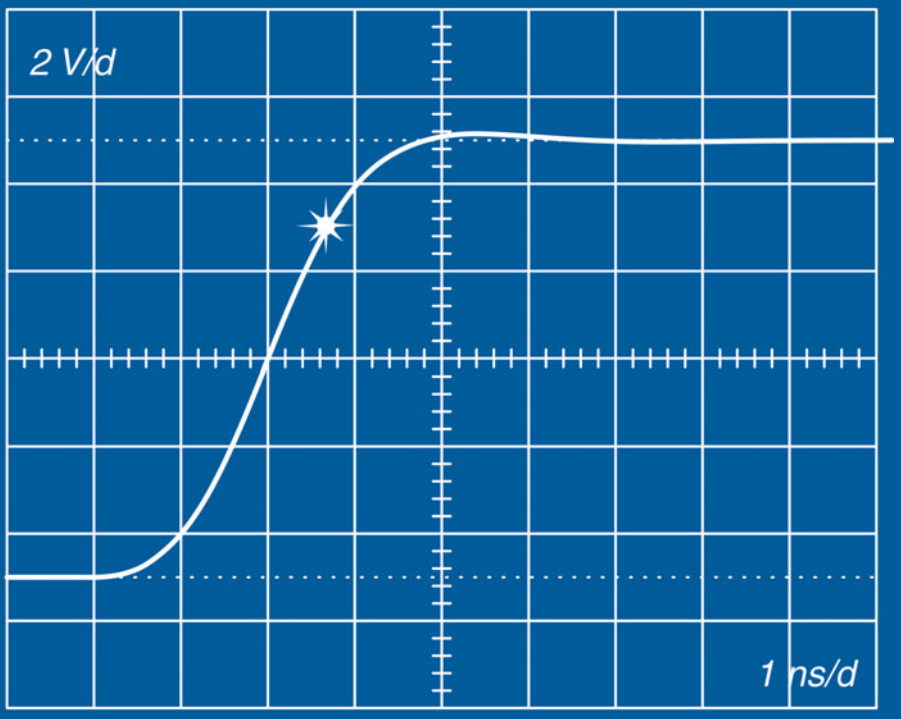
Wideband Amplifiers

By

Peter Starič and Erik Margan



Springer

Most Important Symbols, Units, Constants, and Abbreviations

Note: Owing to historical reasons some symbols have multiple, context-dependent meaning.

$*$	convolution integral (binary oper.); complex conjugate (superscript)
A	ampere, basic SI unit of el. current
A	amplification; amplitude; area, [m^2]
A_v	voltage amplification
A_c	current amplification
a	acceleration, [m/s^2]
AC	alternating current
C	unit of charge, coulomb, [C] = [As]; constant of integration
C	capacitance [F, farad]; contour of integration
C_θ	thermal capacitance, [Ws/K]
c	light propagation velocity, 299 792 458 m/s
CD	critically damped system, all poles real and coincident
DC	direct current
D	circle diameter
d	distance, [m]; wire diameter, [m]
E	electric field strength, [V/m]
E_g	energy gap, [eV] (in semiconductors)
e	Euler (natural) number, 2.718 281 828 459 045…
eV	electron Volt, unit of energy, $1\,\mathrm{eV} = 1.602 \times 10^{-19}\,\mathrm{VAs}$
F	farad, unit of capacitance, [F] = [As/V]
F	force, [N, newton], [N] = [kg m/s^2]
$F(\omega)$	rational function, frequency domain
$F(s)$	rational function in complex freq. domain, $F(s) = P(s)/Q(s)$
f	frequency [Hz, hertz], [Hz] = [1/s]
f_h	upper half-power frequency of a non-peaking amplifier
f_H	upper half-power frequency of a peaking amplifier
$f(t)$	function in time domain (usually impulse response)
G	conductance, $G = 1/R$
$G(s)$	function in complex frequency domain, step resp., $G(s) = F(s)/s$
g_m	mutual transconductance
$g(t)$	function in time domain, usually step response, $g(t) = h(t) * f(t)$
H	unit of inductance, [H, henry], [H] = [Vs/A]
$H(s)$	Heaviside function in complex frequency domain, $H(s) = 1/s$
$h(t)$	Heaviside function, unit step
I	current amplitude [A, ampere], (DC or a complex phasor)
i	index (subscript)
i	instantaneous current [A, ampere]
j	imaginary unit, $\sqrt{-1}$
K	kelvin, absolute temperature, $0\,\mathrm{K} = -273.15\,^\circ\mathrm{C} = -459.67\,^\circ\mathrm{F}$, $\Delta T[1\mathrm{K}] = [1^\circ\mathrm{C}] = [1.8\,^\circ\mathrm{F}]$
k	magnetic field coupling factor of a coil; counting index
k_B	Boltzmann constant, $1.38 \times 10^{-23}\,\mathrm{VAs/K}$
kg	kilogram, basic SI unit of mass
L	inductance [H, henry], [H] = [Vs/A]
$\mathcal{L}\{\}$	direct Laplace transform
$\mathcal{L}^{-1}\{\}$	inverse Laplace transform
l	length of coil winding, [m]
M	mutual inductance [H, henry]; modulus (magnitude)
MFA	maximally flat amplitude
MFED	maximally flat envelope delay
m	coil parameter
m_e	electron mass, $9.1095 \times 10^{-31}\,\mathrm{kg}$
m	meter, basic SI unit of length
N	total number of turns of a coil
n	system order number (n. of poles); specific number of turns (per unit length) of a coil; capacitance ratio

P	electric power [W, watt];
P_n	polynomial of order n
p	pole (polynomial denominator's zero)
Q_m	polynomial of order m
Q_1	transistor label
q_e	electron charge, 1.602×10^{-19} As
R	resistance [Ω, ohm]
R_θ	thermal resistance, [K/W]
r	incremental resistance; radius of a circle
res	residue (*residuum* [lat.])
rad	radian, unit angle, [1 rad] $= \dfrac{360°}{2\pi}$
s	second, basic SI unit of time
S	unit of conductance, siemens, $[1/\Omega]$
s	complex frequency, $\sigma + j\omega$,
s_i	pole or zero of a transfer function
T	time constant, RC or L/R; temperature
t	time [s, second]
t_r	rise time
u	real part of a complex variable; part of a decomposed function
V	volt, unit of electric tension (voltage)
V	voltage amplitude [V, volt], DC voltage or phasor
v	instantaneous voltage [V, volt]; imaginary part of a complex variable; part of a decomposed function; velocity, [m/s]
W	watt, unit of power, [W] = [VA]
X	reactance, the imaginary part of an impedance
x	general variable
Y	admittance (complex), $Y = 1/Z$
y	general function or operation result
Z	impedance (complex), $R + jX$
z	complex variable; polynomial zero

α	current ratio $i_\mathrm{c}/i_\mathrm{e} = \beta/(1+\beta)$
β	transistor current gain, $i_\mathrm{c}/i_\mathrm{b}$
Δ	difference
δ	overshoot [%]; spacing between adjacent turns of a coil [m]
$\delta(t)$	Dirac function, ideal impulse
ε	electrical permittivity, dielectric constant
ε_0	electrical permittivity of vacuum, 8.8542×10^{-12} As/Vm
ε_r	relative electrical permittivity
η_b	bandwidth improvement factor (extension)
η_r	rise time improvement factor (decrease)
θ	pole angle, $\arctan(\sigma/\omega)$
μ	magnetic permeability, $\mu = \mu_0\,\mu_\mathrm{r}$; mobility of charge carriers
μ_0	magnetic permeability of vacuum, $1/(4\pi \times 10^7)$ Vs/Am
μ_r	relative magnetic permeability
$\prod$	product
π	Ludolph number, 3.141 592 653…
$\sum$	sum
σ	real part of a pole
τ	dummy time variable
τ_T	transistor time constant, $1/\omega_\mathrm{T}$
τ_r	rise time
φ	phase angle
χ	frequency ratio, ω/ω_h
Ω	unit of resistance, Ohm
ω	angular frequency, $2\pi f$; imaginary part of a pole
ω_h	upper half-power angular frequency of a non-peaking amplifier
ω_H	upper half-power angular frequency of a peaking amplifier
ω_T	transistor transition frequency at which $\beta(\omega_\mathrm{T}) = 1$

$\Im\{\}$	imaginary part of a complex quantity
$\Re\{\}$	real part of a complex quantity

WIDEBAND AMPLIFIERS

Wideband Amplifiers

by

PETER STARIČ
Jožef Stefan Institute
Milan Vidmar Electric Power Research Institute
Ljubljana, Slovenia

and

ERIK MARGAN
Jožef Stefan Institute
Ljubljana, Slovenia

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN-10 0-387-28340-4 (HB)
ISBN-13 978-0-387-28340-1 (HB)
ISBN-10 0-378-28341-2 (e-book)
ISBN-13 978-0-387-28431-8 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

www.springeronline.com

Printed on acid-free paper

First print 2006
Reprints 2007, 2015

Printed in the Netherlands.

We dedicate this book to all our friends and colleagues in the art of electronics.

Table of Contents

Extra Materials

Available online at: http://extras.springer.com

Appendix 1.1: Simple Poles, Complex Spaces wbapdx11.pdf

Appendix 2.3: Step-responses for 3rd- and 4th-order systems wbapdx23.pdf
Appendix 2.4: Summary of all Inductive Peaking Circuits wbapdx24.pdf

Appendix 3.1: Thermal Analysis wbapdx31.pdf

Appendix 5.1: Improving the CRT Deflection Field wbapdx51.pdf

Appendix 7.1: MFB-3 Transfer Function wbapdx71.pdf
Appendix 7.2: MFB-2 Transfer Function wbapdx72.pdf

Matlab Files: WBAmps Tools and Examples (read Matlab.txt) ..... Matlab_files.zip

Animations (created with Matlab):

Sinewave movie (Fig.1.1.1) ..... sinemovie.gif
Harmonics movie (Fig.1.2.1 and 1.2.2) ..... harmovie.gif
Convolution movie (Fig.1.15.1) ..... convmovie.gif

Acknowledgments

The authors are grateful to ***John Addis***, ***Carl Battjes***, ***Dennis Feucht***, ***Bruce Hoffer*** and ***Bob Ross***, all former employees of Tektronix, Inc., for allowing us to use their class notes, ideas, and publications, and for their help when we had run into problems concerning some specific circuits.

We are also thankful to prof. ***Ivan Vidav*** of the Faculty of Mathematics and Physics in Ljubljana for his help and reviewing of Part 1, and to ***Csaba Szathmary***, a former employee of EMG, Budapest, for allowing us to use some of his measurement results in Part 5.

However, if, in spite of meticulously reviewing the text, we have overlooked some errors, this, of course, is our own responsibility alone.

Peter Starič and Erik Margan

Foreword

With the exception of the tragedy on September 11th, the year 2001 was relatively normal and uneventful — remember, this should have been the year of the *Clarke's* and *Kubrick's* Space Odyssey, mission to Jupiter; it should have been the year of the HAL-9000 computer.

Today, the Personal Computer is as ubiquitous and omnipresent as was HAL on the Discovery spaceship. And the rate of technology development and market growth in electronics industry still follows the famous 'Moore Law', almost four decades after it has been first formulated: in 1965, *Gordon Moore* of Intel Corporation predicted the doubling of the number of transistors on a chip every 2 years, corrected to 18 months in 1967; at that time, the landing on the Moon was in full preparation.

Curiously enough, today no one cares to go to the Moon again, let alone Jupiter. And, in spite of all the effort in digital engineering, we still do not have anything close to 0.1% of HAL's capacity (fortunately?!). Whilst there are many research labs striving to put artificial intelligence into a computer, there are also rumors that this has already happened (with Windows-95, of course!).

In the early 1990s it was felt that digital electronics will eventually render analog systems obsolete. This never happened. Not only is the analog sector vital as ever, the job market demands are expanding in all fields, from high-speed measurement instrumentation and data acquisition, telecommunications and radio frequency engineering, high-quality audio and video, to grounding and shielding, electromagnetic interference suppression and low-noise printed circuit board design, to name a few. And it looks like this demand will be going on for decades to come.

But while the proliferation of digital systems attracted a relatively high number of hardware and software engineers, analog engineers are still rare birds. So, for creative young people, who want to push the envelope, there are lots of opportunities in the analog field.

However, analog electronics did not earn its 'Black-Magic Art' attribute in vain. If you have ever experienced the problems and frustrations from circuits found in too many 'cook-books' and 'sure-working schemes' in electronics magazines, and if you became tired of performing exorcism on every circuit you build, then it is probably the time to try it in a different way: in our own experience, the 'hard' way of doing the correct math first often turns out to be the 'easy' way!

Here is the book **'Wideband Amplifiers'**. The book was intended to serve both as a design manual to more experienced engineers and as a good learning guide to beginners. It should help you to improve your analog design, making better and faster amplifier circuits, especially when time-domain performance is of major concern. We have striven to provide the complete math for every design stage. And, to make learning a joyful experience, we explain the derivation of important math relations from a design engineer point of view, in an intuitive and self-evident manner (rigorous mathematicians might not like our approach). We have included many schematics and performance plots, some practical applications, and a number of computer routines.

However, as with any interesting subject, the greatest problem was never what to include, but rather what to leave out!

In the foreword of a very popular book 'A Brief History of Time', the author *Steven Hawking* wrote that his publisher warned him not to include any math, since the number of readers would be halved by each formula. So he included the $E = m\,c^2$ and bravely cut one half of the world population out.

We went further: there are some 220 formulae in Part 1 only. By estimating the current world population to some $6{\times}10^9$, of which 0.01% could be electronics engineers and assuming an average lifetime interest in the subject of, say, 30 years, if the publisher's rule holds, there ought to be one reader of our book once every:

$$2^{220}/(6 \times 10^9 \times 10^{-4} \times 30 \times 356 \times 24 \times 3600) \approx 3 \times 10^{51} \text{ seconds}$$

or something like $6.6{\times}10^{33}$ × the total age of the Universe!

Now, whatever you might think of it, this book is **not** about math! It is about getting your design to run right first time! Be warned, though, that it will be not enough to just read the book. To have any value, a theory must be put into practice. Although there is no theoretical substitute for hands-on experience, this book should help you to significantly shorten the trial-and-error phase.

We hope that by studying this book thoroughly you will find yourself at the **beginning of a wonderful journey**!

Peter Starič and Erik Margan,

Ljubljana, June 2003

Important Note:

We would like to reassure the Concerned Environmentalists that during the writing of this book, no animal or plant had suffered any harm whatsoever, either in direct or indirect form (excluding the authors, one computer 'mouse' and countless computation 'bugs'!).

More Important Note:

Probably a book without errors still has to be written. We shall be grateful for any error brought to our attention, so that we can make the necessary corrections in the next edition. To report the errors please use one of the e-mail addresses below:

✉ peter.staric@guest.arnes.si
✉ erik.margan@ijs.si

Release Notes (Sept. 2006)

Although the first manuscript of **Wideband Amplifiers** appeared already in spring of 1988, the book has been published only in December 2005.

This is the third reprint, with 5 errors corrected from the second reprint. A list of errors corrected from the previous releases is available on the Springer web page.

Unlike the previous releases which were accompanyied with a compact disk, this release is also available online as an e-book from the Springer web site in the Adobe Portable Document Format (PDF), readable by the Adobe Acrobat™ Reader program, the latest version of which can be downloaded free of charge from http://get.adobe.com/reader/otherversions/). Extra material is also available from the Springer web site http://extras.springer.com.

The PDF format offers numerous links (blue underlined text), which enable easy access to related topics by pointing the 'mouse' cursor on the link and clicking the left 'mouse' button. Returning to the original reading position is possible by clicking the right 'mouse' button and selecting 'Go Back' from the pop up menu (see the AR Help menu for your program version). The Internet and World Wide Web links are in violet (dark magenta) and are accessed by opening the default web browser installed on your computer.

The writing of the text and the math formatting was done by using a program which, in our opinion, is the best for this job: 𝔼𝕏ℙ ™ the Scientific Word Processor, (produced by Simon L. Smith, see http://www.expswp.com/).

The computer algorithms, the development of which we present in Part 6 and 7, are intended as tools for the process of amplifier design and analysis. Written for Matlab™, the Language of Technical Computing (from The MathWorks, Inc., http://www.mathworks.com/), the algorithms have all been revised to conform with the newer version of Matlab, but still retaining downward compatibility as much as possible. We have used Matlab to check all the calculations and draw most of the figures. Before importing them into 𝔼𝕏ℙ the figures were finalized and converted to vector graphics format by using the Adobe Illustrator™ version 8 (see http://www.adobe.com/products/illustrator.html).

All circuit designs were checked using Micro-CAP ™, the Microcomputer Circuit Analysis Program, v. 5 (Spectrum Software, http://www.spectrum-soft.com/). Many circuits have also been checked in practice; unfortunately, in spite of all our effort, we could not include hyperlinks pointing to that ☹.

Peter Starič and Erik Margan

P. Starič, E. Margan

Wideband Amplifiers

Part 1

The Laplace Transform

There is nothing more practical than a good theory!
(William Thompson, Lord Kelvin)

About Transforms

The Laplace transform can be used as a powerful method of solving linear differential equations. By using a time domain integration to obtain the frequency domain transfer function and a frequency domain integration to obtain the time domain response, we are relieved of a few nuisances of differential equations, such as defining boundary conditions, not to speak of the difficulties of solving high order systems of equations.

Although Laplace had used integrals of exponential functions for this same purpose already at the beginning of the 19th century, the method we now attribute to him was effectively developed some 100 years later, based on the Heaviside's operational calculus.

The method is applicable to a variety of physical systems (and even some non physical ones, too!) involving transport of energy, storage and transform, but we are going to use it in a relatively narrow field of calculating the time domain response of amplifier filter systems, starting from a known frequency domain transfer function.

As for any tool, the transform tools, be they Fourier, Laplace, Hilbert, etc., have their limitations. Since the parameters of electronic systems can vary over the widest of ranges, it is important to be aware of these limitations in order to use the transform tool correctly.

Contents:

List of Tables:

List of Figures:

1.0 Introduction

With the advent of television and radar during the Second World War, the behavior of wideband amplifiers in the time domain has become very important [Ref. 1.1]. In today's digital world this is even more the case. It is a paradox that designers and troubleshooters of **digital** equipment still depend on oscilloscopes, which — at least in their fast and low level input part — consist of **analog** wideband amplifiers. So the calculation of the time domain response of wideband amplifiers has become even more important than the frequency, phase, and time delay response.

The emphasis of this book is on the amplifier's time domain response. Therefore a thorough knowledge of time related calculus, explained in Part 1, is a necessary prerequisite for understanding all other parts of this book where wideband amplifier networks are discussed.

The time domain response of an amplifier can be calculated by two main methods: The first one is based on **differential equations** and the second uses the **inverse Laplace transform ($\mathcal{L}^{-1}$ transform)**. The differential equation method requires the calculation of boundary conditions, which — in case of high order equations — means an unpleasant and time consuming job. Another method, which also uses differential equations, is the so called *state variable* calculation, in which a differential equation of order n is split into n differential equations of the first order, in order to simplify the calculations. The state variable method also allows the calculation of non linear differential equations. We will use neither of these, for the simple reason that the Laplace transform and its inverse are based on the system poles and zeros, which prove so useful for network calculations in the frequency domain in the later parts of the book. So most of the data which are calculated there is used further in the time domain analysis, thus saving a great deal of work. Also the use of the $\mathcal{L}^{-1}$ transform does not require the calculation of boundary conditions, giving the result directly in the time domain.

In using the $\mathcal{L}^{-1}$ transform most engineers depend on tables. Their method consists firstly of splitting the amplifier transfer function into partial fractions and then looking for the corresponding time domain functions in the $\mathcal{L}$ transform tables. The sum of all these functions (as derived from partial fractions) is then the result. The difficulty arises when no corresponding function can be found in the tables, or even at an earlier stage, if the mathematical knowledge available is insufficient to transform the partial fractions into such a form as to correspond to the formulae in the tables.

In our opinion an amplifier designer should be self-sufficient in calculating the time domain response of a wideband amplifier. Fortunately, this can be almost always derived from simple rational functions and it is relatively easy to learn the $\mathcal{L}^{-1}$ transforms for such cases. In Part 1 we show how this is done generally, as well as for a few simple examples. A great deal of effort has been spent on illustrating the less clear relationships by relevant figures. Since engineers seek to obtain a first glance insight of their subject of study, we believe this approach will be helpful.

This part consists of four main sections. In the first, the concept of harmonic (e.g., sinusoidal) functions, expressed by pairs of counter-rotating complex conjugate phasors, is explained. Then the Fourier series of periodic waveforms are discussed to obtain the discrete spectra of periodic waveforms. This is followed by the Fourier integral to obtain continuous spectra of non-repetitive waveforms. The convergence problem of the Fourier

integral is solved by introducing the complex frequency variable $s = \sigma + j\omega$, thus coming to direct Laplace transform ($\mathcal{L}$ transform).

The second section shows some examples of the $\mathcal{L}$ transforms. The results are useful when we seek the inverse transforms of simple functions.

The third section deals with the theory of functions of complex variables, but only to the extent that is needed for understanding the inverse Laplace transform. Here the line and contour integrals (Cauchy integrals), the theory of residues, the Laurent series and the $\mathcal{L}^{-1}$ transform of rational functions are discussed. The existence of the $\mathcal{L}^{-1}$ transform for rational functions is proved by means of the Cauchy integral.

Finally, the concluding section deals with some aspects of the $\mathcal{L}^{-1}$ transforms and the convolution integral. Only two standard problems of the $\mathcal{L}^{-1}$ transform are shown, because all the transient response calculations (by means of the contour integration and the theory of residues) of amplifier networks, presented in Parts 2–5, give enough examples and help to acquire the necessary know-how.

It is probably impossible to discuss Laplace transform in a manner which would satisfy both engineers and mathematicians. Professor *Ivan Vidav* said: "*If we mathematicians are satisfied, you engineers would not be, and vice versa*". Here we have tried to achieve the best possible compromise: to satisfy electronics engineers and at the same time not to 'offend' the mathematicians. But, as late colleague, the physicist *Marko Kogoj*, used to say: "*Engineers never know enough of mathematics; only mathematicians know their science to the extent which is satisfactory for an engineer, but they hardly ever know what to do with it!*" Thus successful engineers keep improving their general knowledge of mathematics — far beyond the text presented here.

After studying this part the readers will have enough knowledge to understand all the time domain calculations in the subsequent parts of the book. In addition, the readers will acquire the basic knowledge needed to do the time-domain calculations by themselves and so become independent of $\mathcal{L}$ transform tables. Of course, in order to save time, they will undoubtedly still use the tables occasionally, or even make tables of their own. But they will be using them with much more understanding and self-confidence, in comparison with those who can do $\mathcal{L}^{-1}$ transform only via the partial fraction expansion and the tables of basic functions.

Those readers who have already mastered the Laplace transform **and its inverse**, can skip this part up to Sec. 1.14, where the $\mathcal{L}^{-1}$ transform of a two pole network is dealt with. From there on we discuss the basic examples, which we use later in many parts of the book; the content of Sec. 1.14 should be understood thoroughly. However, if the reader notices any substantial gaps in his/her knowledge, it is better to start at the beginning.

In the last two parts of this book, Part 6 and 7, we derive a set of computer algorithms which reduce the circuit's time domain analysis, performance plotting and pole layout optimization to a pure routine. However attractive this may seem, we nevertheless recommend the study of Part 1: a good engineer must understand the tools he/she is using in order to use them effectively.

1.1 Three Different Ways of Expressing a Sinusoidal Function

We will first show how a sinusoidal function can be expressed in three different ways. The most common way is to express the instantaneous value a of a sinusoid of amplitude A and angular frequency $\omega_1 = 2\pi f_1$, (f_1 = frequency) by the well known formula:

$$a = f(t) = A \sin \omega_1 t \tag{1.1.1}$$

The reason that we have appended the index '1' to ω will become apparent very soon when we will discuss complex signals containing different frequency components. The amplitude vs. time relation of this function is shown in Fig. 1.1.1a. This is the most familiar display seen by using any sinewave oscillator and an oscilloscope.

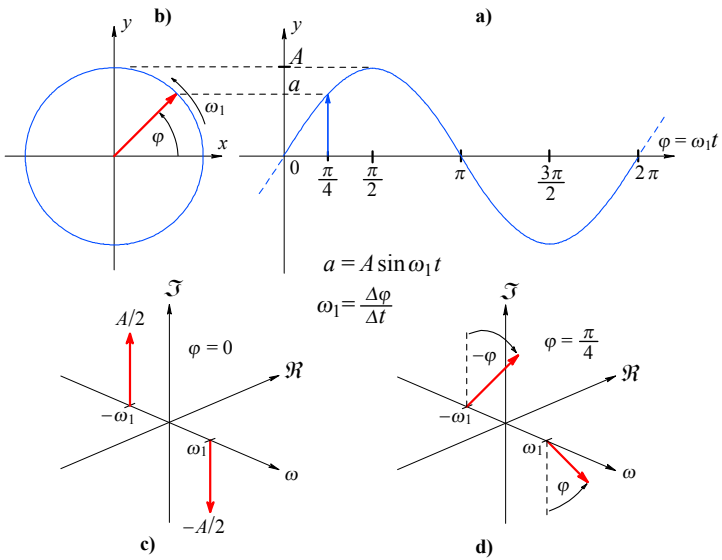


Fig. 1.1.1: Three different presentations of a sine wave: **a)** amplitude in time domain; **b)** a phasor of length A, rotating with angular frequency ω_1; **c)** two complex conjugate phasors of length $A/2$, rotating in opposite directions with angular frequency ω_1, at $\omega_1 t = 0$; **d)** the same as c), except at $\omega_1 t = \pi/4$. To see the movie click on this link: **sinemovie.gif**.

In electrical engineering, another presentation of a sinusoidal function is often used, coming from the vertical axis projection of a rotating phasor A, as displayed in Fig. 1.1.1b, for which the same Eq. 1.1.1 is valid. Here both axes are real, but one of the axes may also be imaginary. In this case the corresponding mathematical presentation is:

$$f(t) = \widehat{A} = A\, \mathrm{e}^{j\omega_1 t} \tag{1.1.2}$$

where $\widehat{A}$ is a complex quantity and $\mathrm{e} = 2.718\,281\ldots$ is the basis of natural logarithms. However, we can also obtain the real quantity a by expressing the sinusoidal function by **two complex conjugate phasors** of length $A/2$ which rotate in opposite directions, as

displayed in a three-dimensional presentation in Fig. 1.1.1c. Here both phasors are shown at $\omega t = 0$ (or $\omega t = 2\pi,\ 4\pi, \ldots$). The sum of both phasors has the instantaneous value a, which is **always real**. This is ensured because both phasors rotate with the same angular frequency $+\omega_1$ and $-\omega_1$, starting as shown in Fig. 1.1.1c, and therefore they are always complex conjugate at any instant. We express a by the well-known *Euler* formula:

$$a = f(t) = A \sin \omega_1 t = \frac{A}{2j}\left(\mathrm{e}^{j\omega_1 t} - \mathrm{e}^{-j\omega_1 t}\right) \tag{1.1.3}$$

The j in the denominator means that both phasors are imaginary at $t = 0$. The sum of both rotating phasors is then zero, because:

$$f(0) = \frac{A}{2j}\,\mathrm{e}^{j\omega_1 0} - \frac{A}{2j}\,\mathrm{e}^{-j\omega_1 0} = 0 \tag{1.1.4}$$

Both phasors in Fig. 1.1.1c and 1.1.1d are placed on the frequency axis at such a distance from the origin as to correspond to the frequency $\pm\,\omega_1$. Since the phasors rotate with time the Fig. 1.1.1d, which shows them at $\varphi = \omega_1 t = \pi/4$, helps us to acquire the idea of a three-dimensional presentation. The understanding of these simple time-frequency relations, presented in Fig. 1.1.1c and 1.1.1d and expressed by Eq. 1.1.3, is essential for understanding both the Fourier transform and the Laplace transform.

Eq. 1.1.3 can be changed to the **cosine** function if the phasor with $+\omega_1$ is multiplied by $j = \mathrm{e}^{j\pi/2}$ and the phasor with $-\omega_1$ by $-j = \mathrm{e}^{-j\pi/2}$. The first multiplication means a **counterclockwise** rotation by 90° and the second a **clockwise** rotation by 90°. This causes both phasors to become real at time $t = 0$, their sum again equaling A:

$$f(t) = j\frac{A}{2j}\,\mathrm{e}^{j\omega_1 t} - (-j)\frac{A}{2j}\,\mathrm{e}^{-j\omega_1 t} = A\cos\omega_1 t \tag{1.1.5}$$

In general a sinusoidal function with a non-zero phase angle φ at $t = 0$ is expressed as:

$$A\sin(\omega t + \varphi) = \frac{A}{2j}\left[\,\mathrm{e}^{j(\omega t+\varphi)} - \mathrm{e}^{-j(\omega t+\varphi)}\,\right] \tag{1.1.6}$$

The need to introduce the frequency axis in Fig. 1.1.1c and 1.1.1d will become apparent in the experiment shown in Fig. 1.1.2. Here we have a unity gain amplifier with a poor loop gain, driven by a sinewave source with frequency ω_1 and amplitude A_1, and loaded by the resistor R_L. If the resistor's value is too low and the amplitude of the input signal is high the amplifier reaches its maximum output current level, and the output signal $f(t)$ becomes distorted (we have purposely kept the same notation A as in the previous figure, rather than introducing the symbol V for voltage). The distorted output signal contains not just the original signal with the same fundamental frequency ω_1, but also a third harmonic component with the amplitude $A_3 < A_1$ and frequency $\omega_3 = 3\,\omega_1$:

$$f(t) = A_1 \sin\omega_1 t + A_3 \sin 3\,\omega_1 t = A_1 \sin\omega_1 t + A_3 \sin\omega_3 t \tag{1.1.7}$$

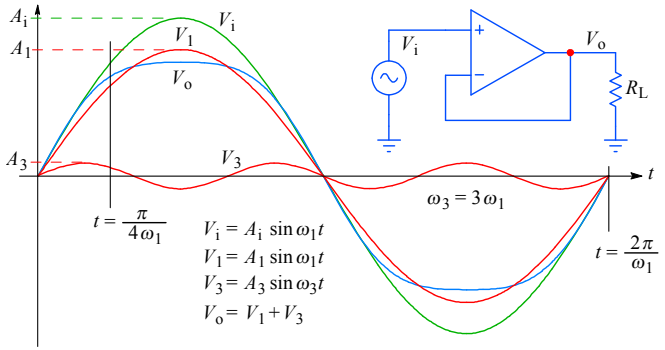


Fig. 1.1.2: The amplifier is slightly overdriven by a pure sinusoidal signal, V_i, with a frequency ω_1 and amplitude A_i. The output signal V_o is distorted, and it can be represented as a sum of two signals, $V_1 + V_3$. The fundamental frequency of V_1 is ω_1 and its amplitude A_1 is somewhat lower. The frequency of V_3 (the third harmonic component) is $\omega_3 = 3\,\omega_1$ and its amplitude is A_3.

Now let us draw the output signal in the same way as we did in Fig. 1.1.1c,d. Here we have two pairs of harmonic components: the first pair of phasors $A_1/2$ rotating with the fundamental frequency $\pm\omega_1$, and the second pair $A_3/2$ rotating with the third harmonic frequency $\pm\omega_3$, which are three times more distant from the origin than ω_1. This is shown in Fig. 1.1.3a, where all four phasors are drawn at time $t = 0$. Fig. 1.1.3b shows the phasors at time $t = \pi/4\,\omega$. Because the third harmonic phasor pair rotates with an angular frequency three times higher, they rotate up to an angle $\pm 3\pi/4$ in the same time.

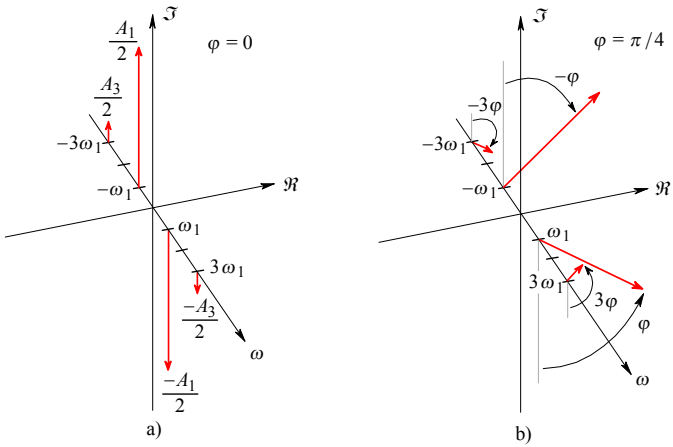


Fig. 1.1.3: The output signal of the amplifier in Fig. 1.1.2, expressed by two pairs of complex conjugate phasors: **a)** at $\omega_1 t = 0$; **b)** at $\omega_1 t = \pi/4$.

Mathematically Eq. 1.1.7, according to Fig. 1.1.2 and 1.1.3, can be expressed as:

$$f(t) = A_1 \sin \omega_1 t + A_3 \sin \omega_3 t$$

$$= \frac{A_1}{2j}\left(e^{j\omega_1 t} - e^{-j\omega_1 t}\right) + \frac{A_3}{2j}\left(e^{j\omega_3 t} - e^{-j\omega_3 t}\right) \qquad (1.1.8)$$

The amplifier output obviously cannot exceed either its supply voltage or its maximum output current. So if we keep increasing the input amplitude the amplifier will clip the upper and lower peaks of the output waveform (some input protection, as well as some internal signal source resistance must be assumed if we want the amplifier to survive in these conditions), thus generating more harmonics. If the input amplitude is very high and if the amplifier loop gain is high as well, the output voltage $f(t)$ would eventually approach a square wave shape, such as in Fig. 1.2.1b in the following section. A true mathematical square wave has an infinite number of harmonics; since no amplifier has an infinite bandwidth, the number of harmonics in the output voltage of any practical amplifier will always be finite.

In the next section we are going to examine a generalized harmonic analysis.

1.2 The Fourier Series

In the experiment shown in Fig. 1.1.2 we have **composed** the output time-function $f(t)$ from sinusoidal waveforms of amplitudes A_1 and A_3. Now, if we have a square wave, as in Fig. 1.2.1b, we would have to deal with many more discrete frequency components. We intend to calculate their amplitudes, assuming that the time function of the square wave is known. This means a **decomposition** of the time function $f(t)$ into the corresponding harmonic frequency components. To do so we will examine the *Fourier* series, named after the French mathematician *Jean Baptiste Joseph de Fourier*[1].

The square wave time function is periodic. A function is periodic if it acquires the same value after its characteristic period $T_1 = 2\pi/\omega_1$, at any instant t:

$$f(t) = f(t + T_1) \tag{1.2.1}$$

Consequently the same is true for $f(t) = f(t + n\,T_1)$, where n is an integer. According to Fourier this square wave can be expressed as a sum of harmonic components with frequencies $f_n = \pm n/T_1$. If $n = 1$ we have the fundamental frequency f_1 with a phasor $A_1/2$, rotating counterclockwise. The phasor f_{-1} with the same length $A_1/2 = A_{-1}/2$ rotates clockwise and forms a complex conjugate pair with the first one. A true square wave would have an infinite number of odd-order harmonics (all even-order harmonics are zero).

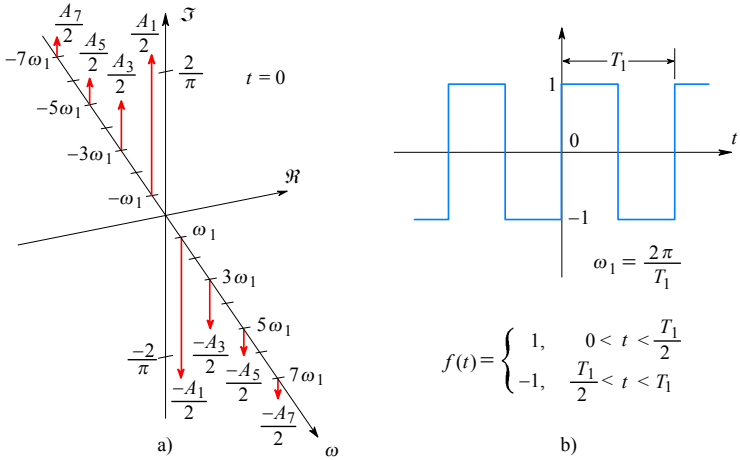


Fig. 1.2.1: A square wave, as shown in **b)**, has an infinite number of odd-order frequency components, of which the first 4 complex-conjugate phasor pairs are drawn in **a)** at time $t = (0 \pm 2\,n\,\pi)/\omega_1$, where n is an integer representing the number of the period.

[1] It is interesting that Fourier developed this method in connection with thermal engineering. As a general in the Napoleon's army he was concerned with gun deformation by heat. He supposed that one side of a straight metal bar is heated and then bent, joining the ends, to form a thorus. Then he calculated the temperature distribution along the circle so formed, in such a way that it would be the sum of sinusoidal functions, each having a different amplitude and a different angular frequency.

In Fig. 1.2.1, we have drawn the complex-conjugate phasor pairs of the first 4 harmonics. Because all the phasor pairs are always complex-conjugate, the sum of any pair, as well as their total sum, is always real. The phasor pairs rotate with different speeds and in opposite directions. Fig. 1.2.2a shows them at time $T_1/8$ to help the reader's imagination. Although this figure looks confusing, the phasors shown have an exact inter-relationship. Looking at the positive ω axis, the phasor with the amplitude $A_1/2$ has rotated in the counterclockwise direction by an angle of $\pi/4$. During the same interval of $T_1/8$ the remaining phasors have rotated: $A_3/2$ by $3\pi/4$; $A_5/2$ by $5\pi/4$; $A_7/2$ by $7\pi/4$; etc. The corresponding complex conjugate phasors on the negative ω axis rotate likewise, but in the opposite (clockwise) direction. The sum of all phasors at any instant t is the instantaneous amplitude of the time domain function. In general, the time function with the fundamental frequency ω_1 is expressed as:

$$f(t) = \sum_{n=-\infty}^{\infty} \frac{A_n}{2} e^{jn\omega_1 t}$$

$$= \cdots + \frac{A_{-n}}{2} e^{-jn\omega_1 t} + \cdots + \frac{A_{-2}}{2} e^{-j2\omega_1 t} + \frac{A_{-1}}{2} e^{-j\omega_1 t}$$

$$+ A_0 + \frac{A_1}{2} e^{j\omega_1 t} + \frac{A_2}{2} e^{j2\omega_1 t} + \cdots + \frac{A_n}{2} e^{jn\omega_1 t} + \cdots \qquad (1.2.2)$$

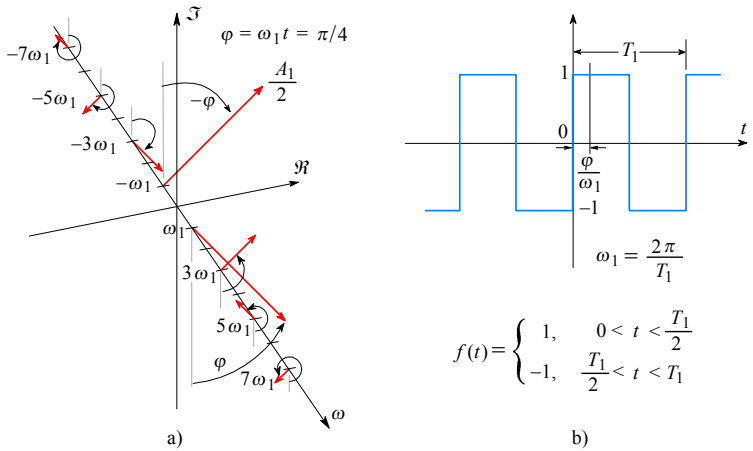


Fig. 1.2.2: As in Fig. 1.2.1, but at an instant $t = (\pi/4 \pm 2n\pi)/\omega_1$; **a)** the spectrum, expressed by complex conjugate phasor pairs, corresponds to the instant $t = \varphi/\omega_1$ in **b)**. To see the movie over one period click this link: **harmovie.gif**. (we apologize for the bad quality resulting from the poor conversion from vector graphics to the bitmap format; readers who have access to Matlab are invited to run the 'harmovie.m' file).

Note that for the square wave all the even frequency components are missing. For other types of waveforms the even coefficients can be non-zero. In general A_i may also be complex, thus containing some non-zero initial phase angle φ_i. In Eq. 1.2.2 we have also introduced A_0, the DC component, which did not exist in our special case. The meaning of A_0 can be understood by examining Fig. 1.2.3a, where the so-called saw tooth waveform is shown, with no DC component. In Fig. 1.2.3b, the waveform has a DC component of magnitude A_0.

Eq. 1.2.2 represents the *complex spectrum* of the function $f(t)$, whilst Fig. 1.2.1 represents the corresponding most significant part of the complex spectrum of a square wave. The next step is the calculation of the magnitudes of the rotating phasors.

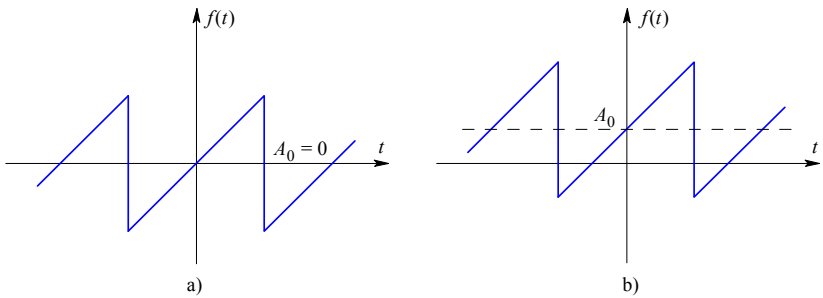


Fig. 1.2.3: a) A waveform without a DC component; **b)** with a DC component A_0.

If we want to measure safely and accurately the diameter of a wheel of a working machine, we must first stop the machine. Something similar can be done with our Eq. 1.2.2, except that here we can mathematically stop the rotation of any single phasor. Suppose we have a phasor $A_k/2$, rotating counterclockwise with frequency $\omega_k = k\,\omega_1$ with an initial phase angle φ_k (at $t = 0$), which is expressed as:

$$\frac{A_k}{2}\,e^{j(\omega_k t+\varphi_k)} = \frac{A_k}{2}\,e^{j\omega_k t}\,e^{j\varphi_k} \tag{1.2.3}$$

Now we multiply this expression by a unit amplitude, clockwise rotating phasor $e^{-j\omega_k t}$ (having the same angular frequency ω_k) to cancel the $e^{j\omega_k t}$ term, [Ref. 1.2]:

$$\frac{A_k}{2}\,e^{j\varphi_k}\,e^{j\omega_k t}\,e^{-j\omega_k t} = \frac{A_k}{2}\,e^{j\varphi_k} \tag{1.2.4}$$

and obtain a non-rotating component which has the magnitude $A_k/2$ and phase angle φ_k **at any time**. With this in mind let us attack the whole time function $f(t)$. The duration of the multiplication must last exactly one whole period and the corresponding expression is:

$$\frac{A_k}{2} = \frac{1}{T}\int_{-T/2}^{T/2} f(t)\,e^{-j\omega_k t}\,dt \tag{1.2.5}$$

Since we have integrated over the whole period T in order to get the average value of that harmonic component, the result of the integration must be divided by T, as in Eq. 1.2.5. If there is a DC component (with $\omega = 0$) in the spectrum, the calculation of it is simply:

$$A_0 = \frac{1}{T}\int_{-T/2}^{T/2} f(t)\,dt \tag{1.2.6}$$

To return to Eq. 1.2.5, let us explain the meaning of the integration Eq. 1.2.5 by means of Fig. 1.2.4.

By multiplying the function $f(t)$ by $e^{-j\omega_k t}$ we have stopped the rotating phasor $A_k/2$, while during the time interval of integration all the other phasors have rotated through an angle of $n\,2\pi$ (where n is an integer), **including** the DC phasor A_0, because it is now multiplied by $e^{-j\omega_k t}$. The result of the integration for all these rotating phasors is zero, as indicated in Fig. 1.2.4a, whilst the phasor $A_k/2$ has stopped, integrating eventually to its full amplitude; the integration for this phasor only is shown in Fig. 1.2.4b.

The understanding of the described effect resulting from the multiplication of $f(t)$ by $e^{-j\omega t}$ is essential for the understanding of the basic principles behind the Fourier series, the Fourier integral and the Laplace transform.

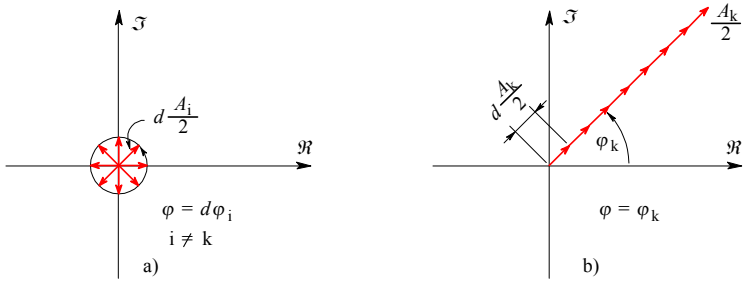


Fig. 1.2.4: a) The integral over the full period T of a rotating phasor is zero; **b)** the integral over a full period T of a non-rotating phasor $d\,A_k/2$, gives its amplitude, $A_k/2$, (the symbol d stands for dt/T — in this figures $dt \rightarrow \Delta t$ such that $\Delta t\,\omega_k = \pi/4$). Note that a stationary phasor retains its initial angle φ_k.

For us the Fourier series represents only a transitional station on the journey towards the Laplace transform. So we will drive through it with a moderate speed "via the Main Street", without investigating some interesting things in the side streets. Nevertheless, it is useful to make a practical example. Since we have started with a square wave, shown in Fig. 1.2.5, let us calculate its complex spectrum components $A_n/2$, assuming that the square wave amplitude is $A = 1$.

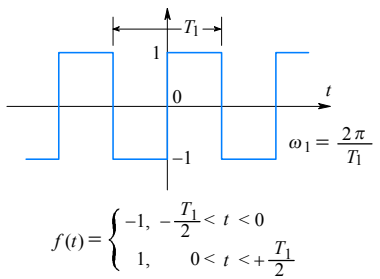


Fig. 1.2.5: A square wave signal.

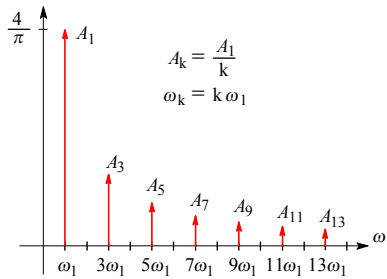


Fig. 1.2.6: The frequency spectrum of a square wave, expressed by **real** values (magnitudes) only.

For a **single period** the corresponding mathematical expression of this function is:

$$f(t) = \begin{cases} -1 & \text{for} \quad -T/2 < t < 0 \\ +1 & \text{for} \quad 0 < t < T/2 \end{cases}$$

According to Eq. 1.2.5 we calculate:

$$\frac{A_n}{2} = \frac{1}{T}\left[\int_{-T/2}^{0}(-1)\,\mathrm{e}^{-j2\pi nt/T}\,dt + \int_{0}^{T/2}(+1)\,\mathrm{e}^{-j2\pi nt/T}\,dt\right]$$

$$= \frac{1}{T}\left(\frac{T}{j2\pi n}\,\mathrm{e}^{-j2\pi nt/T}\Big|_{-T/2}^{0} + \frac{T}{-j2\pi n}\,\mathrm{e}^{-j2\pi nt/T}\Big|_{0}^{T/2}\right)$$

$$= \frac{1}{j2\pi n}\left(1 - \mathrm{e}^{j\pi n} - \mathrm{e}^{-j\pi n} + 1\right) = \frac{-1}{j\pi n}\left(\frac{\mathrm{e}^{j\pi n} + \mathrm{e}^{-j\pi n}}{2} - 1\right)$$

$$= \frac{-1}{j\pi n}\left(\cos\pi n - 1\right) \qquad (1.2.7)$$

The result is zero for $n = 0$ (the DC component A_0) and for any even n. For any odd n the value of $\cos\pi n = -1$, and for such cases the result is:

$$\frac{A_n}{2} = \frac{2}{j\pi n} = \frac{-2j}{\pi n} \qquad (1.2.8)$$

The factor $-j$ in the numerator means that for any positive n the phasor is negative and imaginary (also for $t = 0,\ 2\pi,\ 4\pi,\ 6\pi, \ldots$), whilst for negative n it is positive and imaginary. This is evident from Fig. 1.2.1a.

Let us calculate the first eight phasors by using Eq. 1.2.8. The lengths of phasors in Fig. 1.2.1a and 1.2.2.b correspond to the values reported in Table 1.2.1. All the phasors form complex conjugate pairs and their total sum **always gives a real value**.

Table 1.2.1: The first few harmonics of a square wave

$\pm n$	0	1	3	5	7	9	11	13
$\mp A_n/2$	0	$2j/\pi$	$2j/3\pi$	$2j/5\pi$	$2j/7\pi$	$2j/9\pi$	$2j/11\pi$	$2j/13\pi$

However, a spectrum can also be shown with real values only, e.g., as it appears on the cathode ray tube screen of a spectrum analyzer. To obtain this, we simply sum the corresponding complex conjugate phasor pairs (e.g., $|A_n/2| + |A_{-n}/2| = A_n$) and place them on the abscissa of a two-dimensional coordinate system, as shown in Fig. 1.2.6. Such a non-rotating spectrum has only the positive frequency axis. Although such a presentation of spectra is very useful in the analysis of signals containing several (or many) frequency components, we will continue calculating with the complex spectra, because the phase information is also important. And, of course, the Laplace transform, which is our main goal, is based on a complex variable.

Now let us recompose the waveform using only the harmonic frequency components from Table 1.2.1, as shown in Fig. 1.2.7a. The waveform resembles the square wave but it has an exaggerated overshoot $\delta \simeq 18\,\%$ of the nominal amplitude.

The reason for the overshoot δ is that we have abruptly cut off the higher harmonic components from a certain frequency upwards. Would this overshoot be lower if we take more harmonics? In Fig. 1.2.7b we have increased the number of harmonic components

three times, but the overshoot remained the same. No matter how many, yet for any **finite** number of harmonic components, used to recompose the waveform, the overshoot would stay the same (only its duration becomes shorter if the number of harmonic components is increased, as is evident from Fig. 1.2.7a and 1.2.7b).

This is the *Gibbs*' phenomenon (*Josiah Willard Gibbs*, 1839–1903, [Ref. 1.24]). It tells us that we should not cut off the frequency response of an amplifier abruptly if we do not wish to add an undesirably high overshoot to the amplified pulse. Fortunately, real amplifiers can not have an infinitely steep high frequency roll off, so a gradual decay of high frequency response is always ensured. However, as we shall explain in Part 2 and 4, the overshoot may increase as a result of other effects.

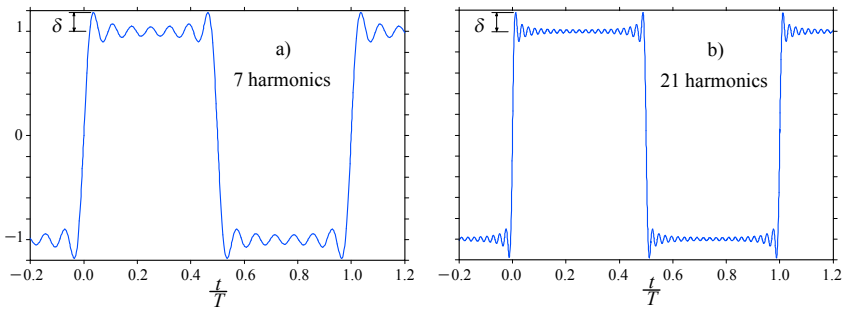


Fig. 1.2.7: The Gibbs' phenomenon; **a)** A signal composed of the first seven harmonics of a square wave spectrum from Table 1.2.1. The overshoot is $\delta \simeq 18\,\%$ of the nominal amplitude; **b)** Even if we take three times more harmonics the overshoot δ is nearly equal in both cases.

In a similar way to that for the square wave, **any periodic signal of finite amplitude and with a finite number of discontinuities within one period, can be decomposed into its frequency components.** As an example the waveform in Fig. 1.2.8 could also be decomposed, but we will not do it here. Instead in the following section we will analyze another waveform which will allow us to generalize the method of frequency analysis.

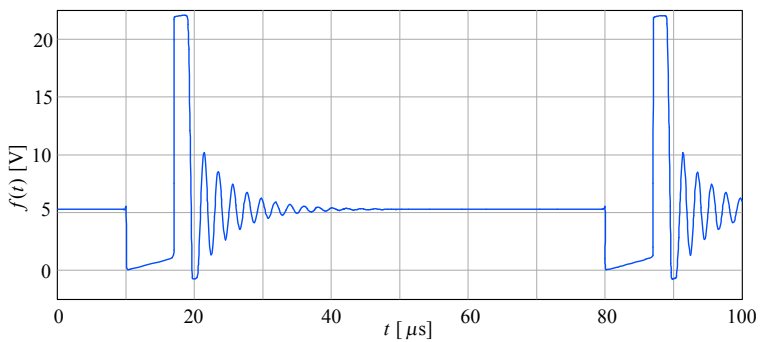


Fig. 1.2.8: An example of a periodic waveform (a typical flyback switching power supply), having a finite number of discontinuities within one period. Its frequency spectrum can also be calculated using the Fourier transform, if needed (e.g., to analyze the possibility of electromagnetic interference at various frequencies), in the same way as we did for the square wave.

1.3 The Fourier Integral

Suppose we have a function $f(t)$ composed of square waves with the duration τ and repeating with a period T, as shown in Fig. 1.3.1. For this function we can also calculate the Fourier series (the corresponding spectrum is shown in Fig. 1.3.2) in the same way as for the continuous square wave case in the previous section.

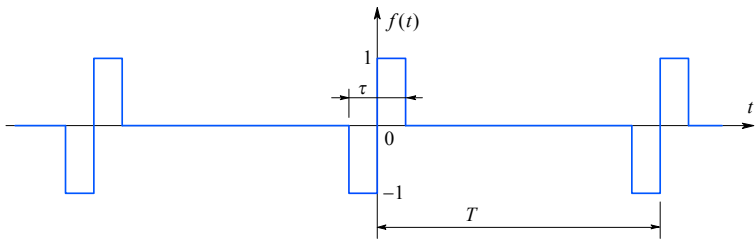


Fig. 1.3.1: A square wave with duration τ and period $T = 5\tau$

The difference between the continuous square wave spectrum and the spaced square wave in Fig. 1.3.1 is that the integral of this function can be broken into two parts, one comprising the length of the pulse, τ, and the zero-valued part between two pulses of a length $T - \tau$. The reader can do this integration for himself, because it is fairly simple. We will only write the result:

$$\frac{A_n}{2} = -j\,\tau\,\frac{\sin^2\left[n\,\omega_1(\tau/4)\right]}{n\,\omega_1(\tau/4)} \tag{1.3.1}$$

where $\omega_1 = 2\pi/T$, assuming that the pulse amplitude is 1 (if the amplitude were A it would simply multiply the right hand side of the equation). For the conditions in Fig. 1.3.1, where $T = 5\tau$ and $A = 1$, the spectrum has the form shown in Fig. 1.3.2, with $\omega_\tau = 2\pi/\tau$.

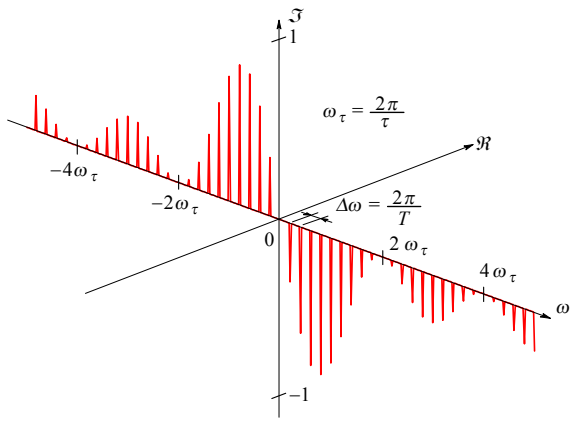


Fig. 1.3.2: Complex spectrum of the waveform in Fig. 1.3.1.

A very interesting question is that of what would happen to the spectrum if we let the period $T \to \infty$? In general a function $f(t)$ can be recomposed by adding all its harmonic components:

$$f(t) = \sum_{n=-\infty}^{\infty} \frac{A_n}{2}\, \mathrm{e}^{jn\omega_1 t} \tag{1.3.2}$$

where A_n may also be complex, thus containing the initial phase angle φ_i. Again, as in the previous section, each discrete harmonic component can be calculated with the integral:

$$\frac{A_n}{2} = \frac{1}{T} \int_{-T/2}^{T/2} f(t)\, \mathrm{e}^{-jn\omega_1 t}\, dt \tag{1.3.3}$$

For the case in Fig. 1.3.1 the integration should start at $t = 0$ and the integral has the form:

$$\frac{A_n}{2} = \frac{1}{T} \int_{0}^{T} f(t)\, \mathrm{e}^{-jn\omega_1 t}\, dt \tag{1.3.4}$$

Insert this into Eq. 1.3.2:

$$f(t) = \sum_{n=-\infty}^{\infty} \left[\frac{1}{T} \int_{0}^{T} f(\tau)\, \mathrm{e}^{-jn\omega_1 \tau}\, d\tau \right] \mathrm{e}^{jn\omega_1 t} \tag{1.3.5}$$

Here we have introduced a dummy variable τ in the integral, in order to distinguish it from the variable t outside the brackets. Now we express the integral inside the brackets as:

$$\int_{0}^{T} f(\tau)\, \mathrm{e}^{-jn\omega_1 \tau}\, d\tau = \int_{0}^{T} f(\tau)\, \mathrm{e}^{-j2\pi n\tau/T}\, d\tau = F\Big(\frac{2n\pi}{T}\Big) = F(n\omega_1) \tag{1.3.6}$$

Thus:

$$f(t) = \sum_{n=-\infty}^{\infty} \frac{1}{T} F(n\omega_1)\, \mathrm{e}^{jn\omega_1 t} = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} \frac{2\pi}{T} F(n\omega_1)\, \mathrm{e}^{jn\omega_1 t}$$

$$= \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} \omega_1 F(n\omega_1)\, \mathrm{e}^{jn\omega_1 t} \tag{1.3.7}$$

where $2\pi/T = \omega_1$. If we let $T \to \infty$ then ω_1 becomes infinitesimal, and we call it $d\omega$. Also $n\omega_1$ becomes a continuous variable ω. So in Eq. 1.3.7 the following changes take place:

$$\sum_{n=-\infty}^{\infty} \Rightarrow \int_{-\infty}^{\infty} \qquad \omega_1 \Rightarrow d\omega \qquad n\omega_1 \Rightarrow \omega$$

With all these changes Eq. 1.3.7 is transformed into Eq. 1.3.8:

$$\boxed{f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega)\, \mathrm{e}^{j\omega t}\, d\omega} \tag{1.3.8}$$

Consequently Eq. 1.3.6 also changes, obtaining the form:

$$F(\omega) = \int_0^{\infty} f(t)\, \mathrm{e}^{-j\omega t}\, dt \tag{1.3.9}$$

In Eq. 1.3.9 $F(\omega)$ has no discrete frequency components but it forms a **continuous** spectrum. Since $T \to \infty$ the DC part vanishes (as it would for **any** pulse shape, not just symmetrical shapes), according to Eq. 1.2.6:

$$A_0 = \lim_{T \to \infty} \frac{1}{T} \int_0^{T} f(t)\, dt = 0 \tag{1.3.10}$$

Eq. 1.3.8 and 1.3.9 are called *Fourier integrals*. Under certain (usually rather limited) conditions, which we will discuss later, it is possible to use them for the calculation of transient phenomena. The second integral (Eq. 1.3.9) is called the *direct Fourier transform*, which we express in a shorter way:

$$\mathcal{F}\left\{f(t)\right\} = F(\omega) \tag{1.3.11}$$

The first integral (Eq. 1.3.8) represents the *inverse Fourier transform* and it is usually written as:

$$\mathcal{F}^{-1}\left\{F(\omega)\right\} = f(t) \tag{1.3.12}$$

In Eq. 1.3.8, $F(\omega)$ means a **firm** spectrum and the factor $\mathrm{e}^{j\omega t}$ means the rotation of each of the corresponding infinite spectrum components contained in $F(\omega)$ with its angular frequency ω, which is a continuous variable. In Eq. 1.3.9 $f(t)$ means the complete time function, containing an infinite number of **rotating** phasors and the factor $\mathrm{e}^{-j\omega t}$ means the rotation 'in the opposite direction' to stop the rotation of the corresponding rotating phasor $\mathrm{e}^{j\omega t}$ contained in $f(t)$, at its particular frequency ω.

Let us now select a suitable time function $f(t)$ and calculate its continuous spectrum. Since we have already calculated the spectrum of a periodic square wave, it would be interesting to display the spectrum of a single square wave as shown in Fig. 1.3.3b. We use Eq. 1.3.9:

$$F(\omega) = \int_{-\tau/2}^{\infty} f(t)\, \mathrm{e}^{-j\omega t}\, dt = \int_{-\tau/2}^{0} (-1)\, \mathrm{e}^{-j\omega t}\, dt + \int_{0}^{\tau/2} (+1)\, \mathrm{e}^{-j\omega t}\, dt \tag{1.3.13}$$

Here we have a single square wave with a 'period' T from $t = -\tau/2$ to $+\infty$. However, we need to integrate only from $t = -\tau/2$ to $t = \tau/2$, because $f(t)$ is zero outside this interval. It is important to note that at the discontinuity where $t = 0$, we have started the second integral. For a function with more discontinuities, between each of them we must write a separate integral. Thus it is obvious that the function $f(t)$ must have a **finite** number of discontinuities for it to be possible to calculate its spectrum.

The result of the above integration is:

$$F(\omega) = \frac{1}{-j\omega}\left(-1 + e^{j\omega\tau/2} + e^{-j\omega\tau/2} - 1\right) = \frac{2}{j\omega}\left(1 - \frac{e^{j\omega\tau/2} + e^{-j\omega\tau/2}}{2}\right)$$

$$= \frac{-2j}{\omega}\left(1 - \cos\frac{\omega\tau}{2}\right) = \frac{-2j}{\omega}\left(2\sin^2\frac{\omega\tau}{4}\right) = \frac{-4j}{\omega}\sin^2\frac{\omega\tau}{4}$$

$$= -j\tau\,\frac{\sin^2\dfrac{\omega\tau}{4}}{\dfrac{\omega\tau}{4}} \qquad (1.3.14)$$

A three-dimensional display of a spectrum, corresponding to this result, is shown in Fig. 1.3.3a. Here the frequency scale has been altered with respect to Fig. 1.2.1a in order to display the spectrum better.

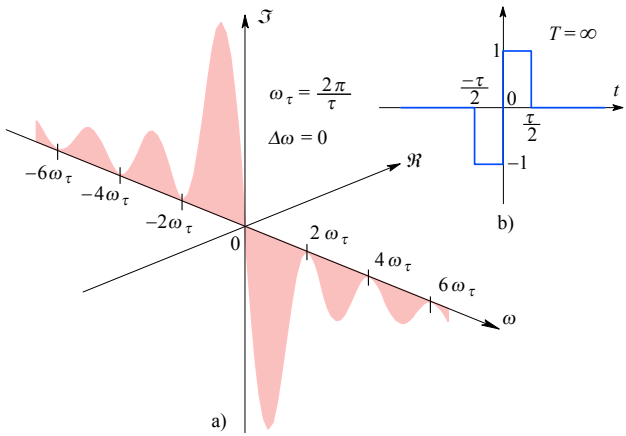


Fig. 1.3.3: a) The frequency spectrum of a single square wave is expressed by complex conjugate phasors. Since the phasors are infinitely many, they merge in a continuous planar form. Also the spectrum extends to $\omega = \pm\infty$. The corresponding waveform is shown in **b)**. Note that all the even frequency components $2\pi n/\tau$ are missing (n is an integer).

By comparing Fig. 1.2.1a and 1.3.3a we may draw the following conclusions:

1. Both spectra contain no even frequency components, e.g., at $\pm 2\omega_\tau$, $\pm 4\omega_\tau$, etc., where $\omega_\tau = 2\pi/\tau$;
2. In both spectra there is no DC component A_0;
3. By comparing Fig. 1.3.2 and 1.3.3 we note that the envelope of both spectra can be expressed by Eq. 1.3.14;
4. By comparing Eq. 1.3.1 and 1.3.14 we note that the discrete frequency $n\,\omega_1$ from the first equation is replaced by the continuous variable ω in the second equation. Everything else has remained the same.

In the above example we have decomposed an aperiodic waveform (also called a *transient*), expressed as $f(t)$, into a **continuous complex** spectrum $F(\omega)$. Before discussing the functions which are suitable for the application of the Fourier integral let us see some common periodic and non-periodic signals. A sustained tone from a trumpet we consider to be a periodic signal, whilst a beat on a drum is a non-periodic signal (in a strict mathematical sense, both signals are non-periodic, because the first one also started out of silence). The transition from silence to sound we call the *transient*. In accordance with this definition, of the waveforms in Fig. 1.3.4 only a) and b) show a periodic waveform, whilst c) and d) display transients.

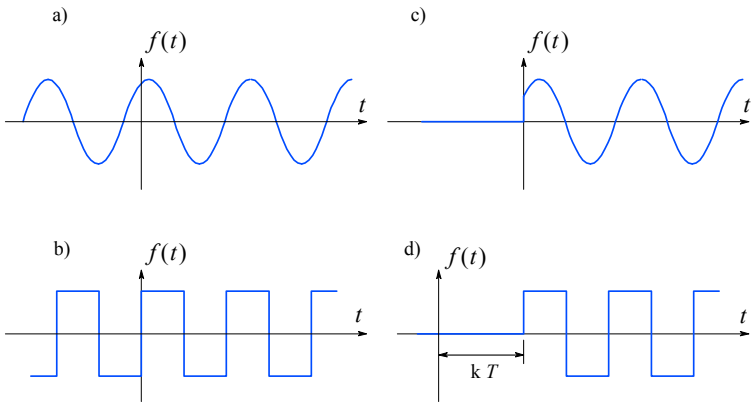


Fig. 1.3.4: a) and **b)** periodic functions, **c)** and **d)** aperiodic functions

The question arises of whether it is possible to calculate the spectra of the transients in Fig. 1.3.4c and 1.3.4d by means of the Fourier integral using Eq. 1.3.8?

The answer is **no**, because the integral in Eq. 1.3.8 does not converge for any of these two functions. The integral is also non-convergent for the most simple step signal, which we intend to use extensively for the calculation of the step response of amplifier networks.

This inconvenience can be avoided if we multiply the function $f(t)$ by a suitable convergence factor, e.g., e^{-ct}, where $c > 0$ and its magnitude is selected so that the integral in Eq. 1.3.2 remains finite when $t \to \infty$. In this way, the problem is solved for $t \geq 0$. In doing so, however, the integral becomes divergent for $t < 0$, because for negative time the factor e^{-ct} has a positive exponent, causing a rapid increase to infinity. But this, too, can be avoided, if we assume that the function $f(t)$ is zero for $t < 0$. In electrical engineering and electronics we can always assume that a circuit is dead until we switch the power on or we apply a step voltage signal to its input and thus generate a transient. The transform where $f(t)$ must be zero for $t < 0$ is called a *unilateral transform*.

For functions which are suitable for the unilateral Fourier transform the following relation must hold [Ref. 1.3]:

$$\lim_{T \to \infty} \int_0^T |f(t)|\, \mathrm{e}^{-ct} dt \; < \; \infty \qquad (1.3.15)$$

where $f(t)$ is a single-valued function of t and c is positive and real.

If so, we can write the direct transform:

$$F(c,\omega) = \int_0^\infty \left[f(t)\,\mathrm{e}^{-ct} \right] \mathrm{e}^{-j\omega t}\,dt \tag{1.3.16}$$

If we want this integral to converge to some finite value for $t \to \infty$, the real constant must be $c \geq \sigma_\mathrm{a}$, where σ_a is the *abscissa of absolute convergence*. The magnitude of σ_a depends on the nature of the function $f(t)$. I.e., if $f(t) = 1$, then $\sigma_\mathrm{a} = 0$, and if $f(t) = \mathrm{e}^{-\alpha t}$ then $\sigma_\mathrm{a} = -\alpha$, where $\alpha > 0$. By applying the convergence factor e^{-ct}, the inverse Fourier transform obtains the form:

$$f(t)\,\mathrm{e}^{-ct} = \frac{1}{2\pi} \int_{-\infty}^\infty F(c,\omega)\,\mathrm{e}^{j\omega t}\,d\omega \qquad \text{for } t \geq 0 \tag{1.3.17}$$

Here we must add all the complex-conjugate phasors with frequencies from $\omega = -\infty$ to $+\infty$. Although the direct Fourier transform in our case was unilateral, **the inverse transform is always bilateral**. Because in Eq. 1.3.16 we have deliberately introduced the convergence factor e^{-ct} we must limit $c \to 0$ after the integral is solved in order to get the required $F(\omega)$.

Since our final goal is the Laplace transform we will stop the discussion of the Fourier transform here. We will, however, return to this topic later in Part 6, where we will discuss the solving of system transfer functions and transient responses using numerical methods, suitable for machine computation. There we will discuss the application of the very efficient Fast Fourier Transform (FFT) algorithm to both frequency and time domain related problems.

1.4 The Laplace Transform

By a slight change of Eq. 1.3.16 and 1.3.17 we may arrive at a general complex Fourier transform [Ref. 1.3]. This is done so that we join the kernel $e^{-j\omega t}$ and the convergence factor e^{-ct}. In this way Eq. 1.3.16 is transformed into:

$$F(c+j\omega) = \int_0^\infty \big[f(t)\big]\, e^{-(c+j\omega)t}\, dt \qquad \text{where} \quad c \geq \sigma_a \tag{1.4.1}$$

The formula for an inverse transform is derived from Eq. 1.3.17 if both sides of the equation are multiplied by e^{ct}. In addition, the simple variable ω is now replaced by a new one: $c + j\omega$. By doing so we obtain:

$$f(t) = \frac{1}{2\pi j} \int_{c-j\infty}^{c+j\infty} F(c+j\omega)\, e^{(c+j\omega)t}\, d(c+j\omega) \qquad \text{for } t \geq 0 \text{ and } c \geq \sigma_a \tag{1.4.2}$$

If in Eq. 1.4.1 and 1.4.2 the *constant* c becomes a *real variable* σ, both equations are transformed into the form called *Laplace transform*. The name is fully justified, since the French mathematician *Pierre Simon de Laplace* had already introduced this transform in 1779, whilst Fourier published his transform 43 years later.

It is a custom to denote the complex variable $\sigma + j\omega$ by a single symbol s, which we also call the *complex frequency* (in some, mostly mathematical, literature this variable is also denoted as p). With this new variable Eq. 1.4.1 can be rewritten:

$$\boxed{F(s) = \mathcal{L}\{f(t)\} = \int_0^\infty f(t)\, e^{-st}\, dt} \tag{1.4.3}$$

and this is called the *direct Laplace transform*, or $\mathcal{L}$ *transform*. It represents the complex spectrum $F(s)$. The above integral is valid for functions $f(t)$ such that the factor e^{-st} keeps the integral convergent. If we now insert the variable s in Eq. 1.4.2, we have:

$$\boxed{f(t) = \mathcal{L}^{-1}\{F(s)\} = \frac{1}{2\,\pi\, j} \int_{c-j\infty}^{c+j\infty} F(s)\, e^{st}\, ds} \tag{1.4.4}$$

This integral is called the *inverse Laplace transform*, or $\mathcal{L}^{-1}$ *transform*.

Like the inverse Fourier transform, Eq. 1.4.4 is a **bilateral transform** too. In the integral Eq. 1.4.3 it is assumed that $f(t) = 0$ for $t < 0$, thus that equation represents the *unilateral transform*. In addition, the real part of the variable s satisfies $\Re\{s\} = \sigma \geq \sigma_a$, where σ_a is the *abscissa of absolute convergence*, as we have already discussed for Eq. 1.3.16 and 1.3.17 [see also Ref. 1.23]. In the integral Eq. 1.4.4 $t \geq 0$, so here too we must have $\sigma \geq \sigma_a$.

The path of integration is parallel with the imaginary axis, as shown in Fig. 1.4.1. The constant c in the integration limits must be properly chosen, in order to ensure the convergence of the integral.

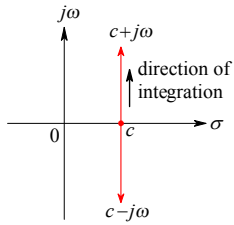


Fig. 1.4.1: The abscissa of absolute convergence — the integration path for Eq. 1.4.4.

The factor e^{-st} in Eq. 1.4.3 is needed to stop the rotation of the corresponding phasor e^{st}; there are infinitely many such phasors in the time function $f(t)$. As our variable is now complex, $s = \sigma + j\omega$, the factor e^{-st} does not mean a simple rotation, but a **spiral rotation** in which the radius is exponentially **decreasing** with t because of σ, the real part of s. This is necessary to cancel the corresponding rotation e^{st}, contained in $f(t)$, with a radius, which, in an exactly equal manner, **increases** with t [Ref. 1.23].

Since in Eq. 1.4.4 the factor e^{st} becomes divergent if the exponent $\Re\{st\} > 1$, the above conditions for the variable σ (and for the constant c) must be met to ensure the convergence of the integral. In the analysis of passive networks these conditions can always be met, as we will show in many examples in the subsequent sections.

Now, because we have reached our goal, the Laplace transform and its inverse, we may ask ourselves what we have accomplished by doing all this hard work.

For the time being we can claim that we have transformed the function of a *real variable* t into a function of a *complex variable* s. This allows us to calculate, using the $\mathcal{L}$ transform, the spectrum function $F(s)$ of a finite transient, defined by the function $f(t)$. Or, more important for us, by means of the $\mathcal{L}^{-1}$ transform we can calculate the time domain function, if the frequency domain function $F(s)$ is known.

Later we will show how we can transform **linear differential equations** in the time domain, by means of the $\mathcal{L}$ transform, into **algebraic equations** in the s domain. Since the algebraic equations are much easier to solve than the differential ones, this means one has a great facility. Once our calculations in the s domain are completed, then by means of the $\mathcal{L}^{-1}$ transform we obtain the corresponding time domain function. In this way we avoid solving directly the differential equations and the calculation of boundary conditions.

1.5 Examples of Direct Laplace Transform

Now let us put our new tools to use and calculate the $\mathcal{L}$ transform of several simple functions. The results may also be used for the $\mathcal{L}^{-1}$ transform and the reader is encouraged to learn the most basic of them by heart, because they are used extensively in the other parts of the book and, of course, in the analysis of the most common electronics circuits.

1.5.1 Example 1

Most of our calculations will deal with the step response of a network. To do so our excitation function will be a simple unit step $h(t)$, or the *Heaviside* function (after *Oliver Heaviside*, 1850–1925) as is shown in Fig. 1.5.1. This function is defined as:

$$f(t) = h(t) = \begin{cases} 0 & \text{for} \quad t < 0 \\ 1 & \text{for} \quad t > 0 \end{cases}$$

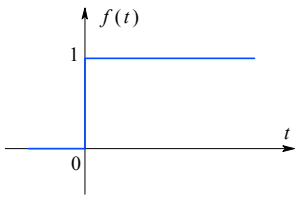


Fig. 1.5.1: Unit step function.

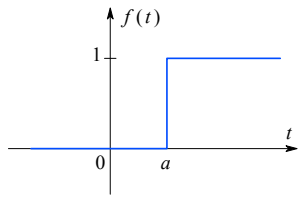


Fig. 1.5.2: Unit step function starting at $t = a$.

As we agreed in the previous section, $f(t) = 0$ for $t < 0$ for all the following functions, and we will not repeat this statement in further examples. At the same time let us mention that for our calculations of $\mathcal{L}$ transform it is not important what is the actual value of $f(0)$, providing it is finite [Ref. 1.3].

The $\mathcal{L}$ transform for the unit step function $f(t) = h(t)$ is:

$$F(s) = \mathcal{L}\{f(t)\} = \int_0^\infty [1]\, \mathrm{e}^{-st}\, dt = \frac{1}{-s}\, \mathrm{e}^{-st} \Big|_{t=0}^{t=\infty} = \frac{1}{s} \tag{1.5.1}$$

1.5.2 Example 2

The function is the same as in Example 1, except that the step does not start at $t = 0$ but at $t = a > 0$ (Fig. 1.5.2):

$$f(t) = \begin{cases} 0 & \text{for} \quad t < a \\ 1 & \text{for} \quad t > a \end{cases}$$

Solution:

$$F(s) = \int_a^\infty [1]\, \mathrm{e}^{-st}\, dt = \frac{1}{-s}\, \mathrm{e}^{-st} \Big|_{t=a}^{t=\infty} = \frac{1}{s}\, \mathrm{e}^{-as} \tag{1.5.2}$$

1.5.3 Example 3

The exponential decay function is shown in Fig. 1.5.3; its mathematical expression:

$$f(t) = \mathrm{e}^{-\sigma_1 t}$$

is defined for $t > 0$, as agreed, and σ_1 is a constant.
Solution:

$$F(s) = \int_0^\infty \mathrm{e}^{-\sigma_1 t}\, \mathrm{e}^{-st}\, dt = \int_0^\infty \mathrm{e}^{-(\sigma_1+s)t}\, dt$$

$$= \frac{-1}{\sigma_1 + s}\, \mathrm{e}^{-(\sigma_1+s)t}\Big|_{t=0}^{t=\infty} = \frac{1}{\sigma_1 + s} \tag{1.5.3}$$

Later, we shall meet this and the following function and also their product very often.

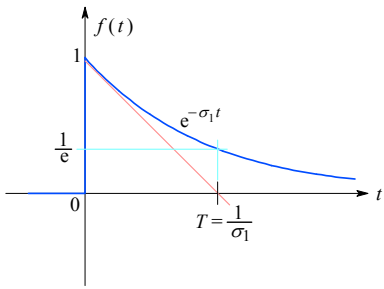


Fig. 1.5.3: Exponential function.

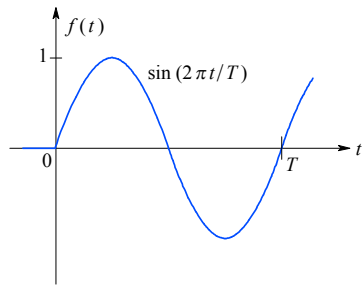


Fig. 1.5.4: Sinusoidal function.

1.5.4 Example 4

We have a sinusoidal function as in Fig. 1.5.4; its corresponding mathematical expression is:

$$f(t) = \sin \omega_1 t$$

where the constant $\omega_1 = 2\pi/T$.
Solution: its $\mathcal{L}$ transform is:

$$F(s) = \int_0^\infty (\sin \omega_1\, t)\, \mathrm{e}^{-st}\, dt \tag{1.5.4}$$

To integrate this function we substitute it using Euler's formula:

$$\sin \omega_1 t = \frac{1}{2\,j}\left(\mathrm{e}^{j\omega_1 t} - \mathrm{e}^{-j\omega_1 t}\right) \tag{1.5.5}$$

Then we have:

$$F(s) = \frac{1}{2\,j}\left(\int_0^\infty \mathrm{e}^{j\omega_1 t}\, \mathrm{e}^{-st}\, dt - \int_0^\infty \mathrm{e}^{-j\omega_1 t}\, \mathrm{e}^{-st}\, dt\right)$$

$$= \frac{1}{2\,j}\left[\int_0^\infty \mathrm{e}^{-(s-j\omega_1)t}\, dt - \int_0^\infty \mathrm{e}^{-(s+j\omega_1)t}\, dt\right] \tag{1.5.6}$$

The solution of this integral is, in a way, similar to that in the previous example:

$$F(s) = \frac{1}{2\,j}\left(\frac{1}{s - j\omega_1} - \frac{1}{s + j\omega_1}\right)$$

$$= \frac{1}{2\,j} \cdot \frac{s + j\omega_1 - s + j\omega_1}{s^2 + \omega_1^2} = \frac{\omega_1}{s^2 + \omega_1^2} \tag{1.5.7}$$

This is a typical function of a continuous wave (CW) sinusoidal oscillator, with a frequency ω_1.

1.5.5 Example 5

Here we have the cosine function as in Fig. 1.5.5, expressed as:

$$f(t) = \cos\omega_1 t$$

Solution: the $\mathcal{L}$ transform of this function is calculated in a similar way as for the sine. According to Euler's formula:

$$\cos\omega_1 t = \frac{1}{2}\left(\mathrm{e}^{j\omega_1 t} + \mathrm{e}^{-j\omega_1 t}\right) \tag{1.5.8}$$

Thus we obtain:

$$F(s) = \frac{1}{2}\left[\int_0^\infty \mathrm{e}^{-(s-j\omega_1)t}\,dt + \int_0^\infty \mathrm{e}^{-(s+j\omega_1)t}\,dt\right] = \frac{1}{2}\left(\frac{1}{s - j\omega_1} + \frac{1}{s + j\omega_1}\right)$$

$$= \frac{1}{2} \cdot \frac{s + j\omega_1 + s - j\omega_1}{s^2 + \omega_1^2} = \frac{s}{s^2 + \omega_1^2} \tag{1.5.9}$$

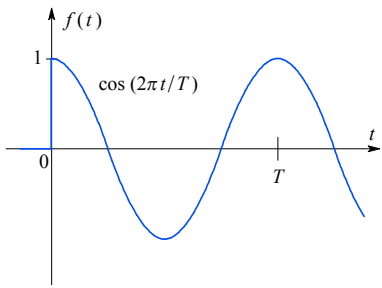


Fig. 1.5.5: Cosine function.

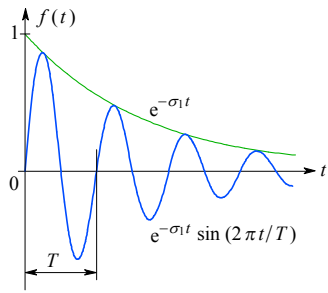


Fig. 1.5.6: Damped oscillations.

1.5.6 Example 6

In Fig. 1.5.6 we have a damped oscillation, expressed by the formula:

$$f(t) = \mathrm{e}^{-\sigma_1 t}\sin\omega_1 t$$

Solution: we again substitute the sine function, according to Euler's formula:

$$F(s) = \frac{1}{2j}\int_0^\infty \mathrm{e}^{-(s+\sigma_1)t}\left(\mathrm{e}^{j\omega_1 t} - \mathrm{e}^{-j\omega_1 t}\right) dt$$

$$= \frac{1}{2j}\int_0^\infty \left[\mathrm{e}^{-(s+\sigma_1-j\omega_1)t} - \mathrm{e}^{-(s+\sigma_1+j\omega_1)t}\right] dt$$

$$= \frac{1}{2j}\left(\frac{1}{s+\sigma_1-j\omega_1} - \frac{1}{s+\sigma_1+j\omega_1}\right) = \frac{\omega_1}{(s+\sigma_1)^2+\omega_1^2} \qquad (1.5.10)$$

An interesting similarity is found if this formula is compared with the result of Example 4. There, for a CW we have in the denominator s^2 alone, whilst here, because the oscillations are damped, we have $(s+\sigma_1)^2$ instead, and σ_1 is the damping factor.

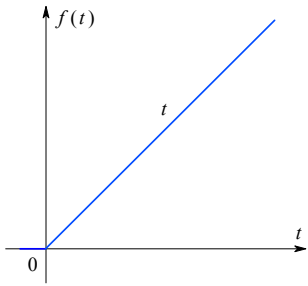


Fig. 1.5.7: Linear ramp $f(t) = t$.

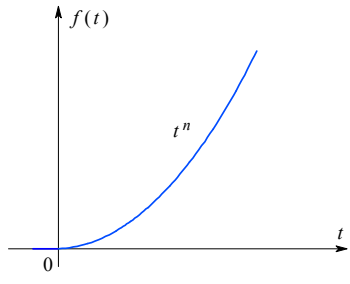


Fig. 1.5.8: Power function $f(t) = t^n$.

1.5.7 Example 7

A linear ramp, as shown in Fig. 1.5.7, is expressed as:

$$f(t) = t$$

Solution: we integrate by parts according to the known relation:

$$\boxed{\int u\,dv = u\,v - \int v\,du}$$

and we assign $t = u$ and $\mathrm{e}^{-s\,t}\,dt = dv$ to obtain:

$$F(s) = \int_0^\infty t\,\mathrm{e}^{-st}\,dt = \frac{t\,\mathrm{e}^{-st}}{-s}\bigg|_{t=0}^{t=\infty} + \frac{1}{s}\int_0^\infty \mathrm{e}^{-st}\,dt$$

$$= 0 - 0 - \frac{1}{s^2}\,\mathrm{e}^{-st}\bigg|_{t=0}^{t=\infty} = \frac{1}{s^2} \qquad (1.5.11)$$

1.5.8 Example 8

Fig. 1.5.8 displays a function which has a general analytical form:

$$f(t) = t^n$$

Solution: again we integrate by parts, decomposing the integrand $t^n\,\mathrm{e}^{-st}$ into:

$$u = t^n \qquad du = n\,t^{n-1}dt \qquad v = \frac{1}{-s}\,\mathrm{e}^{-st} \qquad dv = \mathrm{e}^{-st}\,dt$$

With these substitutions we obtain:

$$F(s) = \int_0^\infty t^n\,\mathrm{e}^{-st}\,dt = \frac{t^n\,\mathrm{e}^{-st}}{-s}\bigg|_{t=0}^{t=\infty} + \frac{n}{s}\int_0^\infty t^{n-1}\,\mathrm{e}^{-st}\,dt$$

$$= \frac{n}{s}\int_0^\infty t^{n-1}\,\mathrm{e}^{-st}\,dt \tag{1.5.12}$$

Again integrating by parts:

$$\frac{n}{s}\int_0^\infty t^{n-1}\,\mathrm{e}^{-st}\,dt = \frac{t^{n-1}\,\mathrm{e}^{-st}}{-s}\bigg|_{t=0}^{t=\infty} + \frac{n\,(n-1)}{s^2}\int_0^\infty t^{n-2}\,\mathrm{e}^{-st}dt$$

$$= \frac{n\,(n-1)}{s^2}\int_0^\infty t^{n-2}\,\mathrm{e}^{-st}\,dt \tag{1.5.13}$$

By repeating this procedure n times we finally arrive at:

$$F(s) = \int_0^\infty t^n\,\mathrm{e}^{-st}\,dt = \frac{n\,(n-1)(n-2)\cdots 3\cdot 2\cdot 1}{s^n}\int_0^\infty t^0\,\mathrm{e}^{-st}\,dt = \frac{n!}{s^{n+1}} \tag{1.5.14}$$

1.5.9 Example 9

The function shown in Fig. 1.5.9 corresponds to the expression:

$$f(t) = t\,\mathrm{e}^{-\sigma_1 t}$$

Solution: by integrating by parts we obtain:

$$F(s) = \int_0^\infty t\,\mathrm{e}^{-\sigma_1 t}\,\mathrm{e}^{-st}\,dt = \int_0^\infty t\,\mathrm{e}^{-(\sigma_1+s)t}\,dt = \frac{1}{(\sigma_1+s)^2} \tag{1.5.15}$$

1.5.10 Example 10

Similarly to Example 9, except that here we have $t^{\,n}$, as in Fig. 1.5.10:

$$f(t) = t^{\,n}\,\mathrm{e}^{-\sigma_1 t}$$

Solution: we apply the procedure from Example 8 and Example 9:

$$F(s) = \int_0^\infty t^{\,n}\,\mathrm{e}^{-\sigma_1 t}\,\mathrm{e}^{-st}\,dt = \int_0^\infty t^{\,n}\,\mathrm{e}^{-(\sigma_1+s)t}\,dt = \frac{n!}{(\sigma_1+s)^{n+1}} \tag{1.5.16}$$

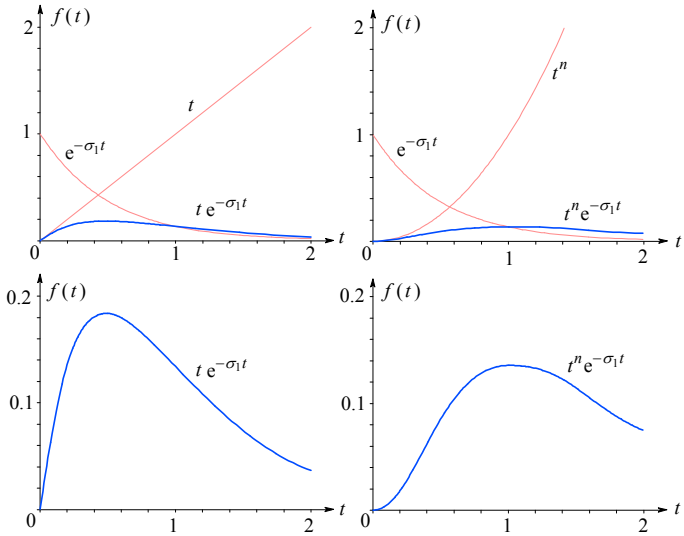


Fig. 1.5.9: Function $f(t) = t\,\mathrm{e}^{-\sigma_1 t}$. **Fig. 1.5.10:** Function $f(t) = t^n\,\mathrm{e}^{-\sigma_1 t}$.

These ten examples, which we frequently meet in practice, demonstrate that the calculation of an $\mathcal{L}$ transform is not difficult. Since the results derived are used often, we have collected them in Table 1.5.1.

