) with $f_0 = 1$ and noting that the function is symmetric about $x_0 = \mathrm{Log}(1) = 0$ for any value of Q_0. It should also be noted that there is an infinity of scales from microscopic to macroscopic, and it is often more natural to think in logarithmic dimensions than in linear ones. In the case of frequency, this can be seen by considering the classic representation of the electromagnetic spectrum as illustrated below.



There is no theoretical minimum frequency on the logarithmic scale; although the lowest electromagnetic frequency that can be encountered in practice is the reciprocal of the age of the universe, about $1/(13.82{\times}10^9 \times 365.2421897 \times 24 \times 60^2) = 2.29{\times}10^{-18}$ Hz by current reckoning[21]. Practical DC electrical systems come nowhere close to that because even the best stop working after a few years. Hence zero frequency is impossible; we employ the concept merely as a mathematical convenience for the purpose of circuit analysis.

Note incidentally, that some writers have been moved to claim that there is a flaw in Maxwell's equations because electrical formulae tend to produce infinities when frequency is set to zero. This is a fallacy of course; the infinities being merely a reflection of the fact that zero frequency is not a property of the Universe. A practical consequence however, is that annoying "divide-by-zero" errors occur when putting zero frequency into (for example) frequency-response calculations. The solution, when calculating a frequency response that needs to appear as though it starts from 0 Hz, is to input a very low frequency instead of zero. For radio-frequency calculations, starting from 1 Hz, instead of 0 Hz, will usually do the trick.

[21] 13.82×10^9 years is the age of the Universe according to the cosmic microwave background survey of the ESA Planck telescope (data released in 2013). 365.2421897 days is the mean tropical year as of Jan. 1st 2000.

31. A proper definition for resonant Q

Now that we have established that Q_0 is an important circuit parameter, we will take the opportunity to have another look at at its definition. The point is that there is something horribly unsatisfying about writing:

$$Q_0 = X_{0L}/R \quad \textbf{OR} \quad Q_0 = -X_{0C}/R$$

It seems more logical that the definition should simultaneously involve both inductance *and* capacitance. If so, then why not write:

$$Q_0 = [\ +\sqrt{(-X_{0C}\ X_{0L})}\ ] / R$$

which is the same as multiplying the two standard definitions and taking the positive square root (i.e., taking the geometric mean of the two definitions)?
Of course:

$$-X_{0C}\ X_{0L} = 2\pi f_0 L / (2\pi f_0 C) = L/C$$

(i.e., the L/C ratio)
hence:

$Q_0 = [\ \sqrt{(L/C)}\ ] / R$ or $Q_0 = R_0 / R$	**31.1**

Here is a definition of resonant Q that properly involves all of the components; and as an added bonus in calculation, does not involve the resonant frequency.

32. Bandwidth in terms of Q

A bandwidth function for the series resonator was derived earlier and given as equation (**29.2**):

$$P / P_0 = R^2 / (R^2 + [2\pi fL - 1/(2\pi fC)]^2)$$

Now that we have a sensible definition for Q_0 however, we can see that it can be used as a substitution for R in the expression above, i.e.(from **31.1**):

$$R = [\ \sqrt{}(L/C)\ ] / Q_0$$

Hence:

$$P/P_0 = \frac{[(L/C) / Q_0^2]}{[(L/C) / Q_0^2] + [2\pi fL - 1/(2\pi fC)]^2}$$

Now, if we forcibly factorise the quantity L/C from the right hand term in the denominator we obtain:

$$P/P_0 = \frac{[(L/C) / Q_0^2]}{[(L/C) / Q_0^2] + (L/C)\{\ [\sqrt{}(C/L)][2\pi fL - 1/(2\pi fC)]\}^2}$$

which, noting that $L/\sqrt{L} = \sqrt{L}$ and $(\sqrt{C})/C = 1/\sqrt{C}$, simplifies to:

$$P/P_0 = \frac{[1 / Q_0^2]}{[1 / Q_0^2] + \{\ f2\pi\sqrt{}(LC) - 1/[f2\pi\sqrt{}(LC)]\ \}^2}$$

We cans substitute for the quantity $2\pi\sqrt{}(LC)$ by noting that the standard series resonance formula can be rearranged thus:

$$2\pi\sqrt{}(LC) = 1/f_0$$

Hence:

$$\boxed{P/P_0 = \frac{[1 / Q_0^2]}{[1 / Q_0^2] + [\ (f / f_0) - (f_0 / f)\ ]^2}} \qquad (\mathbf{32.1})$$

This puts the bandwidth function into a form most similar to a curve known as the Lorentzian line-shape function (next section), but a further simplification is possible by multiplying both numerator and denominator by Q_0^2 :

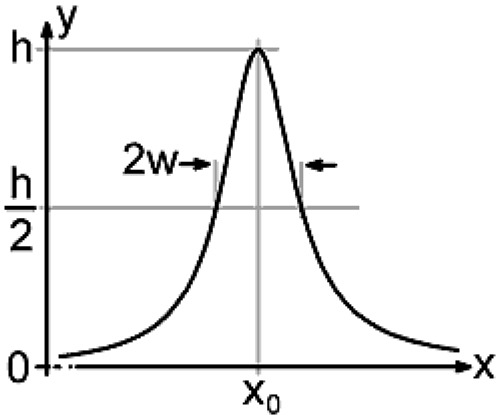
$$\boxed{P/P_0 = \frac{1}{1 + \{\ Q_0[\ (f / f_0) - (f_0 / f)\ ]\ \}^2}} \qquad (\mathbf{32.2})$$

which demonstrates in the clearest possible way that the bandwidth of an LC resonator is dictated entirely by Q_0 and f_0 .

33. Lorentzian line-shape function

The electrical resonance curve is closely related to a simple mathematical function known as the ***Lorentzian*** (or Cauchy) line-shape function, which has the general form:

$$\boxed{y = \frac{h\,w^2}{w^2 + (x - x_0)^2}} \qquad (33.1)$$



where h is the peak height and w is called the half-width. The expression can also be written:

$$\boxed{\frac{y}{h} = \frac{1}{1 + [(x - x_0)/w]^2}} \qquad (33.2)$$

which is the form most similar to equation (**32.2**).

The Lorentzian is regarded as the characteristic signature of natural electromagnetic resonance processes. In particular, the peaks in molecular and atomic spectra in the microwave, optical, x-ray and gamma-ray regions are all of this form when displayed on a linear amplitude (y-axis) scale. The curve is called a line-shape function because the narrow spikes that occur when dense spectra are drawn by a chart-recorder or otherwise displayed are traditionally known as *lines*[22]. It is only when the frequency scale is expanded that the individual peaks resolve into Lorentzians.

In comparing the Lorentzian to the electrical resonance curve, we can first note that the Lorentzian is always exactly symmetric about x_0, and that x_0 can be set to zero. We have noted before (section **30**) that the electrical resonance curve is skewed when plotted against a linear frequency scale, but becomes symmetric to a good approximation when the Q is high. We also noted that the resonance curve can be made perfectly symmetric by plotting it on a logarithmic frequency scale; in which case, since the logarithm of unit frequency is zero, the curve can also be symmetric about $\mathrm{Log}(f) = 0$. In fact, natural resonance processes have such high Q that they appear symmetric on linear, logarithmic, and even reciprocal (wavelength) scales; but to find the relationship between the Lorentzian and the electrical curve, it is obvious that we must identify the x-axis as corresponding to logarithmic frequency, i.e., $x = \mathrm{Log}_{\mathbf{a}}(f)$, where the base a can be chosen arbitrarily. Here we will use Naperian logarithms because it will allow us to use the series-expansion of e to solve the problem. Hence we choose:

$$x = \mathrm{Log}_{\mathbf{e}}(f)$$

which means that:

$$f = e^{\mathbf{x}}$$

and

$$f_0 = e^{\mathbf{x_0}}$$

[22] Early optical spectra were recorded literally as lines, on a photographic plate exposed using a prism or diffraction-grating to split the light into its component colours. In that case, the line-shape function is the optical density (opacity) profile of the photographic emulsion in a direction perpendicular to the line.

Substituting these identities into the electrical resonance curve (32.2) we obtain:

$$P/P_0 = \frac{1}{1 + \{ Q_0 [(e^x / e^{x_0}) - (e^{x_0} / e^x)] \}^2}$$

but, from the rules of logarithms discussed in section **28**:

$$e^x / e^{x_0} = e^{x-x_0}$$

and

$$e^{x_0} / e^x = e^{x_0-x} = e^{-(x-x_0)}$$

Hence:

$$P/P_0 = \frac{1}{1 + [Q_0 (e^{x-x_0} - e^{-(x-x_0)})]^2}$$

The quantity $e^{x-x_0} - e^{-(x-x_0)}$ is related to a function known as the hyperbolic sine (Sinh, pronounced "shine"), which is defined as:

$$\mathrm{Sinh}(x) = (e^x - e^{-x})/2$$

Hence:

$$\boxed{P/P_0 = \frac{1}{1 + [2\, Q_0\, \mathrm{Sinh}(x - x_0)]^2}} \qquad (33.3)$$

The function e^x can be expanded as an infinite series:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \ldots\ldots\ldots\ldots\ldots$$

where an exclamation mark indicates a *factorial number*, the factorials being defined as:

Factorial	0!	1!	2!	3!	4!	n!	(n+1)!
Value	1	1	2×1	3×2×1	4×3×2×1	n(n-1)(n-2)× . . . . . . ×1	(n+1)×n!

It follows that the series for e^{-x} is:

$$e^{x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} - \frac{x^5}{5!} + \ldots\ldots\ldots\ldots$$

and that by subtracting one series from the other we can obtain a series for $e^{x}-e^{-x} = 2\,\mathrm{Sinh}(x)$

$$2\,\mathrm{Sinh}(x) = 2x + \frac{2x^3}{3!} + \frac{2x^5}{5!} + \frac{2x^7}{7!} + \frac{2x^9}{9!} + \ldots\ldots\ldots$$

or

$$\mathrm{Sinh}(x) = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \frac{x^7}{7!} + \frac{x^9}{9!} + \ldots\ldots\ldots$$

Now notice that when the magnitude of x is somewhat less than 1, the magnitudes of the terms in which x is raised to a high power become very small, and so we can make the approximation:

$\mathrm{Sinh}(x) \approx x$ when $|x| < 1$ and $\mathrm{Sinh}(x) \rightarrow x$ when $|x| << 1$

(' ≈ ' means 'approximately equal to' ; ' << ' means 'much less than').

Substituting this into equation (33.3) we get:

$$P/P_0 \approx \frac{1}{1 + [\,2\,Q_0\,(x - x_0)\,]^2}$$

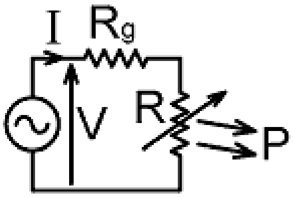
which is a Lorentzian with $w = 1/(2Q_0)$. Hence the electrical resonance curve is Lorentzian when $|x - x_0| << 1$. The electrical resonance curve is, of course, an electromagnetic resonance curve; and like any spectral line, is Lorentzian when the Q of the resonance is reasonably large.

34. Maximum power transfer

So far we have considered generators to be sources of constant RMS voltage. In the case of a physically realisable generator however (or a device or network that is to be considered as a generator); in the absence of a control system to keep it constant, the output voltage will droop as the output current is increased. This means that the generator has an internal impedance, which is somehow distributed throughout its wiring and component parts, but which will be seen from outside as though there is a single impedance in series with an otherwise perfect generator. This impedance is known as the *source impedance* or the generator's *output impedance*, and must often be taken into account when carrying out circuit analysis. In particular, it is necessary to include the source impedance explicitly when determining the characteristics of the parallel resonator bandpass filter; but there are various other connotations relating to power transmission in general. The basic matter is that of the effect that the load impedance has on the amount of power delivered to the load, and is encapsulated in a set of relationships known as the *maximum power-transfer theorem*.

For the special case of a generator with a purely resistive output impedance and a purely resistive load, we can obtain the maximum power-transfer condition using a graphical method. The circuit to be considered is shown below, where R_g is the generator output resistance, and R is the load. V is the off-load generator voltage, i.e., it is the voltage that will be seen at the generator terminals when the load R is disconnected. It should also be obvious by inspection that no power is delivered when $R = 0$ (short-circuit), and also that no power is delivered when R is disconnected (i.e., when $R \to \infty$). Hence we expect a peak in power output at some intermediate value of R, and we can obtain that value in relation to R_g by determining the relationship between P and R and plotting it as a graph.

In the circuit on the right, the power delivered to the load is:



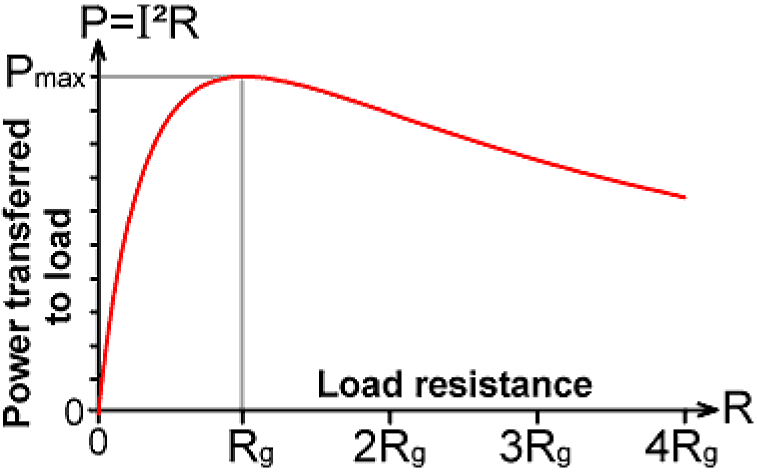
$P = I^2 R$

Where:

$I = V / (R + R_g)$

Hence:

$P = V^2 R / (R + R_g)^2$	34.1

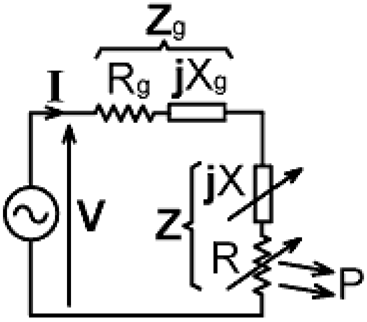


This function is plotted on the right, for constant V, and shows that maximum power output occurs when

$R = R_g$

The result is, of course, well known, but it is by no means the whole story, and its interpretation is subject to various common misconceptions. We can settle all of these issues by deriving the complete maximum power transfer condition (see box below). This requires the use of calculus, which will not be explained here, but those unfamiliar with the technique may still avail themselves of the result.

