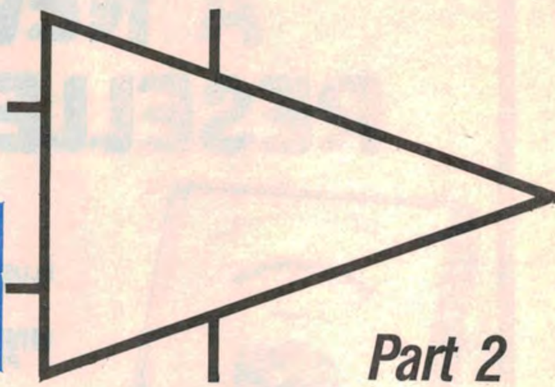
by BRYAN MAHER

OP AMPS Explained



Part 2

Our circuits can become unstable unless we observe certain precautions. This month we take a look at time constants, frequency response, and the criteria necessary to ensure circuit stability.

Every circuit, system, machine, person or thing takes time to change its state, that is, its energy. This time is a characteristic called the *time constant* and has a profound effect on circuits and systems. This is an inescapable rule for the whole universe and everything in it, including operational amplifiers and even ourselves.

A change in voltage output of an operational amplifier implies a change in energy level which *must* take time to accomplish. The circuit is quite incapable of changing its energy level instantaneously, just as we are incapable of accelerating instantaneously from standstill to running at 15km per hour (ask any slip fieldsman!).

To fully describe our circuits we must talk in terms of variable quantities such as voltage, current, frequency, power and energy; and also constant quantities which are resistance, capacitance and inductance. It is quite possible for us to invent more variable quantities, for example "period", and we could give a name to "the rate of change of voltage". But it is a fact not always recognised that in electrical circuits there can exist only three constants, the ones quoted above. There are no more!

You may immediately explode into objections and want to beat the author into the ground because he did not

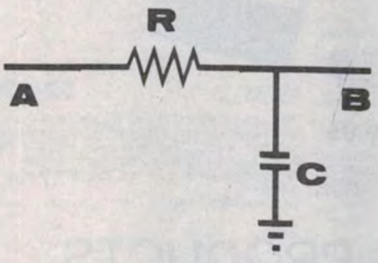


Fig. 1: A time constant.

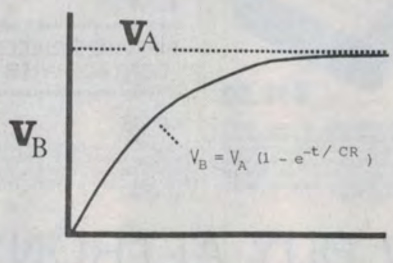


Fig. 2: The rise of voltage at B for a voltage step at A.

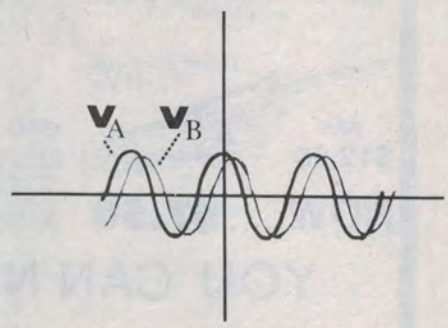


Fig. 3: The oscilloscope traces of voltages at A and B.

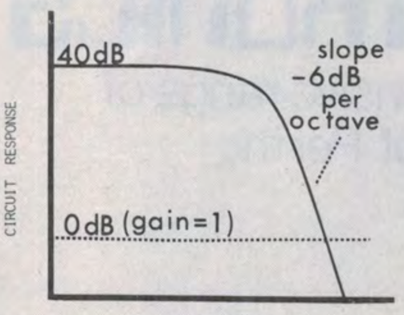


Fig. 5: Bode amplitude response for one time constant.

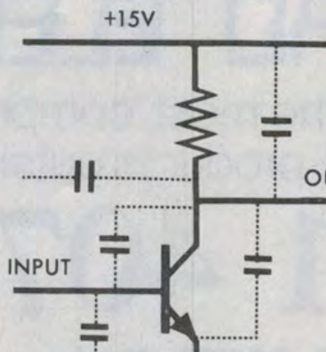


Fig. 6: Stray capacitances.