Table 1.5.1: **Ten frequently met $\mathcal{L}$ transform examples**

No.	$f(t)$	$F(s)$		No.	$f(t)$	$F(s)$
1	$1 \quad (\text{for } t > 0)$	$\dfrac{1}{s}$		6	$\mathrm{e}^{-\sigma_1 t}\sin\omega_1 t$	$\dfrac{\omega_1}{(s^2+\sigma_1^2)+\omega_1^2}$
2	$1 \quad (\text{for } t > a)$	$\dfrac{1}{s}\,\mathrm{e}^{-a\,t}$		7	t	$\dfrac{1}{s^2}$
3	$\mathrm{e}^{-\sigma_1 t}$	$\dfrac{1}{\sigma_1 + s}$		8	t^n	$\dfrac{n!}{s^{n+1}}$
4	$\sin\omega_1 t$	$\dfrac{\omega_1}{s^2+\omega_1^2}$		9	$t\,\mathrm{e}^{-\sigma_1 t}$	$\dfrac{1}{(\sigma_1+s)^2}$
5	$\cos\omega_1 t$	$\dfrac{s}{s^2+\omega_1^2}$		10	$t^n\,\mathrm{e}^{-\sigma_1 t}$	$\dfrac{n!}{(\sigma_1+s)^{n+1}}$

1.6 Important Properties of the Laplace Transform

It is useful to know some of the most important properties of the $\mathcal{L}$ transform:

1.6.1 Linearity (1)

$$\mathcal{L}\{f(t) \pm g(t)\} = \mathcal{L}\{f(t)\} \pm \mathcal{L}\{g(t)\} \tag{1.6.1}$$

Example:

$$\mathcal{L}\{t + \sin\omega_1 t\} = \mathcal{L}\{t\} + \mathcal{L}\{\sin\omega_1 t\} = \frac{1}{s^2} + \frac{\omega_1}{s^2 + \omega_1^2} \tag{1.6.2}$$

1.6.2 Linearity (2)

$$\mathcal{L}\{K\, f(t)\} = K\, \mathcal{L}\{f(t)\} \tag{1.6.3}$$

where K is a real constant.
Example:

$$\mathcal{L}\{4\,\mathrm{e}^{-\sigma_1 t}\sin\omega_1 t\} = 4\,\mathcal{L}\{\mathrm{e}^{-\sigma_1 t}\sin\omega_1 t\} = 4\,\frac{\omega_1}{(s+\sigma_1)^2 + \omega_1^2} \tag{1.6.4}$$

1.6.3 Real Differentiation

$$\mathcal{L}\left\{\frac{df(t)}{dt}\right\} = s\,F(s) - f(0^+) \tag{1.6.5}$$

The transform of a derivative of the function $f(t)$ is obtained if we multiply $F(s)$ by s and subtract the value of $f(t)$ if $0 \leftarrow t$ from the **right** side, denoted by the + sign at $f(0^+)$ (the direction is important because the values $f(0^-)$ and $f(0^+)$ can be different). We will prove this statement by deriving it from the definition of the $\mathcal{L}$ transform:

$$F(s) = \mathcal{L}\{f(t)\} = \int_0^\infty f(t)\,\mathrm{e}^{-st}\,dt \tag{1.6.6}$$

We will integrate by parts by making $f(t) = u$ and $\mathrm{e}^{-st}\,dt = dv$. The result is:

$$\int_0^\infty f(t)\,\mathrm{e}^{-st}\,dt = f(t)\,\frac{1}{-s}\,\mathrm{e}^{-st}\Big|_{t=0}^{t=\infty} + \frac{1}{s}\int_0^\infty \left(\frac{d\,f(t)}{dt}\right)\mathrm{e}^{-st}\,dt$$

$$= \frac{f(0^+)}{s} + \frac{1}{s}\int_0^\infty \left(\frac{d\,f(t)}{dt}\right)\mathrm{e}^{-st}\,dt$$

$$= \frac{f(0^+)}{s} + \frac{1}{s}\,\mathcal{L}\left\{\frac{df(t)}{dt}\right\} \tag{1.6.7}$$

By rearranging, we prove the statement expressed in Eq. 1.6.6:

$$s\int_0^\infty f(t)\,\mathrm{e}^{-st}\,dt - f(0^+) = s\,F(s) - f(0^+)$$

$$= \int_0^\infty \left[\frac{d\,f(t)}{dt}\right]\mathrm{e}^{-st}\,dt = \mathcal{L}\left\{\frac{d\,f(t)}{dt}\right\} \tag{1.6.8}$$

Example:

$$\mathcal{L}\left\{ \frac{d\,(\mathrm{e}^{-\sigma_1 t})}{dt} \right\} = s\,F(s) - f(0^+) = s\,\frac{1}{s+\sigma_1} - 1 = \frac{-\sigma_1}{s+\sigma_1} \tag{1.6.9}$$

We may also check the result by first differentiating the function $\mathrm{e}^{-\sigma_1 t}$:

$$\frac{d\,(\mathrm{e}^{-\sigma_1 t})}{dt} = -\sigma_1\,\mathrm{e}^{-\sigma_1 t} \tag{1.6.10}$$

and then applying the $\mathcal{L}$ transform:

$$\mathcal{L}\{-\sigma_1\,\mathrm{e}^{-\sigma_1 t}\} = -\sigma_1\,\mathcal{L}\{\mathrm{e}^{-\sigma_1 t}\} = \frac{-\sigma_1}{s+\sigma_1} \tag{1.6.11}$$

The result is the same.

By now the advantage of the $\mathcal{L}$ transform against differential equations should have become obvious. In the s domain the derivative of the function $f(t)$ corresponds to $F(s)$ multiplied by s and subtracting the value $f(0^+)$. The reason that t must approach zero from the right (+) side is our prescribing $f(t)$ to be zero for $t < 0$. In other words, we have a unilateral transform.

The higher derivatives are obtained by repeating the above procedure. If for the first derivative we have obtained:

$$\mathcal{L}\{f'(t)\} = s\,F(s) - f(0^+) \tag{1.6.12}$$

then the $\mathcal{L}$ transform of the second derivative is:

$$\begin{aligned}\mathcal{L}\{f''(t)\} &= s\left(\mathcal{L}\{f'(t)\} - f'(0^+)\right)\\ &= s\left(s\,F(s) - f(0^+)\right) - f'(0^+) = s^2\,F(\mathrm{s}) - s\,f(0^+) - f'(0^+)\end{aligned} \tag{1.6.13}$$

By a similar procedure the $\mathcal{L}$ transform of the third derivative is:

$$\begin{aligned}\mathcal{L}\{f'''(t)\} &= s\left(s^2\,F(s) - s\,f(0^+) - f'(0^+)\right) - f''(0^+) =\\ &= s^3\,F(s) - s^2\,f(0^+) - s\,f'(0^+) - f''(0^+)\end{aligned} \tag{1.6.14}$$

Thus the $\mathcal{L}$ transform of the n^{th} derivative is simply:

$$\begin{aligned}\mathcal{L}\left\{f^{(n)}(t)\right\} &= s^n\,F(s) - s^{n-1}f(0^+) - s^{n-2}f'(0^+) - s^{n-3}f''(0^+) - \cdots\\ &\cdots - s^2 f^{(n-3)}(0^+) - s\,f^{(n-2)}(0^+) - f^{(n-1)}(0^+)\end{aligned} \tag{1.6.15}$$

1.6.4 Real Integration

We intend to prove that:

$$\mathcal{L}\left\{ \int_0^t f(\tau)\,d\tau \right\} = \frac{F(s)}{s} \tag{1.6.16}$$

We will derive the proof from the basic definition of the $\mathcal{L}$ transform:

$$\mathcal{L}\left\{\int_0^t f(\tau)\,d\tau\right\} = \int_0^\infty \left[\int_0^t f(\tau)\,d\tau\right] \mathrm{e}^{-st}\,dt \tag{1.6.17}$$

For the integration by parts we assign:

$$u = \int_0^t f(\tau)\,d\tau \qquad du = f(\tau)\,d\tau \qquad v = \frac{1}{-s}\,\mathrm{e}^{-st} \qquad dv = \mathrm{e}^{-st}\,dt \tag{1.6.18}$$

By considering all this we may write the integral:

$$\int_0^\infty \int_0^t f(\tau)\,d\tau\,\mathrm{e}^{-st}\,dt = \left[\int_0^t \frac{1}{-s}\,\mathrm{e}^{-st}\,f(\tau)\,d\tau\right]_{t=0}^{t=\infty} - \int_0^\infty \frac{1}{-s}\,\mathrm{e}^{-st}\,f(t)\,dt \tag{1.6.19}$$

The term between both limits is zero for $t = 0$ because $\int_0^0 \cdots = 0$, and for $t = \infty$ as well because the exponential function $\mathrm{e}^{-\infty} = 0$. Thus only the last integral remains, from which we can factor out the term $1/s$. The result is:

$$\mathcal{L}\left\{\int_0^t f(\tau)\,d\tau\right\} = \frac{1}{s}\int_0^\infty f(t)\,\mathrm{e}^{-st}\,dt = \frac{F(s)}{s} \tag{1.6.20}$$

and thus the statement expressed by Eq. 1.6.16 is proved.

Example:

$$f(t) = \mathrm{e}^{-\sigma_1 t}\sin\omega_1 t \tag{1.6.21}$$

We have already calculated the transform of this function (Eq. 1.5.10) and it is:

$$F(s) = \frac{\omega_1}{(s+\sigma_1)^2 + \omega_1^2}$$

Let us now calculate the integral of this function according to Eq. 1.6.20 by introducing a dummy variable τ:

$$\mathcal{L}\left\{\int_0^t \mathrm{e}^{-\sigma_1\tau}\sin\omega_1\tau\,d\tau\right\} = \frac{F(s)}{s} = \frac{\omega_1}{s\left[(s+\sigma_1)^2 + \omega_1^2\right]} \tag{1.6.22}$$

This expression describes the step response of a network, having a complex conjugate pole pair. We meet such functions very often in the analysis of inductive peaking circuits or in calculating the step response of an amplifier with negative feedback.

We may obtain the transform of multiple integrals by repeating the procedure expressed by Eq. 1.6.16.

By doing so we obtain for the

single integral: $$\mathcal{L}\left\{\int_0^t f(\tau)\,d\tau\right\} = \frac{F(s)}{s}$$

double integral: $$\mathcal{L}\left\{\int_0^t\int_0^{\tau_1} f(\tau)\,d\tau\right\} = \frac{F(s)}{s^2}$$

triple integral: $$\mathcal{L}\left\{\int_0^t\int_0^{\tau_1}\int_0^{\tau_2} f(\tau)\,d\tau\right\} = \frac{F(s)}{s^3}$$

n^{th} integral: $$\mathcal{L}\left\{\underbrace{\int_0^t\cdots\int_0^{\tau_{n-1}}}_{n\text{ integrals}} f(\tau)\,d\tau\right\} = \frac{F(s)}{s^n} \tag{1.6.23}$$

The $\mathcal{L}$ transform of the integral of the function $f(t)$ gives the complex function $F(s)/s$. The function $F(s)$ must be divided by s as many times as we integrate.

Here again we see a great advantage of the $\mathcal{L}$ transform, for **we can replace the integration in the time domain** (often a rather demanding procedure) **by a simple division by s in the (complex) frequency domain.**

1.6.5 Change of Scale

We have the function:

$$\mathcal{L}\{f(a\,t)\} = \int_0^\infty f(a\,t)\,\mathrm{e}^{-st}\,dt \tag{1.6.24}$$

We introduce a new variable $v = a\,t$, and for this $dv = a\,dt$ and also $t = v/a$. Thus we obtain:

$$\mathcal{L}\{f(a\,t)\} = \frac{1}{a}\int_{v=0}^\infty f(v)\,\mathrm{e}^{-\frac{st}{a}}\,dv = \frac{1}{a}\,F\left(\frac{s}{a}\right) \tag{1.6.25}$$

Example: we have the function:

$$f(t) = t\,\mathrm{e}^{-3\,t} \tag{1.6.26}$$

We have already calculated the $\mathcal{L}$ transform of a similar function by Eq. 1.5.15. For the function above the result is:

$$F(s) = \frac{1}{(s+3)^2} \tag{1.6.27}$$

Now let us change the scale tenfold. The new function is:

$$g(t) = f(10\,t) = 10\,t\,\mathrm{e}^{-30\,t} \tag{1.6.28}$$

According to Eq. 1.6.25 it follows that:

$$\mathcal{L}\{g(t)\} = \frac{1}{10}\,F\Big(\frac{s}{10}\Big) = \frac{1}{10\left(\frac{s}{10}+3\right)^2} = \frac{10}{(s+30)^2} \tag{1.6.29}$$

1.6.6 Impulse $\delta(t)$

In Fig. 1.6.1a we have a square pulse A_1 with amplitude 1 and duration $t = 1$. The surface S_1 under the pulse, equal to the time integral of this pulse, is *amplitude* × *time* and thus equal to 1. It is obvious that we may obtain the same time integral if the duration of the pulse is halved and its amplitude doubled (A_2). The pulse A_4 has a four times higher amplitude and its duration is only $t = 0.25$ and still has the same time-integral.

If we keep narrowing the pulse and adjusting the amplitude accordingly to keep the value of the time integral 1, we eventually arrive at a situation where the duration of the pulse becomes infinitely small, $t = \varepsilon \to 0$, and its amplitude infinitely large, $A = (1/\varepsilon) \to \infty$, as shown in Fig. 1.6.1b.

This impulse is denoted $\delta(t)$ and it is called the *Dirac*[2] *function*. Mathematically we express this function as:

$$\delta(t) = f_\varepsilon(t) = \left\{ \begin{array}{cll} 1/\varepsilon & \text{when} & 0 \le t \le \varepsilon \\ 0 & \text{when} & t > \varepsilon \end{array} \right|_{\varepsilon \to 0} \tag{1.6.30}$$

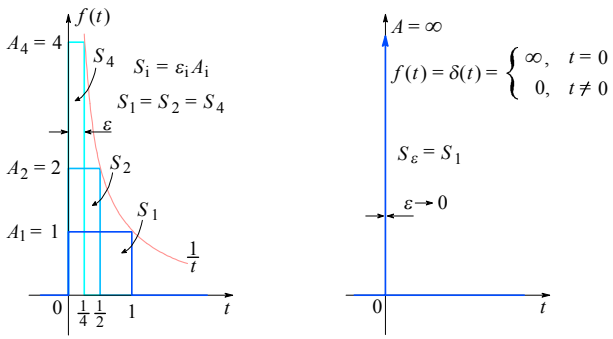


Fig. 1.6.1: The Dirac function as the limiting case of narrowing the pulse width, while keeping the time integral constant: **a)** If the pulse length is decreased, its amplitude must increase accordingly. **b)** When the pulse length $\varepsilon \to 0$ the amplitude is $(1/\varepsilon) \to \infty$.

Let us calculate the $\mathcal{L}$ transform of this function:

$$\mathcal{L}\{f_\varepsilon(t)\} = \int_0^\infty f_\varepsilon(t)\,\mathrm{e}^{-st}\,dt = \int_0^\varepsilon \frac{1}{\varepsilon}\,\mathrm{e}^{-st}\,dt + \int_\varepsilon^\infty (0)\,\mathrm{e}^{-st}\,dt$$

$$= -\frac{1}{s\,\varepsilon}\,\mathrm{e}^{-st}\Big|_{t=0}^{t=\varepsilon} = \frac{1-\mathrm{e}^{-s\,\varepsilon}}{s\,\varepsilon} \tag{1.6.31}$$

[2] *Paul Dirac*, 1902–1984, English physicist, Nobel Prize winner in 1933 (together with Erwin Schrödinger).

Now we express the function $e^{-s\varepsilon}$ in this result by the following series:

$$\frac{1-e^{-s\varepsilon}}{s\varepsilon}=\frac{1-[1-s\varepsilon+(s\varepsilon)^2/2!-(s\varepsilon)^3/3!+\cdots]}{s\varepsilon}=1-\frac{s\varepsilon}{2!}+\frac{(s\varepsilon)^2}{3!}-\cdots$$

and by letting $\varepsilon\to 0$ we obtain:

$$\mathcal{L}\{\delta(t)\}=\lim_{\varepsilon\to 0}\frac{1-e^{-s\varepsilon}}{s\varepsilon}=\lim_{\varepsilon\to 0}\left[1-\frac{s\varepsilon}{2!}+\frac{(s\varepsilon)^2}{3!}-\cdots\right]=1 \qquad (1.6.32)$$

Therefore the magnitude of the spectrum envelope of this function is **one** and it is independent of frequency. This means that the Dirac impulse $\delta(t)$ contains **an infinite number of frequency components**, the amplitude of each component being $A=1$.

1.6.7 Initial and Final Value Theorems

The **initial value theorem** is expressed as:

$$\lim_{0\leftarrow t} f(t)=\lim_{s\to\infty} s\,F(s) \qquad (1.6.33)$$

We have written the notation $0\leftarrow t$ in order to emphasize that t approaches zero from the right of the coordinate system. From real differentiation we know that:

$$\mathcal{L}\left\{\frac{d\,f(t)}{dt}\right\}=\mathcal{L}\{f'(t)\}=\int_0^\infty f'(t)\,e^{-st}\,dt=s\,F(s)-f(0^+) \qquad (1.6.34)$$

The limit of this integral when $s\to\infty$ is zero:

$$\lim_{s\to\infty}\int_0^\infty f'(t)\,e^{-st}\,dt=0 \qquad (1.6.35)$$

If we assume that $f(t)$ is continuous at $t=0$ we may write the limit of the right hand side of Eq. 1.6.33:

$$0=\lim_{s\to\infty} s\,F(s)-f(0^+) \qquad (1.6.36)$$

or, in a form more useful for practical calculations:

$$\boxed{\lim_{s\to\infty} s\,F(s)=f(0^+)+\lim_{0\leftarrow t} f(t)=\lim_{0\leftarrow t} f(t)} \qquad (1.6.37)$$

Even if $f(t)$ is not continuous at $t=0$, this relation is still valid, although the proof is slightly more difficult [Ref. 1.10]. The expression $f(0^+)$ is introduced because we are dealing with a unilateral transform, in which it is assumed that $f(t)=0$ for $t<0$, so to calculate the actual initial value we must approach it from the positive side of the time axis.

For the functions which we will discuss in the rest of the book we can, in a similar way, prove the **final value theorem**, which is stated as:

$$\lim_{t\to\infty} f(t)=\lim_{s\to 0} s\,F(s) \qquad (1.6.38)$$

(note that for some functions, such as $\sin\omega_1 t$ or $\cos\omega_1 t$ or the square wave, this limit does not exist, since the value oscillates with the time integral of the function!).

We repeat the statement from Eq. 1.6.34:

$$\mathcal{L}\{f'(t)\} = \int_0^\infty f'(t)\,\mathrm{e}^{-st}\,dt = s\,F(s) - f(0^+) \tag{1.6.39}$$

Now let $s \to 0$ (using q as an intermediate dummy variable):

$$\lim_{s\to 0}\int_0^\infty f'(t)\mathrm{e}^{-st}\,dt = \int_0^\infty f'(t)\,dt = \lim_{q\to\infty}\int_0^q f'(t)\,dt$$

$$= \lim_{q\to\infty}[f(q) - f(0^+)] = \lim_{t\to\infty} f(t) - f(0^+) \tag{1.6.40}$$

Although the lower limit of the integral is a (simple) zero we have nevertheless written 0^+ in the result, to emphasize the unilateral transform. The limit of the right hand side of Eq. 1.6.39, when $s \to 0$ is:

$$\lim_{s\to 0} s\,F(s) - f(0^+) \tag{1.6.41}$$

By comparing the results of Eq. 1.6.34, 1.6.39 and 1.6.41 we may write:

$$\lim_{t\to\infty} f(t) - f(0^+) = \lim_{s\to 0} s\,F(s) - f(0^+) \tag{1.6.42}$$

or as stated initially:

$$\boxed{\lim_{t\to\infty} f(t) = \lim_{s\to 0} s\,F(s)} \tag{1.6.38, again}$$

The Eq. 1.6.37 and 1.6.38 are extremely useful for checking the results of complicated calculations by the direct or the inverse Laplace transform, as we will encounter in the following parts of the book. Should the check by these two equations fail, then we have obviously made a mistake somewhere.

However, this is a necessary, but not a sufficient condition: if the check was passed we are not guaranteed that other 'sneaky' mistakes will not exist, which may become obvious when we plot the resulting function.

1.6.8 Convolution

We need a process by which we can calculate the response of two systems connected so that the output of the first one is the input of the second one and their individual responses are known. We have two functions [Ref. 1.19]:

$$f(t) = \mathcal{L}^{-1}\{F(s)\} \qquad \text{and} \qquad g(t) = \mathcal{L}^{-1}\{G(s)\} \tag{1.6.43}$$

and we are looking for the inverse transform of the product:

$$y(t) = \mathcal{L}^{-1}\{F(s)\cdot G(s)\} \tag{1.6.44}$$

The product of functions is equal to the product of their Laplace transforms:

$$F(s)\cdot G(s) = \int_0^\infty f(\tau)\,\mathrm{e}^{-s\tau}\,d\tau \cdot \int_0^\infty g(v)\,\mathrm{e}^{-sv}\,dv \tag{1.6.45}$$

In order to distinguish better between $f(t)$ and $g(t)$, we assign the letter u for the argument of f and v for the argument of g; thus, $f(t) \to f(u)$ and $g(t) \to g(v)$. Since both variables are now well separated we may write the above integral also in the form:

$$F(s) \cdot G(s) = \int_0^\infty \left[\int_0^\infty f(u)\, g(v)\, \mathrm{e}^{-s(u+v)}\, dv \right] du \tag{1.6.46}$$

Let us integrate the expression inside the brackets to the variable v. To do so we introduce a new variable τ:

$$\tau = u + v \qquad \text{so} \qquad v = \tau - u \qquad \text{and} \qquad dv = d\tau \tag{1.6.47}$$

We consider the variable τ in the inner integral to be a (variable) parameter. From the above expressions it follows that $v = 0$ if $\tau = u$. By considering all this we may transform Eq. 1.6.46 into:

$$F(s) \cdot G(s) = \int_0^\infty \left[\int_\tau^\infty f(u)\, g(\tau - u)\, \mathrm{e}^{-s(u+\tau-u)}\, d\tau \right] du \tag{1.6.48}$$

We may also change the sequence of integration. Thus we may choose a fixed t_1 and first integrate from $\tau = 0$ to $\tau = t_1$. In the second integration we integrate from $u = 0$ to $u = \infty$. Then the above expression obtains the form:

$$F(s) \cdot G(s) = \int_0^\infty \left[\int_0^{t_1} f(u)\, g(\tau - u)\, du \right] \mathrm{e}^{-s\tau}\, d\tau \tag{1.6.49}$$

Now we can return from u back to the usual time variable t:

$$F(s) \cdot G(s) = \int_0^\infty \underbrace{\left[\int_0^{t_1} f(t)\, g(\tau - t)\, dt \right]}_{y(t)} \mathrm{e}^{-s\tau}\, d\tau \tag{1.6.50}$$

The expression inside the brackets is the function $y(t)$ which we are looking for, whilst the outer integral is the usual Laplace transform. Thus we define the convolution process, denoted by $g(t) * f(t)$, as:

$$\boxed{y(t) = \mathcal{L}^{-1}\{F(s) \cdot G(s)\} = g(t) * f(t) = \int_0^{t_1} f(t)\, g(\tau - t)\, dt} \tag{1.6.51}$$

The operator symbolized by the asterisk $(*)$ means 'convolved with'. *Convolutio* is the Latin word for folding. The German name for convolution is *die Faltung* and this also means folding. Obviously:

$$g(\tau - t) = g(-t)\Big|_{\tau=0}$$

and

$$g(\tau - t) = g(0)\Big|_{\tau=t}$$

This means that the function is 'folded' in time around the ordinate, from the right to the left side of the coordinate system. At the end of this part, after we master the network analysis in Laplace space, we will make an example (Fig. 1.15.1) in which this 'folding' and the convolution process will be explicitly shown, step by step.

In general we convolve whichever of the two functions is simpler. We may do so because the convolution is commutative:

$$g(t) * f(t) = f(t) * g(t) \tag{1.6.52}$$

The main properties of the Laplace transform are listed in Table 1.6.1.

Table 1.6.1: **The main properties of Laplace transform**

Property	$f(t)$	$F(s)$
Real Differentiation	$\frac{d\,f(t)}{dt}$	$s\,F(s) - f(0^+)$
Real Integration	$\int_0^{t_1} f(t)\,dt$	$\frac{F(s)}{s}$
Time-Scale Change	$f(at)$	$\frac{1}{a}\,F\left(\frac{s}{a}\right)$
Impulse function	$\delta(t)$	1
Initial Value	$\lim_{0 \leftarrow t} f(t)$	$\lim_{s \to \infty} s\,F(s)$
Final Value	$\lim_{t \to \infty} f(t)$	$\lim_{s \to 0} s\,F(s)$
Convolution	$\int_0^{t_1} f(t)\,g(\tau - t)\,dt$	$F(s) \cdot G(s)$

1.7 Application of the $\mathcal{L}$ transform in Network Analysis

1.7.1 Inductance

As we have discussed the fundamentals of $\mathcal{L}$ transform, we will now apply it in the network analysis. From basic electrical engineering we know that the instantaneous voltage $v(t)$ across an inductance L, through which a current $i(t)$ flows, as in Fig. 1.7.1a, is:

$$v(t) = \frac{di}{dt} L \tag{1.7.1}$$

By assuming time $t > 0$ and $i(0^+) = 0$, then, according to Eq. 1.6.5, the $\mathcal{L}$ transform of the above equation is the voltage across the inductance in the s domain:

$$V(s) = s\,L\,I(s) \tag{1.7.2}$$

where $I(s)$ is the current in the s domain. The inductive reactance is then:

$$\boxed{\frac{V(s)}{I(s)} = s\,L} \tag{1.7.3}$$

Here $s = \sigma + j\omega$ and thus it is complex; it can lie anywhere in the s plane. In the **special case** when $\sigma = 0$, and considering only the positive $j\omega$ axis, s degenerates into $j\omega$. Then the inductive reactance becomes the familiar $j\omega L$, as is known from the usual 'phasor' analysis of networks.

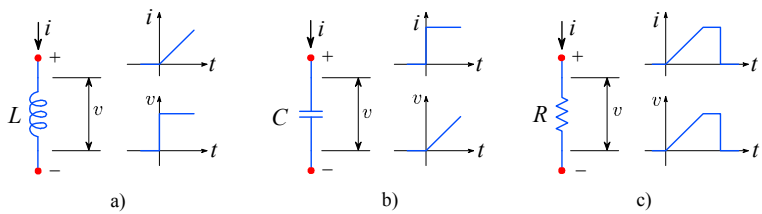


Fig. 1.7.1: The instantaneous voltage v as a function of the instantaneous current i: **a)** on an inductance L; **b)** on a capacitance C; **c)** on a resistance R.

1.7.2 Capacitance

From basic electrical engineering we also know that the instantaneous voltage $v(t)$ across a capacitance through which a current $i(t)$ flows during a time $t \geq 0$ is:

$$v(t) = \frac{q(t)}{C} = \frac{1}{C}\int_0^t i\,dt \tag{1.7.4}$$

as shown in Fig. 1.7.1b. Here $q(t)$ is the instantaneous charge on the capacitor C. By applying Eq. 1.6.20 we may calculate the voltage on the capacitor in the s domain:

$$V(s) = \frac{1}{s\,C}\,I(s) \tag{1.7.5}$$

The capacitive reactance in the s domain is:

$$\boxed{\frac{V(s)}{I(s)} = \frac{1}{s\,C}} \tag{1.7.6}$$

Here, too, s degenerates to $j\omega$ if $\sigma = 0$. In this case the capacitive reactance becomes simply $1/j\omega C$.

1.7.3 Resistance

For a resistor (Fig. 1.7.1c) the instantaneous voltage is simply:

$$v(t) = R\,i(t) \tag{1.7.7}$$

and, as there are no time-derivatives the same holds in the s domain, with the corresponding values $V(s)$ and $I(s)$:

$$V(s) = R\,I(s) \tag{1.7.8}$$

yielding:

$$\boxed{\frac{V(s)}{I(s)} = R} \tag{1.7.9}$$

1.7.4 Resistor and capacitor in parallel

By applying the Eq. 1.7.3, 1.7.6 and 1.7.9 we may transform a differential equation in the t domain into an algebraic equation in the s domain. Thus we may express an impedance $Z(s)$ or an admittance $Y(s)$ of more complicated networks by simple algebraic equations. Let us express a parallel combination of a resistor and a capacitor in s domain as shown in the upper part of Fig. 1.7.2. The impedance is:

$$Z(s) = \frac{1}{s\,C + \dfrac{1}{R}} = R\,\frac{-\left(-\dfrac{1}{RC}\right)}{s - \left(-\dfrac{1}{RC}\right)} = R\,\frac{-s_1}{s - s_1} = R\,G_1(s) \tag{1.7.10}$$

where the (real) pole is at $s_1 = \sigma_1 = -1/RC$ and G_1 represents the frequency dependence. The pole of a function is that particular value of the argument for which the function denominator is equal to zero and, consequently, the function value goes to infinity.

Now let us apply a current step, $I(s) = 1\,\mathrm{V}/R$, to our network expressed in the s domain as $1/(sR)$, according to Eq. 1.5.1. We introduced the factor $1/R$ in order to get a voltage of 1 V on our RC combination when $t \to \infty$. The corresponding function is then:

$$\begin{aligned} F(s) = V(s) = I(s) \cdot Z(s) &= \frac{1}{sR} \cdot R\,G_1(s) = \frac{1}{s}\,G_1(s) \\ &= \frac{1}{RC} \cdot \frac{1}{s} \cdot \frac{1}{s - \left(-\dfrac{1}{RC}\right)} \end{aligned} \tag{1.7.11}$$

From $1/s$ a second pole at $s = 0$ is introduced, as is drawn in Fig. 1.7.2b and c. To obtain the time domain function of the voltage across our impedance, we apply Eq. 1.5.3 and 1.6.20. First we discuss only the function $G_1(s)$. According to Eq. 1.5.3:

$$\mathcal{L}\{e^{\sigma_1 t}\} = \frac{1}{s - \sigma_1} \tag{1.7.12}$$

or inversely:

$$\mathcal{L}^{-1}\left\{\frac{1}{s - \sigma_1}\right\} = e^{\sigma_1 t} \tag{1.7.13}$$

By comparing Eq. 1.7.11 with Eq. 1.7.13 we see that $\sigma_1 = -1/RC$:

$$\mathcal{L}^{-1}\left\{\frac{1}{s - (-1/RC)}\right\} = e^{-t/RC} \tag{1.7.14}$$

From Eq. 1.6.20 we concluded that the division in the s domain corresponds to the real integration in the t domain. By considering this together with Eq. 1.5.1, we obtain:

$$v_o(t) = f(t) = \mathcal{L}^{-1}\{F(s)\} = \mathcal{L}^{-1}\left\{\frac{1/RC}{s\,[s - (-1/RC)]}\right\} = \frac{1}{RC}\int_0^t e^{-t/RC}\,dt$$

$$= \frac{1}{RC}\left[-RC\left(e^{-t/RC}\right)\Big|_0^t\right] = -e^{-t/RC} - (-1) = 1 - e^{-t/RC} \tag{1.7.15}$$

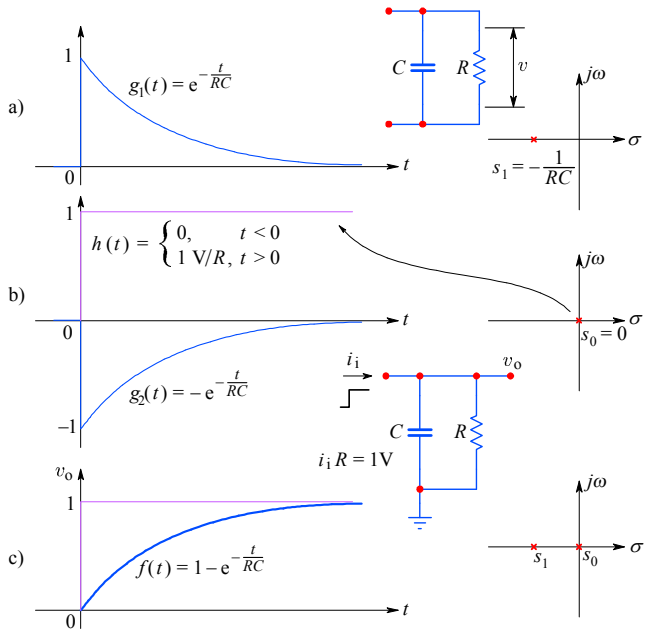


Fig. 1.7.2: The course of mathematical operations for a parallel RC network excited by a unit step current i_i. The t domain functions are on the left, the s domain functions are on the right. **a)** The self-discharge network function is equal to the impulse function $g_1(t)$. **b)** The unit step in t domain, $h(t)$, is represented by a pole at the origin (s_0) in the s domain. The function $g_2(t)$ is the reaction of the network to the unit step excitation. **c)** The output voltage is the sum of both functions, $v_o = f(t) = h(t) + g_2(t)$.

From this simple example we obtain the idea of how to use tables of $\mathcal{L}$ transforms to obtain the response in the t domain, which should otherwise be calculated by differential equations. In addition to this we may state a very important conclusion for the s-domain:

$$\boxed{\Big(\textit{output function}\Big) = \Big(\textit{excitation function}\Big) \times \Big(\textit{network function}\Big)}$$

In our case it was:

excitation function $$H(s) = \frac{1}{sR}$$

network function $$R\,G_1(s) = R \cdot \frac{\dfrac{1}{RC}}{s + \dfrac{1}{RC}} \qquad \left(\begin{array}{c}\text{also named}\\ \text{'}\textit{impulse response}\text{'}\end{array}\right)$$

output function $$F(s) = \frac{1}{s} \cdot \frac{\dfrac{1}{RC}}{s + \dfrac{1}{RC}}$$

However, in general, especially for more complicated networks, the calculation of the corresponding function in the t domain is not as easy as shown above. Of course, one may always apply the formula for the inverse Laplace transform (Eq. 1.4.4):

$$f(t) = \mathcal{L}^{-1}\{F(s)\} = \frac{1}{2\,\pi\,j} \int_{c-j\infty}^{c+j\infty} F(s)\,\mathrm{e}^{st}\,ds$$

but it would not be fair to leave the reader to grind through this integral of his $F(s)$ with the best of his/her knowledge. In essence the above expression is a *contour integral*. **Knowledge of contour integration is a necessary prerequisite for calculating the inverse Laplace transform.** We will discuss this in the following section. After studying it the reader will realize that the calculation of the step-response in the t domain by contour integration is — although a little more difficult than in the above example of the simple RC circuit — still a relatively simple procedure.

1.8 Complex Line Integrals

In order to learn how to calculate contour integrals the first step is the calculation of complex line integrals. Both require a knowledge of the basics of complex variable theory, also called *the theory of analytic functions*. We will discuss only that part of this theory which is relevant to the inverse Laplace transform of rational functions (which are important in the calculation of amplifier step and impulse response). The reader who would like to know more about the complex variable theory, should study at least one of the books listed at the end of Part 1 [Ref. 1.4, 1.9, 1.10, 1.11, 1.12, 1.13, 1.14, 1.15], of which [Ref. 1.10 and 1.13] (in English), [Ref. 1.4] (in German), and [Ref. 1.11] (in Slovenian) are especially recommended.

The definition of an analytical function is:

In a certain domain (which we are interested in) a function $f(z)$ is analytical if it is:

1) continuous;
2) single valued (at each argument value); and
3) has a derivative at any selected point z, independently of from which side we approach that point.

From the calculus we know that a definite integral of a function of a **real** variable, such as $y = f(x)$, is equal to the area A between the function (curve) and the real axis x and between both limits x_1 to x_2. An example is shown in Fig. 1.8.1, where the integral of a simple function $1/x$, integrated from x_1 to x_2 is displayed.

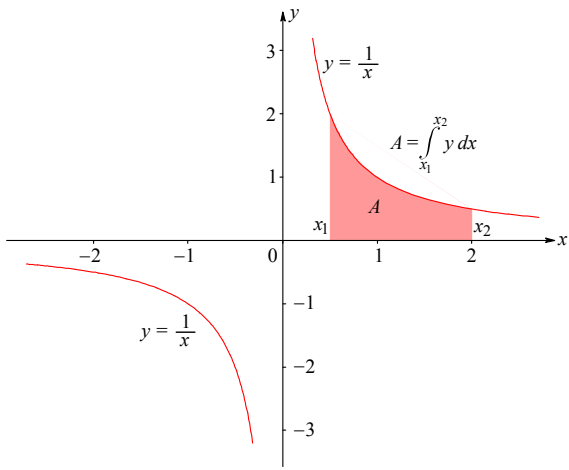


Fig. 1.8.1: The integral of a real function $y = 1/x$ between the limits x_1 and x_2 corresponds to the area A.

The corresponding mathematical expression is:

$$A = \int_{x_1}^{x_2} \frac{1}{x}\, dx = F(x_2) - F(x_1) = \ln x_2 - \ln x_1 = \ln \frac{x_2}{x_1} \qquad (1.8.1)$$

The area above the x axis is counted as positive and the area below the x axis (if any) as negative. The area A in Fig. 1.8.1 represents the difference of the integral values at the upper limit, $F(x_2)$ and the lower limit, $F(x_1)$. As shown in Fig. 1.8.1 the integration path was from x_1 along the x axis up to x_2.

For a comparison let us now calculate a similar integral, but with a **complex** variable $z = x + j\,y$:

$$W = \int_{z_1}^{z_2} \frac{1}{z}\,dz = F(z_2) - F(z_1) = \ln z_2 - \ln z_1 = \ln \frac{z_2}{z_1} \tag{1.8.2}$$

So far we can not see any difference between Eq. 1.8.1 and 1.8.2 (a close investigation of the result would show that it may be multi-valued in the case the path from z_1 to z_2 circles the pole one or more times; but we will not discuss such cases). The whole integration procedure is the same in both cases. The difference in the result of the second equation becomes apparent when we express the complex variable z in the exponential form:

$$z_1 = |z_1|\,\mathrm{e}^{j\,\theta_1} \qquad \text{and} \qquad z_2 = |z_2|\,\mathrm{e}^{j\,\theta_2} \tag{1.8.3}$$

then:

$$\ln \frac{z_2}{z_1} = \ln \frac{|z_2|\,\mathrm{e}^{j\,\theta_2}}{|z_1|\,\mathrm{e}^{j\,\theta_1}} = \ln \frac{|z_2|}{|z_1|}\,\mathrm{e}^{j(\theta_2-\theta_1)} = \ln \frac{|z_2|}{|z_1|} + j(\theta_2 - \theta_1) = u + j\,v \tag{1.8.4}$$

where: $$u = \ln \frac{|z_2|}{|z_1|} \qquad \text{and} \qquad v = \theta_2 - \theta_1 \tag{1.8.5}$$

and also: $$|z_i| = \sqrt{x_i^2 + y_i^2} \qquad \text{and} \qquad \theta_i = \arctan \frac{y_i}{x_i} \tag{1.8.6}$$

Obviously the result of Eq. 1.8.2, as shown in Eq. 1.8.4, is complex. It can not be plotted as simply as the integral of Fig. 1.8.1, since for displaying the complex function of a complex argument we would need a 4D graph, whilst the present state of technology allows us to plot only a 2D projection of a 3D graph, at best.

We can, however, restrict the z argument's domain, as in Fig. 1.8.2, by making its real part a constant, say, $x = c$ and then make plots of $F(c + j\,y) = 1/(c + j\,y)$ for some selected value of c. In Fig. 1.8.2 we have chosen $c = 0$ and $c = 0.5$, whilst the imaginary part was varied from $-j\,3$ to $+j\,3$.

In this way we have plotted two graphs, labeled A and B. The graph A belongs to $c = 0$ and lies in the $\Im\{s\} \times \Im\{F(s)\}$ plane; it looks just like the one in Fig. 1.8.1, but changed in sign, owing to the following rationalization of the function's denominator:

$$\frac{1}{j\,y} = \frac{j}{j^2\,y} = \frac{j}{-1 \cdot y} = -\frac{j}{y}$$

The graph B belongs to $c = 0.5$ and is a 3D curve, twisting in accordance with the phase angle of the function. To aid the 3D view the three projections of B have also been plotted.

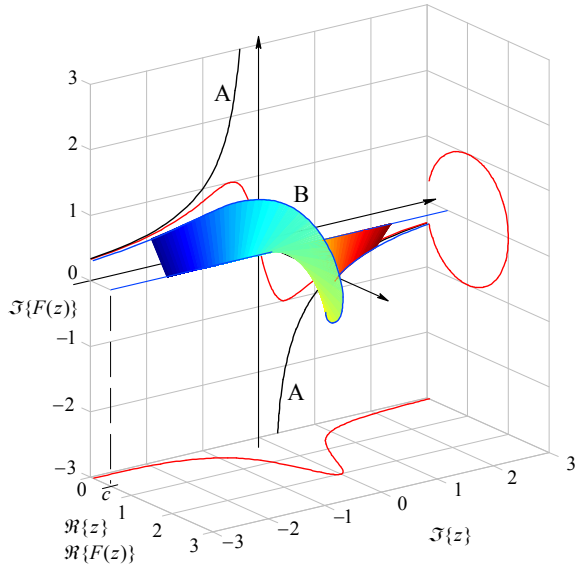


Fig. 1.8.2: By reducing the complex domain $x + jy$ to $c + jy$, where c is a constant, we can plot the complex function $F(c + jy)$ in a 3D graph. Here we have $c = 0$ (graph A) and $c = 0.5$ (graph B). Also shown are the three projections of B. The twisted surface is the integral of $F(c + jy)$ for $c = 0.5$ and y in the range $-j2 < y < j2$. See Appendix 1 (web only) for details.

Let us determine a few characteristic points of the graph B:

a) for the first point on the left we have $c = 0.5$ and $y = -3$, thus:

$$F(c + jy) = \frac{1}{0.5 - j3} = \frac{0.5 + j3}{(0.5 - j3)(0.5 + j3)}$$

$$= \frac{0.5 + j3}{0.5^2 + 3^2} = \frac{0.5 + j3}{0.25 + 9} = \frac{0.5}{9.25} + j\frac{3}{9.25} = 0.0541 + j0.3243$$

b) next, let us have $c = 0.5$ and $y = -0.5$, thus:

$$F(c + jy) = \frac{1}{0.5 - j0.5} = \frac{0.5 + j0.5}{(0.5 - j0.5)(0.5 + j0.5)}$$

$$= \frac{0.5 + j0.5}{0.5^2 + 0.5^2} = \frac{0.5 + j0.5}{0.25 + 0.25} = \frac{0.5}{0.5} + j\frac{0.5}{0.5} = 1 + j$$

(here both the real and the imaginary part are 1 — this is the top point of the curve).

c) an obvious choice is $c = 0.5$ and $y = 0$, thus:

$$F(c + jy) = \frac{1}{0.5 - j0} = \frac{1}{0.5} = 2$$

(here the real part is 2, the imaginary part is 0 and this is the rightmost point on the curve; also, it is its only real value point).

For positive imaginary values, $F(z)$ is the complex conjugate of the values above.

Now that we have some idea of how $F(z)$ looks, let us return to our integral problem. If the integration path is parallel to the imaginary axis, $-j2 < y < +j2$, and displaced by $x = \Re\{z\} = c = 0.5$, the result of integration would be the surface indicated in Fig. 1.8.2. But for an arbitrary path, with x not constant, we should make many such plots as above and then trace the integration path to appropriate curves. The area bounded by the integration path and its trace on those curves would be the result we seek.

For a detailed treatment of complex function plotting see Appendix 1 (web only).

Returning to the result of Eq. 1.8.4 we may draw an interesting conclusion:

The complex line integral depends only on the initial value z_1 and the final value z_2, which represent both limits of the integral.

The result of the integration is independent of the actual path beneath these limits, providing that the path lies on the same side of the pole.

All the significant differences between an integral of a real function and the line integral of a complex function are listed in Table 1.8.1.

The x axis is the argument's domain for a real integral, whilst for a complex integral it is the whole z plane. Do not confuse the z plane (the complex plane, $x + jy$, with the diagram's z axis (vertical axis), which here is $F(z) = F(x + jy)$. We recommend the readers to ponder over Fig. 1.8.2 and try to acquire a clear idea of the differences between both types of integral, since this is necessary for the understanding of the discussion which follows.

Table 1.8.1 Differences between real and complex integration

	real variable	complex variable
integral	$\int_{x_1}^{x_2} \frac{1}{x}\, dx$	$\int_{z_1}^{z_2} \frac{1}{z}\, dz$
independent variable	x	$z = x + j\,y$
dependent variable	$y = \frac{1}{x}$	$w = \frac{1}{z} = u + j\,v = \frac{x}{x^2 + y^2} - j\,\frac{y}{x^2 + y^2}$
integration path	from x_1 to x_2 along the x axis	from $z_1 = x_1 + j\,y_1$ to $z_2 = x_2 + j\,y_2$ anywhere in the z plane*
result	$\ln \frac{x_2}{x_1}$ **(real)**	$\ln \frac{z_2}{z_1} = \ln \frac{\lvert z_2 \rvert}{\lvert z_1 \rvert} + j\,(\theta_2 - \theta_1)$ **(complex)**

* except through the pole, where $z = 0$

To understand the theory better let us give a few examples:

1.8.1 Example 1

We have a function $f(z) = 3z$ which we shall integrate from $2j$ to $1 + j$:

$$\int_{2j}^{1+j} 3z\,dz = \frac{3\,z^2}{2}\bigg|_{2j}^{1+j} = \frac{3}{2}\left[(1+j)^2 - (2\,j)^2\right] = 6 + 3\,j$$

1.8.2 Example 2

The integration limits are the same as in the previous example, whilst the function is different, $f(z) = 1 + z^2$:

$$\int_{2j}^{1+j} \left(1 + z^2\right) dz = z\bigg|_{2j}^{1+j} + \frac{z^3}{3}\bigg|_{2j}^{1+j} = \frac{1}{3} + \frac{7}{3}\,j$$

1.8.3 Example 3

The same function as in Example 1, except that both limits are interchanged:

$$\int_{1+j}^{2\,j} 3\,z\,dz = \frac{3\,z^2}{2}\bigg|_{1+j}^{2\,j} = \frac{3}{2}\left[(2\,j)^2 - (1+j)^2\right] = -6 - 3\,j$$

We see that although the function under the integral is complex, the same rules apply for integration as for a function of a real variable. The last example shows us that if the limits of the integral of a complex function are exchanged the result of the integration changes the sign.

As already mentioned, the result of the integration of a complex function is independent of the actual path of integration between the limits z_1 and z_2 (see Fig. 1.8.3), provided that no pole lies between the extreme paths L_1 and L_2. Thus for all the paths shown the result of integration is the same. This means that the function in the area between L_1 and L_2 is *analytic*. When at least one pole of the function lies between L_1 and L_2, the integral along the path L_1 is in general no more equal to the integral along the path L_2. In Fig. 1.8.4 we show such a case, in which the function is non-analytic (or *non-regular*) inside a small area between z_1 and z_2 (in the remaining area the function is analytic).

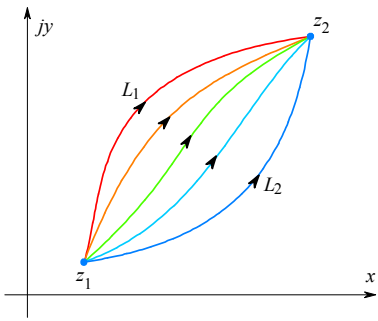


Fig. 1.8.3: A line integral from z_1 to z_2 along the line L_1, L_2, or any other line lying between these two yields the same result because between L_1 and L_2 the function has no pole.

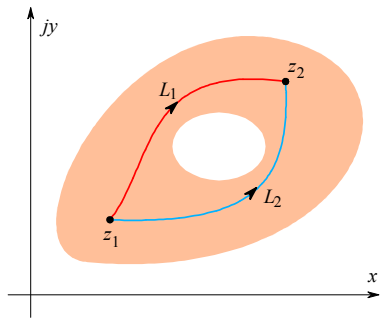


Fig. 1.8.4: Here the function has a non-analytic domain area between L_1 and L_2. Now the integral along the path L_1 is **not** equal to the integral along the path L_2.

Let us prove the above statement by two simple examples.

1.8.4 Example 4

We will again take the function $f(z) = 1/z$ and integrate along a part of a circle with the radius of 1, from $-j$ to $+1$ as is drawn in Fig. 1.8.5 (the pole z_0 lies at $z = 0$). We first calculate the integral:

$$\int_{-j}^{1} \frac{1}{z}\, dz \quad \text{along the path } L_1$$

On the circle with radius $|z| = 1$ it is:

$$z = (1) \cdot \mathrm{e}^{j\theta} \quad \text{and} \quad dz = j\,\mathrm{e}^{j\theta} d\theta$$

When we integrate along L_1 the angle θ goes from $-\pi/2$ to 0. Thus it is:

$$\int_{-\pi/2}^{0} \frac{j\,\mathrm{e}^{j\theta}}{\mathrm{e}^{j\theta}}\, d\theta = j\,\theta \Big|_{-\pi/2}^{0} = j\,\frac{\pi}{2}$$

1.8.5 Example 5

Here everything is the same as in the previous example, except that we will integrate along the path L_2 of Fig. 1.8.5. In this case the angle θ goes from $3\pi/2$ to 0:

$$\int_{3\pi/2}^{0} \frac{j\,\mathrm{e}^{j\theta}}{\mathrm{e}^{j\theta}}\, d\theta = j\,\theta \Big|_{3\pi/2}^{0} = -j\,\frac{3\pi}{2}$$

In Example 4, the integration path goes counterclockwise (which in mathematics is the positive sense) and we obtain a positive result. But in Example 5, in which the integration path goes clockwise, the result is negative, and, moreover, it has a different value, because the integration path lies on the other side of the pole, even if the limits of integration remain the same as in Example 4.

1.8.6 Example 6

We would like to see whether there is any difference in the result of Example 4 if we choose not to integrate along the circle, but instead along a straight line from $-j$ to $+1$ (L_3 in Fig. 1.8.6):

$$\int_{-j}^{1} \frac{1}{z}\, dz = \ln 1 - \ln(-j) = -\ln \mathrm{e}^{-j\pi/2} = j\,\frac{\pi}{2}$$

because $\ln 1 = 0$ and $-j = \mathrm{e}^{-j\pi/2}$. The result is the same as in Example 4.

In general if we consider Fig. 1.8.7, the integral along the path L_a or L_b, or any path in between, is always equal to $j\pi/2$. Similarly, the integral along the path L_c or L_d, or any path in between, is equal to $-j\,3\pi/2$.

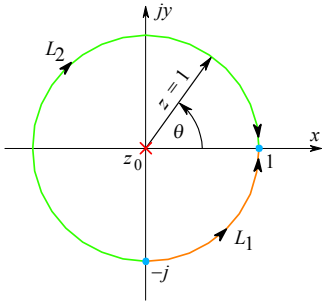


Fig. 1.8.5: The integral along the path L_1 is not equal to the integral along the path L_2 because the function has a pole which lies between both paths

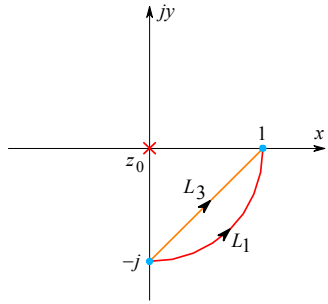


Fig. 1.8.6: The integral along the straight path L_3 is the same as the integral along the circular path L_1

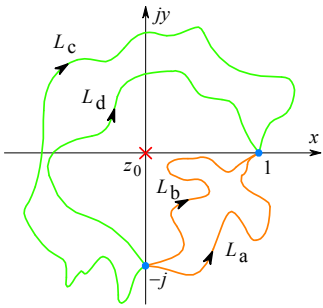


Fig. 1.8.7: The integrals along the paths L_a and L_b are equal to $j\pi/2$. However, those along L_c and L_d are equal to $-j3\pi/2$.

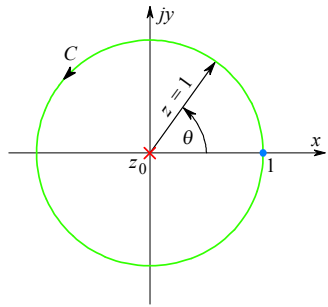


Fig. 1.8.8: The integral along the circular path C around the pole is $2\pi j$. See Sec.1.9.

1.9 Contour Integrals

Let us take again our familiar function $f(z) = 1/z$ and calculate the integral along the full circle C (Fig. 1.8.8), where $|z| = 1$. We use the same notation for z and dz as we did in Example 4 and start the integration at $\theta = 0$, going counterclockwise (positive by definition):

$$\int_0^{2\pi} \frac{dz}{z} = j\int_0^{2\pi} \frac{e^{j\theta}\,d\theta}{e^{j\theta}} = j\int_0^{2\pi} d\theta = 2\pi j = \oint_C \frac{dz}{z} \tag{1.9.1}$$

The resulting integral along the circle C is called *the contour integral*; the arrow in the symbol indicates the direction of encircling the pole (at $z = 0$).

Now let us move the pole from the origin to the point $a = x_a + j\,y_a$. The corresponding function is then $f(z) = 1/(z - a)$. The first attempt would be to integrate along the contour C as shown in Fig. 1.9.1. Inside this contour the domain of the function is analytic, except for the point a. Unfortunately C is a random contour and can not be expressed in a convenient mathematical way. Since a is the only pole inside the contour C, we may select another, simpler integration path. As we have already mastered the integration around a circular path, we select a circle C_c with the radius ε that lies inside the contour C. From Fig. 1.9.1 it is evident that:

$$\varepsilon = |z - a| \quad \text{or} \quad z - a = \varepsilon\, e^{j\theta} \tag{1.9.2}$$

Thus:

$$z = \varepsilon\, e^{j\theta} + a \tag{1.9.3}$$

where the angle θ can have any value in the range $0 \ldots 2\pi$. Furthermore it follows that:

$$dz = j\varepsilon\, e^{j\theta} d\theta \tag{1.9.4}$$

The contour integral around the pole a is then:

$$\oint_{C_c} \frac{dz}{z - a} = \int_0^{2\pi} \frac{j\varepsilon\, e^{j\theta} d\theta}{\varepsilon\, e^{j\theta}} = j\int_0^{2\pi} d\theta = 2\,\pi\,j \tag{1.9.5}$$

The result is the same as we have obtained for the function $f(z) = 1/z$, in which the pole was at the origin of the z plane.

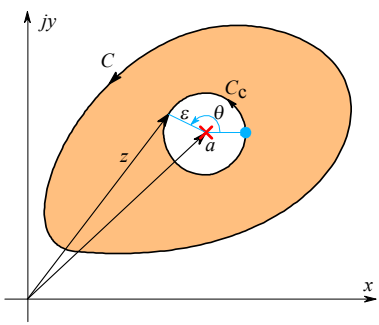


Fig. 1.9.1: Contour integral around the pole at a.

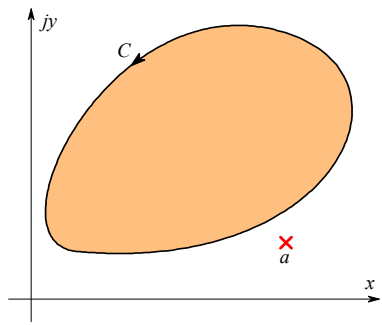


Fig. 1.9.2: The integral around the contour C is zero because a, the only pole of the function, lies outside the contour.

We look again at Fig. 1.8.3, where the integral around the path L_1 is equal to the integral around the path L_2 because there is no pole between L_1 and L_2. It would be interesting to make the integral from z_1 to z_2 along the path L_2 and then back again from z_2 to z_1 along the path L_1, making a closed loop (contour) integral:

$$\underbrace{\int_{z_1}^{z_2} f(z)\,dz}_{\text{along } L_2} + \underbrace{\int_{z_2}^{z_1} f(z)\,dz}_{\text{along } L_1} = \underbrace{\int f(z)\,dz}_{\text{along } L_2 + L_1} = 0 \tag{1.9.6}$$

Since both integrals have the same magnitude, by exchanging the limits of the second integral, thus making it negative, their sum is zero. This statement affords us the conclusion that the integral around the contour C in Fig. 1.9.2, which encircles an area where the function is analytic, is zero (the only pole a in the vicinity lies outside the contour of integration). This is expressed as:

$$\oint_C f(z)\,dz = 0 \tag{1.9.7}$$

The expressions in Eq. 1.9.6 and 1.9.8 were derived by the French mathematician *Augustine Louis Cauchy* (1788–1857). In all the calculations so far we have integrated in a counterclockwise sense, having the integration field, including the pole, always on the **left** side. In the case of a clockwise direction, let us again take Eq. 1.9.1 and integrate clockwise from 2π to 0:

$$\int_{2\pi}^{0} \frac{dz}{z} = j\int_{2\pi}^{0} \frac{e^{j\theta}\,d\theta}{e^{j\theta}} = j\int_{2\pi}^{0} d\theta = -2\pi j = \oint_C \frac{dz}{z} \tag{1.9.8}$$

Note that the sign of the result changes if we change the direction of encircling. So we may write in general:

$$\oint_C f(z)\,dz = -\oint_C f(z)\,dz \tag{1.9.9}$$

1.10 Cauchy's Way of Expressing Analytic Functions

Let us take a function $f(z)$ which is analytic inside a contour C. There are no regulations for the nature of $f(z)$ outside the contour, where $f(z)$ may also have poles. So this function is analytic also at the point a (inside C) where its value is $f(a)$.

Now we form another function:

$$g(z) = \frac{f(z)}{z-a} \tag{1.10.1}$$

This function is also analytic inside the contour C, except at the point a, where it has a pole, as shown in Fig. 1.10.1. Let us take the integral around the closed contour C:

$$\oint_C \frac{f(z)}{z-a}\, dz \tag{1.10.2}$$

which is similar to the integral in Eq. 1.9.5, except that here we have $f(z)$ in the numerator. Because at the point a the function under the integral is not analytic, the path of integration must avoid this point. Therefore we go around it along a circle of the radius ε, which can be made as small as required (but not zero).

For the path of integration we shall use the required contour C **and** the circle C_c. To make the closed contour the complete integration path will start at point 1 and go **counterclockwise** around the contour C to come back to the point 1; then from the point 1 to the point 2 along the dotted line; then **clockwise** around the circle C_c back to the point 2; and finally from the point 2 back to the point 1 along the dotted line. In this way, the contour of integration is closed. The integral from the point 1 to 2 and back again is zero. Thus there remain only the integrals around the contour C and around the circle C_c.

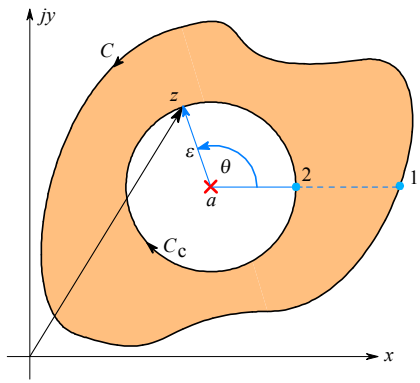


Fig. 1.10.1: Cauchy's method of expressing analytic functions (see text).

Since around the complete integration path the domain on the left hand side of the contour was always analytical, the resulting integral must be zero. Thus:

$$\oint_{C-C_c} \frac{f(z)}{z-a}\,dz = \oint_{C} \frac{f(z)}{z-a}\,dz + \oint_{C_c} \frac{f(z)}{z-a}\,dz = 0 \qquad (1.10.3)$$

and so it follows that:

$$\oint_{C} \frac{f(z)}{z-a}\,dz = \oint_{C_c} \frac{f(z)}{z-a}\,dz \qquad (1.10.4)$$

Here we have changed the second integral sign by reversing the sense of encircling.

Similarly as in Eq. 1.9.2 and 1.9.4 we write:

$$z-a = \varepsilon\, e^{j\theta} \qquad dz = j\varepsilon\, e^{j\theta} d\theta \qquad \text{and} \qquad \frac{dz}{z-a} = j\,d\theta$$

Nothing would change if in Eq. 1.10.4 we write:

$$f(z) = f(a) + [f(z) - f(a)] \qquad (1.10.5)$$

thus obtaining:

$$\oint_{C} \frac{f(z)}{z-a}\,dz = j\,f(a)\int_0^{2\pi} d\theta + j\int_0^{2\pi} [f(z) - f(a)]\,d\theta \qquad (1.10.6)$$

The integration must go from 0 to 2π in order to encircle the point a in the required direction. The value of the first integral on the right is:

$$j\,f(a)\int_0^{2\pi} d\theta = 2\pi j\,f(a) \qquad (1.10.7)$$

and we will prove that the second integral is zero. Its magnitude is:

$$M < 2\pi \max\{|f(z) - f(a)|\} \qquad (1.10.8)$$

The function $f(z)$ is continuous everywhere inside the field bordered by C and C_c; therefore the point z can be as close to the point a as desired. Consequently $|f(z) - f(a)|$ may also be as small as desired. The radius of the circle C_c inside the contour C is $\varepsilon = |z - a|$, and in Eq. 1.9.5 we have already observed that the value of the integral is independent of ε. If we take the limit $\varepsilon \to 0$ we obtain:

$$\lim_{\varepsilon \to 0} \int_0^{2\pi} [f(z) - f(a)]\,d\theta = 0 \qquad (1.10.9)$$

Thus:

$$2\pi j\,f(a) = \oint_{C} \frac{f(z)\,dz}{z-a} \qquad (1.10.10)$$

and:

$$\boxed{f(a) = \frac{1}{2\pi j}\oint_{C} \frac{f(z)\,dz}{z-a}} \qquad (1.10.11)$$

where the point a may be any point inside the contour C.

Eq. 1.10.11 is of essential importance for the inverse Laplace transform; we name it *Cauchy's expression for an analytic function*. By means of this integral it is possible to calculate the value of an analytic function at any desired point (say, a) if all the values on the contour surrounding this point are known. Thus if:

$$g(z) = \frac{f(z)}{z-a}$$

then the value $f(a)$ is called the *residue* of the function $g(z)$ for the pole a.

To make the term 'residue' clear let us make a practical example. Suppose $g(z)$ is a rational function of two polynomials:

$$g(z) = \frac{P(z)}{Q(z)} = \frac{z^m + b_{m-1}z^{m-1} + b_{m-2}\,z^{m-2} + \cdots + b_1 z + b_0}{z^n + a_{n-1}z^{n-1} + a_{n-2}\,z^{n-2} + \cdots + a_1 z + a_0} \tag{1.10.12}$$

where b_i and a_i are real constants and $n > m$. Eq. 1.10.12 represents a general form of a frequency response of an amplifier, where z can be replaced by the usual $s = \sigma + j\omega$ and b_0/a_0 is the DC amplification (at frequency $s = 0$). Instead of the sums, the polynomials $P(z)$ and $Q(z)$, and thus $g(z)$, may also be expressed in the product form:

$$g(z) = \frac{(z-z_1)(z-z_2)\cdots(z-z_m)}{(z-p_1)(z-p_2)\cdots(z-p_n)} \tag{1.10.13}$$

In this equation, $z_1, z_2, \ldots, z_m$ are the roots of the polynomial $P(z)$, so they are also the zeros of $g(z)$. Similarly, $p_1, p_2, \ldots, p_n$ are the roots of the polynomial $Q(z)$ and therefore also the poles of $g(z)$. Both statements are valid if $p_\mathrm{i} \neq z_\mathrm{i}$ for any i that can be applied to Eq. 1.10.13 (if $z - z_1$ were equal to, say, $z - p_3$, there would be no pole at p_3, because this pole would be canceled by the zero z_1). Now we factor out the term with one pole, i.e., $1/(z-p_2)$ and write:

$$g(z) = \frac{(z-z_1)(z-z_2)\cdots(z-z_m)}{(z-p_1)(z-p_3)\cdots(z-p_n)} \cdot \frac{1}{(z-p_2)} = f(z)\frac{1}{(z-p_2)} \tag{1.10.14}$$

where:

$$f(z) = \frac{(z-z_1)(z-z_2)\cdots(z-z_m)}{(z-p_1)\,(z-p_3)\cdots(z-p_n)} \tag{1.10.15}$$

If we focus only on $f(z)$ and let $z \to p_2$, we obtain the residue of the function $g(z)$ for the pole p_2 and this residue is equal to $f(p_2)$. Since we have taken the second pole we have appended the index '2' to the residue. By performing the suggested operation we obtain:

$$\mathrm{res}_2\, g(z) = \lim_{z \to p_2} (z-p_2)\, g(z) = f(p_2) \tag{1.10.16}$$

The word 'residue' is of Latin origin and means the *remainder*. However, since a remainder may also appear when we divide a polynomial by another, we shall keep using the expression 'residue' in order to avoid any confusion. Also in our further practical calculations we will simply write, say, res_2 instead of the complete expression $\mathrm{res}_2\, F(s)$.

The reader could obtain a rough idea of a residue by the following similarity: suppose we have a big circus tent, the canvas of which is supported by, say, four poles. If one of the poles is removed, the canvas sags. The height of the canvas above the ground where we have removed the pole is something similar to a residue for that pole. However, in this comparison two important facts are different: first, in the complex variable theory our 'canvas' as well as 'ground' are complex and, second, the poles are infinitely high (actually $+\infty$ on one side of the pole and $-\infty$ on the other; see Appendix 1, web only).

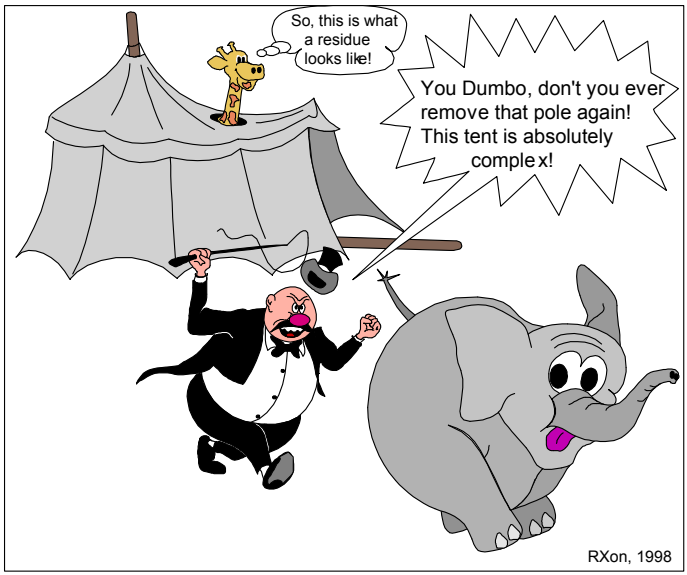


Fig. 1.10.2: Obtaining a residue is easy, just remove a pole! But be sure to put it back again, otherwise you will have to deal with the manager!

In the following examples we shall see that the calculation of residues is a relatively simple matter.

From now on we shall replace the variable z by our familiar complex variable $s = \sigma + j\omega$. Also, in order to distinguish more easily the functions of complex frequency from functions of time, we shall write the former with capitals, like $F(s)$ or $G(s)$ and the later with small letters, like $f(t)$ or $g(t)$.

To prove that the calculation of residues is indeed a simple task let us calculate two examples.

1.10.1 Example 1

Let us take a function: $$F(s) = \frac{(s+2)(s+3)}{(s+4)(s+5)(s+6)}$$

We need to calculate the three residues of $F(s)$ for the poles at $s=-4$, $s=-5$ and $s=-6$:

$$\text{res}_1 = \lim_{s\to -4}(s+4)\,\frac{(s+2)(s+3)}{(s+4)(s+5)(s+6)} = \frac{(-4+2)(-4+3)}{(-4+5)(-4+6)} = 1$$

and in a similar way:

$$\text{res}_2 = \lim_{s\to -5}(s+5)\,\frac{(s+2)(s+3)}{(s+4)(s+5)(s+6)} = -6$$

$$\text{res}_3 = \lim_{s\to -6}(s+6)\,\frac{(s+2)(s+3)}{(s+4)(s+5)(s+6)} = 6$$

An interesting fact here is that since all the poles are real, all the residues are real as well; in other words, a real pole causes the residue of that pole to be real.

1.10.2 Example 2

Our function is:

$$F(s) = \frac{(s+2)\,\mathrm{e}^{st}}{3\,s^2+9\,s+9}$$

Here we must consider that the variable of the function $F(s)$ is only s and not t. First we tackle the denominator to find both roots, which are the poles of our function:

$$3\,s^2+9\,s+9 = 3\,(s^2+3\,s+3) = 0$$

Thus:

$$s_{1,2} = \sigma_1 \pm j\,\omega_1 = -\frac{3}{2} \pm \sqrt{\left(\frac{3}{2}\right)^2 - 3} = -\frac{3}{2} \pm j\,\frac{\sqrt{3}}{2}$$

and by expressing the function $F(s)$ with both poles we have:

$$F(s) = \frac{(s+2)\,\mathrm{e}^{st}}{3\,(s-s_1)(s-s_2)}$$

We shall carry out a general calculation of the two residues and then introduce the numerical values for σ_1 and ω_1.

$$\text{res}_1 = \lim_{s\to s_1}(s-s_1)\,\frac{(s+2)\,\mathrm{e}^{st}}{3\,(s-s_1)(s-s_2)} = \frac{(s_1+2)\,\mathrm{e}^{st}}{3\,(s_1-s_2)}$$

$$= \frac{(\sigma+j\,\omega+2)\,\mathrm{e}^{\sigma_1 t}\,\mathrm{e}^{j\omega_1 t}}{3\,(\sigma_1+j\,\omega_1-\sigma_1+j\,\omega_1)} = \frac{(\sigma+j\,\omega+2)\,\mathrm{e}^{\sigma_1 t}\,\mathrm{e}^{j\omega_1 t}}{6\,j\,\omega_1}$$

We now set $\sigma_1 = -3/2$ and $\omega_1 = \sqrt{3}\,/2$ to obtain the numerical value of the residue:

$$\text{res}_1 = \frac{\left(-3/2 + j\,\sqrt{3}\,/2 + 2\right)\mathrm{e}^{-3\,t/2}\,\mathrm{e}^{j\sqrt{3}\,t/2}}{6\,j\,\sqrt{3}}$$

$$= \frac{1 + j\sqrt{3}}{12\, j\sqrt{3}}\, \mathrm{e}^{-3t/2}\, \mathrm{e}^{j\sqrt{3}\,t/2} = \frac{\sqrt{3} - j}{12\sqrt{3}}\, \mathrm{e}^{-3t/2}\, \mathrm{e}^{j\sqrt{3}\,t/2}$$

In a similar way we calculate the second residue:

$$\mathrm{res}_2 = \lim_{s \to s_2} (s - s_2)\, \frac{(s+2)\, \mathrm{e}^{st}}{3\,(s - s_1)(s - s_2)} = \frac{(s_2 + 2)\, \mathrm{e}^{st}}{3\,(s_2 - s_1)}$$

$$= \frac{\sqrt{3} + j}{12\sqrt{3}}\, \mathrm{e}^{-3t/2}\, \mathrm{e}^{-j\sqrt{3}\,t/2}$$

Since both poles are complex conjugate, both residues are complex conjugate as well. In rational functions, which will appear in the later sections, all the poles will be either real, or complex conjugate, or both. Therefore **the sum of all residues of these functions (that is, the time function) will always be real.**

1.11 Residues of Functions with Multiple Poles, the Laurent Series

When a function contains multiple poles it is not possible to calculate the residues in the way shown in the previous section. As an example let us take the function:

$$G(s) = \frac{F(s)}{(s-a)^n} \tag{1.11.1}$$

To calculate the residue we first expand $F(s)$ into a *Taylor* series [Ref. 1.4, 1.11]:

$$F(s) = (s-a)^n\, G(s) \tag{1.11.2}$$

$$= \frac{F(a)}{0!} + \frac{F(a)'(s-a)}{1!} + \frac{F(a)''(s-a)^2}{2!} + \cdots + \frac{F(a)^{(n-1)}(s-a)^{n-1}}{(n-1)!} + \cdots$$

Now we divide all the fractions in this equation by $(s-a)^n$ (considering that $0! = 1$ by definition):

$$G(s) = \frac{F(s)}{(s-a)^{\,n}} \tag{1.11.3}$$

$$= \frac{F(a)}{(s-a)^n} + \frac{F'(a)}{(s-a)^{n-1}} + \frac{F''(a)}{2!(s-a)^{n-2}} + \cdots + \frac{F^{(n-1)}(a)}{(n-1)!\,(s-a)} + \cdots$$

The values $F(a)$, $F'(a)$, $F''(a)/2!$, …, $F^{(n-1)}(a)/(n-1)!$, … are constants and we write them as A_{-n}, $A_{-(n-1)}$, $A_{-(n-2)}$, …, A_{-1}, A_0, A_1, A_2, … .

We may now express the function $G(s)$ as:

$$G(s) = \frac{A_{-n}}{(s-a)^n} + \frac{A_{-(n-1)}}{(s-a)^{n-1}} + \frac{A_{-(n-2)}}{(s-a)^{n-2}} + \cdots + \frac{A_{-1}}{(s-a)}$$

$$+ A_0 + A_1(s-a) + A_2(s-a)^2 + \cdots \tag{1.11.4}$$

The sum of all fractions from the above function we call the *principal part* and the rest is the *analytic part* (also known as the *regular part*).

Eq. 1.11.4 is named the *Laurent series*, after the French mathematician *Pierre-Alphonse Laurent*, 1813–1854, who in 1843 described "a series with negative powers".

A general expression for the Laurent series is:

$$\boxed{F(s) = \sum_{n=-m}^{+\infty} A_n\,(s-a)^n} \tag{1.11.5}$$

where m and n are integers.

Let us calculate the contour integral of the above function:

$$\oint_C F(s)\,ds = \oint_C \sum_{n=-m}^{+\infty} A_n\,(s-a)^n\,ds \tag{1.11.6}$$

We shall integrate each part of the series separately.

Again, $A_{-n}, \ldots, A_{-2}, A_{-1}, A_0, A_1, A_2, \ldots$ are constants and they may be put in front of the integral sign. In general we need to know the solution of the integral:

$$\oint_C (s-a)^n\,ds \tag{1.11.7}$$

where n is an integer either positive or negative. As in Eq. 1.9.2 we write again:

$$s-a = \varepsilon\,\mathrm{e}^{j\theta} \qquad \Rightarrow \qquad ds = j\,\varepsilon\,\mathrm{e}^{j\theta}d\theta \qquad \Rightarrow \qquad (s-a)^n = \varepsilon^n\,\mathrm{e}^{j\,n\,\theta}$$

where the radius ε is considered to be constant. Thus:

$$\oint_C (s-a)^n\,ds = \int_0^{2\pi} \left(\varepsilon^n\,\mathrm{e}^{j\,n\,\theta}\right) j\,\varepsilon\,\mathrm{e}^{j\theta}d\theta = j\,\varepsilon^{n+1}\int_0^{2\pi} \mathrm{e}^{j(n+1)\theta}\,d\theta \tag{1.11.8}$$

If $n \neq -1$, the result of this integration is:

$$\frac{j\,\varepsilon^{n+1}}{j\,(n+1)}\,\mathrm{e}^{j(n+1)\theta}\Big|_{\theta=0}^{\theta=2\pi} = \frac{\varepsilon^{n+1}}{(n+1)}\left[\mathrm{e}^{j(n+1)2\pi} - 1\right] = 0 \tag{1.11.9}$$

because $\mathrm{e}^{j\,k\,2\pi} = 1$, for any positive or negative integer k, including 0. For $n = -1$, we derive from Eq. 1.11.8:

$$\oint_C \frac{ds}{s-a} = \int_0^{2\pi} \frac{j\,\varepsilon\,\mathrm{e}^{j\theta}\,d\theta}{\varepsilon\,\mathrm{e}^{j\theta}} = 2\pi j \tag{1.11.10}$$

In order that the result corresponds to the Laurent series we must add the constant factor with $n = -1$ and this is A_{-1}. Eq. 1.11.8 to 1.11.10 prove that the contour integration for the complete Laurent series $G(s)$ yields only:

$$\boxed{\oint_C G(s)\,ds = A_{-1}2\pi j} \tag{1.11.11}$$

Thus from the whole series (Eq. 1.11.4) only the part with A_{-1} remained after the integration. If we return to Eq. 1.11.3 we conclude that:

$$A_{-1} = \frac{1}{2\pi j}\oint_C G(s)\,ds = \operatorname{res}\frac{F(s)}{(s-a)^n} = \frac{F^{(n-1)}(a)}{(n-1)!}$$

$$= \lim_{s\to a}\frac{1}{(n-1)!}\left[\frac{d^{(n-1)}}{ds^{(n-1)}}\,(s-a)^n\,G(s)\right] \tag{1.11.12}$$

is the residue of the function $G(s) = F(s)/(s-a)^n$ for the pole a. The following examples will show how we calculate the residues for multiple poles in practice.

1.11.1 Example 1

We take a function:

$$G(s) = \frac{F(s)}{(s-a)^3}$$

Our task is to calculate the general expression for the residue of the triple pole $(n = 3)$ at $s = a$. According to Eq. 1.11.12 it is:

$$\text{res} = \lim_{s \to a} \frac{1}{(3-1)!} \left[\frac{d^{(3-1)}}{ds^{(3-1)}} (s-a)^3 G(s) \right]$$

$$= \frac{1}{2} \left[\frac{d^2}{ds^2} (s-a)^3 G(s) \right]_{s \to a}$$

1.11.2 Example 2

Here we shall calculate with numerical values.

We intend to find the residues for the double pole at $s = -2$ and for the single pole at $s = -3$ of the function:

$$F(s) = \frac{5}{(s+2)^2 (s+3)}$$

Solution:

$$\text{res}_1 = \lim_{s \to -2} \frac{1}{(2-1)!} \left[\frac{d^{(2-1)}}{ds^{(2-1)}} (s+2)^2 \frac{5}{(s+2)^2 (s+3)} \right]$$

$$= \left[\frac{d}{ds} \frac{5}{(s+3)} \right]_{s \to -2} = \left[\frac{-5}{(s+3)^2} \right]_{s \to -2} = -5$$

$$\text{res}_2 = \lim_{s \to -3} \left[(s+3) \frac{5}{(s+2)^2 (s+3)} \right] = \left[\frac{5}{(s+2)^2} \right]_{s \to -3} = 5$$

It is important to remember the required order of the operations: first we multiply by the expression containing the multiple pole and then find the derivative. To do it the opposite way is wrong! Finally, we insert the numerical value for the pole.

1.12 Complex Integration Around Many Poles: the Cauchy–Goursat Theorem

So far we have calculated a contour integral around one pole (simple or multiple). Now we will integrate around more poles, either single or multiple.

Cheese is a regular part of French meals. So we may imagine that the great mathematician Cauchy observed a slice of Emmentaler cheese like that in Fig. 1.12.1 (the characteristics of this cheese is big holes) on his plate and reflected in the following way:

Suppose all that is cheese is an analytic (regular) domain R of a function $F(s)$. In the holes are the poles $s_1, \ldots, s_5$. We are not interested in the domain outside the cheese. How could we 'mathematically' encircle the cheese around the crust and around the rims of all the holes, so that the cheese is always on the left side of the contour?

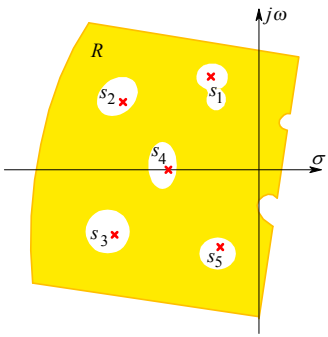


Fig. 1.12.1: The cheese represents a regular (analytic) domain R of a function which has one simple pole in each hole.

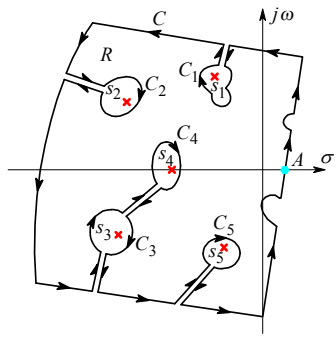


Fig. 1.12.2: Encircling the poles by contours $C, C_1, \ldots, C_5$, so that the regular domain of the function is always on the left side.

Impossible? No! If we take a knife and make a cut from the crust towards each hole without removing any cheese, we provide the necessary path for the suggested contour, as shown in Fig. 1.12.2.

Now we calculate a contour integral starting from the point A in the suggested (counterclockwise) direction until we come to the cut towards the first pole, s_1. We follow the cut towards contour C_1, follow it around the pole and then go along the cut again, back to the crust. We continue around the crust up to the cut of the next pole and so on, until we arrive back to point A and close the contour. Since we have not removed any cheese in making the cuts, the paths from the crust to the corresponding hole and back again cancel out in this integration path. As we have proved by Eq. 1.9.5:

$$\int_a^b F(s)\,ds + \int_b^a F(s)\,ds = 0$$

Therefore, only the contour C around the crust and the small contours $C_1, \ldots, C_5$ around the rims of the holes containing the poles are what we must consider in the integration around the contour in Fig. 1.12.2. The contour C was taken **counterclockwise**, whilst the contours $C_1, \ldots, C_5$ were taken **clockwise**.

We write down the complete contour integral:

$$\oint_C F(s)\,ds + \oint_{C_1} F(s)\,ds + \cdots + \oint_{C_5} F(s)\,ds = 0 \tag{1.12.1}$$

The result of integration is zero because along this circuitous contour of integration we have had the regular domain always on the left side. By changing the sense of encircling of the contours $C_1, \ldots, C_5$ we may write Eq. 1.12.1 also in the form:

$$\oint_C F(s)\,ds = \oint_{C_1} F(s)\,ds + \cdots + \oint_{C_5} F(s)\,ds \tag{1.12.2}$$

When we changed the sense of encircling, we changed the sign of the integrals; this allows us to put them on the right hand side with a positive sign. Now all the integrals have positive (counterclockwise) encircling. Therefore the integral encircling all the poles is equal to the sum of the integrals encircling each particular pole.

By observing this equation we realize that the right hand side is the sum of residues for all the five poles, multiplied by $2\pi j$. Thus for the general n-pole case the Eq. 1.12.2 may also be written as:

$$\boxed{\oint_C F(s)\,ds = 2\pi j\,[\,\mathrm{res}_1 + \cdots + \mathrm{res}_n] = 2\pi j \sum_{i=1}^{n} \mathrm{res}_i} \tag{1.12.3}$$

Eq. 1.12.2 and 1.12.3 are called the *Cauchy–Goursat theorem*; they are of essential importance for the inverse Laplace transform.

1.13 Equality of the Integrals $\oint_C F(s)\,e^{st}\,ds$ and $\int_{c-j\infty}^{c+j\infty} F(s)\,e^{st}\,ds$

The reader is invited to examine Fig. 1.13.1, where the function $|F(s)| = |1/s|$ was plotted. The function has one simple pole at the origin of the complex plane. The resulting surface has been cut between $-j$ and 1 to expose an arbitrarily chosen integration path L between $s_1 = x_1 + j\,y_1 = 0 - j0.5$ and $s_2 = x_2 + j\,y_2 = 0.5 + j0$ (see the integration path in the plot of the s domain in Fig. 1.13.2).

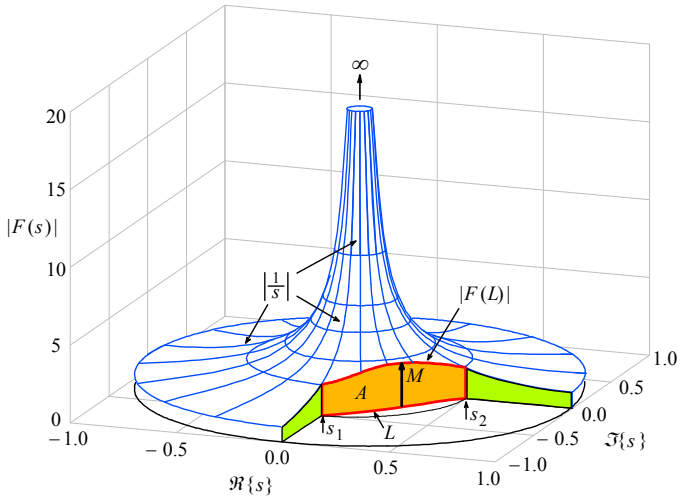


Fig. 1.13.1: The complex function magnitude, $|F(s)| = |1/s|$. The resulting surface has been cut between $-j$ and $+1$ to expose an arbitrarily chosen integration path L, starting at $s_1 = 0 - j0.5$ and ending at $s_2 = 0.5 + j0$. On the path of integration the function $|F(s)|$ has a maximum value M.

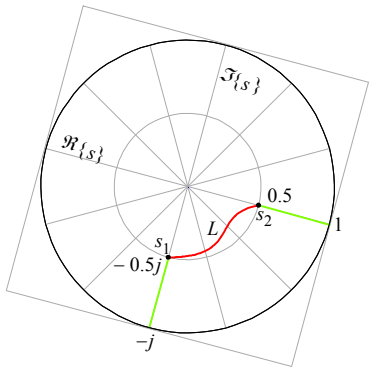


Fig. 1.13.2: The complex domain of Fig. 1.13.1 shows the arbitrarily chosen integration path L, which starts at $s_1 = 0 - j0.5$ and ends at $s_2 = 0.5 + j0$.

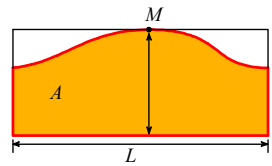


Fig. 1.13.3: The area A of Fig. 1.13.1 has been laid flat in order to show that it must be smaller than or at best equal to the area of the rectangle $M{\times}L$.

Let us take a closer look at the area A between s_1, s_2, $|F(s_1)|$ and $|F(s_2)|$, shown in Fig 1.13.3. The area A corresponds to the integral of $F(s)$ from s_1 to s_2 and it can be shown that it is always smaller than, or at best equal to, the rectangle $M{\times}L$:

$$\underbrace{\left|\int_{s_1}^{s_1} F(s)\,ds\right| = \left|\int_{s_1}^{s_2} \frac{ds}{s}\right| \le \int_{s_1}^{s_2} \frac{|ds|}{|s|}}_{\text{along the path } L} \le \int_{s_1}^{s_2} M\,|ds| = ML \qquad (1.13.1)$$

Here M is the greatest value of $|F(s)|$ for this particular path of integration L, as shown in Fig. 1.13.3, in which the resulting 3D area between s_1, s_2, $|F(s_1)|$ and $|F(s_2)|$ was stretched flat. So:

$$\left|\int_{s_1}^{s_2} F(s)\,ds\right| \le \int_{s_1}^{s_2} \left|F(s)\,ds\right| \le ML \qquad (1.13.2)$$

Eq. 1.13.2 is an essential tool in the proof of the inverse $\mathcal{L}$ transform via the integral around the closed contour.

Let us now move to network analysis, where we have to deal with rational functions of the complex variable $s = \sigma + j\,\omega$. These functions have a general form:

$$F(s) = \frac{s^m + b_{m-1}\,s^{m-1} + \cdots + b_1 s + b_0}{s^n + a_{n-1}\,s^{n-1} + \cdots + a_1 s + a_0} \qquad (1.13.3)$$

where $m < n$ and both are positive and real. Since we can also express $s = R\,e^{j\theta}$ (as can be derived from Fig. 1.13.4), we may write Eq. 1.13.3 also in the form:

$$F(s) = \frac{R^m\,e^{j\,m\,\theta} + b_{m-1}\,R^{m-1}\,e^{j(m-1)\theta} + \cdots + b_1 R\,e^{j\,\theta} + b_0}{R^n\,e^{j\,n\,\theta} + a_{n-1}\,R^{n-1}\,e^{j(n-1)\theta} + \cdots + a_1 R\,e^{j\,\theta} + a_0} \qquad (1.13.4)$$

According to Eq. 1.13.2 and 1.13.4 we have:

$$\left|F(s)\right| = \left|\frac{R^m\,e^{j\,m\,\theta} + \cdots + b_0}{R^n\,e^{j\,n\,\theta} + \cdots + a_0}\right| \le \frac{K}{R^{n-m}} = M \qquad (1.13.5)$$

where K is a real constant and M is the maximum value of $\left|F(s)\right|$ within the integration interval, according to Fig. 1.13.1 and 1.13.3 (in [Ref. 1.10, p. 212] the interested reader can find the complete derivation of the constant K).

$$s_1 = \sigma_1 + j\,\omega_1 = R\,e^{j\theta}$$
$$\sigma_1 = R\cos\theta \qquad \omega_1 = R\sin\theta$$
$$R = \sqrt{(\sigma_1 + j\,\omega_1)(\sigma_1 - j\,\omega_1)}$$
$$= \sqrt{\sigma_1^2 + \omega_1^2}$$
$$\theta = \arctan\frac{\omega_1}{\sigma_1}$$

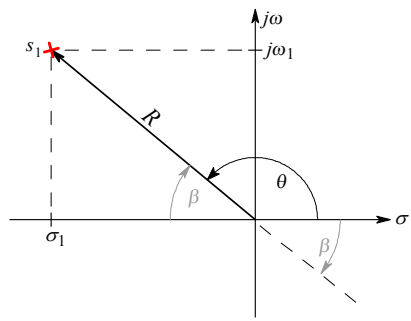


Fig. 1.13.4: Cartesian and polar representations of a complex number (note: $\tan\theta$ is equal for the counterclockwise defined θ from the positive real axis and for the clockwise defined $\beta = \theta - \pi$).

Let us draw the poles of Eq. 1.13.3 in the complex plane to calculate the integral around an inverted 'D' shaped contour, as shown in Fig. 1.13.5 (for convenience only three

poles have been drawn there). Since Eq. 1.13.3 is assumed to describe a real passive system, all poles must lie either on the left side of the complex plane or at the origin. As we know, the integral around the closed contour embracing all the poles is equal to the sum of residues of the function $F(s)$:

$$\oint_L F(s)\,ds = \int_{\sigma_a - j\omega_1}^{\sigma_a + j\omega_1} F(s)\,ds + \int_\Gamma F(s)\,ds = 2\pi j \sum_{i=1}^{n} \text{res}_i \tag{1.13.6}$$

The contour has two parts: the straight line from $\sigma_a - j\,\omega_1$ to $\sigma_a + j\,\omega_1$, where σ_a is a constant (which we will define more exactly later) and the arc $\Gamma = R\,\gamma$, where γ is the arc angle and R is its radius. According to Eq. 1.13.2, the line integral along the path L is:

$$\left| \oint_L F(s)\,ds \right| \le ML \tag{1.13.7}$$

where M is the maximum value of the integral (magnitude!) on the path L. In our case:

$$M = \frac{K}{R^{n-m}} \quad \text{and} \quad L = \gamma\,R = \Gamma \tag{1.13.8}$$

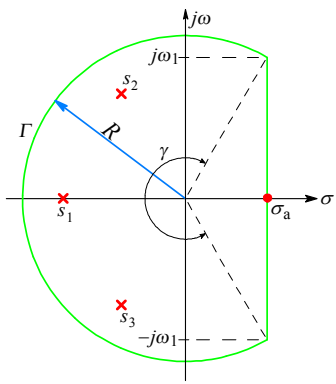


Fig. 1.13.5: The integral along the inverted 'D' shaped contour encircling the poles is equal to the sum of the residues of each pole. This contour is used to prove the inverse Laplace transform, where the integral along the arc vanishes if $R \to \infty$, provided that the number of poles exceeds the number of zeros by at least 2 (in this example no zeros are shown).

For a very large R we may write:

$$\left| \int_\Gamma F(s)\,ds \right| \le \frac{K}{R^{n-m}}\,\gamma\,R = \frac{K\gamma}{R^{n-m-1}} \tag{1.13.9}$$

If $R \to \infty$:

$$\lim_{R\to\infty} \frac{K\gamma}{R^{n-m-1}} = 0 \quad \textbf{only if} \quad n - m \ge 2 \tag{1.13.10}$$

and this procedure is called *Jordan's lemma*.

If the condition of Eq. 1.13.10 holds, only the straight part of the contour counts because if $R \to \infty$ then also $\omega_1 \to \infty$, thus changing the limits of the integral along the straight path accordingly. If we make these changes to Eq. 1.13.6, it shrinks to:

$$\int_{\sigma_a - j\infty}^{\sigma_a + j\infty} F(s)\, ds = \oint_L F(s)\, ds = 2\pi j \sum_{i=1}^{n} \text{res}_i \tag{1.13.11}$$

The function $F(s)$ may also contain the factor e^{st}, where $\Re(s) \geq \sigma_a$ and $t \geq 0$. In this case the constant σ_a, which is called *the abscissa of absolute convergence* [Ref. 1.3, 1.5, 1.8], must be small enough to ensure the convergence of the integral. The factor e^{st} is always present in the inverse $\mathcal{L}$ transform. Let us write this factor down and let us divide Eq. 1.13.11 by the factor $2\pi j$. In this way the integral obtains the form:

$$f(t) = \mathcal{L}^{-1}\{F(s)\} = \frac{1}{2\pi j} \int_{\sigma_a - j\infty}^{\sigma_a + j\infty} F(s)\, e^{st}\, ds = \sum \text{res}\left\{F(s)\, e^{st}\right\} \tag{1.13.12}$$

and this is the formula for the inverse $\mathcal{L}$ transform [Ref. 1.3, 1.5, 1.8]. The above integral is convergent for $t \geq 0$, which is the usual constraint in passive network analysis. This constraint will also apply to all derivations which follow.