The maximum power transfer theorem:
In the circuit shown on the right, the power P delivered to the load **Z** is:



$P = |\mathbf{I}|^2 R$

where $|\mathbf{I}| = |\mathbf{V}| / |(\mathbf{Z} + \mathbf{Z}_g)|$

Hence:

$P = |\mathbf{V}|^2 R / |(\mathbf{Z} + \mathbf{Z}_g)|^2 = |\mathbf{V}|^2 R / |(R+R_g + \mathbf{j}[X+X_g])|^2$

$\boxed{P = |\mathbf{V}|^2 R / [(R + R_g)^2 + (X + X_g)^2]}$

There are two maximum power transfer conditions to be obtained here, one being the value of load *reactance*, and the other being the value of load *resistance*. For changes in either of these variables, there will be a peak in the graph of power versus the variable, and the peak will of course occur at the point where the gradient of the curve is zero. Hence, for the reactance condition, maximum power transfer occurs when $\partial P/\partial X= 0$, and for the resistance condition, maximum power transmission occurs where $\partial P/\partial R= 0$ (where ∂ is known as "partial d" or "curly d" and indicates a partial differential; i.e., differentiation of one variable with respect to another is carried out with all other variables held constant). In order to carry out these differentiations on the expression above, we can use the quotient rule:

If $y = N/D$ **then** $dy/dx = (D\, dN/dx - N\, dD/dx) / D^2$

Hence if we let $N = |\mathbf{V}|^2 R$

and

$D = (R+R_g)^2 + (X+X_g)^2$

$\quad = R^2 +R_g^2 +2RR_g +X^2 +X_g^2 +2XX_g$

then:

$\partial N/\partial X= 0$, $\partial D/\partial X=2X+2X_g$, $\partial N/\partial R=|\mathbf{V}|^2$ and $\partial D/\partial R=2R+2R_g$

Hence:

$\partial P/\partial X = [0 - |\mathbf{V}|^2 R (2X+2X_g)] / D^2$

$\qquad = -2|\mathbf{V}|^2 R (X+X_g) / D^2$

therefore:

$\partial P/\partial X= 0$ **when** $X = -X_g$

(maximum power transfer occurs when the power factor is 1)

and

$$\partial P/\partial R = \{\, |\mathbf{V}|^2[\,(R+R_g)^2+(X+X_g)^2\,] - |\mathbf{V}|^2\, R\,(2R+2R_g)\,\}/\, D^2$$

$$= |\mathbf{V}|^2\,[\,(X+X_g)^2 + R^2 + R_g^2 + 2RR_g - 2R^2 - 2RR_g\,]\,/\,D^2$$

i.e.,

$$\partial P/\partial R = |\mathbf{V}|^2\,[\,R_g^2 + (X+X_g)^2 - R^2\,]\,/\,D^2$$

therefore:

$$\partial P/\partial R = 0 \quad \text{when} \quad R_g^2 + (X+X_g)^2 - R^2 = 0$$

i.e.,

$$\boxed{\partial P/\partial R = 0 \ \ \textbf{when} \ \ R = \sqrt{[\,R_g^2+(X+X_g)^2\,]}}$$

Notice that this latter maximum power transfer condition is a magnitude: it is the same as;

$$R = |\, R_g + \mathbf{j}(X+X_g)\,|$$

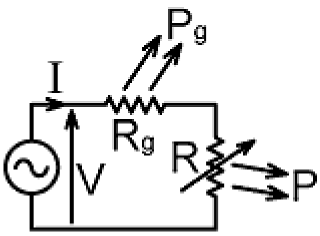
i.e., maximum power transfer occurs when the load resistance is equal to the magnitude of the impedance formed by the source resistance and the total reactance. This also means that if the source impedance is purely resistive, then maximum power transfer occurs when the magnitude of the load impedance is equal to the source resistance. Observe also that when the unity power-factor condition $X=-X_g$ is satisfied, the $(X+X_g)$ term disappears and maximum power transfer occurs when $R = R_g$. Thus the overall maximum power transfer condition occurs when $R = R_g$ *and* $X = -X_g$, i.e, when

$$\boxed{\mathbf{Z} = \mathbf{Z}_g^*}$$

The condition obtained when the load impedance is the complex conjugate of the source impedance is known as a conjugate match.

We can address the most common misconception regarding impedance matching by stating that, although unity power-factor ($X=-X_g$) is always desirable, is it is not necessary and not always desirable that the load resistance should be equal to the source resistance. The reason can be understood by considering the poor generator, which must dissipate power in its internal resistance, and will therefore get hot.

If we assume that power-factor correction will normally be carried out, then there is no need to consider the reactances in the system, and we can analyse the power dissipated in the generator using the case where both the source impedance and the load are purely resistive. Thus:



$P_g = I^2 R_g$

where the current is, as defined earlier: $I=V/(R+R_g)$. Hence:

$$P_g = V^2 R_g / (R+R_g)^2 \qquad \textbf{34.2}$$

We will plot this function shortly; but when doing so it will be interesting to use the comparison between the load power and the power wasted in the generator as a measure of the power transmission efficiency. We can define efficiency as:

Transmission efficiency = Power delivered / Total power generated

and here we will give it the symbol η (Greek lower case 'eta'). Thus:

$\eta = P / (P + P_g)$

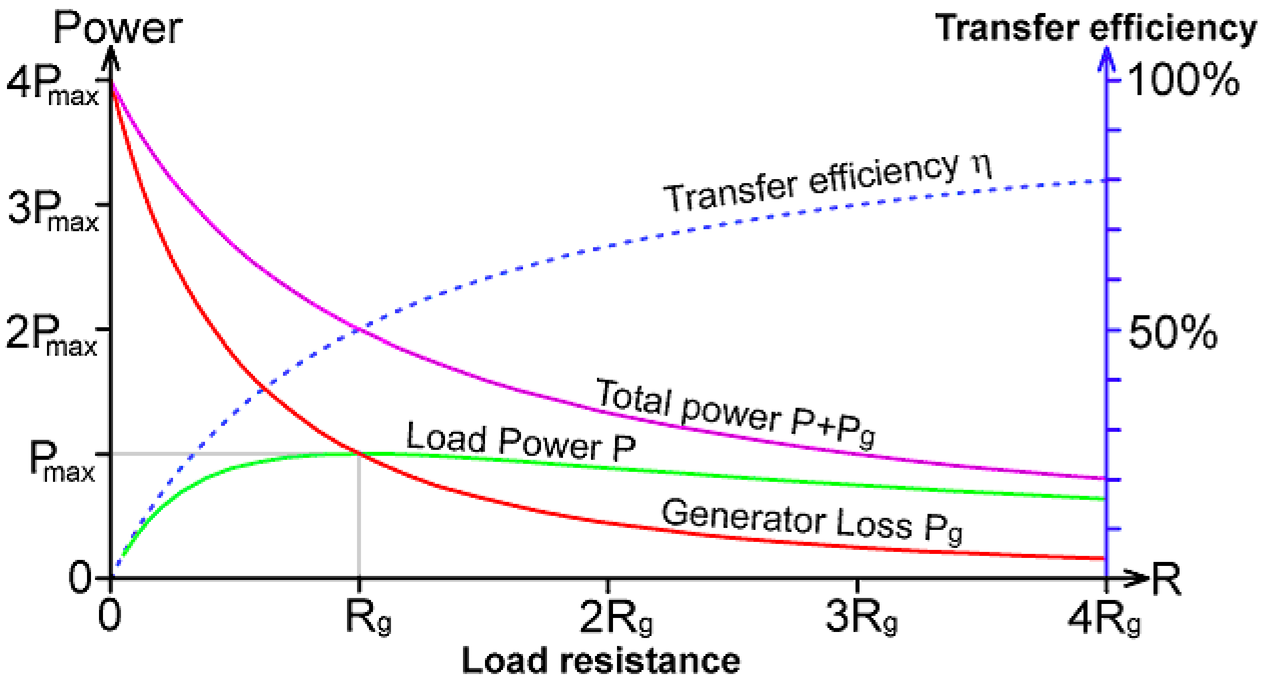
Now, substituting the definitions of P (**34.1**) and P_g (**34.2**) into this expression we get:

$\eta = V^2 R/(R+R_g)^2 / \{ [V^2 R/(R+R_g)^2] + [V^2 R_g/(R+R_g)^2] \}$

i.e.,

$$\eta = R / (R + R_g)$$

Shown plotted below for comparison are: P, the power delivered to the load; P_g, the power dissipated as heat in the generator; $P+P_g$, the total power generated; and η, the ratio of power delivered to power generated, expressed as a percentage.

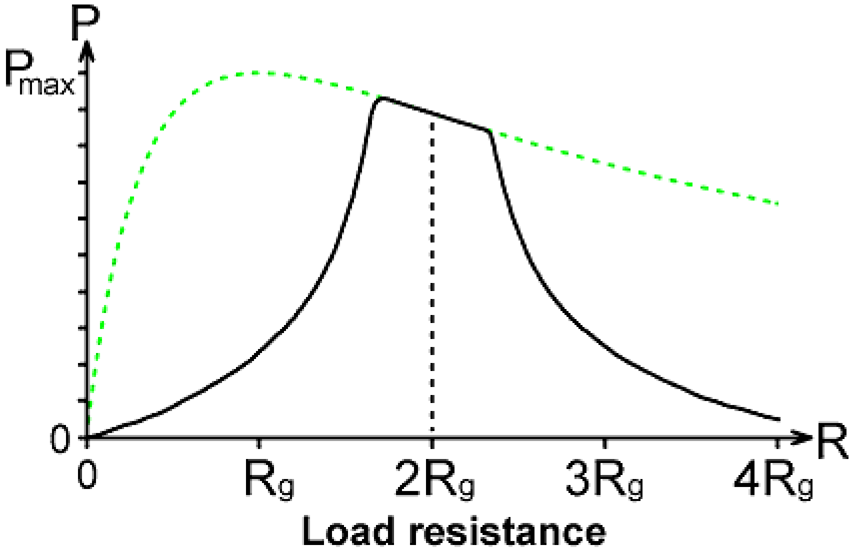


The tabulated results below show the various power levels as a proportion of the maximum deliverable power P_{max}, and are applicable to any power-factor corrected generator-load system.

Load R / R_g	**Total Power** $(P+P_g) / P_{max}$	**Power Loss** P_g / P_{max}	**Load Power** P / P_{max}	**Load power** / dB	**Efficiency** $R / (R+R_g)$
0	.... 4.00	.... 4.00	0.00	-∞	0.00
1/16	.... 3.76	.... 3.54	. 0.22	-6.55	. 0.06
1/8	.... 3.56	.... 3.16	.. 0.40	-4.03	. 0.11
1/4	.... 3.20	.... 2.56	... 0.64	-1.94	.. 0.20
1/2	.... 2.67	.... 1.78	.... 0.89	-0.51	... 0.33
1	.... 2.00	...... 1.00	..... 1.00	0.00	..... 0.50
2	...... 1.33	.. 0.44	.... 0.89	-0.51	.... 0.67
4	.... 0.80	. 0.16	... 0.64	-1.94	.... 0.80
8	.. 0.44	0.05	.. 0.40	-4.03	.... 0.89
16	. 0.26	0.01	. 0.22	-6.55	.... 0.94

Notice in the graph above that as the load resistance R is increased and becomes greater than source resistance R_g, the power delivered to the load tails off gently. The reason for this behaviour is that, as the current drawn from the generator reduces, the output voltage increases; and so the system possesses a self-regulating property when lightly loaded. When the load is twice the source resistance, the power delivered is still 89% of the maximum possible, a droop in output of only 0.51 dB.

The major advantage of light loading however is seen in the transfer efficiency. When a conjugate match is achieved, the efficiency is only 50%, but it rises to 67% (2/3) when $R=2R_g$, and 80% (4/5) when $R=4R_g$. This means that light loading, when compared to conjugate matching, gives a reduction in generator dissipation and power input for a given power output. In radio practice, of course, the generator is a radio transmitter; and if the transmitter is designed for light loading it can have smaller heat-sinks and reduced battery or mains power consumption in comparison to a transmitter designed for conjugate matching. Consequently, the figure often referred to as the "output impedance" of a radio transmitter (often 50 Ω) is usually nothing of the sort, it is instead (and should be called) the *preferred load-resistance*, or alternatively the *design load resistance*. The preferred load resistance of a broadband transistor power amplifier is usually higher than the output impedance, and attempting to provide such an amplifier with a conjugate match will result in excessive internal dissipation, overheating, and possibly catastrophic failure. Fortunately, most modern amplifiers are provided with protection circuitry to prevent over-dissipation, and this circuitry gives the transmitter a loading characteristic that makes it appear that the source resistance is higher that it really is. This loading characteristic will be different from the power transfer curve derived above because it is caused by the action of non-linear circuit elements (level detectors etc.), and so the load resistance that corresponds to the middle of the permitted operation window is known as the *pseudo output-impedance* (or, if you like, the *pseudo source-resistance*). The diagram below shows what the power transfer curve might look like with the operation window centred on twice the source resistance (unprotected transfer-function shown dotted).



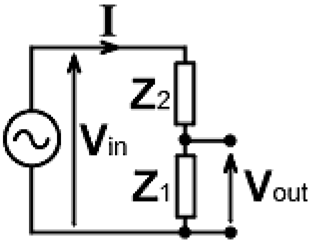
Notice that the protection circuitry also operates when the load resistance is higher than the preferred value. This is not usually necessary for the protection of push-pull transistor power amplifiers (the most common type of output stage in modern practice); but it helps to ensure that any harmonic suppression filter after the amplifier will function correctly, and it occurs because the load impedance is traditionally detected using a bridge circuit (often called a reflectometer or SWR bridge, but really an impedance bridge) balanced for a particular value of resistance. An interesting discussion of the conditions that provoke transistor failure is given by Bob Pearson[23].

If the protection circuitry is correctly designed and adjusted, the pseudo output-impedance should be the same as the preferred load-resistance. When determining the effect of source impedance on the Q of antenna systems and bandpass filters however, it is the *true output-impedance*, not the preferred load-resistance that must be used. Unfortunately, this quantity is often impossible to obtain from the manufacturer's data; but, as we will see shortly; despite the efforts of the protection circuitry to disguise it, it can be measured with the aid of two dummy load resistors of different value.

[23] **"How Big is a Bad SWR?"** Bob Pearson, G4FHU, Rad Com, March 1993, p64-65, April 1993, p62-63. The greatest danger for push-pull transistor amplifiers lies in low load resistance; which, for a given value of SWR, is more pernicious than residual reactance. SWR is a poor matching criterion because it does not indicate whether the magnitude of the load impedance is too high (harmless) or too low (harmful).

35. The potential divider.

The potential divider is the simplest *three-terminal* electrical network. We have made some use of its properties already, but there comes a point when it is useful to characterise it formally. Here we will do so for the general case, which is that of defining the voltage at the intersection of two impedances. Referring to the diagram:



$\mathbf{V}_{\mathrm{out}} = \mathbf{I}\,\mathbf{Z}_1$

where, since $\mathbf{V}_{\mathrm{in}} = \mathbf{I}\,(\mathbf{Z}_1+\mathbf{Z}_2)$:

$\mathbf{I} = \mathbf{V}_{\mathrm{in}} / (\mathbf{Z}_1 + \mathbf{Z}_2)$

hence

$$\mathbf{V}_{\mathrm{out}} = \mathbf{V}_{\mathrm{in}}\,\mathbf{Z}_1 / (\mathbf{Z}_1 + \mathbf{Z}_2) \tag{35.1}$$

and if we multiply the right-hand side by $\mathbf{Z}_2 / \mathbf{Z}_2$:

$$\mathbf{V}_{\mathrm{out}} = \mathbf{V}_{\mathrm{in}}\,(\mathbf{Z}_1 \,//\, \mathbf{Z}_2) / \mathbf{Z}_2 \tag{35.2}$$

Note that $\mathbf{Z}_2$ is the sum of the source impedance and any additional impedance placed in series with the generator. $\mathbf{V}_{\mathrm{in}}$ is the off-load generator voltage.