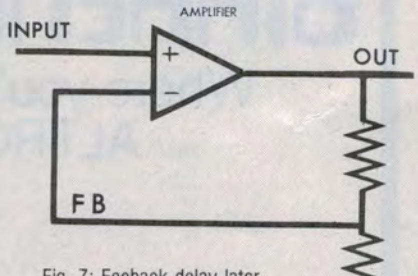


Fig. 7: Feeback delay later than Vin.

mention "conductance" and there are those other names too like "impedance" and even "admittance"! The answer to your resistance (pardon the pun) to your author's statement becomes clear when we look into the meanings of those constants. In electrical systems we can only do two things using our constant quantities: we can either *lose* the electrical form of some of our energy or we can *store* it. The loss of the electrical form of some energy is associated with resistance, and it is the idea of "loss" that is important.

It is not important whether we call it "resistance in ohms" or the inverse term "conductance in Siemens". Also it does not matter if this "loss element" is a physical resistor bought over the counter or the radiation resistance of a transmitting antenna (a non-visible quantity) or the output resistance Ro of a feedback amplifier (a defined relation: change in voltage/change in current). Whenever we lose the electrical form of energy we have the one basic constant and please yourself what you call it. The energy lost is of course i^2R (or V^2/R or $Vi\cos\Theta$ where Θ is the phase angle between the voltage and current waveforms).

The other two constant quantities, capacitance and inductance, are quite incapable of energy loss and form two

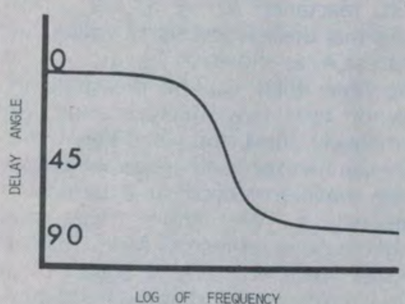


Fig. 4: Bode phase diagram for one time constant.

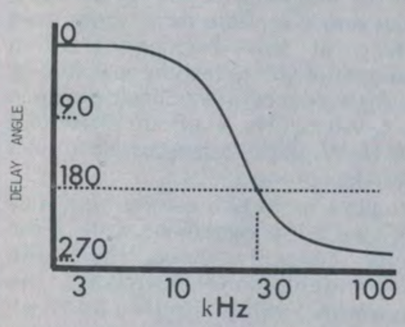


Fig. 8a: Bode phase diagram for three time constants.

Fig. 8b, c, d: How complex numbers are plotted

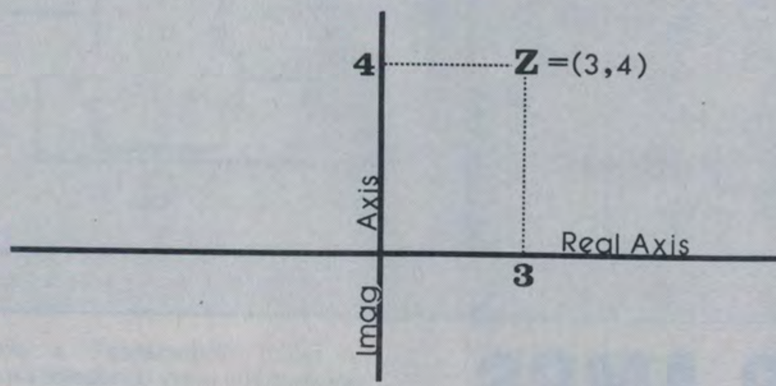


Fig. 8b: A single complex number Z = (3, 4) represented on a two axis co-ordinate system.

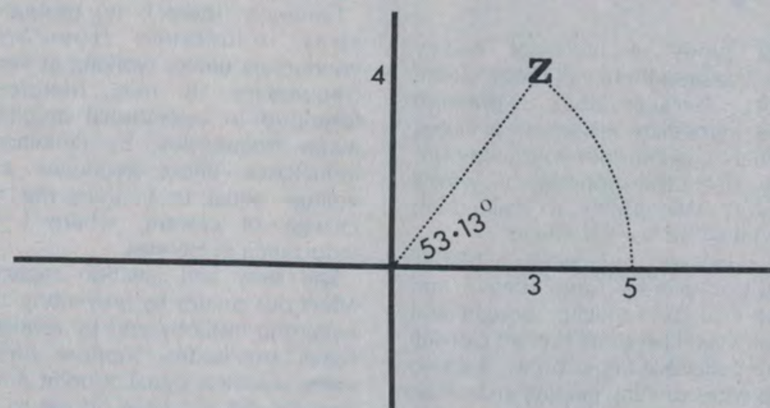


Fig. 8c: The complex number Z = (3, 4) is just the real number 5 rotated anticlockwise 53.13 degrees because the 3 and 4 and the origin form a right angle triangle. The 5 comes from Pythagoras and $5^2 = 3^2 + 4^2$. The angle comes from 4/3 = Tan (53.13°).