In the condition written in Eq. 1.13.10 we see that the order of the denominator's polynomial must exceed the order of the numerator by at least **two**, otherwise we could not prove the inverse $\mathcal{L}$ transform by the method derived above. This means that the number of poles must exceed the number of zeros by at least two. However, in network theory we often deal with the input functions called *positive real functions* [Ref. 1.16]. The degree of the denominator in these functions may exceed the degree in the numerator by **one** only. To prove the inverse $\mathcal{L}$ transform for such a case, we must reach for another method. The proof is possible by using a rectangular contour [Ref. 1.5, 1.13, 1.17]:

When the degree of the denominator exceeds the degree of the numerator by one only, Eq. 1.13.5 is reduced to:

$$|F(s)| \leq \frac{K}{R} = M \tag{1.13.13}$$

so to prove the inverse Laplace transform we use:

$$F(s) = \frac{1}{s - s_p} = \frac{1}{s - \sigma_p} \tag{1.13.14}$$

This is a single-pole function, with the pole on the negative real axis (for our calculations it is not essential that the pole lies on the real axis, but in the theory of real passive networks, a single-pole **always** lies either on the negative σ axis or at the origin of the complex plane).

The pole and the rectangular contour with the sides Γ_1, Γ_2, Γ_3 and Γ_4 are shown in Fig. 1.13.6. We will integrate around this rectangular contour. At the same time we let both $\sigma_1 \to \infty$ and $\omega_1 \to \infty$. Next we will prove, considering these limits, that the line integrals along the sides Γ_2, Γ_3 and Γ_4 are all equal to zero.

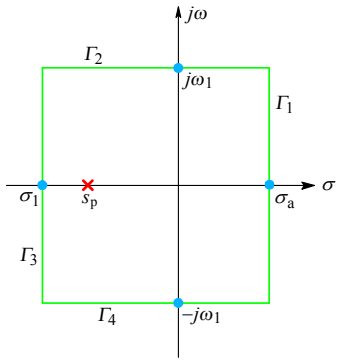


Fig. 1.13.6: By using a rectangular contour as shown it is possible to prove the inverse Laplace transform by means of the contour integral, even if the number of poles exceeds the number of zeros by only one. In this integral, encircling the single simple pole, we let $\sigma_1 \to \infty$ and $\omega_1 \to \infty$, so that the integrals along Γ_2, Γ_3 and Γ_4 vanish.

The proof must show that:

$$\oint_{\Gamma} F(s)\,\mathrm{e}^{st}\,ds = \lim_{\omega_1 \to \infty} \int_{\sigma_a - j\omega_1}^{\sigma_a + j\omega_1} F(s)\,\mathrm{e}^{st}\,ds = 2\,\pi\,j \sum \operatorname{res}\left\{F(s)\,\mathrm{e}^{st}\right\} \tag{1.13.15}$$

Here we will include the factor e^{st} (which always appears in the inverse $\mathcal{L}$ transform) at the very beginning, because it will help us in making the integral along Γ_3 convergent. Let us start with the integral along the side Γ_2, where ω_1 is constant:

$$\left|\int_{\Gamma_2} F(s)\,\mathrm{e}^{st}\,ds\right| = \left|\int_{\sigma_a}^{-\sigma_1} F(\sigma + j\,\omega_1)\,\mathrm{e}^{(\sigma + j\omega_1)t}\,d\sigma\right|$$

$$\leq \int_{-\sigma_1}^{\sigma_a} \frac{K}{\sigma_1}\,\mathrm{e}^{\sigma t}\,d\sigma$$

$$= \frac{K}{\sigma_1} \cdot \frac{1}{t}\left(\mathrm{e}^{\sigma_a t} - \mathrm{e}^{-\sigma_1 t}\right) \to 0\Big|_{\sigma_1 \to \infty} \tag{1.13.16}$$

Since we are calculating the absolute value, we can exchange the limits of the last integral. The integral along Γ_4 is almost equal:

$$\left|\int_{\Gamma_4} F(s)\,\mathrm{e}^{st}\,ds\right| = \left|\int_{-\sigma_1}^{\sigma_a} F(\sigma + j\,\omega_1)\,\mathrm{e}^{(\sigma + j\omega_1)t}\,d\sigma\right|$$

$$\leq \int_{-\sigma_1}^{\sigma_a} \frac{K}{\sigma_1}\,\mathrm{e}^{\sigma t}\,d\sigma$$

$$= \frac{K}{\sigma_1} \cdot \frac{1}{t}\left(\mathrm{e}^{\sigma_a t} - \mathrm{e}^{-\sigma_1 t}\right) \to 0\Big|_{\sigma_1 \to \infty} \tag{1.13.17}$$

In the integral along Γ_3, σ_1 is constant:

$$\left| \int_{\Gamma_3} F(s)\, \mathrm{e}^{st}\, ds \right| = \left| \int_{-j\omega_1}^{j\omega_1} F(-\sigma_1 + j\omega)\, \mathrm{e}^{(-\sigma_1 + j\omega)t}\, d\omega \right|$$

$$\leq \int_{-\omega_1}^{\omega_1} \frac{K}{\sigma_1}\, \mathrm{e}^{-\sigma_1 t}\, d\omega$$

$$= \frac{K}{\sigma_1}\, \mathrm{e}^{-\sigma_1 t}(\omega_1 + \omega_1) \;\rightarrow 0 \Big|_{\substack{\sigma_1 \rightarrow \infty \\ \omega_1 \rightarrow \infty}} \qquad (1.13.18)$$

Since the integrals along Γ_2, Γ_3 and Γ_4 are all equal to zero if $\sigma_1 \rightarrow \infty$ and $\omega_1 \rightarrow \infty$, only the integral along Γ_1 remains, which, in the limit, is equal to the integral along the complete rectangular contour and, in turn, to the sum of the residues of the poles of $F(s)$:

$$\lim_{\omega_1 \rightarrow \infty} \int_{\sigma_a - j\omega_1}^{\sigma_a + j\omega_1} F(s)\, \mathrm{e}^{st}\, ds = \int_{\sigma_a - j\infty}^{\sigma_a + j\infty} F(s)\, \mathrm{e}^{st}\, ds = \oint_{\Gamma} F(s)\, \mathrm{e}^{st}\, ds$$

$$= 2\pi j \sum \mathrm{res} \left\{ F(s)\, \mathrm{e}^{st} \right\} \qquad (1.13.19)$$

If this equation is divided by $2\pi j$, we again obtain the Eq. 1.13.12 which is the inverse Laplace transform of the function $F(s)$.

Although there was only a single pole in our $F(s)$ in Eq. 1.13.14 the result obtained is valid in the general case, when $F(s)$ has n poles and $n-1$ zeros.

Thus we have proved the $\mathcal{L}^{-1}$ transform by means of a contour integral for positive real functions. As in Eq. 1.13.12, here, too, the *abscissa of absolute convergence* σ_a must be chosen so that $\Re\{s\} \geq \sigma_a$ and also $t \geq 0$ in order to ensure the convergence of the integral. However, we may also integrate along a straight path, where $\sigma < \sigma_a$, provided that all the poles remain on the left side of the path.

From all the complicated equations above the reader must remember only one important fact, which we will use very frequently in the following sections: **By means of the $\mathcal{L}^{-1}$ transform of $F(s)$, the complex transfer function of a linear network, we obtain the real time function, $f(t)$, as the sum of the residues of all the poles of the complex frequency function $F(s)\, \mathrm{e}^{st}$.**

Let us put this in the symbolic form:

$$\boxed{f(t) = \mathcal{L}^{-1}\{F(s)\} = \sum \mathrm{res} \left\{ F(s)\, \mathrm{e}^{st} \right\}} \qquad (1.13.20)$$

1.14 Application of the Inverse Laplace Transform

In the following parts of the book we will very frequently need the inverse Laplace transform of two-pole and three-pole systems, in which the third pole at the origin, $1/s$, is the $\mathcal{L}$ transform of the unit step function. Therefore it would be useful to perform our first example of the $\mathcal{L}^{-1}$ transform calculation on such a network function.

Fig. 1.14.1 shows a typical two-pole network. Our task is to calculate the voltage on the resistor R as a function of time for $t > 0$. First we will apply an input current i_i in the form of an impulse $\delta(t)$, and next the input current will have a unit step form. Both results will be used in many cases in the following parts of the book. In the same way as for $F(s)$ and $f(t)$ we will label the voltages and currents with capitals (V, I) when they are the functions of frequency and with small letters (v, i) when they are functions of time.

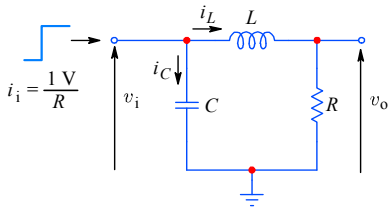


Fig. 1.14.1: A simple RLC circuit, driven by a current step, often found in electrical and electronics networks.

The input current is composed of two components, the current through the capacitor I_C, and through the inductor I_L (and the resistor R) and V_i is the input voltage:

$$I_\mathrm{i} = I_C + I_L = V_\mathrm{i}\, s\, C + \frac{V_\mathrm{i}}{s\, L + R} = V_\mathrm{i} \frac{s^2 L\, C + s\, R\, C + 1}{s\, L + R} \tag{1.14.1}$$

Correspondingly:

$$\boxed{\frac{V_\mathrm{i}}{I_\mathrm{i}} = \frac{s\, L + R}{s^2 L\, C + s\, R\, C + 1}} \tag{1.14.2}$$

This is a typical *input function* [Ref. 1.16], in this case it has the form of an (input) impedance, Z_i. The characteristics of an input function is that the number of poles exceeds the number of zeros by **one** only. The output voltage V_o is:

$$V_\mathrm{o} = V_\mathrm{i}\, \frac{R}{s\, L + R} \tag{1.14.3}$$

and so:

$$\frac{V_\mathrm{o}}{I_\mathrm{i}} = \frac{R}{s\, L + R} \cdot \frac{s\, L + R}{s^2 L\, C + s\, R\, C + 1} = \frac{R}{s^2 L\, C + s\, R\, C + 1} \tag{1.14.4}$$

The result is the *transfer function* of the network (from input to output, but is expressed as the output to input ratio). Since the dimension of Eq. 1.14.4 is (complex) Ohms it is also named the *transimpedance*. In general we will assume that the input current is $1\,\mathrm{V}/R$, in order to obtain a *normalized* transfer function:

$$\boxed{V_\mathrm{o} = \frac{1\,[\mathrm{V}]}{s^2 L\, C + s\, R\, C + 1} = G(s)} \tag{1.14.5}$$

In our later applications of the circuit in Fig. 1.14.1 the denominator of Eq. 1.14.5 must have **complex** roots (although, in general, the roots can also be real). Now let us calculate both roots of the denominator from its canonical form:

$$s^2 + s\,\frac{R}{L} + \frac{1}{L\,C} = 0 \tag{1.14.6}$$

with the roots:

$$s_{1,2} = \sigma_1 \pm j\,\omega_1 = -\frac{R}{2\,L} \pm \sqrt{\frac{R^2}{4\,L^2} - \frac{1}{L\,C}} \tag{1.14.7}$$

In special cases, some of which we shall analyze in the later parts of the book, the roots may also be double and real.

Expressing the transfer function, Eq. 1.14.5, by its roots, we obtain:

$$G(s) = \frac{1}{L\,C} \cdot \frac{1}{(s - s_1)(s - s_2)} \tag{1.14.8}$$

From the $\mathcal{L}^{-1}$ transform of this function we obtain the system's impulse response in the time domain, $g(t) = g\{\delta(t)\}$. The factor $1/LC$ is the system resonance, ω_1^2, which in a different network may take a different form (in the general normalized second-order case it is equal to the product of the two poles, $s_1 s_2$). Thus, we put $K = 1/LC$:

$$g(t) = \mathcal{L}^{-1}\{G(s)\} = \mathcal{L}^{-1}\left\{\frac{K}{(s - s_1)(s - s_2)}\right\}$$

$$= \frac{K}{2\pi j}\oint_C \frac{\mathrm{e}^{st}}{(s - s_1)(s - s_2)}\,ds = K\sum \mathrm{res}\left\{\frac{\mathrm{e}^{st}}{(s - s_1)(s - s_2)}\right\} \tag{1.14.9}$$

The contour of integration in Eq. 1.14.9 must encircle both poles.

Since the network in Fig. 1.14.1 is passive, both poles s_1 and s_2 lie on the left side of the complex plane. As an example, in Fig. 1.14.2 we have drawn the magnitude of $G(s)$ for the special case of a 2nd-order *Butterworth* network, for which the absolute values of the pole components are equal, $|\sigma| = |\omega| = 1/\sqrt{2} = 0.707$.

In this figure the Laplace transformed system transfer function, $\big|G(s)\big|$, is represented by the surface over the complex plane, peaking to infinity over the poles.

If we intersect the magnitude function by a vertical plane along the $j\omega$-axis, the surface edge (curve) at the intersection represents the frequency response $|G(j\,\omega)|$ of the function $|G(s)|$. The response is shown in a linear scale, and for the negative values of ω as well. In later sections we shall draw the frequency response graphs with a logarithmic scale for the frequency (positive only). Likewise, the magnitude will be logarithmic as well.

Eq. 1.14.9 has two residues:

$$\mathrm{res}_1 = \lim_{s \to s_1}(s - s_1)\,\frac{\mathrm{e}^{st}}{(s - s_1)(s - s_2)} = \frac{\mathrm{e}^{s_1 t}}{s_1 - s_2}$$

$$\mathrm{res}_2 = \lim_{s \to s_2}(s - s_2)\,\frac{\mathrm{e}^{st}}{(s - s_1)(s - s_2)} = \frac{\mathrm{e}^{s_2 t}}{s_2 - s_1} \tag{1.14.10}$$

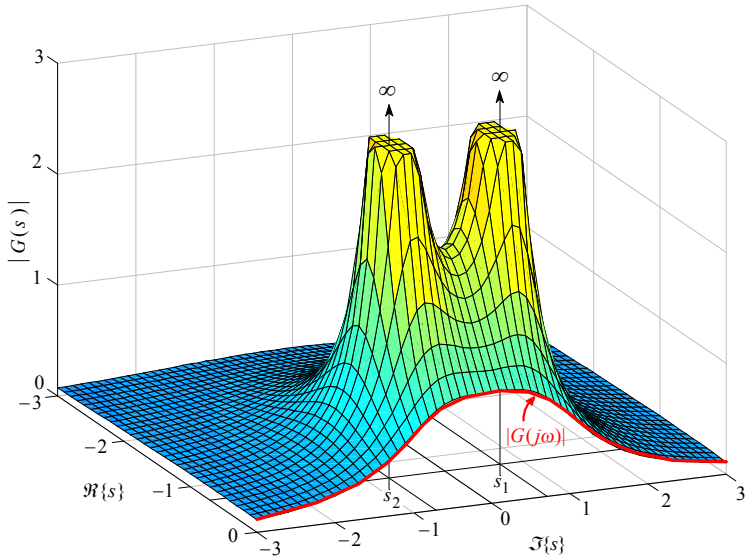


Fig. 1.14.2: The magnitude of the system transfer function, Eq. 1.14.8, for $s_{1,2} = (-1 \pm j)/\sqrt{2}$ and $K = 1$. For $\Re(s) = 0$, the surface $|G(s)|$ is reduced to the frequency response's magnitude curve, $|G(j\omega)|$. The height at $s_{1,2}$ is ∞, but was limited to 3 in order to see $|G(j\omega)|$ in detail.

The corresponding time function is the sum of both residues:

$$g(t) = \text{res}_1 + \text{res}_2 = \frac{K}{s_1 - s_2}\left(\mathrm{e}^{s_1 t} - \mathrm{e}^{s_2 t}\right) \tag{1.14.11}$$

Now we insert $\sigma_1 + j\omega_1$ for s_1 and $\sigma_1 - j\omega_1$ for s_2:

$$g(t) = \frac{K}{\sigma_1 + j\omega_1 - \sigma_1 + j\omega_1}\left[\mathrm{e}^{(\sigma_1 + j\omega_1)t} - \mathrm{e}^{(\sigma_1 - j\omega_1)t}\right] \tag{1.14.12}$$

We factor out $\mathrm{e}^{\sigma_1 t}$ and rearrange the denominator to obtain:

$$g(t) = \frac{K}{2\,j\,\omega_1}\,\mathrm{e}^{\sigma_1 t}\left[\mathrm{e}^{j\omega_1 t} - \mathrm{e}^{-j\omega_1 t}\right] = \frac{K}{\omega_1}\,\mathrm{e}^{\sigma_1 t}\,\frac{\mathrm{e}^{j\omega_1 t} - \mathrm{e}^{-j\omega_1 t}}{2\,j} \tag{1.14.13}$$

Since:

$$\frac{\mathrm{e}^{j\omega_1 t} - \mathrm{e}^{-j\omega_1 t}}{2\,j} = \sin\omega_1 t \tag{1.14.14}$$

then:

$$g(t) = \frac{K}{\omega_1}\,\mathrm{e}^{\sigma_1 t}\sin\omega_1 t \tag{1.14.15}$$

But K can also be expressed with σ_1 and ω_1:

$$K = \frac{1}{LC} = s_1 s_2 = \sigma_1^2 + \omega_1^2 \tag{1.14.16}$$

so:

$$g(t) = \frac{\sigma_1^2 + \omega_1^2}{\omega_1}\,\mathrm{e}^{\sigma_1 t}\sin\omega_1 t = \frac{\sqrt{\sigma_1^2 + \omega_1^2}}{\sin\theta}\,\mathrm{e}^{\sigma_1 t}\sin\omega_1 t \tag{1.14.17}$$

where θ is the angle between a pole and the positive σ axis, as in Fig. 1.13.4.

In our example, $\sigma_1^2 + \omega_1^2 = 1$ (Butterworth case), so Eq. 1.14.17 can be simplified:

$$g(t) = \frac{1}{\omega_1}\, e^{\sigma_1 t} \sin\omega_1 t = \frac{1}{\sin\theta}\, e^{\sigma_1 t} \sin\omega_1 t \qquad (1.14.18)$$

Note that Eq. 1.14.13 and Eq. 1.14.17 are valid for any complex pole pair, not just for Butterworth poles. This completes the calculation of the impulse response.

The next case, in which we are interested more often, is the step response. In Example 1, Sec. 1.5, we have calculated that the unit step function in the time domain corresponds to $1/s$ in the frequency domain. To obtain the step response in the time domain, we need only to multiply the frequency response by $1/s$ and calculate the inverse $\mathcal{L}$ transform of the product. So by multiplying $G(s)$ by $1/s$ we obtain a new function:

$$F(s) = \frac{1}{s}\, G(s) = \frac{K}{s\,(s - s_1)(s - s_2)} \qquad (1.14.19)$$

To calculate the step response in the time domain we use the $\mathcal{L}^{-1}$ transform:

$$f(t) = \mathcal{L}^{-1}\{F(s)\} = \mathcal{L}^{-1}\, \frac{K}{s\,(s - s_1)\,(s - s_2)} =$$

$$= \frac{K}{2\,\pi\, j} \oint_C \frac{e^{st}}{s\,(s - s_1)(s - s_2)}\, ds = K \sum \mathrm{res} \left\{ \frac{e^{st}}{s\,(s - s_1)(s - s_2)} \right\} \qquad (1.14.20)$$

The difference between Eq. 1.14.9 and Eq. 1.14.20 is that here we have an additional pole $s_0 = 0$, because of the factor $1/s$. Thus here we have three residues:

$$\begin{aligned}
\mathrm{res}_0 &= \lim_{s \to 0} s\, \frac{e^{st}}{s\,(s - s_1)(s - s_2)} = \frac{1}{s_1\, s_2} \\
\mathrm{res}_1 &= \lim_{s \to s_1} (s - s_1)\, \frac{e^{st}}{s\,(s - s_1)(s - s_2)} = \frac{e^{s_1 t}}{s_1\,(s_1 - s_2)} \\
\mathrm{res}_2 &= \lim_{s \to s_2} (s - s_2)\, \frac{e^{st}}{s\,(s - s_1)(s - s_2)} = \frac{e^{s_2 t}}{s_2\,(s_2 - s_1)}
\end{aligned} \qquad (1.14.21)$$

In the double-pole case (coincident pole pair, $s_1 = s_2$) the calculation is different (remember Eq. 1.11.12) and it will be shown in several examples in Part 2. The time domain function is the sum of all three residues ($K/s_1\, s_2$ is factored out):

$$f(t) = \frac{K}{s_1\, s_2} \left(1 + \frac{s_2}{s_1 - s_2}\, e^{s_1 t} + \frac{s_1}{s_2 - s_1}\, e^{s_2 t} \right) \qquad (1.14.22)$$

By expressing $s_1 = \sigma_1 + j\,\omega_1$ and $s_2 = \sigma_1 - j\,\omega_1$ in each of the residues we obtain:

$$\begin{aligned}
&\frac{K}{s_1\, s_2} = \frac{K}{(\sigma_1 + j\,\omega_1)(\sigma_1 - j\,\omega_1)} = \frac{K}{\sigma_1^2 + \omega_1^2} = 1 && \text{(see Eq. 1.14.16)} \\
&\frac{s_2\, e^{s_1 t}}{s_1 - s_2} = \frac{(\sigma_1 - j\,\omega_1)\, e^{(\sigma_1 + j\,\omega_1) t}}{\sigma_1 + j\,\omega_1 - \sigma_1 + j\,\omega_1} = \frac{\sigma_1 - j\,\omega_1}{2\,j\,\omega_1}\, e^{(\sigma_1 + j\,\omega_1) t} \\
&\frac{s_1\, e^{s_2 t}}{s_2 - s_1} = \frac{(\sigma_1 + j\,\omega_1)\, e^{(\sigma_1 - j\,\omega_1) t}}{\sigma_1 - j\,\omega_1 - \sigma_1 - j\,\omega_1} = \frac{\sigma_1 + j\,\omega_1}{-2\,j\,\omega_1}\, e^{(\sigma_1 - j\,\omega_1) t}
\end{aligned} \qquad (1.14.23)$$

We put these results into Eq. 1.14.22 and obtain:

$$f(t) = 1 + \frac{\sigma_1 - j\omega_1}{2\,j\,\omega_1}\,\mathrm{e}^{(\sigma_1 + j\omega_1)t} + \frac{\sigma_1 + j\omega_1}{-2\,j\,\omega_1}\,\mathrm{e}^{(\sigma_1 - j\omega_1)t} \tag{1.14.24}$$

By factoring out $\mathrm{e}^{\sigma_1 t}$, and with a slight rearranging, we arrive at:

$$f(t) = 1 + \mathrm{e}^{\sigma_1 t}\left[\frac{\sigma_1}{\omega_1}\left(\frac{\mathrm{e}^{j\omega_1 t} - \mathrm{e}^{-j\omega_1 t}}{2\,j}\right) - \left(\frac{\mathrm{e}^{j\omega_1 t} + \mathrm{e}^{-j\omega_1 t}}{2}\right)\right] \tag{1.14.25}$$

Since $(\mathrm{e}^{j\omega_1 t} - \mathrm{e}^{-j\omega_1 t})/2\,j = \sin\omega_1 t$ and $(\mathrm{e}^{j\omega_1 t} + \mathrm{e}^{-j\omega_1 t})/2 = \cos\omega_1 t$ we can simplify Eq. 1.14.25 into the form:

$$\boxed{f(t) = 1 + \mathrm{e}^{\sigma_1 t}\left(\frac{\sigma_1}{\omega_1}\sin\omega_1 t - \cos\omega_1 t\right)} \tag{1.14.26}$$

We could now numerically calculate the response, but we want to show two things:

1) how the formula relates to the physical circuit behavior;

2) explain an error, all too often ignored (even by experienced engineers!).

We can further simplify the sine–cosine term by using the vector sum of the two phasors (this relation can be found in any mathematics handbook):

$$A\sin\alpha + B\cos\alpha = \sqrt{A^2 + B^2}\,\sin(\alpha + \theta) \qquad \text{where} \qquad \theta = \arctan(B/A)$$

By putting $A = \sigma_1/\omega_1$ and $B = -1$ we arrive at:

$$f(t) = 1 + \sqrt{1 + \left(\frac{\sigma_1}{\omega_1}\right)^2}\,\mathrm{e}^{\sigma_1 t}\sin(\omega_1 t + \theta) \tag{1.14.27}$$

where:

$$\theta = \arctan\left(\frac{-\omega_1}{\sigma_1}\right)$$

For the Butterworth case, the square root is equal to $\sqrt{2}$, but in the general case it is:

$$\sqrt{1 + \left(\frac{\sigma_1}{\omega_1}\right)^2} = \frac{\sqrt{\sigma_1^2 + \omega_1^2}}{\omega_1} = \left|\frac{1}{\sin\theta}\right| \tag{1.14.28}$$

Note that for any value of σ_1 and ω_1 their square can never be negative, which is reflected in the absolute value notation at the end; on the other hand, it is important to preserve the correct sign of the phase shifting term in $\sin(\omega_1 t + \theta)$. By putting Eq. 1.14.28 back into Eq. 1.14.27 we obtain a relatively simple expression:

$$\boxed{f(t) = 1 + \left|\frac{1}{\sin\theta}\right|\mathrm{e}^{\sigma_1 t}\sin(\omega_1 t + \theta)} \tag{1.14.29}$$

If we now insert the numerical values for σ_1, ω_1 and θ and plot the function for t in the interval from 0 to 10, the resulting graph will be obviously **wrong!** What happened?

Let us check our result by applying the rule of initial and final value from Sec. 1.6. We will use Eq. 1.14.29 and Eq. 1.14.8, considering that $K = s_1 s_2$ (Eq. 1.14.16).

1. Check the initial value in the frequency-domain, $s \to \infty$:

$$f(0) = \lim_{s\to\infty} s\,F(s) = \lim_{s\to\infty} \frac{s_1 s_2}{(s - s_1)(s - s_2)} = 0 \tag{1.14.30}$$

which is correct. But in the time-domain at $t = 0$:

$$f(0) = 1 + \frac{1}{|\sin\theta|}\, e^{\sigma_1 0} \sin(\omega_1 0 + \theta) = 2 \tag{1.14.31}$$

which is **wrong!**

2. Check the final value for $t \to \infty$:

$$f(\infty) = \lim_{t\to\infty} \left[1 + \frac{1}{|\sin\theta|}\, e^{\sigma_1 t} \sin(\omega_1 t + \theta)\right]$$

$$= s\,F(0) = \frac{s_1 s_2}{(0 - s_1)(0 - s_2)} = 1 \tag{1.14.32}$$

and at least this one is correct in both the time and frequency domain. Note that in both checks the pole at $s = 0$ is canceled by the multiplication of $F(s)$ by s.

Considering the error in the initial value in the time domain, many engineers wrongly assume that they have made a **sign** error and change the time domain equation to:

$$f(t) = 1 - \left|\frac{1}{\sin\theta}\right| e^{\sigma_1 t} \sin(\omega_1 t + \theta) \qquad \textbf{(wrong!)}$$

Although the step response plot will now be correct, a careful analysis shows that the negative sign is completely unjustified! Instead we should have used:

$$\theta = \pi + \arctan\left(\frac{-\omega_1}{\sigma_1}\right) \tag{1.14.33}$$

The reason for the added π lies in the *tangent* function, which repeats with a period of π radians (and not 2π, as the *sine* and *cosine* do). This results in a lost sign since the *arctangent* can not tell between angles in the first quadrant from those in third, and angles in the second quadrant from those in fourth. See Appendix 2.3 (web only) for more of such cases in 3rd- and 4th-order systems.

A graphical presentation of the step response solution, given by Eq. 1.14.29 and with the correct initial phase angle, Eq. 1.14.33, is displayed in Fig. 1.14.3.

The physical circuit behavior can be explained as follows:

The system resonance term, $\sin(\omega_1 t)$, is first shifted by θ, the characteristic angle of the pole, becoming $\sin(\omega_1 t + \theta)$ (the time shift is θ/ω_1). At resonance the voltage and

current in reactive components are each others' derivatives (a sine–cosine relationship, see Eq. 1.14.26), the initial phase angle θ reflects their impedance ratios.

The amplitude of the shifted function is then corrected by the absolute value of the function at $t = 0$, which is $|1/\sin\theta|$. Thus the starting value is equal to -1, and in addition the slope is precisely identical to the initial slope of the exponential damping function, $e^{\sigma_1 t}$, so that their product has **zero** initial slope.

This product is the system reaction to the unit step excitation, $h(t)$, which sets the final value for $t \to \infty$ ($s_0 = 0$). By summing the residue at s_0 ($\text{res}_0 = 1$) with the reaction function gives the final result, the step response $f(t)$.

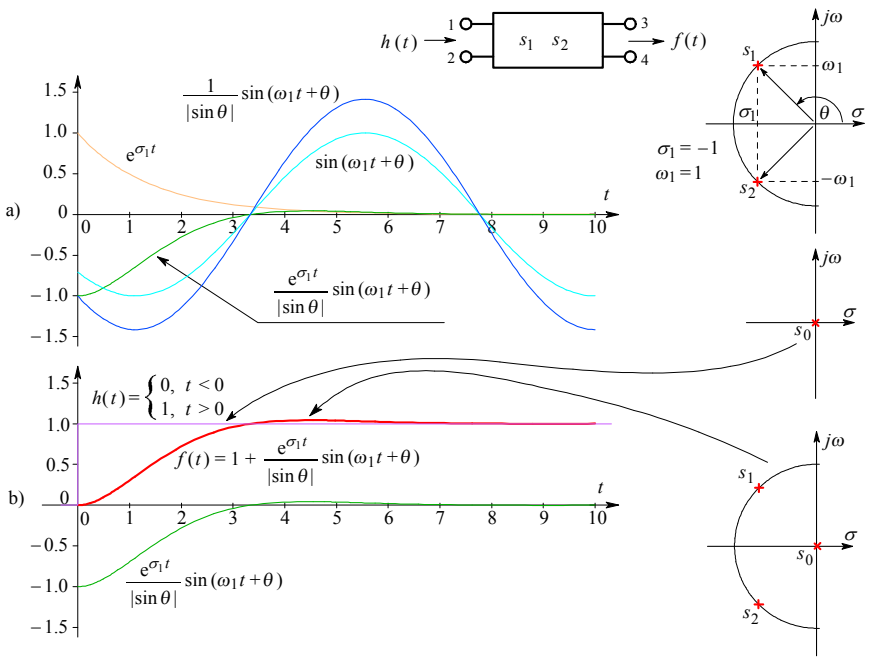


Fig. 1.14.3: Step by step graphic representation of the procedures used in the calculation of the step response of a system with two complex conjugate poles.

We have purposely presented the complete calculation of the step response for the RLC circuit in every detail for two reasons:

1) to show how the step response is calculated by means of the $\mathcal{L}^{-1}$ transform and the theory of residues; and

2) because we shall meet such functions very frequently in the following parts.

1.15 Convolution

In network analysis we often encounter a cascade of networks, so that the output of the preceding network is driving the input of the following one. The output of the later network is therefore a response to the response of the preceding network. We need a procedure to solve such problems. In the time domain this is done by the convolution integral [Ref. 1.2]. Fig. 1.15.1 displays the complete procedure of convolution.

In Fig. 1.15.1a there are two networks:

The network A has a Bessel pole pair with the following data:

$$s_{1,2} = \sigma_1 \pm j\,\omega_1 = -1.500 \pm j0.866; \quad \text{the pole angle is } \theta_1 = \pm 150°.$$

In addition, owing to the input unit step function, we have a third pole $s_0 = 0$.

The network B has a Butterworth pole pair with the following data:

$$s_{3,4} = \sigma_2 \pm j\,\omega_2 = -0.7071 \pm j0.7071; \quad \text{the pole angle is } \theta_2 = \pm 135°.$$

Bessel and Butterworth poles are discussed in detail in Part 4 and Part 6.

According to Eq. 1.14.29 the step response of the network A is:

$$f(t) = 1 + \frac{1}{|\sin\theta_1|}\,\mathrm{e}^{\sigma_1 t}\sin(\omega_1 t + \theta_1)$$

and, according to Eq. 1.14.17, the impulse response of the network B is:

$$g(t) = \frac{1}{|\sin\theta_2|}\,\mathrm{e}^{\sigma_2 t}\sin\omega_2 t$$

Both functions are shown in Fig. 1.15.1a. We will convolve $g(t)$ because it is easier to do so. This convolving (folding) is done by time reversal about $t = 0$, obtaining $g(\tau - t)$. The reversion interval τ has to be chosen so that $g(t \geq \tau) = 0$ (or at least very close to zero), otherwise the convolution integral would not converge to the correct final value. The output function $y(t)$ is then the convolution integral:

$$y(t) = \int_0^{t_{\max}} \underbrace{\left[1 + \frac{1}{|\sin\theta_1|}\,\mathrm{e}^{\sigma_1 t}\sin(\omega_1 t + \theta_1)\right]}_{f(t)} \underbrace{\frac{1}{|\sin\theta_2|}\,\mathrm{e}^{\sigma_2(\tau - t)}\sin\omega_2(\tau - t)}_{g(\tau - t)} dt \qquad (1.15.1)$$

To solve this integral requires a formidable effort and the reader may be assured that we shall not attempt to solve it here, because — as we will see later — there is a more elegant method of doing so. We have written the complete integral merely to give the reader an example of the convolution based on the functions which we have already calculated. Nevertheless, it is a challenge for the reader who wants to do it by himself (for the construction of diagrams in Fig. 1.15.1, this integral has been solved!).

In Fig. 1.15.1b we first convolve the function $g(t)$ and introduce the time constant τ to obtain $g(\tau - t)$. Next, in Fig. 1.15.1c the function $g(\tau - t)$ is shifted right along the time axis to the position $t = 1$, obtaining $g(\tau - t + 1)$. The area A_1 under the product of the two signals is the value of the convolution integral for the interval $0 \leq t \leq 1$.

In a similar fashion, in Fig. 1.15.1d the function $g(\tau - t)$ is shifted to $t = 2$. Here the value of the convolution integral for the interval $0 \leq t \leq 2$ is equal to the area A_2.

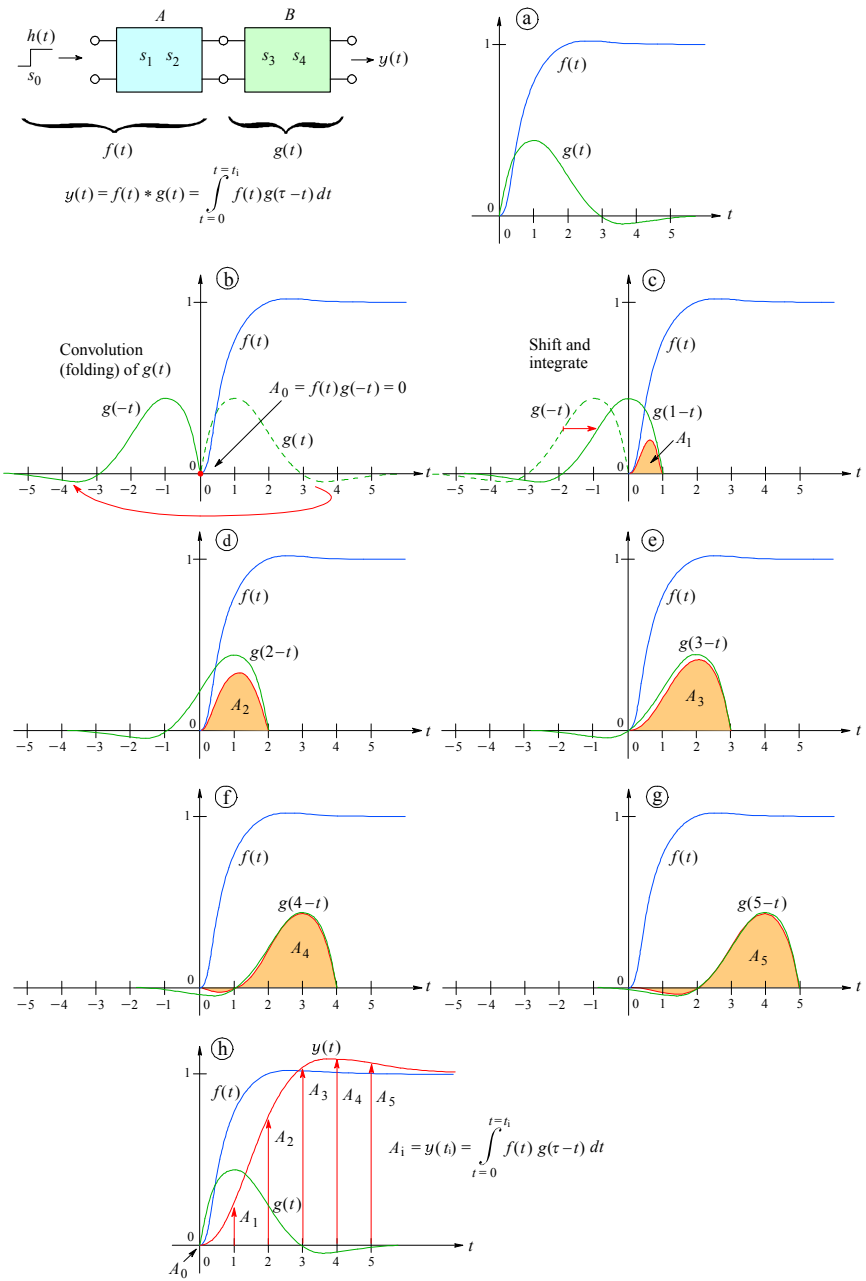


Fig. 1.15.1: Graphic presentation of the mathematical course of the convolution integral, $f(t)*g(t)$. See the text for the description. To see the movie click on this link: **convmovie.gif** (we apologize for the bad quality resulting from the poor conversion from vector graphics to bitmap format; readers who have access to Matlab are invited to run the 'convmovie.m' file).

In Fig. 1.15.1e, the function $g(\tau - t)$ is shifted to $t = 3$ to obtain the area A_3 and in Fig. 1.15.1f, the function $g(\tau - t)$ is shifted to $t = 4$, resulting in the area A_4, which is in part negative, owing to the shape of $g(\tau - t)$.

In Fig. 1.15.1g, $t = 5$ and the area A_5 is obtained. Since $f(t)$ has nearly reached its final value and $g(t)$ is almost zero for $t > 5$, any further shifting changes A_t only slightly.

Finally, in Fig. 1.15.1h the values of $A_1, \ldots, A_5$ are inserted to point to the particular values of the output function $y(t)$. For comparison, the input of network B, $f(t)$, is also drawn. Although $f(t)$ has almost no overshoot, the Butterworth poles in the network B cause a large undershoot in $g(t)$, which results in an overshoot in the output signal $y(t)$.

Important note: In the last plot of Fig. 1.15.1 the system response $y(t)$ is plotted as if the network B had a unity gain. The impulse response of a unity gain system is characterized by the whole area under it being equal to 1; consequently, its peak amplitude would be very small compared to $f(t)$, so there would not be much to see. Therefore for $g(t)$ we have plotted its ideal impulse response. The normalization to a unity-gain is accomplished by dividing the ideal impulse-response by its own time integral (numerically, each instantaneous amplitude sample is divided by the sum of all samples). See Part 6 and Part 7 for more details.

From Eq. 1.6.51, it has become evident that convolution in the time domain corresponds to a simple frequency domain multiplication. This is also shown in Fig. 1.15.2. The upper half of the figure is the s domain whilst the bottom half is the t domain. Instead of making the convolution $g(t) * f(t)$ in the t domain, which is difficult, (see Eq. 1.15.1), we rather perform a simple multiplication $G(s) \cdot F(s)$ in the s domain. Then, by means of the $\mathcal{L}^{-1}$ transform we obtain the function $y(t)$ which we are looking for.

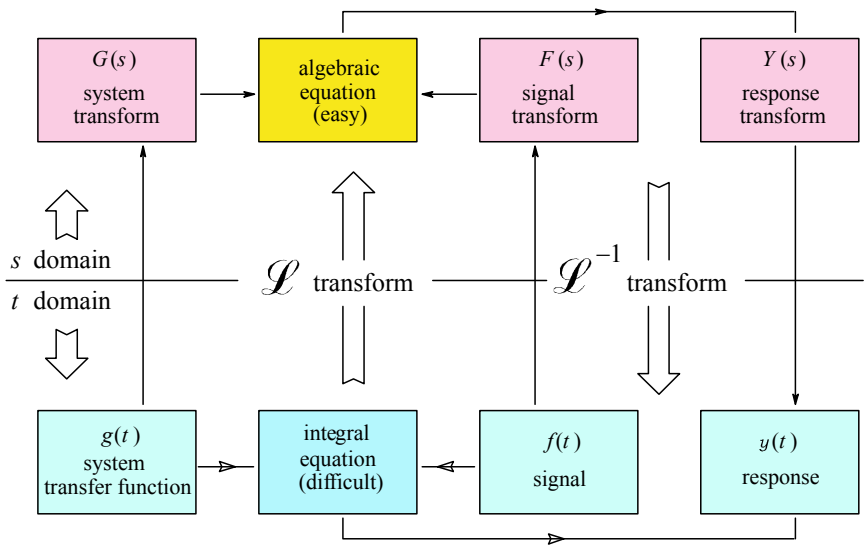


Fig. 1.15.2: Equivalence of the system response calculation in the time domain, $f(t) * g(t)$, and the frequency domain, $F(s) \cdot G(s)$. For analytical work the transform route is the easy way. For computer use the direct method is preferred.

By taking the transform route we need only to calculate the sum of all the residues (five in the case shown in Fig. 1.15.1), which is far less difficult than the calculation of the integral in Eq. 1.15.1.

The mathematical expression, which applies to this case, this is:

$$y(t) = \mathcal{L}^{-1}\{F(s) \cdot G(s)\}$$

$$= \sum \text{res} \underbrace{\left[\frac{s_1\, s_2}{s\,(s - s_1)(s - s_2)} \right]}_{F(s)} \underbrace{\left[\frac{s_3\, s_4}{(s - s_3)(s - s_4)} \right]}_{G(s)} \mathrm{e}^{st} \qquad (1.15.2)$$

Here the numerators of both fractions have been normalized by introducing the products $s_1 s_2$ and $s_3 s_4$ respectively, to replace the constant K (according to Eq. 1.14.16) in the Eq. 1.14.19 and 1.14.8. A solution of the above equation can be found in Part 2, Sec. 2.6.

Fig. 1.15.2 also reveals another very important possibility. If the input signal $f(t)$ is known and a certain output signal $y(t)$ is desired, we can synthesize (not always!) the intermediate network $G(s)$ by taking the $\mathcal{L}$ transform of both time functions and calculating their quotient:

$$G(s) = \frac{Y(s)}{F(s)} \qquad (1.15.3)$$

where $Y(s) = \mathcal{L}\{y(t)\}$ and $F(s) = \mathcal{L}\{f(t)\}$.

Eq. 1.15.1 has convinced us that the calculation of convolution is not an easy task, even for relatively simple functions. By using a PC computer, the convolution in time domain can be calculated **numerically**. Several good mathematical programs exist (we have been using Matlab™ [Ref. 1.18]), which simplify the convolution calculation to a matter of pure routine. This is explained in detail in Part 6 and Part 7.

Résumé of Part 1

So far we have discussed the Laplace transform and its inverse, only to the extent which the reader needs for understanding the rest of the book.

Since we shall calculate many practical examples of the $\mathcal{L}^{-1}$ transform in the following chapters, we have discussed extensively only the calculation of the time function of a simple two pole network with a complex conjugate pole pair, excited by the unit step function.

The readers who want to broaden their knowledge of the Laplace transform, can find enough material for further study in the references quoted.

References:

[1.1] *G.E. Valley & H. Wallman,* Vacuum Tube Amplifiers,
MIT Radiation Laboratory Series, Vol. 18, McGraw-Hill, New York, 1948.

[1.2] *R.B. Randall,* Frequency Analysis,
Brüel & Kjær, Nærum, 1987.

[1.3] *M.F. Gardner & J.L. Barnes*, Transients in Linear Systems Studied by Laplace Transform,
Twelfth Printing, John Wiley & Sons, New York, 1956.

[1.4] *O. Föllinger*, Laplace und Fourier Transformation,
AEG–Telefunken, Berlin, 1982.

[1.5] *G. Doetsch*, Introduction to the Theory and Application of the Laplace Transform,
Springer–Verlag, Berlin, 1970.

[1.6] *G. Doetsch*, Anleitung zum praktischen Gebrauch der Laplace Transformation und der
Z–Transformation, R. Oldenburg Verlag, Munich–Vienna, 1985.

[1.7] *T.F. Bogart, Jr.*, Laplace Transforms and Control Systems, Theory for Technology,
John Wiley & Sons, New York, 1982.

[1.8] *M. O'Flynn & E. Moriarthy*, Linear Systems, Time Domain and Transform Analysis,
John Wiley, New York, 1987.

[1.9] *G.A. Korn & T.M. Korn*, Mathematical Handbook for Scientists and Engineers,
McGraw-Hill, New York, 1961.

[1.10] *M.R. Spiegel*, Theory and Problems of Laplace Transforms,
Schaum's Outline Series, McGraw-Hill, New York, 1965.

[1.11] *J. Plemelj*, Teorija analitičnih funkcij,
Slovenska akademija znanosti in umetnosti, Ljubljana, 1953.

[1.12] *M.R. Spiegel*, Theory and Problems of Complex Variable, SI (Metric) Edition,
McGraw-Hill, New York, 1974.

[1.13] *I. Stewart & D. Tall*, Complex Variables,
Cambridge University Press, Cambridge, 1983.

[1.14] *R.W. Churchil & J.W. Brown*, Complex Variables and Applications, Fourth Edition,
International Student Edition, McGraw-Hill, Auckland, 1984.

[1.15] *W. Gellert, H. Küstner, M. Hellwich, & H. Kästner*, The VNR Concise Encyclopedia of
Mathematics, second edition, Van Nostrand Rheinhold, New York, 1992.

[1.16] *E. Van Valkenburg*, Introduction to Modern Network Synthesis,
John Wiley & Sons, New York, 1960.

[1.17] *P. Starič*, Proof of the Inverse Laplace Transform for Positive Real Functions,
Elektrotehniški vestnik, Ljubljana, 1991, pp. 23–27.

[1.18] *J.N. Little & C.B. Moler*, MATLAB User's Guide,
The MathWorks, Inc., South Natick, USA, 1990.

[1.19] *G.E. Hostetter*, Engineering Network Analysis,
Harper & Row, Publishers, New York, 1984.

[1.20] *J. Bednařík & J. Daněk*, Obrazové zesilovače pro televisí a měřicí techniku,
Statní nakladatelství technické literatury, Prague, 1957.

[1.21] *V. Bubeník*, Impulsová Technika,
Statní nakladatelství technické literatury, Prague, 1958.

[1.22] *D.E. Scott*, An Introduction to System Analysis, A System Approach,
McGraw-Hill, New York, 1987.

[1.23] *P. Kraniauskas*, Transforms in Signals and Systems,
Addison-Wesley, Wokingham, 1992.

[1.24] <http://en.wikipedia.org/wiki/Gibbs_phenomenon/>

P. Starič, E. Margan

Wideband Amplifiers

Part 2:

Inductive Peaking Circuits

Complex solutions always have one simple explanation!
(Lunsford's Rule of scientific endeavor)

The Renaissance of Inductance

In Part 2 of Wideband Amplifiers we discuss various forms of inductive peaking circuits.

The topic of inductive peaking actually started with Oliver Heaviside's "Telegrapher's Equation" back in 1890s, in which for the first time an inductance was used to compensate a dominantly capacitive line to extend the bandwidth. The development flourished with radio, TV and radar circuits and reached a peak in oscilloscopes in 1970s.

With the widespread use of modern high speed low power operational amplifiers and digital electronics, the inductance virtually disappeared in signal transmission path, remaining mostly in power supply filtering and later in switching power supplies. By all too many contemporary electronics engineers, the inductance is being considered more as a nuisance, rather than a useful circuit component.

However, the available frequency spectrum is fixed and the bandwidth requirements are continuously rising, especially with modern wireless telecommunications. In our opinion the inductance just waits to be rediscovered by new generations of electronics circuit designers. So we believe that the inclusion of this subject in the book is fully justified.

Contents:

List of Figures:

List of Tables:

2.0 Introduction

In the early days of wideband amplifiers 'suitable coils' were added to the load (consisting of resistors and stray capacitances) in order to extend the bandwidth, causing in most cases a resonance peak in the frequency response. Hence the term *inductive peaking*. Even though later designers of wideband amplifiers were more careful in doing their best to achieve as flat a frequency response as possible, the word 'peaking' remained in use and it is still used today.

In some respect the British engineer *S. Butterworth* might be considered the first to introduce coils in the (then) anode circuits of electronic tubes to construct an amplifier with a maximally flat frequency (low pass) response. In his work **On the Theory of Filter Amplifiers**, published as early as October 1930 [Ref. 2.1], besides introducing the pole placement which was later named after him, he also mentioned: *"The writer has constructed filter units in which the resistances and inductances are wound round a cylinder of length 3in and diameter 1.25 in, whilst the necessary condensers are contained within the core of the cylinder"*. However, it is hard to tell exactly the year when these 'necessary condensers' were omitted to leave only the stray and inter-electrode capacitances of the electronic tubes to form, together with the properly dimensioned coils and load resistances, a **wideband amplifier** with maximally flat frequency response. This was probably done some time in the mid 1930s, when the first electronic voltmeters, oscilloscopes, and television amplifiers were constructed.

The need for wideband and pulse amplifiers was emphasized with the introduction of radar during the Second World War. A book of historical value, *G. E. Valley & H. Wallman*, **Vacuum Tube Amplifiers** [Ref. 2.2] was written right after the war and published in 1948. Apart from details about other types of amplifiers, the most important knowledge about wideband amplifiers, gained during the war in the Radiation Laboratory at Massachusetts Institute of Technology, was made public. In this work the amplifier step response calculation also received the necessary attention.

After the war people who worked in the Radiation Laboratory spread over USA and UK, and many of them started working at firms where oscilloscopes were produced. Many articles were written about wideband amplifiers with inductive peaking, but books which would thoroughly discuss wideband amplifiers were almost non-existent. The reason was probably because the emphasis has shifted from the frequency domain to the time domain, where a gap-free mathematical discussion was considered difficult. Nevertheless, here and there a book on this subject appeared, and one of the most significant was published in 1957 in Prague: *J. Bednařík & J. Daněk*, **Obrazové zesilovače pro televisí a měřicí techniku**, (Video Amplifiers for Television and Measuring Techniques) [Ref. 2.3]. There the authors attempted to present a thorough discussion of all inductive peaking circuits known at that time and also of high frequency resonant amplifiers. Computers were a rare commodity in those days, with restricted access, and equally rare was the programming knowledge; this prevented the authors from executing some important calculations, which were too elaborate to be done by pencil and paper.

An important change in wideband amplifier design, using inductive peaking, was introduced by *E.L. Ginzton*, *W.R. Hewlett*, *J.H. Jasberg*, and *J.D. Noe* in their revolutionary article **Distributed Amplification**, [Ref. 2.4]. This was an amplifier with electronic tubes connected in parallel, where the grid and anode interconnections were made of lumped sections of a delay line. In this way the bandwidth of the amplifier was extended beyond

the limits imposed by the mutual conductance (g_m) divided by stray capacitance (C_{in}) of electronic tubes. For reasons which we will discuss in Part 3, this type of amplification has a rather limited application if transistors are used instead of electronic tubes. The necessary delay in a distributed amplifier was realized using the so-called 'm-derived' T-coils, which did not have a constant input impedance. **The correct T-coil circuit** was developed in 1964 by *C.R. Battjes* [Ref. 2.17] and was used for inductive peaking of wideband amplifiers. Compared with a simple series peaking circuit, a T-coil circuit improves the bandwidth and rise time exactly twofold. For many years the T-coil peaking circuits were considered a trade secret, so the first complete mathematical derivations were published by a pupil of C.R. Battjes only in the 1980s [Ref. 2.5, 2.6, 2.41] and in 1995 by C. R. Battjes himself [Ref. 2.18]. **Transistor inter-stage coupling with T-coils** represented a special problem, which was solved by *R.I. Ross* in late 1960s. This too was considered a proprietary matter and appeared in print some ten to twenty years later [Ref. 2.7, 2.8, 2.9]. Owing to the superb performance of the T-coil circuit we shall discuss it very thoroughly. The transistor inter-stage T-coil coupling will be derived in Part 3.

Here in Part 2, we shall first explain the basic idea of inductive peaking, followed by the discussion of the peaking circuits with poles only: series peaking two-pole, series peaking three-pole, T-coil two-pole, T-coil three-pole, and L+T four-pole circuits. This will be followed by peaking circuits with poles and zeros: shunt peaking two-pole and one-zero circuit, shunt peaking three-pole and two-zero circuit, and shunt–series peaking circuit. For each of the circuits discussed we shall calculate and plot the frequency, phase, envelope delay, and the step response. The emphasis will be on T-coil circuits, owing to their superb performance. All the necessary calculations will be explained as we proceed and, whenever practical, the complete derivations will be given. The exception is the step response of the series peaking circuit with one complex conjugate pole pair, which was already derived and explained in Part 1. Since the complete calculation for the step-responses of four-pole L+T circuits and shunt–series peaking circuits is rather complicated, only the final formulae will be given. Those readers who want to have the derivations for these circuits as well, will be able to do so themselves by learning and applying the principles derived in Part 1 and 2 (some assistance can also be found in Appendix 2.1, 2.2 and in 2.3 on web only).

To the beginners we strongly recommend the study of Sec. 2.2 and 2.3: the circuit examples are simple enough to allow the analysis to be easily followed and learned; the same methods can then be applied to more sophisticated circuits in other sections, in which some of the most basic details are omitted and some equations imported from those two sections.

At the end of Part 2 we shall draw two diagrams, showing the Butterworth (MFA) frequency responses and the Bessel (MFED) step responses, to offer an easy comparison of performance. Finally, in Appendix 2.4 (web only) we give a summary table containing the essential design parameters and equations for all the circuits discussed.

2.1 The Principle of Inductive Peaking

A simple common base transistor amplifier is shown in Fig. 2.1.1. A current step source i_s is connected to the emitter; the time scale has its origin $t = 0$ at the current step transition time and is normalized to the system time constant, RC. The collector is loaded by a resistor R; in addition there is the collector–base capacitance C_{cb}, along with the unavoidable stray capacitance C_s and the load capacitance C_L in parallel. Their sum is denoted by C.

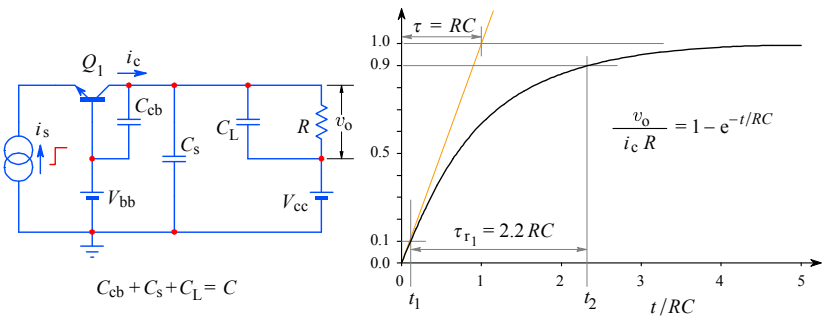


Fig. 2.1.1: A common base amplifier with RC load: the basic circuit and its step response.

Because of these capacitances, the output voltage v_o does not jump suddenly to the value $i_c R$, where i_c is the collector current. Instead this voltage rises exponentially according to the formula (see Part 1, Eq. 1.7.15):

$$v_o = i_c R\left(1 - e^{-t/RC}\right) \tag{2.1.1}$$

The time elapsed between 10 % and 90 % of the final output voltage value ($i_c R$), we call the *rise time*, τ_{r_1} (the index '1' indicates that it is the rise time of a single-pole circuit). We calculate it by inserting the 10 % and 90 % levels into the Eq. 2.1.1:

$$0.1\, i_c R = i_c R\left(1 - e^{-t_1/RC}\right) \quad \Rightarrow \quad t_1 = RC \ln 0.9 \tag{2.1.2}$$

Similarly for t_2:

$$0.9\, i_c R = i_c R\left(1 - e^{-t_2/RC}\right) \quad \Rightarrow \quad t_2 = RC \ln 0.1 \tag{2.1.3}$$

The rise time is the difference between these two instants:

$$\tau_{r_1} = t_2 - t_1 = RC \ln 0.9 - RC \ln 0.1 = RC \ln \frac{0.9}{0.1} = \boxed{2.197\, RC} \approx 2.2\, RC \tag{2.1.4}$$

The value 2.197 *RC* is the reference against which we shall compare the rise time of all other circuits in the following sections of the book.

Since in wideband amplifiers we strive to make the output voltage a replica of the input voltage (except for the amplitude), we want to reduce the rise time of the amplifier as much as possible. As the output voltage rises more current flows through R and less current remains to charge C. Obviously, we would achieve a shorter rise time if we could disconnect R in some way until C is charged to the desired level. To do so let us introduce a switch S between the capacitor C and the load resistor R. This switch is open at time $t = 0$, when the current step starts, but closes at time $t = RC$, as in Fig. 2.1.2. In this way we force all the available current to the capacitor, so it charges linearly to the voltage $i_c R$. When the capacitor has reached this voltage, the switch S is closed, routing all the current to the loading resistor R.

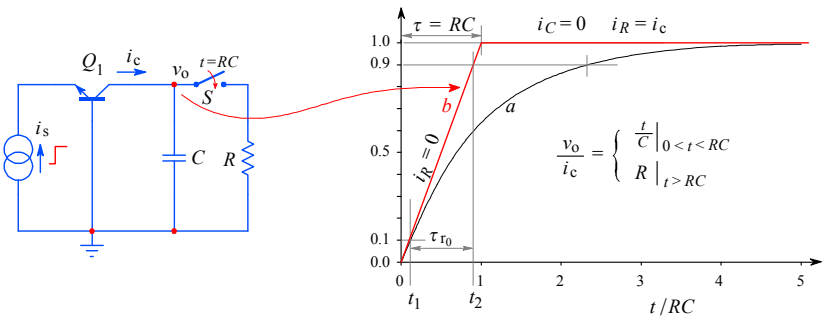


Fig. 2.1.2: A hypothetical ideal rise time circuit. The switch disconnects R from the circuit, so that all of i_c is available to charge C; but after a time $t = RC$ the switch is closed and all i_c flows through R. The resulting output voltage is shown in ***b***, compared with the exponential response in ***a***.

By comparing Fig. 2.1.1 with Fig. 2.1.2 , we note a substantial decrease in rise time τ_{r_0}, which we calculate from the output voltage:

$$v_o = \frac{1}{C}\int_0^{\tau} i_c\,dt = \frac{i_c}{C}\,t\Big|_{t=0}^{t=\tau} = i_c R \tag{2.1.5}$$

where $\tau = RC$. Since the charging of the capacitor is linear, as shown in Fig. 2.1.2, the rise time is simply:

$$\tau_{r_0} = 0.9\,RC - 0.1\,RC = 0.8\,RC \tag{2.1.6}$$

In comparison with Fig. 2.1.1, where there was no switch, the improvement factor of the rise time is:

$$\eta_r = \frac{\tau_{r_1}}{\tau_{r_0}} = \frac{2.20\,RC}{0.8\,RC} = 2.75 \tag{2.1.7}$$

It is evident that the **rise time** (Eq. 2.1.6) **is independent of the actual value of the current i_c, but the maximum voltage $i_c R$** (Eq. 2.1.5) **is not**. On the other hand, the smaller the resistor R the smaller is the rise time. Clearly the introduction of the switch S would mean a great improvement. By using a more powerful transistor and a lower value resistor R we could (at least in principle) decrease the rise time at a will (provided that C remains unchanged). Unfortunately, it is impossible to make a low on-resistance switch,

functioning as in Fig. 2.1.2, which would also suitably follow the signal and automatically open and close in nanoseconds or even in microseconds. So it remains only a nice idea.

But instead of a switch we can insert an appropriate inductance L between the capacitor C and resistor R and so **partially** achieve the effect of the switch, as shown in Fig. 2.1.3. Since the current through an inductor can not change instantaneously, more current will be charging C, at least initially. The configuration of the RLC network allows us to take the output voltage either from the resistor R or from the capacitor C. In the first case we have a *series peaking network*, whilst in the second case we speak of a *shunt peaking network*. Both types of peaking networks are used in wideband amplifiers.

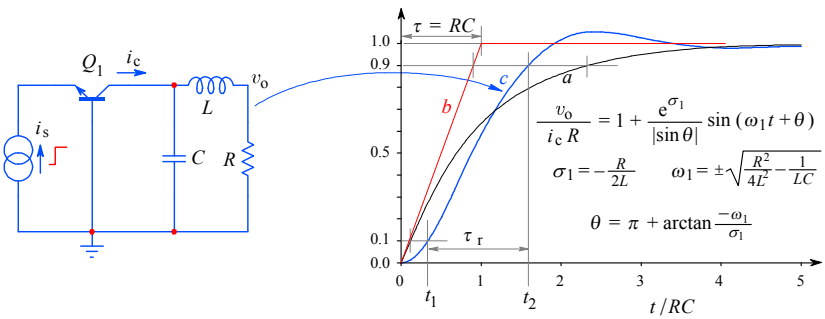


Fig. 2.1.3: A common base amplifier with the series peaking circuit. The output voltage v_o (curve *c*) is compared with the exponential response (*a*, $L = 0$) and the response using the ideal switch (*b*). If we were to take the output voltage from the capacitor C, we would have a shunt peaking circuit (see Sec. 2.7). We have already seen the complete derivation of the procedure for calculating the step response in Part 1, Sec. 1.14. However, the response optimization in accordance with different design criteria is shown in Sec. 2.2 for the series peaking circuit and in Sec. 2.7 for the shunt peaking circuit.

Fig. 2.1.3 is the simplest series peaking circuit. Later, when we discuss T-coil circuits, we shall not just achieve rise time improvements similar to that in Eq. 2.1.7, but in cases in which it is possible (usually it is) to split C into two parts, we shall obtain a substantially greater improvement.

2.2 Two pole series peaking circuit

Besides the series peaking circuit, in this section we shall discuss all the significant mathematical methods which are needed to calculate the frequency, phase and time delay response, the upper half power frequency and the rise time. In addition, we shall derive the most important design parameters of the series peaking circuit, which we will use in the other sections of the book also.

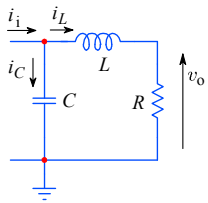


Fig. 2.2.1: A two-pole series peaking circuit.

In Fig. 2.2.1 we have repeated the collector loading circuit of Fig. 2.1.3. Since the inductive peaking circuits are used mostly as collector load circuits, from here on we shall omit the transistor symbol; instead we shall show the input current I_i (formerly I_c) flowing into the network, with the common ground as its drain. At first we shall discuss the behavior of the network in the frequency domain, assuming that I_i is the RMS value of the sinusoidally changing input current i_i. This current is split into two parts: the current through the capacitance I_C, and the current through the inductance I_L. Thus we have:

$$I_i = I_C + I_L = V_i\, j\omega C + \frac{V_i}{j\omega L + R} = V_i \left(j\omega C + \frac{1}{j\omega L + R} \right) \tag{2.2.1}$$

where the input voltage V_i is the product of the driving current I_i and the input impedance Z_i (represented by the expression in parentheses). The output voltage is:

$$V_o = I_L\, R = V_i \frac{R}{j\omega L + R} \tag{2.2.2}$$

From these equations we obtain the transfer function:

$$\frac{V_o}{I_i} = \frac{V_i \dfrac{R}{j\omega L + R}}{V_i \left(j\omega C + \dfrac{1}{j\omega L + R} \right)} = \frac{R}{j\omega C\,(j\omega L + R) + 1}$$

$$= \frac{R}{-\omega^2 LC + Rj\omega C + 1} \tag{2.2.3}$$

Let us set $I_i = 1\,V/R$ and $L = mR^2C$, where m is a dimensionless parameter; also let us substitute $j\omega$ with s. With these substitutions the output voltage $V_o = F(s)$ becomes:

$$F(s) = \frac{1}{s^2\, mR^2C^2 + sRC + 1} = \frac{1}{mR^2C^2} \cdot \frac{1}{s^2 + \dfrac{s}{mRC} + \dfrac{1}{mR^2C^2}} \tag{2.2.4}$$

The denominator roots, which for an efficient peaking must be complex conjugates, as in Fig. 2.2.2, are the poles of $F(s)$:

$$s_{1,2} = \sigma_1 \pm j\omega_1 = -\frac{1}{2mRC} \pm j\sqrt{\frac{1}{mR^2C^2} - \frac{1}{4m^2R^2C^2}} \tag{2.2.5}$$

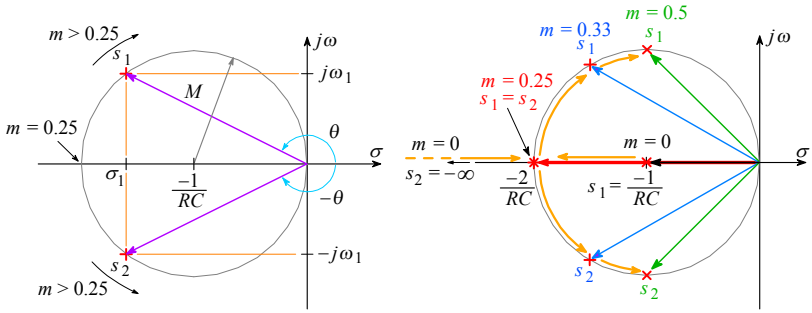


Fig. 2.2.2: The poles s_1 and s_2 in the complex plane. If the parameter $m = 0$, the poles are $s_1 = -1/RC$ and $s_2 = -\infty$. by increasing m, they travel along the real axis towards each other and meet at $s_1 = s_2 = -2/RC$ (for $m = 0.25$). Increasing m further, the poles split into a complex conjugate pair traveling along the circle, the radius of which is $r = 1/RC$ and its center at $\sigma = -r$. The figure on the right shows the four characteristic layouts, which are explained in detail in the text.

With these poles we may write Eq. 2.2.4 also in the following form:

$$F(s) = \frac{1}{mR^2C^2} \cdot \frac{1}{(s - s_1)(s - s_2)} \tag{2.2.6}$$

At DC ($s = 0$) Eq. 2.2.6 shrinks to:

$$F(0) = \frac{1}{mR^2C^2} \cdot \frac{1}{s_1 s_2} \tag{2.2.7}$$

By dividing Eq. 2.2.6 by Eq. 2.2.7, we obtain the **amplitude normalized** transfer function:

$$F(s) = \frac{s_1 s_2}{(s - s_1)(s - s_2)} \tag{2.2.8}$$

We shall need this expression for the calculation of the step response. But for the frequency response $F(j\omega)$ we replace both poles by their components from Eq. 2.2.5 and group the imaginary parts to obtain:

$$F(j\omega) = \frac{\sigma_1^2 + \omega_1^2}{[-\sigma_1 + j(\omega - \omega_1)][-\sigma_1 + j(\omega + \omega_1)]} \tag{2.2.9}$$

We are often interested in the magnitude, $|F(\omega)|$, which we obtain by multiplying $F(j\omega)$ by its own complex conjugate and then taking the root:

$$|F(\omega)| = \sqrt{F(j\omega) \cdot F^*(j\omega)} = \frac{\sigma_1^2 + \omega_1^2}{\sqrt{[\sigma_1^2 + (\omega - \omega_1)^2][\sigma_1^2 + (\omega + \omega_1)^2]}} \tag{2.2.10}$$

The next step is the calculation of the parameter m. Its value depends on the type of poles we want to have, which in turn depend on the intended application of the amplifier. In general, for sine wave signal amplification we prefer the Butterworth poles, whilst for

pulse amplification we prefer the Bessel poles. If high bandwidth is not of primary importance, we can use a 'critically damped' system for a zero overshoot step response. Other types of poles are optimized for use in filters, where our primary goal is to selectively amplify only a part of the spectrum. Poles are discussed in Part 4 (derived from some chosen optimization criteria) and Part 6 (computer algorithms).

2.2.1 Butterworth Poles for Maximally Flat Amplitude Response (MFA)

We shall calculate the actual values of the poles as well as the parameter m, by using Eq. 2.2.5 where we factor out $1/2mRC$. If the square root of Eq. 2.2.11 is imaginary, which is true for $m > 0.25$, we can also factor out the imaginary unit:

$$s_{1,2} = \frac{1}{2mRC}\left(-1 \pm \sqrt{1-4m}\right) = \frac{1}{2mRC}\left(-1 \pm j\sqrt{4m-1}\right) \qquad (2.2.11)$$

We now compare this relation with the normalized 2nd-order Butterworth poles (the reader can find them in Part 4, Table 4.3.1, or by running the BUTTAP computer routine given in Part 6). The values obtained are $\sigma_{1t} = -0.7071$ and $\omega_{1t} = \pm 0.7071$.

> Note: From now on we will append the index 't' to the poles taken from the tables or calculated by a suitable computer program; these values are normalized to the frequency of 1 *radian per second*.
>
> Since both the real and imaginary axis of the Laplace plane have the dimension of frequency, the pole dimension is *radians per second* [rad/s]; however, it has become almost a custom not to write the dimensions.
>
> The sign is also seldom written; instead, most authors leave it to the reader to keep in mind that the poles of unconditionally stable systems always have the real part negative and the imaginary part is either zero or both positive and negative, forming a complex conjugate pair.
>
> To make it easier for the reader, we shall always have the symbols σ and ω signed as required by the mathematical operation to be performed, whilst the numerical values within the symbols will always be negative for σ and positive for ω. For example, we shall express a complex conjugated pole pair $(s_1, s_2) = (s_1, s_1^*)$ as:
>
> $$s_1 = \sigma_1 + j\omega_1 = -0.7071 + j0.7071$$
> $$s_2 = \sigma_2 + j\omega_2 = -0.7071 - j0.7071$$
> $$\Rightarrow \quad s_2 = \sigma_1 - j\omega_1 = s_1^*$$
>
> A real pole will be given as:
>
> $$s_3 = \sigma_3 = -1.000$$
>
> Each σ_i and ω_i will bear the index of the pole s_i (and not their table order number). We shall use the odd index for complex conjugate pair components (with the appropriate $+/-$ sign for the imaginary part).

In order to have the same response, the poles of Eq. 2.2.11 must be proportional to those from the tables, so the ratio of their imaginary to the real part must be the same:

$$\frac{\Im\{s_{1t}\}}{\Re\{s_{1t}\}} = \frac{\Im\{s_1\}}{\Re\{s_1\}} \;\Rightarrow\; \frac{\omega_{1t}}{\sigma_{1t}} = \frac{\omega_1}{\sigma_1} \;\Rightarrow\; \frac{0.7071}{-0.7071} = \frac{\sqrt{4m-1}}{-1} = -1 \qquad (2.2.12)$$

and the same is true for s_2 (except the sign). From the square root of Eq. 2.2.12 it follows that the value of m which satisfies our requirement for the Butterworth poles must be:

$$m = 0.5 \qquad (2.2.13)$$

Thus the inductance is:

$$L = mR^2C = 0.5\,R^2C \qquad (2.2.14)$$

Finally, by inserting the value of m back into Eq. 2.2.11 the poles of our system are:

$$s_{1,2} = \sigma_1 \pm j\omega_1 = \frac{1}{RC}\left(-1 \pm j\right) \qquad (2.2.15)$$

The value $1/RC = \omega_h$ is equal to the upper half power frequency of the non-peaking amplifier of Fig. 2.1.1 (at this frequency, since power is proportional to voltage squared, the voltage gain drops to $1/\sqrt{2} = 0.7071$). If we put $1/RC = 1$ (or $R = 1\,\Omega$ and $C = 1\,\mathrm{F}$, or $R = 500\,\mathrm{k}\Omega$ and $C = 2\,\mu\mathrm{F}$, or any other similar combination, provided that it can be driven by the signal source), we obtain the normalized (denoted by the index 'n') poles:

$$s_{1n,2n} = \sigma_{1n} \pm j\omega_{1n} = -1 \pm j \qquad (2.2.16)$$

If we use normalized poles, we must also normalize the frequency: $j\omega/\omega_h$ instead of $j\omega$.

Note: It is important not to confuse our system with normalized poles (Eq. 2.2.16) with the system having normalized Butterworth poles taken from the table (s_{1t}, $s_{2t} = -0.707 \pm j\,0.707$). Although both are Butterworth-type and both are normalized, they differ in bandwidth:

$$\sqrt{s_{1t}s_{2t}} = 1 \qquad \text{whilst} \qquad \sqrt{s_{1n}s_{2n}} = \sqrt{2} \qquad (2.2.17)$$

This will become evident soon in Sec. 2.2.4, where we shall calculate and plot the magnitude (absolute value) of the frequency response.

2.2.2 Bessel Poles for Maximally Flat Envelope Delay (MFED) Response

From Table 4.4.3 in Part 4 (or by using the BESTAP routine in Part 6), the poles for the 2nd-order Bessel system are $\sigma_{1t} = -1.7544$ and $\omega_{1t} = \pm 1.5000$. Then, as for the Butterworth case above, the ratio of their imaginary to real component is:

$$\frac{\Im\{s_1\}}{\Re\{s_1\}} = \frac{\omega_{1t}}{\sigma_{1t}} \qquad \Rightarrow \qquad \frac{\sqrt{4m-1}}{-1} = \frac{1.5000}{-1.7544} \qquad (2.2.18)$$

Solving for m gives:

$$m = \frac{1}{3} \qquad (2.2.19)$$

So the inductance is:

$$L = 0.3\dot{3}\;R^2C \qquad (2.2.20)$$

and the poles are:

$$s_{1,2} = \frac{1}{RC}\left(-1.5 \pm j\,0.866\right) \qquad (2.2.21)$$

2.2.3 Critical Damping (CD)

In this case both poles are real and equal, so the imaginary part in Eq. 2.2.11 (the square root) must be zero:

$$4m - 1 = 0 \qquad \Rightarrow \qquad m = 0.25 \tag{2.2.22}$$

from which the inductance is:

$$L = 0.25\, R^2 C \tag{2.2.23}$$

resulting in a double real pole:

$$s_{1,2} = -\frac{2}{RC} \tag{2.2.24}$$

In general the parameter m may be calculated with the aid of Fig. 2.2.2, where both poles and the angle θ are shown. If the poles are expressed by Eq. 2.2.11:

$$\tan\theta = \frac{\Im\{s_1\}}{\Re\{s_1\}} = \frac{\omega_1}{\sigma_1} = \frac{\sqrt{4m-1}}{-1} \tag{2.2.25}$$

and from this we obtain:

$$\boxed{m = \frac{1 + \tan^2\theta}{4}} \tag{2.2.26}$$

which is also equal to $1/4\cos^2\theta$, as can be found in some literature. We prefer Eq. 2.2.26 .

Now we have all the data needed for further calculations of the frequency, phase, time delay, and step responses.

2.2.4 Frequency Response Magnitude

We have already written the magnitude in Eq. 2.2.10. Here we will use the normalized frequency ω/ω_h:

$$\boxed{\left|F(\omega)\right| = \frac{\sigma_{1n}^2 + \omega_{1n}^2}{\sqrt{\left[\sigma_{1n}^2 + \left(\frac{\omega}{\omega_h} + \omega_{1n}\right)^2\right]\left[\sigma_{1n}^2 + \left(\frac{\omega}{\omega_h} - \omega_{1n}\right)^2\right]}}} \tag{2.2.27}$$

This is a normalized equation, in magnitude as $\left|F(\omega)\right| = 1$ for $\omega = 0$, and in frequency to the upper half power frequency ω_h of the non-peaking system.

Inserting the pole types of MFA, MFED, and CD, and the frequency in the range $0.1 < (\omega/\omega_h) < 10$, we obtain the diagrams in Fig. 2.2.3 .

2.2.5 Upper Half Power Frequency

An important amplifier parameter is its upper half power frequency, which we shall name ω_H for the peaking amplifier (in contrast to ω_h in the non-peaking case). This is the frequency at which the output voltage V_o drops to $V_{oDC}/\sqrt{2}$, where V_{oDC} is the output voltage at DC ($\omega = 0$), or, if normalized, to $1\,V/\sqrt{2}$. Since the power is proportional to

the square of the voltage, the normalized output power $P_o = (1\,\mathrm{V})^2/2$, which is one half of the output power at DC. We can calculate the upper half power frequency from Eq. 2.2.27, by inserting $\omega = \omega_H$; the result must be $1/\sqrt{2}$:

$$\left|F(\omega_H)\right| = \frac{\sigma_1^2 + \omega_1^2}{\sqrt{\left[\sigma_1^2 + (\omega_H + \omega_1)^2\right]\left[\sigma_1^2 + (\omega_H - \omega_1)^2\right]}} = \frac{1}{\sqrt{2}} \qquad (2.2.28)$$

We shall use the term *upper half power frequency* intentionally, rather than the term *upper −3 dB frequency*, which is commonly found in the literature. Whilst it has become a custom to express the amplifier gain in dB, the dB scale (the log of the output to input power ratio) implies that the driving circuit, which supplies the current I_i to the input, has the same internal resistance as the loading resistor R. This is not the case in most of the circuits which we shall discuss.

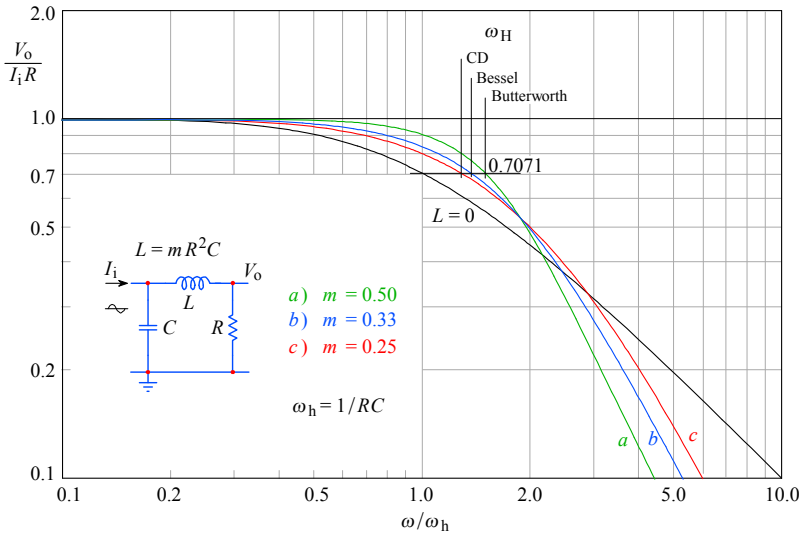


Fig. 2.2.3: Frequency response magnitude of the two-pole series peaking circuit for some characteristic values of m: ***a)*** $m = 0.5$ is the maximally flat amplitude (MFA) response; ***b)*** $m = 0.33$ is the maximally flat envelope delay (MFED) response; ***c)*** $m = 0.25$ is the critical damping (CD) case; the non-peaking case ($m = 0 \Rightarrow L = 0$) is the reference. The bandwidth of all peaking responses is improved compared to the non-peaking bandwidth ω_h at $V_o/I_i R = 0.7071$.

For a series peaking circuit the calculation of ω_H is relatively easy. The calculation becomes progressively more difficult for more sophisticated networks, where more poles and sometimes even zeros are introduced. In such cases it is better to use a computer and in Part 6 we have presented the development of routines which the reader can use to calculate the various response functions.

If we solve Eq. 2.2.28 for ω_H/ω_h we can define [Ref. 2.2, 2.4]:

$$\boxed{\eta_b = \frac{\omega_H}{\omega_h}} \qquad (2.2.29)$$

The value η_b is the cut off frequency improvement factor, defined as the ratio of the system upper half power frequency against that of the non-peaking amplifier (and, since the lower half power frequency of a wideband amplifier is generally very low, usually it is flat down to DC, we may call η_b also the *bandwidth improvement factor*). In Table 2.2.1 at the end of this section the bandwidth improvement factors and other data for different values of the parameter m are given.

2.2.6 Phase Response

We calculate the phase angle φ of the output voltage V_o referred to the input current I_i by finding the phase shift $\varphi_k(\omega)$ of each pole $s_k = \sigma_k \pm j\,\omega_k$ and then sum them:

$$\varphi(\omega) = \sum_{k=1}^{n} \varphi_k(\omega) = \sum_{k=1}^{n} \arctan \frac{\omega \mp \omega_k}{\sigma_k} \tag{2.2.30}$$

In Eq. 2.2.30 we have the ratio of the imaginary part to the real part of the pole, so the pole values may be either exact or normalized. For normalized values we must also normalize the frequency variable as ω/ω_h. Our frequency response function (Eq. 2.2.8) has two complex conjugated poles, therefore the phase response is:

$$\varphi(\omega) = \arctan \frac{\dfrac{\omega}{\omega_h} - \omega_{1n}}{\sigma_{1n}} + \arctan \frac{\dfrac{\omega}{\omega_h} + \omega_{1n}}{\sigma_{1n}} \tag{2.2.31}$$

In Fig. 2.2.4 the phase plots corresponding to the same values of m as in Fig. 2.2.3 are shown:

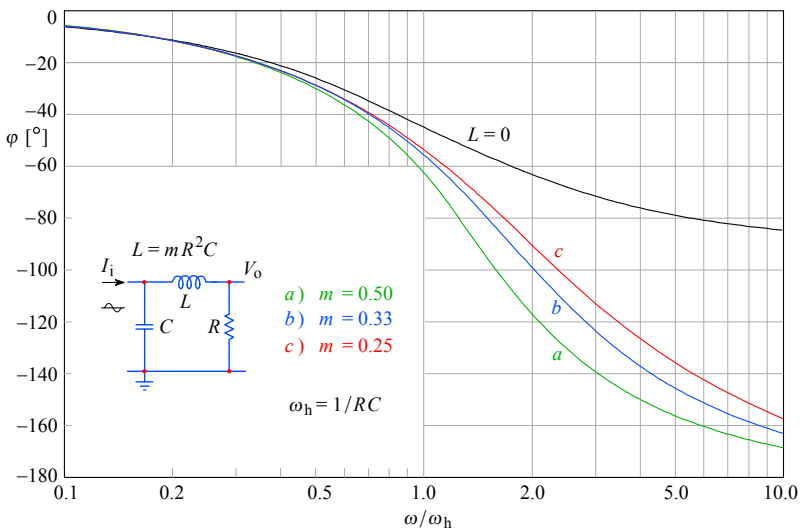


Fig. 2.2.4: Phase response of the series peaking circuit for ***a)*** MFA, ***b)*** MFED and ***c)*** CD case, compared with the non-peaking response ($L = 0$). The phase angle scale was converted from radians to degrees by multiplying it by $180/\pi$. For $\omega \to \infty$ the non-peaking (single-pole) response has its asymptote at 90°, whilst the second-order peaking systems have their asymptote at 180°.

2.2.7 Phase Delay and Envelope Delay

For each pole the *phase delay* (or the *phase advance* for each zero) is:

$$\tau_\varphi = \frac{\varphi}{\omega} \tag{2.2.32}$$

If ω is the positive angular frequency with which the input signal phasor rotates, then the angle φ by which the output signal phasor lags the input is defined in the direction opposite to ω, meaning that, for a phase-delay, φ will be negative, as in Fig. 2.2.4; consequently τ_φ will also be negative. Note that τ_φ has the dimension of time.

Now, τ_φ is obviously frequency dependent, so in order to evaluate the time domain performance of a wideband amplifier on a fair basis we are much more interested in the 'specific' phase delay, known as the *envelope delay* (also *group delay*), which is a frequency derivative of the phase angle as the function of frequency:

$$\tau_\mathrm{e} = \frac{d\varphi(\omega)}{d\omega} \tag{2.2.33}$$

Here, too, a negative result means a delay and a positive result an advance against the input signal. In Fig. 2.2.5 a tentative explanation of the difference between the phase delay and the envelope delay is displayed both in time domain and as a phasor diagram.

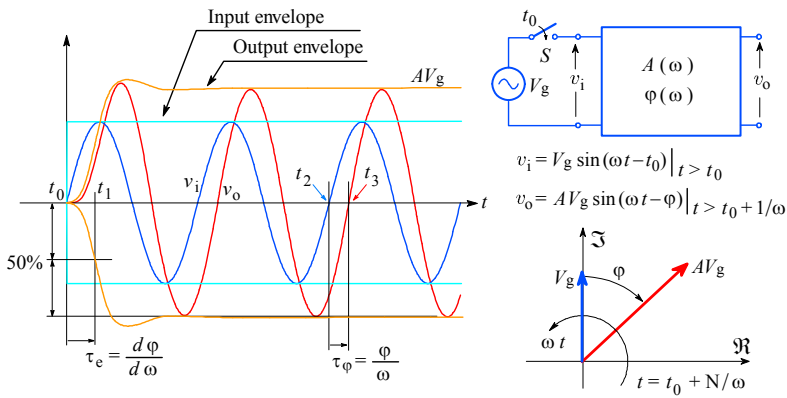


Fig. 2.2.5: Phase delay and envelope delay definitions. The switch S is closed at the instant t_0, applying a sinusoidal voltage with amplitude V_g to the input of the amplifier having a frequency dependent amplitude response $A(\omega)$ and its associated phase response $\varphi(\omega)$. The input signal envelope is a unit step. The output envelope lags the input by $\tau_e = d\varphi/d\omega$, measured from t_0 to t_1, where t_1 is the instant at which the output envelope reaches 50% of its final value. A number of periods later (N/ω), the phase delay can be measured as the time between the input and output zero crossing, indicated by t_2 and t_3, and is expressed as $\tau_\varphi = \varphi/\omega$. Note the phase lag being defined in the opposite direction of the rotation ωt in the corresponding phasor diagram.

In the phase advance case, when zeros dominate over poles, the name suggests that the output voltage will change before input, which is impossible, of course. To see what actually happens we apply a sinewave to two simple RC networks, low pass and high pass, as shown in Fig. 2.2.6. Compare the phase advance case, v_oHP, with the phase delay case, v_oLP. The input signal frequency is equal to the network cutoff, $1/RC$.

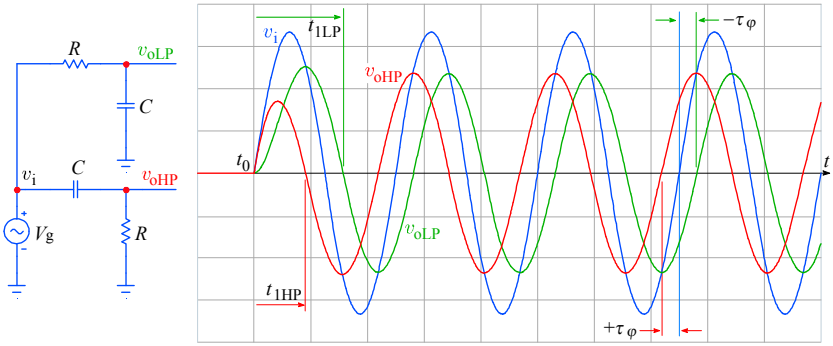


Fig. 2.2.6: Phase delay and phase advance. It is evident that both output signals undergo a phase modulation during the first half period. The time from t_0 to the first 'zero crossing' of the output is shorter for v_{oHP} (t_{1HP}) and longer for v_{oLP} (t_{1LP}). However, both envelopes lag the input envelope. On the other hand, the phase, measured after a number of periods, exhibits an advance of $+\tau_\varphi$ for the high pass network and a delay of $-\tau_\varphi$ for the low pass network.

Returning to the envelope delay for the series peaking circuit, in accordance with Eq. 2.2.33 we must differentiate Eq. 2.2.30. For each pole we have:

$$\frac{d\varphi}{d\omega} = \frac{d}{d\omega}\left[\arctan\frac{\omega \mp \omega_i}{\sigma_i}\right] = \frac{\sigma_i}{\sigma_i^2 + (\omega \mp \omega_i)^2} \tag{2.2.34}$$

and, as for the phase delay, the total envelope delay is the sum of the contributions of each pole (and zero, if any). Again, if we use normalized poles and the normalized frequency, we obtain the normalized envelope delay, $\tau_e\,\omega_h$, resulting in a unit delay at DC.

For the 2-pole case we have:

$$\tau_e\,\omega_h = \frac{\sigma_{1n}}{\sigma_{1n}^2 + \left(\dfrac{\omega}{\omega_h} - \omega_{1n}\right)^2} + \frac{\sigma_{1n}}{\sigma_{1n}^2 + \left(\dfrac{\omega}{\omega_h} + \omega_{1n}\right)^2} \tag{2.2.35}$$

The plots for the same values of m as before, in accordance with Eq. 2.2.35, are shown in Fig. 2.2.7.

For pulse amplification the importance of achieving a flat envelope delay cannot be overstated. A flat delay means that all the important frequencies will reach the output with unaltered phase, preserving the shape of the input signal as much as possible for the given bandwidth, thus resulting in minimal overshoot of the step response (see the next section). Also, a flat delay means that, since it is a phase derivative, the phase must be a linear function of frequency up to the cutoff. This is why Bessel systems are often being referred to as 'linear phase' systems. This property can not be seen in the log scale used here, but if plotted against a linearly scaled frequency it would be seen. We leave it to the curious reader to try it by himself.

In contrast the Butterworth system shows a pronounced delay near the cut off frequency. Conceivably, this will reveal the system resonance upon the step excitation.

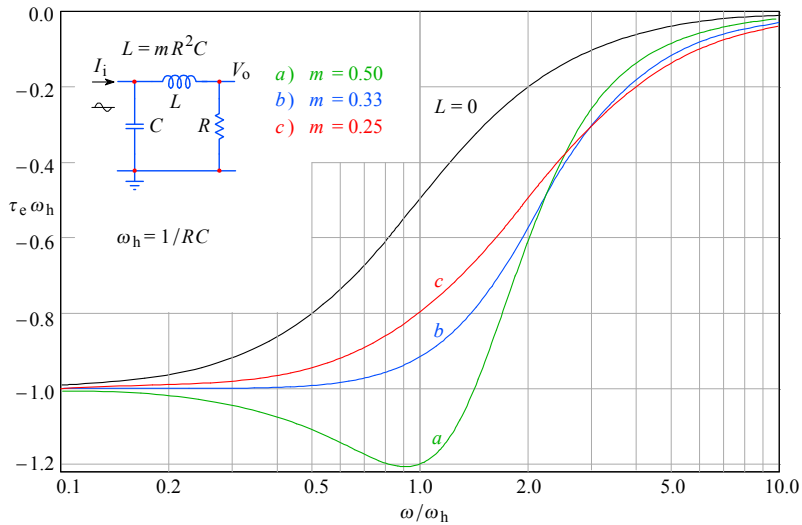


Fig. 2.2.7: Envelope delay of the series peaking circuit for the same characteristic values of m as before: ***a)*** MFA, ***b)*** MFED, ***c)*** CD. Note the MFED plot being flat up to nearly $0.5\,\omega_h$.

2.2.8 Step Response

We have already derived the formula for the step response in Part 1, Eq. 1.14.29:

$$g(t) = 1 + \frac{1}{|\sin\theta|}\,\mathrm{e}^{\sigma_1 t}\sin(\omega_1 t + \theta) \tag{2.2.36}$$

where θ is the pole angle in radians, $\theta = \arctan(\omega_1/\sigma_1) + \pi$ (read the following Note!).