If the impedances are pure resistances, the formula above reverts to:

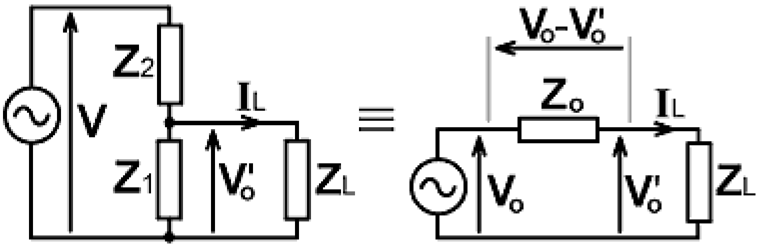
$$V_{\mathrm{out}} = V_{\mathrm{in}}\,R_1 / (R_1 + R_2) \tag{35.3}$$

where R_1 is the resistance across which V_{out} is said to appear.
Alternatively, multiplying by R_2 / R_2 :

$$V_{\mathrm{out}} = V_{\mathrm{in}}\,(R_1 \,//\, R_2) / R_2 \tag{35.4}$$

36. Output impedance of potential divider

The output impedance of a network is defined as that impedance which, when placed in series with a hypothetical perfect generator, accounts for the drop in output voltage that occurs when a load is connected. Shown below is a representation of a potential divider network loaded with an impedance $\mathbf{Z}_L$. If the load is removed, the output voltage is $\mathbf{V}_o$, but when the load is connected, the output drops to a new voltage $\mathbf{V}_o'$ (the single inverted comma is pronounced "prime"). This situation is modelled on the right as a perfect generator with an output $\mathbf{V}_o$ in series with an impedance $\mathbf{Z}_o$, the latter being the output impedance we wish to define.



Using the definitions given in the diagram:

$\mathbf{Z}_o = (\mathbf{V}_o - \mathbf{V}_o') / \mathbf{I}_L$

where:

$\mathbf{I}_L = \mathbf{V}_o' / \mathbf{Z}_L$

Hence, combining these two equations:

$\mathbf{Z}_o = \mathbf{Z}_L (\mathbf{V}_o - \mathbf{V}_o') / \mathbf{V}_o' = \mathbf{Z}_L [(\mathbf{V}_o / \mathbf{V}_o') - 1]$ . . . . . . (36.3)

From equation (35.2) given above:

$\mathbf{V}_o = \mathbf{V} (\mathbf{Z}_1 // \mathbf{Z}_2) / \mathbf{Z}_2$

and by considering $\mathbf{Z}_L$ as part of the potential divider itself:

$\mathbf{V}_o' = \mathbf{V} (\mathbf{Z}_1 // \mathbf{Z}_L // \mathbf{Z}_2) / \mathbf{Z}_2$

Hence

$\mathbf{V}_o / \mathbf{V}_o' = (\mathbf{Z}_1 // \mathbf{Z}_2) / (\mathbf{Z}_1 // \mathbf{Z}_L // \mathbf{Z}_2)$

$= [(1/\mathbf{Z}_1) + (1/\mathbf{Z}_L) + (1/\mathbf{Z}_2)] / [(1/\mathbf{Z}_1) + (1/\mathbf{Z}_2)]$

$= 1 + (1/\mathbf{Z}_L) / [(1/\mathbf{Z}_1) + (1/\mathbf{Z}_2)]$

$= 1 + (\mathbf{Z}_1 // \mathbf{Z}_2) / \mathbf{Z}_L$

Substituting this into (36.3) gives:

$\mathbf{V_o} = Z_L \; \{ \; 1 + [\; (\mathbf{Z}_1 \, // \, \mathbf{Z}_2) \, / \, Z_L \;] - 1 \; \}$

i.e.:

$\boxed{\mathbf{Z_o} = \mathbf{Z}_1 \, // \, \mathbf{Z}_2}$

The output impedance of a potential divider is the parallel combination of the component impedances.

Note that the output impedance of the main generator is part of $\mathbf{Z}_2$. If however, as is often the case, this source impedance is small in comparison to the total $\mathbf{Z}_2$, then it can be neglected.

37. Thévenin's theorem.

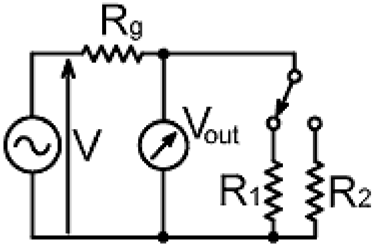
The current in any impedance **Z**, connected to a network consisting of any number of impedances and generators, is the same as though the impedance were connected to a single generator having an output impedance equal to the impedance seen looking back into the network when all generators are replaced by their output impedances, and an output voltage equal to the voltage that appears at the terminals when **Z** is disconnected.

Thévenin's theorem (pronounced "tae-ven-in") arises from the observation that the output impedance of a generator is effectively in parallel with its load, and so the output impedance of an active network is its impedance when all of its generators are replaced by their internal impedances. This is entirely logical when we think of the generator as an ideal generator in series with an impedance, because the ideal generator part of the model is a short-circuit with regard to any power reflected back into the network. Thévenin's theorem simply takes this observation to its logical conclusion by allowing that the network can be represented as a single generator, which may be characterised completely by knowing its output impedance and its off-load voltage (and its output spectrum, but in a linear network we can treat each frequency component separately and so need only consider sine waves).

Notice that in the preceding section we could have obtained the output impedance of the potential divider directly by using Thévenin's theorem; i.e., with the generator replaced by a short-circuit, it is obvious that the impedance looking back into the network is $\mathbf{Z}_1 \, // \, \mathbf{Z}_2$.
This is an extremely useful trick; but even more useful analytically is the technique of replacing a complicated network with a single generator and a series impedance (i.e., the output impedance or source impedance). The replacement network is known as the 'Thévenin equivalent'.

38. Measuring source resistance

In the test setup shown below; the output voltage of a generator is measured with two different load resistances, all other variables being kept constant, and the circuit being constructed in such a way as to minimise stray capacitance and inductance (i.e., using very short wires). It is assumed that the source impedance is purely resistive, this being reasonable in the case of a transistor RF amplifier, but very unreasonable in the case of a tuned valve (tube) RF amplifier. In order to avoid interference from any protection circuitry, the test should be carried out at a low power level (<10% of maximum output). The voltmeter should be capable of measurement at the generator frequency and should have a high input resistance. An oscilloscope with a high-impedance probe is suitable, but ordinary multimeters do not work at radio frequencies. Only the *voltage ratio* needs to be determined accurately, the absolute voltages are immaterial.



Let the output voltages be V_1 when R_1 is connected, and V_2 when R_2 is connected. The source and load resistances form a potential divider. Hence (using equation **35.3**):

$$V_1 = V R_1 / (R_g + R_1)$$

and

$$V_2 = V R_2 / (R_g + R_2)$$

Rearranging both of these expressions to get V on its own and then equating them gives:

$$V = V_1 (R_g + R_1)/R_1 = V_2 (R_g + R_2)/R_2$$

$$R_2V_1 (R_g + R_1) = R_1V_2 (R_g + R_2)$$

$$R_g (R_2V_1 - R_1V_2) = R_1 R_2 (V_2 - V_1)$$

$$\boxed{R_g = R_1 R_2 (V_2 - V_1) / (R_2V_1 - R_1V_2)}$$

If (say) V_1 is factored out of the numerator and denominator, a form is obtained that makes it clear that only the voltage ratio is needed:

$$R_g = R_1 R_2 ([V_2/V_1] - 1) / (R_2 - R_1V_2 /V_1) \qquad \textbf{38.1}$$

A respectable difference between the two load resistors is necessary in order to minimise the effect of measurement errors, but too large a deviation from the preferred load resistance is likely to provoke a transistor power-amplifier's protection circuitry. For a transmitter designed to operate into a 50 Ω load; 25 Ω and 100 Ω dummy-load resistors correspond to the upper and lower 2:1 SWR points and should give an easily discernible output voltage difference. A 25 Ω resistor can be had by connecting two 50 Ω dummy load resistors in parallel with a coaxial T-piece. 100 Ω coaxial resistors are less readily available, but an old-fashioned 75 Ω load will do instead. In the calculation, the actual resistance of the load measured with an accurate resistance meter should be used (rather than the nominal value stamped on the resistor).

Example: The output voltage of a Kenwood TS430S 100 W HF transmitter was measured with two different dummy loads. The measurement frequency was 1.9 MHz, and the test power level was very approximately 1 W. One load was a 75 Ω nominal coaxial resistor measuring 75.1±0.7 Ω, the other was the combination of this resistor and a 50 Ω nominal coaxial resistor in parallel with it, the combination measuring 29.6±0.3 Ω. The voltage ratio was measured using an oscilloscope with a 10 MΩ ×10 probe. The resistors and the probe were attached directly to the antenna socket using coaxial T-pieces (no cables). The measurement was made by attaching and removing the 50 Ω resistor from the T-piece with the transmitter running and noting the change in the peak to peak excursion of the output waveform. Using the following designations: $R_2 = 75.1$ Ω , $R_1 = 29.6$ Ω , the voltage ratio V_2/V_1 was 1.364 ±0.04. Using equation (**38.1**), the source resistance R_g was calculated to be 23.3 Ω. An error analysis (see next section), gave an estimated standard deviation of 3.4 Ω ; i.e., $R_g = 23.3 \pm 3.4$ Ω . Note incidentally, that this determination assumes that the output impedance does not change with power output level. Given that power transistors are non-linear devices, this is not necessarily true.

39. Error analysis

While it would be inappropriate here to delve too deeply into the subject of scientific data analysis; the reader should nevertheless be aware that all physical measurements are meaningless unless they have some kind of error-window or *confidence interval* associated with them. This is not a serious problem when taking a reading with say, a multimeter, because (assuming that the instrument has been calibrated), the manual will say what the measurement accuracy is. A digital multimeter, for example, might have a quoted accuracy of ±0.8% ±1digit (i.e., ±1 in the last decimal place) for its resistance ranges, so if we obtain a resistance reading from this instrument of (say) 75.1 Ω, the actual measurement will have a confidence interval of ±0.6 ±0.1, i.e., the reading should be recorded as 75.1±0.7 Ω. Scientists and engineers normally equate error boundaries stated in this way with the *estimated standard deviation* (ESD) of the measurement; where, on the assumption that errors are scattered randomly according to a 'normal' or Gaussian distribution, a standard deviation represents a region where we have a 68% confidence that the true result will lie. The standard deviation is usually given the symbol σ (Greek lower case 'sigma'), and so if we obtain a measurement $x \pm \sigma$, we have 68% confidence that the true answer lies between $x-\sigma$ and $x+\sigma$. From the properties of the Gaussian error distribution also, we have a 95.5% confidence that the true answer lies between $x-2\sigma$ and $x+2\sigma$, and a 99.7% confidence that the true answer lies between $x-3\sigma$ and $x+3\sigma$ [ref.[24]]. The use of standard deviations rather than 'brick wall' tolerances reflects the reality that there is always a finite probability that the true result will lie outside the stated error range. We can only ever have *absolute confidence* that the magnitude of the true answer lies somewhere between zero and infinity, but we expect only 3 measurements in every 1000 to fall outside $x \pm 3\sigma$.

It is *always* advisable to try to write down an ESD for every measurement made. This is a reasonably straightforward matter where direct measurements are involved, but a difficulty arises in situations where several measurements are made and then put into a formula in order to obtain the required result. The problem is that of working out how much influence the deviation of a particular variable has on the overall result, and how to add the various deviations together in order to arrive at the overall ESD. It is therefore fortuitous that we have been engaged in the study of

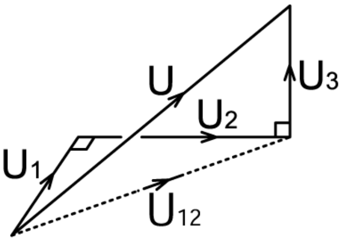
[24] **Data Reduction and Error Analysis for the Physical Sciences**, Philip R Bevington. McGraw-Hill, 1969. Library of Congress cat. card # 69-16942.
2-2: Sample mean and standard deviation. Area under the Gaussian distribution: Table C-2, p308.

vectors, because it turns out that this is a problem of vector addition and magnitudes.

If two or more measurements are made in such a way that the outcome of one has no influence on the outcome of any of the others, the measurement errors are said to be *uncorrelated*. An example of uncorrelated errors is that of readings taken from two separate instruments, where an error or inaccuracy in the reading of one instrument is not related to any error or inaccuracy in the reading of the other. On the other hand, the errors in two measurements made using the same instrument may be correlated, in the sense that if the instrument always reads too high or too low, it will introduce errors in the same direction in both cases. If measurement errors are correlated, then it means that there is some systematic (design, interpretation, or calibration) defect in the measuring process; but if we believe that the measurements have been made to the best of our abilities with the equipment available, then it is usually sensible to assume that any measurement errors are uncorrelated.

Now, if the errors in two or more measurements are uncorrelated, this means that a deviation from the true value in one measured quantity can occur without influencing the deviations in any of the other quantities. If we determine a quantity by applying a formula to a set of measurements, each measurement will contribute a random error to the result, but there is just as much chance that the error due to one measurement will partly cancel the error due to another as there is that a pair of error contributions will both increase or decrease the result. Therefore it will be unduly pessimistic to add the uncertainty contributions of the individual measurements directly. Instead, we should allow for the independence of the uncertainty contributions by regarding each one as a vector pointing in a direction that is at right angles (orthogonal) to all of the others. In effect, by virtue of its randomness, each uncertainty contribution exists in its own dimension, and we can identify its magnitude as its length in that dimension. It follows that the overall uncertainty is the length (i.e., the magnitude) of the vector that results from the addition of a set of orthogonal uncertainty vectors. This situation is represented in the diagram below, where U_1, U_2, and U_3 are the uncertainty contributions to the determined value of an unknown, and U is the overall uncertainty in the result. We can easily find U by successive application of Pythagoras' theorem, as follows:

Let the magnitude of the vector sum of $\mathbf{U}_1$ and $\mathbf{U}_2$ be U_{12}:



$$U_{12} = \sqrt{(U_1^2 + U_2^2)}$$

Then U is the magnitude of the vector sum of $\mathbf{U}_{12}$ and $\mathbf{U}_3$:

$$U = \sqrt{(U_{12}^2 + U_3^2)}$$

but $U_{12}^2 = U_1^2 + U_2^2$

Hence:

$$U = \sqrt{(U_1^2 + U_2^2 + U_3^2)}$$

This process can be extended to find the magnitude of a vector in an arbitrary number of dimensions (we can't make perspective drawings in more than three dimensions, but there is no restriction on the number of dimensions that a vector can have). Hence:

$$U = \sqrt{(U_1^2 + U_2^2 + U_3^2 + \ldots + U_n^2)}$$

Now note that this formula says: "to find the overall uncertainty; calculate the sum of the squares of the *uncertainty contributions* and take the square root." The uncertainty contributions are *not the* *same* as the uncertainties in the measurements made.

Imagine that an unknown quantity x is given by a formula *f* , which is a mathematical function involving measurable quantities (variables) m_1 , m_2 , m_3 , etc. We can express this situation by writing:

$x = f(m_1 , m_2 , m_3 , ...)$

and we can determine x by plugging m_1 , m_2 , m_3 , etc. into the formula. We can also determine the uncertainty contribution due to any one of the variables by changing it and noting the change that occurs in x. The obvious amount by which to change the variable is its standard deviation, hence:

$x+\sigma_{x1} = f(m_1+\sigma_1 , m_2 , m_3 , ...)$

Here we have assumed that a positive change in m_1 will cause a positive change in x. This might not be the case, but since we intend to add the contributions from changes in each of the variables as orthogonal vectors, it makes no difference either way. Now, restoring m_1 to its original value we determine the uncertainty contribution due to m_2.