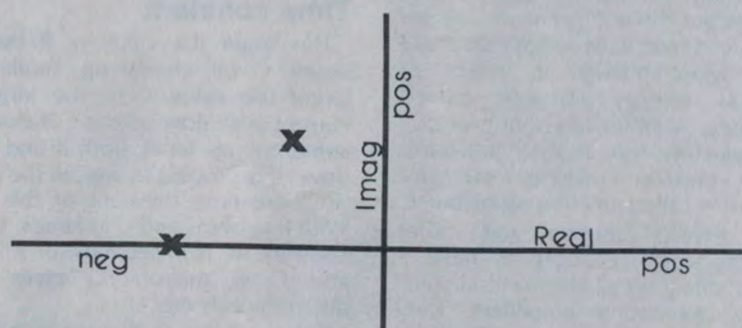


Fig. 8d: A possible set of poles for a three pole system as shown in Fig. 8a. In this case, all poles have negative real parts and the system is stable. The standard mark for a pole is X, as shown here.

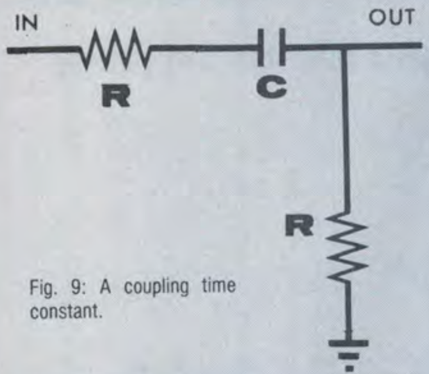


Fig. 9: A coupling time constant.

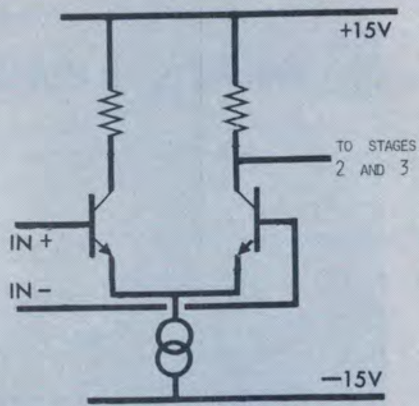


Fig. 10: A long tail pair first stage

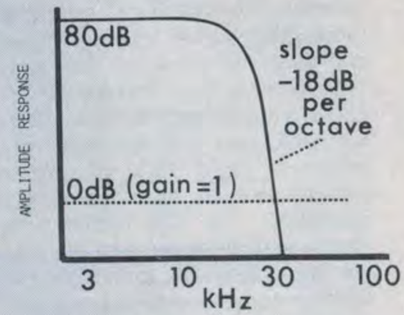


Fig. 11: Bode amplifier response for three time constants.

OP AMPS Explained

different types of electrical energy storage. We hesitate to call these "circuit elements" because that expression conjures immediate visions of a circuit component you can "put your finger on" or buy over the counter, a wrong impression! We prefer to call them "circuit ideas" or "circuit effects".

If we build into our circuit a 1000μF capacitor it will be large, heavy and obvious. A 2.2pF capacitor bought and soldered in will be small but we can still "put our collective finger on it". But two hookup wires running parallel in a circuit may easily have the effect of 5pF capacitance between them, even if we did not wish for that effect. Yes, you say, we can point to those two hookup wires, but we cannot "put our thumb on" that 5pF stray capacitive effect that we wish was not there. We try closing our eyes a few times but it won't go away: we are finally forced to acknowledge that those parallel wires posses an effect of electrical energy storage called capacitance, whether we want it or not! Semiconductors too at their junctions, exhibit similar energy storage mechanisms called junction capacitance.

