Bandwidth Estimation Techniques

1.0 Introduction

Finding the –3 dB bandwidth of an arbitrary linear network can be a difficult problem in general. Consider, for example, the standard recipe for computing bandwidth:

1) Derive the input-output transfer function (using node equations, for example)
2) Set $s = j\omega$;
3) Find the magnitude of the resulting expression;
4) Set the magnitude = $1/\sqrt{2}$ of the "midband" value; and
5) Solve for ω

It doesn't take a great deal of insight to recognize that explicit computation (by hand) of the –3 dB bandwidth using this method is generally impractical for all but the simplest systems. In particular, the order of the denominator polynomial obtained in step 1 above is equal to the number of poles (natural frequencies), which in turn equals the number of degrees of freedom (measured, say, by the number of initial conditions one may independently specify), which in turn equals the number of independent energy storage elements (e.g., *L* or *C*), which in turn can be as large as the number of energy storage elements (phew!). Thus, a network with *n* capacitors might require the equivalent of finding the roots of an *n*th-order polynomial. If *n* exceeds just four, no algebraic closed form solution exists. Even if $n = 2$, it might be labor-intensive to obtain the final numerical result.

Now, machine computation is cheap and getting cheaper all the time, so perhaps the analysis of networks doesn't present much of a problem. However, we are interested in developing *design* insight so that if a simulator tells us that there is a problem, we have some idea of what to do about it. We therefore seek methods that are reasonably simple to apply, yet conveys the desired insight, even if it yields answers that might be approximate. Simulators can then be used to provide final quantitative verification.

Two such approximate methods are open- and short-circuit time constants. The former provides an estimate of the high-frequency rolloff while the latter yields an estimate of the low-frequency rolloff point. These methods are valuable because they identify which elements are responsible for the bandwidth limitation. This information alone is often sufficient to suggest what modifications should be tried next.

2.0 The Method of Open-Circuit Time Constants

The method of open-circuit time constants (OCτ's), also known as zero value time constants, was developed in the mid-1960's at MIT. As we shall see, this powerful technique allows us to estimate the bandwidth of a system almost by inspection, and sometimes with surprisingly good accuracy. More important, and unlike typical circuit simulation programs, *OCτ's identify which elements are responsible for bandwidth limitations*. The great value of this property in the **design** of amplifiers hardly needs expression.

To begin development of this method, let us consider all-pole transfer functions only. Such a system function may be written as follows:

$$\frac{V_o(s)}{V_i(s)} = \frac{a_o}{(\tau_1 s + 1)\ (\tau_2 s + 1) \ldots (\tau_n s + 1)} \tag{1}$$

where the various time constants may or may not be real.

Multiplying out the terms in the denominator leads to a polynomial we shall express as:

$$b_n s^n + b_{n-1} s^{n-1} + \ldots + b_1 s + 1 \tag{2}$$

where the coefficient b_n is simply the product of all of the time constants and b_1 is the sum of all of the time constants. (In general, the coefficient of the s^j term is computed by forming all unique products of the n time constants taken j at a time and summing all $n!/j!(n\text{-}j)!$ such products.)

We now assert that, near the –3 dB frequency, the first-order term typically dominates over the higher-order terms so that (perhaps) to a reasonable approximation, we have:

$$\frac{V_o(s)}{V_i(s)} \approx \frac{a_o}{b_1 s + 1} = \frac{a_o}{\left(\sum_{i=1}^{n} \tau_i\right) s + 1} \tag{3}$$

The bandwidth of our original system in radian frequency as estimated by this first-order approximation is then simply the reciprocal of the effective time constant:

$$\omega_h \approx \frac{1}{b_1} = \frac{1}{\sum_{i=1}^{n} \tau_i} = \omega_{h,\,est} \tag{4}$$

Before proceeding further, we should consider the conditions under which our neglect of the higher-order terms is justified, so let us examine the denominator of the transfer function near our estimate of ω_h. For the sake of simplicity, we start with a second-order polynomial with purely real roots.

Now, at our estimated –3 dB frequency, the original denominator polynomial is:

$$-\tau_1 \tau_2 \omega_{h,\,est}^2 + j\,(\tau_1 + \tau_2)\,\omega_{h,\,est} + 1 \tag{5}$$

Note that the magnitude of the second term is unity (why?). As a consequence, both

$$\tau_1 \omega_{h,\,est} \tag{6}$$

and

$$\tau_2 \omega_{h,\,est} \tag{7}$$

must have magnitudes no greater than one. Thus the product of these terms (which in turn is equal to the magnitude of the leading term of the polynomial) must be small compared to the magnitude of the second (first-order) term. The worst case occurs when the two time constants are equal, and even then the second-order term is only one-fourth as large as the first-order term. Extending these arguments to polynomials of higher order reveals that the estimate of the bandwidth based simply on the coefficient b_1 is generally reasonable since the first-order term generally does dominate the denominator. Furthermore, the bandwidth estimate is usually conservative in the sense that *the actual bandwidth will almost always be at least as high as estimated by this method.*

So far, all we've done is show that a first-order estimate of the bandwidth is possible if one is given the sum of the pole time-constants (= b_1). Alas, such information is almost never available, apparently casting serious doubt on the value of our entire enterprise since the whole point was to avoid things such as direct computation of the pole locations in the first place.

Fortunately, it is possible to relate the desired time-constant sum, b_1, to (more or less) easily computed network quantities. The new recipe is thus as follows: Consider an arbitrary linear network comprising only resistors, sources (dependent or independent), and m capacitors. Then:

1) Compute the effective resistance R_{jo} facing each jth capacitor with all of the other capacitors removed (open-circuited, hence the name);

2) Form the product $\tau_{jo} = R_{jo}C_j$ (the subscript "o" refers to the open-circuit condition) for each capacitor;

3) Sum all m such "open-circuit" time constants.

Remarkably, the sum of open-circuit time constants formed in step 3) is in fact precisely equal to the sum of the pole time constants, b_1, a result proved by Adler (see reference at end of this chapter). Thus, at last, we have:

$$\omega_{h,\,est} = \frac{1}{\sum_{j=1}^{m} R_{jo}C_j} \tag{8}$$

2.1 Some Observations and Interpretations

The method of OCτ's is relatively simple to apply because each time constant calculation involves the computation of just a single resistance, although one must be wary of the impedance-modifying potential of dependent sources (such as the transconductance of a transistor model). In any event, the amount of computation required is typically substantially (indeed, often fantastically) less than that needed for an exact solution.

The greatest value of the technique lies in its identification of those elements implicated in bandwidth limitations, that is, those elements whose associated open-circuit time constants dominate the sum. This knowledge can guide the designer to effect appropriate modifications to circuit values or even suggest wholesale topological changes. In contrast, SPICE and other typical simulators only provide a numerical value for the bandwidth while conveying little or nothing about what the designer can do to alter the performance in a desired direction.

The origin of this property of OCτ's may be regarded intuitively as follows. *The reciprocal of each jth open-circuit time constant is the bandwidth that the circuit would exhibit if that jth capacitor were the only capacitor in the (all-pole) system*. Thus each time constant represents a *local bandwidth degradation term*. The method of OCτ's then states that the linear combination of these individual, local limitations yields an estimate of the total bandwidth. The value of OCτ's derives directly from the identification and approximate quantification of the local bandwidth bottlenecks.

2.2 Accuracy of OCτ's

One must be careful not to place too much faith in the ability of OCτ's to provide accurate estimates of bandwidth in all cases. This situation should hardly be surprising in view of the rather brutal truncation to first order of the denominator polynomial. However there are numerous conditions under which OCτ estimates are fairly reasonable, as seen in Section 2.0, for example.

It should be clear that an OCτ bandwidth estimate is in fact exact for a first-order network since *no* truncation of terms is involved there. Not surprisingly then, the OCτ estimate will be quite accurate if a network of higher order happens to be dominated by one pole (that is, one pole is much lower in frequency than all of the other poles). There are many systems of practical interest, such as operational amplifiers, that are designed to have a dominant single pole, and thus for which OCτ estimates are quite accurate.

Unfortunately, there are so many other conditions under which OCτ's give poor estimates that some caveat is necessary. For example, complex poles quite commonly arise (intentionally or otherwise) in the design of wideband multistage amplifiers. Often the physical origin of these complex poles can be traced to the interaction of the primarily capacitive input impedance of one stage (as in a common-source configuration) with the inductive component of the output impedance of source followers.