$x+\sigma_{x2} = f(m_1 , m_2+\sigma_2 , m_3 , ...)$

and so on. If we work through all of the variables in this way and determine their error contributions, we can obtain an estimate of the standard deviation of x by summing the squares of the contributions and taking the square root:

$\sigma = \sqrt{(} \sigma_{x1}^2 + \sigma_{x2}^2 + \sigma_{x3}^2 + ... + \sigma_{xn}^2)$

Note that there are a number of assumptions inherent in this procedure: firstly, as discussed before, that the uncertainties are uncorrelated; and secondly that we have assumed that the function *f* is linear for changes in any of the variables. The latter condition is normally true to a good approximation for small changes, and the effect of any non-linearity is mitigated by the fact that the object of the exercise is to obtain an *estimate*.

Example:
The output resistance R_g of an RF amplifier was determined by loading the output with two different resistances and noting the change in the output voltage with all other conditions held constant. The applicable formula is equation (**38.1**):

$R_g = R_1 R_2 (N_V - 1) / (R_2 - R_1 N_V)$

Where N_V is the ratio of the output voltages:

$N_V = V_2/V_1$

The voltage measurements were made using an oscilloscope, and it was considered that each measurement had an uncertainty of about 2%. It was also considered that these uncertainties were uncorrelated because they were incurred by different operations; one operation being to set the transmitter carrier level and oscilloscope Y-shift until the waveform just touched the top and bottom of the measuring graticule with the higher value resistor connected, the other being to read the height on the graticule with the lower value resistor connected. The overall uncertainty of the voltage ratio measurement was therefore taken to be the square root of the sum of the squares of the two voltage measurements; i.e., $\sqrt{(2^2+2^2)}=2.8\%$, which was rounded to 3% in view of the approximate nature of the estimate. The actual voltage ratio was 1.364, and 3% of 1.364 is 0.04. Hence:

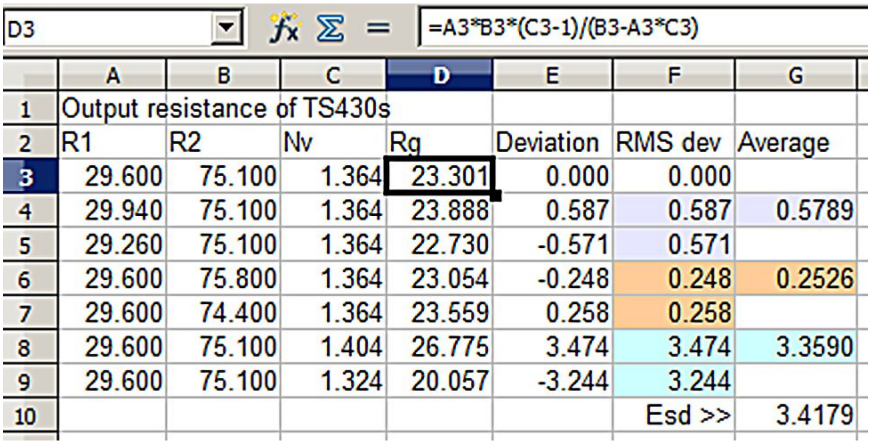
$N_V = 1.364 \pm 0.04$

The resistances were measured using a multimeter known to read correctly within ±0.1 Ω against a standard resistance of 100.0 Ω. The stated accuracy of the instrument was ±0.8% ±1 digit. The measured resistances were $R_1 = 29.6\ \Omega$ and $R_2 = 75.1\ \Omega$. Hence:

$R_1 = 29.6 \pm 0.34\ \Omega$

$R_2 = 75.1 \pm 0.7\ \Omega$

The output impedance R_g was calculated from the formula (**38.1**) using a spreadsheet program and determined to be 23.3 Ω. The output impedance was also calculated with each of the measured values individually incremented and decremented by an amount equal to its estimated standard deviation, and the resulting deviation in R_g was noted. The spreadsheet (**Rg_meas.ods**) is shown below:



D3 =A3*B3*(C3-1)/(B3-A3*C3)

	A	B	C	D	E	F	G
1	Output resistance of TS430s						
2	R1	R2	Nv	Rg	Deviation	RMS dev	Average
3	29.600	75.100	1.364	23.301	0.000	0.000	
4	29.940	75.100	1.364	23.888	0.587	0.587	0.5789
5	29.260	75.100	1.364	22.730	-0.571	0.571	
6	29.600	75.800	1.364	23.054	-0.248	0.248	0.2526
7	29.600	74.400	1.364	23.559	0.258	0.258	
8	29.600	75.100	1.404	26.775	3.474	3.474	3.3590
9	29.600	75.100	1.324	20.057	-3.244	3.244	
10						Esd >>	3.4179

Note that the formula is somewhat non-linear in its behaviour because the deviations caused by incrementing and decrementing a variable are not exactly equal and opposite. The correct way to allow for this effect is to take the average of the deviation magnitude (RMS) for each case. Therefore, the estimated standard deviation in R_g is:

$\sigma = \sqrt{(0.579^2 + 0.253^2 + 3.359^2)} = 3.418\ \Omega$

Hence:

$R_g = 23.3 \pm 3.4\ \Omega$

Notice that the major contributor to the uncertainty in R_g in this case is the uncertainty in N_V. We cannot ignore the effect of the resistance uncertainties however, because if we repeat the experiment with more closely spaced values for R_1 and R_2 , we will find that their contributions to the uncertainty increase dramatically.

Analytical approach to error analysis:
While the error analysis technique just described is perfectly respectable; those who write computer programs will generally prefer an analytical approach. The derivation of an error function from a formula requires the use of calculus. Those who are unfamiliar with calculus may proceed to the next section without losing track of the narrative.

The analytical form of an error function is obtained from the observation that an error in a variable is transmitted through a formula according to the rate of change of the formula with respect to the variable. Thus the error contribution from a variable is the partial derivative of the formula with respect to the variable multiplied by the deviation in the variable. Hence if

$x = f(m_1 , m_2 , m_3 , ...)$

and the ESDs of the measured quantities are $\sigma_1 , \sigma_2 , \sigma_3$, etc.; the contribution that the variable m_1 makes to the ESD of x is given by:

$\sigma_{x1} = (\partial f/\partial m_1) \sigma_1$

and so on (strictly we should take the modulus of the derivative because standard deviations are by definition positive, but it does not matter in this case because orthogonal addition involves squaring of the error contributions). Hence the analytical form of the error function is:

$\sigma = \surd\{ [(\partial f/\partial m_1)\sigma_1]^2 + [(\partial f/\partial m_2)\sigma_2]^2 + [(\partial f/\partial m_3)\sigma_3]^2 + ... \}$

Example:
The output impedance of a generator is obtained from the formula:

$R_g = R_1 R_2 (N_V - 1) / (R_2 - R_1 N_V)$

Differentiation of this function requires the use of the quotient rule:

If $y = N/D$ **then** $dy/dx = (DdN/dx - NdD/dx)/D^2$

where, in this case, the numerator is:

$N = R_1R_2(N_V - 1) = R_1R_2N_V - R_1R_2$

and the denominator is:

$D = R_2 - R_1N_V$

Differentiating the numerator with respect to each of the variables gives:

$\partial N/\partial R_1 = R_2(N_V - 1)$, $\partial N/\partial R_2 = R_1(N_V - 1)$, $\partial N/\partial N_V = R_1R_2$

and differentiating the denominator with respect to each of the variables gives:

$\partial D/\partial R_1 = -N_V$, $\partial D/\partial R_2 = 1$, $\partial D/\partial N_V = -R_1$

Using these results we obtain:

$\partial R_g/\partial R_1 = [\ D(\partial N/\partial R_1) - N(\partial D/\partial R_1)\ ] / D^2$

$= [\ (R_2 - R_1N_V)\ R_2(N_V - 1) - R_1R_2(N_V - 1)(-N_V)\ ] / (R_2 - R_1N_V)^2$

$= [\ (R_2 - R_1N_V)\ R_2(N_V - 1) + R_1R_2(N_V - 1)N_V\ ] / (R_2 - R_1N_V)^2$

$= [\ R_2^2N_V - R_2^2 - R_1R_2N_V^2 + R_1R_2N_V + R_1R_2N_V^2 - R_1R_2N_V\ ] / (R_2 - R_1N_V)^2$

$\partial R_g/\partial R_1 = R_2^2(N_V - 1) / (R_2 - R_1N_V)^2$

$\partial R_g/\partial R_2 = [\ D(\partial N/\partial R_2) - N(\partial D/\partial R_2)\ ] / D^2$

$= [\ (R_2 - R_1N_V)R_1(N_V - 1) - R_1R_2(N_V - 1)\ ] / (R_2 - R_1N_V)^2$

$= [\ R_1R_2N_V - R_1R_2 - R_1^2N_V^2 + R_1^2N_V - R_1R_2N_V + R_1R_2\ ] / (R_2 - R_1N_V)^2$

$\partial R_g/\partial R_2 = R_1^2N_V(1 - N_V\ ) / (R_2 - R_1N_V)^2$

$\partial R_g/\partial N_V = [\ D(\partial N/\partial N_V) - N(\partial D/\partial N_V)\ ] / D^2$

$= [\ (R_2 - R_1N_V)R_1R_2 - R_1R_2(N_V - 1)(-R_1)\ ] / (R_2 - R_1N_V)^2$

$= [\ R_1R_2^2 - R_1^2R_2N_V + R_1^2R_2N_V - R_1^2R_2\ ] / (R_2 - R_1N_V)^2$

$\partial R_g/\partial N_V = R_1R_2(R_2 - R_1) / (R_2 - R_1N_V)^2$

The error function in this case is:

$\sigma = \sqrt{\{\ [(\partial R_g/\partial R_1)\sigma_{R1}]^2 + [(\partial R_g/\partial R_2)\sigma_{R2}]^2 + [(\partial R_g/\partial N_V)\sigma_{Nv}]^2\ \}}$

The derivatives all share a common denominator D^2, and so on writing the expression in full, a factor $(1/D^2)^2$ can be removed from the square root bracket. Hence:

$$\boxed{\sigma = [1/(R_2 - R_1\,N_V)^2]\ \sqrt{\{\ [R_2^2\ (N_V\text{-}1)\ \sigma_{R1}\ ]^2 + [R_1^2\ N_V\,(1\text{-}N_V\ )\ \sigma_{R2}\ ]^2 + [R_1\,R_2\,(R_2\text{-}R_1)\ \sigma_{Nv}\ ]^2\ \}}}$$

In the previous section, we determined $R_g = 23.3\ \Omega$ from the following measurements:

$R_1 = 29.6 \pm 0.34\ \Omega$, $R_2 = 75.1 \pm 0.7\ \Omega$, $N_V = 1.364 \pm 0.04$

These give:

$D^2 = (R_2 - R_1N_V)^2 = 1205.8673$

$\partial R_g/\partial R_1 = R_2^2\ (N_V - 1) / D^2 = 2052.9636 / 1205.8673 = 1.7025$

$\partial R_g/\partial R_2 = R_1^2\, N_V(1 - N_V\,) / D^2 = -435.0100 / 1205.8673 = -0.3607$

$\partial R_g/\partial N_V = R_1\, R_2\, (R_2 - R_1) / D^2 = 101144.68 / 1205.8673 = 83.8771$

$\sigma = \surd\{\ [(\partial R_g/\partial R_1) \times \sigma_{R1}]^2 + [(\partial R_g/\partial R_2) \times \sigma_{R2}]^2 + [(\partial R_g/\partial N_V) \times \sigma_{Nv}]^2\ \}$

$= \surd\{\ [1.7025 \times 0.34]^2 + [0.3607 \times 0.7]^2 + [(83.8771 \times 0.04]^2\ \}$

$= \surd\{\ 0.5789^2 + 0.2525^2 + 3.3551^2\ \}\ \ \Omega$

Note that the error contributions in the expression above are very close to the averages of the deviations calculated by the incremental (spreadsheet) method used previously. Finally we have:

$\sigma = 3.414\ \ \Omega$

$R_g = 23.3 \pm 3.4\ \Omega$

A spreadsheet version of this calculation (which can be used as a template) is given on sheet 2 of the accompanying file: **Rg_meas.ods** .

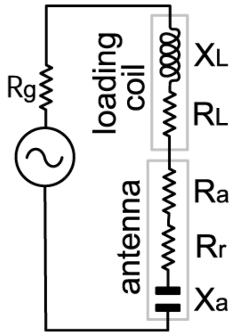
40. Antenna system Q

In previous sections we showed that conjugate matching is not necessarily a good idea, and that radio transmitter manufacturers do not necessarily design power amplifiers to work into a conjugate load. Besides the obvious advantages in terms of output regulation and efficiency however; there is a further reason to load a radio transmitter lightly in cases where the input impedance of an antenna system changes rapidly with frequency or is subject to variable environmental factors. Recall from section **34** that the true maximum power transfer condition occurs when:

$R=\sqrt{[R_g^2+(X+X_g)^2]}$

Consequently, if the antenna system is subject to disturbances that can cause a reactive component to appear *after* matching has been carried out, then the best average maximum power output will be obtained when the load resistance is somewhat higher than the source resistance. Possible causes of transient residual reactance are many and various, including: changing physical environment (of mobile and portable transmitters), wind, rain, component heating, birds, etc., and (as is easily forgotten) *modulation*.

In section **9**, we discussed an electrically-short inductively-loaded vertical antenna system. The equivalent circuit of this antenna is shown on the right; with one extra resistance added, that being the (true) source resistance R_g. With this additional piece of information, it becomes possible to calculate the Q, and hence the bandwidth, of this system.



If we take the same example component values as were used before, we have:

$\mathbf{Z}_L = R_L + \mathbf{j}X_L = 7.5 + \mathbf{j}3000$

$\mathbf{Z}_a = R_a + R_r + \mathbf{j}X_a = 2.5 - \mathbf{j}3000$

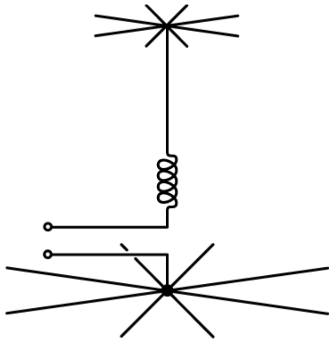
giving an input resistance of 10 Ω for a whip length of about 0.07 λ. The whole system is of course, a series resonator, and we can define the circuit Q as:

$Q_0 = |X_L|/R_{total} = |X_a|/R_{total}$

Since bandwidth is proportional to f_0, problems of excessive Q are likely to occur at low operating frequencies, so let us see what happens if this antenna is built to operate on (say) 1.9 MHz, with the generator source resistance adjusted to be 5 Ω. This will make the total series resistance 15 Ω, and with X_L=3000 Ω, the Q will be 3000/15=200. The -3 dB bandwidth of the antenna will therefore be f_0/Q=9.5 kHz. This is wide enough to accommodate a communications SSB signal (2.7 kHz bandwidth) but there is very little latitude for incidental detuning, and the antenna will exhibit a small variation of input impedance depending on the modulation frequency. Light loading of the generator will help to offset these problems, because it will create a situation where transient detuning forces the generator-load system *closer* to its maximum power transfer point (although detuning won't increase the amount of power delivered, light loading will give better regulation than a conjugate match).