Note: We are often forced to calculate some of the circuit parameters from the trigonometric relations between the real and imaginary components of the pole. The Cartesian coordinates of the pole s_1 in the Laplace plane are σ_1 on the real axis and ω_1 on the imaginary axis. In polar coordinates the pole is expressed as $M\mathrm{e}^{j\theta}$, where M is the modulus (the distance of the pole from the origin of the complex plane):

$$M = \sqrt{(\sigma_1 + j\,\omega_1)(\sigma_1 - j\,\omega_1)} = \sqrt{\sigma_1^2 + \omega_1^2}$$

and its argument (angle) θ, defined so that:

$$\tan\theta = \frac{\Im\{s_1\}}{\Re\{s_1\}} = \frac{\omega_1}{\sigma_1}$$

Now, a mathematically correct definition of the positive-valued angle is counter-clockwise from the positive real axis; so if σ_1 is negative, θ will be greater than $\pi/2$. However, the *tangent* function is defined within the range of $\mp\pi/2$ and then repeats for values between $\pi \pm k\,\pi/2$. Therefore, by taking the *arctangent*, $\theta = \arctan(\omega_1/\sigma_1)$, we loose the information about which half of the complex plane the pole actually lies in and consequently a sign can be wrong. This is bad, because the left (negative) side of the real axis is associated with energy dissipative, that is, resistive circuit action, while the right (positive) side is associated with energy generative action (this is why

unconditionally stable circuits have the poles always in the left half of the complex plane), so it is undesirable to mix up the two cases.

To keep our analytical expressions simple we will keep tracking the pole layout and correct the sign and value of the arctan( ) by adding π radians to the angle θ wherever necessary. But in order to avoid any confusion our computer algorithm should use a different form of equation (see Part 6).

See Appendix 2.3 (on web only) for more details.

To use the normalized values of poles in Eq. 2.2.36 we must also enter the normalized time, t/T, where T is the system time constant, $T = RC$. Thus we obtain:

a) for Butterworth poles (MFA):

$$g_a(t) = 1 + \sqrt{2}\, \mathrm{e}^{-t/T} \sin\left(t/T + 0.785 + \pi\right) \tag{2.2.37}$$

b) for Bessel poles (MFED):

$$g_b(t) = 1 + 2\, \mathrm{e}^{-1.5\, t/T} \sin\left(0.866\, t/T + 0.5236 + \pi\right) \tag{2.2.38}$$

c) for Critical damping (CD) we have a double real pole at s_1, so Eq. 2.2.36 is not valid here, because it was derived for simple poles. To calculate the step response for the function with a double pole, we start with Eq. 2.2.8, insert the same (real!) value $(s_1 = s_2)$ and multiply it by the unit step operator $1/s$. The resulting equation:

$$G(s) = \frac{s_1^2}{s\,(s - s_1)^2} \tag{2.2.39}$$

has the time domain function:

$$g(t) = \mathcal{L}^{-1}\{G(s)\} = \sum \mathrm{res}\, \frac{s_1^2\, \mathrm{e}^{st}}{s\,(s - s_1)^2} \tag{2.2.40}$$

There are two residues, res_0 and res_1; $s = 0$ is a simple pole, so for res_0 we have:

$$\mathrm{res}_0 = \lim_{s \to 0} s \left[\frac{s_1^2\, \mathrm{e}^{st}}{s\,(s - s_1)^2} \right] = 1$$

But s_1 is a double pole, so for res_1 we must use Eq. 1.11.12 in Part 1, which for $n = 2$ results in:

$$\mathrm{res}_1 = \lim_{s \to s_1} \frac{d}{ds} \left[(s - s_1)^2\, \frac{s_1^2\, \mathrm{e}^{st}}{s\,(s - s_1)^2} \right] = \mathrm{e}^{s_1 t}\,(s_1 t - 1)$$

The sum of the residues is then:

$$g(t) = 1 + \mathrm{e}^{\sigma_1 t}\,(\sigma_1 t - 1) \tag{2.2.41}$$

Eq. 2.2.39 has a double **real** pole $s_1 = \sigma_1 = -2/RC$ or, normalized, $\sigma_{1n} = -2$. We insert this in the Eq. 2.2.41 to obtain the CD step response plot (curve c, $m = 0.25$):

$$g_c(t) = 1 - \mathrm{e}^{-2t/T}\,(1 + 2t/T) \tag{2.2.42}$$

The step-response plots of all three cases are shown in Fig. 2.2.8. Also shown is the non-peaking response as the reference ($L = 0$). The MFA overshoot is $\delta = 4.3\,\%$, whilst for the MFED case it is 10 times smaller!

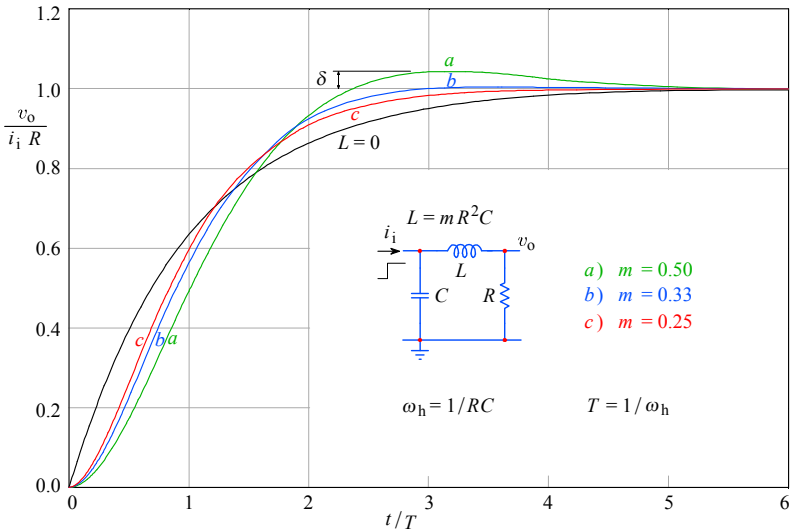


Fig. 2.2.8: Step response of the series peaking circuit for the four characteristic values of m: ***a)*** MFA; ***b)*** MFED; ***c)*** CD. The case $m = 0$ ($L = 0$) is the reference. The MFA overshoot is $\delta = 4.3\,\%$, whilst for MFED it is only $\delta = 0.43\,\%$.

2.2.9 Rise Time

The most important parameter, by which the time domain performance of a wideband amplifier is evaluated, is the rise time. As we have already seen in Fig. 2.1.1, this is the difference between the instants at which the step response crosses the 90 % and 10 % levels of the final value. For the non-peaking amplifier, we have labeled this time as τ_{r_1} and we have already calculated it by Eq. 2.1.4, obtaining the value $\approx 2.20\,RC$. The risetime of a peaking amplifier is labeled τ_R.

To calculate τ_R we use Eq. 2.1.4. For more complex circuits, the step response function can be rather complicated, consequently the analytical calculation becomes difficult, and in such cases it is better to use a computer (see Part 6). The rise time improvement against a non-peaking amplifier is:

$$\boxed{\eta_r = \frac{\tau_{r_1}}{\tau_R}} \qquad (2.2.43)$$

The values for the bandwidth improvement η_b and for the rise time improvement η_r are similar, but in general **they are not equal**. In practice we more often use η_b, the calculation of which is easier. If the step response overshoot is not too large ($\delta < 2\,\%$) we can **approximate** the rise time by starting from the formula for the cut off frequency:

$$\omega_h = 2\pi f_h = \frac{1}{RC} \qquad \text{and furthermore} \qquad f_h = \frac{1}{2\pi RC}$$

where ω_h is the upper half power frequency in [rad/s], whilst f_h is the upper half-power frequency in Hz. We have already calculated the non-peaking risetime τ_{r_1} by Eq. 2.1.4 and found it to be $\approx 2.20\,RC$. From this we obtain $\tau_{r_1} f_h = 2.20/2\pi \approx 0.35$, and this relation we meet very frequently in practice:

$$\boxed{\tau_{r_1} \approx \frac{0.35}{f_h}} \tag{2.2.44}$$

By replacing f_h with f_H in this equation, we obtain (an estimate of) the rise time of the **peaking** amplifier. However, Eq. 2.2.44 is exact only for the single-pole amplifier, where the load is the parallel RC network. For all other cases, Eq. 2.2.44 **can be used as an approximation only if the overshoot $\delta < 2\,\%$**. The overshoot of a Butterworth two-pole network amounts to 4.3 % and it becomes larger with each additional pole(-pair), thus calculating the rise time by Eq. 2.2.43 will result in an excessive error. Even greater error will result for networks with Chebyshev and Cauer (elliptic) system poles. In such cases we must compute the actual system step response and find the risetime from it. For Bessel poles, the error is tolerable since the ω_h-normalized Bessel frequency response closely follows the first-order response up to ω_h. Nevertheless, by using a computer to obtain the rise time from the step response will yield a more accurate result.

2.2.10 Input Impedance

We shall use the series peaking network also as an addition to T-coil peaking. This is possible since the T-coil network has a constant input impedance (the T-coil is discussed in Sec. 2.4, 2.5 and 2.6). Therefore it is useful to know the input impedance of the series peaking network. From Fig. 2.2.1 it is evident that the input impedance is a capacitor C in parallel with the serially connected L and R:

$$Z_i = \frac{1}{j\omega C + 1/(j\omega L + R)} = \frac{j\omega L + R}{1 - \omega^2 LC + j\omega RC} \tag{2.2.45}$$

It would be inconvenient to continue with this expression. To simplify we substitute $L = mR^2C$ and $\omega_h = 1/RC$, obtaining:

$$Z_i = R\,\frac{1 + m\left(\dfrac{j\omega}{\omega_h}\right)}{1 - m\left(\dfrac{\omega}{\omega_h}\right)^2 + \dfrac{j\omega}{\omega_h}} \tag{2.2.46}$$

By making the denominator real and carrying out some further rearrangement we obtain:

$$Z_i = R\,\frac{1 + \left(\dfrac{j\omega}{\omega_h}\right)\left[(m-1) - m^2\left(\dfrac{\omega}{\omega_h}\right)^2\right]}{1 + (1-2m)\left(\dfrac{\omega}{\omega_h}\right)^2 + m^2\left(\dfrac{\omega}{\omega_h}\right)^4} \tag{2.2.47}$$

and the phase angle is:

$$\varphi = \arctan\frac{\Im\{Z_i\}}{\Re\{Z_i\}} = \arctan\left\{\left(\frac{\omega}{\omega_h}\right)\left[(m-1) - m^2\left(\frac{\omega}{\omega_h}\right)^2\right]\right\} \tag{2.2.48}$$

The normalized impedance modulus is:

$$\frac{|Z_i|}{R} = \sqrt{\Re\left\{\frac{Z_i}{R}\right\}^2 + \Im\left\{\frac{Z_i}{R}\right\}^2}$$

$$= \frac{\sqrt{1+\left(\frac{\omega}{\omega_h}\right)^2\left[(m-1)-m^2\left(\frac{\omega}{\omega_h}\right)^2\right]^2}}{1+(1-2m)\left(\frac{\omega}{\omega_h}\right)^2+m^2\left(\frac{\omega}{\omega_h}\right)^4} \qquad (2.2.49)$$

In Fig. 2.2.9 the plots of Eq. 2.2.49 and Eq. 2.2.48 for the same values of m as before are shown:

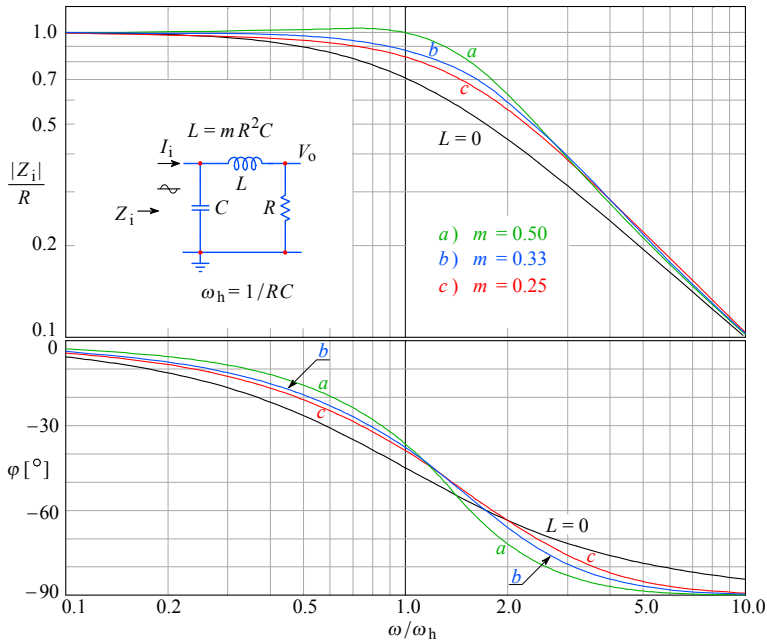


Fig. 2.2.9: Input impedance modulus (normalized) and the associated phase angle of the series peaking circuit for the characteristic values of m. Note that for high frequencies the input impedance approaches that of the capacitance. ***a)*** MFA; ***b)*** MFED; ***c)*** CD.

Table 2.2.1 shows the design parameters of the two-pole series peaking circuit:

Table 2.2.1

response	m	η_b	η_r	δ [%]
MFA	0.50	1.41	1.45	4.30
MFED	0.33	1.36	1.39	0.43
CD	0.25	1.28	1.31	0.00

Table 2.2.1: 2nd-order series peaking circuit parameters summarized: m is the inductance proportionality factor; η_b is the bandwidth improvement; η_r is the risetime improvement; and δ is the step response overshoot.

2.3 Three Pole Series Peaking Circuit

In a practical amplifier we cannot have a pure two-pole series-peaking circuit. The output of the amplifier is always connected to something, be it the next amplifying stage or, say, a cathode ray tube. Any device connected to the output will have at least some capacitance. Therefore the series peaking circuit shown in Fig. 2.3.1 is what we generally encounter in practice. Here we have three independent reactive elements (two capacitors and one inductor), so the circuit has three poles. In order to extract the greatest possible bandwidth from this circuit, the value of the input capacitor C_i, which is in parallel to the loading resistor R, must always be smaller than the loading capacitance C. Since the network is reciprocal, which means we may exchange the input and the output, the condition $C_i < C$ can always be met. As we will see later, the ratio C/C_i depends on the pole pattern selected and it can not be chosen at random.

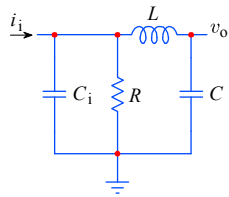


Fig. 2.3.1: The three-pole series peaking circuit.

We shall calculate the network transfer function from the input admittance:

$$Y_i = j\omega C_i + \frac{1}{R} + \frac{1}{j\omega L + \dfrac{1}{j\omega C}} \tag{2.3.1}$$

The input impedance is then:

$$Z_i = \frac{1}{Y_i} = \frac{R(1-\omega^2 LC)}{(1+j\omega C_i R)(1-\omega^2 LC) + j\omega CR} \tag{2.3.2}$$

The input voltage is:

$$V_i = I_i Z_i \tag{2.3.3}$$

and the output voltage is:

$$V_o = a\, V_i = a\, I_i Z_i \tag{2.3.4}$$

where a is the voltage attenuation caused by the elements L and C:

$$a = \frac{\dfrac{1}{j\omega C}}{\dfrac{1}{j\omega C} + j\omega L} = \frac{1}{1-\omega^2 LC} \tag{2.3.5}$$

If we insert Eq. 2.3.2 and Eq. 2.3.5 into Eq. 2.3.4, we obtain:

$$V_o = I_i \frac{R}{1 + j\omega R(C+C_i) - \omega^2 LC - j\omega^3 C_i CLR} \tag{2.3.6}$$

Since $I_i R$ is the voltage at zero frequency, we can obtain the amplitude-normalized transfer function by dividing Eq. 2.3.6 by $I_i R$:

$$F(\omega) = \frac{1}{1 + j\omega R(C + C_i) - \omega^2 LC - j\omega^3 C_i C R L} \tag{2.3.7}$$

Let us now make the following three substitutions:

$$L = m\,R^2(C + C_i) \qquad n = \frac{C}{C + C_i} \qquad \omega_h = \frac{1}{R(C + C_i)} \tag{2.3.8}$$

where ω_h is the upper half power frequency of the non-peaking case ($L = 0$). With these substitutions we obtain the function which is normalized both in amplitude and in frequency (to the non-peaking system cut off):

$$F(\omega) = \frac{1}{1 + j\,\frac{\omega}{\omega_h} - m\,n\left(\frac{\omega}{\omega_h}\right)^2 - j\,m\,n\,(1 - n)\left(\frac{\omega}{\omega_h}\right)^3} \tag{2.3.9}$$

Since the denominator is a 3rd-order polynomial we have three poles, one of which must be real and the remaining two should be complex conjugated (readers less experienced in mathematics can find the general solutions for polynomials of 1st-, 2nd-, 3rd-and 4th-order in Appendix 2.1). Here we shall show how to calculate the required parameters in an easier way. The magnitude is:

$$\left|F(\omega)\right| = \frac{1}{\sqrt{\left(\Re\{F(\omega)\}\right)^2 + \left(\Im\{F(\omega)\}\right)^2}} \tag{2.3.10}$$

By rearranging the real and imaginary parts in Eq. 2.3.9 and inserting them into Eq. 2.3.10, we obtain:

$$\left|F(\omega)\right| = \frac{1}{\sqrt{\left[1 - m\,n\left(\frac{\omega}{\omega_h}\right)^2\right]^2 + \left[\frac{\omega}{\omega_h} - m\,n\,(1 - n)\left(\frac{\omega}{\omega_h}\right)^3\right]^2}} \tag{2.3.11}$$

The squaring of both expressions under the root gives:

$$\left|F(\chi)\right| = \frac{1}{\sqrt{1 + (1 - 2\,m\,n)\chi^2 + m\,n\,[m\,n - 2(1 - n)]\chi^4 + m^2\,n^2(1 - n)^2\chi^6}} \tag{2.3.12}$$

where we have used $\chi = \omega/\omega_h$ in order to be able to write the equation on a single line.

2.3.1 Butterworth Poles (MFA)

The magnitude of the normalized frequency response for a three-pole Butterworth function is:

$$\left|F(\omega)\right| = \frac{1}{\sqrt{1 + \left(\frac{\omega}{\omega_h}\right)^6}} \tag{2.3.13}$$

By comparing Eq. 2.3.13 with Eq. 2.3.12 we realize that the factors at $(\omega/\omega_h)^2$ and at $(\omega/\omega_h)^4$ in Eq. 2.3.12 must be zero if we want the function to correspond to Butterworth poles:

$$\begin{aligned} 1 - 2\,m\,n = 0 \qquad & \text{and} \qquad m\,n - 2(1-n) = 0 \\ \Rightarrow \qquad m = 2/3 \qquad & \text{and} \qquad n = 3/4 \end{aligned} \tag{2.3.14}$$

With these data we can calculate the actual values of Butterworth poles and the upper half power frequency. By inserting m and n into Eq. 2.3.12 and, considering that now the coefficients at $(\omega/\omega_h)^2$ and at $(\omega/\omega_h)^4$ are zero, we obtain the frequency response; its plot is shown in Fig. 2.3.2 as curve *a*.

To calculate the poles we insert the values for m and n into Eq. 2.3.9 and by inserting s instead of $j\omega/\omega_h$, the denominator of Eq. 2.3.9 gets the form:

$$\mathcal{D} = 0.125\,s^3 + 0.5\,s^2 + s + 1 \tag{2.3.15}$$

To obtain the canonical form we divide this equation by 0.125. Then to find the roots we equate it to zero:

$$s^3 + 4\,s^2 + 8\,s + 8 = 0 \tag{2.3.16}$$

The roots of this function are the normalized poles of the function $F(s)$:

$$\begin{aligned} s_{1n}\,, s_{2n} &= \sigma_{1n} \pm j\,\omega_{1n} = -1 \pm j\,\sqrt{3} \\ s_{3n} &= \sigma_{3n} = -2 \end{aligned} \tag{2.3.17}$$

The values are the normalized to ω_h, considering that $\omega_h = 1/R(C_i + C) = 1$.

All Butterworth poles lie on a circle with a radius equal to the system upper half-power frequency; the real pole s_{3n} also lies on the same circle. Now remember that the poles in tables are normalized to $\omega_h = 1$ rad/s, so their radius is equal to 1. This means that (for Butterworth poles only!) s_{3n}/s_{3t} is already the bandwidth improvement ratio, $\omega_H/\omega_h = \eta_b$, and in our case it is equal to 2 (we would obtain the same value from the factor at $(\omega/\omega_h)^6$ of Eq.2.3.12, $1/\sqrt[6]{m^2n^2(1-n)^2} = 2$).

2.3.2. Bessel Poles (MFED)

The 'classical' way of calculating the parameters m and n for Bessel poles is first to derive the formula for the envelope delay, $\tau_e = d\varphi/d\omega$. This is a rational function of ω. By equating the two coefficients in the numerator with the corresponding two in the denominator polynomial we obtain two equations from which both parameters may be calculated. However, this is a lengthy and error-prone procedure. A more direct and easier way is as follows: in the literature [e.g. Ref. 2.10, 2.11], or with an appropriate computer program (as in Part 6, BESTAP), we look for the Bessel 3rd-order polynomial:

$$B_3(s) = s^3 + 6\,s^2 + 15\,s + 15 \tag{2.3.18}$$

The canonical form of the denominator of Eq. 2.3.9, with s instead of $j\omega/\omega_h$, is:

$$\mathcal{D} = s^3 + \frac{s^2}{1-n} + \frac{s}{m\,n\,(1-n)} + \frac{1}{m\,n\,(1-n)} \tag{2.3.19}$$

The functions in Eq. 2.3.18 and Eq. 2.3.19 must be the same. This is only possible if the corresponding coefficients are equal. Thus we may write the following two equations:

$$\frac{1}{1-n} = 6 \qquad \text{and} \qquad \frac{1}{m\,n\,(1-n)} = 15 \tag{2.3.20}$$

This gives the following values for the parameters:

$$m = 0.480 \qquad \text{and} \qquad n = 0.833 \tag{2.3.21}$$

The roots of Eq. 2.3.18 (or Eq. 2.3.19, with the above values for m and n) are the Bessel poles of the function $F(s)$:

$$s_{1n,2n} = \sigma_{1n} \pm j\,\omega_{1n} = -1.8389 \pm j\,1.7544$$
$$s_{3n} = \sigma_{3n} = -2.3222 \tag{2.3.22}$$

Note that the same values are obtained from the pole tables (or by running the BESTAP, Part 6 routine); in general, for Bessel poles normalized to a unit delay, $s_{kn} = s_{kt}$.

With these poles the frequency response, according to Eq. 2.3.11, results in the curve b in Fig. 2.3.2. The Bessel poles are derived from the condition that the transfer function has a unit envelope delay at the origin, so there is no simple way of relating it to the upper half power frequency ω_H. We need to calculate $|F(\omega)|$ numerically for a range, say, $1 < \omega/\omega_h < 3$, using either Eq. 2.3.11 or the FREQW algorithm in Part 6, and find ω_H from it. The bandwidth improvement factor for Bessel poles is given in Table 2.3.1 at the end of this section.

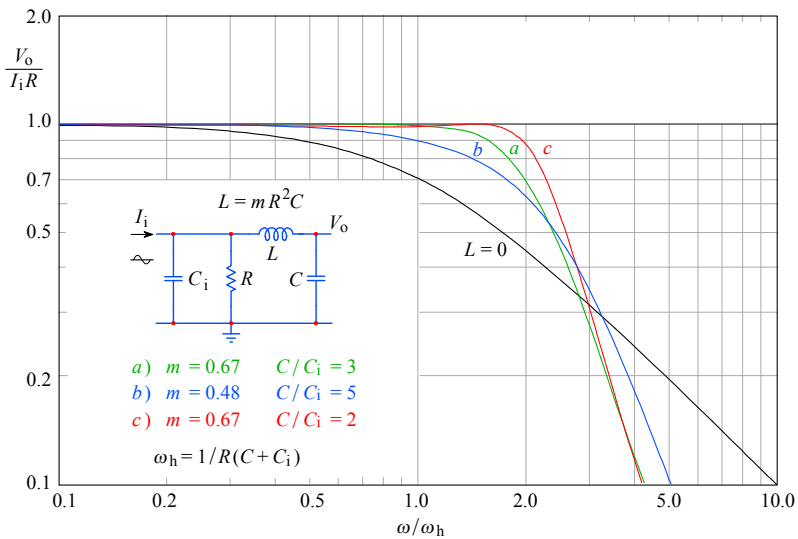


Fig. 2.3.2: Frequency response of the third-order series peaking circuit for different values of m. The correct setting for the required pole pattern is achieved by the input to output capacitance ratio, C/C_i. Fair circuit performance comparison is met by normalization to the total capacitance $C + C_i$. Here we have: ***a)*** MFA; ***b)*** MFED; ***c)*** SPEC, and the non-peaking ($L = 0$) case as a reference. Although being of highest bandwidth, the SPEC case is non-optimal, owing to the slight but notable dip in the range $0.5 < \omega/\omega_h < 1.2$.

2.3.3 Special Case (SPEC)

In practice it is sometimes difficult to achieve the capacitance ratio C/C_i required for Butterworth or for Bessel poles. Let us see what the frequency response would be if we take the capacitance ratio $C/C_\mathrm{i} = 2$, which we shall call a *special case* (SPEC). This makes both parameters equal, $m = n = 0.667$, and the canonical form of the denominator in Eq. 2.3.9, where $(j\omega/\omega_\mathrm{h}) = s$, is then:

$$\mathcal{D} = s^3 + 3\,s^2 + 6.7255\,s + 6.7255 = 0 \tag{2.3.23}$$

Its roots are the required poles:

$$\begin{aligned} s_{1n,2n} &= \sigma_{1n} \pm j\,\omega_{1n} = -0.7500 \pm j\,1.9848 \\ s_{3n} &= \sigma_{3n} = -1.5000 \end{aligned} \tag{2.3.24}$$

The corresponding frequency response is the curve c in Fig. 2.3.2. This gives a bandwidth improvement $\eta_\mathrm{b} = 2.28$, which sounds very fine if there were not a small dip in the range $0.5 < (\omega/\omega_\mathrm{h}) < 1.2$. So we regrettably realize that the ratio C/C_i can not be chosen at random. The aberrations are even greater for the envelope delay and the step response, as we shall see later.

2.3.4 Phase Response

For the calculation of phase response we can use Eq. 2.2.31, but we must also add the influence of the real pole σ_{3n}:

$$\varphi = \arctan\frac{\dfrac{\omega}{\omega_\mathrm{h}} - \omega_{1n}}{\sigma_{1n}} + \arctan\frac{\dfrac{\omega}{\omega_\mathrm{h}} + \omega_{1n}}{\sigma_{1n}} + \arctan\frac{\dfrac{\omega}{\omega_\mathrm{h}}}{\sigma_{3n}} \tag{2.3.25}$$

In Fig. 2.3.3 we have plotted the phase response for different values of parameters m and n. Instead of the parameter n, the ratio C/C_i is given.

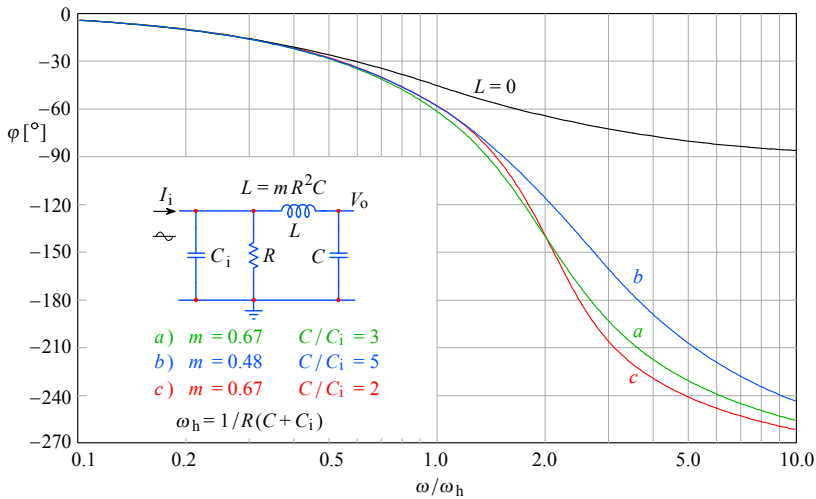


Fig. 2.3.3: Phase response of the third-order series peaking circuit for different values of m: ***a)*** MFA; ***b)*** MFED; ***c)*** SPEC; the non-peaking ($L = 0$) case is the reference.

2.3.5. Envelope-delay

We apply Eq. 2.2.35 to which we add the influence of the real pole σ_{3n}:

$$\tau_e\omega_h = \frac{\sigma_{1n}}{\sigma_{1n}^2 + \left(\frac{\omega}{\omega_h} + \omega_{1n}\right)^2} + \frac{\sigma_{1n}}{\sigma_{1n}^2 + \left(\frac{\omega}{\omega_h} - \omega_{1n}\right)^2} + \frac{\sigma_{3n}}{\sigma_{3n}^2 + \left(\frac{\omega}{\omega_h}\right)^2} \tag{2.3.26}$$

In Fig. 2.3.4 the corresponding plots for different values of parameters m and n are shown; instead of n the ratio C/C_i is given.

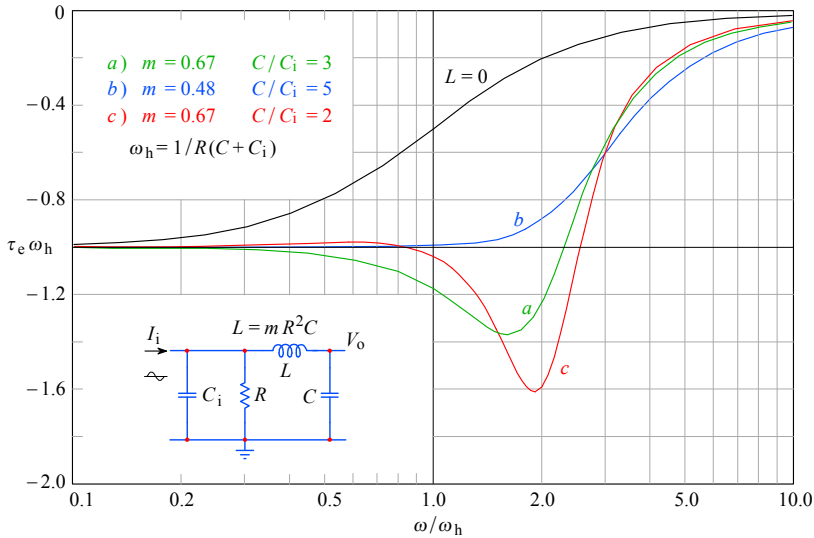


Fig. 2.3.4: Envelope delay of the third-order series peaking circuit for some characteristic values of m: ***a)*** MFA; ***b)*** MFED; ***c)*** SPEC; the non-peaking ($L = 0$) case is the reference. Note the MFED flatness extending beyond ω_h.

2.3.6 Step Response

The calculation is done in a way similar to the case of a two-pole series peaking circuit. Our starting point is Eq. 2.3.9, where we consider that we have two complex conjugate poles s_1 and s_2, and a real pole s_3. The resulting equation must be transformed into a similar form as Eq. 2.2.8. We need a normalized form of equation, so we must multiply the numerator by $-s_1 s_2 s_3$ (see Appendix 2.2). So we obtain a general form:

$$F(s) = \frac{-s_1\, s_2\, s_3}{(s - s_1)(s - s_2)(s - s_3)} \tag{2.3.27}$$

Since we apply a unit step to the network input, the above expression must be multiplied by $1/s$ to obtain a new, fourth-order function:

$$G(s) = \frac{-s_1\, s_2\, s_3}{s\,(s - s_1)(s - s_2)(s - s_3)} \tag{2.3.28}$$

The sum of the residues of $G(s)$ is the step response:

$$g(t) = \mathcal{L}^{-1}\{G(s)\} = \sum_{i=0}^{3} \text{res}_i\,\{G(s)\} \tag{2.3.29}$$

Since the calculation of a three-pole network step response is lengthy, we give here only the final result. The curious reader can find the full derivation in Appendix 2.3 (web only).

$$g(t) = 1 - \frac{\sigma_3}{\omega_1 C}\sqrt{A^2 + \omega_1^2 B^2}\;\mathrm{e}^{\sigma_1 t}\sin(\omega_1 t + \beta) - \frac{\sigma_1^2 + \omega_1^2}{C}\,\mathrm{e}^{\sigma_3 t} \tag{2.3.30}$$

where:

$$A = \sigma_1(\sigma_1 - \sigma_3) - \omega_1^2 \qquad B = 2\,\sigma_1 - \sigma_3$$
$$C = (\sigma_1 - \sigma_3)^2 + \omega_1^2 \qquad \beta = \arctan(-\omega_1 B/A) + \pi \tag{2.3.31}$$

Note that we have written β for the initial phase angle of the resonance function, instead of the usual θ, in order to emphasize the difference between the response phase and the angle of the complex conjugated pole pair (in two-pole circuits they have the same value). We enter the normalized poles from Eq. 2.3.17, 2.3.22, and 2.3.24, and the normalized time $t/R(C_\mathrm{i} + C) = t/T$, obtaining the step responses (plotted in Fig. 2.3.5):

a) For Butterworth poles, where $m = 0.667$ and $n = 0.750$ ($\beta = \pi$ rad):

$$g_a(t) = 1 + 1.155\,\mathrm{e}^{-t/T}\sin(1.732\,t/T + \pi) - \mathrm{e}^{-2\,t/T} \tag{2.3.32}$$

b) For Bessel poles, where $m = 0.480$ and $n = 0.833$ ($\beta = 2.5970$ rad):

$$g_b(t) = 1 + 1.839\,\mathrm{e}^{-1.839\,t/T}\sin(1.754\,t/T + 2.597) - 1.951\,\mathrm{e}^{-2.322\,t/T} \tag{2.3.33}$$

c) For our Special Case, where $m = n = 0.667$ ($\beta = \pi$ rad):

$$g_c(t) = 1 + 0.756\,\mathrm{e}^{-0.75\,t/T}\sin(1.985\,t/T + \pi) - \mathrm{e}^{-1.5\,t/T} \tag{2.3.34}$$

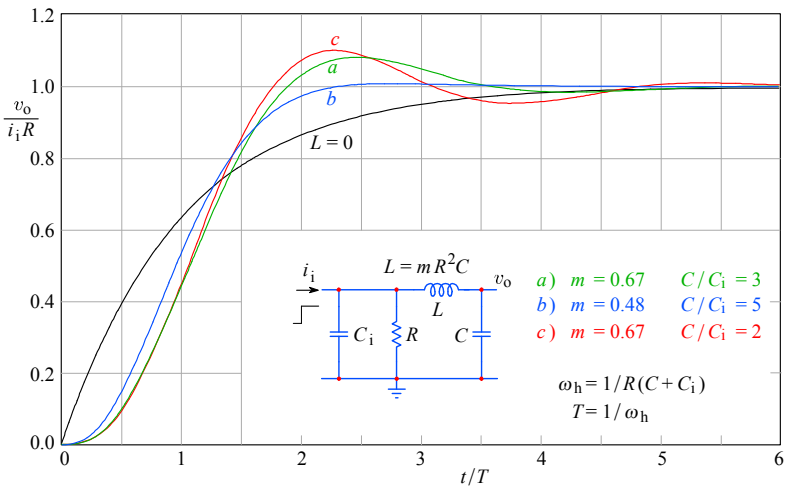


Fig. 2.3.5: Step response of the third-order series peaking circuit for some characteristic values of m: ***a)*** MFA; ***b)*** MFED; ***c)*** SPEC; the non-peaking ($L = 0$) case is the reference. The overshoot of both MFA and SPEC case is too large to be suitable for pulse amplification.

The pole patterns for the three response types discussed are shown in Fig. 2.3.6. Note the three different second-order curves fitting each pole pattern: a (large) horizontal ellipse for MFED, a circle for MFA, and a vertical ellipse for the SPEC case.

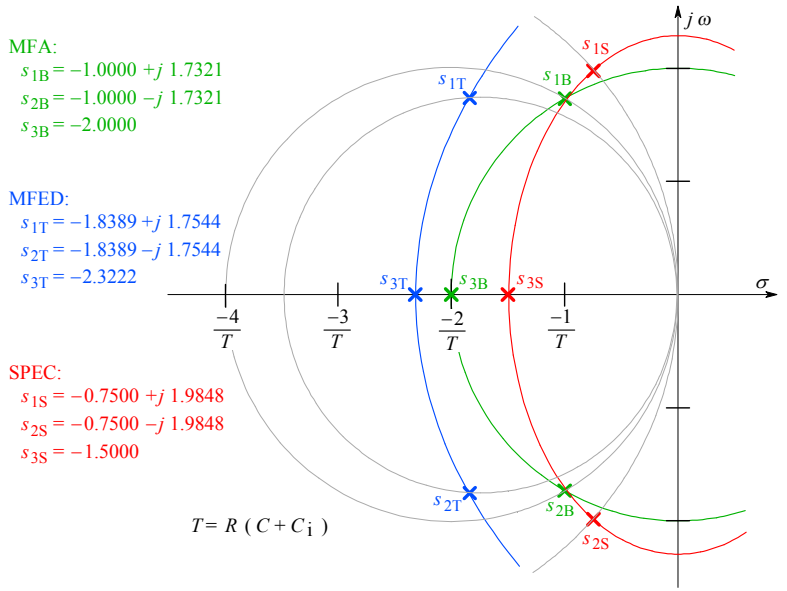


Fig. 2.3.6: Pole patterns of the 3-pole series peaking circuit for the MFA, the MFED, and the SPEC case. The curves on which the poles lie are: a circle with the center at the origin for MFA; an ellipse with both foci on the real axis (the nearer at the origin) for the MFED; and an ellipse with both foci on the imaginary axis for the SPEC case (which is effectively a Chebyshev-type pole pattern). Also shown are the characteristic circles of each complex conjugate pole pair.

Table 2.3.1 resumes the parameters for the three versions of the 3-pole series peaking circuit. Note the high overshoot values for the MFA and the SPEC case, both unacceptable for a pulse amplifier.

Table 2.3.1

response	m	n	η_b	η_r	$\delta\,\%$
a) MFA	0.667	0.750	2.00	1.92	8.1
b) MFED	0.480	0.833	1.75	1.77	0.7
c) SPEC	0.667	0.667	2.28	2.09	10.2

Table 2.3.1: Third-order series peaking circuit parameters.

Table 2.3.1 is based on equal $R(C + C_i)$. From a practical point of view, given the value of the load resistance R, the realizability of the desired response is governed by the ratio C/C_i. If the driving point capacitance C_i is small enough we keep the loading capacitance C as it is and increase C_i as required. Otherwise, we increase C. Since the bandwidth is a function of the sum $C + C_i$, it is clear that an MFED system will be more difficult to realize ($C/C_i = 5$) than an MFA system ($C/C_i = 3$); also its actual bandwidth will be even lower than what can be expected from η_b alone.

2.4 Two-Pole T-coil[1] Peaking Circuit

The circuit schematic of a two-pole T-coil peaking network is shown in Fig. 2.4.1*a*. The main component of this circuit is the center tapped coil L, which is bridged by the capacitance C_b, consisting (ideally) of the coil's self-capacitance [Ref. 2.4, 2.17-2.21]. Since the coils in the equivalent network in Fig. 2.4.1*b* form a letter 'T', we call it a T-coil network. The magnetic coupling factor k between both halves of the coil L and the bridging capacitance C_b must be in a certain relation, dependent on the network poles layout. In addition the relation $R = \sqrt{L/C}$ must hold in order to obtain a constant input impedance $Z_i = R$ at any frequency [Ref. 2.18, 2.21]. This is true if the elements of the network do not have any losses. Owing to losses in a practical circuit, the input impedance may be considered to be constant only up to a certain frequency, which, with a careful design, can be high enough for the application of the T-coil circuit in a wideband amplifier.

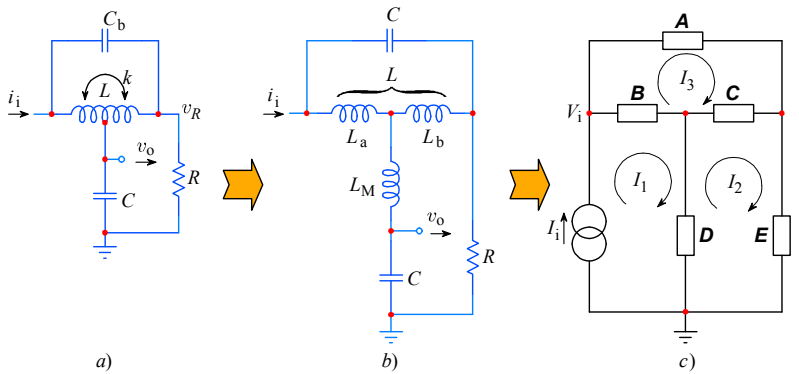


Fig. 2.4.1: ***a)*** The basic T-coil circuit: the voltage output is taken from the center tap node of the inductance L and its two parts are magnetically coupled by the factor $0 < k < 1$; ***b)*** an equivalent circuit, with no magnetic coupling between the coils — it has been replaced by the mutual inductance L_M; ***c)*** a simplified generalized impedance circuit, excited by the current generator I_i, showing the current loops.

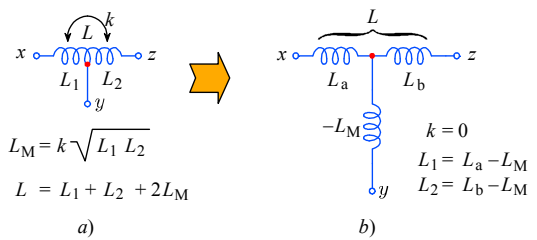


Fig. 2.4.2: Modeling the coupling factor: ***a)*** The T-coil coupling factor k between the two halves L_1 and L_2 of the total inductance L can be represented by ***b)*** an equivalent circuit, having two separate (non-coupled) inductances, in which the magnetic coupling is modeled by the mutual inductance L_M (negative in value), so that $L_1 = L_a - L_M$ and $L_2 = L_b - L_M$.

[1] Networks with tapped coils have been used already in 1948 [Ref. 2.4] and 1954 [Ref. 2.16], but since the bridging capacitance C_b has not been accounted for, the networks described did not have a constant input impedance, as do have the T-coil networks discussed in this and the following three sections.

If the output is taken from the loading resistor R, the network in Fig. 2.4.1*a* behaves as an all pass network. **However, for peaking purposes we take the output voltage from the capacitor *C* and in this application the circuit is a low pass filter.**

The equivalent network in Fig. 2.4.1*b* needs to be explained. We will do this with the aid of Fig. 2.4.2. The original network has a center tapped coil whose inductance L can be calculated by the same general equation for two coupled coils, [Ref. 2.18, 2.28]:

$$L = L_1 + L_2 + 2\,L_\mathrm{M} \tag{2.4.1}$$

where L_1 and L_2 are the inductances of the respective coil parts (which, in general, need not to be equal) and L_M is their mutual inductance. L_M is taken twice, since the magnetic induction from L_1 to L_2 is equal to the induction from L_2 to L_1 and both contribute to the total. If k is the factor of magnetic coupling between L_1 and L_2 the mutual inductance is:

$$L_\mathrm{M} = k\,\sqrt{L_1 L_2} \tag{2.4.2}$$

In the equivalent circuit, with no coupling between the coils, we have:

$$L_\mathrm{a} = L_1 + L_\mathrm{M} \qquad\qquad L_\mathrm{b} = L_2 + L_\mathrm{M} \tag{2.4.3}$$

Then L_1 and L_2 are:

$$L_1 = L_\mathrm{a} - L_\mathrm{M} \qquad\qquad L_2 = L_\mathrm{b} - L_\mathrm{M} \tag{2.4.4}$$

Note the negative sign of L_M, which is a consequence of magnetic coupling; owing to this the driving impedance at the center tap as seen by C is lower than without the coupling. In the symmetrical case, when $L_1 = L_2$, we can calculate the value of L_1 and L_2 from the required coupling k and total inductance L:

$$L_1 = L_2 = \frac{L}{2\,(1+k)} \tag{2.4.5}$$

Thus we have proved that the circuits in Fig. 2.4.1*a* and 2.4.1*b* are equivalent, even though no coupling exists between the coils in the circuit of Fig. 2.4.1*b*.

The corresponding generalized impedance model of the T-coil circuit is shown in Fig. 2.4.1*c*, where the input voltage V_i is equal to the product of the input current and the circuit impedance, $I_\mathrm{i} Z_\mathrm{i}$. The input current splits into I_1 and I_2. The current I_3 flows in the remaining loop. The impedances in the branches are:

$$\begin{aligned} \boldsymbol{A} &= 1/s\,C_\mathrm{b} \\ \boldsymbol{B} &= s\,L_\mathrm{a} \\ \boldsymbol{C} &= s\,L_\mathrm{b} \\ \boldsymbol{D} &= -\,s\,L_\mathrm{M} + 1/s\,C \\ \boldsymbol{E} &= R \end{aligned} \tag{2.4.6}$$

We have written s instead of $j\omega$. With these substitutions the calculation will be much easier. Frankly, from here on, the whole calculation could be done by a suitable computer program, but then some important intermediate results, which we want to explain in detail, would not be shown. So we will do a hand calculation and only at the very end, where the difficulties will increase, shall we use a computer.

We form a system of equations in accordance with the current loops in Fig. 2.4.1*c*:

$$\begin{aligned} V_\mathrm{i} &= I_1(\boldsymbol{B}+\boldsymbol{D}) - I_2\,\boldsymbol{D} - I_3\,\boldsymbol{B} \\ 0 &= -I_1\,\boldsymbol{D} + I_2\,(\boldsymbol{C}+\boldsymbol{D}+\boldsymbol{E}) - I_3\,\boldsymbol{C} \\ 0 &= -I_1\,\boldsymbol{B} - I_2\,\boldsymbol{C} + I_3\,(\boldsymbol{A}+\boldsymbol{B}+\boldsymbol{C}) \end{aligned} \tag{2.4.7}$$

The determinant of the coefficients is:

$$\Delta = \begin{vmatrix} \boldsymbol{B}+\boldsymbol{D} & -\boldsymbol{D} & -\boldsymbol{B} \\ -\boldsymbol{D} & \boldsymbol{C}+\boldsymbol{D}+\boldsymbol{E} & -\boldsymbol{C} \\ -\boldsymbol{B} & -\boldsymbol{C} & \boldsymbol{A}+\boldsymbol{B}+\boldsymbol{C} \end{vmatrix} \tag{2.4.8}$$

with the solution:

$$\begin{aligned} \Delta = {} & (\boldsymbol{B}+\boldsymbol{D})\,[(\boldsymbol{C}+\boldsymbol{D}+\boldsymbol{E})(\boldsymbol{A}+\boldsymbol{B}+\boldsymbol{C}) - \boldsymbol{C}^2] \\ & + \boldsymbol{D}\,[-\boldsymbol{D}(\boldsymbol{A}+\boldsymbol{B}+\boldsymbol{C}) - \boldsymbol{BC}] - \boldsymbol{B}\,[\boldsymbol{DC} + \boldsymbol{B}(\boldsymbol{C}+\boldsymbol{D}+\boldsymbol{E})] \end{aligned} \tag{2.4.9}$$

After multiplication some terms will cancel. Thus the solution is simplified to:

$$\Delta = \boldsymbol{BCA} + \boldsymbol{BDA} + \boldsymbol{BEA} + \boldsymbol{BEC} + \boldsymbol{DCA} + \boldsymbol{DEA} + \boldsymbol{DEB} + \boldsymbol{DEC} \tag{2.4.10}$$

For further calculation we shall need both cofactors Δ_{11} and Δ_{12}. The cofactor for I_1 is:

$$\begin{aligned} \Delta_{11} &= \begin{vmatrix} V_\mathrm{i} & -\boldsymbol{D} & -\boldsymbol{B} \\ 0 & \boldsymbol{C}+\boldsymbol{D}+\boldsymbol{E} & -\boldsymbol{C} \\ 0 & -\boldsymbol{C} & \boldsymbol{A}+\boldsymbol{B}+\boldsymbol{C} \end{vmatrix} \\ &= V_\mathrm{i}\,(\boldsymbol{CA} + \boldsymbol{CB} + \boldsymbol{DA} + \boldsymbol{DB} + \boldsymbol{DC} + \boldsymbol{EA} + \boldsymbol{EB} + \boldsymbol{EC}) \end{aligned} \tag{2.4.11}$$

and in a similar way the cofactor for I_2:

$$\begin{aligned} \Delta_{12} &= \begin{vmatrix} \boldsymbol{B}+\boldsymbol{D} & V_\mathrm{i} & -\boldsymbol{D} \\ -\boldsymbol{D} & 0 & -\boldsymbol{C} \\ -\boldsymbol{B} & 0 & \boldsymbol{A}+\boldsymbol{B}+\boldsymbol{C} \end{vmatrix} \\ &= V_\mathrm{i}\,(\boldsymbol{DA} + \boldsymbol{DB} + \boldsymbol{DC} + \boldsymbol{BC}) \end{aligned} \tag{2.4.12}$$

Let us first find the input admittance, which we would like to be equal to $1/R = 1/\boldsymbol{E}$.

$$\begin{aligned} Y &= \frac{I_1}{V_\mathrm{i}} = \frac{\Delta_{11}}{V_\mathrm{i}\Delta} \\ &= \frac{\boldsymbol{CA} + \boldsymbol{CB} + \boldsymbol{DA} + \boldsymbol{DB} + \boldsymbol{DC} + \boldsymbol{EA} + \boldsymbol{EB} + \boldsymbol{EC}}{\boldsymbol{BCA} + \boldsymbol{BDA} + \boldsymbol{BEA} + \boldsymbol{BEC} + \boldsymbol{DCA} + \boldsymbol{DEA} + \boldsymbol{DEB} + \boldsymbol{DEC}} = \frac{1}{\boldsymbol{E}} \end{aligned} \tag{2.4.13}$$

After eliminating the fractions and canceling some terms, we obtain the expression:

$$\boldsymbol{BCA} + \boldsymbol{BDA} + \boldsymbol{BEA} + \boldsymbol{DCA} - \boldsymbol{ECA} - \boldsymbol{E}^2\boldsymbol{A} - \boldsymbol{E}^2\boldsymbol{B} - \boldsymbol{E}^2\boldsymbol{C} = 0 \tag{2.4.14}$$

Now we put in the values from Eq. 2.4.6, considering that $L_\mathrm{a} = L_\mathrm{b}$, perform all the multiplications, and arrange the terms with the decreasing powers of s. We obtain:

$$s\left[\left(\frac{L_\mathrm{a}^2}{C_\mathrm{b}}-\frac{L\,L_\mathrm{M}}{C_\mathrm{b}}\right)-R^2L\right]+\frac{1}{s}\left(\frac{L}{C\,C_\mathrm{b}}-\frac{R^2}{C_\mathrm{b}}\right)=0 \tag{2.4.15}$$

or, in a general form:

$$s\,\boldsymbol{K}_1+s^{-1}\boldsymbol{K}_2=0 \tag{2.4.16}$$

This expression tells us that the input admittance can indeed be made equal to $1/R$, as we wanted in Eq. 2.4.13. For a constant input admittance circuit, Eq. 2.4.16 must be valid for any s [Ref. 2.21]. This is possible only if both $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ are zero (*Ross' method*):

$$\boldsymbol{K}_1=\frac{L_\mathrm{a}^2}{C_\mathrm{b}}-\frac{L\,L_\mathrm{M}}{C_\mathrm{b}}-R^2L=0 \tag{2.4.17}$$

$$\boldsymbol{K}_2=\frac{L}{C\,C_\mathrm{b}}-\frac{R^2}{C_\mathrm{b}}=0 \tag{2.4.18}$$

From this we obtain the following two relations:

$$\begin{aligned} L&=R^2C\\ L_\mathrm{M}&=\frac{L}{4}-R^2C_\mathrm{b}=R^2\left(\frac{C}{4}-C_\mathrm{b}\right)\end{aligned} \tag{2.4.19}$$

For the symmetrical case, with the tap at the center of the coil, $L_\mathrm{a}=L_\mathrm{b}=L/2$. Since only two parameters, C and R, are known initially, we must obtain another, independent equation in order to calculate the parameters L_M and C_b. For this we can use the transimpedance equation, V_o/I_1 ($I_\mathrm{i}=I_1$, see Fig. 2.4.2.*b*). From Fig. 2.4.1*c* it is evident that the current difference I_1-I_2 flows through branch D. This difference current, multiplied by the impedance $1/sC$, is equal to the output voltage V_o. The transimpedance is then:

$$\frac{V_\mathrm{o}}{I_1}=\frac{1}{s\,C}\cdot\frac{I_1-I_2}{I_1} \tag{2.4.20}$$

The currents are calculated by Cramer's rule:

$$I_1=\frac{\Delta_{11}}{\Delta}\qquad\text{and}\qquad I_2=\frac{\Delta_{12}}{\Delta} \tag{2.4.21}$$

and if we put these expressions into Eq. 2.4.20 we obtain:

$$\frac{V_\mathrm{o}}{I_1}=\frac{1}{s\,C}\cdot\frac{\Delta_{11}-\Delta_{12}}{\Delta_{11}} \tag{2.4.22}$$

Again we make use of the common expressions in Eq. 2.4.6. The difference of both cofactors is:

$$\Delta_{11}-\Delta_{12}=V_\mathrm{i}(\boldsymbol{CA}+\boldsymbol{EA}+\boldsymbol{EB}+\boldsymbol{EC}) \tag{2.4.23}$$

With these expressions, the transimpedance is:

$$\frac{V_\mathrm{o}}{I_1} = \frac{1}{s\,C} \cdot \frac{\boldsymbol{CA + EA + EB + EC}}{\boldsymbol{CA + CB + DA + DB + DC + EA + EB + EC}} \tag{2.4.24}$$

The voltage V_i is a factor of both the numerator and the denominator, so it cancels out. Now we replace the common expressions with those from Eq. 2.4.6, express L_M with Eq. 2.4.19, perform the indicated multiplication, make the long division of the polynomials, and the result is a relatively simple expression:

$$F(s) = \frac{V_\mathrm{o}}{I_1} = \frac{R}{s^2 R^2 C\,C_\mathrm{b} + sRC/2 + 1} \tag{2.4.25}$$

Although the author of this idea, Bob Ross, calculated it 'by hand' [Ref. 2.21], we will not follow his example because this calculation is a formidable work. With modern computer programs (such as *Mathematica* [Ref. 2.34] or similar [Ref. 2.35, 2.38, 2.39, 2.40]), the calculation takes less time than is needed to type in the data.

For those designers who want to construct a distributed amplifier using electronic tubes or FETs (**but not transistors**, as we will see in Part 3!), where the resistor R is replaced by another T-coil circuit and so forth (forming a lumped delay line), it is important to know the transimpedance from the input I_1 to the voltage V_R. The result is:

$$\frac{V_R}{I_1} = R\,\frac{s^2 R^2 C\,C_\mathrm{b} - sRC/2 + 1}{s^2 R^2 C\,C_\mathrm{b} + sRC/2 + 1} \tag{2.4.26}$$

Besides the two poles on the left side of the s-plane, s_{1p} and s_{2p}, this equation also has two symmetrically placed zeros on the right side of the s-plane, s_{1z} and s_{2z}, as shown in Fig. 2.4.3. Since Eq. 2.4.26 has equal powers of s both in the numerator and the denominator it is an all pass response. We shall return to this when we shall calculate the step response.

The poles are the roots of the denominator of Eq. 2.4.25. The canonical form is:

$$s^2 + s\,\frac{1}{2RC_\mathrm{b}} + \frac{1}{R^2 C\,C_\mathrm{b}} = 0 \tag{2.4.27}$$

In general the roots are complex conjugates:

$$s_{1,2} = \sigma_1 \pm j\,\omega_1 = -\frac{1}{4RC_\mathrm{b}} \pm \sqrt{\frac{1}{(4RC_\mathrm{b})^2} - \frac{1}{R^2 C\,C_\mathrm{b}}} \tag{2.4.28}$$

By factoring out $1/4RC_\mathrm{b}$ we obtain a more convenient expression:

$$s_{1,2} = -\frac{1}{4RC_\mathrm{b}} \left(1 \pm \sqrt{1 - \frac{16\,C_\mathrm{b}}{C}} \right) \tag{2.4.29}$$

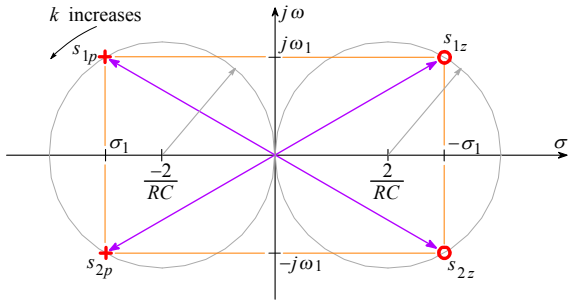


Fig. 2.4.3: The poles (s_{1p} and s_{2p}) and zeros (s_{1z} and s_{2z}) of the all pass transimpedance function corresponding to Eq. 2.4.26 and Fig. 2.4.1*a*. By changing the bridge capacitance C_b and the mutual inductance L_M (by the coupling factor k) according to Eq. 2.4.19, both poles and both zeros travel along the circles shown.

An efficient inductive peaking circuit must have complex poles. By taking the imaginary unit out of the square root, the terms within it exchange signs. Then the pole angle θ can be calculated from the ratio of its imaginary to the real component, as we have done before. From Fig. 2.2.4:

$$\tan\theta = \frac{\Im\{s_1\}}{\Re\{s_1\}} = \frac{\sqrt{\frac{16\,C_b}{C} - 1}}{-1} \tag{2.4.30}$$

This gives a general result:

$$\boxed{C_b = C\,\frac{1+\tan^2\theta}{16}} \tag{2.4.31}$$

The Bessel pole placement is show in Fig. 2.4.4. The characteristic angle θ is measured from the positive real axis.

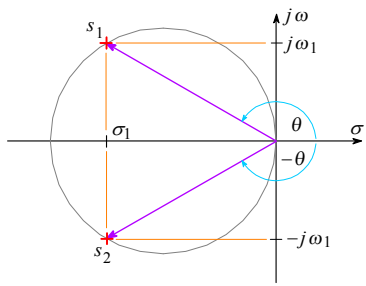


Fig. 2.4.4: The layout of complex conjugate poles s_1 and s_2 of a second-order Bessel transfer function. In this case, the angle is $\theta = 150°$.

By using the pole angle, which we have calculated previously, and Eq. 2.4.31, the corresponding bridging capacitance can be found:

For Bessel poles:

$$\theta = 150° \qquad \tan^2\theta = 1/3 \qquad C_b = C/12 \tag{2.4.32}$$

For Butterworth poles:

$$\theta = 135° \qquad \tan^2\theta = 1 \qquad C_b = C/8 \tag{2.4.33}$$

The corresponding mutual inductance is, according to Eq. 2.4.19:

for Bessel poles:

$$L_M = \frac{R^2\,C}{6} \tag{2.4.34}$$

and for Butterworth poles:

$$L_M = \frac{R^2\,C}{8} \tag{2.4.35}$$

The general expression for the coupling factor is [Ref. 2.21, 2.28, 2.33]:

$$k = \frac{L_M}{\sqrt{L_1\,L_2}} = \frac{L_M}{\sqrt{(L_a - L_M)(L_b - L_M)}} \tag{2.4.36}$$

By considering that $L_a = L_b = L/2 = R^2\,C/2$ we obtain:

$$k = \frac{L_M}{\frac{L}{2} - L_M} = \frac{R^2\left(\frac{C}{4} - C_b\right)}{\frac{R^2C}{2} - R^2\left(\frac{C}{4} - C_b\right)} = \frac{\frac{C}{4} - C_b}{\frac{C}{4} + C_b} \tag{2.4.37}$$

If C_b is expressed by Eq. 2.4.31, we may derive a very interesting expression for the coupling factor k:

$$\boxed{k = \frac{3 - \tan^2\theta}{5 + \tan^2\theta}} \tag{2.4.38}$$

Since $\theta = 150°$ for the Bessel pole pair and $135°$ for the Butterworth pole pair, the corresponding coupling factor is:

for Bessel poles:

$$k = 0.5 \tag{2.4.39}$$

for Butterworth poles:

$$k = 0.3\dot{3} \tag{2.4.40}$$

Let us calculate the parameters k, L_M and C_b for two additional cases. If we want to avoid any overshoot, then both poles must be real and equal. In this case $\theta = 180°$ and the damping of the circuit is critical (CD). The expression under the root of Eq. 2.4.29 must be zero and we obtain:

$$C_b = \frac{C}{16} \qquad L_M = \frac{3\,R^2C}{16} \qquad k = 0.6 \tag{2.4.41}$$

We are also interested in the circuit values for the limiting case in which the coupling factor k and consequently the mutual inductance L_M are zero. Here we calculate C_b from Eq. 2.4.31:

$$C_b = \frac{C}{4}\bigg|_{\substack{k=0\\L_M=0}} \tag{2.4.42}$$

The next task is to calculate the poles for all four cases. We will show only the calculation for Bessel poles; the other calculations are equal.

For the starting expression we use the denominator of Eq. 2.4.25 in the canonical form, which we equate to zero:

$$s^2 + s\,\frac{1}{2RC_\mathrm{b}} + \frac{1}{R^2CC_\mathrm{b}} = 0 \tag{2.4.43}$$

Now we insert $C_\mathrm{b} = C/12$, which corresponds to Bessel poles; the result is:

$$s^2 + s\,\frac{6}{RC} + \frac{12}{R^2C^2} = 0 \tag{2.4.44}$$

By factoring out $1/RC$ the roots (poles of Eq. 2.4.25) are:

$$s_{1,2} = \sigma_1 \pm j\,\omega_1 = \frac{1}{RC}\,(\,-3 \pm j\,\sqrt{3}\,) \tag{2.4.45}$$

In a similar way we calculate the Butterworth poles, where $C_\mathrm{b} = C/8$:

$$s_{1,2} = \sigma_1 \pm j\,\omega_1 = \frac{1}{RC}\,(\,-2 \pm j\,2\,) \tag{2.4.46}$$

For critical damping (CD) the imaginary part of the poles is zero, so $C_\mathrm{b} = C/16$, as found before. The poles are:

$$s_{1,2} = \sigma_1 = -\,\frac{4}{RC} \tag{2.4.47}$$

In the no-coupling case ($k = 0$) the bridging capacitance $C_\mathrm{b} = C/4$, and the poles are:

$$s_{1,2} = \sigma_1 \pm j\,\omega_1 = \frac{1}{RC}\,(\,-1 \pm j\,\sqrt{3}\,) \tag{2.4.48}$$

For all four kinds of poles, the input impedance $Z_\mathrm{i} = R = \sqrt{L/C}$ and it is independent of frequency. Now we have all the necessary data to calculate the frequency, phase, time delay and the step response.

2.4.1 Frequency Response

We can use the amplitude- and frequency-normalized Eq. 2.2.27:

$$\left|F(\omega)\right| = \frac{\sigma_{1\mathrm{n}}^2 + \omega_{1\mathrm{n}}^2}{\sqrt{\left[\sigma_{1\mathrm{n}}^2 + \left(\frac{\omega}{\omega_\mathrm{h}} + \omega_{1\mathrm{n}}\right)^2\right]\left[\sigma_{1\mathrm{n}}^2 + \left(\frac{\omega}{\omega_\mathrm{h}} - \omega_{1\mathrm{n}}\right)^2\right]}}$$

By inserting the values for normalized poles, with $RC = 1$ and $\omega_\mathrm{h} = 1/RC$, we can plot the response for each of the four types of poles, as shown in Fig. 2.4.5.

By comparing this diagram with the frequency-response plot for a simple series peaking circuit in Fig. 2.2.3, we realize that the upper cut off frequency ω_H of the T-coil circuit is exactly twice as much as it is for the two-pole series peaking circuit (by comparing, of course, the responses for the same kind of poles). I.e., for Butterworth poles we had $s_{1\mathrm{n},2\mathrm{n}} = 1 \pm j$ (Eq. 2.2.16) for the series peaking circuit, whilst here we have $s_{1\mathrm{n},2\mathrm{n}} = 2 \pm j\,2$. Thus the bandwidth improvement factor for a two pole T-coil circuit, compared with the single pole (RC) circuit is $\eta_\mathrm{b} = 2.83$ (the ratio of the absolute values of poles). Similarly, for other kinds of poles, the bandwidth improvement is greater too, as reported in Table 2.4.1 at the end of this section. Owing to this property, it is worth

considering the use of a T-coil circuit whenever possible. For the same reason we shall discuss T-coil circuits further in detail.

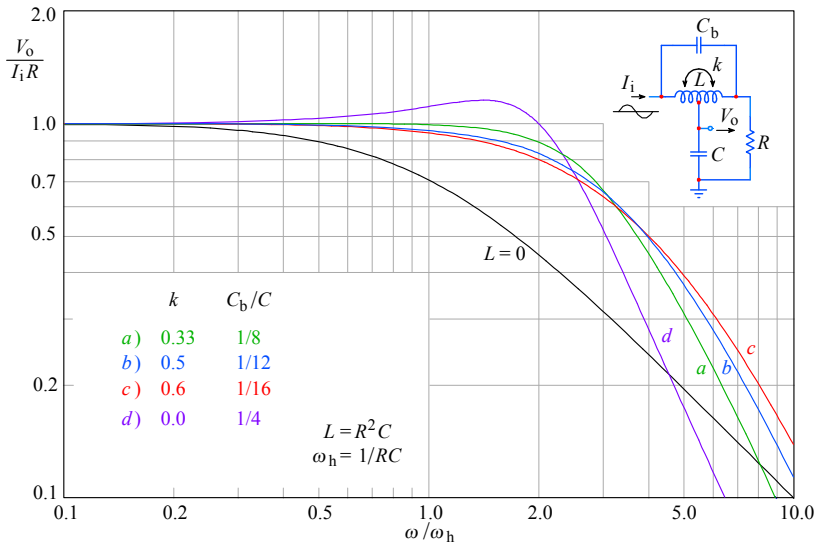


Fig. 2.4.5: The frequency response magnitude of the T-coil, taken from the coil center tap. The curve ***a)*** is the MFA (Butterworth) case, ***b)*** is the MFED (Bessel) case, ***c)*** is the critical damping (CD) case and ***d)*** is the no-coupling ($k = 0$) case. The non-peaking ($L = 0$) case is the reference. The bandwidth extension is notably larger, not only compared with the two-pole series peaking, but also to the three-pole series peaking circuit.

2.4.2 Phase Response

Here we use again the Eq. 2.2.31:

$$\varphi = \arctan \frac{\frac{\omega}{\omega_h} - \omega_{1n}}{\sigma_{1n}} + \arctan \frac{\frac{\omega}{\omega_h} + \omega_{1n}}{\sigma_{1n}}$$

and, by inserting the values for the normalized poles, as we did in the calculation of the frequency response, we obtain the plots shown in Fig. 2.4.6.

2.4.3 Envelope Delay

We use again Eq. 2.2.35:

$$\tau_e\,\omega_h = \frac{\sigma_{1n}}{\sigma_{1n}^2 + \left(\frac{\omega}{\omega_h} - \omega_{1n}\right)^2} + \frac{\sigma_{1n}}{\sigma_{1n}^2 + \left(\frac{\omega}{\omega_h} + \omega_{1n}\right)^2}$$

and, with the pole values as before, we get the Fig. 2.4.7 responses.

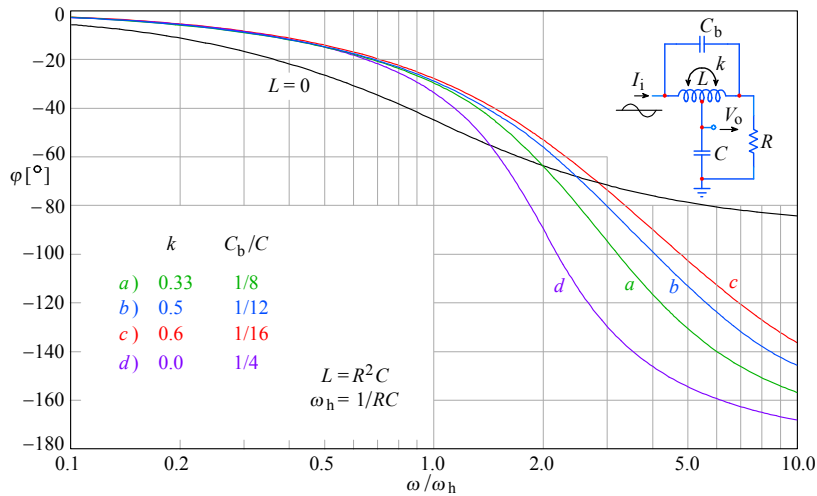


Fig. 2.4.6: The transfer function phase angle of the T-coil circuit, for the same values of coupling and capacitance ratio as for the frequency response magnitude. ***a)*** is MFA, ***b)*** is MFED, ***c)*** is CD and ***d)*** is the no-coupling case. The non-peaking ($L = 0$) case is the reference.

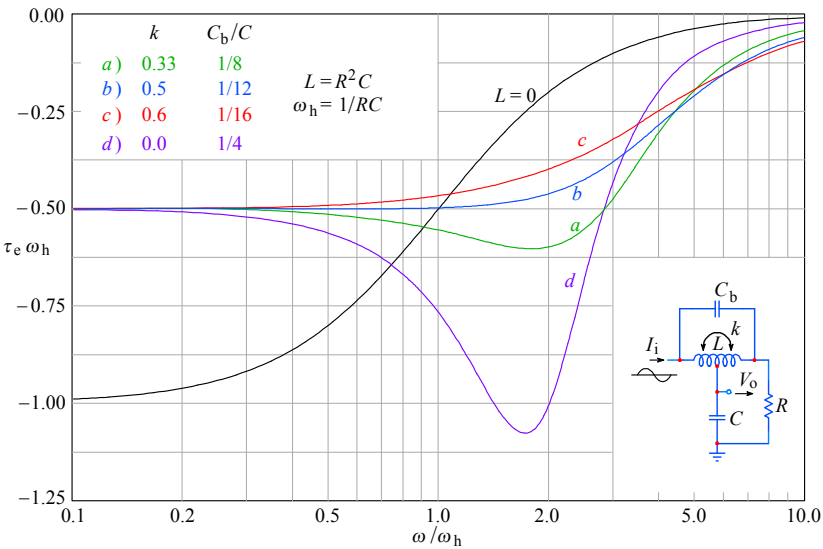


Fig. 2.4.7: The envelope delay of the T-coil: ***a)*** MFA, ***b)*** MFED, ***c)*** CD, ***d)*** $k = 0$. The T-coil circuit delay at low frequencies is exactly one half of that in the $L = 0$ case.

2.4.4 Step Response

We derive the step response from Eq. 2.4.25, as shown in Part 1, Eq. 1.14.29. We take Eq. 2.2.36 for complex poles (MFA, MFED, and the case $k = 0$) and Eq. 2.2.41 for double pole (the CD case). To make the expressions simpler we insert the numerical values of normalized poles and substitute $t/RC = t/T$:

a) for Butterworth poles (MFA), where $k = 0.3\dot{3}$ and $C_\mathrm{b} = C/8$:

$$g_a(t) = 1 + \sqrt{2}\ \mathrm{e}^{-2\,t/T} \sin\left(2\,t/T + 0.7854 + \pi\right) \tag{2.4.49}$$

b) for Bessel poles (MFED), where $k = 0.5$ and $C_\mathrm{b} = C/12$:

$$g_b(t) = 1 + 2\,\mathrm{e}^{-3\,t/T} \sin\left(\sqrt{3}\ t/T + 0.5236 + \pi\right) \tag{2.4.50}$$

c) for Critical Damping (CD), where $k = 0.6$ and $C_\mathrm{b} = C/16$:

$$g_c(t) = 1 - \mathrm{e}^{-4\,t/T}\left(1 + 4\,t/T\right) \tag{2.4.51}$$

d) for $k = 0$ and $C_\mathrm{b} = C/4$:

$$g_d(t) = 1 + \frac{2}{\sqrt{3}}\ \mathrm{e}^{-t/T} \sin\left(\sqrt{3}\ t/T + 1.0472 + \pi\right) \tag{2.4.52}$$

The plots corresponding to these four equations are shown in Fig. 2.4.8. Also shown are the corresponding four pole patterns.

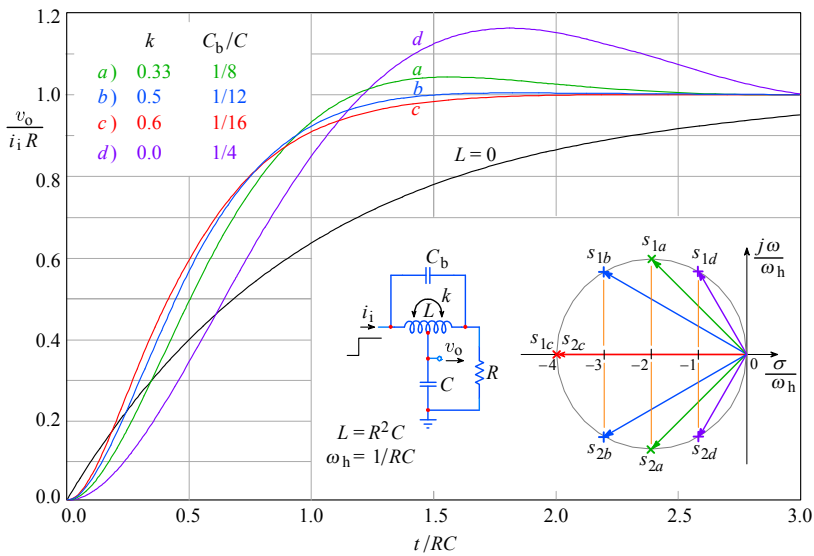


Fig. 2.4.8: The step response of the T-coil circuit. As before, ***a)*** is MFA, ***b)*** is MFED, ***c)*** is CD and ***d)*** is the case $k = 0$. The non-peaking case ($L = 0$) is the reference. The no-coupling case has excessive overshoot, 16.3 %, but MFA overshoot is also high, 4.3 %. Note the pole pattern of the four cases: the closer the poles are to the imaginary axis, the greater is the overshoot. The diameter of the circle on which the poles lie is $4/RC$.

2.4.5 Step Response from Input to v_R

For the application as a delay network, or if we want to design a distributed amplifier with either electronic tubes or FETs we need to know the transmission from the input to the load R, where we have the voltage v_R. For the calculation we use Eq. 2.4.26, which we normalize by making $RC = 1$. Here in addition to the two poles in the left half of the s-plane we also have two symmetrically placed zeros in the right half of the s-plane:

$$s_{1,2} = \sigma_1 \pm j\omega_1 \qquad s_{3,4} = -\sigma_1 \pm j\omega_1 \tag{2.4.53}$$

We shall write Eq. 2.4.26 in the form:

$$F(s) = \frac{s^2R^2CC_b - sRC/2 + 1}{s^2R^2CC_b + sRC/2 + 1} = \frac{(s - s_3)(s - s_4)}{(s - s_1)(s - s_2)} \tag{2.4.54}$$

By multiplication with $1/s$ we obtain the corresponding formula for the step response in the frequency domain:

$$G(s) = \frac{(s - s_3)(s - s_4)}{s\,(s - s_1)(s - s_2)} \tag{2.4.55}$$

and the corresponding time function:

$$g(t) = \mathcal{L}^{-1}\{G(s)\} = \sum \text{res}\,\frac{(s - s_3)(s - s_4)}{s\,(s - s_1)(s - s_2)}\,e^{st} \tag{2.4.56}$$

This operation is performed by contour integration, as explained in Part 1. The integration path must encircle all three poles; however, it is not necessary to encircle the zeros.

Since the function has the poles and zeros arranged symmetrically with respect to both axes we shall express all the components with σ_1 and ω_1, taking care of the polarity of each pole and zero, according to Eq. 2.4.53. We have three residues:

$$\text{res}_0 = \lim_{s\to 0} s\left[\frac{(s - s_3)\,(s - s_4)}{s\,(s - s_1)\,(s - s_2)}\,e^{st}\right] = \frac{s_3\,s_4}{s_1\,s_2} = \frac{\sigma_1^2 + \omega_1^2}{\sigma_1^2 + \omega_1^2} = 1$$

$$\text{res}_1 = \lim_{s\to s_1}(s - s_1)\left[\frac{(s - s_3)\,(s - s_4)}{s\,(s - s_1)\,(s - s_2)}\,e^{st}\right] = \frac{(s_1 - s_3)(s_1 - s_4)}{s_1(s_1 - s_2)}\,e^{s_1 t}$$

$$= \frac{[(\sigma_1 + j\omega_1) - (-\sigma_1 + j\omega_1)][(\sigma_1 + j\omega_1) - (-\sigma_1 - j\omega_1)]}{(\sigma_1 + j\omega_1)[(\sigma_1 + j\omega_1) - (\sigma_1 - j\omega_1)]}\,e^{(\sigma_1 + j\omega_1)t}$$

$$= \frac{2\,\sigma_1}{j\,\omega_1}\,e^{\sigma_1 t}\,e^{j\omega_1 t} \tag{2.4.57}$$

$$\text{res}_2 = \lim_{s\to s_2}(s - s_2)\left[\frac{(s - s_3)\,(s - s_4)}{s\,(s - s_1)\,(s - s_2)}\,e^{st}\right] = \frac{(s_2 - s_3)(s_2 - s_4)}{s_2(s_2 - s_1)}\,e^{s_2 t}$$

$$= \frac{[(\sigma_1 - j\omega_1) - (-\sigma_1 + j\omega_1)][(\sigma_1 - j\omega_1) - (-\sigma_1 - j\omega_1)]}{(\sigma_1 - j\omega_1)[(\sigma_1 - j\omega_1) - (\sigma_1 + j\omega_1)]}\,e^{(\sigma_1 - j\omega_1)t}$$

$$= \frac{2\,\sigma_1}{-\,j\,\omega_1}\,e^{\sigma_1 t}\,e^{-j\omega_1 t}$$

The sum of all three residues is:

$$g(t) = 1 + \frac{2\,\sigma_1}{j\,\omega_1}\,\mathrm{e}^{\sigma_1 t}\,\mathrm{e}^{j\omega_1 t} - \frac{2\,\sigma_1}{j\,\omega_1}\,\mathrm{e}^{\sigma_1 t}\,\mathrm{e}^{-j\omega_1 t}$$

$$= 1 + \frac{4\,\sigma_1}{\omega_1}\,\mathrm{e}^{\sigma_1 t}\left(\frac{\mathrm{e}^{j\omega_1 t} - \mathrm{e}^{-j\omega_1 t}}{2j}\right) = 1 + \frac{4\,\sigma_1}{\omega_1}\,\mathrm{e}^{\sigma_1 t}\sin\omega_1 t \qquad (2.4.58)$$

For critical damping (CD) both zeros and both poles are real. Then, $s_1 = s_2$ and $s_3 = s_4 = -s_1$. There are only two residues, which are calculated in two different ways (because the residue of the double pole must be calculated from the first derivative):

$$\mathrm{res}_0 = \lim_{s\to 0} s\left[\frac{(s-s_3)^2}{s\,(s-s_1)^2}\,\mathrm{e}^{st}\right] = \frac{s_3^2}{s_1^2} = 1 \qquad \text{(because } s_3 = -s_1\text{)}$$

$$\mathrm{res}_1 = \lim_{s\to s_1}\frac{d}{ds}\left[(s-s_1)^2\,\frac{(s-s_3)^2}{s\,(s-s_1)^2}\,\mathrm{e}^{st}\right]$$

$$= \lim_{s\to s_1}\frac{d}{ds}\left[\frac{(s-s_3)^2}{s}\,\mathrm{e}^{st}\right]$$

$$= \lim_{s\to s_1}\left[\mathrm{e}^{st} + s\,t\,\mathrm{e}^{st} - 2\,s_3\,t\,\mathrm{e}^{st} + s_3^2\,\frac{s\,t\,\mathrm{e}^{st} - \mathrm{e}^{st}}{s^2}\right]$$

$$= 4\,s_1 t\,\mathrm{e}^{s_1 t} \qquad \text{(because } s_3 = -s_1\text{)} \qquad (2.4.59)$$

The sum of both residues is the time response sought. We insert the normalized poles and put $t/RC = t/T$ to obtain:

a) for Butterworth poles (MFA), where $k = 0.3\dot{3}$ and $C_\mathrm{b} = C/8$:

$$g_a(t) = 1 - 4\,\mathrm{e}^{-2\,t/T}\sin(2\,t/T) \qquad (2.4.60)$$

b) for Bessel poles (MFED), where $k = 0.5$ and $C_\mathrm{b} = C/12$:

$$g_b(t) = 1 - 4\sqrt{3}\;\mathrm{e}^{-3\,t/T}\sin(\sqrt{3}\;t/T) \qquad (2.4.61)$$

c) for critical damping (CD), where $k = 0.6$ and $C_\mathrm{b} = C/16$:

$$g_c(t) = 1 - 16(t/T)\,\mathrm{e}^{-4\,t/T} \qquad (2.4.62)$$

d) for the case when $k = 0$ and $C_\mathrm{b} = C/4$:

$$g_d(t) = 1 - \frac{4}{\sqrt{3}}\,\mathrm{e}^{-t/T}\sin(\sqrt{3}\;t/T) \qquad (2.4.63)$$

All four plots are shown in Fig. 2.4.9. Note the initial transition owing to the bridge capacitance C_b at high frequencies, the dip where the phase inversion between the high pass and low pass section occurs, and the transition to the final value.

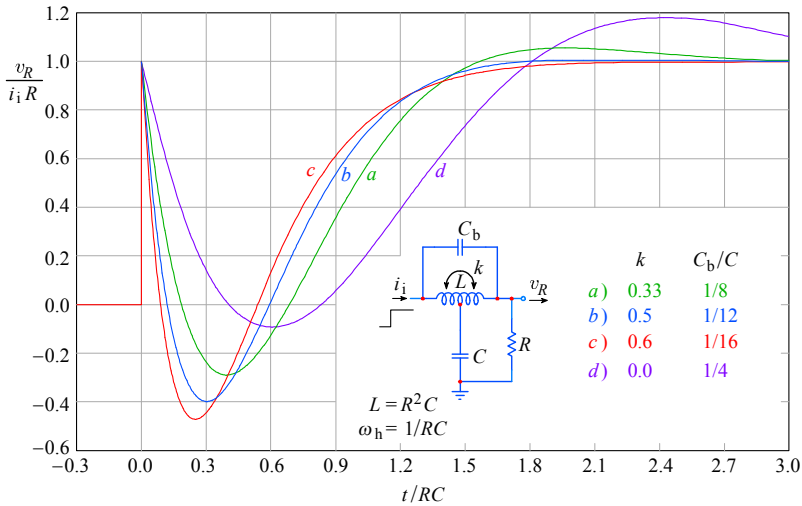


Fig. 2.4.9: The step response of the T-coil circuit, but now with the output from the loading resistor R (this is interesting for cascading stages, as explained later). As before, ***a)*** MFA, ***b)*** MFED, ***c)*** CD, and ***d)*** $k = 0$. The system has the characteristics of an all pass filter.

All the significant data of the T-coil peaking circuit are collected in the Table 2.4.1.

Table 2.4.1

response type	k	C_b/C	η_b	η_r	δ [%]
a) MFA	0.33	1/8	2.82	2.89	4.30
b) MFED	0.50	1/12	2.72	2.79	0.43
c) CD	0.60	1/16	2.57	2.62	0.00
d) $k = 0$	0.00	1/4	2.54	2.68	16.3

Table 2.4.1: Two-pole T-coil circuit parameters

2.4.6 A T-coil application example

One interesting application example of a T-coil all pass network is shown in Fig. 2.4.10. The signal coming out of a TV camera via a 75 Ω cable must be controlled by the monitor and by the vectorscope before it enters the video modulator. The video monitor and the vectorscope should not cause any reflections in the interconnecting cables. Reflections would be caused mostly by the input capacitances of these devices, since their input resistances R_1 , R_2 , $R_3 \gg 75\,\Omega$ and we shall neglect them in our calculations.

To avoid reflections we must connect an impedance matching circuit to each of these inputs and the T-coil circuit can do well, as shown in Fig. 2.4.11. The signal from the TV camera will pass any of three T-coils without reflections. However, the last T-coil in the chain (at the video modulator) must be terminated by the characteristic impedance of the cable, which is 75 Ω. We will take the output for the three devices from their input capacitances C_1, C_2, and C_3. This will cause a slight decrease in bandwidth, but — as we

will see later — the decrease introduced by T-coils will not harm the operation of the total system in any way.

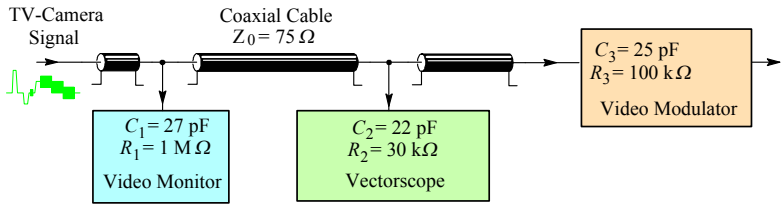


Fig. 2.4.10: An example of a system with different input impedances (TV studio equipment). The signal from a color TV camera is controlled on the monitor screen, the RGB color vectors are measured by the vectorscope, and, finally, the signal is sent to the video modulator for broadcasting. All interconnections are made by a coaxial cable with the characteristic impedance of 75Ω. With long cables, adding considerable delay, the input capacitances can affect the highest frequencies, causing reflections.

On the basis of the data in Fig. 2.4.10 we will calculate the T-coil for each of the three devices. In addition we will calculate the bandwidth at each input. Since the whole system must faithfully transmit pulses, we will consider Bessel poles for all three T-coils.

We use the following four relations:

$$L = R^2 C \qquad \text{(Eq. 2.4.19)}$$

$$C_\mathrm{b} = C/12 \qquad \text{(Eq. 2.4.32)}$$

$$k = 0.5 \qquad \text{(Eq. 2.4.37)}$$

$$\eta_\mathrm{b} = 2.72 \qquad \text{(Table 2.4.1)}$$

$$f_\mathrm{H} = \eta_\mathrm{b} f_\mathrm{h} = \frac{\eta_\mathrm{b}\omega_\mathrm{h}}{2\pi} = \frac{\eta_\mathrm{b}}{2\pi RC} \quad \text{(from Eq. 2.2.29)} .$$

So we calculate:

a) for the monitor,

$$L_1 = 152\,\mathrm{nH}, \quad C_{\mathrm{b}1} = 2.25\,\mathrm{pF}, \quad f_{\mathrm{h}1} = 78.6\,\mathrm{MHz}, \quad f_{\mathrm{H}1} = 231\,\mathrm{MHz};$$

b) for the vectorscope,

$$L_2 = 124\,\mathrm{nH}, \quad C_{\mathrm{b}2} = 1.83\,\mathrm{pF}, \quad f_{\mathrm{h}2} = 96.5\,\mathrm{MHz}, \quad f_{\mathrm{H}2} = 262\,\mathrm{MHz};$$

c) for the video modulator,

$$L_3 = 141\,\mathrm{nH}, \quad C_{\mathrm{b}3} = 2.08\,\mathrm{pF}, \quad f_{\mathrm{h}3} = 84.9\,\mathrm{MHz}, \quad f_{\mathrm{H}3} = 230\,\mathrm{MHz}.$$

The bandwidths are far above the requirement of the system, which is about 6 MHz for either a color or a black and white signal. Fig. 2.4.11 shows the schematic diagram in which the calculated component values are implemented.

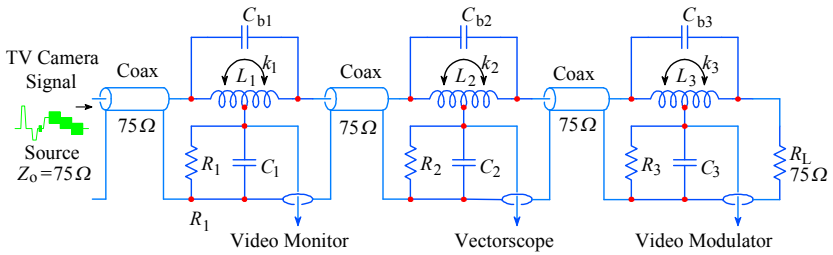


Fig. 2.4.11: Input impedance compensation by T-coil sections prevents signal reflections. Each section of the coaxial cable sees the terminating 75 Ω resistor at the end of the chain. The bandwidth is affected only slightly. The circuit values are shown in the text above.

Since a properly designed T-coil circuit has a constant input impedance, it may be used in connection with a series peaking circuit in order to improve further the system bandwidth, as we shall see in Sec. 2.6. But first we shall examine a 3-pole T-coil system.

2.5 Three-Pole T-coil Peaking Circuit

As in the three-pole series peaking circuit of Fig. 2.3.1, an input capacitance C_i is also always present at the input of the two-pole T-coil circuit, changing it into a three-pole network. This C_i can be a sum of the driving circuit capacitance and the stray input capacitance of the T-coil circuit itself. Actually, as shown in Fig. 2.5.1, the total capacitance of the non-peaking circuit is split by the T-coil into the capacitance of the driving node (C_i) and that of the output stage (C).

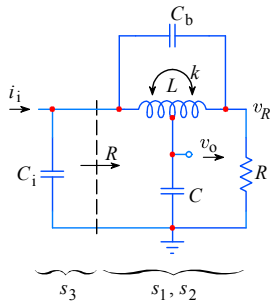


Fig. 2.5.1: The three-pole T-coil network.

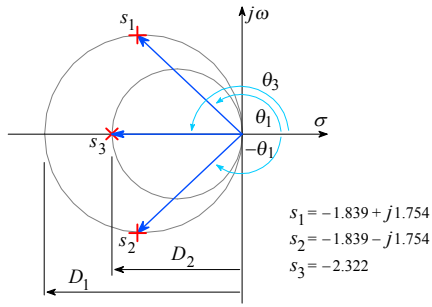


Fig. 2.5.2: Bessel pole layout for the Fig. 2.5.1.

If properly designed, the basic two-pole T-coil circuit will have a constant input impedance R, independent of frequency. This property allows a great simplification of the three-pole network analysis. To both poles of the T-coil, s_1 and s_2, we only need to add the third input pole $s_3 = -1/RC_i$. In order to design an efficient peaking circuit, the tap of the coil must feed a greater capacitance, so $C_i < C$, because the T-coil has no influence on the input pole s_3. Since the network is reciprocal (the current input and voltage output nodes can be exchanged without affecting the response) we can always fulfill this requirement. Also, because of the constant input impedance we can obtain the expression for the transfer function from that of a two-pole T-coil circuit (Eq. 2.4.25) by adding to it the influence of the third pole s_3 (resulting in a simple multiplication of the first-order and second-order transfer function):

$$F(s) = \frac{V_o}{I_1 R} = \frac{1}{\left(s + \frac{1}{RC_i}\right)\left(s^2 R^2 C C_b + s\frac{RC}{2} + 1\right)} \qquad (2.5.1)$$

We shall resist the temptation to perform the suggested multiplication in the denominator, since we would obtain a third-order equation and needlessly complicate the analysis. Owing to the capacitance C_i we have a real pole in addition to the two complex conjugate poles which the T-coil circuit has (in wideband application). With the input capacitance C_i the input impedance is not constant any longer. Its value is equal to R at DC and approaches that of C_i at frequencies beyond ω_H. As before, our task is to select such parameters of the Eq. 2.5.1 that the network will have either Bessel or Butterworth poles. We shall do this by using the trigonometrical relations as indicated in Fig. 2.5.3. The T-coil parameters will carry an index '1' (D_1, θ_1, σ_1, and ω_1).