The reason that the presence of complex poles upsets OCτ estimates is as follows: The coefficient b_1 is the sum of the pole time constants, and thus ignores the imaginary parts of complex poles since they must appear in conjugate pairs. However, the true bandwidth of, say, a two-pole system does depend on both the real and imaginary parts. As a result, gross errors in OCτ estimates are not uncommon if complex poles are present in abundance.

The nature and magnitude of the problem are best illustrated with an example. Consider the simplest possible case, a 2-pole transfer function:

$$H(s) = \frac{1}{\frac{s^2}{\omega_n^2} + \frac{2\zeta s}{\omega_n} + 1} \tag{9}$$

The OCτ bandwidth estimate is found from the coefficient of the *s* term:

$$\omega_h \approx \frac{\omega_n}{2\zeta} \tag{10}$$

while it may be shown that the actual bandwidth is:

$$\omega_h = \omega_n\left[1 - 2\zeta^2 + \sqrt{2 - 4\zeta^2 + 4\zeta^4}\right]^{0.5} \tag{11}$$

In this particular case, we see that the OCτ estimate predicts monotonically increasing bandwidth as the damping ratio ζ approaches zero, while the actual bandwidth asymptotically approaches about $1.55\omega_n$. Thus, it is possible for OCτ estimates to be *optimistic*, in this case, wildly so. At a ζ of about 0.35, OCτ estimates are correct, and for any higher damping ratio, OCτ's are pessimistic. Fortunately, the poles of amplifiers are usually designed to have relatively high damping ratios (to control overshoot and ringing in the step response, and minimize peaking in the frequency response), so for most practical situations, OCτ estimates are pessimistic.

Since it is generally not possible to tell by inspection of a network if complex poles are going to be an issue, one must always keep in mind that the primary value of OCτ's is in identifying those portions of a circuit that control the bandwidth, rather than in providing accurate bandwidth estimates. Circuit simulators will take care of the latter task.

2.3 Other Important Considerations

Although application of open-circuit time constants is reasonably straightforward, there are one or two final issues that deserve consideration. An extremely important idea is that not all capacitors in a network belong in the OCτ calculations. For instance, fairly large coupling capacitors are frequently used in discrete designs to connect the output of one stage to the input of the next without the bias point of one stage upsetting that for the other. Blind application of the OCτ method would lead one to conclude erroneously that the larger this capacitor, the lower the bandwidth (time constants here often correspond to the audio range, suggesting that large bandwidths are not possible). Fortunately, real circuit behavior defies these implications.

The problem stems from the presence of zeros associated with the coupling capacitors. Recall that the assumed form for the system function consists of poles only. Since all zeros are thus assumed at infinitely high frequency, it is hardly surprising that the presence of low-frequency zeros confounds our estimates of bandwidth.

The solution is to *pre-process the network* prior to application of the OCτ method. That is, recognize that the coupling capacitors are effectively short circuits relative to the impedances around them at frequencies near the upper bandwidth limit. Thus, **one must apply OCτ's only to models that are appropriate to the high-frequency regime**.

While it is usually obvious which capacitors are to be ignored (considered short circuits), there are occasions where one is not so sure. In these cases, a simple thought experiment usually suffices to decide the issue. Now OCτ's are concerned only with those capacitors that limit high-frequency gain. As a consequence, the removal (that is, the open-circuiting) of a capacitor that belongs in the OCτ calculation should result in an increase in high-frequency gain. The test, therefore, is to consider exciting the network at some high frequency and imagining what would happen to the gain if the capacitor in question were open-circuited. If the gain would go up, the capacitor belongs in the OCτ calculation since we infer from the thought experiment that the capacitor does indeed limit the high-frequency gain. If the gain would not change (or even decrease, as in the coupling-capacitor case) upon removal, that capacitor should probably be short-circuited. The necessary conclusions can usually be reached without taking pencil to paper.

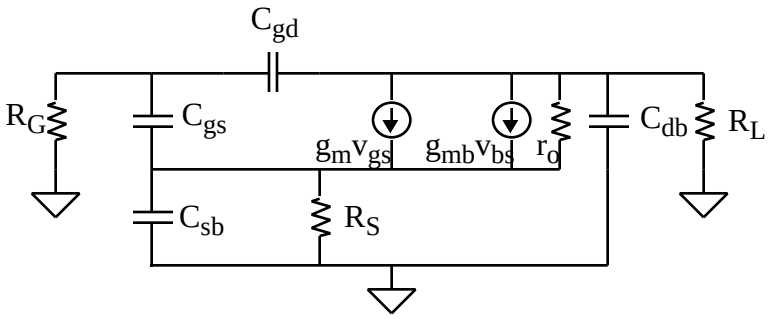
One last issue that deserves some attention concerns the relationship between the individual open-circuit time constants and the time constants of the poles. We have asserted (without formal proof) only that the *sums* of these time constants are equal to each other. **One must therefore resist the temptation to equate an open-circuit time constant with a corresponding pole location**. Indeed, the number of poles may not even equal the number of capacitors (consider the trivial case of two capacitors in parallel). Since the number of open-circuit time constants and the number of poles may be unequal, one clearly cannot expect each OCτ to equal the time constant of a pole in general.

2.4 Some Useful Formulas

When computing open-circuit time constants for transistor amplifiers, care is required because the feedback action of the g_m generator modifies resistances. As a consequence, one should explicitly apply a test source (choose the type that will most directly allow computation of v_{gs}) to derive expressions for the effective resistances. To illustrate a general method, we will derive formulas for the resistances facing C_{gs} and C_{gd}. To simplify the derivations, we will ignore body-effect and output resistance. However, complete formulas including both of those effects are provided at the end of this chapter for reference. Derivations are (surprise!) left as an exercise for the reader.

Consider the following comprehensive model for a MOSFET with external resistances added in series with each terminal (except for the substrate, which is our ground reference):

FIGURE 1. Incremental model for open-circuit resistance calculations



Although this model explicitly includes the back-gate transconductance g_{mb} and output resistance r_o, we will not use them in this first set of derivations. In all of the SPICE runs that follow, however, the complete model will be used.

First, let's compute the resistance facing C_{gs}. Applying a test voltage source v_t (since that choice directly fixes the value of v_{gs}), we exploit superposition (once v_{gs} is known, we may treat the transconductance generator from that point on as an independent current source of value $g_m v_t$) to obtain, when all is said and done:

$$i_t = \frac{v_t}{R_G + R_S} + \frac{g_m R_S v_t}{R_G + R_S} \tag{12}$$

so that the equivalent resistance facing C_{gs} is given by

$$r_{1o} = \frac{R_G + R_S}{1 + g_m R_S} \tag{13}$$

So, r_{1o} is the sum of the resistors divided by $1+ g_m R_S$.

Now, to compute the resistance facing C_{gd}, use a test current source (you may try a test voltage source, but you'll regret it). The intervening algebra is a little involved, but the resistance may be expressed in the following mnemonic form:

$$r_{2o} = r_{left} + r_{right} + g_{m,eff} r_{left} r_{right} \tag{14}$$

where r_{left} is the resistance between the left terminal and ground, r_{right} is defined between the right terminal and ground, and $g_{m,eff}$ is the *effective* transconductance (defined as the ratio of current from the dependent current source to the voltage between the left terminal and ground). For our model, we have

$$r_{right} = R_L \tag{15}$$

$$r_{left} = R_G \tag{16}$$

$$g_{m,\,eff} = g_m \frac{1}{1 + g_m R_S} \tag{17}$$

After a little practice, these equations will help you to zip through bandwidth calculations.

2.5 A Design Example

We've seen that the method of open-circuit time constants promises to simplify design while conveying important insight. Let's now carry out an actual design to see if it lives up to this promise.

Suppose we want an amplifier with the following specifications:

Voltage gain magnitude: > 18dB (or about a factor of 8)

–3dB Bandwidth: > 450MHz (implies a maximum OCτ sum of ≈ 350ps)

Furthermore, assume that we must meet these specifications with a **2kΩ source resistance** driving the input and a **1pF capacitive load** on the output. In a truly practical design, there would usually be additional specifications, such as maximum allowed power consumption, dynamic range, etc., but we'll keep the design space restricted for now.

Further suppose that we are to meet these specifications with transistors from the 0.5μm (drawn) technology described in the handout on MOS physics. To simplify the process, let us use just one size of device, and just one bias current for all transistors. In a better design, of course, one would generally use different biases and different device sizes, but we need to impose some arbitrary constraints if we are to complete our task in finite space!