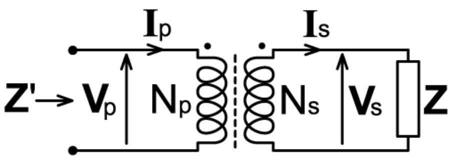
Note incidentally, that the antenna discussed above is not physically small when designed for operation in the 160 m band. The wavelength at 1.9 MHz is c/f =157.8m; and so a 0.07 λ rod will be 11 m long, and consequently far too large for mobile use. To make a mobile antenna, the rod must be shortened; and this will reduce the antenna capacitance (and hence increase the reactance), and sadly for efficiency, will cause the radiation resistance to fall. The larger antenna reactance will necessitate a larger loading reactance, and although this will bring more resistance with it, the increase in reactance will be greater than the increase in total resistance and the Q will rise. A point can be reached where serious curtailment of the modulation bandwidth occurs, although, for this system, it it not predictable using lumped-component theory. The coil can be regarded as a lumped component provided that the whip is long enough to ensure that most of the radiation occurs from the whip rather than from the coil. If that condition applies, then the Q of the antenna system can never be larger than the Q of the loading coil because the total series resistance will always be that of the coil plus a little extra. The maximum tolerable Q (causing some, but not serious, audio degradation) occurs when the antenna system bandwidth is the same as the audio bandwidth, and for SSB on 1.9 MHz this figure is 1900/2.7=704. It is extremely difficult to make a lumped inductor with a Q of greater than about 400, so on 160m the Q limit can be avoided by controlling the length of the coil. If, on the other hand, the whip is of length comparable to or shorter than the coil, then most of the radiation occurs from the coil. In that case, the lumped component description fails completely, and the system is best described as a quarter-wave transmission-line resonator. In the transmission-line regime, the Q and hence the voltage magnification can become enormous, and the usable input power is limited by the tendency for the air around the top of the coil to ionise and become electrically conductive. Coils operated at or slightly below the quarter-wave transmission line resonance frequency are used for artificial lightning experiments, in which context they are known as 'Tesla coils'. In particular, the voltage-magnifier coil connected in series with the output from a step-up transformer is known as the 'Extra Coil'. The transmission-line properties of coils will be discussed in a separate article.

While on the subject of MF and HF mobile antennas; when forced to use a very short whip, it is possible to increase the antenna capacitance artificially (and hence reduce the reactance) by adding a capacitance hat to the antenna (some prongs sticking-out sideways symmetrically; or, if there's a risk that you might poke someone's eye out, an aluminium disk). Reducing the antenna reactance in this way reduces the amount of loading inductance required, and hence allows the coil to be wound with thicker wire for a given size (less resistance). Placing the hat at the top of the antenna moreover, increases the current in the vertical section, and actually increases the radiation resistance slightly (every little helps).



41. Basic impedance transformer

In the previous section it was implied that the generator output impedance could be adjusted, but we have yet to offer any method for doing so. There are numerous options in this respect, but for the present purpose it will be sufficient to have just one: the *transformer*. Transformers are discussed in detail in a separate article[25]; but here we will avail ourselves of the properties of straightforward (but unfortunately mythical) circuit models known as the *perfect* transformer and the *ideal* transformer. An ideal transformer has no losses, and its voltage ratio is the same as its turns ratio. In truth, well-designed transformers can have power-transfer efficiencies of more than 98% within a certain band of frequencies, and so the myth of the ideal transformer is not so far from reality. A perfect transformer is an ideal transformer that has large winding reactances, so that the off-load input current is negligible; which means that its current ratio is the inverse of its voltage ratio. Here we will assume that a tightly-coupled transformer is ideal when operating within its pass-band, on the understanding that it requires a more advanced analysis to determine what the pass-band is. Such a transformer is also approximately perfect when used as part of a low-impedance electrical network.



A transformer loaded with an impedance **Z** is represented on the right. Here N_P is the number of turns in the primary (generator side) winding, and N_S is the number of turns in the secondary (load side) winding. The dots next to the windings indicate either the start or the finish (it doesn't matter how this is designated, as long as it is done consistently), and it it assumed that both coils are wound in the same sense (clockwise or anticlockwise when looking at a particular end of the coil). The dotted line between the coils indicates that the transformer is wound on a magnetic core, the purpose of which (in this instance) is to produce a very tight magnetic coupling between the windings. If all of the magnetic field from the primary winding is captured by the core and linked to the secondary winding (i.e., if there is no magnetic leakage), and if the coils and the core have no heating losses, then all of the power delivered by the generator is transferred to the load. Also, if the inductive reactance of the windings is much larger than the magnitudes of the impedances seen on either side, and the capacitance of both windings is very small, then the secondary voltage will be in phase with the primary voltage, and the secondary current will be in anti-phase with the primary current (i.e., as a current appears to flow into the primary, a current appears to flow out of the secondary). If the number of turns in the secondary winding is greater than the number of turns in the primary, then $\mathbf{V}_S$ will be larger than $\mathbf{V}_P$, and vice versa; and the voltage transformation will be in proportion to the turns ratio, i.e.,

$\mathbf{V}_S = \mathbf{V}_P N_S / N_P$ . . . . (**41.1**)

It follows that if the power produced by the generator is transferred to the load without loss, then the **VI** product will be conserved; which means that if the voltage is stepped up, then the current will be stepped down to keep **VI** very nearly constant (and vice versa). This implies that the transformer performs on the current the inverse of the transformation it performs on the voltage, i.e. (interpreting the currents in the sense of the arrows in the diagram above):

$\mathbf{I}_S = \mathbf{I}_P N_P / N_S$ . . . . (**41.2**)

Now, by definition, the impedance looking into the transformer primary is:

$\mathbf{Z}' = \mathbf{V}_P / \mathbf{I}_P$

[25] Electromagnetic Induction. D W Knight. [available from www.g3ynh.info/]

which gives, using (41.1) and (41.2) as substitutions:

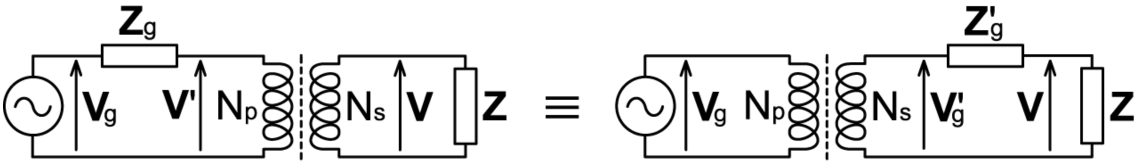
$\mathbf{Z'} = \mathbf{V}_S\ (N_P/N_S)\ /\ [\ \mathbf{I}_S\ (N_S/N_P)\ ]$

and since $\mathbf{Z} = \mathbf{V}_S/\mathbf{I}_S$:

$\mathbf{Z'} = \mathbf{Z}\ (N_P/N_S)^2$	**41.3**

Thus, to a reasonably good approximation; a tightly-coupled transformer having relatively large winding reactances scales an impedance according to the *square of the turns ratio*.

Now let us consider the problem in reverse, and see what a transformer does to the output impedance of a generator. Here we will call the apparent source impedance as seen from the secondary side of the transformer $\mathbf{Z}_g'$, with $\mathbf{Z}_g$ as the *actual* generator output impedance. The relationship between $\mathbf{Z}_g'$ and $\mathbf{Z}_g$ is perhaps guessable; but to derive it mathematically requires a trick, which is that of defining an equivalent circuit with all of the source resistance moved to the secondary side of the transformer. A suitable approach to the derivation is then to write expressions for the voltage **V** across the load using both the original and the equivalent circuits and then equate the two expressions.



For the left-hand circuit, let us define **Z'** as the load impedance seen by the generator, its relationship to to the load **Z** being given by equation (41.3) above:

$\mathbf{Z'} = \mathbf{Z}\ (N_P/N_S)^2$

The voltage **V'** is then the output of a potential divider formed by $\mathbf{Z}_g$ in series with **Z'**, i.e.:

$\mathbf{V'} = \mathbf{V}_g\ \mathbf{Z'}\ /\ (\mathbf{Z}_g + \mathbf{Z'})$

and **V'** is related to the load voltage **V** by the turns ratio, ie.:

$\mathbf{V} = \mathbf{V'}\ N_S\ /\ N_P$

Hence:

$\mathbf{V} = (N_S/N_P)\ \mathbf{V}_g\ \mathbf{Z'}\ /\ (\mathbf{Z}_g + \mathbf{Z'})$

$\mathbf{V} = (N_S/N_P)\ \mathbf{V}_g\ [1 + (\mathbf{Z'}\ /\ \mathbf{Z}_g)\ ]$ . . . . (41.4)

For the right hand circuit, **V** is the output of a potential divider formed by $\mathbf{Z}_g'$ and **Z** :

$\mathbf{V} = \mathbf{V}_g'\ \mathbf{Z}\ /\ (\mathbf{Z}_g' + \mathbf{Z})$

$\mathbf{V} = \mathbf{V}_g'\ [1 + (\mathbf{Z}\ /\ \mathbf{Z}_g')\ ]$

where:

$\mathbf{V}_\mathbf{g}' = (N_S/N_P)\ \mathbf{V}_\mathbf{g}$

Hence:

$\mathbf{V} = (N_S/N_P)\ \mathbf{V}_\mathbf{g}\ [1 + (\mathbf{Z} / \mathbf{Z}_\mathbf{g}')\ ]$

Equating this to expression (41.4) gives:

$1 + (\mathbf{Z}' / \mathbf{Z}_\mathbf{g}) = 1 + (\mathbf{Z} / \mathbf{Z}_\mathbf{g}')$

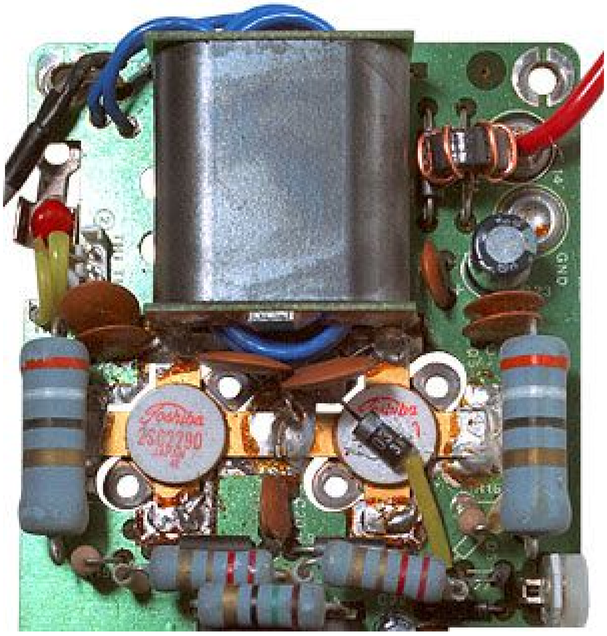
i.e.,

$\mathbf{Z}_\mathbf{g}' = \mathbf{Z}_\mathbf{g}\ \mathbf{Z} / \mathbf{Z}'$

but, by rearrangement of equation (41.4), $\mathbf{Z} / \mathbf{Z}' = (N_S/N_P)^2$, hence:

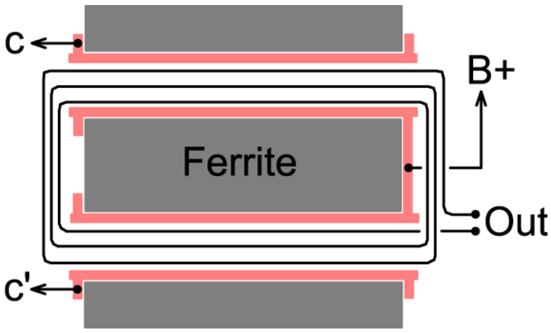
$$\mathbf{Z}_\mathbf{g}' = \mathbf{Z}_\mathbf{g}\ (N_S/N_P)^2 \qquad \mathbf{41.5}$$

Thus a tightly-coupled output transformer scales the source impedance according to the square of the turns ratio, a generator with a low output impedance being converted into a generator with a high output impedance by means of a step-up ($N_S>N_P$) transformer and vice versa.



The broadband output transformer of a fairly typical 100W short-wave radio transmitter (Kenwood TS430s) is shown on the right. The transformer core is a block of ferrite with two hollow channels passing through it (known colloquially as a "pig nose"). The primary winding consists of two short lengths of copper or brass tubing passing through the core and connected together at one end by a strip of copper-laminate board. The secondary winding is a length of PTFE-coated multi-strand silver-plated copper wire threaded through the copper tubes (the reason for the choice of materials is explained in another article[26]). To make a complete turn around the core, a conductor must pass through one hole and back through the other. As shown below diagrammatically, the copper tubes form a centre-tapped single turn, with the DC power supply (B+) connected to the centre tap, and the other ends connected to the collectors of the RF power transistors (a matched pair of 2SC2290s). The transformer in the photograph has four turns and so increases the amplifier output impedance by a factor of 16.

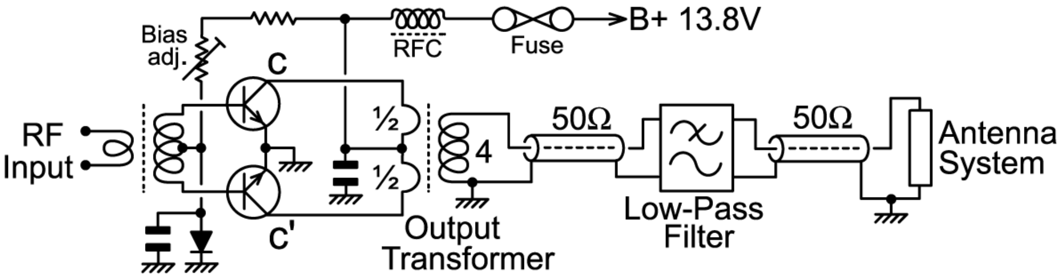
[26] Components and Materials. [www.g3ynh.info/]



There is something more to the use of an output transformer than impedance transformation however, the principal issue being that the transmitter discussed above uses a 13.8V power supply and yet must deliver 100W into a 50Ω load. The required output power and target load impedance defines the output voltage as

$V=\sqrt{(PR)}=\sqrt{(100\times50)}=70.7V$ RMS

i.e., $70.7\times2\sqrt{2}=200V$ peak-to-peak (p-p). A simplified version of the power amplifier circuit is shown below, and we can deduce the minimum allowable transformer step-up ratio by examining it.



This is a so-called *push-pull* amplifier circuit, in which one transistor provides the positive half-cycle of the output waveform, and the other transistor provides the negative half-cycle. When a bipolar transistor is turned hard on, its collector voltage does not go to zero, but stops at some *saturation voltage*, which is usually around 1V. Also, it is not a good idea to drive the transistors close to saturation because this will lead to considerable distortion of the output waveform. Therefore we must assume that the output stage can produce positive and negative half cycles of no more than about 12.5V across half of the primary winding, i.e., 25V per transistor across the whole winding , hence 50V p-p. To obtain 200V p-p (70.7V RMS) therefore, a voltage step-up ratio of 1:4 is required. The fact that this transformation increases the source impedance by a factor of 16 is a secondary consideration; and is of no great concern unless the source impedance begins to approach the design load resistance, the latter situation being associated with low transfer efficiency and poor load regulation as discussed earlier. It follows, that to keep the output impedance as low as possible, a step-up ratio just sufficient to provide the required output voltage is optimal. The actual output impedance (R_g') of the TS430s transmitter (measured at the antenna socket, see the example at the end of section **38**) is about 23Ω (measured 23.3±3.4Ω at 1.9MHz) for a design load resistance of 50Ω. The output impedance of the power amplifier (R_g) is therefore approximately 23/16=1.4Ω. Dye and Granberg[27] give an approximate formula for calculating the output impedance of a transistor power amplifier below 100MHz as:

$R_g = (V_{cc} - V_{sat})^2 / P_{out(max)}$

where V_{cc} is the supply voltage, and $P_{out(max)}$ is the maximum power available from the amplifier. If we assume a saturation voltage V_{sat} of about 1V, this gives:

$R_g = 12.8^2 / 100 = 1.64Ω$.