These stray, junction and other unwanted small capacitances have a profound effect on all electronic circuits, including operational amplifiers. They dictate in no small measure the practical use we can and cannot make of our designs. If we really want a definition of this capacitance, it is the effect wherein a current flowing is proportional to a constant C times the rate of change of voltage. The constant C is the capacitance in Farads. The energy stored at any time is simply $\frac{1}{2}CV^2$.

The second energy storage mechanism is called "inductance", a property possessed by every conductor carrying a current. Although straight wires have the property only in a small way, wires wound into coils have a much greater inductance, the more so if wound on cores or iron, cobalt, nickel or ferrite.

Generally, there is no problem with stray inductance from straight conductors unless working at very high frequencies. It may therefore be forgotten in operational amplifiers at audio frequencies. By definition, the inductance effect produces a back voltage equal to L times the rate of change of current, where L is the inductance in Henries.

The stray and junction capacitances affect our circuits by preventing changes occurring instantly and by reducing AC signal amplitudes. Suppose there was some electrical signal at point A in Fig. 1 and we did not have access to A, but could only look at the voltage at point B. If the voltage at A began to rise current would flow from A through R to charge the capacitor C and the voltage on C would be proportional to such charge.

Time constant

The larger the value of R the more slowly C will charge up. Similarly, the larger the value of C the longer the current must flow before C is charged to some voltage level. Both R and C slow down this process so we call the product "RC" the time constant of this circuit. With R in ohms and C in farads, the time constant RC is in seconds; or R in ohms and C in microfarads gives RC in microseconds etc.

If the voltage at A were to be in the form of a step from 0 to 1 volt, then the voltage at B would rise somewhat more slowly as in Fig. 2. Let's assume that we apply a step voltage V_A at point A in Fig. 1. The voltage at B, V_B, is given by the equation:

$$V_B = V_A(1 - e^{-t/RC})$$

where t is the time in seconds after the voltage at A took that step. After a time equal to RC seconds, V_B will be about two thirds of V_A, the exact value being $(1 - 1/e)V_A$ or $0.632V_A$, where e = 2.718.

Delays

One consequence of the time constant is clear – the system is slow to react. Let us look a little deeper into the circuit's workings. Say a changing input, a sine wave voltage at frequency f, is applied to A. After a time there will also be a sine wave voltage at B, but it will always be a little behind the sine wave at A. That is, A's sine wave will pass through its peak and a while later the sine wave at B will peak. As well, capacitor C provides some path to ground for this signal due to its reactance $Xc = 1/2\pi fC$. This means that there is less signal voltage at B than at A, as shown in Fig. 3.

The time delay can be measured in units of time, say microseconds, or alternatively, for a sine wave signal, the delay can be expressed as some fraction of the waveform cycle at a particular frequency f. We know that one complete cycle represents 360°. Thus, if the sine wave at B is one eighth of a cycle later than at A, at some frequency f, we can just as easily say that it is 45° late. This is called the phase or angle delay.

In most circuits, this phase delay varies with the frequency of the signal. Fig. 4 shows how the phase delay varies from nothing at low frequencies to a maximum of 90° at high frequencies for the single time constant circuit shown in Fig. 1. We call Fig. 4 a Bode phase plot after H. W. Bode, a famous electronics researcher of the 1940s.

To allow us to plot a wide frequency range we use a logarithmic scale. If the circuit contains more than one independent time constant, the maximum delay angle at high frequencies increases in multiples of 90°. For example, a circuit with three independent time constants can have a maximum delay angle of 270° (3 × 90).

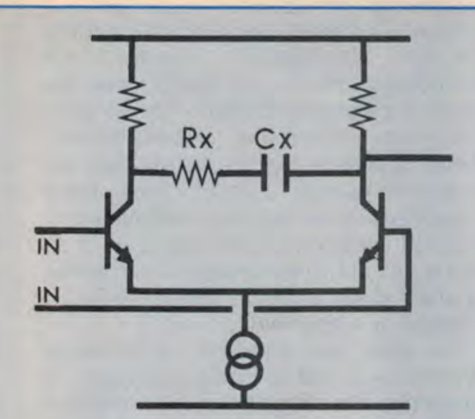


Fig. 12: The addition of Cx and Rx.