Arbitrarily selecting a per-transistor bias current of 3mA, a 150μm wide NMOS transistor in this process technology has the following approximate element values when operating in saturation:

C_{gs} = 220fF

C_{sb} = 130fF

C_{gd} = 45fF

C_{db} = 90fF

r_o = 2kΩ

g_m = 12mS

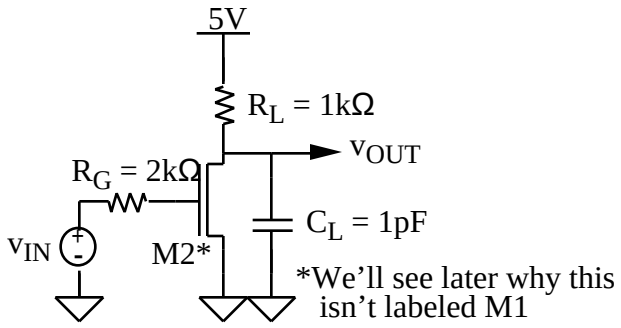
g_{mb} = 1.8mS

Even though some of the capacitances are bias voltage-dependent, we will assume that they are constant at the values shown.

The only way to start a design is, well, to start. Put *something* (almost anything) down. *It's easier to edit than to create,* so virtually any reasonable initial condition is acceptable. A few simple calculations will let you know fairly quickly if you're on the right track, and you can always obsess later about the particulars. So, let's start with the common-source configuration (after all, it provides voltage gain, and has a moderately high input impedance). In all that follows, we'll neglect the details of how biasing is taken care of (since we're focusing on dynamic performance issues), but be aware that any practical design *must* include careful attention to the bias problem.

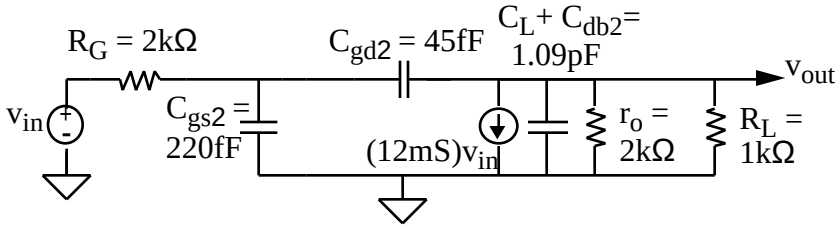
Recalling that (neglecting body effect) the voltage gain from gate to drain of a basic CS amplifier is $-g_m R_L$, and being mindful that we do have to worry about gain loss from the additional loading by the transistor's own output resistance, let's shoot for a $g_m R_L$ product that is 50% larger than the gain specification. With the resulting choice of 12 for $g_m R_L$, we find that we must select R_L = 1kΩ. Our circuit then appears as follows:

FIGURE 2. First-cut design (biasing not shown)



The corresponding incremental model is then:

FIGURE 3. Incremental model of first-pass design



Note that the source-bulk potential is zero, so the back-gate transconductance contributes zero current, and the source-bulk capacitance is shorted out.

From the model, it's easy to see that the low-frequency voltage gain just barely meets specifications:

$$A_V = -g_{m2}(R_L \| r_o) = -8 \tag{18}$$

Now, let's estimate the bandwidth to see just how bad the news there is:

$$\tau_{gs2} = C_{gs2} r_{gs2} = (220fF)(R_G) = 440ps \tag{19}$$

$$\tau_{gd2} = C_{gd2} r_{gd2} = (45fF)(r_{left} + r_{right} + g_{m2} r_{left} r_{right}) \approx 840ps \tag{20}$$

$$\tau_{db2} = C_{db2} r_{db2} = C_{db2}(R_L \| r_o) = 60ps \tag{21}$$

$$\tau_L = C_L(R_L \| r_o) = 670ps \tag{22}$$

$$BW \approx \frac{1}{(0.44ns + 0.84ns + 0.06ns + 0.67ns)} \approx 500Mrps \tag{23}$$

We see that our bandwidth is about 79MHz (SPICE says it's actually somewhat higher, 86MHz), so we're quite a bit shy of our goal of 450MHz.

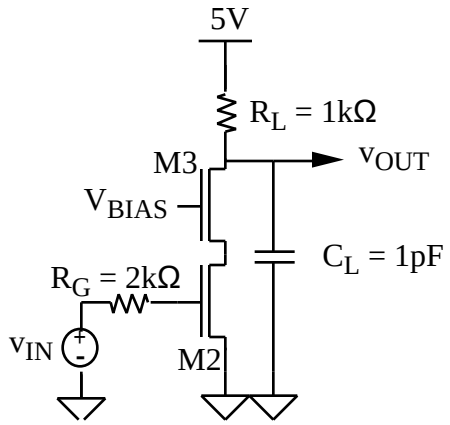
Now, who's the big culprit? From our four calculated time constants, we see that there are two similar-sized ones. The larger of these is associated with the drain-gate capacitance, C_{gd2}, even though that capacitance is numerically the smallest, because its effect is Miller-multiplied by the gain. So, if we are to improve bandwidth, we must figure out how to mitigate the Miller effect problem.

Recall that the Miller effect arises from connecting a capacitance across two nodes that have between them an inverting voltage gain. So, one possible solution would be to distribute the gain among *N* stages, rather than try to get all of our gain out of one stage. You are encouraged to explore this promising option independently.

Another possibility is to isolate (somehow) the offending capacitance so that it no longer appears across a gain stage. We will pursue this approach to attempt to get all of our gain in one stage.

The *cascode* amplifier eliminates the Miller effect precisely by performing this isolation. So, now consider the following circuit:

FIGURE 4. Second pass: cascode amplifier



As usual, the value of V_{BIAS} is not terribly critical. It just has to be high enough to guarantee that M2 stays in saturation, and low enough to guarantee that M3 stays in saturation. A value of 2.3V satisfies these conditions comfortably in this particular case, and is the value used in the SPICE simulations.

Before plunging mindlessly ahead into a pile of computations, let's think about how this circuit works. The input voltage is converted into an output current by transistor M2 (i.e., M2 is a transconductor). Transistor M3 merely transfers this current to the output load resistor. Now, the output is at the drain of M3, while the input is at the gate of M2, and there is no capacitance directly across the two nodes. Hence, there is very little Miller multiplication, and we expect a significant improvement in bandwidth.[1]

The isolation provided by cascoding also has a beneficial effect on the gain. Voltage changes at the drain of M3 have hardly any effect on the drain current of M2. Hence, the output current changes little. An equivalent statement is that the output resistance has increased. In this particular case, the increase is enough to eliminate the effect of r_o for all practical purposes. We would therefore expect a gain very near –12, and SPICE simulations show that it is about –11. If this excess gain holds up as the design evolves, it may be traded off for improved bandwidth, if needed or desired.

Returning to open-circuit time constant estimates of bandwidth, draw the model corresponding to this cascode connection, and calculate the resistance facing each capacitance. Out of laziness, there will be no more incremental models from here on out, so you're on your own now ("some assembly may be required"):

$$\tau_{gs2} = C_{gs2} r_{gs2} = (220fF)(R_G) = 440ps \text{ (unchanged)} \tag{24}$$

$$\tau_{gd2} = C_{gd2} r_{gd2} \approx (45fF)\left(R_G + \frac{1}{g_{m3}} + g_{m2} R_G \frac{1}{g_{m3}}\right) \approx 140ps \text{ (better!!)} \tag{25}$$

1. To complete the argument, note that the gain between the gate and drain of M2 is –1, so that C_{gd2} is not multiplied very much at all.