[27] **Radio Frequency Transistors**, Norm Dye and Helge Granberg. Motorola inc. / Butterworth Heinemann, Newton MA. 1993. ISBN 0-7506-9059-3. Output impedance of a power amplifier: p118.

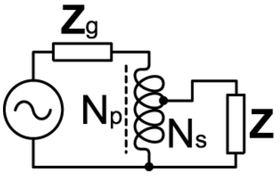
Multiplying this by 16 gives R_g'=26.2Ω, which is within 1σ of the measurement without taking any of the circuitry between the power amplifier and the antenna socket into account.

Notice incidentally, that the power amplifier is shown as feeding into a low-pass filter (LPF) before connection to the antenna system. Such a filter is always necessary with a broad-band transistor power amplifier, because such amplifiers produce relatively high levels of harmonics. The push-pull configuration actually cancels even harmonics, but there are still high levels of odd harmonics (3rd, 5th, 7th etc.) that must be removed (in engineering, the first harmonic is the same as the fundamental). In section **34** we noted that the power amplifier protection circuitry operates when the load impedance is too high, as well as when it is too low. This is not usually necessary for the protection of the amplifier, but the LPF may not provide the required degree of harmonic attenuation when incorrectly terminated; and so the protection circuitry helps to keep spurious emissions within acceptable limits if the load impedance is too high.

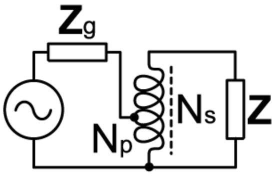
42. Auto transformers

'Auto-transformer' (self-transformer) is just another name for a tapped inductor. The transformation rules for tightly-coupled auto-transformers are identical to those for tightly-coupled transformers with separate windings. The significant functional difference between the two types of transformer is that an auto-transformer does not provide DC isolation between source and load. A more subtle difference is that a transformer with separate windings, by judicious use of electrostatic shielding, can be made in such a way that the coupling between the primary and secondary windings is almost entirely magnetic. An auto-transformer will always exhibit some stray capacitive and resistive (potential divider) coupling, and so if its inductance is part of a filter circuit, the filter may exhibit poor attenuation of signals outside its passband.

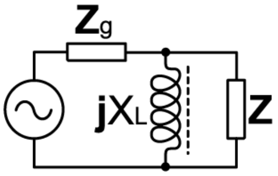
The step-up and step-down auto-transformer configurations are shown below. Also shown is the somewhat redundant 1:1 auto-transformer, otherwise known as an inductor; the point in including it being to draw attention to the inductive reactance that every transformer places in parallel with its load.



Step-down



Step-up



1:1

It should be obvious by inspection of the '1:1' auto-transformer circuit, that the voltage - current relationship for the load seen by the generator is given by:

$$\mathbf{V}/\mathbf{I} = \mathbf{j}X_L \,//\, \mathbf{Z}$$

If the coil has losses moreover, we can represent these as a resistance (R_L say) in series with the coil:

$$\mathbf{V}/\mathbf{I} = (\, R_L + \mathbf{j}X_L \,) \,//\, \mathbf{Z}$$

We can also transform the impedance of the coil into its parallel form (see section **19b**), in which case the load on the generator becomes:

$$\mathbf{V}/\mathbf{I} = R_{Lp} \,//\, \mathbf{j}X_{Lp} \,//\, \mathbf{Z}$$

The implication is that, unless the magnitude if the inductive reactance is very much larger than the magnitude of the load impedance, the transformer will not preserve the load phase relationship. If the load is reactive, the parallel loss component will also alter the load phase relationship slightly.

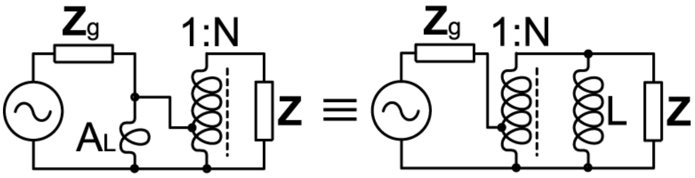
In the previous section, we introduced the idea that an impedance located on one side of a transformer can be transferred to the other side in an equivalent circuit by the act of multiplying it by the turns-ratio squared. So we might represent the inductance of a transformer as a separate inductance L in parallel with the primary side of an ideal transformer (of otherwise infinite inductance), or we might represent it as an inductance L' in parallel with the secondary side. The transformation rule (**33.3**) tells us that:

$$\mathrm{j}2\pi f\, L = \mathrm{j}(N_S/N_P)^2\, 2\pi f\, L'$$

i.e.,

$L = (N_S/N_P)^2 L'$

This is a remarkable result because, not only does it give us the basis for constructing equivalent circuits to serve as models for real transformers, it also tells us something about inductors. The expression can only be true if the inductance of the coil is proportional to the square of the number of turns in it. We can see why by considering the two 1:N auto-transformer equivalent circuits shown below:



In the left-hand circuit, the inductance of the transformer is referred to the primary side, and for reasons of convention is given the symbol A_L. In the right-hand circuit, the inductance is referred to the secondary side and is given the symbol L. From the foregoing discussion, we can immediately write the relationship between L and A_L:

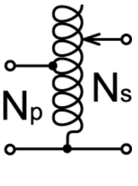
$$\boxed{L = N^2 A_L}$$

We can also interpret L as the inductance of the whole coil, and A_L as the inductance of one turn of the coil.

A_L is known as the *inductance factor*, and depends on the physical dimensions of the coil and the nature of any magnetic core material. It may be interpreted either as the inductance of a one-turn coil, or as the inductance of an auto-transformer referred across a one-turn tap. A_L has the units of inductance (Henrys), but is more informatively given units of inductance / turn² ("Henrys per turn-squared").

Continuously variable auto-transformer:

One of the drawbacks of ferrite or iron-cored transformers as impedance matching devices is that the the transformation ratio can only be altered in a stepwise fashion, by changing windings or tappings one turn at a time (or half a turn if the core has two holes). If the turns in the coils are few, as tends to be the case in radio-frequency applications, then the steps available can be very coarse indeed. It is however possible to make a continuously variable inductor or auto-transformer by rotating a coil about its axis and tapping into it with a rolling contact, the coil end-connections being made by slipping contacts (known, for historical reasons, as "brushes"). Such a device is known colloquially as a "roller coaster", and an example is shown in the photograph on the right.



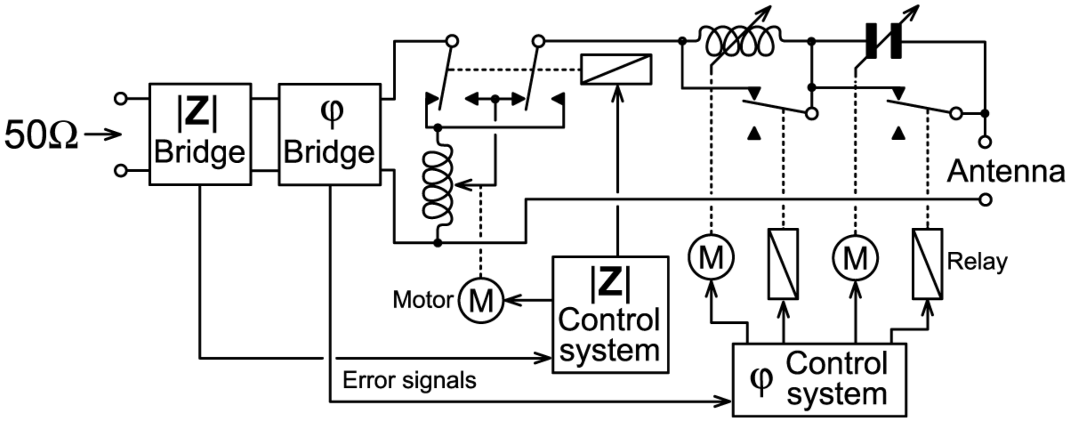
This is the motor-driven variable impedance transformer from a 1957 vintage Collins 180L-3A automatic HF antenna tuner. The tuner is designed to match end-fed wire (Marconi) antennas of 14 to 40 metres in length over a frequency range of 2 MHZ to 25 MHz, and is for use with transmitters with an output of up to 150 W and a preferred load impedance of 52 Ω. An interesting feature of the transformer is that it achieves a continuous transition from step-down to step-up by having an overwind (see diagram right), i.e., the brush contact at one end of the coil goes to a centre-tap, and the end of the coil is left unconnected. The coil has 28 turns, and the input tap is at 14 turns, so a maximum impedance step-up of approximately 4:1 is obtainable.

The disadvantage of the Collins transformer is that the coil does not have a magnetic core. The stray magnetic fields will therefore induce currents (eddy currents) in the surrounding metalwork and give rise to resistive losses. The open magnetic circuit also implies that the impedance transformation obtained will not be exactly proportional to the square of the turns ratio, and due to the absence of a magnetic core the inductance might appear on first consideration to be rather low. The inductance for the whole coil, estimated using Wheeler's long-coil formula[28] is about 20 μH, giving only about 5 μH when referred to the primary side. This will give rise to significant phase shift at lower frequencies, the inductive reactance seen by the transmitter at 2 MHz being something around $2\pi\times2\times10^6\times5\times10^{-6}$=63 Ω. It transpires however, that the choice of primary reactance about equal to the target input impedance at the lowest operating frequency is sensible, because in addition to the impedance transformer, the antenna tuner also has power-factor correction components. In the process of adjusting a reactance in series with the antenna to achieve a resistive input impedance, any phase shift due to the transformer is automatically taken into account. Consequently, it is possible to keep the inductance small, which helps in the avoidance of self-resonance problems at the high-end of the operating frequency range.

[28] Solenoids. D W Knight [www.g3ynh.info/]

43. Prototype Z-matching network

An antenna matching system based loosely on the Collins 180L-3 is depicted in the diagram below. The only major difference is that the transition from step-down to step-up is accomplished by means of a change-over relay. This increases the transformation range in comparison to the overwind method, but also increases the complexity of the control system.



This is the prototype of all antenna tuners in the sense that it approaches the impedance matching problem in the simplest possible way. The object of the exercise in every case is to transform the impedance in its two dimensions: *magnitude* and *phase*, and the most direct approach is to do so using one device that only affects the magnitude and one device that only affects the phase. The magnitude-correcting engine is the variable auto-transformer, and the phase-correcting engine is a series reactance; relays being provided to insert a series coil in the event that the antenna is capacitive, or a series capacitor in the event that the antenna is inductive.

Such a matching unit can, of course, be controlled manually, by the expedient of providing it with control knobs and switches instead of motors and relays. This approach replaces the automatic control system with a human being, but makes no allowance for the fact that humans in general have little aptitude for the task. Here we monitor the load magnitude and phase using bridge circuits, which are the subject of a separate article. The bridges produce error signals, which tell their respective control systems which way to go in event that the error exceeds a certain preset threshold. Not shown on the diagram, but necessary to make the system work, are limit switches, two for each variable device. These tell the control system when a motorised device has hit one of its end-stops: so that the change-over relay can be switched and the motor direction reversed in the case of the impedance transformer; so that the switch-over from coil to capacitor or vice versa can be made in the case of the series reactance network; and as protection against motor burn-out in the event that the load is outside the matching range. The control systems for magnitude and phase are shown as being completely separate; which they are except in respect of common signals, such as the request for a tuning carrier or an instruction to reduce power, which they might send to the transmitter on detecting a matching error. The independence of the two systems is possible because the two chosen matching criteria are independent, i.e., the two matching processes can proceed simultaneously without altering the outcome. The system can even adjust itself when presented with a speech SSB signal, but will reach a solution fastest when the error signals are continuously available. One desirable property of this matching system, and of any properly designed matching system, is that it corrects for the defects of its own components. In this case, when the phase control system adds series inductance (for example), the increasing resistance of the coil will increase the impedance magnitude seen at the input, but the magnitude control system will simply back-off to compensate. Similarly, the inductance of the impedance transformer will cause a positive phase shift, but phase control system will back-off in the capacitive direction to compensate.

While the simple magnitude-phase matching system is entirely practical however, it has never been particularly popular. The reason is that it is difficult to design an efficient and resonance-free variable broadband transformer. The required transformations can just as well be obtained using only variable capacitors and inductors, and this subject is examined in detail in a separate article[29].

[29] Impedance matching. D W Knight. [www.g3ynh.info/]

44. Admittance, conductance, susceptance

The linear circuit analysis technique demonstrated so far consists of breaking the circuit down into two-terminal networks and treating those networks as impedances. This approach has allowed us to attack a wide range of problems; but it results in extremely messy algebra when impedances in parallel are involved. Ultimately, we need a way of dealing with arbitrarily large numbers of impedances in parallel, just as we can already deal with any number of impedances in series; and it transpires that this can be achieved by defining the properties of our component two-terminal networks not in terms of impedance, but in terms of the reciprocal of impedance, this being called ***admittance***. By so doing, we move the problem out of what we so-far think of as its natural space (impedance space) and into what is known as its *reciprocal space*; and the re-definition, trivial though it is in the case of phasors, is known as a *reciprocal-space transformation*.

The reciprocal space transformation is another mathematical invention of James Clerk Maxwell. Its most far-reaching application is in the field of X-ray crystallography, it being the means by which the X-ray diffraction patterns of crystals are traced back to the internal arrangement of atoms. Here however, we need only a simplified version, because the problems we wish to solve are strictly two-dimensional.

The reciprocal of impedance space is known as ***admittance space***. A pair of two-dimensional reciprocal spaces has the property that straight lines in one appear as circles in the other (a correspondence that is used extensively in the article "Impedance Matching", cited earlier); but the real power of the transformation lies in the fact that phasor problems requiring the double-slash product in one space, become problems of *addition* in the other.

Converting an impedance into an admittance is simply a matter of taking the reciprocal. Admittance is usually given the symbol **Y** (and here we put it in bold because it is complex), hence:

$$\mathbf{Y}=1/\mathbf{Z}.$$

Now, if $\mathbf{Z}$=R+$\mathbf{j}$X this gives:

$$\mathbf{Y} = 1/(\mathrm{R} +\mathbf{j}\mathrm{X})$$

which can be put into the a+**j**b form by multiplying the numerator and denominator by the complex conjugate of the denominator, i.e.:

$$\mathbf{Y} = \frac{\mathrm{R} - \mathbf{j}\mathrm{X}}{(\mathrm{R} + \mathbf{j}\mathrm{X})\ (\mathrm{R} - \mathbf{j}\mathrm{X})}$$

hence:

$$\mathbf{Y} = \frac{\mathrm{R} - \mathbf{j}\mathrm{X}}{(\mathrm{R}^2 + \mathrm{X}^2)}$$

This expression can be written:

$$\boxed{\mathbf{Y} = \mathrm{G} +\mathbf{j}\mathrm{B}}$$

where the real part of the admittance, G, is called the ***conductance***, and the imaginary part, B, is called the ***susceptance*** (of the network under consideration). From the above, we can extract definitions for conductance and susceptance which are:

Conductance, $G = R / (R^2 + X^2)$

Susceptance, $B = -X / (R^2 + X^2)$

Now observe that when the impedance of a network is purely resistive, the conductance is 1/R, and G=1/R is the definition of conductance in DC electrical theory. When an impedance is purely reactive, the susceptance B= -1/X (susceptance has no DC counterpart).