Not only does the phase angle vary with frequency but so also does the amount of signal flowing to ground via capacitor C. This is because $Xc = 1/(2\pi fc)$ becomes smaller at higher frequencies. Thus R and Xc form a frequency dependent voltage divider which results in less signal at B as the frequency f rises. We see this in the Bode amplitude response curve of Fig. 5. Logarithmic scales are used not only to allow a wide frequency range but also to give the curve its classic shape.

With a circuit containing more time constants, the frequency roll off occurs at a lower frequency and the curve races off downwards at a steeper slope. A single time constant produces a −6dB per octave slope (−20dB per decade), two equal independent time constants give −12dB per octave (−40dB per decade), and so on. In practice, the frequency rolloff is determined not only by the capacitors we deliberately add to a circuit but by stray capacitance as well.

The villain

The effect of stray capacitance has the nasty habit of turning up everywhere. For example, in Fig. 6 we get stray capacitance from the collector of the transistor to earth, to the +15V rail, to the base via internal collector-base capacitance and to any earlier stages. Each capacitive path contributes to the total stray capacitance but that from the collector to base is the real villain of the story. It appears as C(cb).

Consider what happens when an input signal V is applied to C(cb). Because this signal is applied to the base of the transistor, it appears at the collector, amplified by the stage gain A, as −AV. So the current through C(cb) is a result of the difference in signal potential beteen V and −AV. Thus the current in this stray capacitance is equivalent to (Ic + AIc).

The stray capacitive current flowing to the base is much larger than expected, as if the real collector-base capacitance C(cb) had in parallel with it a fictitious capacitance A.C(cb), making the total collector-base capacitance C(cb) + A.C(cb), or if you like (1+A)C(cb). Because this effect was first explained in 1919 by J. M. Miller, we call this fictitious capacitance A.C(cb) the Miller capacitance or Miller effect. It is fictitious in that you cannot put your finger on it, but its effect is very real.

In high gain amplifier stages, this Miller capacitance is by far the biggest stray capacitance found in any circuit's backyard and is a real villain. It causes an unwanted increase in the phase angle (or time constant) and a corresponding decrease in frequency bandwidth.

Stability

All the above effects – stray and Miller capacitance, frequency response and phase angle delay – have, of course, always been present in all circuits, including our operational amplifiers. Now let's see what we are forced to consider because of them.

Fig. 7 represents our fundamental feedback circuit but now we recognise an effect that has been going on all along; because of unavoidable capacitances in our amplifier block, the output and hence the feedback FB is always later than the input. The phase delay angle of the amplifier is an important consideration when designing feedback circuits.

Take the case of three amplifier stages, each with stray and Miller capacitances. Fig. 8 is the Bode Phase Plot for such a circuit and shows the phase angle delay between the input signal and the feedback signal at various frequencies f. Now, just a cotton-picking-minute!!

At a frequency of about 28kHz we see that the amplifier has delayed its output so much that the feedback signal has rotated a whole 180°! And we remember learning somewhere that a phase rotation of 180° means the same thing as phase reversal – a change of sign. And that means the supposedly negative feedback has reversed in sign (when the frequency is 28kHz) to become *positive* feedback, which will add to the input signal not subtract from it.

This larger signal will then go into the amplifier whereupon it will be amplified again and fed back to add to the input signal and so on ad infinitum. That is obviously a runaway condition!

If we are discussing an audio amplifier the worst we can have is a unit which generates whistles, screeches and howls all on its own initiative. It oscillates its head off!

But worse could be in store in the case of an operational amplifier controlling some large electro-mechanical milling machine. The result could be physically and financially disastrous! Clearly we must take steps to prevent such uncontrolled behaviour – such oscillations. We cannot simply say "don't use 28kHz signals". That will not stop the trouble, because when power is first applied to the system, an abrupt step signal, almost a square wave, is fed to the circuit.

Since a square wave is a signal containing all frequencies (the famous Fourier analysis), all frequencies, including 28kHz, would be applied at switch-on and oscillations would start. Even without switch-on surges, the first sharp interference pulse received by extraneous pick-up would initiate oscillations for the same reasons.

What to do

What causes those oscillations and what should be done to inhibit the conditions allowing them? Thinking back, two factors were involved. Firstly the phase angle delay could exceed 180° at some frequency; secondly the gain around the loop *was greater than one at the frequency which produced 180° angle delay.* Preventing either condition will prohibit oscillations. If a circuit contains enough stray capacitances to make up three or more time constants we cannot prevent the first. So we must prevent the second.