This last equation is a bit approximate because we are neglecting the effect of g_{mb3} and r_o in this calculation (as well several to follow). Because the body effect degrades transconductance, we are somewhat underestimating the true effective resistance ("r_{right}," specifically). A more accurate calculation shows a 0.20ns time constant. The error for this iteration is thus negligible. In any case, the error introduced by neglecting these effects tends to offset the typical pessimism of open-circuit time constants.

$$\tau_{gs3} = C_{gs3}r_{gs3} \approx (220fF)\frac{1}{g_{m3}} = 18ps\ (\text{new}) \tag{26}$$

$$\tau_{gd3} = C_{gd3}r_{gd3} = (45fF)\ (R_L) \approx 45ps\ (\text{new}) \tag{27}$$

$$\tau_{sb3} = C_{sb3}r_{sb3} \approx (130fF)\left(\frac{1}{g_{m3}}\right) \approx 11ps\ (\text{new}) \tag{28}$$

$$\tau_{db3} = C_{db3}r_{db3} \approx (90fF)\ (R_L) \approx 90ps\ (\text{new}) \tag{29}$$

$$\tau_{db2} = C_{db2}r_{db2} \approx (90fF)\left(\frac{1}{g_{m3}}\right) \approx 8ps\ (\text{better}) \tag{30}$$

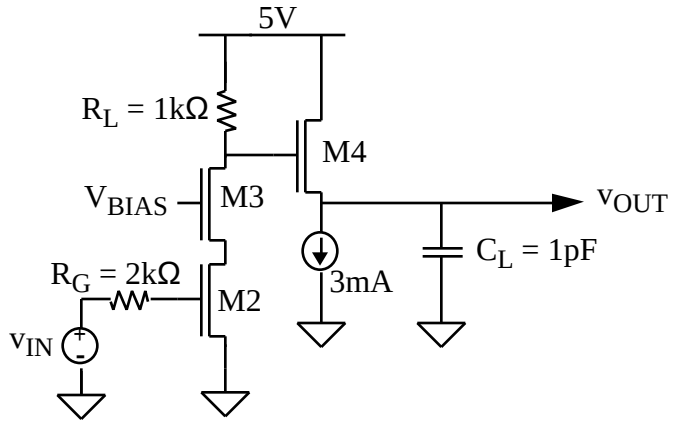
$$\tau_L = C_L r_L = C_L R_L = 1ns\ (\text{worse!}) \tag{31}$$

$$BW \approx \frac{1}{(1.75ns)} \approx 570Mrps \tag{32}$$

With a new (estimated) bandwidth of about 90MHz (SPICE says 109MHz), we can see that the cascode connection has given us a substantial improvement in bandwidth, but we still have a long way to go.

Looking at the new big offender, we see that it involves the load capacitance, C_L. Driving it with such a high (1kΩ) resistance is obviously the problem, so we should be able to reduce that time constant to a small value with a source follower:

FIGURE 5. Third pass: cascode amplifier w/ output source follower



Again, we'll ignore biasing details. Just assume that we put a current source (or just a plain old resistor) in the source leg of M4 to bias it to 3mA. For purposes of time constant calculations, we'll see that the resistance of the bias network is easily made negligible, so it doesn't really matter what we assume.

The source follower does not quite have unity gain because there is a capacitive voltage division between C_{gs4} and C_{sb4}. A careful calculation, verified by SPICE, reveals that the gain has dropped from –11 to –9.5. Fortunately, this value is still in excess of the desired value.

Calculating the time constants for this iteration yields the following list:

$$\tau_{gs2} = 440ps \text{ (unchanged)} \tag{33}$$

$$\tau_{gd2} = 140ps \text{ (unchanged)} \tag{34}$$

$$\tau_{db2} = 8ps \text{ (unchanged)} \tag{35}$$

$$\tau_{sb3} = 11ps \text{ (unchanged)} \tag{36}$$

$$\tau_{gs3} = 18ps \text{ (unchanged)} \tag{37}$$

$$\tau_{gd3} = 45ps \text{ (unchanged)} \tag{38}$$

$$\tau_{db3} = 90ps \text{ (unchanged)} \tag{39}$$

$$\tau_{gd4} = C_{gd4}r_{gd4} \approx (45fF)(R_L) \approx 45ps \text{ (new)} \tag{40}$$

$$\tau_{gs4} = C_{gs4}r_{gs4} \approx (220fF)\left(\frac{1}{g_{m4}}\right) \approx 18ps \text{ (new)} \tag{41}$$

This last equation is a bit more approximate than usual because of the neglect of g_{mb4} with a 1kΩ driving resistance (the correct value is about 0.045ns). A more careful derivation shows that the resistance in Eqn. 41 should be multiplied by about $(1 + g_{mb4}R_L)$, so one may freely neglect this factor only if $g_{mb}R$ is much smaller than unity. Fortunately, this particular time constant is not dominant in this case, so the large percentage error in this term has an insignificant effect on the overall time constant sum.

Continuing:

$$\tau_{sb4} = C_{sb4}r_{sb4} \approx (130fF)\left(\frac{1}{g_{m4}}\right) \approx 11ps \text{ (new)} \tag{42}$$

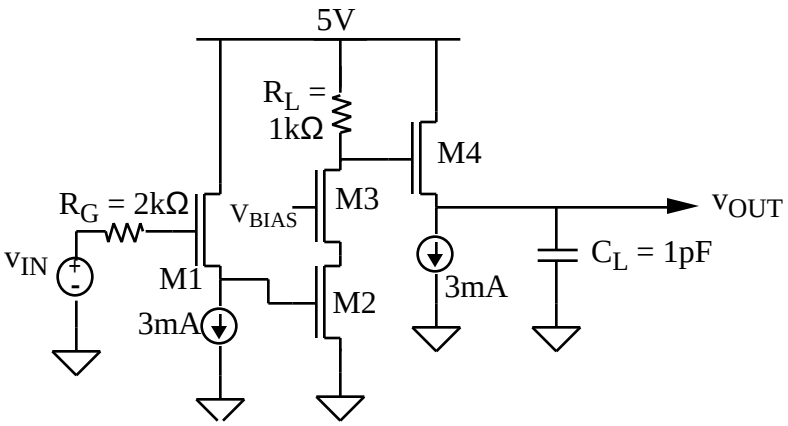
$$\tau_L = C_L r_L \approx (1pF)\left(\frac{1}{g_{m4}}\right) \approx 80ps \text{ (better!!)} \tag{43}$$

$$BW \approx \frac{1}{(0.906ns)} \approx 1.1Grps \tag{44}$$

So, now we're up to about 175MHz (SPICE says 222MHz); we "only" need to pick up another factor of about two in bandwidth.

Looking over our latest list of time constants, we see that τ_{gs2} dominates by far because of the 2kΩ source resistance. One obvious remedy is to add an input buffer to reduce the resistance driving the gate-source capacitance of M2:

FIGURE 6. Fourth pass: cascode amplifier w/ two source followers



A re-computation of the gain reveals that the slight attenuation of the added source follower takes us down to a gain of –8, leaving us with no more margin.

With these changes, we expect to get pretty close to the desired bandwidth because τ_{gs2} is about half the total, and we can probably drop it to near zero. Recomputing the time constants, we get:

$$\tau_{db2} = 8ps \text{ (unchanged)} \tag{45}$$

$$\tau_{sb3} = 11ps \text{ (unchanged)} \tag{46}$$

$$\tau_{gs3} = 18ps \text{ (unchanged)} \tag{47}$$

$$\tau_{gd3} = 45ps \text{ (unchanged)} \tag{48}$$

$$\tau_{db3} = 90ps \text{ (unchanged)} \tag{49}$$

$$\tau_{gd4} = 45ps \text{ (unchanged)} \tag{50}$$

$$\tau_{gs4} = 18ps \text{ (unchanged)} \tag{51}$$

$$\tau_{sb4} = 11ps \text{ (unchanged)} \tag{52}$$

$$\tau_L = 80ps \text{ (unchanged)} \tag{53}$$

$$\tau_{gs1} = C_{gs1}r_{gs1} \approx (220fF)\left(\frac{1}{g_{m1}}\right) \approx 18ps \text{ (new)} \tag{54}$$

Again, this last calculation is in error because $g_{mb1}R_G$ is 3.6, so the time constant really ought to be multiplied by (1 + 3.6) = 4.6, to yield about 83ps. A more careful calculation that also takes r_{o1} into account reveals that the time constant here is in fact about 86ps.

$$\tau_{gd1} = C_{gd1}r_{gd1} \approx (45fF)\ (R_G) \approx 90ps \text{ (new)} \tag{55}$$

$$\tau_{sb1} = C_{sb1}r_{sb1} \approx (130fF)\left(\frac{1}{g_{m1}}\right) \approx 11ps \text{ (new)} \tag{56}$$

$$\tau_{gs2} = C_{gs2}r_{gs2} \approx (220fF)\left(\frac{1}{g_{m1}}\right) \approx 18ps \text{ (better!)} \tag{57}$$

$$\tau_{gd2} = C_{gd2}r_{gd2} \approx (45fF)\left[\frac{1}{g_{m1}} + \frac{1}{g_{m3}} + g_{m2}\left(\frac{1}{g_{m1}}\right)\left(\frac{1}{g_{m3}}\right)\right] \approx 11ps \text{ (better!)} \tag{58}$$

$$BW \approx \frac{1}{(0.474ns)} \approx 2.1Grps \tag{59}$$

The estimated bandwidth has now increased to about 340MHz. Owing to the conservative nature of the estimate, it is reasonable to expect the actual bandwidth to be quite close to the goal. SPICE simulations show that, in fact, the bandwidth is about 540MHz, well in excess of the desired value. If desired, some of this excess bandwidth could be exchanged for increased gain.