From the analysis of a two-pole T-coil circuit we remember that the diameter of the circle on which both poles s_1 and s_2 lie is $D_1 = 4/RC$ (see Fig. 2.4.3 or 2.4.8). The diameter D_2 of the circle which goes through the real pole $s_3 = \sigma_2$ is simply $-1/RC_i$ (the reason why we have drawn the circle through this pole also, will become obvious later, when we shall analyze the four-pole L+T circuits). We introduce a new parameter:

$$n = \frac{C}{C_i} \tag{2.5.2}$$

The ratio of the diameters of these circles going through the poles and the origin is then:

$$\frac{D_2}{D_1} = \frac{\frac{1}{RC_i}}{\frac{4}{RC}} = \frac{\frac{n}{RC}}{\frac{4}{RC}} = \frac{n}{4} \tag{2.5.3}$$

From this we obtain:

$$\frac{C}{C_i} = n = 4\,\frac{D_2}{D_1} \tag{2.5.4}$$

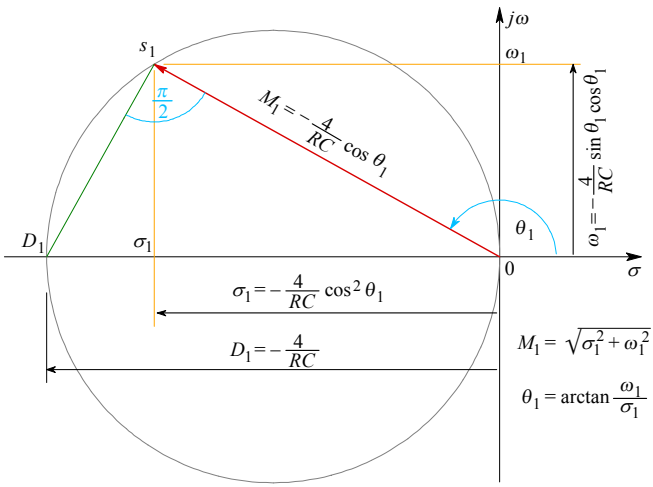


Fig. 2.5.3: The basic trigonometric relations of the main parameters for one of the poles of the T-coil circuit. Knowing one pair of parameters, it is possible to calculate the rest by these simple relations.

Fig. 2.5.3 illustrates some basic trigonometric relations between the polar and Cartesian expression of the poles by taking into account the similarity relationship between the two right angle triangles: $\triangle 0\,\sigma_1\,s_1$ and $\triangle 0\,s_1\,D_1$:

$$\Re\{s_1\} = \sigma_1 = D_1 \cos^2\theta_1 = -\frac{4}{RC}\cos^2\theta_1 \tag{2.5.5}$$

where D_1 is the circle diameter, $D_1 = -4/RC$. Likewise:

$$\Im\{s_1\} = \omega_1 = D_1 \cos\theta_1 \sin\theta_1 = -\frac{4}{RC}\cos\theta_1 \sin\theta_1 \tag{2.5.6}$$

From these equations we can calculate the coupling factor k and the bridging capacitance C_b.

Since:

$$\tan\theta_1 = \frac{\Im\{s_1\}}{\Re\{s_1\}} = \frac{\omega_1}{\sigma_1} \tag{2.5.7}$$

the corresponding coupling factor, according Eq. 2.4.36, is:

$$k = \frac{3 - \tan^2\theta_1}{5 + \tan^2\theta_1}$$

and, as before in Eq. 2.4.31, the bridging capacitance is:

$$C_\mathrm{b} = C\,\frac{1 + \tan^2\theta_1}{16}$$

Next we must calculate the parameter n from the table of poles in Part 4. For Butterworth poles, listed in Table 4.3.1, the values for order $n = 3$ are:

$$\begin{aligned} s_{1t,2t} &= \sigma_{1t} \pm j\,\omega_{1t} = -\frac{1}{2} \pm j\,\frac{\sqrt{3}}{2} \\ s_{3t} &= \sigma_{3t} = -1 \\ \theta_1 &= \pm 120° = 2\pi/3\ \text{[rad]} \end{aligned} \tag{2.5.8}$$

From Eq. 2.5.5 it follows that:

$$D_1 = \frac{\sigma_{1t}}{\cos^2\theta_1} \tag{2.5.9}$$

Since $D_2 = \sigma_{1t}$ the ratio D_2/D_1 is:

$$\frac{D_2}{D_1} = \frac{\sigma_{3t}\cos^2\theta_1}{\sigma_{1t}} = \frac{-1\cdot\cos^2 120°}{-0.5} = 0.5 \tag{2.5.10}$$

Since $n = 4\,D_2/D_1$ we obtain:

$$n = C/C_\mathrm{i} = 2 \qquad \Rightarrow \qquad C_\mathrm{i} = C/2 \tag{2.5.11}$$

Returning to the equations for k and C_b we find $k = 0$ (no coupling !!!) and $C_\mathrm{b} = 0.25\,C$. Just as it was for a two-pole T-coil circuit, here, too, $L = R^2C$. So we have all the circuit parameters for the Butterworth poles.

We can take the values for Bessel poles for order $n = 3$ either from Table 4.4.3 in Part 4, or by running the BESTAP routine (Part 6):

$$\begin{aligned} s_{1t,2t} &= \sigma_{1t} \pm j\,\omega_{1t} = -1.8389 \pm j\,1.7544 \\ s_{3t} &= \sigma_{3t} = -2.3221 \\ \theta_1 &= \pm 136.35° \end{aligned} \tag{2.5.12}$$

In a similar way as before, we obtain:

$$k = 0.3536 \qquad C_\mathrm{b} = 0.12\,C \qquad C_\mathrm{i} = 0.38\,C \tag{2.5.13}$$

To calculate the normalized transfer function we normalize the frequency variable as ω/ω_h, where ω_h is the upper cut off frequency of the non-peaking amplifier ($L = 0$):

$$\omega_\mathrm{h} = \frac{1}{R(C + C_\mathrm{i})} = \frac{1}{RC_\mathrm{c}} \tag{2.5.14}$$

This is important, because if the coil is replaced by a short circuit both capacitances appear in parallel with the loading resistor R. Since $C_\mathrm{i} = C/n$, we may express both capacitances with the total capacitance $C_\mathrm{c} = C + C_\mathrm{i}$ and obtain:

$$C = C_\mathrm{c}\frac{n}{n+1} \qquad \text{and} \qquad C_\mathrm{i} = C_\mathrm{c}\,\frac{1}{n+1} \tag{2.5.15}$$

So far we have used the pole data from tables, since we needed only the ratios of these poles. But, to calculate the frequency, phase, envelope delay, and step response we shall need the actual values of the poles. We have calculated the poles of the T-coil circuit by Eq. 2.4.29, which we repeat here for convenience:

$$s_{1,2} = -\frac{1}{4RC_\mathrm{b}}\left(1 \pm \sqrt{1 - \frac{16\,C_\mathrm{b}}{C}}\right)$$

We shall use the Butterworth poles to explain the procedure. For these poles we have $C_\mathrm{b} = C/4$ and $n = 2$. By inserting these values in the above formula we obtain:

$$s_{1,2} = -\frac{1}{RC}\left(1 \pm j\sqrt{3}\right) \tag{2.5.16}$$

Now let us express the capacitance C by C_c according to Eq. 2.5.15. Then:

$$s_{1,2} = -\frac{1}{RC_\mathrm{c}} \cdot \frac{n+1}{n}\left(1 \pm j\sqrt{3}\right) = -\frac{1}{RC_\mathrm{c}} \cdot \frac{3}{2}\left(1 \pm j\sqrt{3}\right)$$

$$= -\frac{1}{RC_\mathrm{c}}(1.5 \pm j\,2.5981) \tag{2.5.17}$$

The input pole is real:

$$s_3 = -\frac{1}{RC_\mathrm{i}} = -\frac{1}{RC_\mathrm{c}}(n+1) = -\frac{3}{RC_\mathrm{c}} \tag{2.5.18}$$

In a similar way we also calculate the values for Bessel poles and obtain:

$$s_{1,2} = -\frac{1}{RC_\mathrm{c}}(2.8860 \pm j\,2.7532) \qquad \text{and} \qquad s_3 = -\frac{3.6447}{RC_\mathrm{c}} \tag{2.5.19}$$

When calculating the values for the critical damping case (CD) we must consider that the imaginary values of the poles s_1 and s_2 must be zero. This gives $C_\mathrm{b} = C/16$. Here we may choose $n = 2$, and this means that $C_\mathrm{i} = C/2$. The corresponding poles, which are all real, are:

$$s_{1,2} = -\frac{6}{RC_\mathrm{c}} \qquad \text{and} \qquad s_3 = -\frac{3}{RC_\mathrm{c}} \tag{2.5.20}$$

2.5.1 Frequency Response

To calculate the frequency response, we can use Eq. 2.2.27 for a two-pole series peaking circuit and add the effect of the additional input real pole, σ_3. We insert the

normalized poles ($RC_c = 1$) and the normalized frequency ω/ω_h. Thus we obtain the following expression:

$$|F(\omega)| = \frac{(\sigma_{1n}^2 + \omega_{1n}^2)\,\sigma_3}{\sqrt{\left[\sigma_{1n}^2 + \left(\frac{\omega}{\omega_h} + \omega_{1n}\right)^2\right]\left[\sigma_{1n}^2 + \left(\frac{\omega}{\omega_h} - \omega_{1n}\right)^2\right]\left[\sigma_{3n}^2 + \left(\frac{\omega}{\omega_h}\right)^2\right]}} \tag{2.5.21}$$

The plot for all three types of poles is shown in Fig. 2.5.4. By comparing the curve *a*, MFA, where $k = 0$, with the curve *a* in Fig. 2.4.5, where also $k = 0$, we realize that we have achieved a bandwidth extension just by splitting the total circuit capacitance into the input capacitance C_i and the coil loading capacitance C.

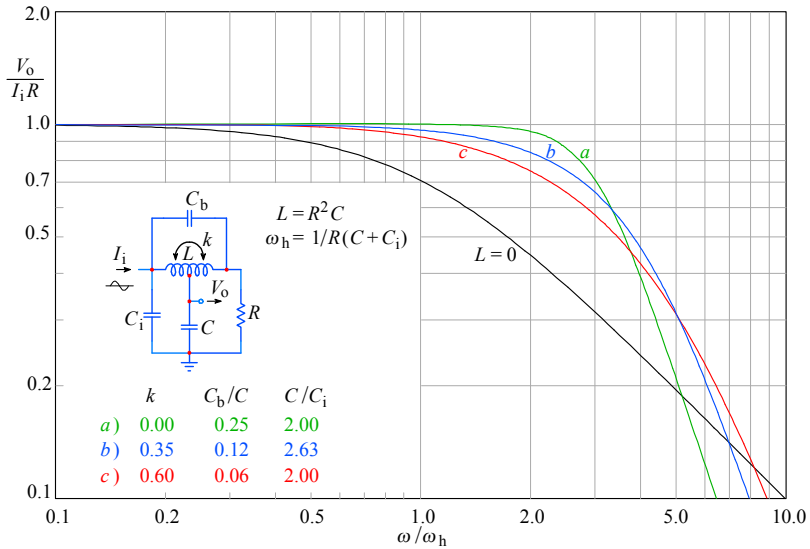


Fig. 2.5.4: Three-pole T-coil network frequency response: ***a)*** MFA; ***b)*** MFED; ***c)*** CD case. The non-peaking response ($L = 0$) is the reference. The MFA bandwidth is larger then that of the two-pole circuit in Fig.2.4.5; in contrast, MFED bandwidth is nearly the same, but the circuit can be realized more easily, owing to the lower magnetic coupling factor required. Note also that, owing to the possibility of separating the total capacitance into a driving and loading part, the reference non-peaking cut off frequency ω_h must be defined as $1/R(C + C_i)$.

2.5.2 Phase Response

For the phase response we can use Eq. 2.3.25 again, and, by inserting the values for the poles we can plot the responses as shown in Fig. 2.5.5:

$$\varphi = \arctan\frac{\frac{\omega}{\omega_h} - \omega_{1n}}{\sigma_{1n}} + \arctan\frac{\frac{\omega}{\omega_h} + \omega_{1n}}{\sigma_{1n}} + \arctan\frac{\frac{\omega}{\omega_h}}{\sigma_{3n}}$$

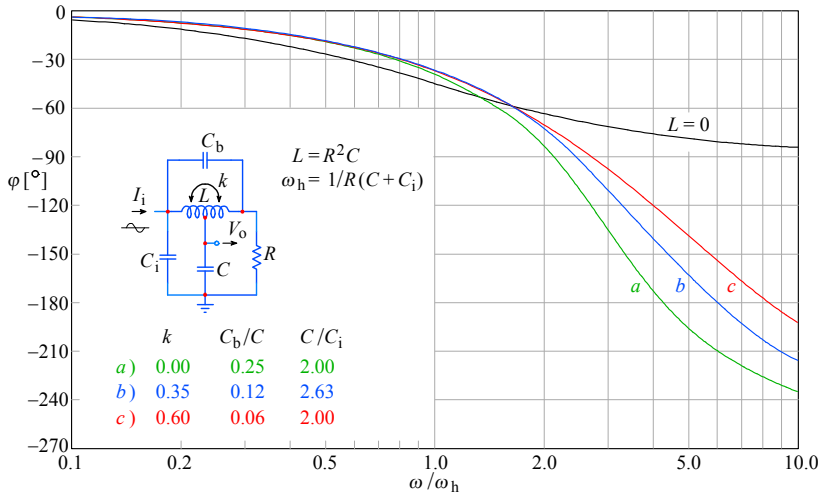


Fig. 2.5.5: Three-pole T-coil network phase response: ***a)*** MFA; ***b)*** MFED; ***c)*** CD case. Note that at high frequencies the 3-pole system phase asymptote is $-270°$ ($3 \times 90°$).

2.5.3 Envelope Delay

We take Eq. 2.3.26 again, and by inserting the values for poles we can plot the envelope delay, as in Fig. 2.5.6:

$$\tau_e\,\omega_h = \frac{\sigma_{1n}}{\sigma_{1n}^2 + \left(\frac{\omega}{\omega_h} - \omega_{1n}\right)^2} + \frac{\sigma_{1n}}{\sigma_{1n}^2 + \left(\frac{\omega}{\omega_h} + \omega_{1n}\right)^2} + \frac{\sigma_{3n}}{\sigma_{3n}^2 + \left(\frac{\omega}{\omega_h}\right)^2}$$

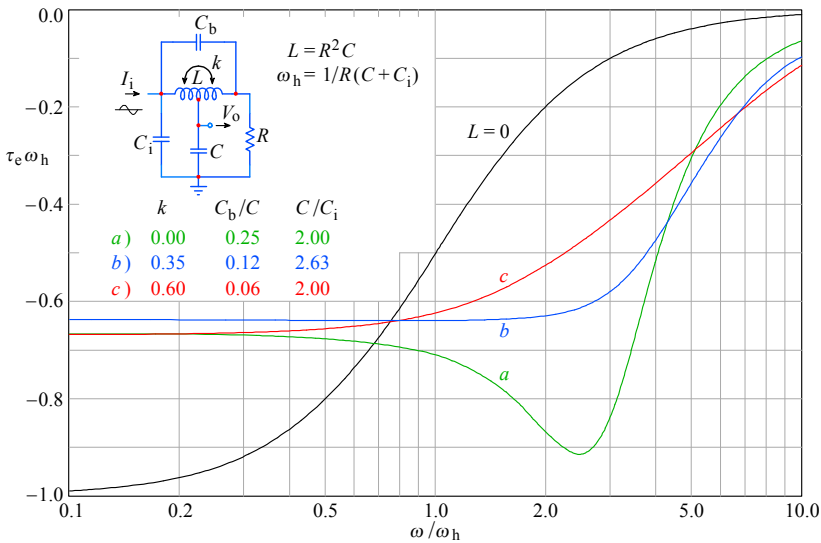


Fig. 2.5.6: Three-pole T-coil network envelope delay: ***a)*** MFA; ***b)*** MFED; ***c)*** CD case. Note that the MFED flatness now extends to nearly $1.5\,\omega_h$.

2.5.4 Step Response

We shall use again Eq. 2.3.27 and Eq. 2.3.28 (see Appendix 2.3, Sec. A2.3.1 for complete derivation). By inserting the values for poles we can calculate the responses and plot them, as shown in Fig. 2.5.7*a* and 2.5.7*b*:

a) from Eq. 2.5.8 for MFA:

$$g_a(t) = 1 + 1.1547\,\mathrm{e}^{-1.5\,t/T}\sin(2.5981\;t/T + 0 + \pi) - \mathrm{e}^{-3\,t/T} \tag{2.5.22}$$

b) from Eq. 2.5.12 for MFED:

$$g_b(t) = 1 + 1.8489\,\mathrm{e}^{-2.886\,t/T}\sin(2.7532\;t/T - 0.5400 + \pi) - 1.9507\,\mathrm{e}^{-3.6447\,t/T} \tag{2.5.23}$$

For the CD case, where we have a double real pole ($s_2 = s_1 = \sigma_1$), the calculation is different (see Eq. 1.11.12 in Part 1). The general expression:

$$F(s) = \frac{-\,s_1^2\,s_3}{(s - s_1)^2(s - s_3)} \tag{2.5.24}$$

(where $s_1 = \sigma_1$ and $s_3 = \sigma_3$) must be multiplied by the unit step operator $1/s$ to obtain the form appropriate for $\mathcal{L}^{-1}$ transform:

$$G(s) = \frac{-\,s_1^2\,s_3}{s\,(s - s_1)^2(s - s_3)} \tag{2.5.25}$$

and the step response is the inverse Laplace transform of $G(s)$, which in turn is equal to the sum of its residues:

$$g(t) = \mathcal{L}^{-1}\{G(s)\} = \sum \mathrm{res}\,\frac{-\,s_1^2\,s_3\,\mathrm{e}^{st}}{s\,(s - s_1)^2\,(s - s_3)} \tag{2.5.26}$$

We have three residues:

$$\mathrm{res}_0 = \lim_{s\to 0} s\left[\frac{-\,s_1^2\,s_3\,\mathrm{e}^{st}}{s\,(s - s_1)^2(s - s_3)}\right] = \frac{-\,s_1^2\,s_3\,\mathrm{e}^{0t}}{s_1^2(-\,s_3)} = 1$$

$$\mathrm{res}_1 = \lim_{s\to s_1}\frac{d}{ds}\left[(s - s_1)^2\,\frac{-\,s_1^2\,s_3\,\mathrm{e}^{st}}{s(s - s_1)^2(s - s_3)}\right] = \lim_{s\to s_1}\frac{d}{ds}\left[\frac{-\,s_1^2\,s_3\,\mathrm{e}^{st}}{s\,(s - s_3)}\right]$$

$$= \lim_{s\to s_1}\left[s_1^2\,s_3\,\frac{s_3(s\,t - 1) - s\,(s\,t - 2)}{s^2\,(s_3 - s)^2}\,\mathrm{e}^{st}\right]$$

$$= s_1^2\,s_3\,\frac{s_3 s_1\,t - s_3 - s_1^2\,t + 2s_1}{s_1^2\,(s_3 - s_1)^2}\,\mathrm{e}^{s_1 t} = s_3\,\frac{s_1(s_1 - s_3)\,t + 2s_1 - s_3}{(s_1 - s_3)^2}\,\mathrm{e}^{s_1 t}$$

$$\mathrm{res}_2 = \lim_{s\to s_3}(s - s_3)\left[\frac{-\,s_1^2\,s_3\,\mathrm{e}^{st}}{s\,(s - s_1)^2(s - s_3)}\right] = \frac{-\,s_1^2}{(s_3 - s_1)^2}\,\mathrm{e}^{s_3 t} \tag{2.5.27}$$

The sum of all three residues is the sought step response:

$$g_c(t) = 1 + s_3 \frac{s_1(s_1 - s_3)\,t + 2s_1 - s_3}{(s_1 - s_3)^2}\,\mathrm{e}^{s_1 t} - \frac{s_1^2}{(s_3 - s_1)^2}\,\mathrm{e}^{s_3 t} \qquad (2.5.28)$$

By inserting $s_1 = \sigma_1$ and $s_3 = \sigma_3$ we obtain:

$$g_c(t) = 1 + \sigma_3 \frac{\sigma_1(\sigma_1 - \sigma_3)\,t + 2\sigma_1 - \sigma_3}{(\sigma_1 - \sigma_3)^2}\,\mathrm{e}^{\sigma_1 t} - \frac{\sigma_1^2}{(\sigma_3 - \sigma_1)^2}\,\mathrm{e}^{\sigma_3 t} \qquad (2.5.29)$$

Finally, we normalize the poles (Eq. 2.5.20), $\sigma_{1n} = -6$ and $\sigma_{3n} = -3$ and normalize the time as t/T, where $T = R(C_i + C)$, to obtain the formula by which the plot *c* in Fig. 2.5.7 is calculated:

$$g_c(t) = 1 + 3\,(1 + 2\,t/T)\,\mathrm{e}^{-6t/T} - 4\,\mathrm{e}^{-3t/T} \qquad (2.5.30)$$

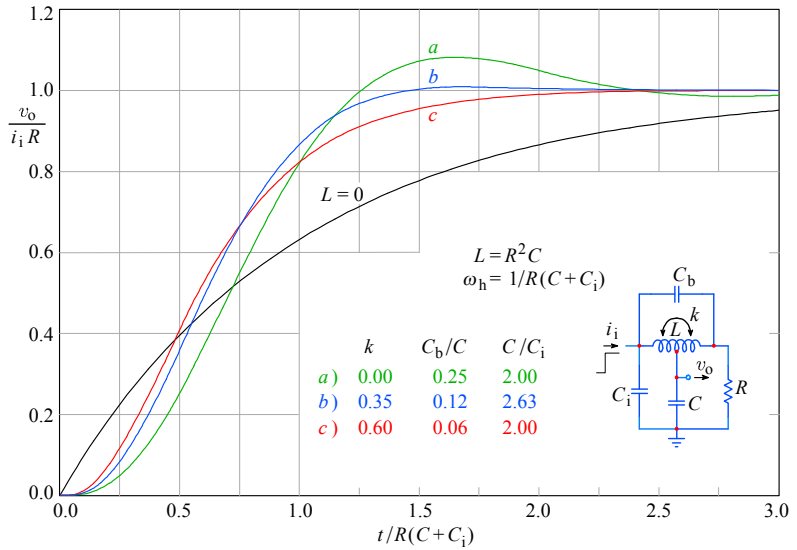


Fig. 2.5.7: The step response of the three-pole T-coil circuit: ***a)*** MFA; ***b)*** MFED; ***c)*** CD. The non-peaking case ($L = 0$) is the reference. Since the total capacitance $C + C_i$ is equal to C of the two-pole T-coil circuit, the MFED rise time is also nearly identical. However, the three-pole circuit is much easier to realize in practice, owing to the lower k required .

2.5.5. Low coupling cases

In the practical realization of a T-coil the toughest problem is to achieve a high coupling factor k. Even $k = 0.5$, as is needed for the two-pole circuit MFED response, is not easy to achieve if we do not want to increase the bridging capacitance C_b excessively. The three-pole T-coil circuits are easy to realize in practice, because of the low coupling required (no coupling for MFA, and only $k = 0.35$ for MFED).

We have also seen that the low coupling factor required is achieved by simply splitting the total circuit capacitance into C and C_i. Therefore it might be useful to further investigate the effects of low coupling. Let us calculate the frequency and the step responses, making three plots each, with a different ratio $C_i/C = n$. In the first group we shall put $k = 0.3\dot{3}$ and in the second group $k = 0.2$.

The corresponding poles are:

Group 1: $k = 0.3\dot{3}$, $C_b = C/8$

a) $n = 2.5$	b) $n = 2$	c) $n = 1.5$
$s_{1n,2n} = -2.8 \pm j2.8$	$s_{1n,2n} = -3 \pm j3$	$s_{1n,2n} = -3.3\dot{3} \pm j3.3\dot{3}$
$s_{3n} = -3.5$	$s_{3n} = -3$	$s_{3n} = -2.5$

The poles are selected so that the sum of $C + C_i$ is the same for all three cases. In this way we have the same upper half power frequency ω_h for any set of poles. This is necessary in order to have the same scale for all three plots. For the above poles we obtain the frequency response as in Fig. 2.5.8 and the step response as in Fig. 2.5.9.

Group 2: $k = 0.2$, $C_b = 0.17\,C$

a) $n = 2$	b) $n = 1.5$	c) $n = 1$
$s_{1n,2n} = -2.5 \pm j2.908$	$s_{1n,2n} = -2.5 \pm j3.227$	$s_{1n,2n} = -3 \pm j3.837$
$s_{3n} = -3$	$s_{3n} = -2.5$	$s_{3n} = -2$

The corresponding frequency response plots are displayed in Fig. 2.5.10 and the step responses in Fig. 2.5.11. From Fig. 2.5.11 it is evident that we have decreased the coupling factor in the second group too much. Nor is a single curve in this figure suitable for the peaking circuit of a pulse amplifier. In curve *a* the overshoot is excessive, whilst the curve *c* exhibits too slow a response. The curve *b* rounds off too soon, reaching the final value with a much slower slope. In a plot with a coarser time scale this curve would clearly show a missing chunk of the step response. Needless to say, it would be very annoying if an oscilloscope amplifier were to have such a step response.

All the important data for the three-pole T-coil peaking circuits are collected in Table 2.5.1. It is worth noting that we achieve a three-pole MFED response with the coupling factor $k = 0.35$ ($\eta_r = 2.78$), whilst for a two-pole T-coil MFED response the $k = 0.5$ was necessary (for a similar $\eta_r = 2.76$). If we are satisfied with a slightly smaller bandwidth it is possible to use the parameters of Group 1, where the coupling factor is $0.3\dot{3}$ only. Such a small coupling factor is much easier to achieve than $k = 0.5$. So for the practical construction of a wideband amplifier we find the three-pole T-coil circuits very convenient.

Table 2.5.1

response type	k	C_b/C	C/C_i	η_b	η_r	δ[%]
MFA	0	0.25	2	2.99	2.89	8.08
MFED	0.35	0.125	2.645	2.75	2.78	0.75
CD	0.60	0.063	2	2.22	2.26	0.00
Group 1, a	0.33	0.125	2.5	2.75	2.77	0.80
Group 1, b	0.33	0.125	2	2.59	2.63	0.00
Group 1, c	0.33	0.125	1.5	2.34	2.39	0.00
Group 2, a	0.2	0.167	2	2.70	2.72	1.85
Group 2, b	0.2	0.167	1.5	2.59	2.62	0.00
Group 2, c	0.2	0.167	1	2.11	2.16	0.00

Table 2.5.1: Three-pole T-coil circuit parameters.

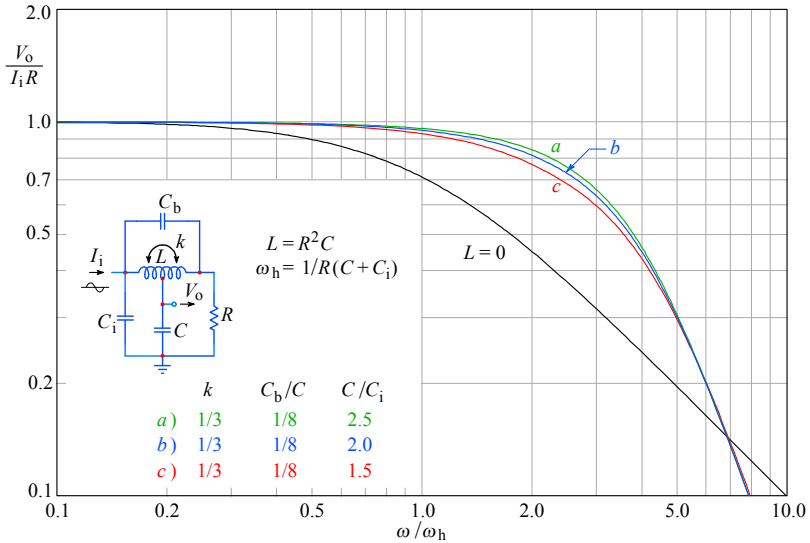


Fig. 2.5.8: Low coupling factor, Group 1: frequency response.

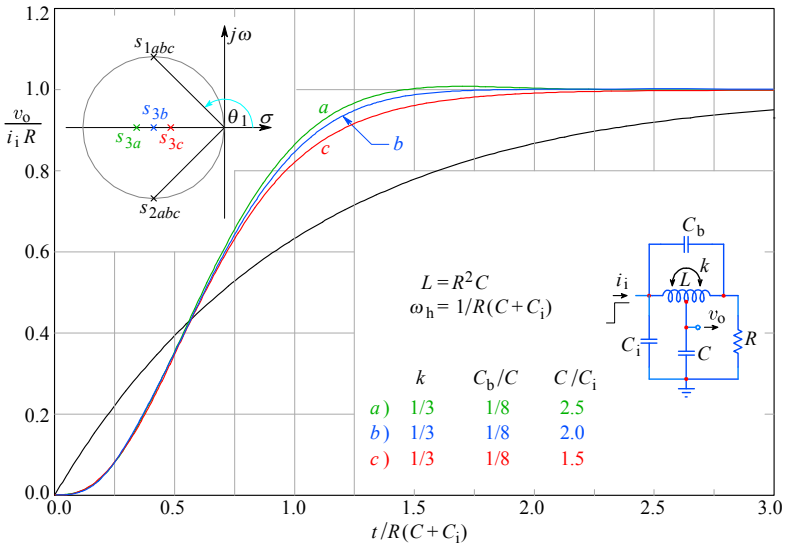


Fig. 2.5.9: Low coupling factor, Group 1: step response. In all three cases the real pole s_3 is placed closer to the origin (becoming dominant) than in the MFA and MFED case, making the responses more similar to the CD case. The characteristic circle of the complex conjugate pole pair has a slightly different diameter in each case.

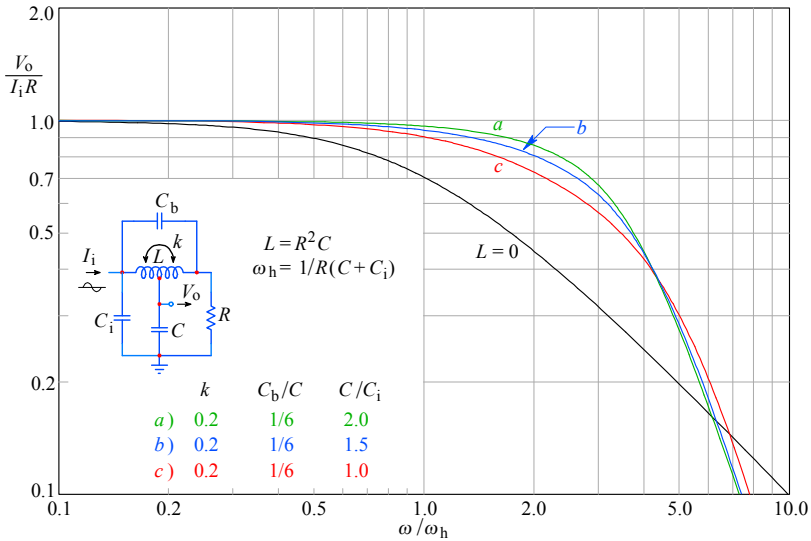


Fig. 2.5.10: Low coupling factor, Group 2: frequency response.

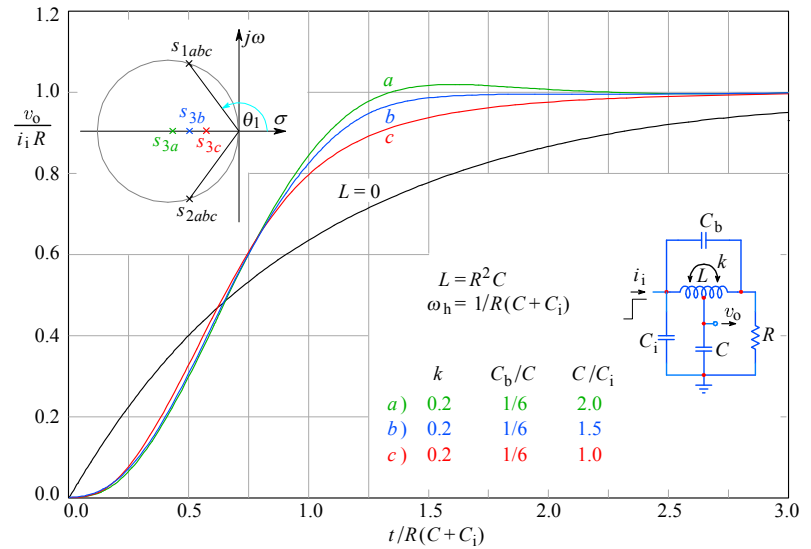


Fig. 2.5.11: Low coupling factor, Group 2: step response. The pole s_{3a} is slightly further away from the real axis than s_{3b} or s_{3c}, therefore causing an overshoot larger than in the MFED case. Both responses *b* and *c* are over-damped, reaching the final value much later than in the MFED case.

2.6 Four-pole L+T Peaking Circuit

Instead of leaving the input capacitance C_i without any peaking coil, as it was in the three-pole T-coil circuit, we can add another coil between C_i and the T-coil. This is the same as adding the series peaking components to the T-coil (it can be done since the input impedance of the T-coil is resistive, if properly designed). In this way, the MFA bandwidth can be extended by over 4 times, compared with the simple RC system. This is substantially more than the 2.75 times found in the two-pole T-coil circuit, where there was only one capacitance. In Fig. 2.6.1 we see such a circuit. The price to pay for such an improvement is a coupling factor $k > 0.5$, which is difficult, but possible, to achieve.

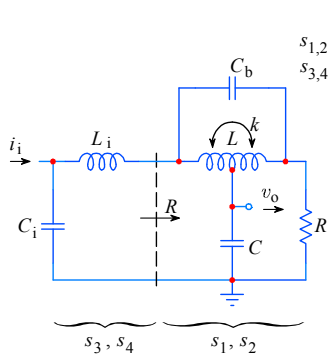


Fig. 2.6.1: The four-pole L+T network.

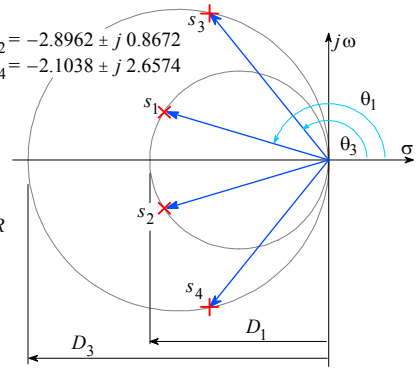


Fig. 2.6.2: The Bessel four-pole pattern of L+T.

Here we utilize the basic property of a T-coil circuit — its constant and real input impedance, presenting the loading resistor to the input series peaking section. Since the input capacitance C_i and the input inductance L_i form a letter 'L' (inverted), we call the network in Fig. 2.6.1 the L+T circuit. This is a four-pole network and **its input impedance is not constant**, but it is similar to the series peaking system, which we have already calculated (Eq. 2.2.44 — 2.2.48, plots in Fig. 2.2.9).

The transfer function of the L+T network is simply the product of the transfer function for a two-pole series peaking circuit (Eq. 2.2.4) and the transfer function for a two-pole T-coil circuit (Eq. 2.4.25). Explicitly:

$$F(s) = \frac{\dfrac{1}{mR^2C_i^2}}{s^2 + \dfrac{s}{mRC_i} + \dfrac{1}{mR^2C_i^2}} \cdot \frac{\dfrac{R}{R^2CC_b}}{s^2 + \dfrac{s}{RC_b} + \dfrac{1}{R^2CC_b}} \tag{2.6.1}$$

Both polynomials in the denominator are written in the canonical form. It would be useless to multiply them, because this would make the analysis very complicated. If we replace R in the later numerator by 1, a normalized equation results.

The T-coil section has two poles, which we can rewrite from Eq. 2.4.28:

$$s_{1,2} = \sigma_1 \pm j\,\omega_1 = -\frac{1}{4\,RC_b} \pm \sqrt{\frac{1}{(4\,RC_b)^2} - \frac{1}{R^2CC_b}} \tag{2.6.2}$$

whilst the input section L has two poles, rewritten from Eq. 2.2.5:

$$s_{3,4} = \sigma_3 \pm j\omega_3 = -\frac{1}{2mRC_\mathrm{i}} \pm \sqrt{\frac{1}{4m^2R^2C_\mathrm{i}^2} - \frac{1}{mR^2C_\mathrm{i}^2}} \qquad (2.6.3)$$

The T-coil circuit extends the bandwidth twice as much as the series peaking circuit, so in order to extend the bandwidth of the L+T-network as much as possible, the T-coil tap must be connected to whichever capacitance is greater. Thus $C_\mathrm{i} < C$. Therefore, the circle with the diameter D_1 and the angle θ_1, corresponding to the T-coil circuit poles $s_{1,2}$, are smaller than the circle with the diameter D_3 and the angle θ_3 corresponding to the poles $s_{3,4}$ of the L branch of the circuit.

Our task is to calculate the circuit parameters for the Bessel pole pattern shown in Fig. 2.6.2, which gives an MFED response. The corresponding values for Bessel poles, shown in Table 4.4.3 in Part 4, order $n = 4$, are:

$$s_{1t,2t} = \sigma_{1t} \pm j\,\omega_{1t} = -2.8962 \pm j\,0.8672 \quad \Rightarrow \quad \theta_1 = 163.33°$$
$$s_{3t,4t} = \sigma_{3t} \pm j\,\omega_{3t} = -2.1038 \pm j\,2.6574 \quad \Rightarrow \quad \theta_3 = 128.37°$$

As indicated in Fig. 2.5.3, the circle diameters are:

$$D_1 = \frac{|s_{1t}|}{\cos\theta_1} = \frac{\sqrt{\sigma_{1t}^2 + \omega_{1t}^2}}{\cos\theta_1} \quad \text{and} \quad D_3 = \frac{|s_{3t}|}{\cos\theta_3} = \frac{\sqrt{\sigma_{3t}^2 + \omega_{3t}^2}}{\cos\theta_3} \qquad (2.6.4)$$

The diameter ratio is:

$$\frac{D_3}{D_1} = \frac{\sqrt{\sigma_{3t}^2 + \omega_{3t}^2}\,\cos\theta_1}{\sqrt{\sigma_{1t}^2 + \omega_{1t}^2}\,\cos\theta_3} = \frac{\sqrt{2.1038^2 + 2.6574^2}}{\sqrt{2.8962^2 + 0.8672^2}} \cdot \frac{\cos 163.33°}{\cos 128.37°} = 1.7304 \qquad (2.6.5)$$

From Fig. 2.2.2 it is evident that the diameter of the circle, on which the poles of the series peaking circuit lie, is $2/RC_\mathrm{i}$. But from Fig. 2.4.3, in the case of a two-pole T-coil circuit, the circle diameter is $4/RC$. Furthermore it is:

$$\frac{C}{C_\mathrm{i}} = n \quad \text{or} \quad C_\mathrm{i} = \frac{C}{n} \qquad (1.6.6)$$

From this we derive:

$$\frac{D_3}{D_1} = \frac{\frac{2}{RC_\mathrm{i}}}{\frac{4}{RC}} = \frac{\frac{2n}{RC}}{\frac{4}{RC}} = \frac{n}{2} \quad \Rightarrow \quad n = 2\,\frac{D_3}{D_1} = 2 \cdot 1.7304 = 3.4608 \qquad (2.6.7)$$

As for the three-pole T-coil analysis, here, too, we express the upper half power frequency of the uncompensated circuit (without coils) as a function of the total capacitance C_c:

$$\omega_\mathrm{h} = \frac{1}{R\,(C + C_\mathrm{i})} = \frac{1}{RC_\mathrm{c}}$$

We can define the capacitors in relation to their sum, like in Eq. 2.5.15:

$$C = C_\mathrm{c}\,\frac{n}{n+1} \qquad \text{and} \qquad C_\mathrm{i} = C_\mathrm{c}\,\frac{1}{n+1}$$

With $n = 3.4608$ form Eq. 2.6.7 we obtain:

$$C = \frac{C_c}{1.2890} \qquad \text{and} \qquad C_i = \frac{C_c}{4.4608} \tag{2.6.8}$$

Now we can calculate all other parameters of the L+T circuit and also the actual values of the poles:

a) Coupling factor (Eq. 2.4.36):

$$k = \frac{3 - \tan^2\theta_1}{5 + \tan^2\theta_1} = \frac{3 - \tan^2 163.33°}{5 + \tan^2 163.33°} = 0.5718$$

b) Bridging capacitance (Eq. 2.4.31):

$$C_b = C\,\frac{1 + \tan^2\theta_1}{16} = C\,\frac{1 + \tan^2 163.33°}{16} = 0.0681\,C$$

c) The parameter m (Eq. 2.2.26):

$$m = \frac{1 + \tan^2\theta_3}{4} = \frac{1 + \tan^2 128.37°}{4} = 0.6488$$

d) Input inductance L_i (Eq. 2.2.14):

$$L_i = mR^2C_i = 0.6488\,R^2C_i$$

e) Real part of the pole s_1, $\Re\{s_1\} = \sigma_1$, (Eq. 2.5.5 and Fig. 2.5.3):

$$\sigma_1 = -\frac{4}{R\,C}\cos^2\theta_1 = -\frac{4 \cdot 1.2890}{R\,C_c}\cos^2 163.33° = -\frac{4.7317}{R\,C_c}$$

f) Imaginary part of the pole s_1, $\Im\{s_1\} = \omega_1$, (Eq. 2.5.6 and Fig. 2.5.3):

$$\omega_1 = -\frac{4}{RC}\cos\theta_1\sin\theta_1 = -\frac{4 \cdot 1.2890}{RC_c}\cos 163.33°\sin 163.33° = \frac{1.4167}{RC_c}$$

g) Real part of the pole s_3, $\Re\{s_3\} = \sigma_3$, (Eq. 2.5.5 and Fig. 2.5.3):

$$\sigma_3 = -\frac{2}{RC_i}\cos^2\theta_3 = -\frac{2 \cdot 4.4608}{RC_c}\cos^2 128.37° = -\frac{3.4376}{RC_c}$$

h) Imaginary part of the pole s_3, $\Im\{s_3\} = \omega_3$, (Eq. 2.5.6 and Fig. 2.5.3):

$$\omega_3 = -\frac{2}{RC_i}\cos\theta_3\sin\theta_3 = -\frac{2 \cdot 4.4608}{RC_c}\cos 128.37°\sin 128.37° = \frac{4.3419}{RC_c}$$

As above, we can calculate the parameters for the MFA response from normalized ($RC_c = 1$) values of the 4th-order Butterworth system (Table 4.3.1 , BUTTAP, Part 6):

$$s_{1n,2n} = \sigma_{1n} \pm j\omega_{1n} = -4.1213 \pm j\,1.7071 \qquad \text{and} \qquad \theta_1 = 157.50°$$
$$s_{3n,4n} = \sigma_{3n} \pm j\omega_{3n} = -1.7071 \pm j\,4.1213 \qquad \text{and} \qquad \theta_3 = 112.50°$$

The L+T network parameters for some other types of poles are given in Table 2.6.1 at the end of this section.

For Butterworth poles (and only for these!) it is very easy to calculate the upper half power frequency ω_H: it is equal to the radius of the circle centered at the origin, on which all four poles lie, which in turn is equal to the absolute value of any one of the four poles. If we use the normalized pole values, the circle radius is also equal to the factor of bandwidth improvement, η_b. By dividing this value by RC_c, we obtain ω_H. We can use any one of the four poles, e.g. s_{1n}:

$$\eta_b = \frac{\omega_H}{\omega_h} = |s_{1n}| = \sqrt{\sigma_{1n}^2 + \omega_{1n}^2} = \sqrt{4.1213^2 + 1.7071^2} = 4.4609 \qquad (2.6.9)$$

and this is a really impressive bandwidth improvement.

2.6.1 Frequency Response

Let us compose a general expression for the frequency response for our L+T circuit. The formula with normalized values is very similar to Eq. 2.5.21, except that here we have two pairs of poles, $\sigma_{1n} \pm j\omega_{1n}$ and $\sigma_{3n} \pm j\omega_{3n}$:

$$|F(\omega)| = \frac{(\sigma_{1n}^2 + \omega_{1n})^2}{\sqrt{\left[\sigma_{1n}^2 + \left(\frac{\omega}{\omega_h} + \omega_{1n}\right)^2\right]\left[\sigma_{1n}^2 + \left(\frac{\omega}{\omega_h} - \omega_{1n}\right)^2\right]}} \cdot \frac{(\sigma_{3n}^2 + \omega_{3n})^2}{\sqrt{\left[\sigma_{3n}^2 + \left(\frac{\omega}{\omega_h} + \omega_{3n}\right)^2\right]\left[\sigma_{3n}^2 + \left(\frac{\omega}{\omega_h} - \omega_{3n}\right)^2\right]}} \qquad (2.6.10)$$

Since we have inserted the normalized poles, the frequency, too, had to be normalized as ω/ω_h. Fig. 2.6.3 shows the frequency response for *a*) MFA and *b*) MFED and also for two other pole placements, reported in [Ref. 2.28], the data of which are:

The curve *c*) corresponds to the poles of Group C of [Ref. 2.28]:

$$s_{1n,2n} = \sigma_{1n} \pm j\omega_{1n} = -3.3252 \pm j0.5863 \quad \text{and} \quad \theta_1 = 170.00°$$
$$s_{3n,4n} = \sigma_{3n} \pm j\omega_{3n} = -1.7071 \pm j4.1213 \quad \text{and} \quad \theta_3 = 112.50°$$

The curve *d*) corresponds to the poles of Group A of [Ref. 2.28]:

$$s_{1n,2n} = \sigma_{1n} \pm j\omega_{1n} = -3.8332 \pm j1.7874 \quad \text{and} \quad \theta_1 = 155.00°$$
$$s_{3n,4n} = \sigma_{3n} \pm j\omega_{3n} = -2.1024 \pm j5.0013 \quad \text{and} \quad \theta_3 = 112.80°$$

Whilst *c* and *d* offer an improvement in η_b and η_r, their step response is far from optimum.

In Fig. 2.6.4, the plot *e*) is the Chebyshev response, $\Delta\varphi = \pm 0.05°$ [Ref. 2.24, 2.30]:

$$s_{1n,2n} = \sigma_{1n} \pm j\omega_{1n} = -3.7912 \pm j1.8656 \quad \text{and} \quad \theta_1 = 153.80°$$
$$s_{3n,4n} = \sigma_{3n} \pm j\omega_{3n} = -2.4861 \pm j4.6755 \quad \text{and} \quad \theta_3 = 118.00°$$

The plot *f*) is the Gaussian frequency response (to -12 dB) [Ref. 2.24, 2.30]:

$$s_{1n,2n} = \sigma_{1n} \pm j\,\omega_{1n} = -3.3835 \pm j\,2.0647 \quad \text{and} \quad \theta_1 = 148.83°$$

$$s_{3n,4n} = \sigma_{3n} \pm j\,\omega_{3n} = -3.4150 \pm j\,6.2556 \quad \text{and} \quad \theta_3 = 118.60°$$

The plot *g*) corresponds to a double pair of Bessel poles, with the following data:

$$s_{1n,2n} = s_{3n,4n} = \sigma_{1n} \pm j\,\omega_{1n} = -4.5000 \pm j\,2.5981 \quad \text{and} \quad \theta_1 = 150.00°$$

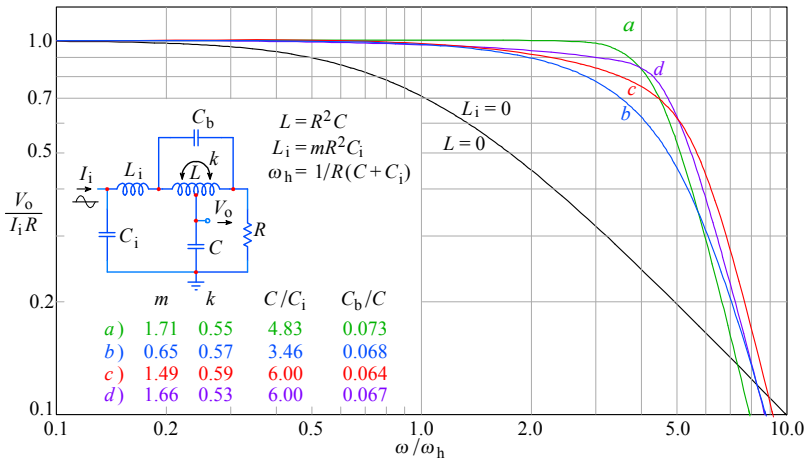


Fig. 2.6.3: Four-pole L+T peaking circuit frequency-response: ***a)*** MFA; ***b)*** MFED; ***c)*** Group C; ***d)*** Group A. In the non-peaking reference case both inductances are zero.

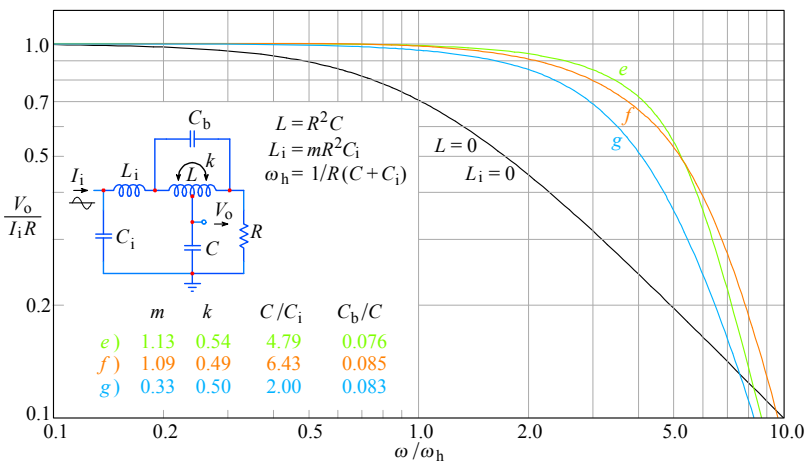


Fig. 2.6.4: Some additional frequency response plots of the four-pole L+T peaking circuit: ***e)*** Chebyshev with 0.05° phase ripple, ***f)*** Gaussian to −12dB, ***g)*** double 2nd-order Bessel. Note the lower bandwidth of *g*) ($2.9\,\omega_h$) compared with *b*) in Fig. 2.6.3 ($3.47\,\omega_h$). This clearly shows that the bandwidth of a cascade of identical stages is lower than if the stages have staggered poles.

Note: The lower bandwidth ($\omega_H = 2.9\,\omega_h$) of the system with repeated poles, plot *g*, compared with the staggered pole placement Fig. 2.6.3, plot *b* ($\omega_H = 3.47\,\omega_h$), clearly shows that using repeated poles is not a clever idea! See also the step response plots.

All these groups of poles can be found in the tables [Ref. 2.30]. Here we can see the extreme adaptability of the calculation method based on the trigonometric relations as shown in Fig. 2.5.2, 2.5.3, 2.6.2, and the corresponding formulae. We call this method *geometrical synthesis*. By this method, the calculation of circuit parameters for the inductive peaking amplifier with any suitable pole placement is very easy and we will use it extensively throughout the rest of the book.

2.6.2 Phase Response

We use Eq. 2.2.30 for each of the four poles:

$$\varphi = \arctan\frac{\dfrac{\omega}{\omega_h} - \omega_{1n}}{\sigma_{1n}} + \arctan\frac{\dfrac{\omega}{\omega_h} + \omega_{1n}}{\sigma_{1n}} + \arctan\frac{\dfrac{\omega}{\omega_h} - \omega_{3n}}{\sigma_{3n}} + \arctan\frac{\dfrac{\omega}{\omega_h} + \omega_{3n}}{\sigma_{3n}} \qquad (2.6.11)$$

The phase response plots for the first four groups of the poles are shown in Fig. 2.6.5. Although the vertical scale ends at $-300°$, the phase asymptote at high frequencies is $-360°$ for all 4th-order responses.

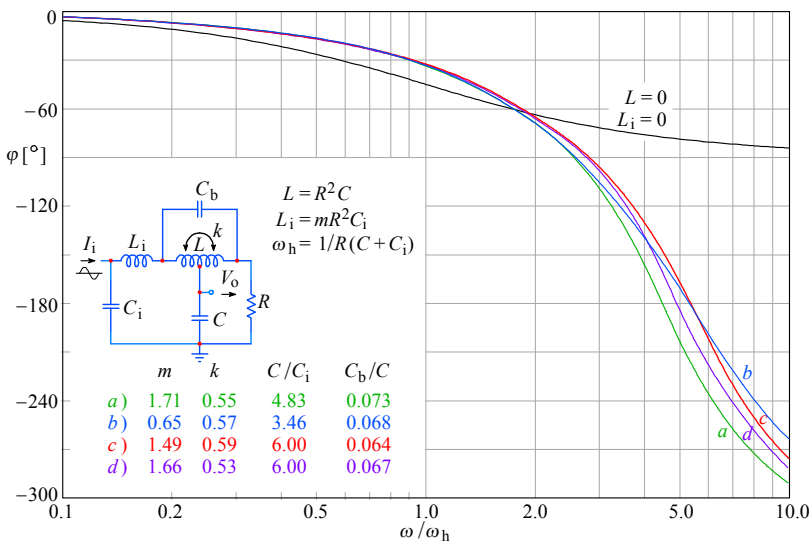


Fig. 2.6.5: Four-pole L+T peaking circuit phase response: ***a)*** MFA; ***b)*** MFED; ***c)*** Group C; ***d)*** Group A. The non-peaking case, in which both inductors are zero, has a 90° maximum phase shift; all other cases, being of 4th-order, have a 360° maximum phase shift.

2.6.3 Envelope Delay

We apply Eq. 2.2.34 for each of the four poles:

$$\tau_e\omega_h = \frac{\sigma_{1n}}{\sigma_{1n}^2 + \left(\frac{\omega}{\omega_h} - \omega_{1n}\right)^2} + \frac{\sigma_{1n}}{\sigma_{1n}^2 + \left(\frac{\omega}{\omega_h} + \omega_{1n}\right)^2} + \frac{\sigma_{3n}}{\sigma_{3n}^2 + \left(\frac{\omega}{\omega_h} - \omega_{3n}\right)^2} + \frac{\sigma_{3n}}{\sigma_{3n}^2 + \left(\frac{\omega}{\omega_h} + \omega_{3n}\right)^2} \tag{2.6.12}$$

The corresponding plots for the first four groups of poles are displayed in Fig. 2.6.6. Note that the value of the delay at low frequency is slightly different for each pole group. This is owed to a different normalization for each circuit.

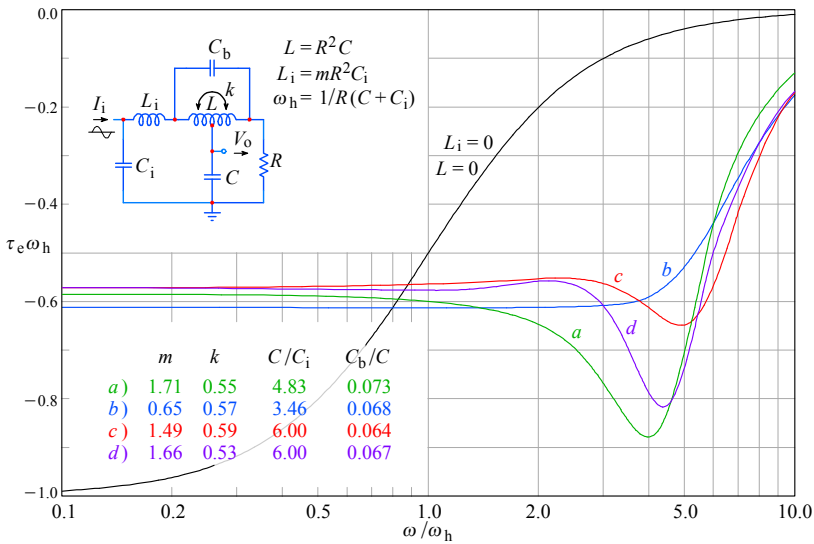


Fig. 2.6.6: Four-pole L+T peaking circuit envelope delay: ***a)*** MFA; ***b)*** MFED; ***c)*** Group C; ***d)*** Group A. For the non-peaking case at DC. the envelope delay is 1; all other cases have a larger bandwidth and consequently a lower delay. Note the MFED flatness up to nearly 3.3 ω_h.

2.6.4 Step Response

A general expression of a four-pole normalized complex frequency response is:

$$F(s) = \frac{s_1\, s_2\, s_3\, s_4}{(s - s_1)(s - s_2)(s - s_3)(s - s_4)} \tag{2.6.13}$$

The $\mathcal{L}$ transform of the step response is obtained by multiplying this function by the unit step operator $1/s$, resulting in a new, five-pole function:

$$G(s) = \frac{s_1\, s_2\, s_3\, s_4}{s\,(s - s_1)(s - s_2)(s - s_3)(s - s_4)} \tag{2.6.14}$$

and to obtain the step response in the time domain, we calculate the $\mathcal{L}^{-1}$ transform:

$$g(t) = \mathcal{L}^{-1}\{G(s)\} = \sum_{i=0}^{4} res_i\,\{G(s)\} \tag{2.6.15}$$

The analytical calculation is a pure routine of algebra, but it would require some 8 pages to present each step. Readers who are interested in the details, can find it in Appendix 2.3 (web only). Here we will write only the result:

$$g(t) = 1 + \frac{K_1}{\omega_1}\,e^{\sigma_1 t}\sqrt{(\sigma_1 A - \omega_1^2 B)^2 + \omega_1^2(A + \sigma_1 B)^2}\;\sin(\omega_1 t + \theta_1)$$
$$+ \frac{K_3}{\omega_3}\,e^{\sigma_3 t}\sqrt{(\sigma_3 C + \omega_3^2 B)^2 + \omega_3^2(C - \sigma_3 B)^2}\;\sin(\omega_3 t + \theta_3) \tag{2.6.16}$$

where:

$$A = (\sigma_1 - \sigma_3)^2 - (\omega_1^2 - \omega_3)^2 \qquad K_1 = \frac{\sigma_3^2 + \omega_3^2}{A^2 + \omega_1^2 B^2}$$

$$B = 2\,(\sigma_1 - \sigma_3)$$

$$C = (\sigma_1 - \sigma_3)^2 + (\omega_1^2 - \omega_3)^2 \qquad K_3 = \frac{\sigma_1^2 + \omega_1^2}{C^2 + \omega_3^2 B^2}$$

$$\theta_1 = \arctan\frac{-\omega_1(A + \sigma_1 B)}{\sigma_1 A - \omega_1^2 B} \qquad \theta_3 = \arctan\frac{-\omega_3(C - \sigma_3 B)}{\sigma_3 C + \omega_3^2 B} \tag{2.6.17}$$

Note: The angles θ_1 and θ_3 calculated by the *arctangent* will not always give a correct result. Depending on the pole pattern, one or both will require an addition of π radians, as we show in Appendix 2.3. In the following relations we give the correct values.

By inserting the normalized values for poles, and the normalized time t/T, where $T = R(C_i + C)$, we obtain the following step response functions:

a) MFA response (Butterworth poles)

$$g(t) = 1 + 2.4142\,e^{-4.1213\,t/T}\sin(1.7071\,t/T + 0.7854 + \pi)$$
$$+ 0.9968\,e^{-1.7071\,t/T}\sin(4.1213\,t/T - 0.7854 + \pi)$$

b) MFED response (Bessel poles)

$$g(t) = 1 + 5.6632\,e^{-4.7317\,t/T}\sin(1.4167\,t/T + 0.4866 + \pi)$$
$$+ 1.6484\,e^{-3.4376\,t/T}\sin(4.3419\,t/T + 1.5389)$$

c) Response for the poles of Group A

$$g(t) = 1 + 2.7233\,e^{-3.8332\,t/T}\sin(1.7874\,t/T + 0.6807 + \pi)$$
$$+ 0.7587\,e^{-2.1024\,t/T}\sin(5.0013\,t/T - 1.2250 + \pi)$$

d) Response for the poles of Group C

$$g(t) = 1 + 5.9875\,e^{-3.3252\,t/T}\sin(0.5863\,t/T + 0.2843 + \pi)$$
$$+ 0.7310\,e^{-1.7284\,t/T}\sin(3.8475\,t/T - 1.1920 + \pi)$$

e) Chebyshev poles with $0.05°$ phase ripple

$$g(t) = 1 + 3.0807\,\mathrm{e}^{-3.7912\,t/T}\sin(1.8565\,t/T + 0.6915 + \pi) + \\ + 0.9744\,\mathrm{e}^{-2.4861\,t/T}\sin(4.6775\,t/T - 1.4289 + \pi)$$

f) Gaussian response to − 12 dB

$$g(t) = 1 + 2.8084\,\mathrm{e}^{-3.3835\,t/T}\sin(2.0467\,t/T + 0.5403 + \pi) + \\ + 0.5098\,\mathrm{e}^{-3.4150\,t/T}\sin(6.2556\,t/T + 1.0598)$$

g) Double 2nd-order Bessel pole pairs

$$g(t) = 1 - \mathrm{e}^{-4.5\,t/T}\left[2\sin(2.5981\,t/T + 0.5236) + \right. \\ \left. + 4\sin(2.5981\,t/T)\cos(0.5236) - 10.3923\,(t/T)\cos(2.5981\,t/T + 0.5236)\right]$$

The last response was calculated by convolution. We will not repeat it here, because it is a very lengthy procedure and it has already been published [Ref. 2.5]. As with any function with repeating poles, the resulting step response is slow compared with the function with the same number of poles but in an optimized pattern. Fig. 2.6.7 shows the step responses of *a*), *b*), *c*) and *d*) ; Fig. 2.6.8 shows the step responses of *e*), *f*) and *g*) .

The data for the four-pole L+T peaking circuit are given in Table 2.6.1.

Table 2.6.1

response type	k	m	n	C_b/C	η_b	η_r	δ [%]
MFA	0.5469	1.7071	4.8283	0.0732	4.46	4.02	10.9
MFED (4th-order Bessel)	0.5718	0.6488	3.4608	0.0681	3.47	3.46	0.90
Group A	0.5333	1.6647	6.0000	0.1667	4.40	4.08	1.90
Group C	0.5901	1.4877	6.0000	0.0644	4.72	4.15	6.20
Chebyshev 0.05°	0.5358	1.1343	4.7902	0.2088	4.09	3.52	3.56
Gaussian to −12 dB	0.4904	1.0888	6.4357	0.1554	3.71	3.43	0.47
Double 2nd-order Bessel	0.5000	0.3333	2.0000	0.0833	2.92	2.96	0.44

Table 2.6.1: Four-pole L+T peaking circuit parameters.

Thus we have concluded the section on four-pole L+T peaking networks. Here we have discussed the geometrical synthesis in a very elementary way, which can be briefly explained as follows:

If the main capacitance C loading the T-coil network tap is known and the loading resistor R is selected upon the required gain, we can — based on the pole data and the geometrical relations of their real and imaginary parts — calculate all the remaining circuit parameters for the complete L+T network.

As we shall see later in the book, the same procedure can be used to calculate the circuit parameters for a multi-stage amplifier by implementing the peaking networks described so far.

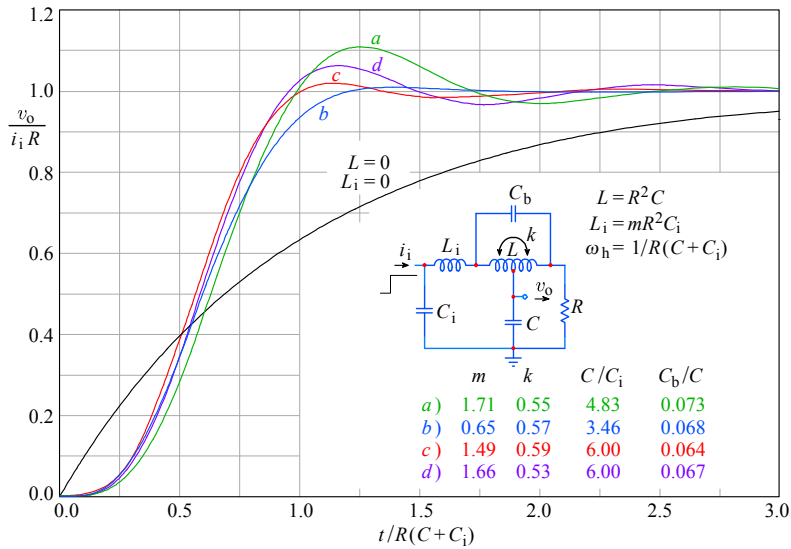


Fig. 2.6.7: Four-pole L+T circuit step response: ***a)*** MFA; ***b)*** MFED; ***c)*** Group C; ***d)*** Group A.

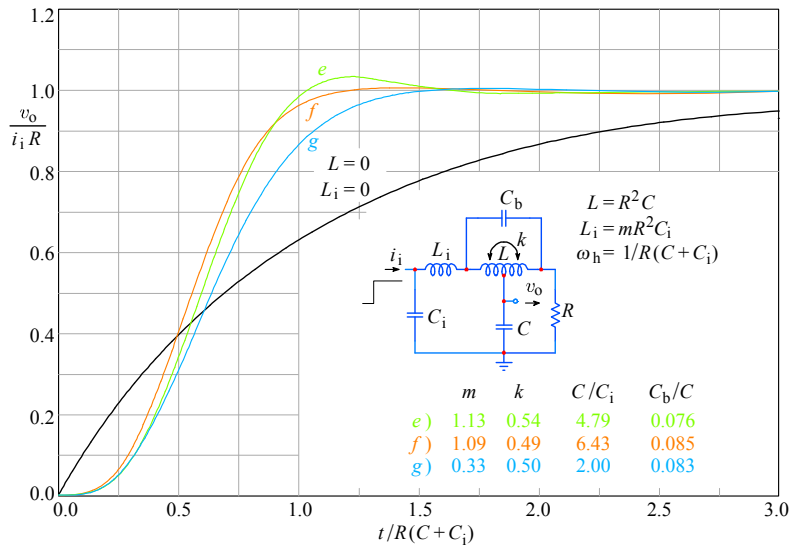


Fig. 2.6.8: Some additional four-pole L+T circuit step responses: ***e)*** Chebyshev 0.05°; ***f)*** Gaussian to − 12dB; ***g)*** double 2nd-order Bessel. Again, the step response confirms that repeating the poles is not optimal; compare the rise times of *g*) and *b*) in Fig. 2.6.7.

2.7 Two-Pole Shunt Peaking Circuit

In some cases, when a single amplifying stage is sufficient, we can use a very simple and efficient shunt peaking circuit, shown in Fig. 2.7.1. This is equivalent to the Fig. 2.1.3, but with the output taken from the capacitor C. Because shunt peaking networks are very simple to make and their bandwidth extension and risetime improvement surpass their series peaking counterparts, they have found very broad application in single stage amplifiers, e.g., in TV receivers.

As the following analysis will show, the two-pole shunt peaking circuit in Fig. 2.7.1 also has one zero. Likewise, the three-pole shunt peaking circuit, which we will discuss in Sec. 2.8, has two zeros. These zeros prevent us from using the geometrical synthesis method for shunt peaking circuits.

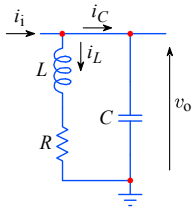


Fig. 2.7.1: A shunt peaking network. It has two poles and one zero.

If we were to compensate the zeros (by intentionally adding a network containing poles coincident with those zeros), we could still use the geometrical synthesis, but that would spoil the optimum performance of the amplifier. Whilst there are no restrictions in using the shunt peaking circuit in a multi-stage amplifier, the total bandwidth and rise time improvement of such amplifier is lower than if the complete amplifier were designed on the basis of the geometrical synthesis. Sometimes a shunt peaking circuit amplifier, designed independently, may be used as an addition to a multi-stage amplifier with series peaking and T-coil peaking circuits in order to shape the starting portion of the amplified pulse to achieve a more symmetrical step response.

The output voltage V_o of the network in Fig. 2.7.1 is:

$$V_o = I_i Z = I_i \frac{\dfrac{R + j\omega L}{j\omega C}}{R + j\omega L + \dfrac{1}{j\omega C}} \tag{2.7.1}$$

This gives the input impedance:

$$Z(\omega) = \frac{R + j\omega L}{1 + j\omega RC - \omega^2 LC} \tag{2.7.2}$$

We introduce the parameters m and ω_h:

$$L = m\,R^2 C \qquad \text{and} \qquad \omega_h = \frac{1}{RC}$$

We insert these parameters into Eq. 2.7.2 to obtain:

$$Z(\omega) = \frac{R + j\omega mR^2C}{1 + \dfrac{j\omega}{\omega_h} - \omega^2 mR^2C^2} = R\,\frac{1 + m\,\dfrac{j\omega}{\omega_h}}{1 + \dfrac{j\omega}{\omega_h} - m\left(\dfrac{\omega}{\omega_h}\right)^2} \tag{2.7.3}$$

2.7.1 Frequency Response

To obtain the normalized frequency response magnitude, we normalize the impedance ($R = 1$), then square the real and imaginary parts in both the numerator and the denominator, and take a square root of the whole fraction:

$$\left|F(\omega)\right| = \sqrt{\frac{1 + m^2\left(\dfrac{\omega}{\omega_h}\right)^2}{1 + (1 - 2m)\left(\dfrac{\omega}{\omega_h}\right)^2 + m^2\left(\dfrac{\omega}{\omega_h}\right)^4}} \tag{2.7.4}$$

We shall first find the value of the parameter m for the MFA response. In this case the factors at $(\omega/\omega_h)^2$ in the numerator and in the denominator must be equal [Ref. 2.4]:

$$1 - 2m = m^2 \qquad \Rightarrow \qquad m = \sqrt{2} - 1 = 0.4141 \tag{2.7.5}$$

If we put this value into Eq. 2.7.4 we obtain:

$$\left|F(\omega)\right| = \sqrt{\frac{1 + 0.1716\left(\dfrac{\omega}{\omega_h}\right)^2}{1 + 0.1716\left(\dfrac{\omega}{\omega_h}\right)^2 + 0.1716\left(\dfrac{\omega}{\omega_h}\right)^4}} \tag{2.7.6}$$

The corresponding plot is shown in Fig. 2.7.2, curve *a*.

For the MFED response, we have to first find the envelope delay.

2.7.2 Phase Response And Envelope Delay

We calculate the value of the parameter m for the MFED response from the envelope delay response, which we derive from the phase angle $\varphi(\omega)$:

$$\varphi(\omega) = \arctan\frac{\Im\{F(\omega)\}}{\Re\{F(\omega)\}}$$

where $F(\omega)$ can be derived from Eq. 2.7.3 by making the denominator real. This is done by multiplying both the numerator and the denominator by the complex conjugate value of the denominator: $1 - m(\omega/\omega_h) - j\,(\omega/\omega_h)$.

The result is:

$$F(\omega)=\frac{\left[1+jm\left(\frac{\omega}{\omega_h}\right)\right]\left[1-m\left(\frac{\omega}{\omega_h}\right)^2-j\left(\frac{\omega}{\omega_h}\right)\right]}{\left[1-m\left(\frac{\omega}{\omega_h}\right)^2\right]^2+\left(\frac{\omega}{\omega_h}\right)^2}=\frac{\mathcal{N}}{\mathcal{D}} \qquad (2.7.7)$$

Next we multiply the brackets in the numerator and separate the real and imaginary parts:

$$\mathcal{N}=1+j\left[(m-1)\left(\frac{\omega}{\omega_h}\right)-m^2\left(\frac{\omega}{\omega_h}\right)^3\right] \qquad (2.7.8.)$$

By dividing the imaginary part of $F(\omega)$ by its real part, $\mathcal{D}$ cancels from the phase:

$$\boxed{\varphi=\arctan\frac{\Im\{\mathcal{N}\}}{\Re\{\mathcal{N}\}}=\arctan\left[(m-1)\left(\frac{\omega}{\omega_h}\right)-m^2\left(\frac{\omega}{\omega_h}\right)^3\right]} \qquad (2.7.9)$$

By inserting $m=0.4141$ (Eq. 2.7.5), we would get the phase response of the MFA case, as plotted in Fig. 2.7.3, curve *a*. But for the MFED response, the correct value of m must be found from the envelope delay, so we must calculate $d\varphi/d\omega$ from Eq. 2.7.9:

$$\tau_e=\frac{d\,\varphi}{d\,\omega}=\frac{d}{d\,\omega}\left\{\arctan\left[(m-1)\left(\frac{\omega}{\omega_h}\right)-m^2\left(\frac{\omega}{\omega_h}\right)^3\right]\right\}$$

$$=\frac{(m-1)-3\,m^2\left(\frac{\omega}{\omega_h}\right)^2}{1+\left[(m-1)\left(\frac{\omega}{\omega_h}\right)-m^2\left(\frac{\omega}{\omega_h}\right)^3\right]^2}\cdot\frac{1}{\omega_h} \qquad (2.7.10)$$

Let us square the bracket in the denominator, factor out $(m-1)$ and multiply both sides of the equation by ω_h in order to obtain the normalized envelope delay:

$$\boxed{\tau_e\,\omega_h=\frac{(m-1)\left[1-\frac{3m^2}{m-1}\left(\frac{\omega}{\omega_h}\right)^2\right]}{1+(m-1)^2\left(\frac{\omega}{\omega_h}\right)^2-2m^2(m-1)\left(\frac{\omega}{\omega_h}\right)^4+m^4\left(\frac{\omega}{\omega_h}\right)^6}} \qquad (2.7.11)$$

A maximally flat envelope-delay is achieved when both factors at $(\omega/\omega_h)^2$ in the numerator and in the denominator are equal [Ref. 2.4]. Taking the sign into account, we have:

$$\frac{3m^2}{1-m}=(m-1)^2 \qquad \Rightarrow \qquad 3m^2=(m^2-2m+1)(1-m) \qquad (2.7.12)$$

Finally:

$$m^3+3\,m-1=0 \qquad \Rightarrow \qquad m=0.3222 \qquad (2.7.13)$$

The only real solution is $m = 0.3222$. At DC $\tau_e\omega_h = m - 1 \approx -0.68$. If we put m into Eq. 2.7.4 for the frequency response, Eq. 2.7.9 for phase response and Eq. 2.7.11 for envelope delay, we can make the plots *b* in Fig. 2.7.2, Fig. 2.7.3, and FIg. 2.7.4.

Now we have enough data to calculate both poles and the zero for the MFA and MFED case, which we shall also need to calculate the step response. But we still have to find the value of m for the critical damping (CD) case. We can derive it from the fact that for CD all the poles are real and equal. To find the poles, we take Eq. 2.7.3, divide it by R, and replace the normalized frequency $j\,\omega/\omega_h$ with the complex variable s:

$$F(s) = \frac{Z(s)}{R} = \frac{1 + m\,s}{1 + s + m\,s^2} = \frac{s + \dfrac{1}{m}}{s^2 + s\,\dfrac{1}{m} + \dfrac{1}{m}} \tag{2.7.14}$$

We obtain the normalized poles from the denominator of $F(s)$:

$$s_{1n,2n} = \sigma_{1n} \pm j\,\omega_{1n} = -\frac{1}{2m} \pm \frac{1}{2m}\sqrt{1 - 4m} = \frac{1}{2m}\left(-1 \pm j\sqrt{4m - 1}\right) \tag{2.7.15}$$

and the normalized real zero from the numerator:

$$s_{3n} = -\sigma_{3n} = -\frac{1}{m} \tag{2.7.16}$$

Since the poles are usually complex, we have written the complex form in the solution of the quadratic equation (Eq. 2.7.15). However, for CD, the solution must be real, so the expression under the square root must be zero and this gives $m = 0.25$. The curves corresponding to CD in Fig. 2.7.2, 2.7.3, and 2.7.4 are marked with the letter *c*.

Note that, in spite of the higher cut off frequency, all the curves have the same high frequency asymptote as the first-order response.

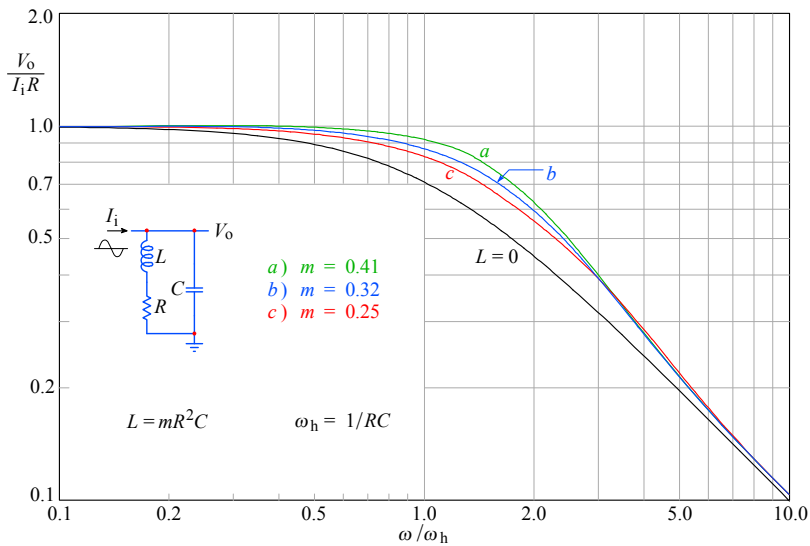


Fig. 2.7.2: Shunt peaking circuit frequency response: ***a)*** MFA; ***b)*** MFED; ***c)*** CD. As usual, the non-peaking case ($L = 0$) is the reference. The system zero causes the high-frequency asymptote to be the same as for the non-peaking system.

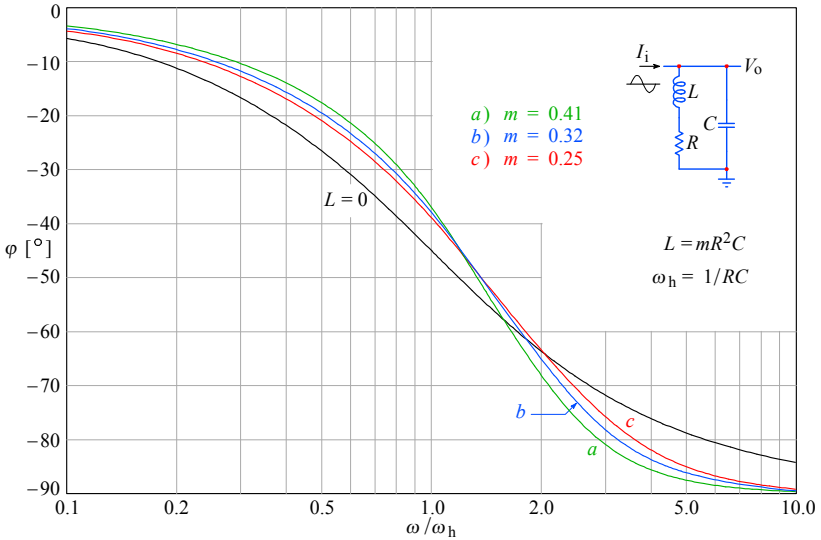


Fig. 2.7.3: Shunt peaking circuit phase response: ***a)*** MFA; ***b)*** MFED; ***c)*** CD. The non-peaking case ($L = 0$) is the reference. The system zero causes the high frequency phase to be $-90°$, the same as for the non-peaking system.

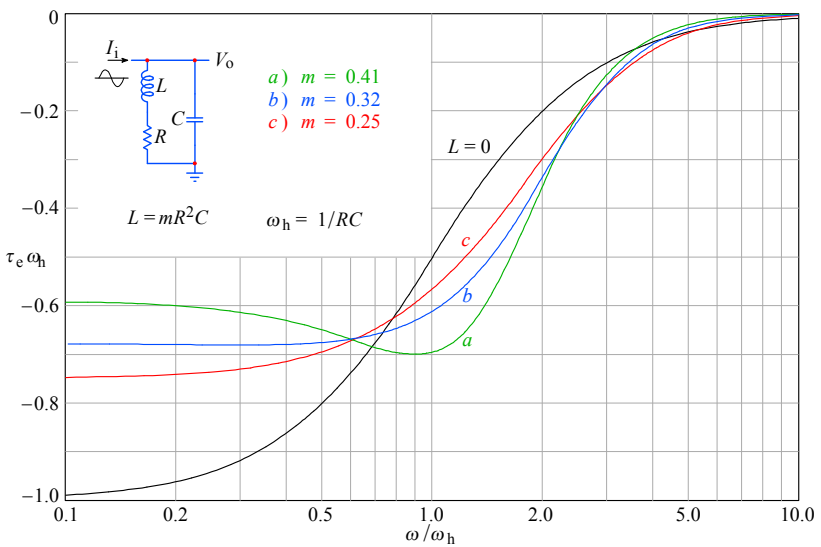


Fig. 2.7.4: Shunt peaking circuit envelope delay: ***a)*** MFA; ***b)*** MFED; ***c)*** CD. The non-peaking case ($L = 0$) is the reference. The peaking systems have a higher bandwidth, and consequently a lower delay at DC than the non-peaking system.

2.7.3 Step Response

For the calculated values of m the poles (Eq. 2.7.15) and the zero (Eq. 2.7.16) are:

a) for MFA response:

the poles $s_{1n,2n} = -1.2071 \pm j\,0.9783$ and the zero $s_{3n} = \sigma_{3n} = -2.4142$

b) for MFED response:

the poles $s_{1n,2n} = -1.5518 \pm j\,0.8340$ and the zero $s_{3n} = \sigma_{3n} = -3.1037$

c) for CD response:

the double pole $s_{1n,2n} = \sigma_{1n} = -2$ and the zero $s_{3n} = \sigma_{3n} = -4$

With these data we can calculate the step response. At first we calculate the MFA and MFED responses, where in both cases we have two complex conjugate poles and one real zero. The general expression for the frequency response is:

$$F(s) = -\frac{s_1 s_2 (s - s_3)}{s_3 (s - s_1)(s - s_2)} \tag{2.7.17}$$

We multiply this equation by the unit step operator $1/s$ and obtain a new function:

$$G(s) = \frac{-s_1 s_2 (s - s_3)}{s\, s_3 (s - s_1)(s - s_2)} \tag{2.7.18}$$

To calculate the step response in the time domain we take the $\mathcal{L}^{-1}$ transform:

$$g(t) = \mathcal{L}^{-1}\{G(t)\} = \sum \text{res}\, \frac{-s_1\, s_2\, (s - s_3)\, \mathrm{e}^{st}}{s\, s_3\, (s - s_1)\, (s - s_2)} \tag{2.7.19}$$

We have three residues:

$$\text{res}_0 = \lim_{s \to 0} s \left[\frac{-s_1\, s_2\, (s - s_3)\, \mathrm{e}^{st}}{s\, s_3\, (s - s_1)(s - s_2)} \right] = \frac{s_1\, s_2\, s_3}{s_1\, s_2\, s_3} = 1$$

$$\text{res}_1 = \lim_{s \to s_1} (s - s_1) \left[\frac{-s_1\, s_2\, (s - s_3)\, \mathrm{e}^{st}}{s\, s_3\, (s - s_1)\, (s - s_2)} \right] = \frac{-s_2\, (s_1 - s_3)}{s_3\, (s_1 - s_2)}\, \mathrm{e}^{s_1 t}$$

$$\text{res}_2 = \lim_{s \to s_2} (s - s_2) \left[\frac{-s_1\, s_2\, (s - s_3)\, \mathrm{e}^{st}}{s\, s_3\, (s - s_1)\, (s - s_2)} \right] = \frac{-s_1 (s_2 - s_3)}{s_3 (s_2 - s_1)}\, \mathrm{e}^{s_2 t} \tag{2.7.20}$$

Since the procedure is the same as for the previous circuits, we shall omit some intermediate expressions. After inserting all the pole components, the sum of residues is:

$$g(t) = 1 - \frac{A + jB}{2jB}\, \mathrm{e}^{\sigma_1 t}\, \mathrm{e}^{j\omega_1 t} - \frac{A - jB}{2jB}\, \mathrm{e}^{\sigma_1 t}\, \mathrm{e}^{-j\omega_1 t} \tag{2.7.21}$$

where $A = \omega_1^2 + \sigma_1(\sigma_1 - \sigma_3)$ and $B = \omega_1 \sigma_3$. After factoring out $-\mathrm{e}^{\sigma_1 t}/B$ we obtain:

$$g(t) = 1 - \frac{\mathrm{e}^{\sigma_1 t}}{B} \left(\frac{A + jB}{2j}\, \mathrm{e}^{j\omega_1 t} + \frac{A - jB}{2j}\, \mathrm{e}^{-j\omega_1 t} \right) \tag{2.7.22}$$

The expression in parentheses can be simplified by sorting the real and imaginary parts:

$$\frac{A+jB}{2j}\,e^{j\omega_1 t}+\frac{A-jB}{2j}\,e^{-j\omega_1 t}=A\,\frac{e^{j\omega_1 t}-e^{-j\omega_1 t}}{2j}+B\,\frac{e^{j\omega_1 t}+e^{-j\omega_1 t}}{2}$$

$$=A\sin\omega_1 t+B\cos\omega_1 t$$

$$=\sqrt{A^2+B^2}\;\sin(\omega_1 t+\beta) \qquad (2.7.23)$$

where:

$$\beta=\arctan\frac{B}{A}$$

Again we have written β in order not to confuse it with the pole angle θ. And here, too, we will have to add π radians to β wherever appropriate, owing to the π period of the *arctangent* function. By entering Eq. 2.7.23 into Eq. 2.7.22 and inserting the poles, we obtain the general expression:

$$\boxed{g(t)=1-\frac{\sqrt{\left[\omega_1^2+\sigma_1(\sigma_1-\sigma_3)\right]^2+\omega_1^2\sigma_3^2}}{\omega_1\sigma_3}\,e^{\sigma_1 t}\sin(\omega_1 t+\beta)} \qquad (2.7.24)$$

where:

$$\boxed{\beta=\arctan\frac{\omega_1\sigma_3}{\omega_1^2+\sigma_1(\sigma_1-\sigma_3)}+\pi} \qquad (2.7.25)$$

We now need the general expression for the step response for the CD case, where we have a double real pole s_1 and a real zero s_3. We start from the normalized frequency response function:

$$F(s)=\frac{-s_1^2\,(s-s_3)}{s_3\,(s-s_1)^2} \qquad (2.7.26)$$

which must be multiplied by the unit step operator $1/s$, thus obtaining:

$$G(s)=\frac{-s_1^2\,(s-s_3)}{s\,s_3\,(s-s_1)^2} \qquad (2.7.27)$$

There are two residues:

$$\mathrm{res}_0=\lim_{s\to 0}\,s\left[\frac{-s_1^2\,(s-s_3)\,e^{st}}{s\,s_3\,(s-s_1)^2}\right]=\frac{s_1^2\,s_3}{s_1^2\,s_3}=1$$

$$\mathrm{res}_1=\lim_{s\to s_1}\frac{d}{ds}\left[(s-s_1)^2\,\frac{-s_1^2\,(s-s_3)\,e^{st}}{s\,s_3\,(s-s_1)^2}\right]=\left[s_1 t\left(1-\frac{s_1}{s_3}\right)-1\right]e^{s_1 t} \qquad (2.7.28)$$

If we express the poles in the second residue with their real and imaginary parts and take the sum of both residues, we obtain:

$$\boxed{g(t)=1+\,e^{\sigma_1 t}\left[\sigma_1 t\left(1-\frac{\sigma_1}{\sigma_3}\right)-1\right]} \qquad (2.7.29)$$

Finally we insert the normalized numerical values for the poles, the zeros and the time variable $t/T=t/RC$. The step response is:

a) for MFA ($\sigma_{1n} = -1.2071$, $\omega_{1n} = \pm 0.9783$, $\sigma_{3n} = -2.4142$)

$$g(t) = 1 + 1.0178\,\mathrm{e}^{-1.2071\,t/T}\sin(0.9783\,t/T - 1.3622 + \pi)$$

b) for MFED ($\sigma_{1n} = -1.5518$, $\omega_{1n} = \pm 0.8340$, $\sigma_{3n} = -3.1037$)

$$g(t) = 1 + 1.1018\,\mathrm{e}^{-1.5518\,t/T}\sin(0.5374\,t/T - 0.9862 + \pi)$$

c) for CD ($\sigma_{1n} = -2$, $\sigma_{3n} = -4$)

$$g(t) = 1 - \mathrm{e}^{-2t/T}(t/T + 1)$$

The step response plots are shown in Fig. 2.7.5. By comparing them with those for the two-pole series peaking circuit in Fig. 2.2.8 we note that the step response derivative at time $t = 0$ is not zero, in contrast to those of the series peaking circuit. Instead the responses look more like the step response of the non-peaking first-order case. The reason for this is in the difference between the number of poles and zeros, which is only 1 in favor of the poles in the shunt peaking circuit.

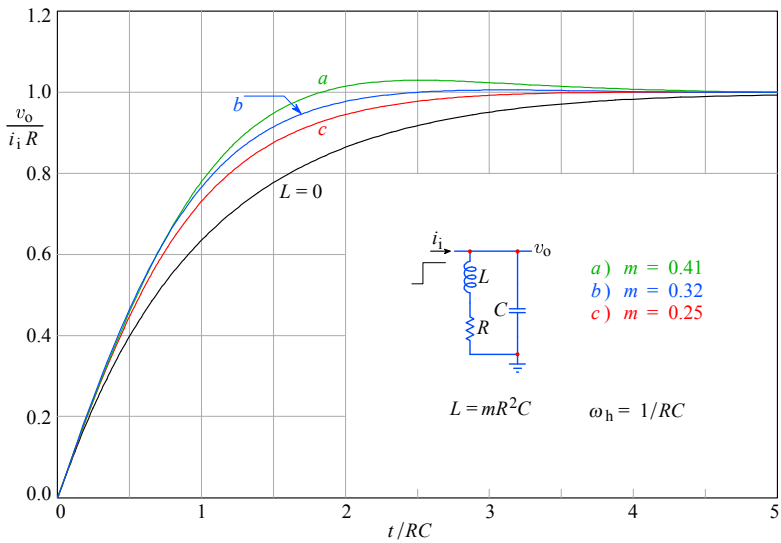


Fig. 2.7.5: Shunt peaking circuit step response: ***a***) MFA; ***b***) MFED; ***c***) CD. The non-peaking case ($L = 0$) is the reference. The difference between the number of poles and the number of zeros is only 1 for the shunt peaking systems, therefore the starting slope of the step response is similar to that of the non-peaking first-order system.

Fig. 2.7.6 shows the pole placements for the three cases. Note the placement of the zero, which is farther from the origin for those systems which have the poles with lower imaginary part.