In the Bode amplitude response curve (Fig. 5), the gain around the loop falls with increasing frequency. We must therefore ensure that the gain falls to less than one before the frequency which causes the 180° phase angle delay. That will be sufficient to ensure stability in any feedback circuit, including our operational amplifiers.

Poles and zeros

Recall from the first article published last month that we used G for the open loop gain of our high gain integrated amplifier, T for the closed loop gain, and H for the feedback factor. The closed loop gain with feedback was given by:

$$T = \frac{G}{1 + GH}$$

Really, the term G should make clear its dependence on frequency; that is, its relation to all its time constants. The time constant has the effect of reducing the gain at higher frequencies, but has no effect at DC.

To accurately calculate G and T at any frequency would submerge us deeper in mathematics than we wish. We would just like to paddle a little. It can be shown that the denominator in the above equation for T decides system stability; so much so that the special equation derived from it:

$(1 + GH) = 0$

is called the *characteristic equation* for the system.

OP AMPS Explained

As we have already seen, three time constants are unavoidable in a 3-stage amplifier system since each stage has Miller capacitance. At some frequency, the phase of the output, and hence the feedback, will be rotated 180°, resulting in instability if the loop gain at that frequency is greater than one. In fact, there are *three* frequencies at which this might happen for a three-stage system. We give these special frequencies a name. We call them *poles*.

Thus the poles of any system are those frequencies at which the system is unstable. You could say that, at pole frequencies, the gain of the closed loop system is infinite, which is really an imaginative description of a system way out of control. In general, a system has as many poles as it has independent low-pass time constants. The poles are related to the reciprocals of these time constants.

Unattainable poles

You are bursting into objection again! If all such systems have poles (which are the frequencies of instability), how come we ever have any stable systems at all? In a four time-constant system, for example, there exist four pole frequencies, any one of which will (theoretically) make the system unstable. The answer is that some or all of those pole frequencies may be unattainable in a practical system.

But hang on! Didn't we just agree that Fourier's theorem tells us that all frequencies appear in sudden step voltages (such as switch-on surges)? The answer to this dilemma lies in the fact that Fourier's theorem refers to *positive* frequencies only.

Imagine the following scenario: suppose we have a four pole system; ie there are four time constants. Yes it has four frequencies of instability. By a mathematical procedure we can find out what those frequencies are. They are, in fact, the four roots of the aforementioned system characteristic equation.

A glimmer of hope appears! Yes of course! Some equations have both positive and negative roots. Now what would "negative" mean in frequency? It would mean a frequency that was not physically realisable! So that's it – if all poles turned out from the mathematics to be negative frequencies, they would not be physically realisable. Thus, the instability condition could not be attained, which is just a fancy way of saying that the system is *stable*.

Complex is simple

Now let's think a bit more about the roots of equations in general. We recall that such roots might be positive or might be negative or might be something else. Yes, there are plenty of equations whose roots are not found amongst the positive or negative real numbers. Because of this, Karl Friedrich Gauss invented another type of number in 1797 – the *complex number*.

These are not really all that mysterious. A complex number can be simply represented as a pair of ordinary real numbers. For example, Z=(3,4) is a complex number. The 3 and the 4 are just ordinary real numbers. We have always been able to draw real numbers along a line (zero in the middle, positive to the right, negative to the left). How shall we draw these complex numbers? Easy – in a plane instead of a line. The single complex number Z=(3,4) is simply marked as a single point 3 units along and 4 units up, as in Fig. 8b. We call the horizontal axis the *real axis*, and here the 3 is called the *real part* of the complex number. The 4 in (3,4) is called the *imaginary part* and is drawn along the vertical or *imaginary axis*.

Now if it should turn out that the roots of our characteristic equation are complex numbers, that would mean the poles are complex frequencies. So what on earth does that mean? Simple! Looking again at Fig. 8b, we see that our complex number Z could have been regarded just as easily as a real number 5 rotated anticlockwise through 53.13° as in Fig. 8c. But to us an angle could mean a phase delay angle, as we saw in Fig. 3, caused by a time delay.

So now we know – a complex frequency is just a picturesque way of depicting the frequency of some signal delayed through some phase delay angle.