Suppose, though, that SPICE were to confirm your worst fears, and you find that the amplifier just doesn't quite make it. Are there any other modifications that you could try?

The answer, of course, is yes. One option was passed over earlier: distribute the gain among several stages. Using two or more stages, it would be a trivial matter to beat the bandwidth specification by a handy margin.

These tricks are by no means the only ones, and we will spend a considerable amount of time exploring some important alternative methods in EE314. However, to whet your appetite and stimulate some thinking, here are some vague allusions to other possibilities.

The method of open-circuit time constants assumes an all-pole transfer function, and gives more accurate answers if all the poles are real. Consider the effect of purposefully violating these assumptions by allowing zeros and/or complex poles. Careful placement of zeros (anti-poles) or complex poles will extend the bandwidth, although the frequency response may no longer be monotonic. One way to form complex poles is through feedback (just think of a two-pole root locus, for example), or through the resonance of inductors (real or

synthetic) with capacitors. The surprisingly large bandwidth of the last circuit in the chain of design iterations is largely due to the formation of complex poles arising from the interaction of the gate-source capacitance of M2 with the *inductive* output impedance of source follower M1.

Zeros can be formed, for example, by capacitors in parallel with source bypass resistors. As we'll see, a judicious choice of capacitor value can cause this zero to cancel a bandwidth limiting pole.

Another possibility, made most practical in differential systems, is to use positive feedback to generate negative capacitances. These negative capacitances can cancel positive ones to yield bandwidth increases. Of course, there is a chance of unstable behavior that must be carefully watched, but this method, called *neutralization*, can yield useful bandwidth improvements. We will turn to a detailed examination of these themes very soon.

2.6 Summary of Open-Circuit Time Constants

We have seen that the method of open-circuit time constants is an extremely valuable tool for designing amplifiers for good dynamic performance mainly because of its ability to identify the problem areas of the circuit. Because of the tremendous insights gained with extremely modest effort we are generally willing to overlook its quantitative limitations, such as the often highly conservative nature of the estimated bandwidth. As long as we take care to apply the method only to models that apply to the high-frequency regime, we are assured of reasonable answers.

As a parting remark, it should be noted that the influence of inductances can be incorporated as well into the method, although with generally unsatisfactory results for reasons that will be explained shortly.

The most intuitive way to understand how one incorporates the effect of inductances on bandwidth is to recall that each time constant term represents an individual, local contribution to the bandwidth limitation; we treat the system at each step of the calculation as if that *j*th reactive element were the sole one. So evidently one treats all of the inductors as *short* circuits when computing the appropriate effective resistances. The L/R time constants are then added to the various RC terms to yield the grand total from which the bandwidth is estimated.

Having said all of this, the presence of explicit inductances and capacitances almost guarantees the formation of troublesome complex pole pairs, often causing the method to yield gross underestimates of bandwidth. This difficulty is exacerbated by the common occurrence of finite zeros. Furthermore, the parasitic inductances in a circuit are often much more difficult to estimate accurately than the capacitances. As a consequence, inductors are rarely taken into consideration. However, one should be aware that while small values of effective resistance minimize the time constant due to capacitances, they tend to maximize the time constant due to the inductances. Hence, if one goes to extremes in reducing the various *R*'s in the quest for ever greater bandwidth, there is often a point not only of diminishing returns but even of reversals. Typically, such considerations become impor-

tant as one pursues bandwidths exceeding, say, 20-50 MHz in discrete designs, where stray inductances of under a few nanohenries are almost impossible to achieve.

Concluding, the method of open-circuit time constants is an indispensable guide in the design of amplifiers. With it one can design intelligently and confidently to satisfy a given bandwidth specification. To be sure, the method has its quantitative shortcomings, but the valuable intuition provided is more than sufficient compensation.

3.0 The Method of Short-Circuit Time Constants

3.1 Introduction

We've already seen how the method of open-circuit time constants allows us to estimate the high-frequency –3 dB point of an arbitrarily complex system by decomposing the bandwidth computation into a succession of first-order calculations. Each of the time constants represents a local bandwidth degradation term, and the sum of these individual degradation terms equals the reciprocal of the overall bandwidth. As we saw, open-circuit time constants are valuable because they identify which elements limit the bandwidth.

Now suppose that, instead of estimating the high frequency –3 dB point, we wanted to find the *low* frequency –3 dB point of an AC-coupled system. How would we calculate how large the coupling capacitors have to be to achieve a specified low-frequency breakpoint? Fortunately, we may invoke a procedure that is analogous to the method of open-circuit time constants. This dual technique is known as the method of short-circuit time constants.

3.2 Background of the Method of Short-Circuit Time Constants

In the method of open-circuit time constants, we assumed that the zeros of the network were all at infinitely high frequency, so that the transfer function consisted only of poles. In the case of short-circuit time constants, we instead assume that all of the zeros are at the *origin*, and that there are as many poles as zeros. Thus, the corresponding system function may be written as follows:

$$\frac{V_o(s)}{V_i(s)} = \frac{ks^n}{(s+s_1)(s+s_2)\ldots(s+s_n)} \tag{60}$$

where the various pole frequencies may or may not be real, and k is simply a constant to fix up the scale factor.

Multiplying out the terms in the denominator leads to a polynomial we shall express as:

$$s^n + b_1 s^{n-1} + \ldots + b_{n-1}s + b_n \tag{61}$$

where the coefficient b_1 is the sum of all of the pole frequencies, and b_n is the product of all of the pole frequencies. (In general, the coefficient of the s^j term is computed by form-

ing all unique products of the n frequencies taken j at a time and summing all $n!/j!(n-j)!$ such products.)

We now assert that, near the low frequency –3 dB breakpoint, the higher-order terms dominate the denominator so that we obtain:

$$\frac{V_o(s)}{V_i(s)} \approx \frac{ks^n}{s^n + b_1 s^{n-1}} = \frac{ks}{s + \sum_{i=1}^{n} s_i} \tag{62}$$

The low frequency –3 dB point of our original system in radian frequency as estimated by this first-order approximation is then simply the sum of the pole frequencies:

$$\omega_l \approx b_1 = \sum_{i=1}^{n} s_i = \omega_{l,\,est} \tag{63}$$

Before proceeding further, we should consider the conditions under which our neglect of the lower-order terms is justified, so let us examine the denominator of the transfer function near our estimate of ω_l. For the sake of simplicity, we consider a second-order polynomial with purely real roots, s_1 and s_2.

Now, at our estimated –3 dB frequency, the original denominator polynomial is:

$$-\omega_{l,\,est}^2 + j\omega_{l,\,est}(s_1 + s_2) + s_1 s_2 \tag{64}$$

Substituting our expression for the estimated –3 dB point, we obtain:

$$-\left[s_1^2 + s_2^2 + 2s_1 s_2\right] + j\left[s_1^2 + s_2^2 + 2s_1 s_2\right] + s_1 s_2 \tag{65}$$

Clearly, the last term is small compared with the magnitudes of the other terms. Thus the neglect of all but the two highest-order terms involves little error. The worst case occurs when the two pole frequencies are equal, and even then the error is not terribly large. Extending these arguments to polynomials of higher order reveals that the estimate of the low-frequency cutoff based simply on the coefficient b_1 is generally reasonable since the higher order terms do in fact dominate the denominator. Furthermore, the low-frequency cutoff estimate is conservative in the sense that *the actual cutoff frequency will almost always be as low as or lower than estimated by this method*

So far, all we've done is show that a first-order estimate of the bandwidth is possible if one is given the sum of the pole frequencies (= b_1). Of course, if we knew the pole frequencies, we could compute this sum directly. Fortunately, as was the case with open-circuit time constants, it is possible to relate the desired pole-frequency sum, b_1, to (more or less) easily computed network quantities.

The recipe is thus as follows: Consider an arbitrary linear network comprising only resistors, sources (dependent or independent), and *m* capacitors. Then:

1) Compute the effective resistance R_{js} facing each jth capacitor with all of the other capacitors *short*-circuited (the subscript "*s*" refers to the short-circuit condition for each capacitor);

2) Compute the "short-circuit frequency" $1/(R_{js}C_j)$;

3) Sum all *m* such short-circuit frequencies.

The sum of the reciprocal short-circuit time constants formed in step 3) turns out to be precisely equal to the sum of the pole frequencies, b_1. Thus, at last, we have:

$$\omega_{l,\,est} = \sum_{j=1}^{m} \frac{1}{R_{js}C_j} \tag{66}$$

3.3 Some Observations and Interpretations

The method of SCτ's is relatively simple to apply for precisely the same reasons that OCτ's are easy to apply, namely, each time constant calculation involves the computation of just a single resistance, although one must again be wary of the impedance-modifying potential of dependent sources. In any event, the amount of computation required still is typically substantially less than that needed for an exact solution.