Admittance, conductance, and susceptance, of course, have units; and the modern unit in this case is the ***Siemens***, which is given the dimension symbol capital S (as opposed to the second, which has a small s). In old textbooks and papers, the unit of admittance is often given as the 'Mho' (Ohm spelt backwards), but in either case, the actual dimensions are in reciprocal Ohms, i.e., /Ω or Ω^{-1}. A pure resistance of 50 Ω therefore corresponds to a conductance of 1/50 Siemens, i.e., 20 milli-Siemens or 20 mS. A pure reactance X=100 Ω corresponds to a susceptance B=10 mS, and so on. "Siemens", incidentally, is a name, like "Jones". The singular of Jones is not Jone, and so Siemens keeps its final s in both singular and plural forms (one Siemens, several Siemens). The plural "Siemenses" is not recommended (but is a lot less embarrassing than the quasi-singular "Siemen").

The double-slash product was previously defined as (**17.5**):

$\mathbf{a} // \mathbf{b} = \mathbf{ab}/(\mathbf{a}+\mathbf{b})$

We can demonstrate that addition is the reciprocal-space counterpart of the double slash operator by transforming the parallel impedance formula; i.e., if:

$\mathbf{Z} = \mathbf{Z}_1 \mathbf{Z}_2 / (\mathbf{Z}_1 + \mathbf{Z}_2)$

then

$\mathbf{Y} = (\mathbf{Z}_1 + \mathbf{Z}_2) / \mathbf{Z}_1 \mathbf{Z}_2$

If we let $\mathbf{Y}_1 = 1/\mathbf{Z}_1$ and $\mathbf{Y}_2 = 1/\mathbf{Z}_2$, then

$\mathbf{Y} = \mathbf{Y}_1 \mathbf{Y}_2 [(1/\mathbf{Y}_1) + (1/\mathbf{Y}_2)]$

Which rearranges to:

$$\boxed{\mathbf{Y} = \mathbf{Y}_1 + \mathbf{Y}_2}$$

i.e., *when two networks are placed in parallel, their admittances are added.*

Recall that the formula for resistances in parallel, $R=R_1R_2/(R_1+R_2)$ is a rearrangement of the expression:

$1/R = (1/R_1) + (1/R_2)$

It should now be apparent, that what the formula really says is:

$G = G_1 + G_2$

The formula for impedances in parallel is of course a rearrangement of:

$1/\mathbf{Z} = (1/\mathbf{Z}_1) + (1/\mathbf{Z}_2)$

and this expression can be extended to cover any number of impedances in parallel by adding more terms, i.e.:

$1/\mathbf{Z} = (1/\mathbf{Z}_1) + (1/\mathbf{Z}_2) + (1/\mathbf{Z}_3) + \ldots\ldots + (1/\mathbf{Z}_n)$

This is a sum of admittances, and may be re-written as:

$\mathbf{Y} = \mathbf{Y}_1 + \mathbf{Y}_2 + \mathbf{Y}_3 + \ldots\ldots + \mathbf{Y}_n$

We can express this result using the double slash notation:

$1/(\mathbf{Z}_1 \,//\, \mathbf{Z}_2 \,//\, \mathbf{Z}_3 \,//\, \ldots \,//\, \mathbf{Z}_n) = \mathbf{Y}_1 + \mathbf{Y}_2 + \mathbf{Y}_3 + \ldots\ldots + \mathbf{Y}_n$

where $\mathbf{Y_k} = 1/\mathbf{Z_k}$ (k being any subscript).

The admittance representation of an electrical circuit is no less authoritative than the impedance representation, and is no more difficult to use. Admittances are phasors, and all of the phasor techniques we have developed in this chapter will work on them. It is however, helpful to remember that a (numerically) large admittance corresponds to a small impedance and vice versa.

Reciprocal-space counterparts

Impedance space	Admittance space
Impedance $\mathbf{Z} = R+\mathbf{j}X = 1/\mathbf{Y}$	Admittance $\mathbf{Y} = G+\mathbf{j}B = 1/\mathbf{Z}$
Resistance $R = G/(G^2+B^2)$	Conductance $G = R/(R^2+X^2)$
Reactance $X = -B/(G^2+B^2)$	Susceptance $B = -X/(R^2+X^2)$
Pure resistance $R = 1/G$	Pure conductance $G = 1/R$
Pure reactance $X = -1/B$	Pure susceptance $B = -1/X$
Inductive reactance $X_L = 2\pi fL$	Inductive susceptance $B_L = -1/(2\pi fL)$
Capacitive reactance $X_C = -1/(2\pi fC)$	Capacitive susceptance $B_C = 2\pi fC$
// operator	+ operator
+ operator	// operator
Straight line	Circle
Circle	Straight line

One further issue of which the reader will need to be aware is that two different definitions of admittance appear in the electrical and electronic literature. Some authors (e.g., Hartshorn[30]) use **Y**=G+**j**B, as is done here; and others (e.g., Langford-Smith[31]) use **Y**=G-**j**B. The 'alternative' definition gives $B=X/(R^2+X^2)$, and thus $B_L=1/(2\pi fL)$, and $B_C= -2\pi fC$.

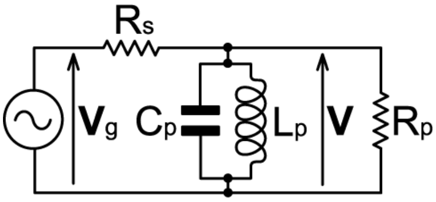
In the next section, we analyse the parallel resonator bandpass filter and determine the relationship between resonant frequency, bandwidth and Q. In the author's first attempt at the derivation, the definition **Y**=G-**j**B was used, and the formula that resulted had it that either Q is negative or f_0 is negative. The change to **Y**=G+**j**B fixed the problem and so, since Q is by definition positive when loss resistance is positive, the other definition is wrong according to the convention that frequency is positive. We may also note a reflection symmetry in the correct choice, in that we have $X_L=2\pi fL$ in impedance space, and $B_C=2\pi fC$ in admittance space, etc.; i.e., inductive reactance and capacitive susceptance are positive, capacitive reactance and inductive susceptance are negative.

[30] Radio-Frequency Measurements by Bridge and Resonance Methods, L. Hartshorn [Principal Scientific Officer, British National Physical Laboratory], Chapman & Hall, 1940 (Vol. X of "Monographs on Electrical Engineering", ed. H P Young). 3rd imp. 1942. Ch. I, section 3: Defines Admittance as Y=G+jB, hence BL=-1/ωL and BC=ωC.

[31] **Radio Designer's Handbook**, Ed. Fritz Langford-Smith. 4th edition. 4th impression (with addenda), Iliffe Publ. 1957 [A later reprint exists (1967) ISBN 0 7506 36351].
Section 4.6(v), p153: Defines inductive susceptance as positive and capacitive susceptance as negative, hence **Y**=G-**j**B. This is contrary to the more convincing derivation given by Hartshorn (above).

45. Parallel resonator BPF

The reader may have noticed that, having determined the relationship between bandwidth and Q for a series resonator, we did not immediately do the same for a parallel resonator, but instead digressed into the subjects of source impedance and impedance transformation. There was a very good reason for doing so, as we will soon see, which is that there is no satisfactory design procedure for parallel-resonant bandpass filters if the source and load impedances cannot be controlled. This situation prevails because, in order to use the resonator as a filter, we need methods for injecting energy into it and extracting energy from it, and the impedances presented by these input and output networks affect the Q.



The prototype band-pass filter is shown on the right. The generator and load coupling scheme used is not the only one possible, but all other schemes are equivalent to this one after suitable transformation. Here we inject energy via a source resistance R_S, which is the sum of the generator output resistance and any additional resistance placed in series with it. R_P is the parallel combination of the resonator dynamic resistance and any load resistance that might be placed across it. Notice that we have provided the model with source and load *resistances* rather than impedances. We are at liberty to do so without affecting the generality of the analysis, because any reactive components in the source and load impedances will turn out to be effectively in parallel with the resonator. This means that these additional reactances will modify the effective values of X_{Cp} and X_{Lp} (i.e., they will change the resonant frequency), but they will not affect the general circuit behaviour provided that they do not exhibit any self-resonances in the analysis frequency range.

If we define $\mathbf{V}_0$ as the output voltage at resonance, then the bandwidth function is $|\mathbf{V}/\mathbf{V}_0|$ plotted against frequency. We can write expressions for $\mathbf{V}$ and $\mathbf{V}_0$ by treating the circuit as a potential divider, thus, noting that $X_{Cp} // X_{Lp} \rightarrow \infty$ at resonance (i.e., when $f = f_0$) we get:

$$\mathbf{V}_0 = \mathbf{V_g}\, R_P / (R_S + R_P)$$

and if we choose the generator voltage as our phase reference we can drop dimensions:

$$V_0 = V_g\, R_P / (R_S + R_P)$$

We will also avail ourselves of a useful property of the potential divider formula (**35.4**), which is that if we multiply it by a unit quantity consisting of the source resistance divided by itself (i.e., R_S/R_S), it becomes a double-slash product:

$$V_0 = V_g\, (R_P // R_S) / R_S$$

Similarly, for the output voltage in general:

$$\mathbf{V} = V_g\, (R_P // \mathbf{j}X_{Cp} // \mathbf{j}X_{Lp}) / [R_S + (R_P // \mathbf{j}X_{Cp} // \mathbf{j}X_{Lp})]$$

and using the associative rule (**18.4**):

$$\mathbf{V} = V_g\, (R_S // R_P // \mathbf{j}X_{Cp} // \mathbf{j}X_{Lp}) / R_S$$

So we can write the ratio $\mathbf{V}/V_0$ as:

$$\mathbf{V}/V_0 = (R_S // R_P // \mathbf{j}X_{Cp} // \mathbf{j}X_{Lp}) / (R_P // R_S)$$

The bandwidth function is the magnitude of this expression; but with all of the components represented as impedances, anyone attempting to expand and simplify it, or isolate part of it as the load, will have a hard time keeping track of all of the intermediate terms. We will therefore convert it into an admittance problem, using the relationship:

$$1/(\mathbf{Z}_1 // \mathbf{Z}_2 // \mathbf{Z}_3 // . . . // \mathbf{Z}_n) = \mathbf{Y}_1 + \mathbf{Y}_2 + \mathbf{Y}_3 + ... + \mathbf{Y}_n$$

Hence:

$$\mathbf{V}/V_0 = \frac{1 / (G_S + G_P + \mathbf{j}B_{Cp} + \mathbf{j}B_{Lp})}{1 / (G_P + G_S)}$$

Where G stands for conductance and B for susceptance, and $G_S=1/R_S$, $G_P=1/R_P$, $B_{Cp}=-1/X_{Cp}$ and $B_{Lp}=-1/X_{Lp}$. The expression above can be re-written:

$$\mathbf{V}/V_0 = \frac{G_S + G_P}{G_S + G_P + \mathbf{j}B_{Cp} + \mathbf{j}B_{Lp}}$$

and the magnitude is:

$$|\mathbf{V}| / V_0 = \frac{\sqrt{[(G_S + G_P)^2]}}{\sqrt{[(G_S + G_P)^2 + (B_{Cp} + B_{Lp})^2]}}$$

i.e.,

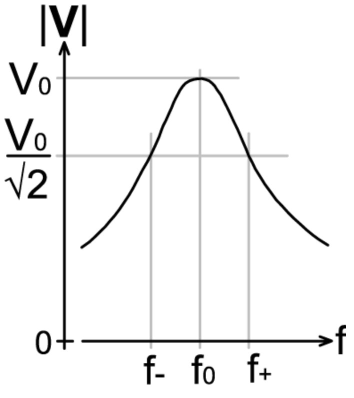
$$\boxed{|\mathbf{V}| / V_0 = \frac{G_S + G_P}{\sqrt{[(G_S + G_P)^2 + (B_{Cp} + B_{Lp})^2]}}}$$

This can be plotted against frequency by substituting $B_{Cp}=2\pi fC_P$ and $B_{Lp}=-1/(2\pi fL_P)$, but we will not bother to do so here because it is identical in appearance to the graph of |**I**|/I0 for a series resonator given in section **29**. We will instead go on to determine the half-power points by noting that, whatever proportion of the parallel resistance R_P is designated as the load, power will always be delivered to it in proportion to $|\mathbf{V}|^2$, so the half-power points occur when $|\mathbf{V}|=V_0/\sqrt{2}$. Hence, at the half-power points we have:

$$\frac{G_S + G_P}{\sqrt{[(G_S + G_P)^2 + (B_{Cp} + B_{Lp})^2]}} = 1/\sqrt{2}$$

Which, upon squaring gives:

$$\frac{(G_S + G_P)^2}{(G_S + G_P)^2 + (B_{Cp} + B_{Lp})^2} = \tfrac{1}{2}$$



and upon inversion gives:

$$\frac{(B_{Cp} + B_{Lp})^2}{(G_S + G_P)^2} + 1 = 2$$

i.e.:

$(B_{Cp} + B_{Lp})^2 / (G_S + G_P)^2 = 1$

and taking the square root:

$(B_{Cp} + B_{Lp}) / (G_S + G_P) = \pm 1$

Thus:

$B_{Lp} + B_{Cp} = \pm(G_S + G_P)$

and if we define the sum G_S+G_P as G_Q (i.e., the conductance that determines the Q):

$B_{Lp} + B_{Cp} = \pm G_Q$

Now, using the substitutions B_{Cp}=$2\pi f C_p$ and B_{Lp}=$-1/(2\pi f L_p)$, we obtain:

$[\ -1/(2\pi f L_p)\ ] + 2\pi f C_p = \pm G_Q$

and by factoring out $1/(2\pi f L_p)$ from the left hand side and re-arranging:

$\pm 2\pi f L_p G_Q = -1 + (2\pi f)^2 L_p C_p$

i.e.,

$[4\pi^2 L_p C_p]f^2 \pm [2\pi L_p G_Q]f -1 = 0$

This is a quadratic equation in f with $a=4\pi^2 L_p C_p$, $b=\pm 2\pi L_p G_Q$, and $c=-1$. It has four solutions as was the case for the series resonator (section **29**), these being the upper and lower bandwidth limits for positive and negative frequencies. To solve it we apply the standard formula:

$f = [-b \pm\sqrt{(b^2 - 4ac)}\ ] / 2a$

Hence:

$f = \{ \pm 2\pi L_p G_Q \pm\sqrt{[(2\pi L_p G_Q)^2 + 4\times 4\pi^2 L_p C_p]}\ \} / (2\times 4\pi^2 L_p C_p)$

and using the substitution L_p=L_p^2/L_p to obtain cancellation of L_p from all but one term:

$f = \{ \pm L_p G_Q \pm\sqrt{[(L_p G_Q)^2 + 4(L_p^2/L_p)C_p]}\ \} / (4\pi L_p C_p)$

i.e.:

$$\boxed{f = \{ \pm G_Q \pm\sqrt{(G_Q^2 + 4C_p/L_p)}\ \} / (4\pi C_p)}$$

Now, since $\sqrt{(G_Q^2 + 4C_p/L_p)}$ will always be larger than G_Q, we can identify the positive frequency upper bandwith limit as:

$$f_+ = \{ [\sqrt{(G_Q^2 + 4C_p/L_p)}] + G_Q \} / (4\pi C_p)$$

and the positive frequency lower bandwidth limit as:

$$f_- = \{ [\sqrt{(G_Q^2 + 4C_p/L_p)}] - G_Q \} / (4\pi C_p)$$

and the bandwidth is:

$$f_w = f_+ - f_- = G_Q/(2\pi C_p)$$

This is the admittance counterpart of the result obtained at this stage in the derivation of the Q of a series resonator (equation **29.3**) and so we will deduce that the bandwidth of the parallel resonator BPF is f_0/Q_0, and use this deduction to find a definition for Q.