Complex numbers are neither positive or negative, but those terms still apply to their real parts. In Figs. 8b and 8c, the real part of Z is 3 which is positive. The criterion for stability is simply stated as:

If the real part of each pole frequency is negative, instability is not physically realisable and the system is stable. If the real part of any pole is positive, the system is unstable; ie, it will oscillate.

In Fig. 8d we see one possible set of poles for the three pole system depicted in Fig. 8a, in this case showing a stable system as all poles have negative real parts. Note that one of the poles only has a real component. However, we can regard ordinary real numbers as simply complex numbers whose imaginary parts happen to be zero. In this sense we can regard all numbers as complex.

Zeros

There is also another type of time constant, illustrated in Fig. 9, called a *coupling time constant*. This will not pass DC; ie, it has zero response at zero frequency. If included in a system, it produces some extra terms on the top line of the expression for the closed loop gain T.

Now for another definition: by a "zero" of a feedback system we mean any particular value of frequency which makes the top line equal to zero in the

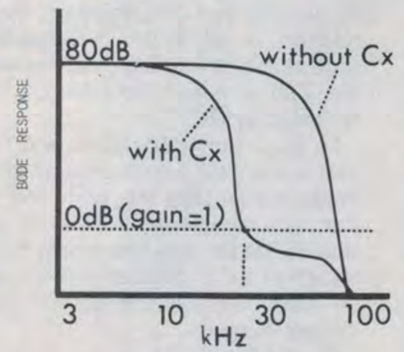


Fig. 13: The effect of Rx and Cx on the Bode response curve.

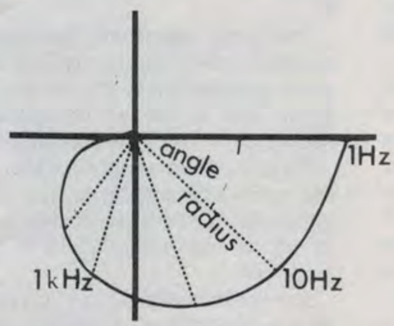


Fig. 14: Construction of Nyquist contour from radius and angle for two time constants.

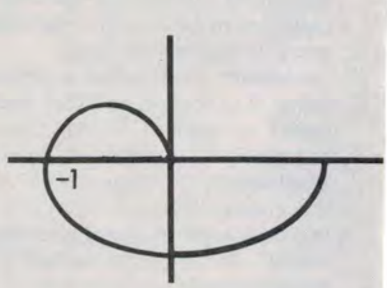


Fig. 15a: Nyquist contour for three pole unstable system.

expression for T. A "zero" is also a complex frequency. Naturally if the top line is zero, the value of closed loop gain T is zero at that frequency. Zeros have a way of cancelling poles of the same frequency.

Compensation circuit

We can now put all our new-found knowledge together. The first stage of an operational amplifier will usually consist of that beautiful differential circuit – the long tail pair as in Fig. 10. Commonly a second and third stage makes up the complete integrated circuit. Unfortunately, each stage will have Miller and other stray capacitance and we are forced to accept at least three time constants, giving Bode phase and gain plots like Figs. 8 and 11.

Because Fig. 11 shows gain still above unity (0dB) when the phase rotation is 180°, it denotes an unstable system. If a long tail pair has an impedance (any combination of resistance, capacitance and inductance) connected between the collectors it acts to reduce the stage gain – the lower the impedance the lower the gain.

Let us connect a resistance Rx and a capacitance Cx in series between the collectors as in Fig. 12. At DC and low frequencies the capacitor acts as an open circuit and both Xc and the gain are high. At medium frequencies, the gain is reduced as Xc becomes lower. This medium frequency gain is maintained all the way up to high frequencies because the impedance of this RC combination cannot fall any lower than Rx.

The Bode response curve of the modified circuit, Fig. 13, shows that the gain is safely down to less than unity (0dB) before the frequency is high enough to cause 180 degrees phase delay. Although the feedback does become positive, it is only when the gain is less than one, so oscillations cannot occur. Our system is stable.

If you feel clumsy looking at two graphs – Bode gain and Bode phase angle – to decide the stability question, we can make life easier by condensing both graphs into one. Beginning with a set of right-angle co-ordinates, we draw a radius line from the central origin with length equal to the Bode gain and angle equal to the Bode phase angle for a particular frequency, as shown in Fig. 14. By repeating this procedure for several different frequencies, we can draw a smooth continuous curve to join the tips of all those radius lines.