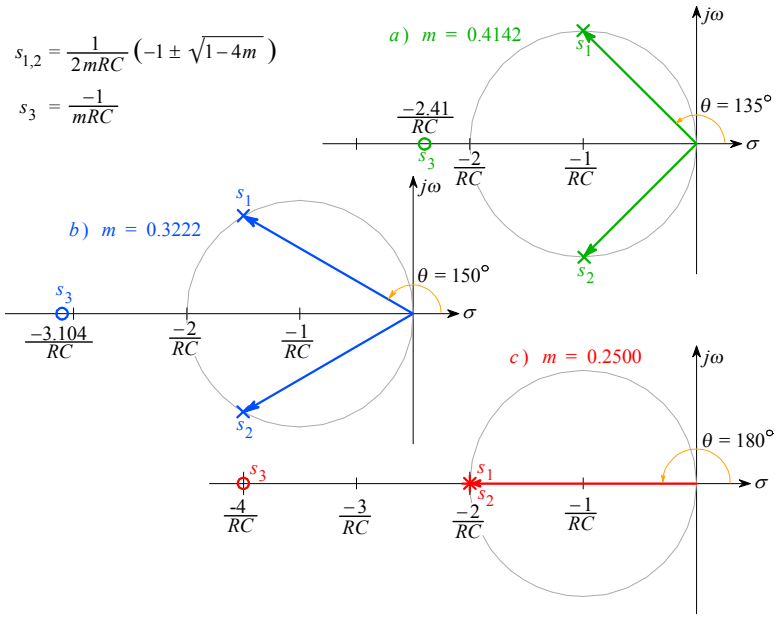


Fig. 2.7.6: Shunt peaking circuit placement of the poles and the zero: ***a)*** MFA; ***b)*** MFED; ***c)*** CD. Note the position of the zero s_3 at the far left of the real axis. Although far from the poles, its influence on the response is notable in each case.

We conclude the discussion with Table 2.7.1, in which all the important two-pole shunt peaking circuit parameters are listed.

Table 2.7.1

response type	m	η_b	η_r	δ [%]
a) MFA	0.4141	1.72	1.81	3.08
b) MFED	0.3222	1.57	1.62	0.41
c) CD	0.2500	1.41	1.43	0.00

Table 2.7.1: Two-pole shunt-peaking circuit parameters

2.8 Three-Pole Shunt Peaking Circuit

If we consider the self capacitance C_L of the coil L the two-pole shunt peaking circuit acquires an additional pole and an additional zero. If the value of C_L can not be neglected it must be in a well defined proportion against other circuit components in order to achieve optimum performance in the MFA or MFED sense. Fig. 2.8.1 shows the corresponding three-pole, two-zero shunt peaking circuit.

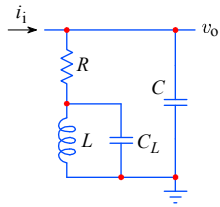


Fig. 2.8.1: The shunt peaking circuit has three poles and two zeros.

The network impedance is:

$$Z(\omega) = \cfrac{1}{j\omega C + \cfrac{1}{R + \cfrac{1}{j\omega C_L + \cfrac{1}{j\omega L}}}}$$

$$= \frac{R + j\omega L - \omega^2 L C_L R}{j\omega C R(1 - \omega^2 L C_L) - \omega^2 L (C + C_L) + 1} \tag{2.8.1}$$

Let us introduce the following parameters:

$$L = mR^2C \qquad \omega_h = \frac{1}{RC} \qquad C_L = n\,C \tag{2.8.2}$$

which we insert into Eq. 2.8.1. Then:

$$Z(\omega) = R\,\frac{1 - mn\left(\frac{\omega}{\omega_h}\right)^2 + jm\left(\frac{\omega}{\omega_h}\right)}{1 - m(1+n)\left(\frac{\omega}{\omega_h}\right)^2 + j\left(\frac{\omega}{\omega_h}\right)\left[1 - mn\left(\frac{\omega}{\omega_h}\right)^2\right]} \tag{2.8.3}$$

2.8.1. Frequency Response

The system transfer function can be obtained easily from $Z(\omega)$. We first replace the normalized imaginary frequency $j(\omega/\omega_h)$ by the complex frequency variable s. Then, by realizing that the output voltage is equal to the product of input current and the system impedance: $V_o = I_i\,Z(\omega)$, we can express the transfer function by normalizing the output to the final value at DC:

$$F(s) = \frac{V_o}{I_i R} = \frac{Z(s)}{R} \tag{2.8.4}$$

With a little rearranging we obtain:

$$F(s) = \frac{s^2 + s\dfrac{1}{n} + \dfrac{1}{m\,n}}{s^3 + s^2\dfrac{1+n}{n} + s\dfrac{1}{m\,n} + \dfrac{1}{m\,n}} \tag{2.8.5}$$

The magnitude $|F(s)| = \sqrt{F(s)\cdot F^*(s)}$ can be obtained more easily from the impedance magnitude. We start from Eq. 2.8.3, square the imaginary and real parts in the numerator and in the denominator and take a square root of the whole fraction:

$$|Z(\omega)| = R\sqrt{\frac{\left[1 - m\,n\left(\dfrac{\omega}{\omega_\mathrm{h}}\right)^2\right]^2 + m^2\left(\dfrac{\omega}{\omega_\mathrm{h}}\right)^2}{\left[1 - m(1+n)\left(\dfrac{\omega}{\omega_\mathrm{h}}\right)^2\right]^2 + \left(\dfrac{\omega}{\omega_\mathrm{h}}\right)^2\left[1 - m\,n\left(\dfrac{\omega}{\omega_\mathrm{h}}\right)^2\right]^2}} \tag{2.8.6}$$

Then we square the brackets and divide by R to obtain a normalized expression:

$$|F(\chi)| = \sqrt{\frac{1 + \left[m^2 - 2\,m\,n\right]\chi^2 + m^2n^2\chi^4}{1 + \left[1 - 2\,m(1+n)\right]\chi^2 + \left[m^2(1+n)^2 - 2\,m\,n\right]\chi^4 + m^2n^2\chi^6}} \tag{2.8.7}$$

and here we have replaced the normalized frequency ω/ω_h with the symbol χ in order to be able to write the equation on a single line.

For the MFA response the numerator and denominator factors at the same powers of $(\omega/\omega_\mathrm{h})$ in Eq. 2.8.7 must be equal [Ref. 2.4]. Thus we have two equations:

$$\begin{aligned} m^2 - 2\,m\,n &= 1 - 2m(1+n) \\ m^2n^2 &= m^2(1+n)^2 - 2\,m\,n \end{aligned} \tag{2.8.8}$$

from which we calculate the values of m and n for the MFA response:

$$m = 0.414 \qquad \text{and} \qquad n = 0.354 \tag{2.8.9}$$

For the MFED response the procedure for calculating the parameters m and n can be similar to that for the two-pole shunt peaking circuit: we would first calculate the formula for the envelope delay and equate the factors at the same powers of $(\omega/\omega_\mathrm{h})$ in the numerator and the denominator, etc. But, with the increasing number of poles, the calculation becomes more complicated. It is much simpler to compare the coefficients of the characteristic polynomial of the complex frequency transfer function Eq. 2.8.5.

The numerical values of the coefficients of the 3rd-order Bessel polynomial, sorted by the falling powers of s, are: 1, 6, 15 and 15 again. Thus, we have two equations:

$$\frac{1+n}{n} = 6 \qquad \text{and} \qquad \frac{1}{m\,n} = 15 \tag{2.8.10}$$

from which we get:

$$n = \frac{1}{5} \qquad \text{and} \qquad m = \frac{1}{3} \tag{2.8.11}$$

Compare these values to those from the work of *V.L. Krejcer* [Ref. 2.4, loc. cit.]. His values for MFED responses are:

$$m = 0.35 \qquad \text{and} \qquad n = 0.22$$

Krejcer also calculated the parameters for a "special" case circuit (SPEC):

$$m = 0.45 \qquad \text{and} \qquad n = 0.22 \tag{2.8.12}$$

By inserting the values of parameters from Eq. 2.8.9 – 2.8.12 into Eq. 2.8.7, we can calculate the corresponding frequency responses. However, for the phase, envelope delay, and step response we also need to know the values of poles and zeros. Since we know all the values of parameters m and n, we can use Eq. 2.8.5 . We equate the denominator $\mathcal{D}$ of $F(s)$ to zero and find the roots, which are the three poles of $F(s)$. Similarly, by equating the numerator $\mathcal{N}$ of $F(s)$ to zero we calculate the two zeros (for readers less experienced in mathematics we have reported the general solutions for polynomials of first, second and third order in Appendix 2.1).

a) MFA response ($m = 0.414$ and $n = 0.354$):

$$\mathcal{D} \Rightarrow \qquad s^3 + 3.825\,s^2 + 6.823\,s + 6.823 = 0$$

The poles: $\quad s_{1n,2n} = \sigma_{1n} \pm j\,\omega_{1n} = -0.850 \pm j\,1.577$

$$s_{3n} = \sigma_{3n} = -2.125$$

$$\mathcal{N} \Rightarrow \qquad s^2 + 2.825\,s + 6.823 = 0$$

The zeros: $\quad s_{4n,5n} = \sigma_{5n} \pm j\,\omega_{5n} = -1.412 \pm j\,2.197$

b) MFED response ($m = 0.333$ and $n = 0.200$):

$$\mathcal{D} \Rightarrow \qquad s^3 + 6\,s^2 + 15\,s + 15 = 0$$

The poles: $\quad s_{1n,2n} = \sigma_{1n} \pm j\,\omega_{1n} = -1.8389 \pm j\,1.7544$

$$s_{3n} = \sigma_{3n} = -2.3222$$

$$\mathcal{N} \Rightarrow \qquad s^2 + 5\,s + 15 = 0$$

The zeros: $\quad s_{4n,5n} = \sigma_{5n} \pm j\,\omega_{5n} = -2.500 \pm j\,2.958$

c) Special case ($m = 0.45$ and $n = 0.22$):

$$\mathcal{D} \Rightarrow \qquad s^3 + 5.545\,s^2 + 10.101\,s + 10.101 = 0$$

The poles: $\quad s_{1n,2n} = \sigma_{1n} \pm j\,\omega_{1n} = -1.035 \pm j\,1.355$

$$s_{3n} = \sigma_{3n} = -3.475$$

$$\mathcal{N} \Rightarrow \qquad s^2 + 4,545\,s + 10.101 = 0$$

The zeros: $\quad s_{4n,5n} = \sigma_{5n} \pm j\,\omega_{5n} = -2.237 \pm j\,2.222$

By inserting the values of m and n in Eq. 2.8.7 we can calculate the frequency response magnitude of the three cases. The resulting plots are shown in Fig. 2.8.2. Note the high frequency asymptote, which is the same as for the non-peaking single-pole case.

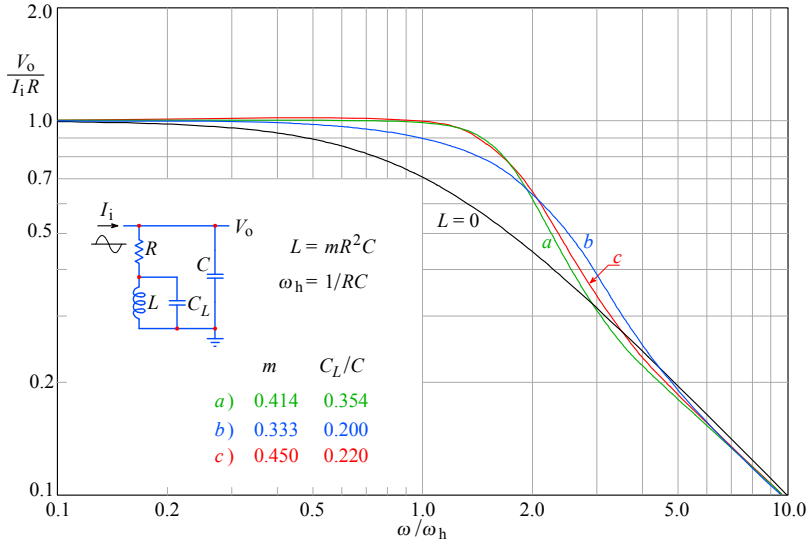


Fig. 2.8.2: Three-pole shunt peaking circuit frequency response: ***a)*** MFA; ***b)*** MFED; ***c)*** SPEC. The non-peaking case ($L = 0$) is the reference. The difference of the number of poles and the number of zeros is only 1 for peaking systems, therefore the ending slope of the frequency response is similar to that of the non-peaking system.

2.8.2 Phase Response

We use Eq. 2.2.30 positive for each pole and negative for each zero and sum them:

$$\varphi = \arctan\frac{\frac{\omega}{\omega_h} - \omega_{1n}}{\sigma_{1n}} + \arctan\frac{\frac{\omega}{\omega_h} + \omega_{1n}}{\sigma_{1n}} + \arctan\frac{\frac{\omega}{\omega_h}}{\sigma_{3n}} + \\ - \arctan\frac{\frac{\omega}{\omega_h} - \omega_{5n}}{\sigma_{5n}} - \arctan\frac{\frac{\omega}{\omega_h} + \omega_{5n}}{\sigma_{5n}} \tag{2.8.13}$$

By entering the numerical values of poles and zeros we obtain the phase response equations for each case. In Fig. 2.8.3 the corresponding plots are shown.

2.8.3 Envelope Delay

We use Eq. 2.2.34, adding a term for each pole and subtracting for each zero:

$$\tau_e\omega_h = \frac{\sigma_{1n}}{\sigma_{1n}^2 + \left(\frac{\omega}{\omega_h} + \omega_{1n}\right)^2} + \frac{\sigma_{1n}}{\sigma_{1n}^2 + \left(\frac{\omega}{\omega_h} - \omega_{1n}\right)^2} + \frac{\sigma_{3n}}{\sigma_{3n}^2 + \left(\frac{\omega}{\omega_h}\right)^2} \\ - \frac{\sigma_{5n}}{\sigma_{5n}^2 + \left(\frac{\omega}{\omega_h} + \omega_{5n}\right)^2} - \frac{\sigma_{5n}}{\sigma_{5n}^2 + \left(\frac{\omega}{\omega_h} - \omega_{5n}\right)^2} \tag{2.8.14}$$

Again we insert the numerical values for poles and zeros in Eq. 2.8.14 to plot the envelope delay as shown in Fig. 2.8.4. As we have explained in Fig. 2.2.6, there is an envelope advance (owed to system zeros) in the high frequency range.

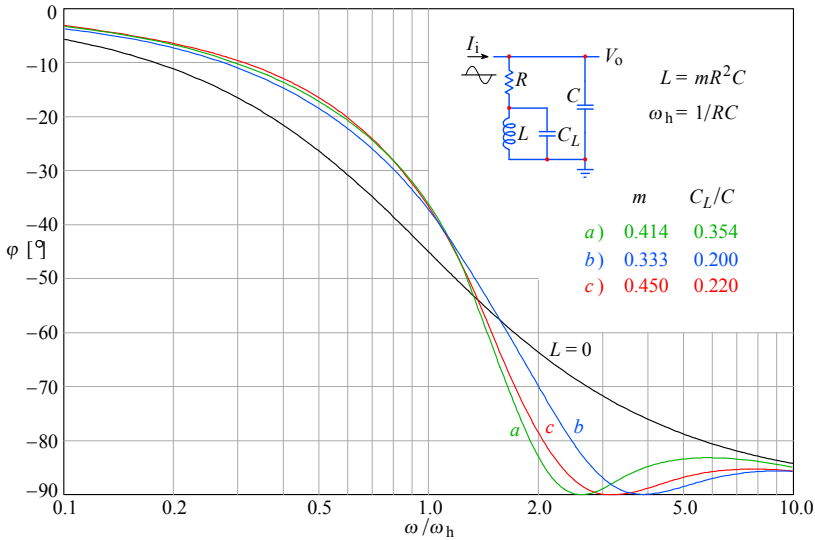


Fig. 2.8.3: Three-pole shunt peaking circuit phase response: ***a)*** MFA; ***b)*** MFED; ***c)*** SPEC. The non-peaking case ($L = 0$) is the reference. The system zeros cause the phase response bouncing up at the $-90°$ boundary and then returning back.

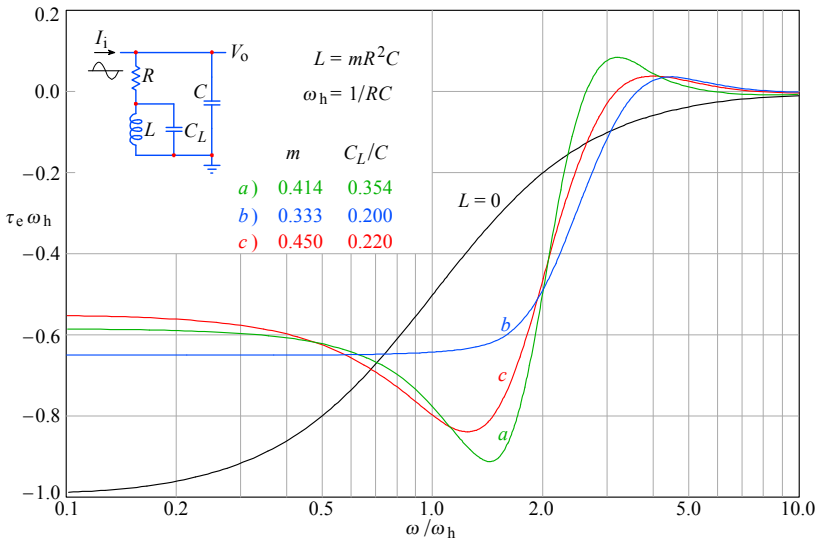


Fig. 2.8.4: Three-pole shunt peaking circuit envelope delay: ***a)*** MFA; ***b)*** MFED; ***c)*** SPEC. The non-peaking ($L = 0$) case is shown as the reference. Note the envelope advance in the high frequency range, owed to system zeros.

2.8.4 Step Response

For the step response we use the general transfer function for three poles and two zeros, which we shall reshape to suit a solution in the Laplace Transform Tables.

$$F(s) = \frac{-\,s_1 s_2\, s_3\,(s - s_4)(s - s_5)}{s_4\, s_5\,(s - s_1)(s - s_2)(s - s_3)} \tag{2.8.15}$$

We multiply this function by the unit step operator $1/s$ and obtain a new equation:

$$G(s) = \frac{1}{s} \cdot \frac{-\,s_1 s_2\, s_3\,(s - s_4)(s - s_5)}{s_4\, s_5\,(s - s_1)(s - s_2)(s - s_3)} \tag{2.8.16}$$

The step response $g(t)$ is fully derived in Appendix 2.3 (web only). The result is:

$$\boxed{g(t) = 1 - \frac{K_1}{\omega_1}\, \mathrm{e}^{\sigma_1 t/T} \sqrt{A^2 + \omega_1^2 B^2}\ \sin(\omega_1 t/T + \beta) - K_3\, \mathrm{e}^{\sigma_3 t/T}} \tag{2.8.17}$$

Besides the usual time normalization $T = RC$, here we have:

$$A = [\sigma_1(\sigma_1 - \sigma_3) - \omega_1^2][\sigma_1^2 - \omega_1^2 + \sigma_5^2 + \omega_5^2 - 2\,\sigma_5\,\sigma_1] - 2\,\sigma_5\,\omega_1^2\,(2\,\sigma_1 - \sigma_3)$$

$$B = (2\,\sigma_1 - \sigma_3)[\sigma_1^2 - \omega_1^2 + \sigma_5^2 + \omega_5^2 - 2\,\sigma_5\sigma_1] + 2\,\sigma_5[\sigma_1(\sigma_1 - \sigma_3) - \omega_1^2]$$

$$\beta = \arctan\left(\frac{-\,\omega_1 B}{A}\right) + \pi$$

$$K_1 = \frac{\sigma_3}{(\sigma_5^2 + \omega_5^2)[(\sigma_1 - \sigma_3)^2\ + \omega_1^2]}$$

$$K_3 = \frac{(\sigma_1^2 + \omega_1^2)[(\sigma_3 - \sigma_5)^2 + \omega_5^2]}{(\sigma_5^2 + \omega_5^2)[(\sigma_3 - \sigma_1)^2 + \omega_1^2]} \tag{2.8.18}$$

By inserting the numerical values of the poles and zeros we obtain the following relations:

a) MFA response ($m = 0.414$ and $n = 0.354$):

$$g(t) = 1 + 0.5573\,\mathrm{e}^{-0.850\,t/T}\,\sin(1.577\,t/T + 0.7741 + \pi) - 0.6104\,\mathrm{e}^{-2.125\,t/T}$$

b) MFED response ($m = 0.333$ and $n = 0.200$):

$$g(t) = 1 + 0.8054\,\mathrm{e}^{-1.839\,t/T}\,\sin(1.754\,t/T - 0.1772 + \pi) - 1.1420\,\mathrm{e}^{-2.322\,t/T}$$

c) Special case ($m = 0.45$ and $n = 0.22$):

$$g(t) = 1 + 0.8814\,\mathrm{e}^{-1.035\,t/T}\,\sin(1.355\,t/T + 1.0333 + \pi) - 0.2429\,\mathrm{e}^{-3.475\,t/T}$$

The plots of these responses can be seen in Fig. 2.8.5. Because the difference between the number of poles and zeros is one only, the initial slope of the response is the same as for the non-peaking response.

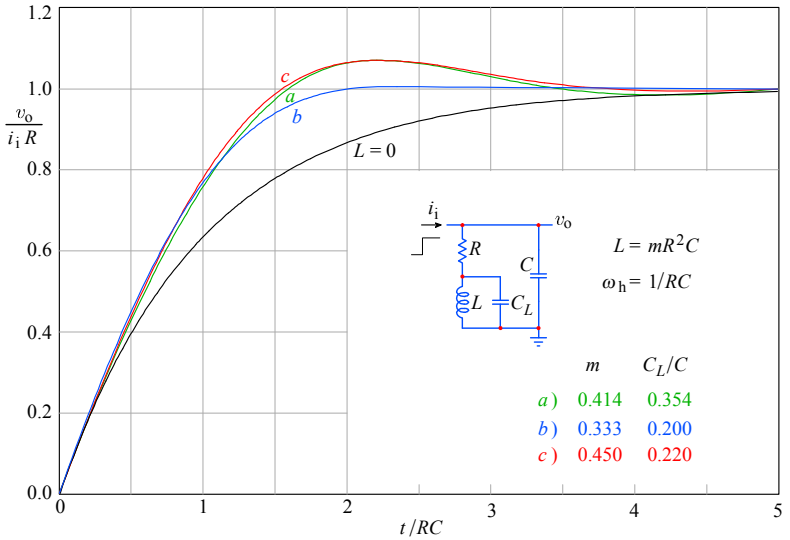


Fig. 2.8.5: Three-pole shunt peaking circuit step response: ***a)*** MFA; ***b)*** MFED; ***c)*** SPEC. The non-peaking case ($L = 0$) is the reference. The initial slope is similar to the non-peaking response, since the difference between the number of system poles and zeros is one only.

We conclude the discussion of the three-pole and two-zero shunt peaking circuit by the Table 2.8.1, which gives all the important circuit parameters.

Table 2.8.1

response type	m	n	η_b	η_r	δ [%]
a) MFA	0.414	0.354	1.84	1.85	7.1
b) MFED	0.333	0.200	1.72	1.74	0.37
c) SPEC	0.450	0.220	1.84	1.91	7.0

Table 2.8.1: Three-pole shunt-peaking circuit parameters

2.9 Shunt–Series Peaking Circuit

In those cases in which the amplifier capacitance may be split into two parts, C_i and C, we can combine the shunt and the series peaking to form a network, shown in Fig. 2.9.1, named the shunt–series peaking circuit. The bandwidth of the shunt–series circuit is increased further than can be achieved by each system alone.

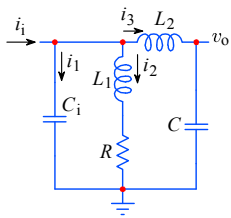


Fig. 2.9.1: The shunt–series peaking circuit.

Although the improvement of the bandwidth and rise time in a shunt–series peaking circuit exceeds that of a pure series or pure shunt peaking circuit, the improvement factors just barely reach the values offered by the three-pole T-coil circuit, which is analytically and practically much easier to deal with; not to speak of the improvement offered by the L+T network, which is substantially greater. This circuit has been extensively treated in literature [Ref. 2.4, 2.25, 2.26]. The calculation of the step response for this circuit can be found in Appendix 2.3, so we shall give only the essential relations.

We start the analysis by calculating the input impedance:

$$Z_i = \frac{V_i}{I_i} = \frac{V_i}{I_1 + I_2 + I_3} \tag{2.9.1}$$

where:

$$I_1 = \frac{V_i}{\frac{1}{s\,C_i}} \qquad I_2 = \frac{V_i}{R + s\,L_1} \qquad I_3 = \frac{V_i}{\frac{1}{s\,C} + sL_2} \tag{2.9.2}$$

By introducing this into Eq. 2.9.1 and eliminating the double fractions we get:

$$Z_i = \frac{(R + s\,L_1)(1 + s^2 L_2 C)}{s\,C_i(R + s\,L_1)(1 + s^2 L_2 C) + s^2 L_2 C + 1 + s\,C(R + s\,L_1)} \tag{2.9.3}$$

The output voltage is:

$$V_o = V_i \frac{\frac{1}{s\,C}}{s\,L_2 + \frac{1}{s\,C}} = I_i\,Z_i \frac{1}{s^2 L_2 C + 1} \tag{2.9.4}$$

We insert Eq. 2.9.3 for Z_i, cancel the $s^2 L_2 C + 1$ terms and extract R from the numerator:

$$V_o = I_i R \frac{1 + s\,L_1/R}{s\,C_i(R + s\,L_1)(1 + s^2 L_2 C) + s^2 L_2 C + 1 + s\,C(R + s\,L_1)} \tag{2.9.5}$$

We divide this by $I_i R$ to get the transfer function normalized in amplitude. Also we multiply all the terms in parentheses and rearrange the result to obtain the canonical form

(divide by the coefficient at the highest power of s) first in the numerator (because it is easy) and then in the denominator:

$$\frac{V_o}{I_i R} = \frac{\dfrac{L_1}{R}\left(s+\dfrac{R}{L_1}\right)}{(s\,C_i R + s^2\,C_i L_1)(1+s^2 L_2 C) + s^2 L_2 C + 1 + s\,C(R+s\,L_1)} \tag{2.9.6}$$

$$= \frac{\dfrac{L_1}{R}\left(s+\dfrac{R}{L_1}\right)}{s^4 L_1 L_2 C C_i + s^3 L_2 C C_i R + s^2\,C_i L_1 + s^2 L_2 C + s^2\,L_1 C + s\,C R + s\,C_i R + 1}$$

$$= \frac{\dfrac{1}{L_1 L_2\,C\,C_i}\cdot\dfrac{L_1}{R}\left(s+\dfrac{R}{L_1}\right)}{s^4 + s^3\dfrac{R}{L_1} + s^2\,\dfrac{L_2 C + L_1 C_i + L_1 C}{L_1 L_2\,C\,C_i} + s\,\dfrac{R\,(C_i + C)}{L_1 L_2\,C\,C_i} + \dfrac{1}{L_1 L_2\,C\,C_i}}$$

Since we would like to know how much we can improve the bandwidth with respect to the non-peaking circuit (inductances shorted), let us normalize the transfer function to $\omega_h = 1/R(C_i + C)$ by putting $R = 1$ and $C_i + C = 1$. To simplify the expressions, we introduce the following parameters:

$$n = \frac{C}{C + C_i} \qquad m_1 = \frac{L_1}{R^2(C + C_i)} \qquad m_2 = \frac{L_2}{R^2(C + C_i)} \tag{2.9.7}$$

and by using the normalization we have:

$$C \Rightarrow n \qquad C_i \Rightarrow (1-n) \qquad L_1 \Rightarrow m_1 \qquad L_2 \Rightarrow m_2 \tag{2.9.8}$$

With these expressions the frequency response Eq. 2.9.6 becomes:

$$F(s) = \frac{\dfrac{1}{m_1 m_2\,n\,(1-n)}\,m_1\left(s+\dfrac{1}{m_1}\right)}{s^4 + s^3\dfrac{1}{m_1} + s^2\,\dfrac{m_2\,n + m_1}{m_1 m_2\,n\,(1-n)} + s\,\dfrac{1}{m_1 m_2\,n\,(1-n)} + \dfrac{1}{m_1 m_2\,n\,(1-n)}} \tag{2.9.9}$$

Now we compare this with the generalized four-pole one-zero transfer function:

$$F(s) = \frac{(-1)^4\,s_1 s_2 s_3 s_4}{(s-s_1)(s-s_2)(s-s_3)(s-s_4)}\cdot\frac{s-s_5}{-s_5} \tag{2.9.10}$$

From the numerator it is immediately clear that the zero is:

$$s_5 = -\frac{1}{m_1} \tag{2.9.11}$$

and the product of the poles is:

$$s_1 s_2 s_3 s_4 = \frac{1}{m_1 m_2\,n\,(1-n)} \tag{2.9.12}$$

Next we transform the denominator of Eq. 2.9.10 into a canonical form:

$$(s-s_1)(s-s_2)(s-s_3)(s-s_4) = s^4 + as^3 + bs^2 + cs + d \tag{2.9.13}$$

where:

$$\begin{aligned} a &= -(s_1 + s_2 + s_3 + s_4) \\ b &= s_1 s_2 + s_1 s_3 + s_1 s_4 + s_2 s_3 + s_2 s_4 + s_3 s_4 \\ c &= -(s_1 s_2 s_3 + s_1 s_2 s_4 + s_1 s_3 s_4 + s_2 s_3 s_4) \\ d &= s_1 s_2 s_3 s_4 \end{aligned} \tag{2.9.14}$$

By comparing the coefficients at equal powers of s, we note that:

$$a = \frac{1}{m_1} \qquad b = \frac{m_2\, n + m_1}{m_1 m_2\, n\,(1-n)} \qquad c = d = \frac{1}{m_1 m_2\, n\,(1-n)} \tag{2.9.15}$$

For the MFED response the coefficients of the fourth-order Bessel polynomial (which we obtain by running the BESTAP routine in Part 6) have the following numerical values:

$$a = 10 \qquad b = 45 \qquad c = 105 \qquad d = 105 \tag{2.9.16}$$

So, from a:

$$m_1 = 0.1 \tag{2.9.17}$$

From b and c:

$$b = (m_2\, n + m_1)\, c \qquad \Rightarrow \qquad m_2 = \frac{\dfrac{b}{c} - m_1}{n} \tag{2.9.18}$$

From c or d:

$$\begin{aligned} n(1-n) = \frac{1}{c\, m_1 m_2} \qquad &\Rightarrow \qquad n - n^2 - \frac{n}{105 \cdot 0.1 \cdot \left(\dfrac{45}{105} - 0.1\right)} = 0 \\ &\Rightarrow \qquad n - n^2 - \frac{n}{3.45} = 0 \\ &\Rightarrow \qquad n\left(1 - n - \frac{1}{3.45}\right) = 0 \end{aligned} \tag{2.9.19}$$

And, since $n \neq 0$:

$$n = 1 - \frac{1}{3.45} = 0.7101 \tag{2.9.20}$$

With this we calculate m_2:

$$m_2 = \frac{\dfrac{b}{c} - m_1}{n} = \frac{\dfrac{45}{105} - 0.1}{0.7101} = 0.4627 \tag{2.9.21}$$

The component values for the MFED transfer function will be:

$$\begin{aligned} C &= n(C_\mathrm{i} + C) = 0.7101\,(C_\mathrm{i} + C) \\ C_\mathrm{i} &= (1-n)(C + C_\mathrm{i}) = 0.2899\,(C_\mathrm{i} + C) \\ L_1 &= m_1 R^2 (C + C_\mathrm{i}) = 0.1\, R^2 (C_\mathrm{i} + C) \\ L_2 &= m_2\, R^2 (C + C_\mathrm{i}) = 0.4627\, R^2 (C_\mathrm{i} + C) \end{aligned} \tag{2.9.22}$$

The MFED poles s_{1-4} (BESTAP routine, Part 6) and the zero s_5 (Eq. 2.9.11) are:

$$\begin{aligned} s_{1n,2n} &= s_{1t,2t} = -2.8962 \pm j\,0.8672 \\ s_{3n,4n} &= s_{3t,4t} = -2.1038 \pm j\,2.6574 \\ s_{5n} &= -10.000 \end{aligned} \tag{2.9.23}$$

For the MFA we can use the same procedure as in Sec.2.3.1, but since we have a system of 4th-order we would get an 8th-order polynomial and, consequently, a complicated set of equations to solve. Instead we shall use a simpler approach (which, by the way, can be used in any other case). We must first consider that our system will have a bandwidth larger than the normalized Butterworth system. Let η_b be the proportionality factor between each normalized Butterworth pole and the shunt–series peaking system pole:

$$s_k = \eta_b\, s_{kt} \tag{2.9.24}$$

The normalized 4th-order Butterworth system poles (see Part 6, BUTTAP routine) are:

$$\begin{aligned} s_{1t,2t} &= -0.3827 \pm j\,0.9239 \\ s_{3t,4t} &= -0.9239 \pm j\,0.3827 \end{aligned} \tag{2.9.25}$$

and the values of the characteristic polynomial coefficients are:

$$1.0000 \quad 2.6131 \quad 3.4142 \quad 2.6131 \quad 1.0000 \tag{2.9.26}$$

The polynomial coefficients a, b, c and d of the shunt-series peaking system are then:

$$\begin{aligned} a &= -\eta_b\,(s_{1t} + s_{2t} + s_{3t} + s_{4t}) = \eta_b \cdot 2.6131 \\ b &= \eta_b^2(s_{1t}s_{2t} + s_{1t}s_{3t} + s_{1t}s_{4t} + s_{2t}s_{3t} + s_{2t}s_{4t} + s_{3t}s_{4t}) = \eta_b^2 \cdot 3.4142 \\ c &= -\eta_b^3(s_{1t}s_{2t}s_{3t} + s_{1t}s_{2t}s_{4t} + s_{1t}s_{3t}s_{4t} + s_{2t}s_{3t}s_{4t}) = \eta_b^3 \cdot 2.6131 \\ d &= \eta_b^4 s_{1t}s_{2t}s_{3t}s_{4t} = \eta_b^4 \end{aligned} \tag{2.9.27}$$

Since the coefficients a, b, c and d are the same as in Eq. 2.9.15, we get four equations form which we will extract the values of factors η_b, m_1, m_2 and n:

$$\begin{aligned} \eta_b \cdot 2.6131 &= \frac{1}{m_1} & \eta_b^2 \cdot 3.4142 &= \frac{m_2\,n + m_1}{m_1 m_2\,n\,(1-n)} \\ \eta_b^3 \cdot 2.6131 &= \frac{1}{m_1 m_2\,n\,(1-n)} & \eta_b^4 &= \frac{1}{m_1 m_2\,n\,(1-n)} \end{aligned} \tag{2.9.28}$$

From the last two equations we immediately find the value of η_b:

$$\eta_b^3 \cdot 2.6131 = \eta_b^4 \quad \Rightarrow \quad \boxed{\eta_b = 2.6131} \tag{2.9.29}$$

Effectively, the pole multiplication factor is equal to the MFA bandwidth extension. From the first equation of 2.9.28 we can now calculate m_1:

$$m_1 = \frac{1}{2.6131\,\eta_b} = \frac{1}{\eta_b^2} = 0.1464 \tag{2.9.30}$$

From the last equation of 2.9.28 we can establish the relationship between m_2 and n:

$$\eta_b^4 = \frac{1}{m_1 m_2\, n\,(1-n)} \qquad \Rightarrow \qquad m_2 = \frac{1}{\eta_b^2\, n\,(1-n)} \tag{2.9.31}$$

Finally, from the second equation we can derive n:

$$\eta_b^2 \cdot 3.4142 = \frac{m_2\, n + m_1}{m_1 m_2\, n\,(1-n)} \tag{2.9.32}$$

$$\Rightarrow \quad \eta_b^2 \cdot 3.4142\; m_1 m_2\, n\,(1-n) = m_2\, n + m_1$$

Here we substitute m_1 with $1/\eta_b^2$ and m_2 with $1/\eta_b^2\, n\,(1-n)$:

$$\Rightarrow \quad 3.4142\, \frac{1}{\eta_b^2} = \frac{n}{\eta_b^2\, n\,(1-n)} + \frac{1}{\eta_b^2}$$

$1/\eta_b^2$ cancels, as well as n:

$$\Rightarrow \quad 3.4142 \;=\; \frac{1}{(1-n)} + 1$$

$$\Rightarrow \quad n \;= 1 - \frac{1}{(3.4142-1)} = 0.5858 \tag{2.9.33}$$

And now we can calculate m_2:

$$m_2 = \frac{1}{\eta_b^2\, n\,(1-n)} = 0.6036 \tag{2.9.34}$$

The MFA poles and the zero are:

$$\begin{aligned} s_{1n,2n} &= -2.4142 \pm j1.0000 \\ s_{3n,4n} &= -1.0000 \pm j2.4142 \\ s_{5n} &= -6.8283 \end{aligned} \tag{2.9.35}$$

and the MFA coefficients are:

$$1.0000 \quad 6.8284 \quad 23.3132 \quad 46.6260 \quad 46.6260 \tag{2.9.36}$$

In addition to the numerical values of parameters m_1 , m_2 and n just calculated, we will also show the results for MFED obtained from two other sources, Braude [Ref. 2.25] and Shea [Ref. 2.26, loc. cit.], to illustrate the possibility of different optimization strategies.

All the design parameters and performance indicators are in Table 2.9.1 at the end of this section.

Let us insert these data into Eq. 2.9.9 and calculate the poles and the zero :

a) MFA by PS/EM ($m_1 = 0.1464$, $m_2 = 0.6036$, and $n = 0.5858$):

$$s^4 + 6.8284\,s^3 + 23.3132\,s^2 + 46.6260\,s + 46.6260 = 0$$

The poles are: $s_{1n,2n} = \sigma_{1n} \pm j\omega_{1n} = -2.4142 \pm j1.0000$

$s_{3n,4n} = \sigma_{3n} \pm j\omega_{3n} = -1.0000 \pm j2.4142$

and the zero: $s_{5n} = \sigma_{5n} = -6.8284$

b) MFED by PS/EM ($m_1 = 0.1000$, $m_2 = 0.4627$, and $n = 0.7101$):

$$s^4 + 10.0000\,s^3 + 44.9933\,s^2 + 104.9863\,s + 104.9863 = 0$$

The poles are: $s_{1n,2n} = \sigma_{1n} \pm j\omega_{1n} = -2.8976 \pm j0.8649$

$s_{3n,4n} = \sigma_{3n} \pm j\omega_{3n} = -2.1024 \pm j2.6573$

and the zero: $s_{5n} = \sigma_{5n} = -10.0000$

c) MFED by Shea ($m_1 = 0.133$, $m_2 = 0.467$ and $n = 0.667$):

$$s^4 + 7.5188\,s^3 + 32.2198\,s^2 + 72.4872\,s + 72.4872 = 0$$

The poles are: $s_{1n,2n} = \sigma_{1n} \pm j\omega_{1n} = -2.1360 \pm j1.0925$

$s_{3n,4n} = \sigma_{3n} \pm j\omega_{3n} = -1.6234 \pm j3.1556$

and the zero: $s_{5n} = \sigma_{5n} = -7.5188$

d) MFED by Braude ($m_1 = 0.122$, $m_2 = 0.511$, and $n = 0.656$):

$$s^4 + 8.1967\,s^3 + 32.4996\,s^2 + 71.0816\,s + 71.0816 = 0$$

The poles are: $s_{1n,2n} = \sigma_{1n} \pm j\omega_{1n} = -2.6032 \pm j0.9618$

$s_{3n,4n} = \sigma_{3n} \pm j\omega_{3n} = -1.4951 \pm j2.6446$

and the zero: $s_{5n} = \sigma_{5n} = -8.1967$

2.9.1 Frequency Response

We shall use the normalized formula which we developed for the 4-pole L+T circuit, (Eq. 2.6.10), to which we must include the influence of the zero. The magnitude of the transfer function (to shorten the expression, we will omit the index 'n' here) is:

$$|F(\omega)| = \frac{\left(\sigma_1^2 + \omega_1^2\right)\left(\sigma_3^2 + \omega_3^2\right)\frac{1}{\sigma_5}\sqrt{\sigma_5^2 + \chi^2}}{\sqrt{\left[\sigma_1^2 + (\chi + \omega_1)^2\right]\left[\sigma_1^2 + (\chi - \omega_1)^2\right]\left[\sigma_3^2 + (\chi + \omega_3)^2\right]\left[\sigma_3^2 + (\chi - \omega_3)^2\right]}} \tag{2.9.37}$$

where again $\chi = \omega/\omega_h$. In Fig. 2.9.2 we have plotted the responses resulting from this equation by inserting the values of the poles and the zero for our MFA and MFED.

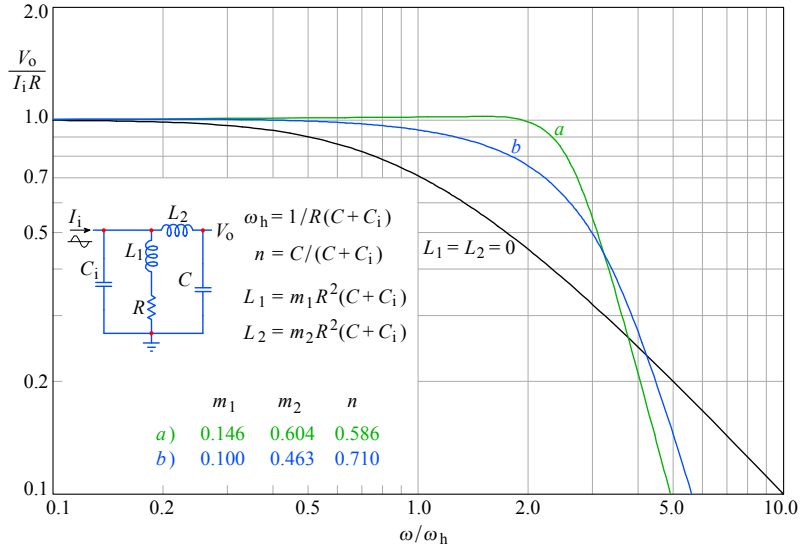


Fig. 2.9.2: The shunt–series peaking circuit frequency-response: ***a)*** MFA; ***b)*** MFED. Note the MFA not being exactly maximally flat, owing to the system zero.

2.9.2 Phase Response

As before, we apply Eq. 2.2.30 for each pole and (negative) for the zero:

$$\varphi(\omega) = \arctan\frac{\frac{\omega}{\omega_h} - \omega_{1n}}{\sigma_{1n}} + \arctan\frac{\frac{\omega}{\omega_h} + \omega_{1n}}{\sigma_{1n}} + \arctan\frac{\frac{\omega}{\omega_h} - \omega_{3n}}{\sigma_{3n}} + \arctan\frac{\frac{\omega}{\omega_h} + \omega_{3n}}{\sigma_{3n}} - \arctan\frac{\frac{\omega}{\omega_h}}{\sigma_{5n}} \qquad (2.9.38)$$

By inserting the values for the poles and the zero from the equations above, we obtain the responses shown in Fig. 2.9.3 .

2.9.3 Envelope Delay

By Eq. 2.2.34, for responses *a*) and *b*) we obtain:

$$\tau_e\omega_h = \frac{\sigma_{1n}}{\sigma_{1n}^2 + \left(\frac{\omega}{\omega_h} + \omega_{1n}\right)^2} + \frac{\sigma_{1n}}{\sigma_{1n}^2 + \left(\frac{\omega}{\omega_h} - \omega_{1n}\right)^2} + \frac{\sigma_{3n}}{\sigma_{3n}^2 + \left(\frac{\omega}{\omega_h} + \omega_{3n}\right)^2} + \frac{\sigma_{3n}}{\sigma_{3n}^2 + \left(\frac{\omega}{\omega_h} - \omega_{3n}\right)^2} - \frac{\sigma_{5n}}{\sigma_{5n}^2 + \left(\frac{\omega}{\omega_h}\right)^2} \qquad (2.9.39)$$

By inserting the values for the poles and the zero from the equations above, we obtain the responses shown in Fig. 2.9.4. Again, as in pure shunt peaking, we have different low frequency delays for each type of poles, owing to the different normalization.

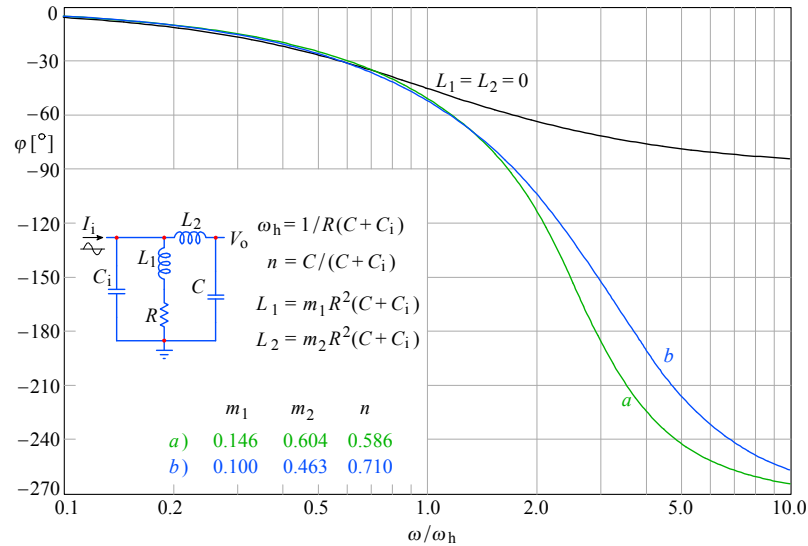


Fig. 2.9.3: The shunt–series peaking circuit phase response: ***a)*** MFA; ***b)*** MFED.

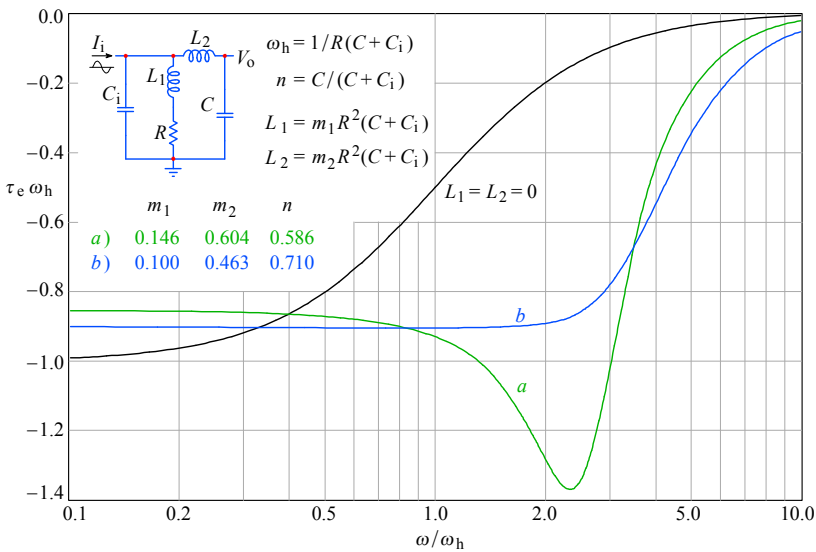


Fig. 2.9.4: The shunt–series peaking circuit envelope delay: ***a)*** MFA; ***b)*** MFED.

2.9.4 Step Response

The normalized general expression for four poles and one zero in the frequency domain is:

$$F(s) = \frac{s_1\, s_2\, s_3\, s_4\, (s - s_5)}{-\, s_5\, (s - s_1)(s - s_2)(s - s_3)(s - s_4)} \tag{2.9.40}$$

To get the step response in the s domain, we multiply $F(s)$ by the unit step operator $1/s$:

$$G(s) = \frac{s_1\, s_2\, s_3\, s_4\, (s - s_5)}{-\, s\, s_5\, (s - s_1)(s - s_2)(s - s_3)(s - s_4)} \tag{2.9.41}$$

The step response in the time domain is obtained by taking the $\mathcal{L}^{-1}$ transform:

$$g(t) = \mathcal{L}^{-1}\{G(s)\} = \sum \text{res}\, \frac{s_1\, s_2\, s_3\, s_4\, (s - s_5)\, e^{st}}{-\, s\, s_5\, (s - s_1)(s - s_2)(s - s_3)(s - s_4)} \tag{2.9.42}$$

This formula requires even more effort than was spent for the L+T network. We shall skip the lengthy procedure (which is presented in Appendix 2.3) and give only the solution, which for all the listed poles and zeros is:

$$g(t) = 1 - \frac{K_1}{\sigma_5\, \omega_1}\, e^{\sigma_1 t}[M \sin(\omega_1 t/T) + \omega_1 N \cos(\omega_1 t/T)]$$

$$- \frac{K_3}{\sigma_5\, \omega_3}\, e^{\sigma_3 t}[P \sin(\omega_3\, t/T) + \omega_3\, Q \cos(\omega_3\, t/T)] \tag{2.9.43}$$

where:

$$\begin{aligned} M &= (\sigma_1 - \sigma_5)[\sigma_1 A - \omega_1^2 B] + \omega_1^2 (A + \sigma_1 B) \\ N &= [\sigma_1 A - \omega_1^2 B] - (\sigma_1 - \sigma_5)(A + \sigma_1 B) \\ P &= (\sigma_3 - \sigma_5)[\sigma_3 C + \omega_3^2 B] + \omega_3^2 (C - \sigma_3 B) \\ Q &= [\sigma_3 C + \omega_3^2 B] - (\sigma_3 - \sigma_5)(C - \sigma_3 B) \end{aligned} \tag{2.9.44}$$

whilst A, B, C, K_1 and K_3 are the same as for the L+T network (Eq. 2.6.17).

The plots in Fig. 2.9.5 and Fig. 2.9.6 were calculated and drawn by using these formulae.

Let us now compare the MFED response with those obtained by Braude and Shea. The step response relation is the same for all three systems (Eq. 2.9.43, 2.9.44), but the pole and zero values are different. As it appears from the comparison of the characteristic polynomial coefficients and even more so from the comparison of the poles and zeros, the three systems were optimized in different ways. This is evident from Fig. 2.9.7.

Although at first glance all three step responses look very similar (Fig. 2.9.6), a closer look reveals that the Braude case has an excessive overshoot. The Shea case has the steepest slope (largest bandwidth), but this is paid for by an extra overshoot and ringing. The Bessel system has the lowest transient slope; however, it has the minimal overshoot and it is the first to settle to $<0.1\%$ of the final amplitude value (in about $2.7\, t/T$).

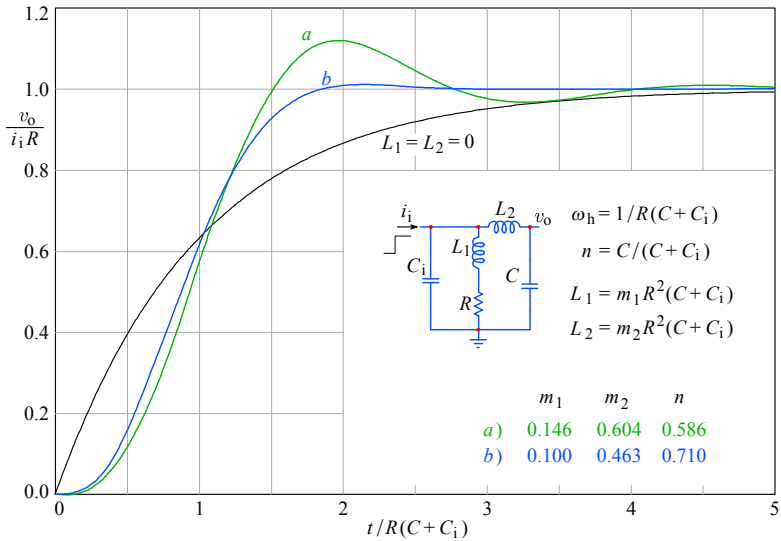


Fig. 2.9.5: The shunt–series peaking circuit step response: ***a)*** MFA; ***b)*** MFED.

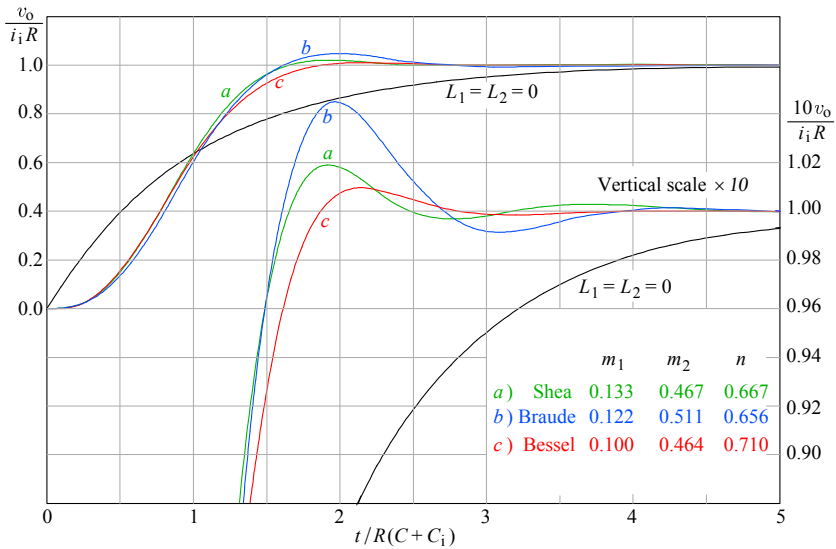


Fig. 2.9.6: The MFED shunt–series peaking circuit step response: ***a)*** by Shea; ***b)*** by Braude; ***c)*** a true Bessel system. The ×10 vertical scale expansion shows the top 10 % of the response. The overshoot in the Braude case is excessive, whilst the Shea version has a prolonged ringing. Although slowest, the Bessel system is the first to settle to < 0.1 % of the final value.

The pole layout in Fig. 2.9.7 confirms the statements above. In the Braude case the two poles with the smaller imaginary part are too far from the imaginary axis to compensate the peaking of the two poles closer, so the overshoot is inevitable. The Shea case has the widest pole spread and consequently the largest bandwidth, but the two poles with the

lower imaginary part are too close to the imaginary axis (this is needed in order to level out the peaks and deeps in the frequency response). As a consequence, whilst the overshoot is just acceptable, there is some long term ringing, impairing the system's settling time. The Bessel system pole layout follows the theoretical requirement. In spite of the presence of the zero (located far from the poles, the farthest of all three systems), the system performance is optimal.

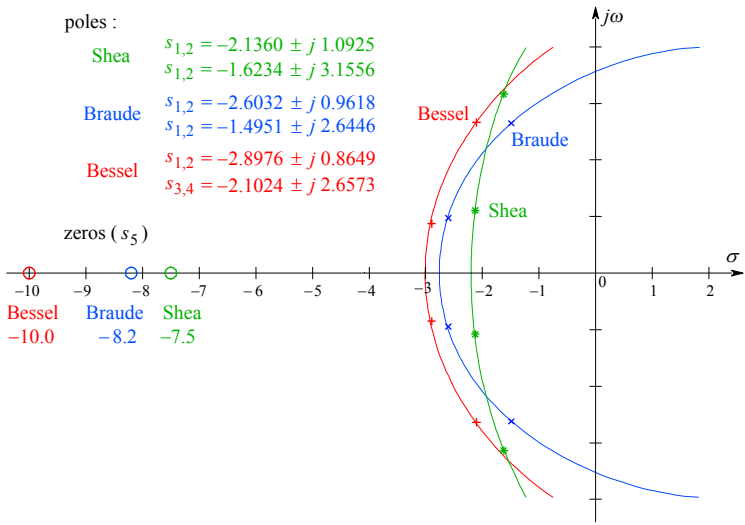


Fig. 2.9.7: The MFED shunt–series peaking circuit pole loci of the three different systems. The zero of each system is too far from the poles to have much influence. It is interesting how a similar step response can be obtained using three different optimization strategies. Strictly speaking, only the Bessel system is optimal.

Let us conclude this section with Table 2.9.1, in which we have collected all the design parameters, in addition to the bandwidth and rise time improvements and the overshoots for the cases discussed.

Table 2.9.1

response type	author	m_1	m_2	n	η_b	η_r	δ [%]
a) MFA	PS/EM	0.1464	0.6036	0.5858	2.61	2.72	12.23
b) MFED	PS/EM	0.1000	0.4627	0.7101	2.18	2.21	0.90
c) MFED	Shea	0.133	0.467	0.667	2.44	2.39	1.86
d) MFED	Braude	0.122	0.511	0.656	2.50	2.36	4.45

Table 2.9.1: Series–shunt peaking circuit parameters.

This completes our discussion of inductive peaking circuits.

We have deliberately omitted the analysis of the combination of a 3-pole + 1-zero shunt peaking , combined with a two-pole series peaking circuit. This network, introduced by *R.L. Dietzold*, was thoroughly analyzed in 1948 in the book *Vacuum Tube Amplifiers*

[Ref. 2.2] (the reader who is interested in following the analysis there should consider that several printing mistakes crept into some of the formulae). In those days that circuit represented the ultimate in inductive peaking circuits. Today we achieve a much better bandwidth and rise time improvement with the L+T circuit, discussed in Sec. 2.6, which is easier to realize in practice and also requires substantially less mathematical work.

With shunt–series peaking it is sometimes not possible to achieve the required ratio of stray capacitances n as in Table 2.9.1; but by adding an appropriate damping resistor across the series peaking coil it is possible to adapt the shunt–series peaking circuit also to an awkward ratio n. This is well described in [Ref. 2.2]. However the bandwidth and the rise time improvement of such circuits may be either similar to that of a three-pole shunt peaking circuit, or even worse than that, so we will not discuss them.

To summarize: in view of the advanced T-coil circuits, the shunt–series peaking circuit may be considered obsolete. This is why we have not discussed it as extensive as all the other inductive peaking circuits. On the other hand, by omitting the shunt–series peaking circuit entirely, the discussion of the inductive peaking circuits would not be complete.

2.10 Comparison of MFA Frequency Responses and of MFED Step Responses

In an actual process of amplifier design the choice of circuits used in different amplifying stages is not just a matter of a designer's personal taste or a simple collection of best performing circuits. Rather, it is a process of carefully balancing the advantages and disadvantages both at the system level and at each particular stage.

To help the designer in making a decision, now that we have analyzed the frequency responses and step responses of the most important types of inductive peaking circuits, we compare their performance in the following two plots.

We have drawn all the MFA frequency responses in Fig. 2.10.1 and all the MFED step responses in Fig. 2.10.2.

On the basis of both figures we conclude that T-coil circuits surpass all the other types of inductive peaking circuits.

In addition we have collected all the data for the circuit parameters corresponding to both figures in the table in Appendix 2.4 (web only). The table contains the circuit schematic diagram, the relations between the component values, the normalized pole and zero values, the formulae for the frequency responses and the step responses, as well as the bandwidth and the rise time enhancement and the step response overshoot.

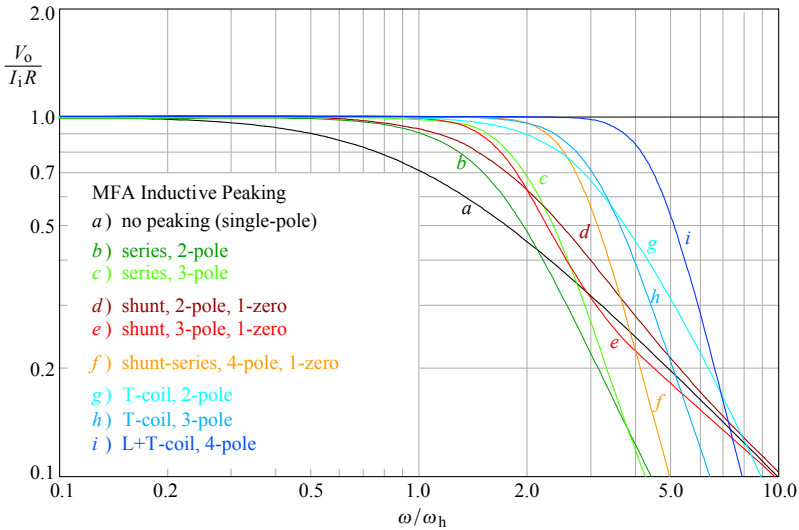


Fig. 2.10.1: MFA frequency responses of all the circuit configurations discussed. By far the 4-pole T-coil response ***i)*** has the largest bandwidth.

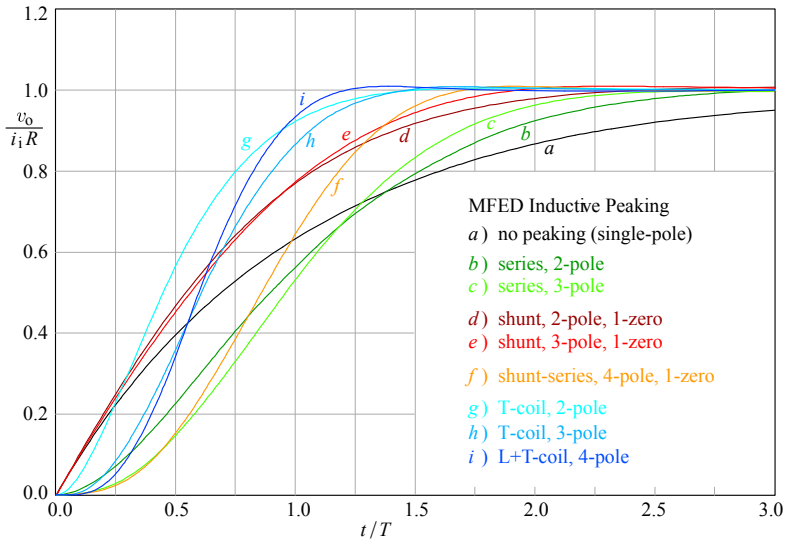


Fig. 2.10.2: MFED step responses of all the circuit configurations discussed. Again, the 4-pole T-coil step response ***i)*** has the steepest slope, but the 3-pole T-coil response ***h)*** is close.

2.11 The Construction of T-coils

Most of the 'know how' concerning the construction of T-coils is a classified matter of different firms, mostly Tektronix, Inc., so we shall discuss only some basic facts about how to make T-coils.

Fig. 2.11.1, made originally by *Carl Battjes* (although with different basic circuit parameters as the reference), shows the performance sensitivity on each component's tolerance of an L+T circuit designed nominally for the MFED response.

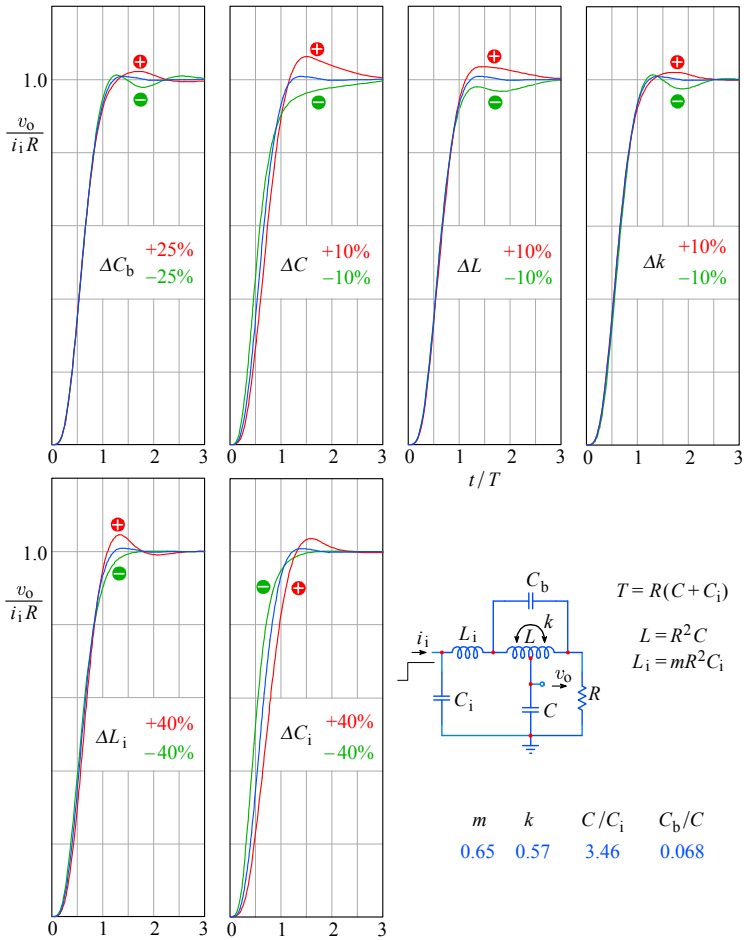


Fig. 2.11.1: Four-pole L+T peaking circuit step response sensitivity on component tolerances. Such graphs were drawn originally by *Carl Battjes* for a class lecture at Tektronix, but with another set of parameters as the reference. The responses presented here were obtained using the MicroCAP–5 circuit analysis program. [Ref.2.36].

These figures prove that the inductance L, the coupling factor k, and the loading capacitance C must be kept within close tolerances in order to achieve the desired performance, whilst the tolerances of the bridging capacitance C_b of the T-coil, the input coil L_i, and input capacitance C_i are less critical. Therefore, the construction of a properly calculated T-coil is not a simple matter. In some respect it resembles a digital AND function: only if all the parameters are set correctly will the result be an efficient peaking circuit. There is not much room for compromise here.

In the serial production of wideband amplifiers there are always some tolerances of stray capacitances, so the T-coils must be made adjustable. Usually the coils are wound on a simple polystyrene cylindrical coil form, with a threaded ferrite core inside. By adjusting the core the required inductance can be set. However, the coupling factor k depends only on the coil *length* to *diameter* ratio (l/D) [Ref. 2.33] **and it is independent of whether the coil has a ferrite core inside or not**. The relation between the coupling factor k and the ratio l/D is shown in the diagram in Fig. 2.11.2, which is valid for the center tapped cylindrical coils.

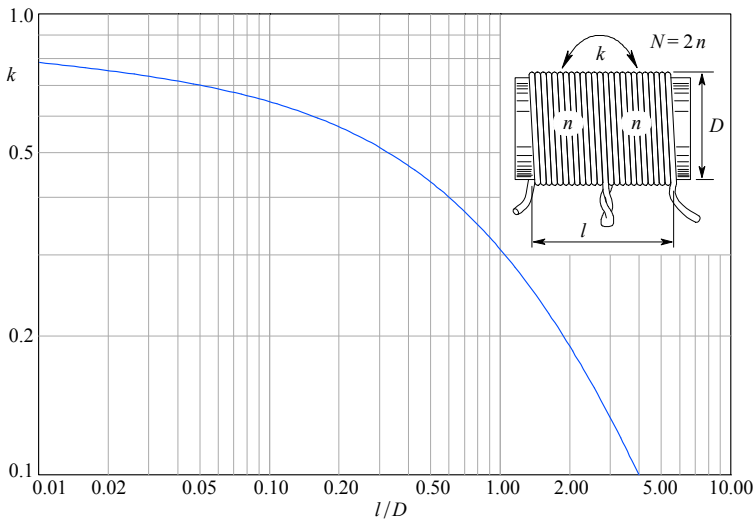


Fig. 2.11.2: T-coil coupling factor k as a function of the coil's length to diameter (l/D) ratio.

The coil inductance L depends on the number of turns, the length to diameter ratio, set by the coil form on which the coil is wound, and on the ferrite core if any is used; both the coil form and the core can be obtained from different manufacturers, together with the formulae for calculating the required number of turns. However, these formulae are often given as some sort of 'cookery book receipts', with the key parameters usually in numerical form for a particular product. As such, they are satisfactory for building general purpose coils, but they do not offer the understanding needed to perform the optimization procedure within a set of possible solutions.

The reader is therefore forced to look for more theoretical explanations in standard physics and electronics text books.

In those text books the following relation can be found:

$$L = \frac{\mu_0 N^2 A}{l} \tag{2.11.1}$$

but this is **valid only for a single layer coreless coil with a homogeneous magnetic field** (such as a tightly wound toroid or a long coil). The parameters represent:

- L inductance in henrys (after *Joseph Henry*, 1791-1878), [1 H = 1 Vs/A]; the inductance of 1 H produces a self-induced voltage of 1 V when a current flowing through it changes at a rate of 1 A/s;
- μ_0 the free space magnetic permeability, $\mu_0 = 4\pi \cdot 10^{-7}$ Vs/Am;
- N the total number of turns;
- A the area encircled by one wire turn, measured from the wire central path; for a cylindrical coil, $A = \pi R^2 = \pi D^2/4$, where R is the radius and D is the diameter, both in meters [m];
- l the total coil length in meters [m]; if the turns are wound adjacent to one another with a wire of a diameter d, then $l = Nd$.

The main problem with the Eq. 2.11.1 is the term 'homogeneous'; this implies a uniform magnetic field, entirely contained within the coil, with no appreciable leakage outwards. Toroidal coils are not easy to build and can not be made adjustable, so in practice cylindrical coils are widely used. For a cylindrical coil the magnetic field is of the same form as that of a stick magnet: the field lines close outside the coil and at both ends the field is non-homogeneous. Because of this, the inductance is reduced by a form factor ζ, which is a function of the ratio D/l (Fig. 2.11.3).

An important note for T-coil production: the form factor, and consequently the inductance, increase with D and decrease with l, in contrast to the coupling factor k. This additionally restricts our choice of D and l.

Also, if the coil is going to be adjustable the relative permeability of the core material, μ_r , must be taken into account; however, only a part of the magnetic field will be contained within the core, so we introduce the average permeability, $\overline{\mu_r}$, reflecting that only a part of the turns will encircle the core. The relative permeability of the air is 1 and that of the ferromagnetic core material can be anything up to several hundred. However, since the field path in air will be much longer than inside the core, the average permeability will be rather low. Note also that the core material is 'slow', i.e., its permeability has an upper frequency limit, often lower than our bandwidth requirement.

Finally, if the bridging capacitance C_b of the T-coil network has to be precisely known, we must take into account the coil's self capacitance, C_s, which appears in parallel with the coil, with a value equivalent to a series connection of capacitances between adjacent turns. Owing to C_s the inductance will appear lower when measured, so C_s should also be measured and the actual inductance value calculated from the two measurements. If the turns are tightly wound the relative permittivity ε_r of the wire isolation lacquer must be considered. Its value is several times larger than for air. The lacquer thickness is also influencing C_s. If C_s is too large it can easily be reduced by increasing the distance between the turns by a small amount, δ, but this will also cause additional field leakage and reduce the inductance slightly. To compensate, the number of turns can be increased; because the inductance increases with N^2, it will outperform the slight decrease resulting from the larger length l.

Multi-layer coils are less suitable for use in wideband amplifiers, because of their high capacitance between the adjacent layers.

Fortunately wideband amplification does not require large inductance values. Also, since the inductances are always in series with relatively large resistive loads (almost never less than 50 Ω), the wire resistance and the skin effect can usually be neglected.

With all these considerations the inductance becomes:

$$L = \zeta \frac{\overline{\mu_r} \mu_0 \pi D^2 l}{4 (d + \delta)^2} \tag{2.11.2}$$

The Fig. 2.11.3 shows the value of ζ as a function of the ratio D/l. The actual function is found through elliptic integration of the magnetic field flux density, which is too complex to be solved analytically here. But a fair approximation, fitting the experimental data to better than 1%, can be obtained using the following relation:

$$\zeta = \frac{a}{a + \left(\frac{D}{l}\right)^b} \tag{2.11.3}$$

where:

$$a = 2$$

$$b = \sin \frac{\pi}{3}$$

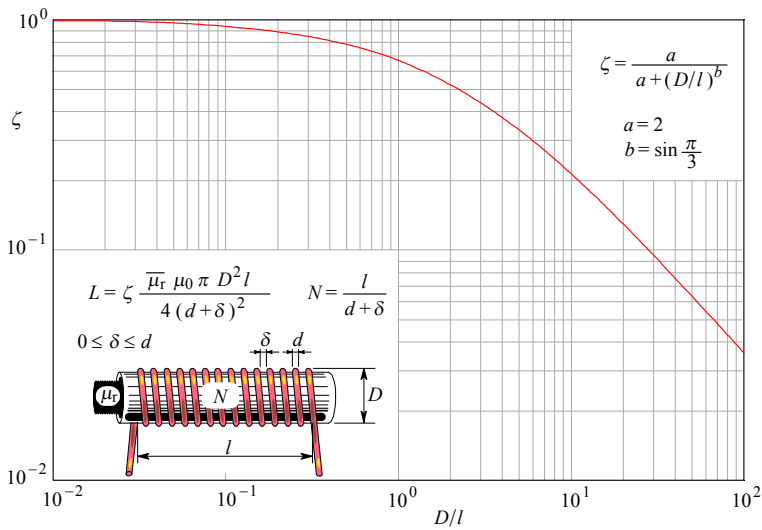


Fig. 2.11.3: The ζ factor as a function of the coil diameter to length ratio, D/l. The equation shown in the upper right corner fits experimental data to better than 1%.

Inductances are susceptible to external fields, mainly from the power supply, or other nearby inductances. The influence of a nearby inductance can be minimized by a perpendicular orientation of coil axes. Otherwise, the circuit should be appropriately

shielded, but the shield will act as a shorted single-turn inductance, lowering the effective coil inductance if it is too close.

In modern miniaturized, bandwidth hungry circuits the coil dimensions become critical, and one possible solution is to construct the coil in a planar spiral form on two sides of a printed circuit board or even within an integrated circuit. This gives the possibility of more tightly controlled parameter tolerances, but there is no free lunch: the price to pay is in many weeks or even months of computer simulation before a satisfactory solution is found by trial and error, since the exact mathematical relations are extremely complicated (a good example of how this is done can be found in the excellent article by *J.Van Hese* of Agilent Technology [Ref. 2.37], where the finite element numerical analysis method is used).

The following figures show a few examples of planar coils made directly on the surface of IC chips, ceramic hybrids, or double-sided conventional PCBs.

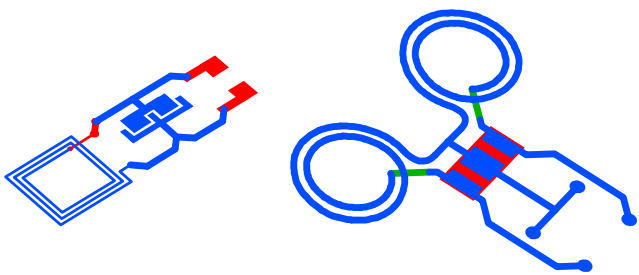


Fig. 2.11.4: Examples of coil structures made directly on an IC chip (left) and on a hybrid circuit (right).

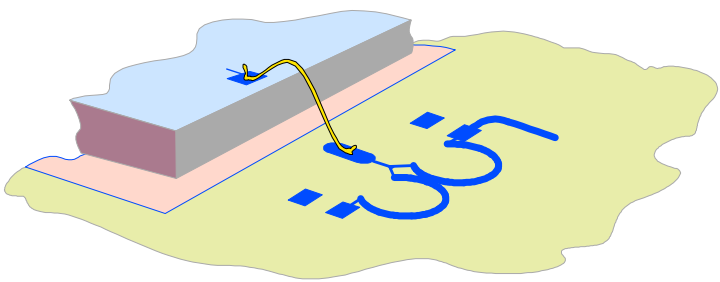


Fig. 2.11.5: A possible compensation of the bonding inductance of an IC chip, mounted on a hybrid circuit, by the negative inductance present at the T-coil center tap.

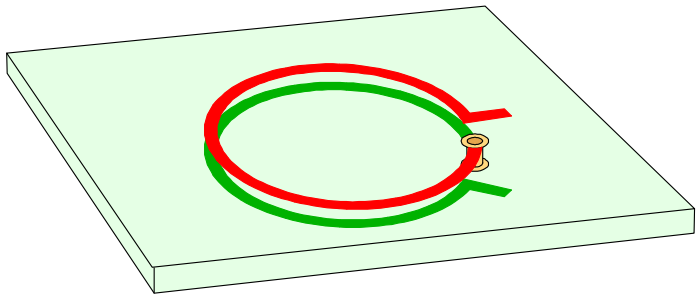


Fig. 2.11.6: A planar T-coil with a high coupling factor, realized on a conventional double-sided PCB. Multi-turn spiral structures are also possible, but need at least a three-layer board for making the inner to outer turn connections.

References:

[2.1] *S. Butterworth*, On the Theory of Filter Amplifiers,
Experimental Wireless & Wireless Engineer, Vol. 7, October, 1930, pp. 536–471.

[2.2] *G.E. Valley & H. Wallman*, Vacuum tube amplifiers,
MIT, Radiation Laboratory Series, Vol. 18, McGraw-Hill, New York, 1948.

[2.3] *J. Bednařík & J. Daněk*, Obrazové zesilovače pro televisi a měřicí techniku
(Video Amplifiers for Television and Measuring Techniques),
Statní nakladatelství technické literatury, Prague, 1957.

[2.4] *E.L. Ginzton, W.R. Hewlett, J.H. Jasberg, J.D. Noe*, Distributed Amplification,
Proc. I.R.E., Vol. 36, August, 1948, pp. 956–969.

[2.5] *P. Starič*, An Analysis of the Tapped-Coil Peaking Circuit for Wideband/Pulse Amplifiers,
Elektrotehniški vestnik, 1982, pp. 66–79.

[2.6] *P. Starič*, Three- and Four-Pole Tapped Coil Circuits for Wideband/Pulse Amplifiers,
Elektrotehniški vestnik, 1983, pp. 129–137.

[2.7] *P. Starič*, Application of T-coil Interstage Coupling in Wideband/Pulse Amplifiers,
Elektrotehniški vestnik, 1990, pp. 143–152.

[2.8] *D.L. Feucht*, Handbook of Analog Circuit Design,
Academic Press, Inc. San Diego, 1990.

[2.9] *J.L. Addis*, Good Engineering and Fast Vertical Amplifiers, Part 4, section 14,
Analog Circuit Design, edited by *J. Williams,*
Butterworth-Heinemann, Boston, 1991.

[2.10] *M.E. Van Valkenburg*, Introduction to Modern Network Synthesis,
John Wiley, New York, 1960.

[2.11] *G. Daryanani*, Principles of Active Network Synthesis and Design,
John Wiley, New York, 1976.

[2.12] *T.R. Cuthbert*, Circuit Design using Personal Computers,
John Wiley, 1983.

[2.13] *G.J.A. Byrd*, Design of Continuous and Digital Electronic Systems,
McGraw-Hill (UK), London, 1980.

[2.14] *P. Starič*, Interpolacija med Butterworthovimi in Thomsonovimi poli
(Interpolation between Butterworth's and Thomson's poles),
Elektrotehniški vestnik, 1987, pp. 133–139.

[2.15] *G.A. Korn & T.M. Korn*, Mathematical Handbook for Scientists and Engineers,
McGraw-Hill, New York, 1961.

[2.16] *F.A. Muller*, High-Frequency Compensation of RC Amplifiers,
Proceedings of the I.R.E., August, 1954, pp. 1271–1276.

[2.17] *C.R. Battjes*, Technical Notes on Bridged T-coil Peaking,
(Internal Publication), Tektronix, Inc., Beaverton, Ore., 1969.

[2.18] *C.R. Battjes*, Who Wakes the Bugler?, Part 2, section 10,
The Art and Science of Analog Circuit Design, edited by *J. Williams,*
Butterworth-Heinemann, Boston, 1995.

[2.19] *N.B. Schrock*, A New Amplifier for Milli-Microsecond Pulses,
Hewlett-Packard Journal, Vol. 1, No. 1., September, 1949.

[2.20] *W.R. Horton, J.H. Jasberg, J.D. Noe*, Distributed Amplifiers: Practical Consideration and
Experimental Results, Proc. I.R.E. Vol. 39, pp. 748–753.

[2.21] *R.I. Ross*, Wang Algebra Speeds Network Computation of Constant Input Impedance Networks,
(Internal Publication), Tektronix, Inc., Beaverton, Ore. 1968.

[2.22] *R.J. Duffin*, An Analysis of the Wang Algebra of the Networks,
Trans. Amer. Math. Soc., 1959, pp. 114–131.

[2.23] *S.P. Chan*, Topology Cuts Circuit Drudgery,
Electronics, November 14, 1966, pp. 112–121.

[2.24] *A.I. Zverev*, Handbook of Filter Synthesis,
John Wiley, New York, 1967.

[2.25] *G.B. Braude, K.V. Epaneshnikov, B.J. Klymushev*, Calculation of a Combined Circuit for the Correction of Television Amplifiers, (in Russian),
Radiotekhnika, T. 4, No. 6. Moscow, 1949, pp. 24–33.

[2.26] *L.J. Giacoletto*, Electronics Designer's Handbook,
Second Edition, McGraw-Hill, New York, 1977.

[2.27] *B. Orwiller*, Vertical Amplifier Circuits,
Tektronix, Inc., Beaverton, Ore., 1969.

[2.28] *D.E. Scott*, An Introduction to Circuit Analysis, A System Approach,
McGraw-Hill, New York, 1987.

[2.29] *J.L. Addis*, Mutual Inductance & T-coils,
(Internal Publication), Tektronix, Inc., Beaverton, Ore. 1977.

[2.30] *A.B. Williams*, Electronic Filter Design Handbook,
McGraw-Hill, New York 1981.

[2.31] *C.R. Battjes*, Amplifier Risetime and Frequency Response,
Class Notes, Tektronix, Inc., Beaverton, Ore. 1969.

[2.32] *W.E. Thomson*, Networks with Maximally Flat Delay,
Wireless Engineer, Vol. 29, 1952, pp 536–541.

[2.33] *F.W. Grover*, Inductance Calculation, (Reprint)
Instrument Society of America, Research Triangle Park, N.C. 27 709, 1973.

[2.34] Mathematica, Wolfram Research, Inc., 100 Trade Center Drive, Champaign, Illinois,
http://www.wolfram.com/

[2.35] Macsyma, Symbolics, Inc. 8 New England Executive Park, East Burlington, Massachusetts, 01803,
http://www.scientek.com/macsyma/main.htm
See also Maxima (free version, GNU Public Licence):
http://www.ma.utexas.edu/users/wfs/maxima.html

[2.36] MicroCAP–5, Spectrum Software, Inc.,
http://www.spectrum-soft.com/

[2.37] *J. Van Hese*, Accurate Modeling of Spiral Inductors on Silicon for Wireless RF-IC Designs,
http://www.techonline.com/ ↦ Feature Articles ↦ Feature Archive ↦ November 20, 2001
See also: http://www.agilent.com/, and:
L. Knockaert, *J. Sercu* and *D. Zutter*, "Generalized Polygonal Basis Functions for the Electromagnetic Simulation of Complex Geometrical Planar Structures," IMS-2001

[2.38] *J.N. Little* and *C.B. Moller,* The MathWorks, Inc.: MATLAB-V For Students
(with the Matlab program on a compact disk), Prentice-Hall, 1998,
http://www.mathworks.com/

[2.39] Derive, http://education.ti.com/product/software/derive/

[2.40] MathCAD, http://www.mathcad.com/

[2.41] *P. Starič*, Inductive peaking circuits,
Electronics & Wireless World, 1988, pp. 471–474.

Appendix 2.1
General Solutions for 1st-, 2nd-, 3rd- and 4th-order polynomials

---

First-order polynomial: $a\,x + b = 0$

Canonical form: $x + \frac{b}{a} = 0$

Solution: $x = -\frac{b}{a}$

Second-order polynomial: $a\,x^2 + b\,x + c = 0$

Canonical form: $x^2 + \frac{b}{a}\,x + \frac{c}{a} = 0$

Solutions: $x_{1,2} = \frac{-\,b \mp \sqrt{b^2 - 4\,a\,c}}{2\,a}$

Third-order polynomial, canonical form:

$$x^3 + a\,x^2 + b\,x + c = 0$$

Solutions:

By substituting:

$$K = \sqrt{a^2 - 3b}$$

$$M = 4a^3c - a^2b^2 - 18abc + 4b^3 + 27c^2$$

$$N = 2a^3 - 9ab + 27c$$

the real solution is:

$$x_1 = -\frac{a}{3} - \frac{2}{3}\,K\,sin\,\frac{atan\,\frac{j\,N}{3\sqrt{3M}}}{3}$$

and the two complex conjugate solutions are:

$$x_2 = -\frac{a}{3} + K\,sin\,\frac{atan\,\frac{j\,N}{3\sqrt{3M}}}{3} - \frac{\sqrt{3}}{3}\,K\,cos\,\frac{atan\,\frac{j\,N}{3\sqrt{3M}}}{3}$$

$$x_3 = -\frac{a}{3} + K\,sin\,\frac{atan\,\frac{j\,N}{3\sqrt{3M}}}{3} + \frac{\sqrt{3}}{3}\,K\,cos\,\frac{atan\,\frac{j\,N}{3\sqrt{3M}}}{3}$$

Note: If preferred, here is a purely algebraic (non-trigonometric) result, obtained by using the new symbolic calculus capability of Matlab (Version 5.3 for Students). The command lines needed are simply:

```
syms x a b c       % define x, a, b and c as symbols
r = solve( x^3 + a*x^2 + b*x + c ) ;
```

The real solution is:

```
r(1) =
   1/6*
      (36*a*b-108*c-8*a^3+12*
         (12*b^3-3*b^2*a^2-54*b*c*a+81*c^2+12*c*a^3)^(1/2)
      )^(1/3)-
   6*(1/3*b-1/9*a^2)/
      (36*a*b-108*c-8*a^3+12*
         (12*b^3-3*b^2*a^2-54*b*c*a+81*c^2+12*c*a^3)^(1/2)
      )^(1/3)-
   1/3*a
```

The two complex-conjugate solutions are:

```
r(2) =
   -1/12*
      (36*a*b-108*c-8*a^3+12*
         (12*b^3-3*b^2*a^2-54*b*c*a+81*c^2+12*c*a^3)^(1/2)
      )^(1/3)+
   3*(1/3*b-1/9*a^2)/
      (36*a*b-108*c-8*a^3+12*
         (12*b^3-3*b^2*a^2-54*b*c*a+81*c^2+12*c*a^3)^(1/2)
      )^(1/3)-
   1/3*a+1/2*i*3^(1/2)*
   (1/6*
      (36*a*b-108*c-8*a^3+12*
         (12*b^3-3*b^2*a^2-54*b*c*a+81*c^2+12*c*a^3)^(1/2)
      )^(1/3)+
   6*(1/3*b-1/9*a^2)/
      (36*a*b-108*c-8*a^3+12*
         (12*b^3-3*b^2*a^2-54*b*c*a+81*c^2+12*c*a^3)^(1/2)
      )^(1/3)
   )
r(3) =
    -1/12*
      (36*a*b-108*c-8*a^3+12*
         (12*b^3-3*b^2*a^2-54*b*c*a+81*c^2+12*c*a^3)^(1/2)
      )^(1/3)+
   3*(1/3*b-1/9*a^2)/
      (36*a*b-108*c-8*a^3+12*
         (12*b^3-3*b^2*a^2-54*b*c*a+81*c^2+12*c*a^3)^(1/2)
      )^(1/3)-
   1/3*a-1/2*i*3^(1/2)*
   (1/6*
      (36*a*b-108*c-8*a^3+12*
         (12*b^3-3*b^2*a^2-54*b*c*a+81*c^2+12*c*a^3)^(1/2)
      )^(1/3)+
   6*(1/3*b-1/9*a^2)/
      (36*a*b-108*c-8*a^3+12*
         (12*b^3-3*b^2*a^2-54*b*c*a+81*c^2+12*c*a^3)^(1/2)
      )^(1/3)
   )
```

Fourth-order equation, canonical form:

$$x^4 + a\,x^3 + b\,x^2 + c\,x + d = 0$$

Solutions:
The roots are identical to the roots of two lower order equations:

$$x^2 + (a + A)\frac{x}{2} + \left(y + \frac{ay - c}{A}\right) = 0$$

where:

$$A = \pm\sqrt{8y + a^2 - 4b}$$

and y is any real root of the third-order equation:

$$8y^3 - 4by + (2ac - 8d)y + d(4b - a^2) - c^2 = 0$$

As has been proven by the French mathematician *Evariste Galois* (1811-1832), the solutions of polynomials of order 5 or higher can not be expressed analytically as rational functions of the polynomial coefficients. In such cases, the roots can be found by numerical computation methods (users of Matlab can try the ROOTS routine, which calculates the roots from polynomial coefficients by numerical methods; see also the POLY routine, which finds the coefficients from the roots).

Appendix 2.2

Normalization of complex frequency response functions
or
Why do some expressions have strange signs?

Do not be afraid of mathematics!
It is probably the only rational product of the human mind!
(E. Margan)

A generalized expression of a non-normalized complex frequency response function of an all pole system (no zeros) of the order N can be written in form of a product of poles of the characteristic polynomial:

$$F(s) = \frac{1}{\prod_{i=1}^{N} (s - s_i)} \tag{A2.2.1}$$

At DC ($s = 0$) the system has a gain factor equal to the product of negated poles:

$$F(0) = \frac{1}{\prod_{i=1}^{N} (0 - s_i)} = \frac{1}{\prod_{i=1}^{N} (-s_i)} \tag{A2.2.2}$$

We would like to compare different systems on a fair basis. This is why we introduce parametric normalization in our equations. We have already seen the frequency being normalized to ω/ω_h. This enabled us to compare the cut off frequencies of different systems, which have some components (or even just one) equal, thus helping us to decide which circuit configuration is either better, simpler, or more economical to build.

Obviously the system's gain is one such parameter which would influence the comparison in frequency if not taken into account. This situation is best resolved if we also normalize the DC gain of every system to some predefined value, preferably unity. Mathematically, we will have a normalized expression by simply dividing Eq. A2.2.1 by Eq. A2.2.2:

$$F_\mathrm{n}(s) = \frac{F(s)}{F(0)} = \frac{\dfrac{1}{\prod_{i=1}^{N} (s - s_i)}}{\dfrac{1}{\prod_{i=1}^{N} (-s_i)}} = \frac{\prod_{i=1}^{N} (-s_i)}{\prod_{i=1}^{N} (s - s_i)} \tag{A2.2.3}$$

The numerator of the last term in Eq. A2.2.3 can be written so that the signs are collected together in a separate product, defining the sign of the total:

$$F_{\mathrm{n}}(s) = \frac{(-1)^N \prod_{i=1}^{N} s_i}{\prod_{i=1}^{N} (s - s_i)} \tag{A2.2.4}$$

This means that **all odd order functions must be multiplied by -1** in order to have a correctly normalized expression. But please, note that the sign defining expression $(-1)^N$ is ***not*** the consequence of all our poles lying in the left half of the complex plane, as is sometimes wrongly explained in literature!

In Eq. A2.2.4 the poles still retain their actual value, be it positive or negative. The term $(-1)^N$ is just the consequence of the mathematical operation (subtraction) required by the function: s must acquire the exact value of s_i, **sign included**, if the function is to have a pole at s_i:

$$s - s_i = 0 \Big|_{s = s_i} \quad \Rightarrow \quad F(s_i) = \pm\infty \tag{A2.2.5}$$

In some literature the sign is usually neglected because we are all too often interested in the frequency response magnitude, which is the absolute value of $F(s)$, or $|F(s)|$. However, as amplifier designers we are interested mainly in the system's time domain performance. If we calculate it by the inverse Laplace transform we must have the correct sign of the transfer function, and consequently the signs of the residues at each pole.

In addition it is important to note that a system with zeros must have the product of zeros normalized in the same way (even if some of the systems with zeros do not have a DC response, such as high pass and band pass systems). If our system has poles p_i and zeros z_k, the normalized transfer function is:

$$F_{\mathrm{n}}(s) = \frac{(-1)^N \prod_{i=1}^{N} p_i}{\prod_{i=1}^{N} (s - p_i)} \cdot \frac{\prod_{k=1}^{M} (s - z_k)}{(-1)^M \prod_{k=1}^{M} z_k} \tag{A2.2.6}$$

From Eq. A2.2.6 the resulting sign factor is obviously:

$$(-1)^N \cdot (-1)^{-M} = (-1)^{N-M} \tag{A2.2.7}$$

which is, incidentally, also equal to $(-1)^{N+M}$, but there is nothing mystical about that, really.