$$f_0/Q_0 = G_Q/(2\pi C_p)$$

$$2\pi f_0 C_p = B_{Cp0} = Q_0\ G_Q$$

$$Q_0 = B_{Cp0} / G_Q$$

Now let $R_Q=1/G_Q$, where R_Q is "the resistance that determines the Q ". Also observe that $B_{Cp0}= -1/X_{Cp0}$, and at resonance $-X_{Cp0}=X_{Lp0}$. Hence:

$$\boxed{Q_0 = -R_Q / X_{Cp0} = R_Q / X_{Lp0}}$$

or, in keeping with the definition of resonant Q given in section **31**:

$$\boxed{Q_0 = R_Q / \sqrt{(L_p/C_p)}}\ \boxed{\mathbf{45.1}}$$

R_Q is simply the parallel combination of the source resistance, the load resistance, and the dynamic resistance of the resonator, i.e.:

$$R_Q = R_S\ //\ R_{p0}\ //\ R_{Load}$$

This result gives us the theoretical information we need in order to be able to design parallel resonant bandpasss filters. Firstly, we may observe that the source and load impedances are effectively in parallel with the resonator, which is why any minor source and load reactances can be lumped with the resonator reactances and cause only a detuning effect (if such reactances are very large however, they will cause a significant change in the dynamic resistance and the problem is best re-analysed from scratch). The source, load, and dynamic resistances however, are critical in determining the Q, and we need to obtain high values for all of them in order to obtain a high Q. We can of course adjust the source and load resistances using transformers; and as we shall see shortly, we can replace the resonator coil with a transformer so that the inductor and the transformer become one and the same. Before we look at such coupling schemes however, we must draw attention to a particularly misleading inference of the formula, which is that high Q can be obtained by making the ratio L_p/C_p as small as possible. This suggestion has appeared in at least one amateur radio publication, but it is a fallacy. If the reactive components are of reasonable quality, the parallel form L/C ratio (L_p/C_p) is only slightly different from the series form L/C ratio, and as we showed in section **21**, imaginary resonance can occur if the L/C ratio becomes too low. The imaginary resonance condition is entirely a function of the series (loss) resistances of the coil and the capacitor. It is nothing to do with the source and load resistances because avoidance of imaginary resonance is a matter of ensuring that the +90° component of the coil current at resonance is sufficiently large to cancel the -90° component of the capacitor current (or vice versa, but in practice coils are more lossy than capacitors). Consequently, the design procedure for a parallel resonator BPF is to make the L/C ratio large enough to obtain a good strong resonance (without making the inductance so large that stray capacitance and coil self-capacitance prevent the target maximum frequency from being reached), and then to make R_{p0} even larger (by minimising loss resistances) in order to obtain a useful working Q.

46. Unloaded Q of parallel resonator

The expression for Q obtained in the previous section is sometimes referred to as the *loaded Q* of the resonator, because it is the Q that results when the source and load impedances are taken into account. We may also imagine that the resonator has an *unloaded Q*, which is that which obtains when the source and load are disconnected. It is not immediately obvious why we should wish to employ such a concept, because it is impossible to use the resonator without coupling to it in some way; but it is nevertheless useful because it sets an upper limit on the Q that can be obtained in a practical circuit. It is obviously obtained by substituting the dynamic resistance in place of R_Q in equation (**45.1**), i.e.,

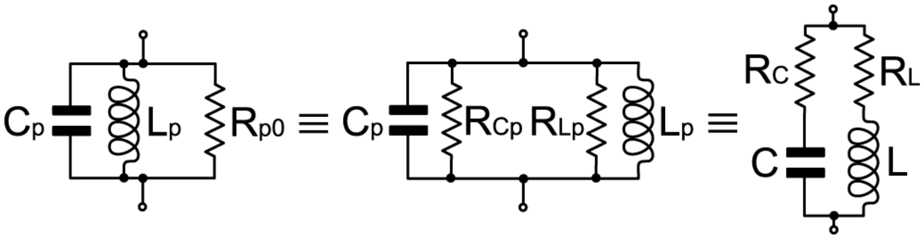
$$Q_{0u} = R_{p0} / \sqrt{(L_p/C_p)}$$

but it would be a lot more useful intuitively if we could express it in terms of the coil and capacitor impedances in their series (R+**j**X) forms. We can do so by using the series-to-parallel transformation (section **19b**); and using the definition of Q from section **31** as precedent, we expect a result in the form:

$$Q_{0u} = [\sqrt{(L/C)}] / R \qquad . . . (\textbf{46.1})$$

the point being to find out what is meant by R in this case.

The translation from parallel to series form is indicated in the set of equivalent circuits shown below:



Here we identify R_{p0} as $R_{Cp}//R_{Lp}$, i.e.:

$$R_{p0} = R_{Cp} R_{Lp} / (R_{Cp} + R_{Lp})$$

and an expansion in terms of the series forms of the impedances has already been given as equation (**20.5**):

$$R_{p0} = \frac{(R_C^2 + X_C^2)\ (R_L^2 + X_L^2)}{R_L(R_C^2 + X_C^2) + R_C(R_L^2 + X_L^2)}$$

The unloaded Q is defined as:

$$Q_{0u} = R_{p0} / \sqrt{(-X_{Cp}\ X_{Lp})}$$

and, from the series-to-parallel transformation (equation **19.3b**), we have:

$$X_{Cp} = (R_C^2 + X_C^2) / X_C \qquad . . . (\textbf{46.2})$$

and

$X_{Lp} = (R_L^2 + X_L^2) / X_L \quad ...$ (**46.3**)

Putting all of this together we have:

$$Q_{0u} = \frac{(R_C^2 + X_C^2)\,(R_L^2 + X_L^2) \,/\, [\, R_L(R_C^2 + X_C^2) + R_C(R_L^2 + X_L^2)\,]}{\sqrt{[\, -(R_C^2 + X_C^2)\,(R_L^2 + X_L^2) \,/\, X_C\, X_L\,]}}$$

and noting that $\sqrt{(-X_C\, X_L\,)}=\sqrt{(L/C)}$, this rearranges to:

$$Q_{0u} = \frac{[\sqrt{(L/C)}]\ \{\sqrt{[\,(R_C^2 + X_C^2)\,(R_L^2 + X_L^2)\,]}\ \}}{[\, R_L(R_C^2 + X_C^2) + R_C(R_L^2 + X_L^2)\,]}$$

So, at this point we have extracted $\sqrt{(L/C)}$ as required by equation (**46.1**), and the resistance by which $\sqrt{(L/C)}$ must be divided in order to obtain Q_{0u} is:

$$R=\sqrt{\left[\frac{\left[R_L\left(R_C^2+X_C^2\right)+R_C\left(R_L^2+X_L^2\right)\right]^2}{\left(R_C^2+X_C^2\right)\left(R_L^2+X_L^2\right)}\right]} \qquad (\mathbf{46.4})$$

Which, upon expanding the numerator gives:

$$R^2 = \frac{R_L^2(R_C^2+X_C^2)^2}{(R_C^2+X_C^2)(R_L^2+X_L^2)} + \frac{R_C^2(R_L^2+X_L^2)^2}{(R_C^2+X_C^2)(R_L^2+X_L^2)} + \frac{2R_LR_C(R_C^2+X_C^2)(R_L^2+X_L^2)}{(R_C^2+X_C^2)(R_L^2+X_L^2)}$$

The simplification we require here comes from noting that the terms $(R_C^2+X_C^2)$ and $(R_L^2+X_L^2)$ occur in the expressions for X_{Cp} and X_{Lp} given above (equations **46.2** and **46.3**), and that at resonance $-X_{Cp}=X_{Lp}$. Hence:

$$(R_C^2 + X_C^2) / (-X_C) = (R_L^2 + X_L^2) / X_L$$

i.e.:

$$(R_C^2 + X_C^2) / (R_L^2 + X_L^2) = -X_C / X_L \quad \text{and} \quad (R_L^2 + X_L^2) / (R_C^2 + X_C^2) = X_L /-X_C$$

Hence:

$$R^2 = R_L^2(-X_C/X_L) + R_C^2(X_L/-X_C) + 2R_LR_C$$

w

hich can be factorised:

$$R^2 = \{\, R_L[\sqrt{(-X_C/X_L)}] + R_C[\sqrt{(X_L/-X_C)}]\ \}^2$$

Hence:

$$R = R_L[\sqrt{(-X_C/X_L)}] + R_C[\sqrt{(X_L/-X_C)}] \quad ...\ (46.5)$$

strictly ±R, but resistance is positive, allowing us to ignore the negative solution) and so:

$$\boxed{Q_{0u} = [\sqrt{(L/C)}]\ /\ \{\ R_L[\sqrt{(-X_C/X_L)}] + R_C[\sqrt{(X_L/-X_C)}]\ \}}$$

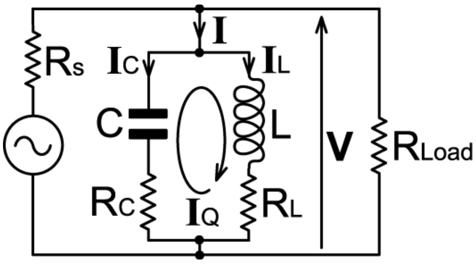
This is an exact solution provided that R_L and R_C do not vary; and once again may be assumed exact for normal engineering purposes because R_L and R_C will not vary significantly in the vicinity of the resonant frequency. Note however that X_L=-X_C to an extremely good approximation when the L/C ratio is reasonably large, and this relationship is exact when R_L=R_C. Hence, for most practical purposes:

$$\boxed{Q_{0u} = [\sqrt{(L/C)}]\ /\ (R_L + R_C)}\boxed{\mathbf{46.6}}$$

Which means that the unloaded Q of the parallel resonator is the same as that of the series resonator, it is the square root of the L/C ratio divided by the total series resistance. In other words, we can estimate the unloaded Q of the parallel resonator by considering it to be a series resonator connected as a loop.

47. Current magnification

There is another way to determine the unloaded Q of a parallel resonator, which stems from the observation, that just as a series resonator exhibits the phenomenon of voltage magnification, the parallel resonator exhibits *current magnification*. In effect, the parallel resonator is a series resonator connected in a different way, because its characteristics at resonance are principally determined by a large circulating current, and the current it draws from the generator is small in comparison (Q times smaller than the circulating current in fact).



In the diagram on the right, the current **I** flowing into the resonator is $\mathbf{I}_L+\mathbf{I}_C$, where:

$$\mathbf{I}_C = \mathbf{V} / (R_C + \mathbf{j}X_C) = \mathbf{V}(R_C - \mathbf{j}X_C)/(R_C^2+X_C^2)$$

and

$$\mathbf{I}_L = \mathbf{V} / (R_L + \mathbf{j}X_L) = \mathbf{V}(R_L - \mathbf{j}X_L)/(R_L^2+X_L^2)$$

but at reasonance **I** is real, which means that the imaginary parts of $\mathbf{I}_L$ and $\mathbf{I}_C$ add up to zero, and the total current at resonance becomes:

$$I_0 = V \left[\frac{R_C}{(R_C^2+X_C^2)} + \frac{R_L}{(R_L^2+X_L^2)} \right]$$

(V and I_0 are now in phase and will be treated as real). Putting the expression onto a common denominator yields:

$$I_0 = V \left[\frac{R_C(R_L^2+X_L^2) + R_L(R_C^2+X_C^2)}{(R_C^2+X_C^2)\,(R_L^2+X_L^2)} \right] \qquad \textbf{47.1}$$

where the term inside the square brackets is the reciprocal of the dynamic resistance (see equation **20.5**), i.e., $I_0=V/R_{p0}$.

Now, the current circulating in the resonator can be determined from either branch as the total current flowing in the branch, less the current drawn from the generator. Hence the circulating current is simply the imaginary part of the current in the branch. The resonant condition (and the concept of circulation) also implies that the circulating current is of the same magnitude for both branches, but of opposite sign. Hence, if we call the circulating current $\mathbf{I}_Q$, then (taking the imaginary parts of the expressions for $\mathbf{I}_C$ and $\mathbf{I}_L$ above) we have:

$$\mathbf{I}_Q = \mathbf{j}VX_L/(R_L^2+X_L^2) = -\mathbf{j}VX_C/(R_C^2+X_C^2)$$

and the magnitudes are:

$$|\mathbf{I}_Q| = VX_L/(R_L^2+X_L^2) = V(-X_C)/(R_C^2+X_C^2)$$

We can also create a definition involving both branches by taking the geometric mean:

$|\mathbf{I}_Q| = V\sqrt{\{(-X_CX_L) / [(R_C^2+X_C^2)(R_L^2+X_L^2)]\}}$

which allows us to extract the L/C ratio:

$|\mathbf{I}_Q| = V\sqrt{\{(L/C) / [(R_C^2+X_C^2)(R_L^2+X_L^2)]\}}$ . . . . (47.2)

Now, let us define the unloaded Q of the resonator as the ratio of the circulating current to the through current:

$Q_{0u} = |\mathbf{I}_Q| / I_0$

Which can be expanded using equations (47.1) and (47.2):

$$Q_{0u} = \sqrt{\left[\frac{(L/C) / [(R_C^2+X_C^2)(R_L^2+X_L^2)]}{[R_C(R_L^2+X_L^2)+R_L(R_C^2+X_C^2)]^2 / [(R_C^2+X_C^2)(R_L^2+X_L^2)]^2}\right]}$$

and rearranged:

$$Q_{0u} = [\sqrt{(L/C)}]\sqrt{\left[\frac{(R_C^2+X_C^2)(R_L^2+X_L^2)}{[R_C(R_L^2+X_L^2)+R_L(R_C^2+X_C^2)]^2}\right]}$$

The rightmost square-root bracket is simply the reciprocal of R as defined in equation (46.4); hence:

$Q_{0u} = [\sqrt{(L/C)}] / R$

and we have proved that the current-magnification definition for unloaded Q is identical to that obtained on the assumption that Q is the magnitude of the resonant frequency divided by the bandwidth of the resonator.

The only residual issue is that of why the exact expression for R is (as given by equation 46.5):

$R = R_L[\sqrt{(-X_C/X_L)}]+R_C[\sqrt{(X_L/-X_C)}]$

rather than simply $R=R_L+R_C$. This however can be understood by noting that the (real) current flowing through the resonator will be very slightly biased in favour of the branch with the lowest resistance. This difference is very small for practical resonators of moderate unloaded Q, and may normally be ignored.

48. Controlling loaded Q

As determined earlier (**45.1**), the Q of a parallel resonator can be given as:

$$Q_0 = (R_{\mathbf{Src}} // R_{\mathbf{p0}} // R_{\mathbf{Load}}) / \sqrt{(L_p/C_p)}$$

where, with a slight change from the previous notation, $R_{\mathbf{Src}}$ is the output resistance of the energy injection network (the source), $R_{\mathbf{p0}}$ is the dynamic resistance of the resonator, and $R_{\mathbf{Load}}$ is the input resistance of the network to which energy is being delivered. The inductance and capacitance are defined in their parallel forms so that the dynamic resistance can be treated as a separate parallel element.

To be continued . . . . .