Again, the greatest value of the technique lies in its identification of those elements implicated in bandwidth limitations. *The reciprocal of each jth short-circuit time constant is the low frequency –3 dB breakpoint that the circuit would exhibit if that jth capacitor were the only capacitor in the system*.The method of SCτ's then states that the linear combination of these individual, local limitations yields an estimate of the overall –3 dB point. The value of SCτ's derives directly from the identification and approximate quantification of the local degradation terms.

Although the development so far has considered only capacitances, inductances also can be incorporated into the method, although their presence often complicates significantly the decision of which reactive elements really belong in the computation.

The most intuitive way to understand how one incorporates the effect of inductances on bandwidth is to recall that each reciprocal time constant term represents an individual, local contribution to the low-frequency cutoff; we treat the system at each step of the calculation as if that jth reactive element were the sole limiting one. So evidently one treats all of the inductors as *open* circuits when computing the appropriate effective resistances. The *R*/*L* frequencies are then added to the various 1/*RC* frequencies to yield the total estimated low-frequency cutoff point.

3.4 Accuracy of SCτ's

As with OCτ's, one must be careful not to place too much faith in the ability of SCτ's to provide accurate estimates of ω_l in all cases because of the truncation of the denominator polynomial. This caveat notwithstanding, it should be clear that an SCτ estimate is in fact exact for a first-order network since *no* truncation of terms is involved there. Not surprisingly then, the SCτ estimate will be quite accurate if a network of higher order happens to be dominated by one pole (here, that means that one pole is much *higher* in frequency than all of the other poles).

3.5 Other Important Considerations

Although application of short-circuit time constants is pretty straightforward, there are one or two fine points that merit discussion. As with OCτ's, not all capacitors in a network belong in the SCτ calculations. For instance, the capacitors in a transistor model almost never belong. Blind application of the SCτ method would lead to curious (erroneous) results (and a whole heap of extra calculations).

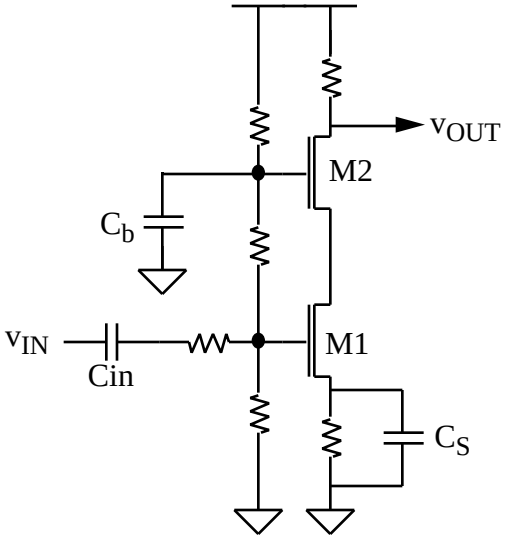
The problem is easily understood if you remember that we assumed that all the zeros are at the origin, and that the number of poles equals the number of zeros, so that the gain in the limit of infinitely high frequency is flat, not zero. We violate these assumptions rather severely if we include all the stuff that causes high-frequency rolloff (i.e., all the stuff that OCτ's worry about). The solution is to *pre-process the network* prior to application of the SCτ method. That is, recognize that all the capacitors that limit high frequency gain are effectively open circuits relative to the impedances around them at frequencies near ω_l. Thus, **one must apply SCτ's only to models that are appropriate to the low-frequency regime**.

While it is usually obvious which capacitors are to be ignored (considered open circuits), there are occasions where one is not so sure. In these cases, a simple thought experiment usually suffices to decide the issue. Now SCτ's are concerned only with those capacitors that limit low-frequency gain. As a consequence, the removal (that is, the open-circuiting) of a capacitor that belongs in the SCτ calculation should result in a decrease in low-frequency gain. The test, therefore, is to consider exciting the network at some low frequency and imagining what would happen to the gain if the capacitor in question were taken out of the circuit (open-circuited). If the gain would decrease, the capacitor belongs in the SCτ calculation since we infer from the result of the thought experiment that the capacitor does indeed limit the low-frequency gain. If the gain would not change (or even increase) upon removal, that capacitor should be open-circuited and left out of the computation. Again, as with OCτ's, the necessary conclusions can usually be reached without taking pencil to paper.

To underscore these issues, let's consider a specific example, the cascode amplifier. As seen in the accompanying schematic, there are three capacitors. The input coupling capacitor, C_{in}, removes any DC from the input signal to prevent upsetting the bias of the amplifier. Source bypass capacitor C_E is chosen to short the source of M1 to ground at all signal frequencies to restore the gain lost by the source degeneration resistor. Bias bypass capac-

itor C_b guarantees that the gate of M2 is an incremental ground at high frequencies to keep the open-circuit time constant sum small.

FIGURE 7. Cascode amplifier



Let's use our thought-experiment technique to deduce which of these three capacitors belongs in the SCτ calculation.

If we begin with C_{in}, we note that the low-frequency gain does decrease (to zero, in fact) if we take it out of the circuit. Hence, it belongs in the calculation. Similarly, C_S belongs in the calculation because its removal also reduces the low-frequency gain.

What about C_b? What happens to the low-frequency gain if we take it out of the circuit? The answer can be either trivial or too deep to fathom, depending on how you approach the question. The easiest way to get to the answer is to recognize that M1 is a device that converts an incoming voltage to an incremental drain current. All M2 does is take this current and pass it on to the output load resistor. Therefore, whether or not the gate of M2 is an incremental ground is irrelevant, and the removal of C_b will therefore have essentially no effect on the low frequency gain. Thus, C_b does *not* belong in the calculation.

The importance of not blindly applying the method cannot be overemphasized. In fact, one of the earliest expositions of the method erroneously includes C_b.[2]

One last issue that deserves some attention concerns the relationship between the reciprocals of the individual short-circuit time constants and the pole frequencies. We have asserted (again without formal proof) only that the *sums* of these frequencies are equal to each other. Therefore, just as with open-circuit time constants, one must resist the tempta-

2. P. E. Gray, C. L. Searle, *Electronic Principles*, Wiley, 1969, pp. 542-546.

tion to equate each reciprocal short-circuit time constant with a corresponding pole frequency. Since the number of short-circuit time constants and the number of poles may not even be equal, one cannot expect each SCτ to equal the time constant of a pole in general.

3.6 Summary and Concluding Remarks

We have seen that the method of short-circuit time constants shares with its dual, the method of open-circuit time constants, a number of advantages and disadvantages. It is an extremely valuable tool for designing amplifiers because of its ability to identify the problem areas of the circuit. Because of the tremendous insights gained with extremely modest effort we are generally willing to overlook its quantitative limitations, such as the often highly conservative nature of the estimated low-frequency cutoff point. As long as we take care to apply the method only to models that apply to the low-frequency regime, we are assured of reasonable answers.

Concluding, the method of short-circuit time constants helps one *design* circuits to satisfy a given low frequency cutoff specification. Despite the quantitative shortcomings of the method, the valuable intuition provided and the labor saved are more than sufficient compensation.

4.0 For Further Reading

For a proof of the equality of the sum of open-circuit and pole time constants see pp.531-535 of *Electronic Principles* by P.E. Gray and C.L. Searle, Wiley and Sons, 1969.

By the way, this work has been extended to allow the exact computation of *all* of the poles of a network. It involves the computation of various cross-products of open- and *short*-circuit time constants to obtain the coefficients of all the powers of *s* in the denominator of the transfer function. Originally developed by Cochrun and Grabel, it was simplified by Rosenstark, but the method is still sufficiently cumbersome (from a hand-calculation viewpoint) that the insight-to-work ratio is usually unfavorably small. However, it occasionally proves useful (especially if you choose to automate the procedure by writing your own code). For more information, see Cochrun and Grabel's paper, "A Method for the Determination of the Transfer Function of Electronic Circuits," IEEE Trans. Ckt. Theory, v. CT-20, no. 1, Jan. 1973, pp. 16-20, and Rosenstark's book, *Feedback Amplifier Principles*, Macmillan, 1986, pp. 67-77.