P. Starič, E. Margan

Wideband Amplifiers

Part 3:

Wideband Amplifier Stages With Semiconductor Devices

The only way to find the limits of what is possible
is by pushing towards the impossible!
Arthur C. Clarke

Back To Basics

This part deals with some elementary amplifier configurations which can serve as building blocks of multi-stage amplifiers described in Part 4 and 5, together with the inductive peaking circuits described in Part 2.

Today two schools of thought prevail amongst amplifier designers: the first one (to which most of the more experienced generation belongs) says that cheap operational amplifiers can never fulfill the conflicting requirements of good wideband design; the other (mostly the fresh forces) says that the analog IC production technology advances so fast that, by the time needed to design a good wideband amplifier, the new opamps on the market will render it obsolete.

Both of them are right, of course!

An important point, however, is that very few amplifier designers have a silicon chip manufacturing facility next door. Those who have, often discover that component size reduction solves half of the problems, whilst packing the components close together produces a nearly equal number of new problems.

Another important point is that computer simulation tells you only a part of what will be going on in the actual circuit. Not because there would be anything wrong with the computer, its program or the circuit modeling method used, but because designers, however experienced they are, can not take everything into account right from the start; and, in order to be able to complete the simulation in the foreseeable future, many things are left out intentionally.

A third important point is that by being satisfied with the performance offered by *LEGO*-tronics (playing with general purpose building blocks, as we call it — and we really do not mean anything bad by that!), one intentionally limits oneself to a performance which, in most cases, is an order of magnitude below of what is achievable by the current 'state of the art' technology. Not to speak of there being only a limited amount of experience to be gained by playing just with the outside of those nice little black boxes.

A wise electronics engineer always takes some time to build a discrete model of the circuit (the most critical part, at least) in order to evaluate the influence of strays and parasitics and find a way of improving it. Even if the circuit will eventually be put on a silicon chip those strays will be scaled down, but will not disappear.

That is why we think that it is important to go back to basics.

Contents:

List of Figures:

List of Tables:

3.0 Introduction: *A Farewell to Exact Calculations*

The inductive peaking circuits discussed in Part 2 should be applied to some amplifier stages in order to have any sense. Today the amplifiers are made almost exclusively using semiconductor devices (**bipolar junction transistors — BJTs** and **field effect transistors — FETs**). The number of books describing semiconductor amplifiers is enormous [e.g., Ref. 3.7, 3.12, 3.13, 3.14, 3.15], whilst books discussing wideband amplifiers in necessary depth are quite rare.

Wideband amplifiers with BJTs and FETs are found in a great diversity of electronic products, in measuring instruments and oscilloscopes in particular. They require a different design approach from the 'classical' low frequency amplifier circuits, where the inter-electrode and stray capacitances are less important. In order to improve the wideband performance many special circuits were invented [e.g., Ref. 3.1, Ch.8 and 3.34, Ch.7] and to discuss thoroughly all of them in a book like this would not be possible for several reasons. Here we shall analyze only some basic circuits (the common emitter and the common base amplifiers) and some of the most used wideband circuits (e.g., the differential cascode, the common source follower, and the f_T doubler configurations; later in Part 5 we shall meet several more complex circuits).

In the first section we analyze the common emitter amplifier. Since the base pole ω_h (set by the base spread resistance r_b and the total input capacitance $C_\pi + C_M$), represents the most prevalent bandwidth limit in the transconductance equation, we discuss how to reduce the input capacitance. In this type of amplifier, as well as in all other types that we intend to discuss, the analysis becomes extremely complicated if all the parameters are considered. Since it is our intention to acquire a clear picture of the basic facts, an analysis with all the minute details would needlessly fog the view. The market is still waiting (probably in vain) for a transistor with $\pm 1\,\%$ tolerances in electrical parameters. Also, it is very difficult to specify the stray inductances and capacitances of the wiring with comparable precision. Therefore we shall simplify our expressions and neglect all the parameters which have either little influence or which must be solved numerically for a specific case. After the basic picture is acquired, the reader who wants to make a more precise analysis, can use a computer with a suitable program such as SPICE [Ref. 3.28], PSpice [Ref. 3.29], or MicroCAP [Ref. 3.30], to name a few. Be warned, however, that the result will be as good as allowed by the models of semiconductor devices and, of course, it will be influenced by the user's ability to correctly model the stray components which depend on layout and which are not explicitly shown either in the initial circuit schematic diagram or included in any semiconductor device model.

The nature of the HF impedance in the emitter circuit changes drastically if we look at it from the base, and vice versa. Since it is useful to know the possible transformations from base to emitter circuit and back, we discuss all the transformations of resistive, capacitive, and inductive impedances. Next we analyze the common base circuit. The cascode circuit, which is effectively a combination of a common emitter and a common base configuration, is often used in wideband amplifiers, so it deserves a thorough analysis, too. The same is valid for the differential cascode amplifier. We also discuss the emitter peaking in a cascode amplifier. The invention of the f_T doubler has made possible a bandwidth extension of almost twofold within a single stage and we

discuss it next. This is followed by an analysis of the JFET source follower, which is commonly used as the input stage of oscilloscopes and other measuring instruments. Such a stage can have the input impedance **negative** at high frequencies when the JFET source is loaded by a capacitor (which is always the case), and we show how to compensate this very undesirable property.

In Part 2 we have analyzed the T-coil peaking circuit with a purely capacitive tap to ground impedance. However, if the T-coil circuit is used for a transistor interstage coupling, the tap to ground impedance ceases to be purely capacitive. This fact requires a new analysis, which we deal with in the last section.

Probably, the reader will ask how accurately we need to model the active devices in our circuits to obtain a satisfying approximation. In 1954, *Ebers* and *Moll* [Ref. 3.9] had already described a relatively simple nonlinear model, which, over the years, was improved by the same authors, and by *Gummel* and *Poon* [Ref. 3.10] in 1970. Modern computer programs for circuit simulation allow the user to trade simulation speed for accuracy by selecting models with different levels of complexity (e.g., an older version of MicroCAP [Ref. 3.30] had 3 EM and 2 GP models for the BJT, the most complex GP2 using a total of 51 parameters). To simplify the circuit analysis we shall use the linearized high frequency EM2 model, explained in Sec. 3.1.

All these models look so simple and perform so well, that it seems as if anyone could have created them. Nothing could be farther from the truth. It takes lots of classical physics (*Boltzmann*'s transport theory, *Gauss*' theorem, *Poisson*'s equation, the charge current mean density integral calculus, the complicated p–n junction boundary conditions, *Maxwell*'s equations, ...), as well as quantum physics (*Fermi* levels, *Schrödinger*'s equation, the *Pauli* principle, charge generation, injection, recombination and photon–electron and phonon–electron scattering phenomena, to name just a few important topics) to be judiciously applied in order to find well defined special cases and clever approximations that would, within limits, provide a model simple enough for everyday use. Of course, if pushed too far the model fails, and there is no other way to the solution but to rework the physics neglected. In our analysis we shall try not to go that far.

It cannot be overstressed that in our analysis we are dealing with **models** of semiconductor devices! As *Philip Darrington*, former Wireless World editor, put it in one of his editorials, *"the map is not the territory"*, just as the schematic diagram is not the circuit. As in any branch of science, we build a (theoretical) model, analyze it, and then compare with the real thing; if it fits, we have had a good nose there, or perhaps we have simply been lucky; if it does not fit we go back to the drawing board.

In the macroscopic world, from which all our experience arises, most models are quite simple, since the size ratio of objects, which can still be perceived directly with our senses, to the atomic size, where some odd phenomena begin to show up, is some 6 orders of magnitude; thus the world appears to us to be smooth and continuous. However, in the world of ever shrinking semiconductor devices we are getting ever closer to the quantum phenomena (e.g., the dynamic range of our amplifiers is limited by thermal noise, which is ultimately a quantum effect). But long before we approach the quantum level we should stay alert: even if we forget that the oscilloscope probe loads our circuit with a shunt capacitance of some 10–20 pF and a serial inductance of about 70–150 nH of the ground lead, the circuit will not forget, and sometimes not forgive, either!

3.1 Common Emitter Amplifier

Fig. 3.1.1a shows a common emitter amplifier (the name reflects that the emitter is the common reference for both input and output signals), whilst Fig. 3.1.1b represents its small signal equivalent circuit (the EM2 model, see [Ref. 3.4 and 3.9]). If a signal source amplitude is much smaller than the base bias voltage, resulting in an output signal small enough compared to the supply voltage, we can assume that all the equivalent circuit components are linear (not changing with the signal). However, when the transistor has to handle large signals, the equivalent model becomes rather complicated and the analysis is usually carried out by a suitable computer program.

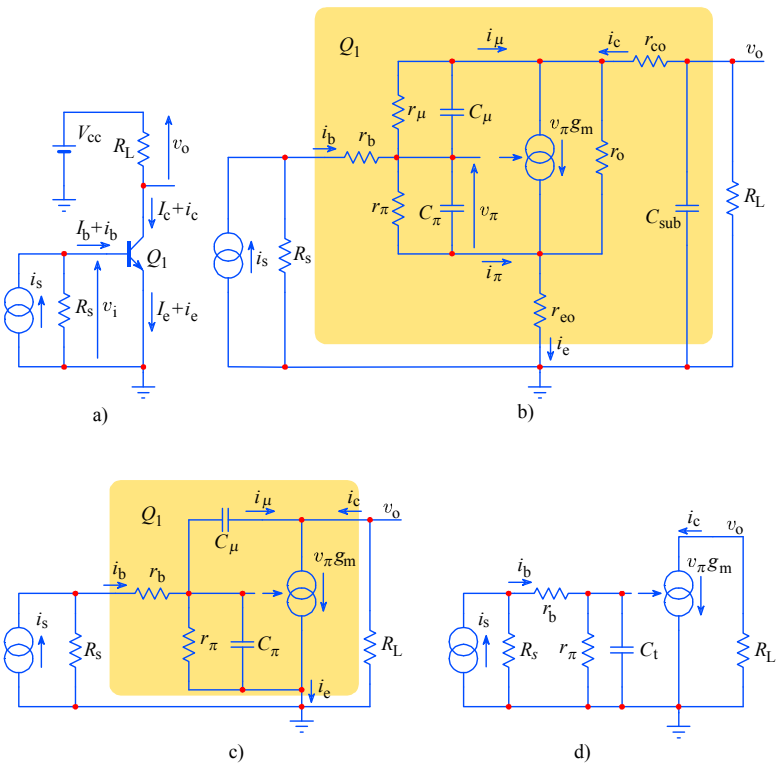


Fig. 3.1.1: The common emitter amplifier: a) circuit schematic diagram – the current and voltage vectors are drawn for the NPN type of transistor; b) high frequency small signal equivalent circuit; the components included in the Q_1 model are explained in the text; c) simplified equivalent circuit; d) an oversimplified circuit where $C_t = C_\pi + AC_\mu = \text{constant}$.

During the first steps of circuit design we can usually neglect the nonlinear effects and obtain a satisfactory performance description by using a first order approximation of the transistor model, Fig. 3.1.1c. Some of the circuit parameters can even be estimated by an oversimplified model of Fig. 3.1.1d. By assuming a certain

operating point OP, set by the DC bias conditions, we can explain the meaning of the model components. On the basis of these explanations it will become clear why and when we may neglect some of them, thus simplifying the model and its analysis.

g_m transistor mutual conductance in *siemens* [S = 1/Ω]: $g_m = \dfrac{i_c}{v_\pi} \approx 1/r_e$

where i_c is the instantaneous collector current;

v_π is the voltage across r_π (see below);

and r_e is the dynamic emitter resistance: $r_e = \dfrac{V_T}{I_e}$ [Ω] (*ohm*);

I_e is the d.c. emitter current [A] (*ampere*);

V_T is the p-n junction 'thermal' voltage $= \dfrac{k_B\, T}{q}$ [V] (*volt*);

where q is the electron charge $= 1.602 \cdot 10^{-19}$ [C] (*coulomb* = [As]);

T is the absolute temperature [K] (*kelvin*);

k_B is the Boltzmann constant $= 1.38 \cdot 10^{-23}$ [J/K] (*joule/kelvin*) (= [VAs/K]).

r_o equivalent collector–emitter output resistance, representing the variation of the collector–emitter potential due to collector signal current at a specified operating point (its value is nearly always much greater than the load):

$$r_o = \frac{v_{ce}}{i_c} = \frac{d\,V_{ce}}{d\,I_c}\bigg|_{OP} = \frac{V_A + V_{ce}}{I_c}$$

where V_A is the *Early* voltage (see Fig. 3.4.11);

In wideband amplifiers: $r_o \gg R_L$

r_μ collector to base resistance, representing the variation of the collector to base potential due to base signal current at some specified DC operating point condition OP(V_{cc}, I_b); in wideband amplifiers its value is always much larger than the source resistance or the base spread resistance:

$$r_\mu = \frac{v_{cb}}{i_b} = \frac{d\,V_{cb}}{d\,I_b}\bigg|_{OP} = \frac{V_A + V_{ce}}{I_b}$$

$$r_\mu \gg R_s + r_b$$

r_π base to emitter input resistance (forward biased BE junction), representing the variation of the base to emitter potential owed to the base signal current at a specified operating point:

$$r_\pi = \frac{v_\pi}{i_b} = \frac{d\,V_{be}}{d\,I_b}\bigg|_{OP} = \beta\,\frac{V_T}{I_c} = \frac{\beta}{g_m} \approx \beta\, r_e$$

r_b base spread resistance (resistance between the base terminal and the base–emitter junction); value range: $10\,\Omega < r_b < 200\,\Omega$.

r_{co} presumably constant collector resistance of internal bonding and external lead; approximate value $< 1\,\Omega$.

r_{eo} presumably constant emitter resistance of internal bonding and external lead; approximate value $< 1\,\Omega$.

C_μ collector–base (reverse biased) junction capacitance; usually the value is small ($<10\,\text{pF}$); however, in the common emitter circuit it is effectively amplified by the voltage gain (the *Miller* effect, see Eq. 3.1.9).

C_π base–emitter junction effective capacitance; depends on DC bias:

$$C_\pi = g_m\,\tau_T - C_\mu$$

where τ_T is the characteristic time constant, derived from the 'transition' frequency, f_T, the frequency at which $\beta(f_T) = 1$ (see below):

$$\tau_T = \frac{1}{\omega_T} = \frac{1}{2\pi f_T}$$

C_{sub} collector to substrate capacitance; must be accounted for only in integrated circuits; in discrete devices it can be neglected.

β the transistor current gain, the collector to base current ratio; the gain frequency dependence is modeled by the characteristic time constant τ_T:

$$\beta = \frac{i_c}{i_b} = \beta_0\,\frac{1}{1+\beta_0\,s\,\tau_T} \qquad \text{where:} \qquad \beta_0 = \frac{I_c}{I_b}$$

It is important to realize that some of those resistances are only 'equivalent resistances', or 'incremental resistances', which can be represented by a tangent to the voltage–current relation curvature at a particular bias point, and as such they are highly nonlinear for large signals. Also, the capacitances are only in part a consequence of the actual p–n junction geometry; they are dominantly volumes in the semiconductor material in which there are energy gaps capable of charge trapping, storing and releasing. In turn the gap energy is voltage dependent, so the effective capacitances are also voltage dependent and therefore also nonlinear.

With bias conditions encountered in wideband amplifier applications, the collector to base resistance r_μ and the collector to emitter resistance r_o are several orders of magnitude larger than the source resistance R_s and the load resistance R_L. In order to simplify the analysis we shall neglect r_μ and r_o; that is why they have not been drawn in Fig. 3.1.1c.

Likewise the DC power supply voltages are also not drawn, because their sources (should!) represent a short circuit for the signal, so we have simply tied the loading resistor and the bias voltages to the ground. Remember that it is the duty of the circuit designer — that is you, yourself! — to provide good power supply bypassing by both adding appropriate capacitors and using wide and short, low resistance, low inductance PCB traces.

The resistors r_{co} and r_{eo} represent the external leads and internal wires, and since their value is usually less than $1\,\Omega$, they are also neglected.

In general we can assume that small signal transistors work at an internal junction temperature of 300 K (27 °C or 80 °F, roughly the 'room temperature' increased by a few degrees owing to some small power dissipation caused by DC bias). Output

transistors work at higher temperatures, depending on the output signal amplitude, the load and the power efficiency of the amplifying stage.

It is interesting to note that $k_B T/q$ has the dimension of voltage (J=VAs, C=As and K cancels) and it has been named the junction 'thermal' voltage, its value $V_T = k_B\,T/q = 26\,\text{mV}$ at $T = 300\,\text{K}$. By assuming that the collector current I_c is almost equal to the emitter current I_e (actually $\times\,\beta/(\beta+1)$), we obtain a well known relation for the effective emitter resistance:

$$r_e = \frac{V_T}{I_e} = \frac{26\,\text{mV}}{I_e} \approx \frac{26\,\text{mV}}{I_c} \tag{3.1.1}$$

The collector to base capacitance C_μ depends on the collector to base voltage V_{cb}. In normal operating conditions (CB junction reverse biased), the corresponding relation [Ref. 3.4, 3.20] is:

$$C_\mu(V_{cb}) = \frac{C_{\mu 0}}{\left[1 + \dfrac{V_{cb}}{V_{jc}}\right]^{m_c}} \tag{3.1.2}$$

This equation is valid under the assumption that there is no charge in the collector to base depletion layer. The meaning of the symbols are:

$C_{\mu 0}$ = junction capacitance [F] (*farad*) (when $V_{cb} = 0$ V)

V_{cb} = collector to base voltage [V] (*volts*)

V_{jc} = collector to base barrier potential $\simeq$ 0.6–0.8 V for silicon transistors

m_c = collector voltage potential gradient factor
(0.5 for abrupt junctions and 0.33 for graded junctions)

Obviously, C_μ decreases inversely with collector voltage. For small signals (amplifier input stage) V_{cb} does not change very much, so we can assume C_μ to be constant, or, in other words, 'linear'. However, in middle stage and output transistors, V_{cb} changes considerably. As already mentioned, in such cases the computer aided circuit simulation is mandatory (after the initial linearized approximation has been found acceptable). Fortunately most transistor manufacturers provide the diagrams showing the dependence of C_μ from V_{cb}. The reader can find a very good review for 25 of the most commonly used transistors in [Ref. 3.21, p. 556].

The input base emitter capacitance C_π strongly depends on the emitter current, respectively, on the transconductance g_m. Since we can not directly access the internal base junction to measure C_π and C_μ separately, we calculate C_π from the total equivalent input capacitance C_t (see Fig. 3.1.1d), from which we first subtract C_μ:

$$C_\pi = C_t - C_\mu = g_m\,\tau_T - C_\mu = \frac{1}{2\pi f_T\,r_e} - C_\mu \tag{3.1.3}$$

where f_T is the frequency at which the (frequency dependent) current amplification factor $\beta = i_c/i_b$ is reduced to 1. Because C_μ is usually small compared to C_t we can simplify Eq. 3.1.3 to obtain:

$$C_\pi \approx \frac{1}{2\,\pi f_T\, r_e} \tag{3.1.4}$$

The product $r_e\, C_\pi$ is called the *transistor time constant* $\tau_T = 1/\omega_T$, where $\omega_T = 2\,\pi\, f_T$. The value $s_1 = -\omega_T = -1/(r_e C_\pi)$ represents the dominant pole of the amplifier and thus it is the main bandwidth limiting factor. In our further discussions we shall find the way to drastically reduce the influence of C_π, at the expense of the amplification factor.

The next problem is to calculate the input impedance. Here we must consider the *Miller effect* [Ref. 3.7, 3.12] owed to capacitance C_μ (in practice, there is also a CB leads stray capacitance, parallel to the junction capacitance, that has to be taken into account). Therefore we first calculate the input admittance looking right into the internal $r_b C_\mu$ junction in Fig. 3.1.1c. The current i_μ flowing through C_μ is:

$$i_\mu = (v_\pi - v_o)\, C_\mu\, s \tag{3.1.5}$$

where v_π is the voltage across r_π. The output voltage is[1]:

$$v_o = -\, i_c R_L = -\, g_m\, v_\pi\, R_L \tag{3.1.6}$$

By inserting Eq. 3.1.6 into Eq. 3.1.5, we obtain:

$$i_\mu = (1 + g_m\, R_L)\, s\, C_\mu\, v_\pi \tag{3.1.7}$$

and from this the junction input admittance:

$$\frac{i_\mu}{v_\pi} = (1 + g_m\, R_L)\, C_\mu\, s \tag{3.1.8}$$

From this equation it follows that the junction input impedance, owed to capacitance C_μ, is a capacitance with the value:

$$C_M = (1\ + g_m\, R_L)\, C_\mu = (1 + A_v)\, C_\mu \tag{3.1.9}$$

where $g_m\, R_L = A_v = v_o/v_i$, the voltage gain.

The capacitance C_M is called the *Miller capacitance*, after the *Miller effect* — bandwidth reduction with increasing voltage gain (the Miller effect is probably named after *John Milton Miller*, [Ref. 3.36]).

[1] Note the negative sign in Eq. 3.1.6: actually, the output voltage is $v_o = V_{cc} - i_c R_L$. However, since we have agreed to replace the supply voltage with a short circuit, we are left with the negative part only.

When the voltage gain is large the effect of C_μ (and, respectively, C_M) becomes prevalent. However, there are ways of reducing the effect of $A_\mathrm{v}\,C_\mu$ (by lowering the voltage gain or by using other circuit configurations); we discuss it in later sections.

The complete input impedance that the signal source would see at the base is:

$$Z_\mathrm{b} = r_\mathrm{b} + \frac{1}{\dfrac{1}{r_\pi} + sC_\pi + sC_\mathrm{M}} = r_\mathrm{b} + \frac{r_\pi}{1 + s(C_\pi + C_\mathrm{M})\,r_\pi} \tag{3.1.10}$$

3.1.1 Calculation of voltage amplification (based on Fig. 3.1.1d)

On the basis of the analysis made so far, we come to the conclusion that the two capacitances C_π and C_μ (effectively C_M) are connected in parallel in the base circuit. We can simply add them together and consider their sum to be C_t. This has been drawn in Fig. 3.1.1d. This equivalent circuit is appropriate for the calculation of both input impedance and the transimpedance. But since we have removed any connection between the output and input (where C_μ is effective), this circuit **is not suitable** for the calculation of output impedance. Therefore when calculating the voltage gain we must also consider the pole $s_2 \simeq -1/R_\mathrm{L}C_\mu$ on the collector side, according to Fig. 3.1.1c (in general, some collector to ground stray capacitance must be added to C_μ, but for the time being, we shall write only C_μ).

According to Fig. 3.1.1d, thus neglecting the pole s_2, but including the source impedance R_s, we have:

$$\frac{v_\mathrm{i} - v_\pi}{R_\mathrm{s} + r_\mathrm{b}} = v_\pi\left(\frac{1}{r_\pi} + s\,C_\mathrm{t}\right) \tag{3.1.11}$$

From this we can express v_π as:

$$v_\pi = v_\mathrm{i}\frac{r_\pi}{r_\pi + R_\mathrm{s} + r_\mathrm{b} + s\,C_\mathrm{t}\,r_\pi(R_\mathrm{s} + r_\mathrm{b})} \tag{3.1.12}$$

The output voltage is:

$$v_\mathrm{o} = -\,g_\mathrm{m}\,v_\pi\,R_\mathrm{L} = -\,g_\mathrm{m}\,R_\mathrm{L}\frac{r_\pi}{r_\pi + R_\mathrm{s} + r_\mathrm{b} + s\,C_\mathrm{t}\,r_\pi(R_\mathrm{s} + r_\mathrm{b})}\,v_\mathrm{i} \tag{3.1.13}$$

and the voltage amplification is:

$$A_\mathrm{v} = \frac{v_\mathrm{o}}{v_\mathrm{i}} = -\,g_\mathrm{m}\,R_\mathrm{L}\frac{r_\pi}{r_\pi + R_\mathrm{s} + r_\mathrm{b} + s\,C_\mathrm{t}\,r_\pi(R_\mathrm{s} + r_\mathrm{b})} \tag{3.1.14}$$

We would like to separate the frequency dependent part of A_v from the frequency independent part, in a normalized form, as:

$$A_\mathrm{v}(s) = A_0\,\frac{-\,s_1}{s - s_1} \tag{3.1.15}$$

where A_0 is the DC gain and s_1 is the system's pole.

To achieve this separation we first divide both the numerator and the denominator of Eq. 3.1.14 by $C_t\, r_\pi (R_s + r_b)$:

$$A_v = -\, g_m\, R_L\, \frac{-\,\dfrac{r_\pi}{C_t\, r_\pi (R_s + r_b)}}{s - \left(-\,\dfrac{r_\pi + R_s + r_b}{C_t\, r_\pi (R_s + r_b)}\right)} \tag{3.1.16}$$

To make the two ratios equal we must multiply the numerator by $(r_\pi + R_s + r_b)/r_\pi$ and then multiply the whole by the reciprocal:

$$A_v = -\, g_m\, R_L\, \frac{-\,\dfrac{r_\pi}{C_t\, r_\pi (R_s + r_b)} \cdot \dfrac{r_\pi + R_s + r_b}{r_\pi}}{s - \left(-\,\dfrac{r_\pi + R_s + r_b}{C_t\, r_\pi (R_s + r_b)}\right)} \cdot \frac{r_\pi}{r_\pi + R_s + r_b} \tag{3.1.17}$$

We rearrange this to obtain:

$$A_v = -\,\frac{g_m\, R_L\, r_\pi}{r_\pi + R_s + r_b} \cdot \frac{-\,\dfrac{r_\pi + R_s + r_b}{C_t\, r_\pi (R_s + r_b)}}{s - \left(-\,\dfrac{r_\pi + R_s + r_b}{C_t\, r_\pi (R_s + r_b)}\right)} \tag{3.1.18}$$

and by comparing this with Eq. 3.1.15 the DC gain is:

$$A_0 = -\,\frac{g_m\, R_L\, r_\pi}{r_\pi + R_s + r_b} \tag{3.1.19}$$

and the pole s_1 is:

$$s_1 = -\,\frac{R_s + r_b + r_\pi}{(R_s + r_b)\, r_\pi\, C_t} \tag{3.1.20}$$

Since s_1 is a simple real pole it is equal to the system's upper half power frequency:

$$\begin{aligned} \omega_h = |s_1| &= \frac{R_s + r_b + r_\pi}{(R_s + r_b)\, r_\pi} \cdot \frac{1}{C_t} \\ &= \left(\frac{1}{R_s + r_b} + \frac{1}{r_\pi}\right) \cdot \frac{1}{C_\pi + (1 + g_m\, R_L)\, C_\mu} \end{aligned} \tag{3.1.21}$$

and it can be seen that it is **inversely proportional to all the components**: R_s, r_b, r_π, C_π, C_μ, g_m, and R_L.

If $R_s + r_b \ll r_\pi$ and if R_L is very small the approximate value of the pole is:

$$|s_1| \approx \frac{1}{r_\pi\, C_\pi} = \frac{1}{\beta_0} \cdot \frac{g_m}{C_\pi} = \frac{\omega_T}{\beta_0} \tag{3.1.22}$$

where $\beta_0 = I_c / I_b$, the DC current amplification factor.

Before more sophisticated circuits were invented, the common emitter amplifier was used extensively (with many amplifier designers having hard times and probably cursing both C_π and C_μ). In 1956 *G. Bruun* [Ref. 3.22] thoroughly analyzed this type of amplifier with the added shunt–series inductive peaking circuit. In view of modern wideband amplifier circuits, this reference is only of historical value today. Nevertheless, the common emitter stage represents a good starting point for the discussion of more efficient wideband amplifier circuits.

3.2 Transistor as an Impedance Converter

In the previous section we have realized that the amplification factor is frequency dependent, decreasing with frequency above some upper frequency limit (asymptotically to $-20\,\mathrm{dB}/10f$, just like a first-order low pass system). This can help us to derive different impedance transformations from the emitter to the base circuit and back [Ref. 3.1, 3.2]. Knowing the possible transformations is extremely useful in the wideband amplifier design. We are going to show how the nature of the impedance changes with these transformations. A capacitive impedance may become inductive and positive impedances may occasionally become negative!

3.2.1 Common base small signal transistor model

As we explained in Sec. 3.1, if the voltage gain is not too high the base emitter capacitance C_π is the dominant cause of the frequency response rolling off at high frequencies. By considering this we can make a simplified small signal high frequency transistor model, as shown in Fig. 3.2.1, for the common base configuration, where i_c, i_e and i_b are the collector-, emitter-, and base-current respectively. For this figure the DC current amplification factor is:

$$\alpha_0 = \frac{I_\mathrm{c}}{I_\mathrm{e}} \tag{3.2.1}$$

Initially we have written $g_\mathrm{m} \approx 1/r_\mathrm{e}$. A better expression for mutual conductance is:

$$g_\mathrm{m} = \frac{\alpha_0}{r_\mathrm{e}} = \frac{\beta_0}{(1+\beta_0)\,r_\mathrm{e}} \tag{3.2.2}$$

where β_0 is the common emitter DC current amplification factor. If $\beta_0 \gg 1$ then $\alpha_0 \simeq 1$, so the collector current I_c is almost equal to the emitter current I_e, and $g_\mathrm{m} \simeq 1/r_\mathrm{e}$. This simplification is often used in practice.

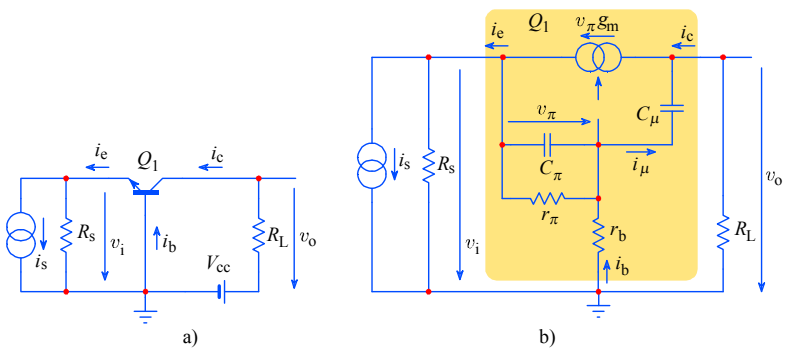


Fig. 3.2.1: The common base amplifier: a) circuit schematic diagram; b) high frequency small signal equivalent circuit.

For the moment let us assume that the base resistance $r_b = 0$ and consider the low frequency relations. The input resistance is:

$$r_\pi = \frac{v_\pi}{i_b} \tag{3.2.3}$$

where v_π is the base to emitter voltage. Since the emitter current is:

$$i_e = i_b + i_c = i_b + \beta_0\, i_b = i_b\,(1 + \beta_0) \tag{3.2.4}$$

then the base current is:

$$i_b = \frac{i_e}{1 + \beta_0} \tag{3.2.5}$$

and consequently:

$$r_\pi = \frac{v_\pi\,(1 + \beta_0)}{i_e} = r_e\,(1 + \beta_0) \approx \beta_0\, r_e \tag{3.2.6}$$

The last simplification is valid if $\beta_0 \gg 1$. To obtain the input impedance at high frequencies the parallel connection of C_π must be taken into account:

$$Z_b = \frac{(1 + \beta_0)\, r_e}{1 + (1 + \beta_0)\, s\, C_\pi\, r_e} \tag{3.2.7}$$

The base current is:

$$i_b = \frac{v_\pi}{Z_b} = v_\pi\, \frac{1 + (1 + \beta_0)\, s\, C_\pi\, r_e}{(1 + \beta_0)\, r_e} \tag{3.2.8}$$

Therefore v_π is:

$$v_\pi = i_b\, \frac{(1 + \beta_0)\, r_e}{1 + (1 + \beta_0)\, s\, C_\pi\, r_e} \tag{3.2.9}$$

Consequently, the collector current is:

$$i_c = g_m\, v_\pi = \frac{\beta_0}{1 + \beta_0} \cdot \frac{1}{r_e}\, v_\pi = \frac{\alpha_0}{r_e}\, v_\pi \tag{3.2.10}$$

If we put Eq. 3.2.9 into Eq. 3.2.10 we obtain:

$$\begin{aligned} i_c &= i_b\, \frac{\beta_0}{1 + \beta_0} \cdot \frac{1}{r_e} \cdot \frac{(1 + \beta_0)\, r_e}{1 + s\,(1 + \beta_0)\, r_e\, C_\pi} \\ &= i_b\, \frac{1}{\dfrac{1}{\beta_0} + s \left(\dfrac{\beta_0 + 1}{\beta_0} \right) r_e\, C_\pi} \\ &\approx i_b\, \frac{1}{\dfrac{1}{\beta_0} + s\, \tau_T} = i_b\, \beta(s) \end{aligned} \tag{3.2.11}$$

In the very last expression we assumed that $\beta_0 \gg 1$ and $\tau_T = r_e\, C_\pi = 1/\omega_T$, where $\omega_T = 2\pi f_T$ is the angular frequency at which the current amplification factor β decreases to unity. The parameter τ_T, and consequently ω_T, depend on the internal configuration and structure of the transistor. Fig. 3.2.2 shows the frequency dependence of β and the equivalent current generator.

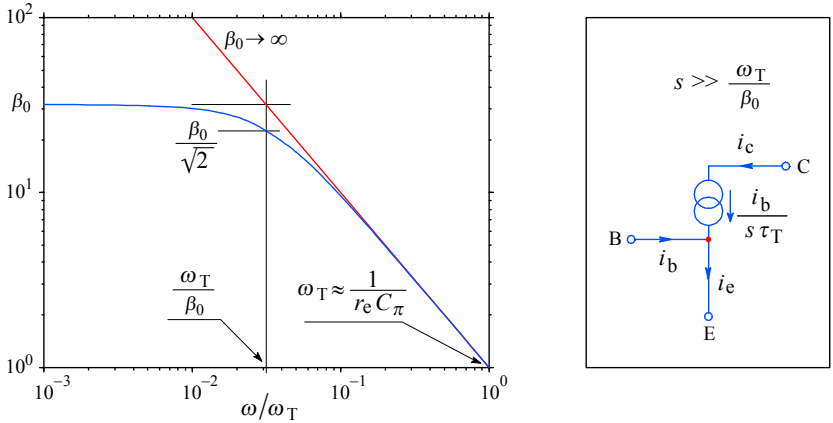


Fig. 3.2.2: a) The transistor gain as a function of frequency, modeled by the Eq. 3.2.11; b) the equivalent HF current generator.

In order to correlate Fig. 3.2.2 with Eq. 3.2.11 we rewrite it as:

$$\frac{i_c}{i_b} \approx \beta_0 \frac{-\frac{\omega_T}{\beta_0}}{s - \left(-\frac{\omega_T}{\beta_0}\right)} = \beta_0 \frac{-s_1}{s - s_1} \tag{3.2.12}$$

where s_1 is the pole at $-\omega_T/\beta_0$. This relation will become useful later, when we shall apply one of the peaking circuits (from Part 2) to the amplifier. At very high frequencies, or if $\beta_0 \gg 1$, the term $s\,\tau_T$ prevails. In this case, from Eq. 3.2.11:

$$\frac{i_c}{i_b} = \beta(s) \approx \frac{1}{s\,\tau_T} = \frac{1}{j\,\omega\,r_e\,C_\pi} \tag{3.2.13}$$

Obviously β is decreasing with frequency. By definition, at $\omega = \omega_T$ the current ratio $i_c/i_b = 1$; then the capacitance C_π can be found as:

$$C_\pi \approx \frac{1}{\omega_T\,r_e} \tag{3.2.14}$$

This simplified relation represents the $-20\,\mathrm{dB}/10f$ asymptote in Fig. 3.2.2a.

3.2.2 The conversion of impedances

We can use the result of Eq. 3.2.11 to transform the transistor internal (and the added external) impedances from the emitter to the base circuitry, and vice versa. Suppose we have the impedance Z_e in the emitter circuit, as displayed in Fig. 3.2.3a, and we are interested in the corresponding base impedance Z_b:

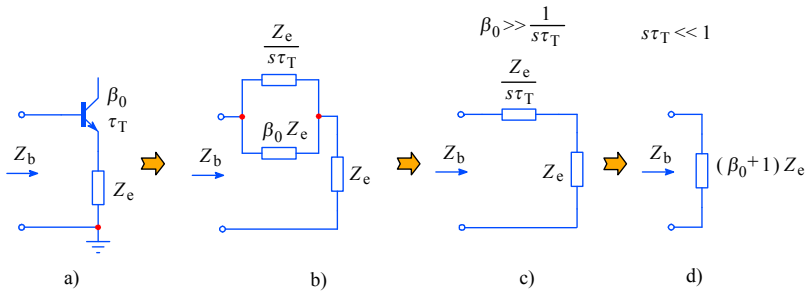


Fig. 3.2.3: Emitter to base impedance conversion: a) schematic; b) equivalent circuit; c) simplified for high β_0; d) simplified for low frequencies.

We know that:

$$Z_b = \beta(s)\, Z_e + Z_e = \left[\beta(s) + 1\right] Z_e \tag{3.2.15}$$

If we insert $\beta(s)$ according to Eq. 3.2.11, we obtain:

$$Z_b = \frac{Z_e}{\dfrac{1}{\beta_0} + s\,\tau_T} + Z_e \tag{3.2.16}$$

The admittance of the first part of this equation is:

$$Y = \frac{\dfrac{1}{\beta_0} + s\,\tau_T}{Z_e} = \frac{1}{\beta_0\, Z_e} + \frac{s\,\tau_T}{Z_e} \tag{3.2.17}$$

and this represents a parallel connection of impedances $\beta_0\, Z_e$ and $Z_e / s\,\tau_T$. By adding the series impedance Z_e, as in Eq. 3.2.16, we obtain the equivalent circuit of Fig. 3.2.3b. At medium frequencies and with a high value of β_0 we can assume that $1/\beta_0 \ll s\,\tau_T$, so we can delete the impedance $\beta_0 Z_e$ and simplify the circuit, as in Fig. 3.2.3c. On the other hand, at low frequencies, where $s\,\tau_T \ll 1$, we can neglect the $Z_e / s\,\tau_T$ component and get a very basic equivalent circuit, displayed in Fig. 3.2.3d.

Eq. 3.2.11 is equally useful when we want to transform the impedance from the base into the emitter circuit as shown in Fig. 3.2.4a. In this case we have:

$$Z_e = \frac{Z_b}{\beta(s) + 1} \tag{3.2.18}$$

Again we calculate the admittance, which is:

$$Y_e = \frac{\beta(s)+1}{Z_b} = [\beta(s)+1]\,Y_b = \beta(s)\,Y_b + Y_b \tag{3.2.19}$$

The first part of this admittance is:

$$Y = \frac{\beta(s)}{Z_b} = \frac{Y_b}{\frac{1}{\beta_0} + s\,\tau_T} = \frac{1}{Z_b} \cdot \frac{1}{\frac{1}{\beta_0} + s\,\tau_T} \tag{3.2.20}$$

and the impedance is:

$$Z = \frac{Z_b}{\beta_0} + s\,\tau_T\,Z_b \tag{3.2.21}$$

Thus the transformed impedance Z_e is composed of three elements: the series connection of Z_b/β_0 and $s\,\tau_T\,Z_b$, in parallel with the impedance Z_b.

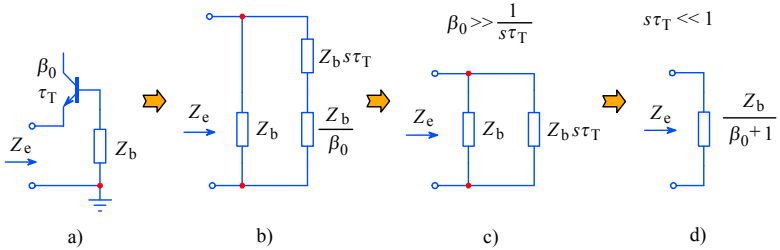


Fig. 3.2.4: Base to emitter impedance conversion: a) schematic; b) equivalent circuit; c) simplified for high β_0 or for $f \simeq f_T$; d) simplified for low frequencies.

The equivalent emitter impedance is shown in Fig. 3.2.4b.

As in the previous example, for some specific conditions the circuit can be simplified. At medium frequencies and a high β_0, we can assume $\beta_0 \gg 1/s\,\tau_T$ and therefore neglect the impedance Z_b/β_0, as in Fig. 3.2.4c. At low frequencies, where $s\,\tau_T \ll 1$, the impedance $Z_b/(\beta_0+1)$ prevails and we can neglect the parallel impedance Z_b, as in Fig. 3.2.4d.

3.2.3 Examples of impedance transformations

The most interesting examples of impedance transformations are the emitter to base transformation of a capacitive emitter impedance and the base to emitter transformation of an inductive base impedance.

In the first case we have $Z_e = 1/s\,C$, where C is the emitter to ground capacitance.

To obtain the base impedance we apply Eq. 3.2.5:

$$Z_b = \frac{\beta(s)+1}{s\,C} = \left[\frac{1}{\frac{1}{\beta_0}+s\,\tau_T}+1\right]\frac{1}{s\,C} = \frac{\frac{1}{\beta_0}+s\,\tau_T+1}{\left(\frac{1}{\beta_0}+s\,\tau_T\right)s\,C}$$

$$= \frac{s\,\tau_T+\left(1+\frac{1}{\beta_0}\right)}{s^2\tau_T\,C+\frac{s\,C}{\beta_0}} \qquad (3.2.22)$$

The inverse of Z_b is the admittance:

$$Y_b = \frac{s^2\tau_T\,C+\frac{s\,C}{\beta_0}}{s\,\tau_T+\left(1+\frac{1}{\beta_0}\right)} \qquad (3.2.23)$$

Let us synthesize this expression by a simple continued fraction expansion [Ref. 3.27]:

$$\frac{s^2\tau_T\,C+\frac{s\,C}{\beta_0}}{s\,\tau_T+\left(1+\frac{1}{\beta_0}\right)} = s\,C-\frac{s\,C}{s\,\tau_T+\left(1+\frac{1}{\beta_0}\right)} \qquad (3.2.24)$$

The fraction on the right is a negative admittance with the corresponding impedance:

$$Z_b' = -\frac{s\,\tau_T+\left(1+\frac{1}{\beta_0}\right)}{s\,C} = -\frac{\tau_T}{C}-\frac{1+\frac{1}{\beta_0}}{s\,C} \qquad (3.2.25)$$

It is evident that this impedance is a series connection of a negative resistance:

$$R_n = -\frac{\tau_T}{C} = -r_e\frac{C_\pi}{C} \qquad (3.2.26)$$

and a negative capacitance:

$$C_n = -\frac{C}{1+\frac{1}{\beta_0}} = -\frac{\beta_0}{1+\beta_0}C = -\alpha_0\,C \qquad (3.2.27)$$

By adding the positive parallel capacitance C, as required by Eq. 3.2.24, we obtain the equivalent circuit which is shown in Fig. 3.2.5. Since we deal with an active circuit (transistor) it is quite normal to encounter negative impedances. The complete base admittance is then:

$$Y_b = s\,C-\frac{1}{\frac{\tau_T}{C}+\frac{1}{s\,\alpha_0\,C}} \qquad (3.2.28)$$

By rearranging this expression and substituting $s = j\omega$ we can separate the real and imaginary parts, obtaining:

$$Y_\mathrm{b} = \Re\{Y_\mathrm{b}\} + j\Im\{Y_\mathrm{b}\} = G_\mathrm{b} + j\omega C_\mathrm{b}$$

$$= -\frac{\frac{\tau_\mathrm{T}}{C}}{\tau_\mathrm{T}^2 + \frac{1}{\omega^2\alpha_0^2}} - j\omega C\,\frac{\tau_\mathrm{T}^2 - \frac{\alpha_0 - 1}{\omega^2\alpha_0^2}}{\tau_\mathrm{T}^2 + \frac{1}{\omega^2\alpha_0^2}} \qquad (3.2.29)$$

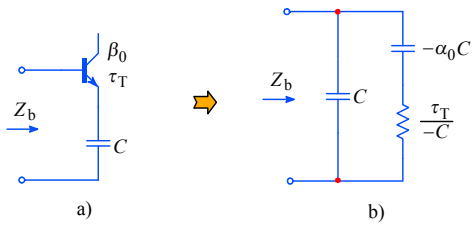


Fig. 3.2.5: A capacitive emitter load is reflected into the base (junction) with **negative** components.

The negative input (base) conductance G_b can cause ringing on steep signals or even continuous oscillations if the signal source impedance has an emphasized inductive component. We shall thoroughly discuss this effect and its compensation later, when we shall analyze the emitter–follower (i.e., common collector) and the JFET source–follower amplifiers.

Now let us derive the emitter impedance Z_e in the case in which the base impedance is inductive ($s\,L$). Here we apply Eq. 3.2.18:

$$Z_\mathrm{e} = \frac{s\,L}{\beta(s)+1} = \frac{s\,L}{\frac{1}{\frac{1}{\beta_0} + s\,\tau_\mathrm{T}} + 1} = \qquad (3.2.30)$$

$$= \frac{s\,L\left(\frac{1}{\beta_0} + s\,\tau_\mathrm{T}\right)}{1 + \frac{1}{\beta_0} + s\,\tau_\mathrm{T}} = \frac{s^2 L\,\tau_\mathrm{T} + \frac{s\,L}{\beta_0}}{s\,\tau_\mathrm{T} + \left(1 + \frac{1}{\beta_0}\right)} \qquad (3.2.31)$$

By continued fraction expansion we obtain:

$$\frac{s^2 L\,\tau_\mathrm{T} + \frac{s\,L}{\beta_0}}{s\,\tau_\mathrm{T} + \left(1 + \frac{1}{\beta_0}\right)} = s\,L - \frac{s\,L}{s\,\tau_\mathrm{T} + \left(1 + \frac{1}{\beta_0}\right)} \qquad (3.2.32)$$

The negative part of the result can be inverted to obtain the admittance:

$$Y'_\mathrm{e} = -\frac{s\,\tau_\mathrm{T} + \left(1 + \frac{1}{\beta_0}\right)}{s\,L} = -\frac{\tau_\mathrm{T}}{L} - \frac{1 + \frac{1}{\beta_0}}{s\,L} \qquad (3.2.33)$$

This means we have two parallel impedances. The first one is a **negative** resistance:

$$R_x = -\frac{L}{\tau_T} \tag{3.2.34}$$

and the second one is a **negative** inductance:

$$L_x = -\frac{L}{1+\frac{1}{\beta_0}} = -\frac{\beta_0}{1+\beta_0} L = -\alpha_0 L \tag{3.2.35}$$

As required by Eq. 3.2.32, we must add the inductance L in series, thus arriving at the equivalent emitter impedance shown in the figure below:

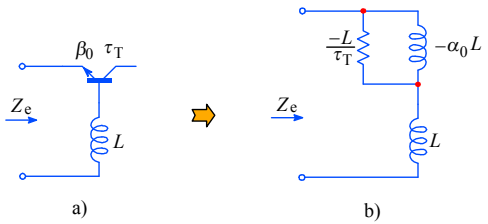


Fig. 3.2.6: The inductive source is reflected into the emitter with **negative** components.

We have just analyzed an important aspect of a common base amplifier, with an inductance (i.e., long lead) between the base and ground. The negative resistance, as given by Eq. 3.2.34, may become the reason for ringing or oscillations if the driving circuit seen by the emitter has a capacitive character. We shall discuss this problem more thoroughly when we shall analyze the cascode circuit.

In a way similar to those used for deriving the previous two results, we can transform other impedance types from emitter to base, and vice versa. The Table 3.2.1 displays the six possible variations and the reader is encouraged to derive the remaining four, which we have not discussed.

Note that all the three transformations for the common base circuit in the table apply to the base–emitter **junction** to ground only. In order to obtain the correct base **terminal** to ground impedance the transistor base spread resistance r_b must be added in series to the circuits shown in the table.

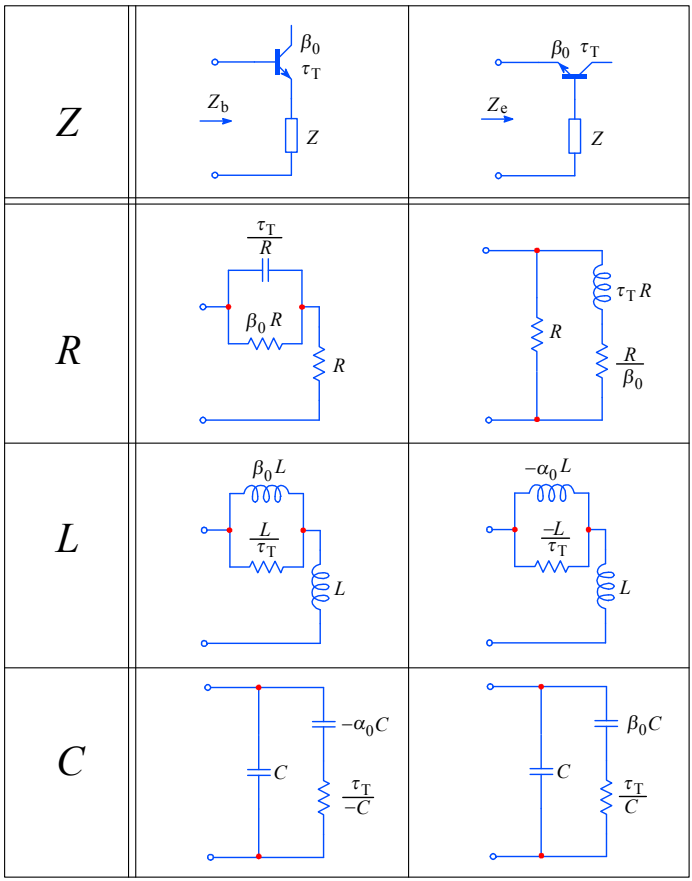


Table 3.2.1: The Table of impedance conversions [Ref. 3.8].

3.2.4 Transformation of combined impedances

The Table 3.2.1 can also help us in transforming impedances consisting of two or more components.

Example 1:

Suppose we have a parallel $R_b C_b$ combination in the base circuit, as shown in Fig. 3.2.7a. What is the emitter impedance Z_e if the common base transistor has a current amplification factor β_0 and the time constant τ_T ($= 1/\omega_T$)?

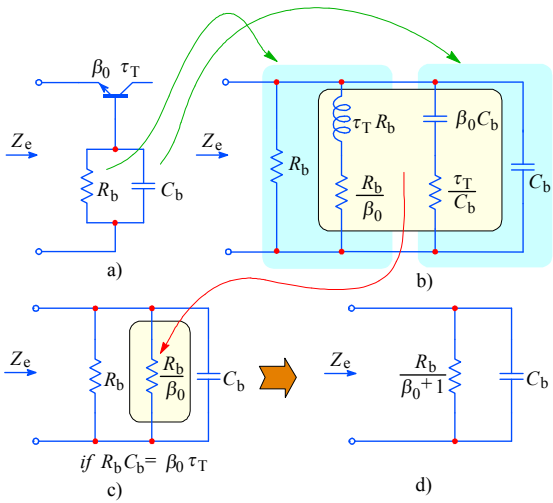


Fig. 3.2.7: Base to emitter RC network transformation: a) schematic; b) equivalent circuit; c) if $R_b C_b = \beta_0 \tau_T$ the middle components form a resistance; d) final equivalent circuit.

From the Table 3.2.1 we first transform the resistance R_b from base to emitter and obtain what is shown on the left half of Fig. 3.2.7b. Then we transform the capacitance C_b and obtain the right half of Fig. 3.2.7b. If we want the transformed network to have the smallest possible influence in the emitter circuit, we can apply the principle of constant resistance network (L and C cancel each other when $RC = L/R$, [Ref. 3.27]). To do so let us focus on both middle branches of the transformed network, where we select such values of R_b and C_b that:

$$\sqrt{\frac{R_b\,\tau_T}{C_b\,\beta_0}} = \frac{R_b}{\beta_0} \tag{3.2.36}$$

which is resolved as:

$$R_b\,C_b = \tau_T\,\beta_0 \tag{3.2.37}$$

With such values of R_b and C_b the middle two branches obtain the form of a resistor with the value R_b/β_0, as shown in Fig. 3.2.7c, which allows us to further simplify the complete circuit to that in Fig. 3.2.7d.

To acquire a feeling for practical values, let us make a numerical example. Our transistor has:

$$\beta_0 = 80 \qquad f_T = 600\,\text{MHz} \qquad R_b = 47\,\Omega$$

where R_b is the external base resistance, as in Fig. 3.2.7.

What should be the value of the capacitance C_b, connected in parallel with R_b, which would fulfill the requirement expressed by Eq. 3.2.36?

We start by calculating the transistor time constant:

$$\tau_T = \frac{1}{\omega_T} = \frac{1}{2\,\pi\,f_T} = \frac{1}{2\,\pi\cdot 600\cdot 10^6} = 265\,\text{ps} \tag{3.2.38}$$

Then we calculate the capacitance by using Eq. 3.2.37:

$$C_b = \frac{\tau_T\,\beta_0}{R_b} = \frac{265\cdot 10^{-12}\cdot 80}{47} = 451\,\text{pF} \tag{3.2.39}$$

The equivalent parallel resistance, R_q, according to Fig. 3.2.7d, is:

$$R_q = \frac{R_b}{1+\beta_0} = \frac{47}{1+80} = 0.58\,\Omega \tag{3.2.40}$$

The time constant of the equivalent circuit, τ_q, is:

$$\tau_q = R_q\,C_b = 0.56\cdot 451\cdot 10^{-12} = 261.58\,\text{ps} \tag{3.2.41}$$

whilst the base time constant is:

$$\tau_b = R_b\,C_b = 47\cdot 451\cdot 10^{-12} = 21.197\,\text{ns} \tag{3.2.42}$$

and the ratio of time constants is:

$$\frac{\tau_b}{\tau_q} = \beta_0 + 1 \tag{3.2.43}$$

We shall consider these results in the design of the common base amplifier and of the cascode amplifier.

Example 2:

By using the same principles as we have used above, we shall take another example, which is also very important for the wideband amplifier design. We shall transform a parallel combination R_eC_e, as shown in Fig. 3.2.8a, from emitter to base. With the data from Table 3.2.1, we can draw the transformed base network separately for R_e and C_e and then connect them in parallel. This is shown in Fig. 3.2.8b. Now we

focus only on the middle part of the circuit, which is drawn in Fig. 3.2.8c. If we select such values of $R_e C_e$ that:

$$R_e C_e = \tau_T \tag{3.2.44}$$

and if we consider that $\alpha_0 \approx 1$, then the admittance Y of the middle part of the circuit becomes zero, because in this case:

$$R_e = -\frac{\tau_T}{C_e} \tag{3.2.45}$$

and:

$$\frac{\tau_T}{R_e} = -\alpha_0 C_e \approx -C_e \tag{3.2.46}$$

and the parallel connection of a positive and an equal negative admittance gives zero admittance:

$$Y = \frac{1}{R_e + \dfrac{1}{s\dfrac{\tau_T}{R_e}}} + \frac{1}{-\dfrac{\tau_T}{C_e} - \dfrac{1}{s\,C_e}} = 0 \Bigg|_{R_e C_e = \tau_T} \tag{3.2.47}$$

A zero admittance is an infinite impedance. So in this case the only components that remain are the parallel connection of C_e and $(\beta_0 + 1)R_e$, as in Fig. 3.2.9d.

Note that this transformation was carried out by taking the **internal base junction** as the viewing point. The actual input impedance at the external base terminal will be equal to the parallel RC combination of Fig. 3.2.8d to which the series connected base spread resistance r_b must be added.

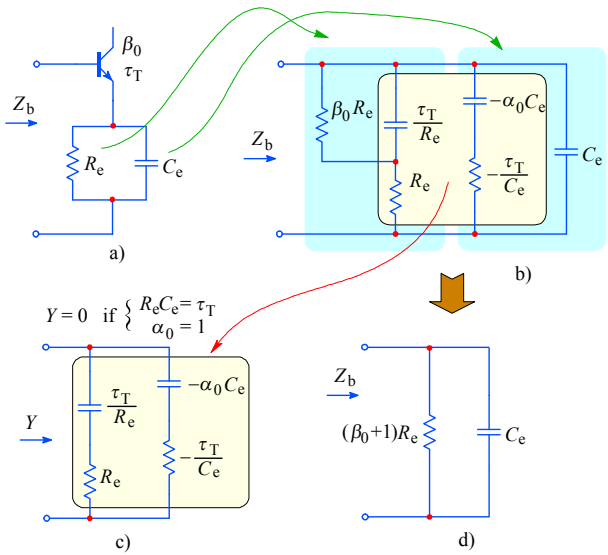


Fig. 3.2.8: Transformation of the emitter RC network as seen from the base: a) schematic; b) equivalent circuit; c) if $RC = \tau_T$ and $\alpha_0 = 1$, the sum of the middle frame component's admittances is zero; d) final equivalent circuit.

The transformation in Fig. 3.2.8 allows us to trade the gain of a common emitter circuit for the reduced input capacitance. Instead of C_π (large) with the emitter grounded, the input capacitance is now equal to the capacitance C_e (small) which we have inserted in the emitter circuit. Of course, we still have to add the base to collector capacitance C_μ or Miller capacitance C_M. As we shall see in the Sec. 3.4, where we shall discuss the cascode amplifier, the gain is reduced in proportion to R_L/R_e. Since in a wideband amplifier stage we almost never exceed the voltage gain of ten, we can always apply the above transformation.

For a numerical example let us use the same transistor as before ($\beta_0 = 80$, $f_T = 600\,\text{MHz}$). According to Eq. 3.2.38 the corresponding τ_T is 265 ps. Let us say that on the basis of the desired current gain of the common-emitter stage we select an emitter resistor $R_e = 100\,\Omega$. What is the value of the parallel emitter capacitance C_e which would give the input impedance according to Fig. 3.2.8d?

Since $R_e C_e = \tau_T = 265\,\text{ps}$, it follows that:

$$C_e = \frac{\tau_T}{R_e} = \frac{256 \cdot 10^{-12}}{100} = 2.65\,\text{pF} \quad \text{only!} \tag{3.2.48}$$

If the stage has an emitter current $I_e = 10\,\text{mA}$, then:

$$r_e = \frac{26\,\text{mV}}{10\,\text{mA}} = 2.6\,\Omega \tag{3.2.49}$$

Without the R_eC_e network in the emitter, the base to emitter capacitance C_π would define the bandwidth and its estimated value would be (Eq. 3.1.4):

$$C_\pi = \frac{1}{2\,\pi\,r_e\,f_T} = \frac{1}{2\,\pi\,\cdot 2.6 \cdot 600 \cdot 10^6} = 102\ \text{pF} \tag{3.2.50}$$

By introducing the $R_e C_e$ network in the emitter circuit, we have reduced the base to emitter capacitance, seen by the input current, by $102/2.65 = 38.5$ times! Of course, to our 2.65 pF we must add in parallel the capacitance $C_M = C_\mu\,(1 + A_v)$ and the base spread resistance r_b in series with this network to obtain a more accurate input impedance. Since now the collector to base capacitance C_μ has become significant, especially because it is 'magnified' $(1 + A_v)$ times, we must look for some methods to reduce its effect as much as possible. We shall discus this in the following sections.

Owing to R_e the input resistance is increased to:

$$R_i \simeq (1 + \beta_0)\,R_e = (1 + 80)\,100 = 8100\,\Omega \tag{3.2.51}$$

Here, too, we have neglected the base resistance r_b; it must be added to the value above to obtain a more accurate figure.

In Fig. 3.2.9a the transistor stage with R_eC_e is shown again and in Fig. 3.2.9b is its small signal equivalent input circuit.

In wideband amplifiers we usually make the emitter network with $C_e \leq 20\,\text{pF}$. In order to match $R_e C_e = \tau_T$ the capacitor C_e is often made adjustable, because τ_T in commercially available transistors have rather large tolerances.

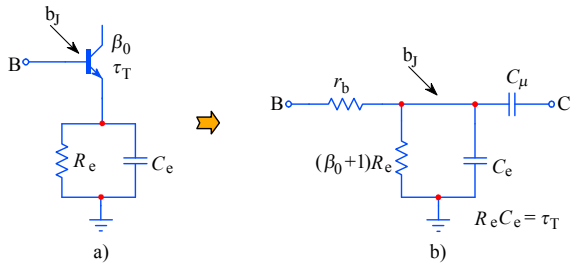


Fig. 3.2.9: R_eC_e network transformation: a) schematic; b) equivalent circuit. The symbol b_J represents the internal base junction.

With an appropriate R_L in the collector (not shown in Fig. 3.2.9) we might now calculate the (decreased) voltage amplification A_v owed to the R_eC_e network in the emitter circuit of the common emitter stage and consider the decreased value of the Miller capacitance C_M. Since we shall not use exactly such amplifier configuration we leave this as an exercise to the reader.

But for the application in the cascode amplifier, which we are going to discuss in Sec. 3.4, it is important to know the transconductance i_o/v_i of the amplifier with the R_eC_e network. The corresponding circuit is drawn again in Fig. 3.2.10a and Fig. 3.2.10b shows the equivalent small signal circuit.

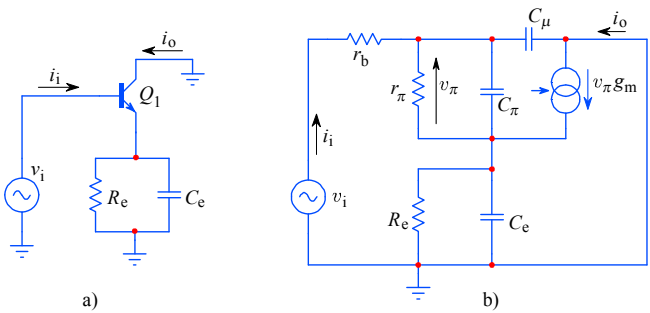


Fig. 3.2.10: Common collector amplifier: a) schematic; b) equivalent small signal circuit.

If we neglect the resistance r_b and the capacitance C_μ the following relation is valid for the remaining circuit:

$$v_i = i_i\,(z_\pi + Z_e) + i_o\,Z_e \tag{3.2.52}$$

where:

$$i_o = g_m\,v_\pi \qquad \text{and} \qquad v_\pi = i_i\,z_\pi$$

therefore:

$$i_o = g_m\,i_i\,z_\pi \tag{3.2.53}$$

The impedances z_π and Z_e are:

$$z_\pi = \frac{r_\pi}{1 + s\,C_\pi\,r_\pi} \qquad \text{and} \qquad Z_e = \frac{R_e}{1 + s\,C_e\,R_e} \tag{3.2.54}$$

We can rewrite Eq. 3.2.52 as:

$$v_i = i_i\,(z_\pi + Z_e + g_m\,z_\pi\,Z_e) \tag{3.2.55}$$

and the input current is:

$$i_i = \frac{v_i}{z_\pi + Z_e + g_m\,z_\pi\,Z_e} \tag{3.2.56}$$

The output current can be obtained by inserting Eq. 3.2.56 back into Eq. 3.2.53:

$$i_o = \frac{g_m\,z_\pi\,v_i}{z_\pi + Z_e + g_m\,z_\pi\,Z_e} \tag{3.2.57}$$

The transadmittance is:

$$\frac{i_o}{v_i} = \frac{g_m\,z_\pi}{z_\pi + Z_e + g_m\,z_\pi\,Z_e} \tag{3.2.58}$$

We can divide the numerator and denominator by $g_m\,z_\pi\,Z_e$:

$$\frac{i_o}{v_i} = \frac{1}{Z_e} \cdot \frac{1}{\dfrac{1}{g_m\,Z_e} + \dfrac{1}{g_m\,z_\pi} + 1} \tag{3.2.59}$$

Now we insert the expressions for z_π and Z_e from Eq. 3.2.54 and replace g_m by $1/r_e$:

$$\frac{i_o}{v_i} = \frac{1}{R_e} \cdot \frac{1 + s\,C_e R_e}{\dfrac{r_e}{R_e}(1 + s\,C_e R_e) + \dfrac{r_e}{r_\pi}(1 + s\,C_\pi r_\pi) + 1} \tag{3.2.60}$$

and with a slight rearrangement we obtain:

$$\frac{i_o}{v_i} = \frac{1}{R_e} \cdot \frac{1 + s\,C_e R_e}{\dfrac{r_e}{R_e} + \dfrac{r_e}{r_\pi} + 1 + s\,(C_e + C_\pi)\,r_e} \tag{3.2.61}$$

Because $(r_e/R_e) \ll 1$ and $(r_e/r_\pi) \ll 1$ we can neglect them, so:

$$\frac{i_o}{v_i} = \frac{1}{R_e} \cdot \frac{1 + s\,C_e R_e}{1 + s\,(C_e + C_\pi)\,r_e} \tag{3.2.62}$$

This equation can be simplified if we 'tune' the emitter network so that:

$$\begin{aligned} & C_e R_e = (C_e + C_\pi)\,r_e \\ \Rightarrow \quad & C_e(R_e - r_e) = C_\pi\,r_e \\ \Rightarrow \quad & C_e R_e \approx C_\pi\,r_e \end{aligned} \tag{3.2.63}$$

The transadmittance can thus be expressed simply as:

$$\boxed{\frac{i_\mathrm{o}}{v_\mathrm{i}} = \frac{1}{R_\mathrm{e}}} \tag{3.2.64}$$

Here we must not forget that at the beginning of our analysis we have neglected the resistance r_b, which, together with the transformed input capacitance C_e and the collector to base capacitance C_μ, makes a pole at:

$$s_1 = -1/(C_\mathrm{e} + C_\mu)\, r_\mathrm{b} \tag{3.2.65}$$

The magnitude of s_1 is equal to the upper half power frequency: $|s_1| = \omega_\mathrm{h}$. This pole makes the stage frequency dependent in spite of Eq. 3.2.64. We have also neglected the input resistance $\beta_0 R_\mathrm{e}$, but, since it is much larger than r_b, we shall not consider its influence (with it, the bandwidth would increase slightly). By introducing the pole s_1 back into Eq. 3.2.64, we obtain a more accurate expression for the transadmittance:

$$\boxed{\frac{i_\mathrm{o}}{v_\mathrm{i}} = \frac{1}{R_\mathrm{e}} \cdot \frac{-\dfrac{1}{(C_\mathrm{e} + C_\mu)\, r_\mathrm{b}}}{s - \left[-\dfrac{1}{(C_\mathrm{e} + C_\mu)\, r_\mathrm{b}}\right]}} \tag{3.2.66}$$

3.3 Common-Base Amplifier

In the previous sections we have realized that the collector to base capacitance C_μ has a very undesirable effect on the stage bandwidth (Miller effect). But in the common base configuration the base is effectively grounded and any current through C_μ is fed to the ground, not affecting the base current (actually, owing to the physical construction of the CB junction C_μ is spread across the whole base resistance r_b, so part of that current would nevertheless reach the base, which we shall analyze later).

The common base circuit is drawn in Fig. 3.3.1a and its small signal equivalent circuit in Fig. 3.3.1b. In wideband amplifiers the loading resistor R_L is much smaller than the collector to base resistance r_μ, so we shall neglect the later. In order to make the expressions still simpler, at the beginning of our analysis we shall also not take into account the base resistance r_b. However, we shall have to include r_b later, when we shall discuss the input impedance.

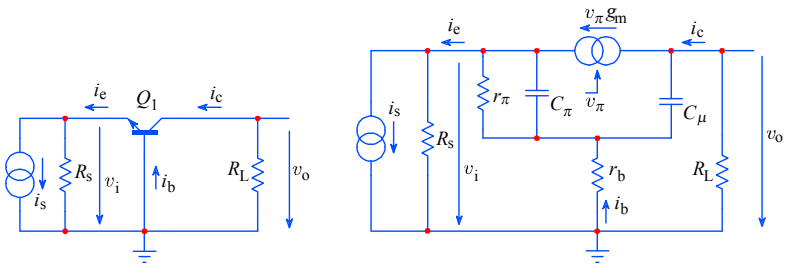


Fig. 3.3.1: Common base amplifier: a) schematic; b) equivalent small signal model.

The main characteristics of the common base stage are a very low input impedance, a very high output impedance, the current amplification factor $\alpha_0 \approx 1$, and, with the correct value of the loading resistor R_L, the possibility of achieving higher bandwidths. The last property is owed to a near elimination of the Miller effect, since C_μ is now grounded and does not affect the input A_v times. Thus C_μ is effectively in parallel with the loading resistor R_L and — because we can make the time constant $R_L C_\mu$ relatively small — the bandwidth of the stage may be correspondingly large.

Another very useful property of the common base stage is that the collector to base breakdown voltage V_{cbo} is highest when the base is connected to ground and the higher reverse voltage reduces C_μ further (Eq. 3.1.2). Owing to all the listed properties the common base stage is used almost exclusively for wideband amplifier stages where large output signals are expected.

Following the current directions in Fig. 3.3.1, the input emitter current is:

$$i_e = \frac{v_\pi}{z_\pi} + g_m\, v_\pi \tag{3.3.1}$$

where:

$$z_\pi = \frac{r_\pi}{1 + s\, C_\pi\, r_\pi} \tag{3.3.2}$$

From these two equations it follows that:

$$i_e = v_\pi \left(g_m + \frac{1}{r_\pi} + s\,C_\pi \right) \tag{3.3.3}$$

The output collector current is:

$$i_c = g_m\, v_\pi \tag{3.3.4}$$

If we put Eq. 3.3.3 into Eq. 3.3.4 we obtain:

$$\frac{i_c}{i_e} = \frac{g_m}{g_m + \dfrac{1}{r_\pi} + s\,C_\pi} \approx \frac{g_m}{g_m + s\,C_\pi} = \frac{\alpha_0}{1 + s\,C_\pi\, r_e} \tag{3.3.5}$$

since $g_m \approx 1/r_e$ and $\alpha_0 = \beta_0/(\beta_0 + 1) \approx 1$. This equation has the pole at $-1/C_\pi\, r_e$, which lies extremely high in frequency, because r_e is normally very low. Since the output pole $-1/R_L C_\mu$, which we shall consider next, becomes prevalent we can neglect $s\,C_\pi r_e$ and assume that $i_c \approx i_e$. The output voltage is:

$$v_o = -\,i_c\, Z_L = -\,i_c\, \frac{R_L}{1 + s\,C_\mu\, R_L} \tag{3.3.6}$$

With the simplifications considered above we can write the expression for the transimpedance, which is:

$$\boxed{\frac{v_o}{i_e} \approx -\,\frac{R_L}{1 + s\,C_\mu\, R_L}} \tag{3.3.7}$$

Since the capacitance C_μ is in parallel with the loading resistor R_L, we can improve the performance by applying any of the inductive peaking circuits from Part 2. In practice we never consider only the 'pure' capacitance C_μ, because some stray capacitances are always present and must be taken into account. Also, if the transistor Q_1 is a part of an integrated circuit, we must consider the collector to substrate capacitance C_S. In such a case we use Eq. 3.1.2 with the exponent $m_c = 0.5$.

3.3.1 Input Impedance

We shall calculate the input impedance of the common base stage by taking into account the base resistance r_b, which — as we shall realize very soon — represents a very nasty obstacle in achieving a wide bandwidth. We shall make our derivation on the basis of Table 3.2.1, from which we have drawn Fig. 3.3.2. This figure represents the equivalent small signal input circuit owed to r_b.

The input admittance of the circuit in Fig. 3.3.2a is:

$$Y_e = \frac{1}{r_b} + \frac{1}{\dfrac{r_b}{\beta_0} + s\,\tau_T\, r_b} \tag{3.3.8}$$

Within the frequency range of interest the value r_b/β_0 in the second fraction is small and can be neglected. The simplified input admittance is thus:

$$Y_e \approx \frac{1}{r_b} + \frac{1}{s\,\tau_T\, r_b} \tag{3.3.9}$$

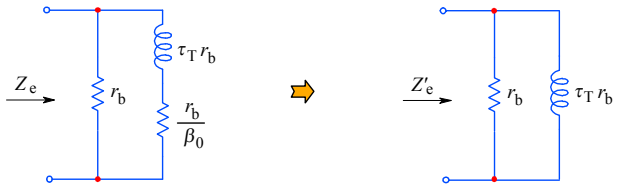


Fig. 3.3.2: Common base amplifier input impedance: a) r_b transformed to Z_e; b) within the frequency range of interest, r_b/β_0 can be neglected.

The real part represents a resistance:

$$\boxed{R_e \approx r_b} \tag{3.3.10}$$

and the imaginary part is an inductance:

$$\boxed{L_e \approx r_b \tau_T} \tag{3.3.11}$$

Normally, if the amplifier is built with discrete components, there is always some lead inductance L_s which must be added in series in order to obtain the total impedance.

In the next section, where we shall discuss the cascode amplifier, we shall find that the inductance L_e, together with the capacitance C_μ of the common emitter current driving stage, forms a parallel resonant circuit which may cause ringing in the amplification of steep signals. In most cases the resistance R_e is too large to damp the ringing effectively enough by itself, so additional circuitry will be required.

Eq. 3.3.10 and Eq. 3.3.11, respectively, disclose the fact that the annoying inductance L_e and the resistance R_e are directly proportional to the base spread resistance r_b. When using this type of amplifier for the output stages, where the amplitudes are large (e.g., in oscilloscopes), we must use more powerful transistors, mostly in TO5 case type. In this case the internal transistor connections are relatively long and its total active area is large, the corresponding r_b is large as well. In order to decrease R_e and L_e we must select transistors which have low r_b. To decrease base spread resistance as much as possible and also to decrease the *transition time* (the time needed by the current carriers to pass the base width), the firm RCA has developed (already in the late 1960s) the so called *overlay transistor*. A typical overlay transistor is the 2N3866. Such transistors are essentially integrated circuits, composed of many identical small transistors connected in parallel.

3.4 Cascode Amplifier

If the common emitter amplifier of Fig. 3.2.1a is used as a driver of the common base amplifier of Fig. 3.3.1a, a *cascode amplifier* [Ref. 3.3, 3.7, 3.12] is obtained. The name *cascode* springs from the times when electronic tubes were the circuit building blocks. The anode of the first tube (the equivalent of the common emitter stage) was loaded by the **cathode** input of the second tube with a grounded grid (the equivalent of the common base stage). Both electronic tubes were therefore connected in **cascade**, hence the compound word **cascode**.

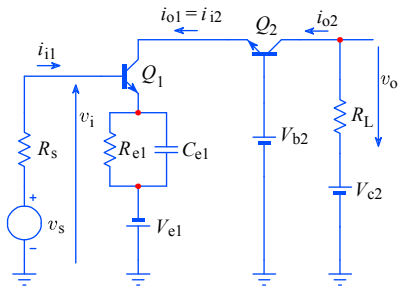


Fig. 3.4.1: Cascode amplifier schematic

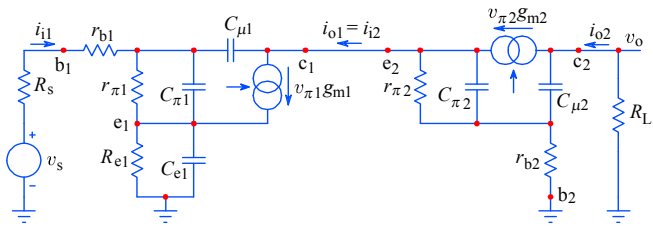


Fig. 3.4.2: Equivalent small signal model of a cascode amplifier. The components belonging to the common emitter circuit bear the index '1' and those of the common base circuit bear the index '2'.

3.4.1 Basic Analysis

A transistor cascode amplifier is drawn in Fig. 3.4.1 and Fig. 3.4.2 shows its small signal equivalent circuit. All the components that belong to transistor Q_1 bear the index '1' and all that belong to transistor Q_2 bear the index '2'.

For the emitter network of Q_1 we select the values such that $R_{e1} C_{e1} = \tau_{\Gamma 1}$.

In order to simplify the initial analysis, we shall first neglect R_s, r_{b2}, and $C_{\mu 1}$. Later we shall reintroduce these elements one by one to get a closer approximation.

We have already derived the equations needed for each part of the combined circuit: for the common emitter stage we have Eq. 3.2.66 and for the common base we

have Eq. 3.3.7. We only need to multiply these two equations to get the voltage gain of our cascode amplifier:

$$A_\mathrm{v} = \frac{i_\mathrm{o1}}{v_\mathrm{i}} \cdot \frac{v_\mathrm{o}}{i_\mathrm{o1}} = \frac{v_\mathrm{o}}{v_\mathrm{i}} \approx \frac{1}{R_\mathrm{e1}} \cdot \frac{1}{1 + s\,C_\mathrm{e1}\,r_\mathrm{b1}} \cdot \left(- \frac{R_\mathrm{L}}{1 + s\,C_{\mu 2}\,R_\mathrm{L}} \right) \qquad (3.4.1)$$

Here we have approximated $\alpha_{02} \approx 1$, and therefore $i_\mathrm{o2} = i_\mathrm{o1}$. The first fraction, multiplied by R_L from the third fraction, is the DC voltage amplification and the remainder represents the frequency dependence:

$$A_\mathrm{v} \approx - \frac{R_\mathrm{L}}{R_\mathrm{e1}} \cdot \frac{1}{(1 + s\,C_\mathrm{e1}\,r_\mathrm{b1})\,(1 + s\,C_{\mu 2}\,R_\mathrm{L})} \qquad (3.4.2)$$

Obviously, the frequency dependence is a second-order function. There are two poles: the pole at the input is $-1/\,C_\mathrm{e1} r_\mathrm{b1} = -1/\tau_\mathrm{T1}$ whilst the pole $-1/C_{\mu 2}\,R_\mathrm{L} = -1/\tau_\mathrm{T2}$ is on the output side. As we shall see later, it is possible to apply the peaking technique on both sides.

In an ideal case the common base stage input (emitter) impedance is very low. Because of this low load the first stage voltage gain $A_\mathrm{v1} \ll 1$, so $C_{\mu 1}$ would not be amplified by it. And if we could neglect r_b2 the capacitance $C_{\mu 2}$ would appear in parallel to the loading resistor R_L, and therefore it would neither be multiplied by the second stage's voltage gain A_v2. Both $C_{\mu 1}$ and $C_{\mu 2}$ are relatively small, so it is obvious that the cascode amplifier has, potentially, a much greater bandwidth in comparison with a simple common emitter amplifier (for the same total voltage gain). The price we pay for this improvement is the additional transistor Q_2.

Of course, in practice things are not so simple, and in addition we should not neglect the inevitable stray capacitances. Those should be added to $C_{\mu 1}$ and $C_{\mu 2}$. Also, $(R_\mathrm{s} + r_\mathrm{b1})$ with $C_{\mu 1}$ will affect the behavior of Q_1 and r_b2 with $C_{\mu 2}$ will affect the behavior of Q_2, as we shall see in the following analysis.

3.4.2 Damping of the Q_2 Emitter

Owing to the base spread resistance r_b2 the Q_2 input (emitter) has an inductive component with the inductance $L_\mathrm{e2} = \tau_\mathrm{T2}\,r_\mathrm{b2}$ in parallel with r_b2, as already shown in Table 3.2.1 and also in Fig. 3.3.2. The equivalent input impedance of the transistor Q_2 was derived in Eq. 3.3.9 to Eq. 3.3.11.

As shown in the simplified circuit in Fig. 3.4.3, the inductance L_e2 and the collector to base capacitance $C_{\mu 1}$ of Q_1 form a series resonant circuit, damped by r_b2 in parallel with L_e2 (and a series emitter resistance r_b2/β_2, which is very small, so it was neglected). The other end of $C_{\mu 1}$ is connected to the base of Q_1, where we must consider the following two effects:

- at very high frequencies the input signal goes directly through r_b1 and $C_{\mu 1}$;
- at lower frequencies, the signal is inverted and amplified by Q_1 and the internal base junction can then be treated as the virtual ground.

In an actual cascode amplifier Q_2 operates at a much higher collector voltage than Q_1, and since the collector currents are nearly equal this means a higher power dissipation for Q_2. By using for Q_2 a transistor capable of higher power dissipation, we shall probably have to accept its higher r_{b2} as well. This increases the inductance L_{e2} and lowers the damping of the series resonance formed by $C_{\mu 1}$ and L_{e2}, resulting in a large peaking near the upper cut off frequency.

To prevent this from happening, an additional impedance Z_d, consisting of a resistor R_d in parallel with a capacitor C_d, is connected between the collector of Q_1 and the emitter of Q_2. If R_d is made equal to r_{b2}, then C_d can be chosen so, that it cancels L_{e2}; the result is a resistive input impedance of the Q_2 emitter: $Z_{e2} \approx r_{b2}$.

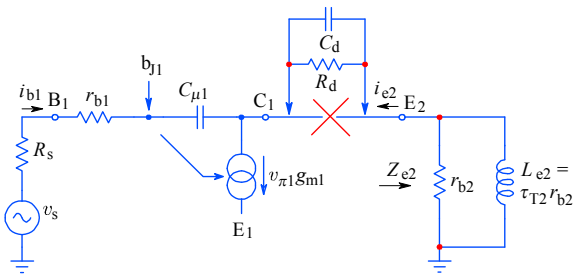


Fig. 3.4.3: Parasitic resonance damping of the cascode amplifier. Two current paths must be considered: at highest frequencies, for i_{b1}, $C_{\mu 1}$ represents a non-inverting cross-talk path; at lower frequencies, for i_{c1}, $C_{\mu 1}$ provides a negative feedback loop, thus it can be viewed as if being connected to a virtual ground (Q_1 base junction b_{J1}). The parasitic resonance, formed by $C_{\mu 1}$ and L_{e2} is only partially damped by r_{b2}; the required additional damping is provided by inserting R_d and C_d between Q_1 collector and Q_2 emitter.

So let us put:

$$R_d = r_{b2} = \sqrt{\frac{L_{e2}}{C_d}} \tag{3.4.3}$$

The value of C_d is then:

$$C_d = \frac{\tau_{T2}}{r_{b2}} = \frac{1}{2\pi f_{T2}\, r_{b2}} \tag{3.4.4}$$

To get a feeling for actual values let us have two equal transistors with parameters such as in the examples in Sec. 3.2.4 ($f_T = 600\,\text{MHz}$, $r_{b2} = 47\,\Omega$):

$$R_d = r_{b2} = 47\,\Omega \qquad C_d = \frac{1}{2\pi \cdot 600 \cdot 10^6 \cdot 47} = 5.6\,\text{pF} \tag{3.4.5}$$

The input impedance of the emitter circuit of transistor Q_2 now becomes:

$$Z_{e2} \approx r_{b2} \tag{3.4.6}$$

which is resistive at all frequencies (approximately so, because we have allowed ourselves a simplification). The corresponding equivalent circuit is shown in Fig. 3.4.4. The task of the impedance Z_d is actually twofold: it must damp the inductive input circuit of transistor Q_2, and as we shall see later, it can be a good choice for providing the thermal stabilization of the cascode stage.

Since we have introduced Z_d into the collector circuit of Q_1 we must now account for the Q_1 Miller capacitance:

$$C_{M1} \approx C_{\mu 1}\left(1+\frac{Z_{e2}}{Z_{e1}}\right) = C_{\mu 1}\left(1+\frac{r_{b2}}{R_{e1}}\right) \tag{3.4.7}$$

where Z_{e2} is the Q_2 compensated emitter input impedance and Z_{e1} is the impedance of the emitter circuit of Q_1. With this consideration the gain, Eq. 3.4.2, becomes:

$$A_v \approx -\frac{R_L}{R_{e1}} \cdot \frac{1}{\left[1+s\,(C_{e1}+C_{M1})\,r_{b1}\right](1+s\,C_{\mu 2}\,R_L)} \tag{3.4.8}$$

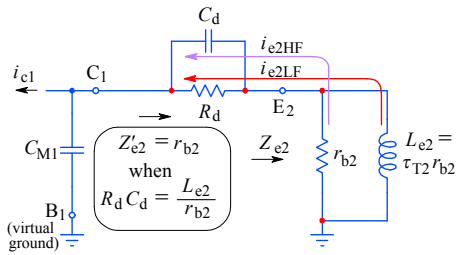


Fig. 3.4.4: With damping the simplified Q_2 input impedance becomes (approximately) resistive. Note the high and low frequency paths.

The collector to base capacitance of the transistor Q_1 allows very high frequency signals from the input bypassing this transistor and directly flowing into the emitter of transistor Q_2. Transistor Q_1 amplifies, inverts, and delays the low frequency signals. In contrast, all of what comes through $C_{\mu 1}$ is *non-delayed*, *non-amplified*, and *non-inverted*, causing a pre-shoot [Ref. 3.1] in the step response, as shown in Fig. 3.4.5. The Q_1 collector current, i_{c1}, is the sum of $i_{\mu 1}$ and $v_{\pi 1} g_{m1}$. Note that both the pre-shoot owed to $i_{\mu 1}$ and the overshoot of $v_{\pi 1} g_{m1}$ are reduced in v_{c2} by the Q_2 pole $(1/C_{\mu 2} R_L)$.

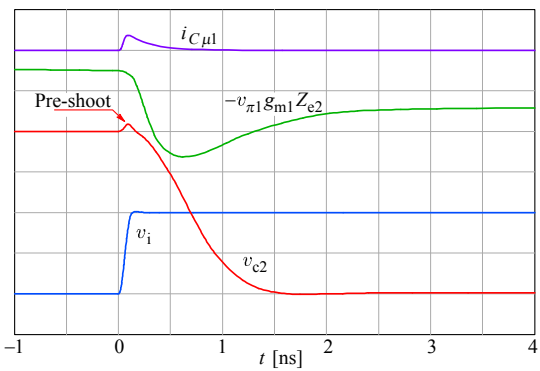


Fig. 3.4.5: The step response v_{c2} has a pre-shoot owed to the signal cross-talk through $C_{\mu 1}$ (arbitrary vertical units, but corresponding to $A_v = -2$).

So far we have excluded $C_{\mu 2}$ from our analysis. When included, its effect on bandwidth is severe, owing to the non-zero r_{b2} and the Miller effect. But it also affects the emitter input impedance of Q_2 since $C_{M2} = C_{\mu 2}(1 + A_v)$ appears in parallel to r_{b2} and is consequently transformed into the emitter in accordance with Table 3.2.1, in the same way as in Fig. 3.2.7. If A_v is relatively high the pronounced resonance owed to $C_{\mu 2}$ can cause long ringing, even if the bandwidth is lower than the resonant frequency.

The damping by $C_d R_d$ works only up to a certain frequency, because they add their own parasitic L and C, which we were trying to avoid in the firts place, and which also reduce the bandwidth. Instead, *John Addis* [Ref. 3.26] suggests an alternative approach by modifying the base impedance: the Q_2 base is connected to the bias voltage through a resistor R_A of up to a 100 Ω and grounded by a small capacitor $C_A \approx C_{\mu 2}$. In Fig. 3.4.6 we compare the voltage gain, the phase, and the group delay for the two cases: $R_A = 0$ and $R_A = 33\,\Omega$, respectively (the value of 33 Ω was optimized to the transistor model used in the simulation). The change in the Q_2 emitter impedance (b curves) is exposed by the lower drive stage current i_{c1} near the cut off frequency.

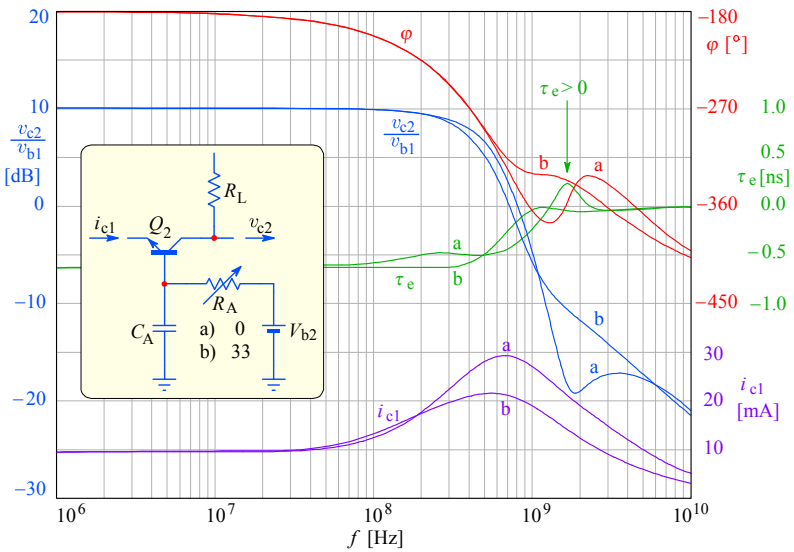


Fig. 3.4.6: The compensation method of Q_2 as suggested by John Addis. a) With $R_A = 0$, the frequency response has a notch at the resonance and a phase-reversed cross-talk, which makes the group delay τ_e positive, causing a potential instability. b) with $R_A = 33\,\Omega$ and $C_A = 3\,\text{pF}$ (the values suit the particular transistor model used for simulation) the frequency response slope is corrected to the $-6\,\text{dB}/2f$, the phase is smoothed and, although the bandwidth is reduced slightly, the group delay linearity is extended and the undesirable positive region is reduced to negative. The Q_2 emitter impedance is increased near cut off, as can be deduced from the lower i_{c1} peak.

To analyze this type of compensation of the Q_2 emitter impedance we must consider the equivalent circuit in Fig. 3.4.7. Here the capacitance $C_{\mu 2}$ is seen by the base as the Miller capacitance C_{M2} (remembering that $A_v = R_L/R_{e1}$):

$$C_{M2} = C_{\mu 2}(1 + A_v) \tag{3.4.9}$$

which appears in parallel with the base spread resistance r_{b2}.

The external compensation network, the parallel connection of R_A and C_A, is added in series and the total can then be transformed by the same rule as in Fig. 3.2.7. The base impedance is:

$$Z_{b2} = \frac{1}{\frac{1}{r_{b2}} + sC_{M2}} + \frac{1}{\frac{1}{R_A} + sC_A} \tag{3.4.10}$$

If $R_A \approx r_{b2}$ and $C_A \approx C_{\mu 2}$ then Z_{b2} becomes:

$$Z_{b2} = \frac{2R_A}{1 + sC_A R_A} = \frac{R_s}{1 + sC_s R_s} \tag{3.4.11}$$

which appears at the Q_2 emitter as a parallel connection of $R_s = 2R_A$ and $C_s = C_A/2$.

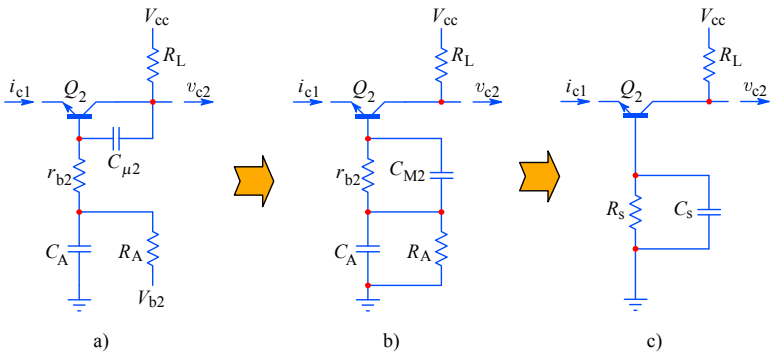


Fig. 3.4.7: The Q_2 emitter input impedance compensation by the added base network R_A and C_A (see also Fig. 3.2.7), making it capacitive at high frequencies. By altering the valuse slightly it can also be used to compensate the stray base inductance.