This curve is called the Nyquist Contour after H. Nyquist who established stability criteria for amplifier circuits during the 1930s.

Making it easier

Looking at this more closely, we observe that a loop gain of one and a bode angle of 180° results in a radius line one unit long to the left of origin. This tip of this radius line is labelled –1 on all Nyquist diagrams (since it is to the left of the vertical axis). If our Nyquist Contour cuts the negative axis outside that –1 point, it means that the gain is greater than one at a frequency causing a delay angle of 180° – ie, the circuit is unstable. Conversely, if the contour cuts the negative axis inside the –1 point, the gain is less than one and the system is stable.

The beauty of the Nyquist Contour is that if we draw the curve and mark the –1 point, we know whether or not that system is stable – just at a glance, as in Figs. 15a and 15b. All Nyquist contours enclosing the –1 point indicate unstable systems, while those not enclosing it belong to stable systems.

Some examples

The 301A integrated amplifier is a popular example of the many manufactured for operational amplifier use. You, the user, are meant to add the compensation capacitor Cx yourself. The circuit is such that if you want to make the closed loop gain = 1, you should use 33pF for Cx and connect it between pins 1 and 8. Of course, you must also install the correct values for Ri and Rf as discussed last month.

Let's assume, however, that you want to make the closed loop gain = 10. As well as choosing the appropriate values for Ri and Rf, you must now use a 3.3pF capacitor for Cx as in Fig. 16. Actually, it's quite easy to calculate the value of the compensation capacitor for the LM301AH (and the LM308AH). All you have to do is apply the formula:

Cx ≥ 30pFl (closed loop gain).

Naturally, we call this use of a single capacitor "single pole compensation". We could, if we chose, improve on the system using two capacitors and one resistor in a "two-pole compensation" scheme. The relevant data sheet should be consulted for this circuit, the improvement being noticed only when a response faster than about 20μs is desired (see Fig. 17).

For very high closed loop gains, 300 or higher, the value calculated for Cx is 0.10pF or less. This tiny capacitance approximates the stray capacitance between the legs of the integrated circuit, so no further capacitance need be added. But at lower gains we must insert capacitor Cx as calculated for these integrated circuits. Too small a Cx risks oscillation; too large a value slows down the response. The correct value gives the best bandwidth possible for a particular closed loop again.

In some op amps, the compensation capacitor Cx is built into the integrated circuit and we need not add it. Examples are the 741, 747, 740A, LH0022CH, AD545 and AD544. These integrated circuits, and many more, will give stable operation over a frequency bandwidth from DC to 10kHz or 100kHz. Others on the market extend the bandwidth to higher values. For example, the AD50XJ and LH0032 operate from DC to over 50MHz.

Next month we will consider noise and distortion in operational amplifiers and what to do about it.

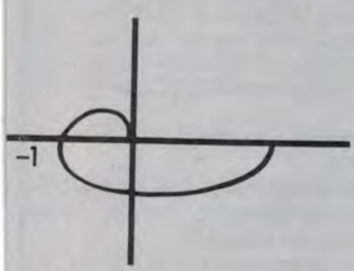


Fig. 15b: The Nyquist contour for a three pole stable system.

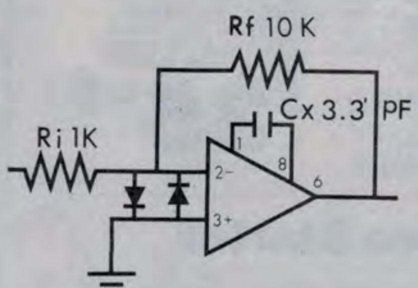


Fig. 16: The use of the LM308AH as a compensated operational amplifier, Gain = 10.

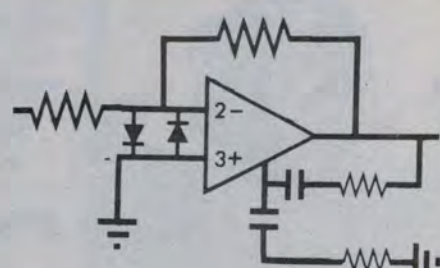


Fig. 17: Use of the LM725AH in a two pole compensation scheme, Gain = 10.