5.0 Risetime, Delay and Bandwidth

5.1 Introduction

The method of open-circuit time constants allows one to estimate the overall bandwidth from local RC products. In this section, we develop a number of ways to estimate bandwidth from time-domain parameters. In this connection, one occasionally encounters vari-

ous rules of thumb, such as “bandwidth times risetime equals 2.2,” “risetimes add quadratically” or “buy low, sell high.” As useful as they are, however, they aren’t entirely reliable. To identify when these rules of thumb hold, we now turn to their formal derivation.

We start by deriving a rule that appears trivial, obvious and irrelevant: The total delay of a cascade of systems is the sum of the individual delays. The reason for starting here is to introduce some analytical techniques and insights that have far broader applicability.

As always, don’t worry too much about all the mathematical minutiae. The derivations are provided simply for completeness. Those interested primarily in the application of these relationships may skip the intervening math and simply take note of the results.

5.2 Delay of Systems in Cascade

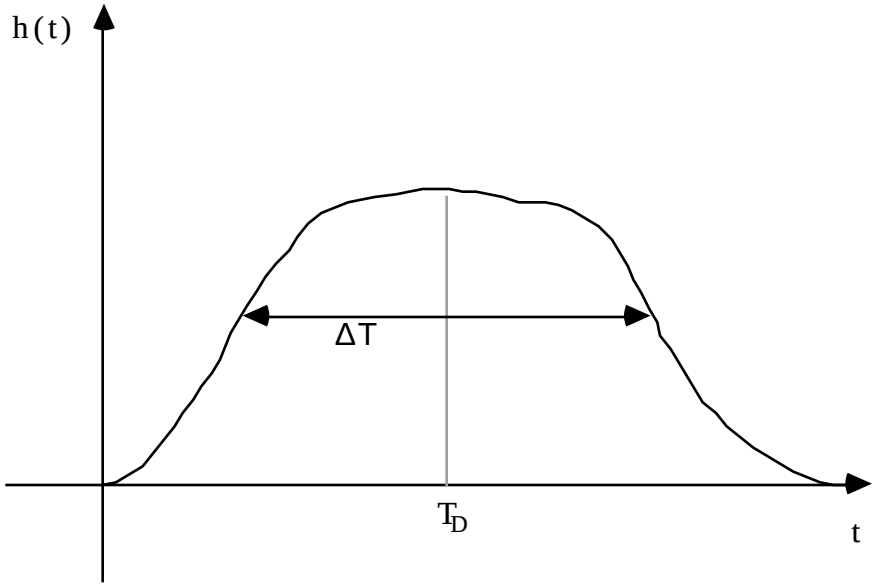
We shall see that many analytical advantages accrue from defining delay (and later, risetime) in terms of *moments of the impulse response*. As seen in Figure 9, one delay measure is the time it takes for the impulse response to reach its “center of mass,” that is, the normalized value of its first moment:

$$T_D \equiv \frac{\int_{-\infty}^{\infty} t h(t)\,dt}{\int_{-\infty}^{\infty} h(t)\,dt} \tag{67}$$

This quantity is also known as the Elmore delay in some of the literature, after the fellow who first used this moment-based approach[3].

3. W. C. Elmore, “The Transient Response of Damped Linear Networks with Particular Regard to Wideband Amplifiers,” *J. Appl. Phys.*, vol. 19, pp. 55-63, January, 1948.

FIGURE 8. Illustrative impulse response



This particular measure of time delay derives much of its utility from the fact that it is readily related to a number of Fourier transform identities, allowing us to exploit the full power of linear system theory. Specifically, the first moment is:

$$\int_{-\infty}^{\infty} th(t)\,dt = -\frac{1}{j2\pi}\frac{d}{df}H(f)\Big|_{f=0} \tag{68}$$

while the normalization factor is simply the DC gain:

$$\int_{-\infty}^{\infty} h(t)\,dt = H(0) \tag{69}$$

so that

$$T_D \equiv \frac{\int_{-\infty}^{\infty} th(t)\,dt}{\int_{-\infty}^{\infty} h(t)\,dt} = -\frac{1}{j2\pi H(0)}\frac{d}{df}H(f)\Big|_{f=0} \tag{70}$$

Using this definition, one finds that the Elmore delay for a single-pole low-pass system is just the pole time constant, τ.

Now that Monsieur Fourier has graciously helped us out by providing a definition of time delay purely in terms of his transforms, the derivation becomes straightforward. Specifically, now consider two systems with impulse responses $h_1(t)$ and $h_2(t)$ with corresponding Fourier transforms $H_1(f)$ and $H_2(f)$. From basic linear system theory, we know that the Fourier transform of these two systems in cascade is just the product of the individual transforms so that $H_{tot} = H_1H_2$. The overall time delay is therefore

$$T_{D,\,tot} = -\frac{1}{j2\pi H_1(0)\,H_2(0)}\frac{d}{df}H_1H_2\bigg|_{f=0} \tag{71}$$

which we may expand to obtain

$$T_{D,\,tot} = -\frac{1}{j2\pi H_1(0)\,H_2(0)}\left[H_2(0)\frac{dH_1}{df}\bigg|_{f=0} + H_1(0)\frac{dH_2}{df}\bigg|_{f=0}\right] \tag{72}$$

from which we immediately (well okay, maybe not quite *immediately*) note that

$$T_{D,\,tot} = T_{D1} + T_{D2} \tag{73}$$

which was to be shown.

We see that use of this particular definition of time delay has led us to the intuitively satisfying result that the overall delay of a cascade of systems is simply the sum of the individual delays.

5.3 Risetime of Systems in Cascade

Deriving a risetime addition rule presents a somewhat more significant challenge. In particular, it turns out that developments based on the conventional 10%-90% definition of risetime are almost certainly doomed to fail because of the analytical difficulties involved with combining exponentials of differing time constants. Since this measure of risetime is arbitrary anyway we might as well seek an alternative (but still arbitrary) definition of risetime that permits tractable analysis.

Just as we employed the first moment of the impulse response in defining the time delay, we find the second moment useful in defining the risetime. Referring again to Figure 9, note that the quantity ΔT is a measure of the duration of the impulse response and hence also a measure of the risetime of the step response (since the step response is the integral of the impulse response). Specifically, ΔT is twice the "radius of gyration" about the "center of mass" (T_D) of $h(t)$. Recalling some dusty relationships from first-year calculus, we find that

$$\left(\frac{\Delta T}{2}\right)^2 \equiv \left[\frac{\int_{-\infty}^{\infty} t^2 h(t)\,dt}{\int_{-\infty}^{\infty} h(t)\,dt} - (T_D)^2\right] \tag{74}$$

Again this definition allows the use of Fourier transform identities. In particular

$$\int_{-\infty}^{\infty} t^2 h(t)\,dt = -\frac{1}{(2\pi)^2}\frac{d^2}{df^2}H(f)\bigg|_{f=0} \tag{75}$$

so that

$$t_{rise}^2 = (\Delta T)^2 = 4\left[\frac{-\frac{1}{(2\pi)^2}\frac{d^2}{df^2}H(f)\Big|_{f=0}}{H(0)} - \left(\frac{1}{j2\pi H(0)}\frac{d}{df}H(f)\Big|_{f=0}\right)^2\right] \tag{76}$$

where we have made use of the equation for delay developed in Section 6.0.

Simplifying (!), we obtain

$$t_{rise}^2 = \frac{4}{(2\pi)^2 H(0)}\left[-\frac{d^2}{df^2}H(f)\bigg|_{f=0} - \frac{1}{H(0)}\left(\frac{d}{df}H(f)\Big|_{f=0}\right)^2\right] \tag{77}$$

The Elmore risetime for a single-pole low-pass system is 2τ.

Proceeding as in Section 6.0, now consider two systems, each with its own risetime. Then

$$t_{rise\,(tot)}^2 = \frac{4}{(2\pi)^2 H_1(0)\,H_2(0)}\left[-\frac{d^2}{df^2}H_1H_2\Big|_{f=0} - \frac{1}{H_1(0)\,H_2(0)}\left(\frac{d}{df}H_1H_2\Big|_{f=0}\right)^2\right] \tag{78}$$

which, after a small algebraic miracle, leads to the desired result at last:

$$t_{rise\,(tot)}^2 = t_{rise1}^2 + t_{rise2}^2 \tag{79}$$

Thus we see that the *squares* of the individual risetimes add linearly to yield the square of the overall risetime. Stated alternatively, the individual risetimes add in root-sum-squared (RSS) fashion to yield the overall risetime:

$$t_{rise\,(tot)} = \sqrt{t_{rise1}^2 + t_{rise2}^2} \tag{80}$$

Now that we've derived these results we should spend some time discussing conditions under which the foregoing formulas may yield unsatisfactory estimates of delay or risetime. In particular consider what happens to the calculated delay and risetime if the integral of $h(t)$ is nearly zero. This situation might arise, for example, if $h(t)$ oscillates more or less evenly about zero. Since the integral of $h(t)$ appears as a normalizing factor in the denominator of our expressions for delay and risetime, inappropriate values for these quantities may result.

To handle this difficulty one might propose a modification of our definitions so that the delay and risetime depend on the moments of the *square* (or perhaps the absolute value) of $h(t)$. "It is left as an exercise for the reader" to show that such modifications result in exceedingly unpleasant expressions that are cumbersome to use and interpret. Thus the simple expressions presented here are understood to apply best when the impulse response is unipolar (or equivalently, the step response is monotonic). If the individual systems satisfy this requirement the relationships derived here will hold well. The greater the departure from the step response monotonicity condition, the less appropriate the use of these formulas.

5.4 A (very short) Application of the Risetime Addition Rule

Aside from permitting one to predict the risetime of a cascade of systems, the risetime addition rule may be used to extend the limits of instrumentation. Consider, for example, trying to measure the risetime of a system whose bandwidth is about the same as that of the instrumentation. Specifically, assume that an oscilloscope with a known risetime of 5 nanoseconds displays a value of 6 nanoseconds for the risetime of a system under test. Using the risetime addition rule we can infer that the true system risetime is about 3.3 nanoseconds, saving us the trouble and expense of trying to make this measurement with equipment that is faster still.

5.5 Bandwidth-Risetime Relations

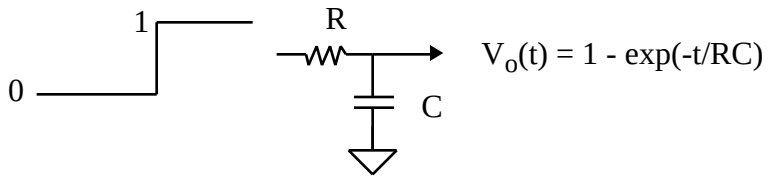
We now take up the problem of examining the rule of thumb that led us to this endeavor in the first place:

$$\omega_{-3dB} t_{rise} \approx 2.2 \tag{81}$$

where ω_{-3dB} is the –3dB bandwidth in radians per second, and t_{rise} is the 10%-90% risetime in response to a step.

Where does this rule come from? Consider our old friend, the simple RC low-pass filter:

FIGURE 9. RC low-pass filter and step response



Given the equation for the response to a unit voltage step, it is straightforward to compute the 10-90% risetime:

$$t_{rise} = RC\ln\left(\frac{0.9}{0.1}\right) \approx 2.2RC \tag{82}$$

Note that this value is about 10% higher than the Elmore risetime computed earlier.

In addition to the risetime, we already know that the –3dB bandwidth (in radians per second) is simply 1/*RC*. Hence, the bandwidth-risetime product is in fact about 2.2, as the rule states.

Since it was derived for a first-order case, should we expect the rule to hold generally for systems of arbitrary order? Well, let's look at a couple of other cases. Consider, for example, the step response of a two-pole system:

$$V_o(t) = 1 - \frac{1}{\sqrt{1-\zeta^2}} e^{-\zeta\omega_n t} \sin\left(\sqrt{1-\zeta^2}\,\omega_n t + \Phi\right) \tag{83}$$

where

$$\Phi = \tan^{-1}\left[\frac{\sqrt{1-\zeta^2}}{\zeta}\right] \tag{84}$$

The –3 dB bandwidth of this system is given by:

$$\omega_h = \omega_n\left(1 - 2\zeta^2 + \sqrt{2 - 4\zeta^2 + 4\zeta^4}\right)^{0.5} \tag{85}$$

Let's use these formulas to explore what happens as we change ζ. For the extreme case of a damping ratio of zero, the risetime and bandwidth are:

$$t_r\big|_{\zeta=0} = \frac{1}{\omega_n}\left[\sin^{-1}0.9 - \sin^{-1}0.1\right] \approx \frac{1.02}{\omega_n} \tag{86}$$

$$\omega_h\big|_{\zeta=0} \approx 1.55\omega_n \tag{87}$$

so that the corresponding bandwidth-risetime product is:

$$\omega_h t_r\Big|_{\zeta = 0} \approx 1.6 \tag{88}$$

or about 72% of the value obtained for the first-order case.

For a reasonably well-damped system, we might expect closer agreement with the first-order result. As a specific example, if we set $\zeta = 1/\sqrt{2}$, the risetime and bandwidth are

$$t_r\Big|_{\zeta = \frac{1}{\sqrt{2}}} \approx \frac{2.14}{\omega_n} \tag{89}$$

and

$$\omega_h\Big|_{\zeta = \frac{1}{\sqrt{2}}} = \omega_n \tag{90}$$

so that

$$\omega_h t_r\Big|_{\zeta = \frac{1}{\sqrt{2}}} \approx 2.14 \tag{91}$$

or a value within a few percent of the first-order result.

Note that the product of bandwidth and Elmore risetime is 2.0 for a single-pole system.

In general, the bandwidth-risetime product will be in the range of 2–2.2 if the system is well damped (or more precisely, if the impulse response is unipolar so that the step response is monotonic, for the same reasons that prevailed in the moment-based expressions for risetime and time delay), and that the product will decrease if the system is not very well damped. However, even in the case of no damping at all, we have seen that the bandwidth-risetime product is still not that far off.

Since most systems of practical interest are generally well damped, we can expect the bandwidth-risetime product to be about 2.2. Therefore, measurement of the step response risetime is often an expedient way to obtain a reasonably accurate estimate of the bandwidth since only one experiment has to be performed, and because step excitations are often easier to generate than sinewaves.[4]

4. At low frequencies, anyway.

5.6 Open-Circuit Time Constants, Risetime and Bandwidth

As we've seen, bandwidth and risetime have a roughly constant product (at least for systems that are "well-behaved"). In addition, the risetimes of cascaded systems increase in root-sum-squared fashion. From these two relationships, we can deduce a bandwidth shrinkage law. It is instructive to compare the results of this exercise with the bandwidth shrinkage law derived rigorously in texts on high-frequency amplifier design.

Consider a cascade of N identical amplifiers, each of which is single-pole with a time constant τ. Combining the risetime addition rule with the bandwidth-risetime relationship yields

$$BW \approx \frac{1}{\sqrt{\sum_{1}^{N} \tau^2}} = \frac{1}{\tau\sqrt{N}} \tag{92}$$

Compare that approximate result with the more exact (but still approximate) relationship (see the chapter on high-frequency amplifier design):

$$BW \approx \frac{\sqrt{\ln 2}}{\tau\sqrt{N}} \approx \frac{0.833}{\tau\sqrt{N}} \tag{93}$$

As can be seen, the functional dependence on N is the same; the equations differ only by a relatively small multiplicative factor.[5]

Note that the method of open-circuit time constants would predict quite a different result. Since the effective time constant is found there by summing all the individual time constants, the OCτ-estimated bandwidth would be

$$BW \approx \frac{1}{\tau N} \tag{94}$$

The difference is significant, and underscores yet again how the use of open-circuit time constants can lead to extremely pessimistic estimates of bandwidth if a single pole does not dominate the transfer function.

6.0 Summary

The method of open- and short-circuit time constants allows us to estimate rapidly the upper and lower –3dB frequencies, almost by inspection of the network. As long as the

5. It should be mentioned that one consequence of the difference between the Elmore and 10-90% risetimes is that Elmore somewhat underestimates the 10-90% risetime growth. A better estimate for identical stages in cascade is about $1.1\sqrt{n}$.

circuit satisfies the assumptions well, the methods yield reasonably accurate answers. More important, however, is the valuable design insight provided.

Another way to estimate bandwidth is from a measurement of risetime. We've seen that moments of the impulse response allow us to exploit the power of linear system theory to show that delays add linearly and that risetimes add in root-sum-squared fashion. Furthermore, we've seen that the product of bandwidth and risetime is roughly constant, and approximately equal to 2.2. For all of these relations, accuracy is greatest when the step response is monotonic. That is, as long as the step response has negligible overshoot and/or ringing, the results derived here will hold well. If these conditions are not well satisfied, then all bets are off. Therefore, do not fall into the trap of believing that these rules of thumb are exact and universally applicable.

As long as we keep this caveat in mind, we can use these relationships to extend significantly the boundaries of our instrumentation, or make quantitative inferences about frequency domain performance from time domain measurements (or vice-versa) when the necessary conditions are well satisfied (as they often, but not always, are).