3.4.3 Thermal Compensation

Amplifier's thermal stability is preferably solved by using a differential configuration. However, in order to illustrate the potential problems, we discuss a single transistor stage. The well known relation for the transistor's base–emitter voltage is:

$$V_{be} = \frac{k_B T_j}{q_e} \ln\left(\frac{I_e}{I_s} + 1\right) \tag{3.4.12}$$

where T_j is the absolute temperature of the p-n junction, which causes the increase of V_{be} in a linear proportion. Unfortunately, the situation is complicated by the 'saturation current' $I_s \approx 10^{-14}$ A (for a typical silicon transistor at room temperature), which approximately doubles for every 8 K increase in temperature. The combined influence results in a V_{be} temperature coefficient of ≈ -2 mV/K (see Appendix 3.1 on the web). All other terms are constants (as *James M. Bryant* of Analog Devices likes to joke, we can not change the Boltzmann constant k_B because *Ludwig Boltzmann* is already dead, and neither can we change the electron charge q_e because both *Charles Augustine de Coulomb* and *Joseph John Thompson* are dead, too!).

When we apply the bias and the supply voltage to a transistor, the power dissipated by the transistor depends on the collector current I_c and the voltage V_{ce} across the transistor:

$$P_D = I_c V_{ce} \tag{3.4.13}$$

If we know the ambient temperature T_a and the thermal resistance from junction to ambient $R_{\theta ja}$ [K/W] (*kelvin* per *watt*), we can calculate the junction temperature T_j:

$$T_j = T_a + R_{\theta ja} P_D \tag{3.4.14}$$

In a properly designed amplifier, a certain time after power-up an equilibrium is reached: all the heat generated by the transistor dissipates into the ambient and the transistor obtains a stable, but higher temperature.

If we now apply a voltage step to the base the collector current I_c increases, changing the power dissipation (depending on the chosen DC bias it can either increase or decrease). A little later, depending on the thermal capacitance of the transistor's case $C_{\theta c}$ and the thermal resistance $R_{\theta ja}$, a new temperature equilibrium is reached. As shown in Fig. 3.4.8, the change occurs gradually, depending on the transistor thermal time constant. But the supply voltage is shared between the transistor and the loading resistance, whilst the same current flows through both of them, therefore the choice of the DC bias point governs whether the transistor power dissipation will increase or decrease with the signal. This is known as the 'thermal distortion'.

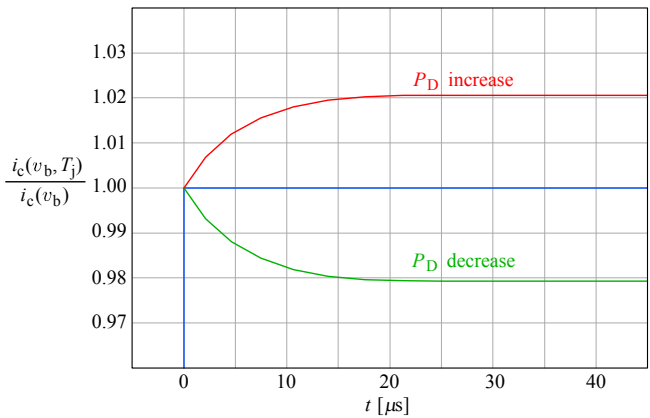


Fig. 3.4.8: The collector current step response is distorted by a long term thermal drift following the transient. Depending on the chosen DC bias point, the output can either increase or decrease relative to the ideal response. In addition to the junction thermal time constant, there can also be a slower one, owed to the transistor case temperature change.

Although the dynamic emitter resistance r_e is also temperature dependent (see Eq. 3.1.1), which affects the gain (see Eq. 3.2.61), in all wideband circuits we put (for bandwidth reasons!) a much larger external degeneration resistance R_e in series, reducing the gain temperature dependence. So in practice all thermal changes occur because of V_{be} and I_s.

In a multi-stage amplifier different transistors will have different operating points, thus different temperatures and temperature dependence. Also, they have different time constants, which cause longer and shorter drifts, and we can not expect all effects to cancel. Consequently even long oscillations in the step response can occur.

Before we look for the remedy for the problem of how to cancel, or at least how to substantially reduce the thermal distortion, let us take a look at Fig. 3.4.9, which shows a simple common emitter stage, and the way in which the power dissipation is shared between the load and the transistor as a function of the collector current. As usual, we use capital letters for the applied DC voltages, loading resistor, etc., and small letters for the instantaneous signal voltages and currents. The transistor's power dissipation is:

$$P_D = v_{ce}\, i_c = v_{ce}\,\frac{v_L}{R_L} = v_{ce}\,\frac{V_{cc} - v_{ce}}{R_L} = v_{ce}\,\frac{V_{cc}}{R_L} - \frac{v_{ce}^2}{R_L} \qquad (3.4.15)$$

Since v_{ce} cannot exceed V_{cc} if the collector load is purely resistive, the right-hand vertical axis is normalized to V_{cc}. The left-hand vertical axis is normalized to the maximum load power, which is simply $P_{Lmax} = V_{cc}^2/R_L$ (corresponding to $v_{ce} = 0$ and thus $P_D = 0$). The transistor's power dissipation, however, follows an inverse parabolic function with a maximum at $v_{ce} = V_{cc}/2$:

$$P_{Dmax} = \left(\frac{V_{cc}}{2}\right)^2 \frac{1}{R_L} = \frac{V_{cc}^2}{4\,R_L} \qquad (3.4.16)$$

This middle point is the optimum DC bias point for a transistor stage. If excited by small signals, the transistor power dissipation follows the parabola close to its broad and relatively flat top and thus it does not change very much. This means that the transistor's temperature does not change very much either.

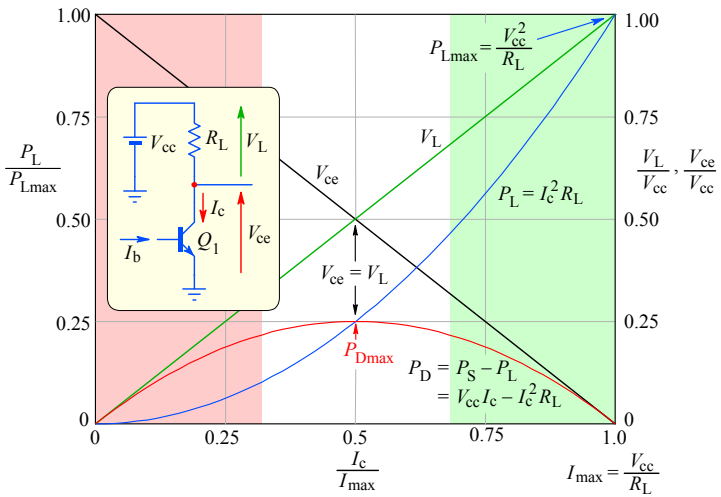


Fig. 3.4.9: The optimum bias point is when the voltage across the load is equal to the voltage across the transistor. This is optimal both in the sense of thermal stability, as well as in the available signal range sense.

If other design requirements force us to move the bias point far from the top of the parabola, the bias with $V_{ce} < V_{cc}/2$ (or $I_c > I_{max}/2$) is preferred, in contrast to the range $V_{ce} > V_{cc}/2$, because the latter situation is unstable. However, in wideband amplifiers we can hardly avoid it, because we want to have a low R_L, a high I_c and a high V_{cb} (to reduce C_μ) and all three conditions are required for high bandwidth.

The typical temperature coefficient of a base–emitter p–n junction voltage ($\approx 0.6\,V$) is approximately $-2\,mV/K$ for silicon transistors, so we can explain the instability in the following way:

When the circuit is powered up the transistor conducts a certain collector current, which heats the transistor, increasing the transistor base–emitter p–n junction temperature. If the base is biased from a voltage source (low impedance, which in wideband amplifiers is usually the case), the temperature increase will, owing to the negative temperature coefficient, decrease the base–emitter voltage. In turn, the base current increases and, consequently, both the emitter and the collector current increase, which further increases the dissipation and the junction temperature. The load voltage drop would also increase with current and therefore reduce the collector voltage, thus lowering the transistor power dissipation. But with a low R_L, the change in the drop of load voltage will be small, so the transistor power dissipation increase will be reduced only slightly.

The effect described is cumulative; it may even lead to a thermal runaway and the consequent destruction of the transistor if the top of the parabola exceeds the maximum permissible power dissipation of the transistor (which is often the case, since we want low R_L and high V_{cc} and I_e, as stated above). In a similar way, on the basis of the -2 mV/K temperature dependence of V_{be}, the reader can understand why the bias point for $V_{ce} < V_{cc}/2$ is thermally stable.

According to Eq. 3.4.16, to have the transistor thermally stable means having resistance R_L (or $R_L + R_e$) such that at the bias point the voltage drop across them is equal to $V_{cc}/2$. In general, this principle is successfully applied in differential amplifiers: when one transistor is excited so that its bias point is pushed to one side of the parabola the bias point of the other transistor is moved exactly to the same dissipation on the opposite side of the parabola. As a result the temperature becomes lower but equal in both transistors. Thus in the differential amplifier both transistors can always have similar temperatures, even if the temperature changes by the signal (provided that we remain within the linear range of excitation). Also, the thermal drift can be minimized by using current source biasing in the emitters of a differential amplifier, as well as using emitter degeneration resistors, which minimize variations in ΔV_{be}, even if transistor parameters vary considerably.

In our cascode circuit of Fig. 3.4.1 the transistor Q_1 already has an emitter resistor R_{e1} as dictated by the required current gain, and we do not want to change it. However, we can add a resistor, which we label R_θ, in the collector circuit of Q_1 to make $V_{ce1} \approx I_{c1}\,(R_{e1} + R_\theta) \approx V_{e2}/2$, where V_{e2} is the voltage at the emitter of the transistor Q_2. Suppose now that the emitter current is $I_{e1} \approx I_{c1} = 50\,mA$ and the Q_2 base voltage $V_{b2} = +15\,V$. Then the emitter voltage of transistor Q_2 is:

$$V_{e2} = V_{b2} - V_{be2} \approx 15 - 0.6 = 14.4\,V \qquad (3.4.17)$$

where V_{be2} is the base–emitter voltage (about $0.6\,V$ for a silicon transistor).

With $R_{e1} = 20\,\Omega$ the value of the thermal compensation resistor is:

$$R_\theta = \frac{V_{e2}/2 - I_{c1}\,R_{e1}}{I_{c1}} = \frac{14.4/2 - 0.05 \cdot 20}{0.05} = 124\,\Omega \qquad (3.4.18)$$

Such a resistor should be used instead of R_d as calculated before to achieve both the ringing suppression and the thermal compensation. But by inserting these $124\,\Omega$ instead of $R_d = 47\,\Omega$, the Miller capacitance C_M would increase to 27.8 pF, decreasing the amplifier bandwidth too much. Fortunately the compensating resistor R_θ needs to correct the behavior of the amplifier only at very low frequencies. So we can bridge that part of it which is not needed for the suppression of ringing by an appropriate capacitor; let us call it C_θ, as shown in Fig. 3.4.10. By doing so, we prevent an excessive increase of Miller capacitance.

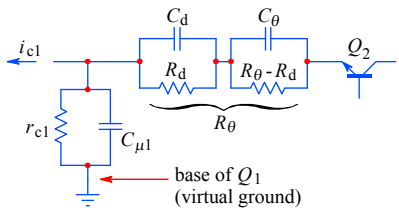


Fig. 3.4.10: The modified compensation network: R_dC_d provide the HF damping, whilst $R_\theta C_\theta$ provide thermal compensation.

The question is how to calculate the proper value of C_θ? The obvious way would be to calculate the thermal capacity of the transistor's die and case mass (and an eventual heat sink) and all the thermal resistances (junction to case, case to heath sink, heath sink to air), as is usually done for high power output stages.

Bruce Hofer [Ref.3.8] suggests the following — more elegant — procedure, based on Fig. 3.4.10. The two larger time constants in this figure must be equal:

$$(R_\theta - R_d)\,C_\theta = r_{c1}\,C_{M1} \qquad (3.4.19)$$

Here r_{c1} is the dynamic collector resistance of transistor Q_1, derived from Fig. 3.4.11 as $\Delta V_{ce}/\Delta I_c$. In this figure V_A is the *Early voltage*:

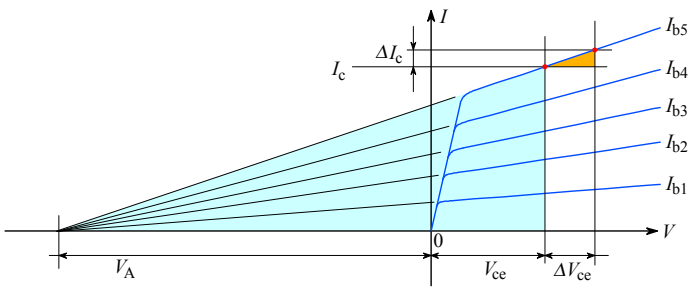


Fig. 3.4.11: The dynamic collector resistance r_{c1} is derived from the $I_c(V_{ce}, I_b)$ characteristic and the Early voltage V_A.

The meaning of the Early voltage can be derived from Fig. 3.4.11, where several curves of the collector current I_c vs. collector to emitter voltage V_{ce} are drawn, with base current I_b as the parameter. With increasing collector voltage the collector current increases even if the base current is kept constant. This is because the collector to base depleted area widens on the account of the (active) base width as the collector voltage increases. This in turn causes the diffusion gradient of the current carriers in the base to increase, hence the increased collector current.

By extending the lines of the collector current characteristics back, as shown in Fig. 3.4.11, all the lines intersect the abscissa at the same virtual voltage point V_A (negative for NPN transistors), called the *Early voltage* (after *J.M. Early*, [Ref. 3.11]). From the similarity of triangles we can derive the collector's dynamic resistance:

$$r_{c1} = \frac{\Delta V_{ce}}{\Delta I_c} = \frac{V_c - (-V_A)}{I_c} \tag{3.4.20}$$

Since the voltage gain of the common emitter stage is low, C_{M1} will be only slightly larger than $C_{\mu 1}$. If we now suppose that transistor Q_1 has an $r_{c1} = 0.5 \cdot 10^6\,\Omega$ and $C_{\mu 1} = 3\,\text{pF}$, the value of C_θ should be:

$$C_\theta = \frac{r_{c1}\, C_{M1}}{R_\theta - R_d} = \frac{0.5 \cdot 10^6 \cdot 3 \cdot 10^{-12}}{124 - 47} = 19.5\,\text{nF} \tag{3.4.21}$$

In practice, we can take the closest standard values, e.g., $C_\theta = 22\,\text{nF}$ and $R_\theta - R_d = 124 - 47 = 77\,\Omega \approx 75\,\Omega$. Since, in general, a wideband amplifier has several amplifying stages, each one having its own temperature and damping problems, these values can be varied substantially in order to achieve the desired performance. Thermal problems tend to be more pronounced towards the output stages where the signal amplitudes are high. Here experimentation should have the last word. Nevertheless, the values obtained in this way represent a good starting point.

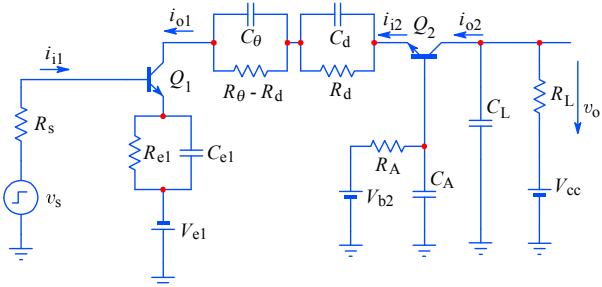


Fig. 3.4.12: The compensated cascode amplifier. If $R_A C_A$ compensation is used, $R_d C_d$ may become unnecessary. The main bandwidth limiting time-constants are now $(R_s + r_{b1})C_{e1}$ and $(C_{\mu 2} + C_L)R_L$.

In order to achieve a good thermal compensation, the same half-voltage principle (Fig. 3.4.9) can be applied to both Q_1 and Q_2, although this will rarely be possible. On the other hand, a cascode amplifier can always be implemented as a differential amplifier, simplifying the problem of thermal compensation and, as a bonus, doubling the gain. We shall see this later in Sec.3.7. There are, as we are going to see in following sections, still some points for possible improvement of the cascode amplifier.

All the discussion on thermal compensation in this section was based on the simple single stage cascode amplifier shown in Fig. 3.4.12.

It is however important to note that a multistage DC coupled amplifier must necessarily be built in a differential configuration. Then many of the effects described would be either compensated automatically or at least reduced significantly. Not all, unfortunately. Often the biggest problems will be caused by the first stage, since its output will be amplified by all the remaining gain of the chain. But the stages following it might also have their share, because of the increased signal amplitude. And the output stage is usually the one to drive the load, thus it is required to handle most of the power. It is therefore necessary to pay attention to thermals at each stage.

In addition, in an integrated circuit it is also necessary to check the adjacent circuits, such as current mirrors, constant current bias circuits and voltage references; they will all be affected by every thermal gradient across the chip. They will also tend to have their own particular thermal time constant, and changes in those can in turn affect the main signal chain. Consequently a complex system can be difficult to control and compensate properly.

Finally, in amplifiers employing DC feedback the thermal problems are usually reduced to only those of the first stage, but still the output stage could be prone to thermal runaway under heavy load.

Some of these problems will be mentioned again in discussions of differential circuit configurations in further sections and chapters.

3.5 Emitter Peaking in a Cascode Amplifier

Here we shall examine the possibility of improving the cascode amplifier bandwidth by applying the emitter peaking technique.

Let us return to the basic cascode amplifier of Fig. 3.4.1 and Fig. 3.4.2, but with a little modification: we shall assume a current signal source in parallel with the source resistance R_s. We shall also assume that there is an output capacitance in parallel with the load resistance R_L such that its value is $C_o = C_{\mu 2} + C_L$. To simplify the analysis we shall disregard the damping and thermal compensation described in the previous section. We shall basically follow the steps of Carl Battjes [Ref. 3.1], to show a different approach to the cascode amplifier design.

3.5.1 Basic Analysis

The transimpedance of the amplifier in Fig. 3.5.1 is:

$$\frac{v_o}{i_s} = \frac{i_{b1}}{i_s} \cdot \frac{i_{c1}}{i_{b1}} \cdot \frac{i_{c2}}{i_{c1}} \cdot \frac{v_o}{i_{c2}} \tag{3.5.1}$$

where:

$$\frac{i_{b1}}{i_s} = \frac{R_s}{R_s + Z_i} \qquad \text{and} \qquad \frac{i_{c1}}{i_{b1}} = \beta(s) = \frac{1}{\dfrac{1}{\beta_0} + s\,\tau_{T1}} \tag{3.5.2}$$

with Z_i being the input impedance looking into the base of transistor Q_1 and R_s the source resistance. At higher frequencies, when the input capacitance of transistor Q_1 prevails (see Eq. 3.2.12 and [Ref.3.1]), we have:

$$\frac{i_{c1}}{i_{b1}} \approx \frac{1}{s\,\tau_{T1}} \qquad \text{and} \qquad \tau_{T1} = \frac{1}{2\pi f_{T1}} \tag{3.5.3}$$

Further, for Q_2 we have:

$$\frac{i_{c2}}{i_{c1}} = \alpha_2 \approx 1 \qquad \text{and} \qquad \frac{v_o}{i_{c2}} = \frac{R_L}{1 + s\,R_L\,C_o} = \frac{R_L}{1 + s\,\tau_L} \tag{3.5.4}$$

where $\tau_L = R_L\,C_o$ and $C_o = C_{\mu 2} + C_L$.

We shall temporarily neglect the base spread resistance r_{b1} and calculate the input impedance Z_i at the base–emitter junction of Q_1 by Eq. 3.2.15:

$$Z_i = \big[\beta(s) + 1\big]\, Z_{e1} = \left(\frac{1}{s\,\tau_{T1}} + 1\right) Z_{e1} = \frac{1 + s\,\tau_{T1}}{s\,\tau_{T1}}\, Z_{e1} \tag{3.5.5}$$

The emitter peaking technique basically involves the introduction of a zero at $s_z = -\omega_R = -1/\tau_R$ in the emitter network of Q_1 to cancel the pole of the Q_2 collector load at $s_p = -\omega_L = -1/\tau_L$. For an efficient peaking τ_R must be lower than τ_L, but still above the time constant τ_{T1} of Q_1:

$$\tau_{T1} < \tau_R < \tau_L \tag{3.5.6}$$

The complete emitter circuit should look as in Fig. 3.5.1a, in which $\tau_R = RC$ and its HF impedance is:

$$Z_{e1} = \frac{R_{e1}\,(1 + s\,\tau_R)}{(1 + s\,\tau_{T1})\,(1 + s\,\tau_L)} = \frac{R_{e1}(1 + s\,\tau_R)}{s^2\,\tau_{T1}\tau_L + s\,(\tau_{T1} + \tau_L) + 1} \tag{3.5.7}$$

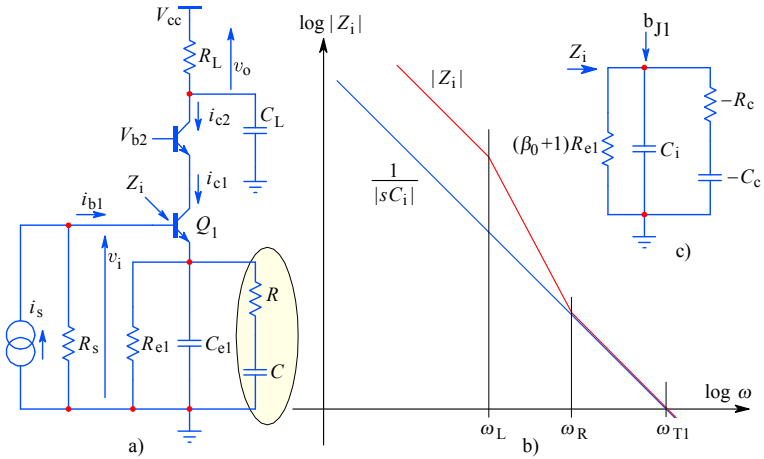


Fig. 3.5.1: Emitter peaking in cascode amplifiers. a) By adding a zero at $\omega_R = 1/RC$ to the emitter network of Q_1, we modify the input impedance Z_i at the base junction of Q_1; b) the frequency dependent asymptote plot of Z_i; c) the equivalent schematic has negative components, too. See the explanation in Sec. 3.5.2.

By introducing Eq. 3.5.7 back into Eq. 3.5.5 the input impedance can be expressed as:

$$Z_i = \frac{1 + s\,\tau_{T1}}{s\,\tau_{T1}} \cdot \frac{R_{e1}\,(1 + s\,\tau_R)}{(1 + s\,\tau_{T1})\,(1 + s\,\tau_L)} = \frac{R_{e1}\,(1 + s\,\tau_R)}{s\,\tau_{T1}(1 + s\,\tau_L)} \tag{3.5.8}$$

Now we put Eq. 3.5.2 and Eq. 3.5.8 into Eq. 3.5.1:

$$\frac{v_o}{i_s} = \frac{R_s\,R_L}{\left[R_s + \dfrac{R_{e1}\,(1 + s\,\tau_R)}{s\,\tau_{T1}(1 + s\,\tau_L)}\right] s\,\tau_{T1}(1 + s\tau_L)}$$

$$= \frac{R_s\,R_L}{s^2\,R_s\,\tau_{T1}\,\tau_L + s\,(R_s\,\tau_{T1} + R_{e1}\,\tau_R) + R_{e1}} \tag{3.5.9}$$

Next we put the denominator into the canonical form and equate it to zero:

$$s^2 + s\,\frac{R_s\,\tau_{T1} + R_{e1}\,\tau_R}{R_s\,\tau_{T1}\,\tau_L} + \frac{R_{e1}}{R_s\,\tau_{T1}\,\tau_L} = s^2 + a\,s + b = 0 \tag{3.5.10}$$

where we set the coefficients:

$$a = \frac{R_s\,\tau_{T1} + R_{e1}\,\tau_R}{R_s\,\tau_{T1}\,\tau_L} \quad \text{and} \quad b = \frac{R_{e1}}{R_s\,\tau_{T1}\,\tau_L} \tag{3.5.11}$$

The general solution of Eq. 3.5.10 is:

$$s_{1,2} = -\frac{a}{2} \pm \sqrt{\frac{a^2}{4} - b} \tag{3.5.12}$$

An efficient peaking must have complex poles, so the expression under the square root must be negative, therefore: $b > a^2/4$. We can then extract the negative sign as the imaginary unit and write Eq. 3.5.12 in the form:

$$s_{1,2} = -\frac{a}{2} \pm j\sqrt{b - \frac{a^2}{4}} \tag{3.5.13}$$

From Eq. 3.5.13 we can calculate the tangent section of the pole angle θ:

$$\tan\theta = \frac{\Im\{s_1\}}{\Re\{s_1\}} = \frac{\sqrt{b - \frac{a^2}{4}}}{\frac{a}{2}} = \sqrt{\frac{4\,b}{a^2} - 1} \tag{3.5.14}$$

It follows that:

$$1 + \tan^2\theta = \frac{4\,b}{a^2} \tag{3.5.15}$$

Now we insert the expressions from Eq. 3.5.11 for a and b and obtain:

$$1 + \tan^2\theta = \frac{4\,\dfrac{R_{e1}}{R_s\,\tau_{T1}\,\tau_L}}{\left(\dfrac{R_s\,\tau_{T1} + R_{e1}\,\tau_R}{R_s\,\tau_{T1}\,\tau_L}\right)^2} = \frac{4\,R_{e1}\,R_s\,\tau_{T1}\,\tau_L}{\left(R_s\,\tau_{T1} + R_{e1}\,\tau_R\right)^2} \tag{3.5.16}$$

By taking the square root the result is:

$$\sqrt{1 + \tan^2\theta} = \frac{2\sqrt{R_{e1}\,R_s\,\tau_{T1}\,\tau_L}}{R_s\,\tau_{T1} + R_{e1}\,\tau_R} \tag{3.5.17}$$

Finally we solve this for τ_R and obtain:

$$\boxed{\tau_R = R\,C = 2\sqrt{\frac{R_s\,\tau_{T1}\,\tau_L}{R_{e1}\left(1 + \tan^2\theta\right)}} - \frac{R_s\,\tau_{T1}}{R_{e1}}} \tag{3.5.18}$$

The admittance Y_{e1} of the emitter circuit in Fig. 3.5.1 is:

$$\begin{aligned} Y_{e1} &= \frac{1}{R_{e1}} + s\,C_{e1} + \frac{1}{R + \dfrac{1}{s\,C}} \\ &= \frac{s^2\,C\,R\,C_{e1}\,R_{e1} + s\left(C\,R + C\,R_{e1} + C_{e1}\,R_{e1}\right) + 1}{R_{e1}\left(1 + s\,C\,R\right)} \end{aligned} \tag{3.5.19}$$

The emitter impedance Z_{e1} is the inverse value of Y_{e1} and it must be equal to Eq. 3.5.7:

$$Z_{e1} = \frac{R_{e1}\left(1 + s\,CR\right)}{s^2\,CR\,C_{e1}R_{e1} + s\left(CR + CR_{e1} + C_{e1}R_{e1}\right) + 1}$$

$$= \frac{R_{e1}\left(1 + s\,\tau_R\right)}{s^2\,\tau_{T1}\,\tau_L + s\left(\tau_{T1} + \tau_L\right) + 1} \tag{3.5.20}$$

The coefficients of s^2 and s respectively must be equal in both fractions, therefore:

$$CR\,C_{e1}R_{e1} = \tau_{T1}\,\tau_L \tag{3.5.21}$$

and:

$$CR + R_{e1}(C + C_{e1}) = \tau_{T1} + \tau_L \tag{3.5.22}$$

The value of R_{e1} is constrained by the DC current amplification R_L/R_{e1}. Thus we need the expressions for C, C_{e1}, and R. By using Eq. 3.5.18, 3.5.21, and 3.5.22 we obtain:

$$C_{e1} = \frac{\tau_{T1}\,\tau_L}{R_{e1}\,\tau_R} \tag{3.5.23}$$

and:

$$C = \frac{\tau_{T1} + \tau_L - \tau_R - \tau_{T1}\,\dfrac{\tau_L}{\tau_R}}{R_e} \tag{3.5.24}$$

where τ_R should be calculated by Eq. 3.5.18. Once the value of C is known we can easily calculate the value of the resistor $R = \tau_R/C$. Of course, τ_R is determined by the angle θ of the poles selected for the specified type of response.

Fig. 3.5.2a and 3.5.2b show the normalized pole loci in the complex plane. As seen already in examples in Part 1 and Part 2, to achieve the maximally flat envelope delay response (MFED), a single stage 2nd-order function must have the pole angle $\theta = \pm 150°$. The original circuit has two real poles s_{T1} and s_L, but when the emitter peaking zero s_R is brought close to s_L the poles form a complex conjugate pair.

The frequency response is altered as shown in Fig. 3.5.2c and the bandwidth is extended. The emitter current increase $i_{e1}(\omega)/I_{e1}$ owing to the introduced RC network has two break points: the lower is owed to $R_{e1}(C + C_{e1})$ and the upper is owed to RC. If the break point at ω_R is brought exactly over ω_L they cancel each other, and the final response is shaped by the break point ω_{T1} of the transistor and the second break point in the emitter peaking network, ω_C. The peaking can thus be adjusted by R and C.

Let us consider an example with these data: $f_{T1} = 2000\,\text{MHz}$, $R_s = 60\,\Omega$, $R_e = 20\,\Omega$, $C_o = 9\,\text{pF}$, $R_L = 390\,\Omega$. We want to make the amplifier with such an emitter peaking network which will suit the Bessel pole loci (MFED), where the pole angle $\theta = \pm 150°$. First we calculate both time constants:

$$\tau_{T1} = \frac{1}{2\pi f_{T1}} = \frac{1}{2\pi \cdot 2000 \cdot 10^6} = 79.58\,\text{ps} \tag{3.5.25}$$

and:

$$\tau_L = R_L\,C_o = 390 \cdot 9 \cdot 10^{-12} = 3.51\,\text{ns} \tag{3.5.26}$$

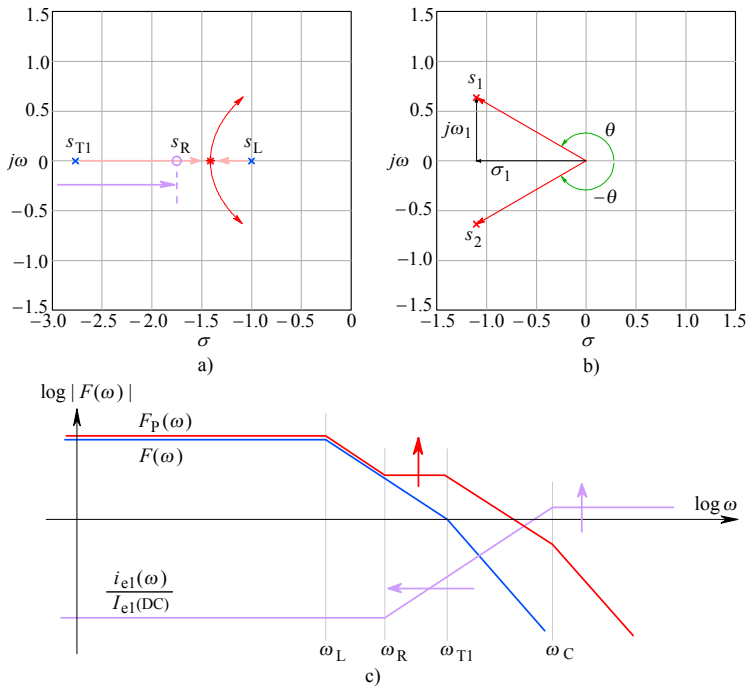


Fig. 3.5.2: Emitter peaking: a) two real poles travel towards each other when the emitter network zero goes from $-\infty$ towards s_L, forming eventually a complex conjugate pair; b) poles for Eq. 3.5.14 for the 2nd-order Bessel (MFED) response. c) frequency response asymptotes — the bandwidth is extended to ω_{T1} if $\omega_R = \omega_L$.

By using Eq. 3.5.18 we then calculate the third time constant:

$$\tau_R = 2\sqrt{\frac{R_s\,\tau_{T1}\,\tau_L}{R_{e1}\,(1+\tan^2\theta)}} - \frac{R_s\,\tau_{T1}}{R_{e1}} = \tag{3.5.27}$$

$$= 2\sqrt{\frac{60\cdot 79.58\cdot 10^{-12}\cdot 3.51\cdot 10^{-9}}{20\cdot(1+\tan^2 30°)}} - \frac{60\cdot 79.58\cdot 10^{-12}}{20} = 1.35\,\text{ns}$$

Now we can calculate C_{e1} using Eq. 3.5.23:

$$C_{e1} = \frac{\tau_{T1}\,\tau_L}{R_{e1}\,\tau_R} = \frac{79.58\cdot 10^{-12}\cdot 3.51\cdot 10^{-9}}{20\cdot 1.35\cdot 10^{-9}} = 10.35\,\text{pF} \tag{3.5.28}$$

According to Eq. 3.5.24 the value of the capacitor C is:

$$C = \frac{\tau_{T1} + \tau_L - \tau_R - \tau_{T1}\dfrac{\tau_L}{\tau_R}}{R_e} =$$

$$= \frac{79.58 \cdot 10^{-12} + 3.51 \cdot 10^{-9} - 1.35 \cdot 10^{-9} - 79.58 \cdot 10^{-12} \cdot \dfrac{3.51 \cdot 10^{-9}}{1.35 \cdot 10^{-9}}}{20}$$

$$= 101.6\,\text{pF} \tag{3.5.29}$$

Finally we calculate the value of the resistor R, from the time constant τ_R:

$$R = \frac{\tau_R}{C} = \frac{1.35 \cdot 10^{-9}}{101.6 \cdot 10^{-12}} = 13.29\ \Omega \tag{3.5.30}$$

3.5.2 Input Impedance Compensation

The introduction of the peaking elements R and C affects the HF input impedance in an unfavorable way. We have calculated the input impedance at the base emitter junction in Eq. 3.5.8, which we rewrite as:

$$Z_i = \frac{R_{e1}\,(1 + s\,\tau_R)}{s\,\tau_{T1}\,(1 + s\,\tau_L)} = \frac{s\,\tau_R\,R_{e1} + R_{e1}}{s^2\tau_{T1}\,\tau_L + s\,\tau_{T1}} \tag{3.5.31}$$

Here we have an additional pole at $s_L = -\omega_L = -1/\tau_L$ and a zero at $s_R = -\omega_R = -1/\tau_R$, both spoiling the purely capacitive character of the input impedance, which we would like to have (frankly, we would prefer the input capacitance to be zero as well, but this is not feasible). We have seen the Bode plot of the input impedance and its configuration already in Fig. 3.5.1b and 3.5.1c. At very high frequencies, where s becomes dominant, the input impedance obtains a simple capacitive character:

$$Z_i = \frac{R_{e1}\,\tau_R}{s\,\tau_{T1}\,\tau_L} \quad \Rightarrow \quad C_i = \frac{\tau_{T1}\,\tau_L}{R_{e1}\,\tau_R} \tag{3.5.32}$$

Our objective is to keep such an input impedance (at the base-emitter junction of Q_1) at lower frequencies also. In other words, at lower frequencies the plot of the input impedance should correspond to the $|1/sC_i|$ line in Fig. 3.5.1b. All other impedances that appear in the input circuit should be canceled by an appropriate compensating network. To find these impedances, we perform a 'continued fraction expansion' synthesis of the input admittance Y_i as derived from the right side of Eq. 3.5.8. Thus:

$$Y_i = \frac{1}{Z_i} = \frac{s^2\tau_{T1}\,\tau_L + s\,\tau_{T1}}{s\,\tau_R\,R_{e1} + R_{e1}} = s\,\frac{\tau_{T1}\,\tau_L}{R_{e1}\,\tau_R} + \frac{s\,\tau_{T1} - s\,\dfrac{\tau_{T1}\,\tau_L}{\tau_R}}{s\,\tau_R\,R_{e1} + R_{e1}} \tag{3.5.33}$$

The first fraction we recognize to be the input admittance $s\,C_i$. The second fraction can be inverted and, by canceling out s, we obtain the impedance:

$$Z_i' = \frac{s\,\tau_R\,R_{e1} + R_{e1}}{s\,\tau_{T1}\left(1 - \dfrac{\tau_L}{\tau_R}\right)} = \frac{R_{e1}\,\tau_R^2}{\tau_{T1}(\tau_R - \tau_L)} + \frac{R_{e1}\tau_R}{s\,\tau_{T1}(\tau_R - \tau_L)}$$

$$= -\,R_c - \frac{1}{s\,C_c} \tag{3.5.34}$$

This means a resistor $-R_c$ and a capacitor $-C_c$ connected in series, and this combination is in parallel with the input capacitance C_i. The values are negative because $\tau_R < \tau_L$ as was required in Eq. 3.5.6. On the basis of these results we can draw the equivalent input impedance circuit corresponding to Fig. 3.5.1c. The expression for the capacitance C_c is:

$$C_c = \frac{\tau_{T1}(\tau_L - \tau_R)}{R_{e1}\,\tau_R} \tag{3.5.35}$$

From Eq. 3.5.33, as well as from our previous analysis, we can derive that $R_c C_c = \tau_R$ and obtain a simpler expression for R_c:

$$R_c = \frac{\tau_R}{C_c} \tag{3.5.36}$$

Let us now continue our example of the emitter peaking cascode amplifier with the data $R_{e1} = 20\,\Omega$, $\tau_{T1} = 79.58\,\text{ps}$, $\tau_R = 1.51\,\text{ns}$, and $\tau_L = 3.51\,\text{ns}$, and calculate the values of C_i, C_c, and R_c. The input capacitance C_i, without C_M, is:

$$C_i = \frac{\tau_{T1}\,\tau_L}{R_{e1}\,\tau_R} = \frac{79.58 \cdot 10^{-12} \cdot 3.51 \cdot 10^{-9}}{20 \cdot 1.35 \cdot 10^{-9}} = 10.35\,\text{pF} \tag{3.5.37}$$

The value of the capacitance C_c is:

$$C_c = \frac{\tau_{T1}(\tau_L - \tau_R)}{R_{e1}\,\tau_R} = \frac{79.58 \cdot 10^{-12} \cdot (3.51 - 1.35) \cdot 10^{-9}}{20 \cdot 1.35 \cdot 10^{-9}} = 6.37\,\text{pF} \tag{3.5.38}$$

and the resistor R_c has a resistance of:

$$R_c = \frac{\tau_R}{C_c} = \frac{1.35 \cdot 10^{-9}}{3.51 \cdot 10^{-12}} = 385\,\Omega \tag{3.5.39}$$

The next step is to compensate the series connected $-C_c$ and $-R_c$. This can be done by connecting in parallel an equal combination with positive elements. The admittance of such a combination is zero and thus the impedance becomes infinity. The mathematical proof for this operation is:

$$Y_i' = \frac{1}{-R_c - \dfrac{1}{s\,C_c}} + \frac{1}{R_c + \dfrac{1}{s\,C_c}} = 0$$

and:

$$Z_i' = \frac{1}{Y_i'} \Rightarrow \infty \tag{3.5.40}$$

By doing so, only the input capacitances $C_i + C_M$ and the input resistance $(1+\beta)R_{e1}$ remain effective at the (junction) input.

The impedance Z_i as given by Eq. 3.5.8 is effective between the base–emitter junction and the ground. Unfortunately, no direct access is possible to the junction, because from there to the base terminal we have the base spread resistance r_{b1}. This means that r_{b1} must be subtracted from R_c to get the proper value of the compensating resistor. Supposing that $r_{b1} = 25\,\Omega$, the proper compensating resistor is simply:

$$R_c' = R_c - r_{b1} = 385 - 25 = 360\,\Omega \tag{3.5.41}$$

The complete input circuit is shown in Fig. 3.5.3; the input impedance components which are reflected from the emitter to base, are shown in the shaded area.

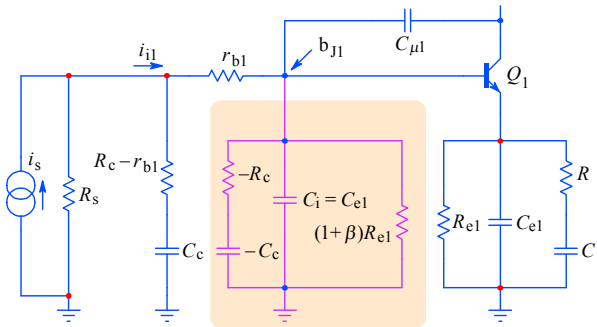


Fig. 3.5.3: The impedances in the emitter are reflected into the base junction of Q_1. The emitter peaking components R and C are reflected into **negative** elements $-R_c$ and $-C_c$, which must be compensated by adding externally an equal and positive R_c and C_c; for proper compensation r_{b1} must be subtracted from R_c.

The compensation of the input impedance is mandatory if we intend to apply an inductive peaking network at the input of that amplifying stage. The equivalent input capacitance, which will be seen as the load by the peaking circuit, is the capacitance from the transistor base–emitter junction to ground, $C_i + C_M$.

However, as mentioned before, these capacitances are seen with the base resistance r_{b1} in series. Also there is a parasitic base lead inductance, in addition to the length of PCB trace and pads having their own stray inductance and capacitance.

Therefore the inductive peaking circuit at the input of a transistor amplifying stage does not see a pure capacitance as its load. Special ‘tricks of trade’ must be applied, e.g., a modified T-coil peaking at the amplifier input, which we shall discuss in the next section.

3.6 Transistor Interstage T-coil Peaking

In Part 2 we have shown that the greatest bandwidth extension is achieved by using T-coil peaking. The analysis was based on the assumption that the T-coil tap was loaded by a pure capacitance. Unfortunately, for a transistor amplifier this is not the case. In case of a cascode amplifier the emitter network, formed by the parallel connection of $R_e C_e = \tau_T$, is reflected into the base circuit as a parallel connection of $(1+\beta)R_e$ and C_e, paralleled also by the Miller capacitance C_M; to this the series base spread resistance r_b must be added.

In Fig. 3.6.1a we draw such a stage [Ref. 3.5]. Since in the following analysis we do not need the transistor Q_2, we shall consider its emitter input as an ideal ground for the Q_1 collector current signal (of course, the value of the Miller capacitance C_M has to be calculated by considering the actual Q_2 emitter impedance). Thus we can drop the index '1', as all the parameters will belong to Q_1 only. To further simplify the analysis, we shall neglect the damping and thermal compensation impedances Z_d and Z_θ, as well as the emitter peaking. The resistor R_L is the T-coil loading resistor, which is also the load of the driving stage and we shall assume that its other end is also connected to an ideal AC ground. Fig. 3.6.1b shows the T-coil loaded by the equivalent small signal, high frequency input impedance of the transistor Q_1.

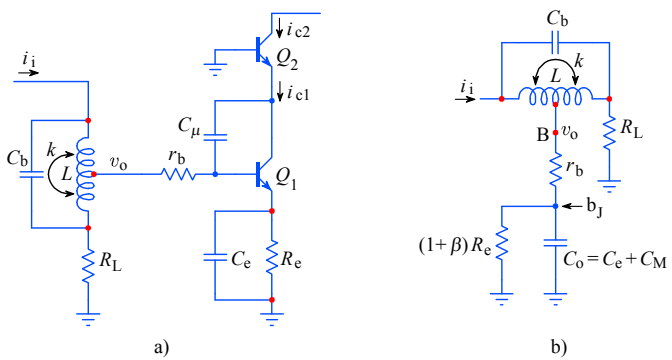


Fig. 3.6.1: a) The cascode amplifier with a T-coil interstage network. b) The T-coil loaded by the equivalent small signal, high frequency input impedance.

To prevent any confusion we must stress that C_b is the T-coil bridging capacitance, and not the base capacitance which is represented by $C_o = C_e + C_M$. As we know from the symmetrical T-coil analysis in Part 2, Sec. 2.4:

$$L = L_a + L_b = R_L^2\, C_o \tag{3.6.1}$$

Since the input shunt resistance $(1+\beta)R_e$ is usually much higher than r_b, we shall neglect it also, thus arriving at the circuit in Fig. 3.6.2a. Fig. 3.6.2b shows the equivalent T-coil circuit in which we have replaced the magnetic field coupling factor k with the negative mutual inductance $-L_M$ and the coil branches by their equivalent inductances L_a and L_b. Finally, in Fig. 3.6.2c we have replaced the branch impedances by symbols ***A*** to ***E*** to determine the three current loops I_1, I_2, I_3.

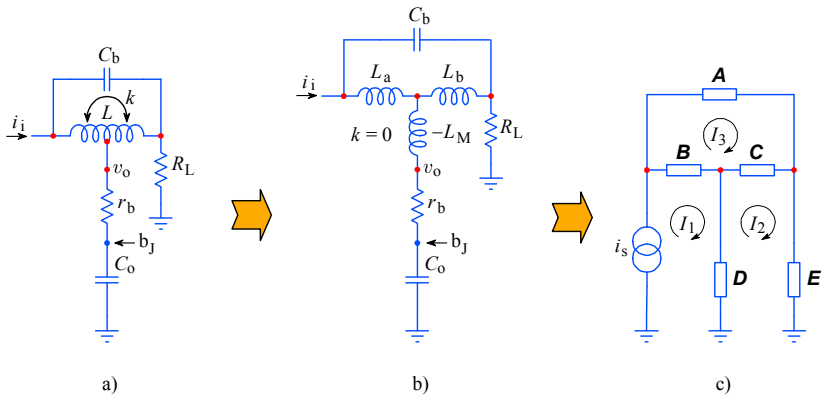


Fig. 3.6.2: a) The T-coil loaded by the simplified input impedance; b) the equivalent T-coil circuit in which k is substituted by $-L_M$; c) the equivalent branch impedances and the three current loops.

By comparing Fig. 2.4.1b,c, Sec. 2.4 with Fig. 3.6.2b,c we see that they are almost equal, except that in the branch ***D***, we have the additional series resistance r_b. Let us list all these impedances again, but now including r_b:

$$\begin{aligned} \boldsymbol{A} &= \frac{1}{s\,C_b} \\ \boldsymbol{B} &= s L_a \\ \boldsymbol{C} &= s L_b \\ \boldsymbol{D} &= -s\,L_M + r_b + \frac{1}{s\,C_o} \\ \boldsymbol{E} &= R_L \end{aligned} \qquad (3.6.2)$$

The general analysis of the branches, Eq. 2.4.6 – 2.4.13, showed that the input impedance of the T-coil network is equal to its loading impedance $Z_i = \boldsymbol{E} = R_L$. As we shall soon see, r_b between the T-coil tap and C_o spoils this nice property; we shall have to compensate it. The analysis here is similar to that in Sec. 2.4, so we do not have to repeat it. Here we give the final result, Eq. 2.4.14, for convenience:

$$\boldsymbol{BCA} + \boldsymbol{BDA} + \boldsymbol{BEA} + \boldsymbol{DCA} - \boldsymbol{ECA} - \boldsymbol{E}^2\boldsymbol{A} - \boldsymbol{E}^2\boldsymbol{B} - \boldsymbol{E}^2\boldsymbol{C} = 0 \qquad (3.6.3)$$

By entering all substitutions from Eq. 3.6.2, performing all the required multiplications and arranging the terms in the decreasing powers of s, we obtain:

$$s\left[\left(\frac{L_a L_b}{C_b} - \frac{L\,L_M}{C_b}\right) - R_L^2\,L\right] + \frac{L\,r_b}{C_b} + \frac{R_L}{C_b}(L_a - L_b) + \frac{1}{s}\left(\frac{L}{C_o\,C_b} - \frac{R_L^2}{C_b}\right) = 0 \qquad (3.6.4)$$

or, more simply:

$$s\,K_1 + K_2 + s^{-1}K_3 = 0 \qquad (3.6.5)$$

The difference between Eq. 3.6.4, 3.6.5 and Eq. 2.4.15, 2.4.16 is in the middle term. Again, if we want to have an input impedance independent of the frequency s, then each of the coefficients K_1, K_2, and K_3 must be zero [Ref. 3.5]:

$$K_1 = \frac{-\,L\,L_\mathrm{M}}{C_\mathrm{b}} + \frac{L_\mathrm{a}\,L_\mathrm{b}}{C_\mathrm{b}} - R_\mathrm{L}^2\,L = 0$$

$$K_2 = \frac{L\,r_\mathrm{b}}{C_\mathrm{b}} + \frac{R_\mathrm{L}}{C_\mathrm{b}}\,(L_\mathrm{a} - L_\mathrm{b}) = 0$$

$$K_3 = \frac{L}{C_\mathrm{o}\,C_\mathrm{b}} - \frac{R_\mathrm{L}^2}{C_\mathrm{b}} = 0 \qquad (3.6.6)$$

So we have the three equations from which we can calculate the parameters L_a, L_b, and L_M. By considering Eq. 3.6.1 we obtain:

$$L_\mathrm{a} = \frac{L}{2}\left(1 - \frac{r_\mathrm{b}}{R_\mathrm{L}}\right) = \frac{R_\mathrm{L}^2\,C_\mathrm{o}}{2}\left(1 - \frac{r_\mathrm{b}}{R_\mathrm{L}}\right) \qquad (3.6.7)$$

$$L_\mathrm{b} = \frac{L}{2}\left(1 + \frac{r_\mathrm{b}}{R_\mathrm{L}}\right) = \frac{R_\mathrm{L}^2\,C_\mathrm{o}}{2}\left(1 + \frac{r_\mathrm{b}}{R_\mathrm{L}}\right) \qquad (3.6.8)$$

$$L_\mathrm{M} = \frac{L}{4}\left(1 - \frac{r_\mathrm{b}^2}{R_\mathrm{L}^2}\right) - R_\mathrm{L}^2\,C_\mathrm{b} = \frac{R_\mathrm{L}^2\,C_\mathrm{o}}{4}\left(1 - \frac{r_\mathrm{b}^2}{R_\mathrm{L}^2}\right) - R_\mathrm{L}^2\,C_\mathrm{b} \qquad (3.6.9)$$

Two interesting facts become evident from Eq. 3.6.7 and 3.6.8. First, $L_\mathrm{a} < L_\mathrm{b}$, and this means that **the coil tap is not at the coil's center** any longer, but it is moved towards the coil's signal input node. Secondly, R_L must always be larger than r_b, otherwise L_a becomes negative. But we reach the limit of realizability long before that, since we know from Part 2 that $L_1 = L_\mathrm{a} - L_\mathrm{M}$ (and also $L_2 = L_\mathrm{b} - L_\mathrm{M}$).

In Eq. 3.6.9 we have two unknowns, L_M and C_b; therefore we need a fourth equation to calculate them. Similarly as we did in Part 2, Sec. 2.4, we shall use the transimpedance equation for this purpose. The procedure is well described from Eq. 2.4.20 to 2.4.24 and we write the last one again:

$$\frac{V_\mathrm{o}}{I_1} = \frac{1}{s\,C_0} \cdot \frac{\boldsymbol{CA + EA + EB + EC}}{\boldsymbol{CA + CB + DA + DB + DC + EA + EB + EC}} \qquad (3.6.10)$$

If we insert the substitutions from Eq. 3.6.2, we obtain the following result:

$$F(s) = \frac{V_\mathrm{o}}{I_1} = \frac{R_\mathrm{L}}{s^2\,R_\mathrm{L}^2\,C_\mathrm{o}\,C_\mathrm{b} + s\,C_\mathrm{o}\,\dfrac{R_\mathrm{L} + r_\mathrm{b}}{2} + 1} \qquad (3.6.11)$$

In a similar way, for the transimpedance from the input to R_L we would obtain:

$$\frac{V_\mathrm{R}}{I_\mathrm{i}} = R_\mathrm{L}\,\frac{s^2\,R_\mathrm{L}^2\,C_\mathrm{o}\,C_\mathrm{b} - s\,C_\mathrm{o}\,\dfrac{R_\mathrm{L} - r_\mathrm{b}}{2} + 1}{s^2\,R_\mathrm{L}^2\,C_\mathrm{o}\,C_\mathrm{b} + s\,C_\mathrm{o}\,\dfrac{R_\mathrm{L} + r_\mathrm{b}}{2} + 1} \qquad (3.6.12)$$

Since we have the factor $(R_L - r_b)$ in the numerator and a different factor $(R_L + r_b)$ in the denominator, this means that **the two zeros are not symmetrically placed** in relation to the two poles in the s-plane. Therefore Eq. 3.6.12 does not describe an all pass network and the input impedance is not simply R_L as before. This represents the basic obstacle to using T-coils in a transistor distributed amplifier, because the T-coil load can not be replaced by another T-coil network (for comparison see [Ref. 2.18 and 2.19]).

Eq. 3.6.11 has two poles, which we calculate from the canonical form of the denominator:

$$s^2 + s\,\frac{R_L + r_b}{2\,R_L^2\,C_b} + \frac{1}{R_L^2\,C_o C_b} = 0 \tag{3.6.13}$$

and both poles are:

$$s_{1,2} = -\frac{R_L + r_b}{4\,R_L^2\,C_b} \pm \sqrt{\left(\frac{R_L + r_b}{4\,R_L^2\,C_b}\right)^2 - \frac{1}{R_L^2\,C_o C_b}} \tag{3.6.14}$$

or, by extracting the common factor:

$$s_{1,2} = \frac{1 + r_b/R_L}{4\,C_b R_L}\left(-1 \pm \sqrt{1 - \frac{16\,C_b}{C_o\,(1 + r_b/R_L)^2}}\right) \tag{3.6.15}$$

An efficient inductive peaking must have complex poles. For Bessel poles, as was in Fig. 3.5.2b, the pole angles $\theta_{1,2} = \pm 150°$ and with this pole arrangement we obtain the MFED response. If the poles are complex the tangent of the pole angle is the ratio of the imaginary to the real component of Eq. 3.6.15:

$$\tan\theta = \frac{\Im\{s_1\}}{\Re\{s_1\}} = \sqrt{\frac{16\,C_b}{C_o\,(1 + r_b/R_L)^2} - 1} \tag{3.6.16}$$

By solving this equation for C_b we obtain:

$$C_b = C_o\,\frac{1 + \tan^2\theta}{16}\left(1 + \frac{r_b}{R_L}\right)^2 \tag{3.6.17}$$

Compared to the symmetrical T-coil, here we have the additional factor $(1 + r_b/R_L)^2$. For Bessel poles $\theta = 150° = 5\pi/6$ and $\tan^2\theta = 1/3$, thus for a single stage case:

$$C_b = \frac{C_o}{12}\left(1 + \frac{r_b}{R_L}\right)^2 \tag{3.6.18}$$

If we replace C_b in Eq. 3.6.9 with Eq. 3.6.18, the mutual inductance is:

$$L_M = R_L^2\,C_o\left[\frac{1}{4}\left(1 - \frac{r_b^2}{R_L^2}\right) - \frac{1}{12}\left(1 + \frac{r_b}{R_L}\right)^2\right] \tag{3.6.19}$$

With this we can calculate the coupling factor k between the coil L_1 and L_2 [Ref. 3.23]:

$$k = \frac{L_M}{\sqrt{L_1\,L_2}} = \frac{L_M}{\sqrt{(L_a - L_M)\,(L_b - L_M)}} \tag{3.6.20}$$

Now we have all the equations needed for the T-coil transistor interstage coupling.

3.6.1 Frequency response

To calculate the frequency response we apply Eq. 3.6.11, in which we shall replace the coil's bridging capacitance C_b by Eq. 3.6.18, since we shall discuss only the MFED response. Then we put it into the canonical form by factoring out the common factor:

$$F(s) = \frac{R_L^2\, C_o^2 \left(1 + \frac{r_b}{R_L}\right)^2}{12} \cdot \frac{R_L}{s^2 + s\,\frac{6}{R_L\, C_o}\left(1 + \frac{r_b}{R_L}\right) + \frac{12}{R_L^2\, C_o^2}\left(1 + \frac{r_b}{R_L}\right)} \tag{3.6.21}$$

The denominator of the second fraction has two roots:

$$s_{1,2} = \frac{1}{R_L\, C_o\, (1 + r_b/R_L)}\,(-3 \pm j\sqrt{3}\,) = \sigma_1 \pm j\,\omega_1 \tag{3.6.22}$$

Sometimes we prefer the normalized form of the roots and in this case $R_L C_o = 1/\omega_h = 1$. To emphasize the normalization, we add the subscript 'n', so $s_{1,2n} = \sigma_{1n} \pm j\,\omega_{1n}$. By applying the normalized poles of Eq. 3.6.22 to Eq. 2.2.27, which is a generalized second-order magnitude function, we obtain:

$$|F(\omega)| = \frac{\sigma_{1n}^2 + \omega_{1n}^2}{\sqrt{\left[\sigma_{1n}^2 + \left(\frac{\omega}{\omega_h} + \omega_{1n}\right)^2\right]\left[\sigma_{1n}^2 + \left(\frac{\omega}{\omega_h} - \omega_{1n}\right)^2\right]}} \tag{3.6.23}$$

By comparing the Bessel poles for a simple T-coil (Eq. 2.4.42) with Eq. 3.6.22, we notice that in the denominator we have an additional factor $(1 + r_b/R_L)$. Therefore it would be interesting to make several frequency responses with different ratios r_b/R_L, as listed in Table 3.6.1:

Table 3.6.1

r_b/R_L	σ_{1n}	ω_{1n}	Note
0.00	− 3.0	− 1.732	symmetrical T-coil
0.25	− 2.4	− 1.386	-
0.50	− 2.0	− 1.155	-

The corresponding frequency-response plots are drawn in Fig. 3.6.3, together with the three non-peaking responses ($L = 0$) as references. The bandwidth improvement factor for all three cases is $\eta_b = 2.72$. That is because the base resistance r_b decreases the bandwidth also for the non-peaking stage, where $L = 0$. We can prove this if we multiply the denominator of Eq. 3.6.22 by R_L:

$$s_{1,2} = \frac{1}{C_o\,(R_L + r_b)}\,(-3 \pm j\sqrt{3}\,) = \omega_h\,(-3 \pm j\sqrt{3}\,) \tag{3.6.24}$$

where $\omega_h = 1/C_o(R_L + r_b)$ is the non-peaking bandwidth considering r_b. So we shall obtain three different curves for three different ratios r_b/R_L, also for $L = 0$.

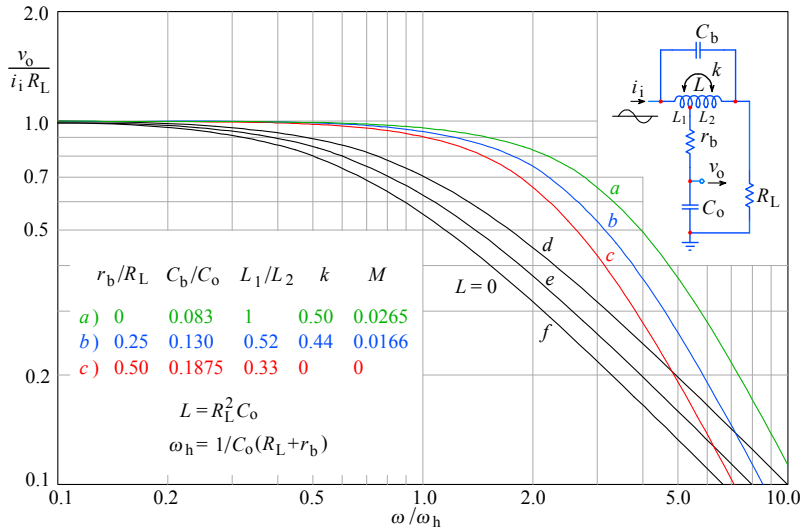


Fig. 3.6.3: MFED frequency response of the T-Coil transistor interstage coupling circuit for three different values of r_b: a) $r_b = 0$; b) $r_b = 0.25\,R_L$; c) $r_b = 0.5\,R_L$. For comparison the three reference cases ($L = 0$): d), e), and f), which correspond to the same three r_b/R_L ratios, are drawn. The bandwidth improvement factor of the peaking system remains 2.72 times over the non-peaking reference for each value of r_b.

From the analysis above, we can draw an important result:

The upper half-power frequency of a non-peaking transistor amplifier must be calculated by taking into account the sum $R_L + r_b$ (and not just R_L).

3.6.2 Phase Response

To calculate the phase response we insert our poles into Eq. 2.2.31:

$$\varphi = \arctan\frac{\omega/\omega_h + \omega_{1n}}{\sigma_{1n}} + \arctan\frac{\omega/\omega_h - \omega_{1n}}{\sigma_{1n}} \qquad (3.6.25)$$

In Fig. 3.6.4 the phase plots for the same three ratios of r_b/R_L as in the frequency response are shown, along with the three references ($L = 0$).

3.6.3 Envelope Delay

The envelope delay is calculated using Eq. 2.2.35:

$$\tau_e\,\omega_h = \frac{\sigma_{1n}}{\sigma_{1n}^2 + (\omega/\omega_h + \omega_{1n})^2} + \frac{\sigma_{1n}}{\sigma_{1n}^2 + (\omega/\omega_h - \omega_{1n})^2} \qquad (3.6.26)$$

and the responses are drawn in Fig. 3.6.5, for the three different ratios r_b/R_L, in addition to the three references ($L = 0$).

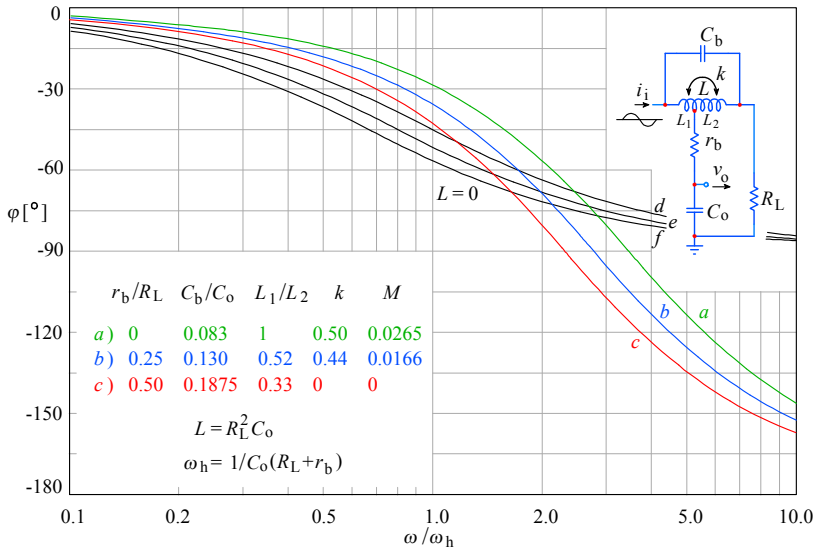


Fig. 3.6.4: MFED phase response of the T-Coil transistor interstage coupling circuit compared with the references ($L = 0$), for the same three values of r_b/R_L.

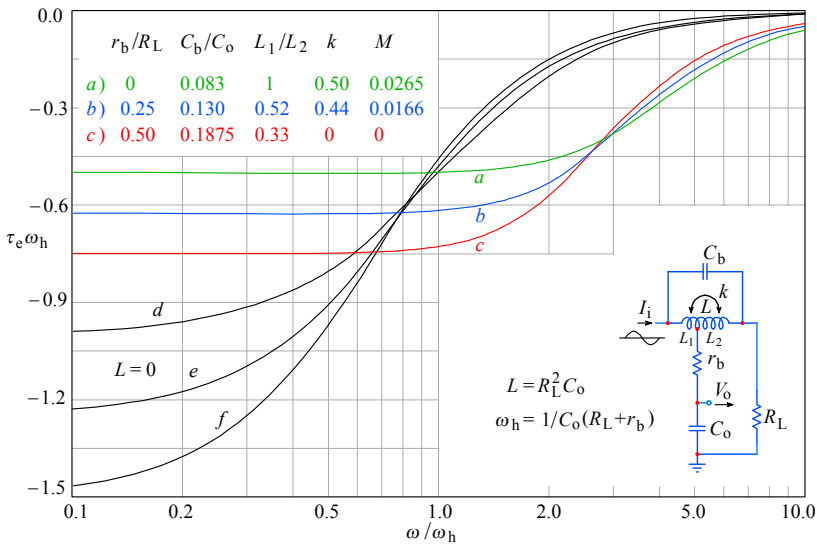


Fig. 3.6.5: MFED envelope delay response of the T-Coil transistor interstage coupling circuit compared with the references ($L = 0$) for the same three values of r_b/R_L.

3.6.4 Step Response

For plotting the step response we can use Eq. 2.4.47 (which was fully derived in Part 1, Eq. 1.14.18):

$$g(t) = 1 + \frac{1}{|\sin\theta|}\, e^{\sigma_1 t} \sin(\omega_1 t + \theta + \pi) \tag{3.6.27}$$

where θ is the pole angle according to Fig. 3.5.2b. By inserting the pole angle $\theta = 150°$ or $5\pi/6$, as required by the 2nd-order Bessel system, we obtain:

$$g(t) = 1 + 2\, e^{-3\,t/T} \sin\left[\sqrt{3}\ t/T + \left(1 + \frac{5}{6}\right)\pi\right] \tag{3.6.28}$$

However, here we must use $T = 1/\omega_h = C_o\,(R_L + r_b)$, as in Eq. 3.6.22 and 3.6.24. In Fig. 3.6.6 the step-response plots are drawn for three different ratios r_b/R_L as well as the three reference cases with $L = 0$.

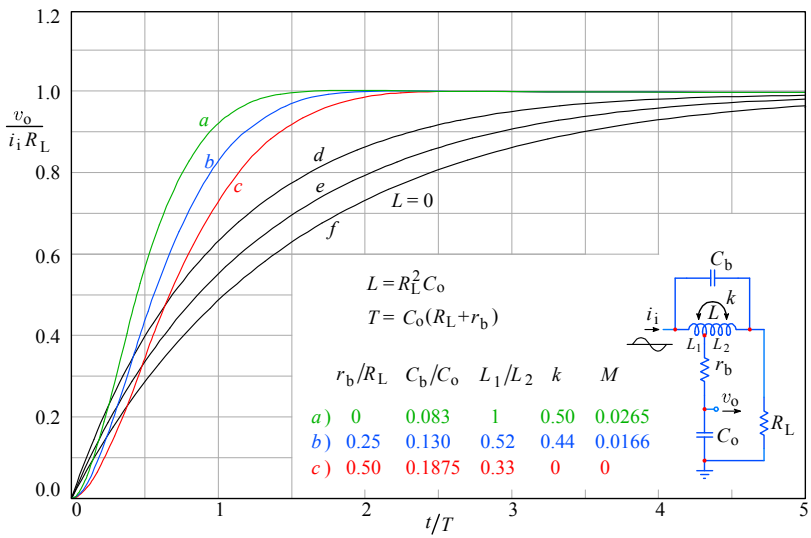


Fig. 3.6.6: MFED step-response of the T-coil transistor interstage coupling circuit, compared to the references ($L = 0$), for the same three values of r_b/R_L.

Thus we have completed the analysis of the basic case of a transistor T-coil interstage coupling. The reader who would like to have more information should study [Ref. 3.5]. In order to simplify the analysis, we have purposely neglected the transistor input resistance $\beta\,(r_\pi + 1)$ and also stray inductance L_s of the tap to transistor base terminal. In the next steps we shall discuss both of them.

3.6.5 Consideration of the transistor's input resistance

Fig. 3.6.7 shows the basic configuration of the transistor input circuit. We have also drawn the base-lead inductance L_s which will be discussed in the next section.

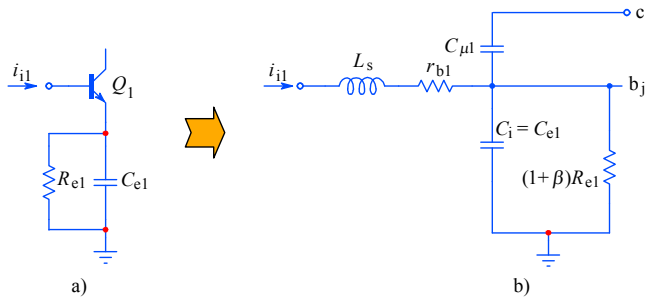


Fig. 3.6.7: The complete transistor input impedance: a) schematic; b) equivalent circuit, in which we include also the base lead inductance L_{b1}. The presence of the shunt resistance $(1 + \beta)\, R_{e1}$ requires a modified interstage T-coil circuit.

The resistance from the base–emitter junction to ground is:

$$R_i = (1 + \beta)\, R_e \tag{3.6.29}$$

as we have derived in Eq. 3.2.15. The effect of this resistance may be canceled if we insert an appropriate resistor R_s from the end of coil L_1 to the start of coil L_2, as shown in Fig. 3.6.8. It is essential that the resistor R_s is inserted on the 'left' side of L_2 (at the T-coil tap node), because in this case the bridging capacitance of C_b (self-capacitance of the coil) and the magnetic field coupling (k) are utilized. With the resistor placed on the 'right' side of L_2 (at the $R_L\, C_b$ node) , that would not be the case.

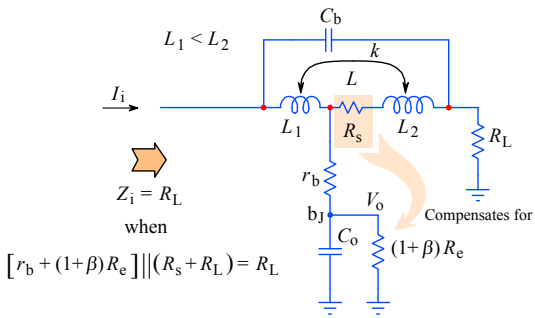


Fig. 3.6.8: The resistance R_s in series with L_2 is inserted near the T-coil tap to compensate the error in the impedance seen by the input current at low frequencies, owed to the parallel connection of $R_L \,||\, [\, r_b + (1 + \beta) R_e]$.

At very high frequencies we can replace all capacitors by short circuit and all inductors by open circuit. In this case the input resistance of the T-coil circuit is R_L. But

at very low frequencies the capacitors represent an open circuit and the inductors a short circuit. The transistor input resistance is then effectively in parallel with R_L. It is the task of the series resistor R_s to prevent this reduction of resistance. The idea is that:

$$(r_b + R_i) \,||\, (R_s + R_L) = R_L \tag{3.6.30}$$

If we solve this for R_s we obtain:

$$R_s = \frac{R_L^2}{R_i + r_b - R_L} \tag{3.6.31}$$

The introduction of this resistor spoils all the expressions from our previous analysis, and, to be exact, everything we derived to determine the basic T-coil parameters should be calculated anew, considering the additional parameter R_s. Since in practice the value of this resistor is very small, no substantial changes in other circuit parameters may be expected and the additional effort, which would be required by an exact analysis, would be worthless.

A sneaky method for implementing this compensation, while at the same time decreasing the stray capacitance (and also to create difficulties to the competition to copy the circuit) is to make the coil L_2 using an appropriate resistive wire.

3.6.6 Consideration of the base lead stray inductance

Fig. 3.6.9a shows the T-coil with the base lead stray inductance L_s at the tap. From Fig. 3.6.9b we realize that the **positive** inductance of the base lead L_s actually decreases the **negative** mutual inductance L_M of the T-coil. To retain the same conditions as in Fig. 3.6.2c at the beginning of the basic T-coil analysis, the coupling factor must be increased, thus increasing the mutual inductance to $L_M + L_s$.

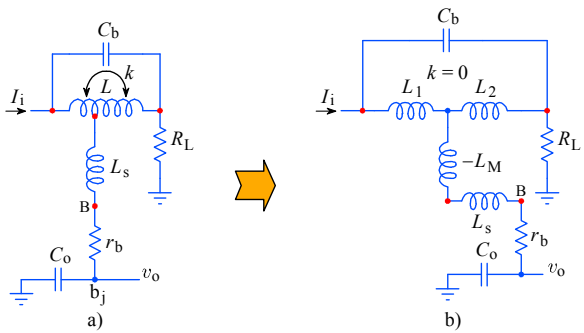


Fig. 3.6.9: a) The base lead inductance L_s decreases the value of mutual inductance, as indicated by the equivalent circuit in b). This can be compensated by recalculating the circuit with an increased coupling factor.

We shall mark the new T-coil circuit parameters with a prime (′) to distinguish them from the original parameters of the transistor interstage T-coil:

$$L'_M = L_M + L_s \tag{3.6.32}$$

Because the inductance from r_b to either end of the coil L is now increased by L_s, both inductances L_a and L_b must be decreased by the value of L_s:

$$L'_a = L_a - L_s \quad \text{and} \quad L'_b = L_b - L_s \tag{3.6.33}$$

By considering all these changes, the new (larger) value of the coupling factor is

$$k' = \frac{L'_M}{\sqrt{(L_a - L_M - 2\,L_s)\,(L_b - L_M - 2\,L_s)}} \tag{3.6.34}$$

Since it is sometimes difficult to achieve the required coupling factor we must take care that the base lead stray inductance L_s is as small as possible.

3.6.7 Consideration of the collector to base spread capacitance

So far we have considered only the lumped collector to base capacitance C_μ. However, in a real transistor the capacitance C_μ is spread along the base resistance r_b, as is drawn in Fig. 3.6.10a [Ref. 3.4]. For the analysis of such a circuit we should know the actual geometry involved; unless we are designing the transistor by ourselves, this would be difficult to find out. So we shall, rather, approximate this by splitting C_μ into three parts, $C_{\mu 1}$, $C_{\mu 2}$, and $C_{\mu 3}$, as suggested in Fig. 3.6.10b, adding also a constant value C_s to account for the external leads and PCB strays.

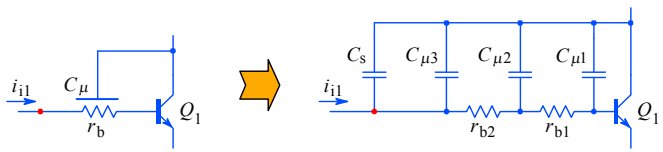


Fig. 3.6.10: a) The base–collector reverse capacitance C_μ is actually spread across the base resistance r_b. b) A good approximation is achieved by splitting r_b in half and C_μ in three parts, adding also the external stray capacitance C_s.

Those readers who are interested in the results and further suggestions, should study [Ref. 3.20].

3.6.8 The 'Folded' Cascode

While we are still speaking about cascode amplifiers, let us examine the 'folded' cascode circuit, Fig. 3.6.11. This circuit is a handy solution in cases of a limited supply voltage, a situation commonly encountered in modern integrated circuits and battery supplied equipment.

The first thing to note is that the collector DC currents can be different, since the bias conditions for Q_1 are set by the input base voltage and R_{e1} , whilst for Q_2 the bias is set by $V_{cc} - V_{b2}$ and R_{cc} .

Another interesting point is that R_{cc} (or a current source in its place) must supply the current for both transistors. Therefore when a signal is applied at the input the currents in Q_1 and Q_2 will be in anti-phase, i.e., when i_{o1} increases, i_{i2} decreases and vice-versa. Consequently, it is easier to achieve good thermal balance with such a circuit, than with the original cascode.

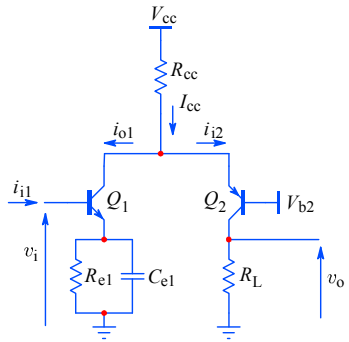


Fig. 3.6.11: The 'folded' cascode is formed by a complementary, NPN and PNP, transistor pair, connected in the otherwise usual cascode configuration. Since thermionic devices are not produced in complementary pairs, this circuit can not be realized with electronic tubes.

In an integrated circuit it is always more difficult to make fast PNP transistors, because of the lower mobility of the virtual positive charge (a vacant charge region, left behind an accelerated electron, can exist for a considerable time before another slow electron comes near enough to be trapped by it). It could then be advantageous to change the supply polarity and use a PNP type for Q_1 and an NPN type for Q_2.

In all other respects, the circuit presents problems identical to those of the original cascode, so all the solutions already discussed apply here as well.

3.7 Differential Amplifiers

In Sec. 3.4.2 we have explained that the instability of the transistor's DC bias depends on the ambient temperature and the heat generated internally, as a consequence of its power dissipation. The current amplification factor β_0 also depends on temperature. These effects multiply in a multi-stage DC amplifier. They can be greatly reduced using a symmetrical differential amplifier.

The basic differential amplifier is shown in Fig. 3.7.1. The input voltage of transistor Q_1 is v_{i1} and that of Q_2 is $v_{i2} = -v_{i1}$ (we are assuming these symmetrical driving voltages in order to eliminate any common mode voltages, thus simplifying the initial analysis). The emitters of both transistors are connected together and fed via the resistor R_{ee} from the voltage $V_{ee} < 0$. If we assume the circuit to be entirely symmetrical, e.g., $Q_1 = Q_2$ and $R_{L1} = R_{L2}$, and if both input voltages v_{i1} and v_{i2} are zero, the DC Output voltages are also equal, $V_{o1} = V_{o2}$, independently of the ambient temperature.

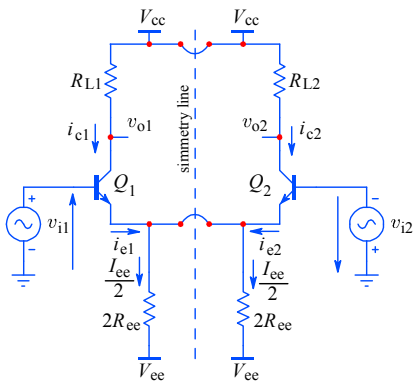


Fig. 3.7.1: The differential amplifier. We simplify the initial analysis by assuming $v_{i2} = -v_{i1}$, $R_{L1} = R_{L2}$ and $Q_1 = Q_2$ (all parameters).

The name *differential amplifier* suggests that we are interested in the amplification of voltage differences. In general, if one signal input voltage, say, v_{i1}, goes positive, the other input voltage v_{i2} goes negative by the same amount (we have accounted for this by drawing the polarity of the voltage generator v_{i2} in Fig. 3.7.1 opposite to v_{i1}). This means that any increase of the emitter current in transistor Q_1 is accompanied by an equal decrease in the emitter current in transistor Q_2. So the current through the resistor R_{ee} and the voltage at the emitter node are not changed. Therefore we can consider the emitter node as a virtual ground. The difference of the input voltages is:

$$v_{i1} - (-v_{i2}) = v_{i1} + v_{i2} \tag{3.7.1}$$

In a similar way as the input voltages, the signal output voltages v_{o1} and v_{o2} go up and down for an equal amount, however, we must account for the signal inversion in

the common emitter amplifier. If the voltage amplification of the input voltage difference is A_{vd} (which we can take directly from Eq. 3.1.14, where we discussed a simple common-base amplifier), the output signal voltage difference is:

$$v_{o2} - (-v_{o1}) = v_{o1} + v_{o2} = -A_{vd}\,(v_{i1} + v_{i2}) \tag{3.7.2}$$

An attentive reader will note that we have added the subscript 'd' to denote the differential mode gain.

In the case of both input voltages being equal and of the same polarity, both output voltages will also be equal and of same polarity; however, the output signal's polarity is the inverse of the input signal's polarity (owing to the 180° phase inversion of each common emitter amplifier stage). If the symmetry of the circuit were perfect, the output voltage difference would be zero, provided that the common mode excitation at the input remains well within the linear range of the amplifier. Such operation is named *common-mode amplification*, A_{vc} (here we have added the subscript 'c').

For the common mode signal the excursion of both output voltages with respect to their DC value is:

$$v_{o1} = v_{o2} = -A_{vc}\,\frac{v_{i1} + v_{i2}}{2} \approx -\frac{R_{L1}}{2\,R_{ee}}\,(v_{i1} + v_{i2}) \tag{3.7.3}$$

A good idea to visualize the common mode operation is to 'fold' the circuit across the symmetry line and consider it as a 'single ended' amplifier with both transistors and both loading resistors in parallel.

Since we are more interested in the differential mode amplification, we have pulled the expression for the common mode amplification, so to say, 'out of the hat'. The analysis so far is more intuitive than exact. The reason we did not bother to make a corresponding derivation of exact formulae is that the simple circuit shown in Fig. 3.7.1 is almost never used as a wideband amplifier owing to its large input capacitance. A basic differential amplifier in cascode configuration, with a constant current generator instead of the resistor R_{ee}, is drawn in Fig. 3.7.2. The reader who wants to study the full analysis of the basic low-frequency differential amplifier according to Fig. 3.7.1 should look in [Ref. 3.7]. We shall return briefly to the differential amplifier in Part 5.

3.7.1 Differential cascode amplifier

The analysis of the circuit in Fig. 3.7.2, a differential cascode amplifier, with the same rigor, considering all the complex impedances, would quickly run out of control owing to its complexity (remember Table 3.2.1, which should be applied to each transistor). However, if we take the emitter node to be a virtual ground, each half of the differential amplifier can be analyzed separately; actually, owing to the symmetry, the analysis of only one arm is needed, which has been done already in Sec. 3.4. So here we can focus on other problems which are peculiar to differential amplifiers only.

First of all, no differential amplifier is perfectly symmetrical, even if all of its transistors are on the same chip. The lack of a perfect symmetry causes the common mode input signal $(v_{i1} + v_{i2})/2$ to appear partly as a differential output signal. For the

same reason there is still some temperature drift, although greatly reduced in comparison with the single ended amplifier. The appearance of common mode signals at the output is especially annoying in electrobiological amplifiers (electrocardiographs and electroencephalographs). In these amplifiers, very small input signal differences (of the order of several μV) must be amplified in the presence of large (up to 1V) common mode signals from power lines, owing to capacitive pickup. The level of ability of the differential amplifier to reject the common mode signal is called *common mode rejection ratio*, CMRR $= A_{vc}/A_{vd}$, generally expressed in decibels (dB). Since this is out of scope of this book, we shall not pursue these effects in detail.

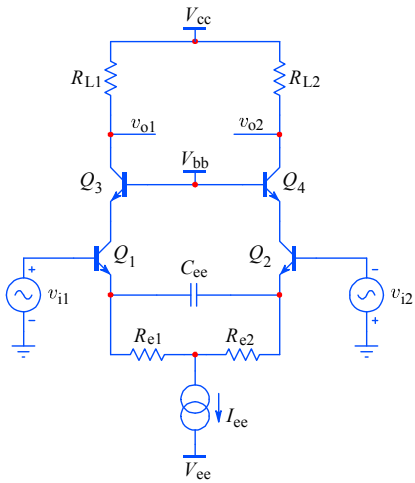


Fig. 3.7.2: The basic circuit of the differential cascode amplifier.

The optimum thermal stability of the differential cascode circuit could again be obtained by adjusting the quiescent currents in both halves of the differential amplifier to values such that the voltage drop on each loading resistor is equal to the voltage $(V_{cc} - V_{bb} - V_{be})/2$, (see Eq. 3.4.19 and the corresponding explanation). However, as has been said for the simple cascode amplifier, the requirements for large bandwidth will prevent this from being realized. We would want to have low R_L, high V_{cc} and V_{bb}, and high I_{ee} to maximize the bandwidth. So the thermal stability will have to be established in a different way.

Differential amplifiers are particularly suitable for compensation of many otherwise insolvable errors. This is achieved by cross-coupling and adding anti-phase signals, so that the errors cancel. For example, the pre-shoot of the simple cascode amplifier, which is owed to capacitive feed through, can be effectively eliminated if two capacitors with the same value as $C_{\mu 1,2}$ are connected from the Q_1 emitter to the Q_2 collector, and vice-versa.

Similarly, by cross-coupling diodes or transistors we can achieve nonlinearity cancellation, leakage current compensation, better gain stability, DC stability, etc. In integrated circuits, even production process variations can be compensated in this way. Some such examples are given in Part 5.

3.7.2 Current source in the emitter circuit

To improve the common mode rejection, instead of having a large resistor R_{ee} and a correspondingly high voltage V_{ee} (resulting in an unnecessarily high power dissipation $P \approx V_{ee}^2/R_{ee}$) the common practice is to use high *incremental* resistance. This is usually done by a special connection of active devices in a configuration called *current source*. Such generators can have a high incremental resistance even when working with a relatively low V_{ee}.

The ideal current generator should be independent as much as possible from the applied voltage and from ambient temperature. A simple current sink, composed of two equal transistors Q_5 and Q_6 is shown in Fig. 3.7.3a. (We have given them the indices '5' and '6' in order to avoid any confusion with the former figures). The circuit is named the *current mirror* [Ref. 3.31], because the collector current I_{c6} is in some respect a 'mirror image' of I_R, as shown in Fig. 3.7.3b, where the current symmetry analysis is performed by normalizing the currents to those of the base. Current mirrors are in widespread use in integrated circuits in which complex multi-stage biassing is controlled from a single point.

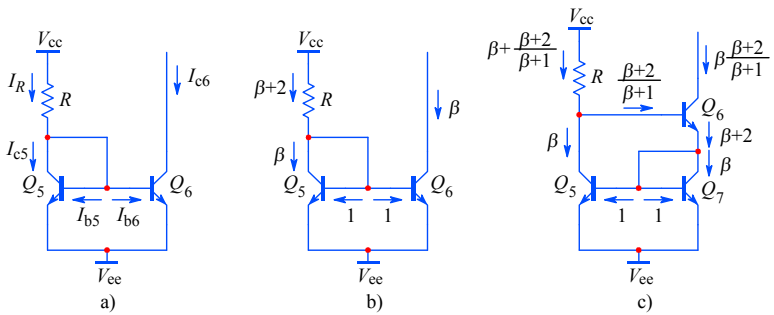


Fig. 3.7.3: a) The basic current mirror. b) Current symmetry analysis with the currents normalized to those of the base. c) The symmetry is improved in the Wilson mirror.

The current I_R is:

$$I_R = I_{c5} + I_{b5} + I_{b6} = I_{c5} + \frac{I_{c5}}{\beta_5} + \frac{I_{c6}}{\beta_6} \tag{3.7.4}$$

If both transistors are identical, then $\beta_5 = \beta_6 = \beta$ and $I_{c5} = I_{c6} = \beta I_b$. In this case:

$$I_R = I_{c5}\left(1 + \frac{2}{\beta}\right) \tag{3.7.5}$$

and the collector current is:

$$I_{c5} = \frac{I_R}{1 + \dfrac{2}{\beta}} \tag{3.7.6}$$

If β is very large then:

$$I_{c6} \approx I_R = \frac{V_{cc} - V_{be}}{R} \tag{3.7.7}$$

In general the collector current of a transistor is [Ref. 3.4]:

$$I_c = I_s \, e^{\frac{V_{be}}{V_T}} \left(1 + \frac{V_{ce}}{V_A + V_{ce}}\right) \tag{3.7.8}$$

where:

I_s = the collector saturation current (approx. 10^{-12} to 10^{-14} A);

V_A = Early voltage (usually between 100 and 150 V, see Fig. 3.4.11);

$V_T = k_B T/q$ (as defined at the beginning of Sec. 3.1).

Eq. 3.7.8 can be written simply from the geometric relations taken from Fig. 3.4.11. For a common silicon transistor the Early voltage is at least 100 V. Suppose both silicon transistors in Fig. 3.7.3a are identical and subject to the same temperature variations (on the same chip). The collector–emitter voltage of transistor Q_5 is the same as the base–emitter voltage, $V_{ce5} = V_{be5} \approx 0.65\,\text{V}$. In contrast, the collector–emitter voltage of transistor Q_6 is higher, say, $V_{ce6} = 15\,\text{V}$. The ratio of both collector currents is then:

$$\frac{I_{c6}}{I_{c5}} = \frac{1 + \dfrac{V_{ce6}}{V_A + V_{ce6}}}{1 + \dfrac{V_{ce5}}{V_A + V_{ce5}}} = \frac{1 + \dfrac{15}{100 + 15}}{1 + \dfrac{0.6}{100 + 0.6}} = 1.237 \tag{3.7.9}$$

By using more sophisticated circuits it is possible to make either $I_{c5} = I_{c6}$ or $I_R = I_{c6}$, or even to make any desired ratio between any of them [Ref. 3.31, 3.32, 3.33]. This is important in order to decrease the power dissipation in the resistor R. The power dissipated by Q_6 will then be the product of the desired current and the collector–emitter voltage set by the desired common mode range of the differential amplifier.

Since we are interested in wideband aspects of differential amplifiers, we shall not discuss further particularities here. However, current mirrors can also be used to convert the output signal of the differential amplifier into a single ended push pull drive, as is often done in modern operational amplifiers, and we shall return to this subject later in Part 5, with a discussion of the *Wilson* mirror [Ref. 3.32], Fig. 3.7.3c.

Before closing the analysis of current sinks, let us calculate the incremental collector resistance r_o of transistor Q_6, which can be derived simply from Fig. 3.4.11:

$$r_o = \frac{\Delta V_{ce6}}{\Delta I_{c6}} = \frac{V_A + V_{ce6}}{I_{c6}} \tag{3.7.10}$$

Returning to our differential amplifier example of Fig. 3.7.2 with $V_{ee} = -15\,\text{V}$, suppose we require a differential amplifier current $I_{ee} = I_{e1} + I_{e2} = I_{c6} = 0.03\,\text{A}$. By assuming the Early voltage $V_A = 135\,\text{V}$ and $V_{ce6} \approx V_{ee}$, the incremental collector resistance is:

$$r_o = \frac{135 + 15}{0.03} = 5\,\text{k}\Omega \tag{3.7.11}$$

If we were to replace the current generator by a simple resistor $R_{ee} = r_o$, the voltage V_{ee} in Fig. 3.7.2 would have to be:

$$(I_{e1} + I_{e3})\, R_{ee} = 0.03\,\text{A} \cdot 5000\,\Omega = 150\,\text{V} \tag{3.7.12}$$

which is 10 times more. Correspondingly, the power dissipation in the resistor R_{ee} would also be 10 times greater, or 4.5 W, compared to 0.45 W for Q_6.

A high incremental resistance r_o is also important for achieving high CMRR, because it gives the differential amplifier a higher immunity to power supply voltage variations (which is also a common mode signal).

A simple way of improving the current generator, thus achieving even greater CMRR factors, is shown in Fig. 3.7.4, where negative feedback, provided by the Q_5 gain, is used to stabilize the collector current of Q_6 and increase the incremental resistance, whilst a low voltage Zener diode (named after its inventor, the American physicist *Clarence M. Zener*, 1905-1993) reduces the V_{be} thermal drift of Q_5, owing to an almost equal, but opposite thermal coefficient.

In this circuit any increase in Q_6 collector current I_{c6} is sensed by its voltage drop on R_2, increasing V_{b5}, which in turn increases I_{e5}, thus reducing I_{b6} and therefore also I_{c6}. The reduction feedback factor is nearly equal to the Q_5 current gain β. Effectively, the output resistance is increased from r_o of Eq. 3.7.10 to about βr_o. Note that this circuit does not rely on identical transistor parameters, so it can be used in discrete circuits.

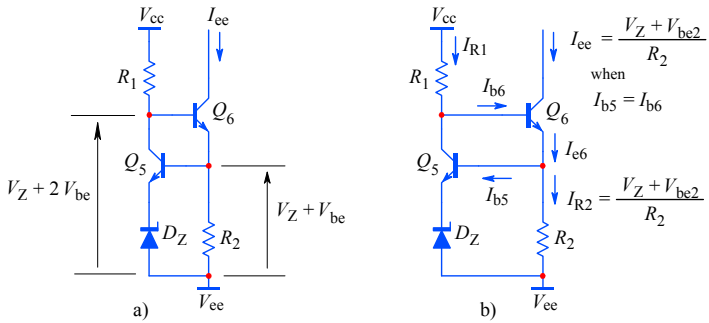


Fig. 3.7.4: Improved current generator: a) voltage drops; b) current analysis.

The circuits shown and only briefly discussed here should give the reader a starting point in current control design. Many more circuits, either simple or more elaborate, can be found in the references quoted.

3.8 The f_T Doubler

Let us return to the differential cascode amplifier of Fig. 3.7.2, where the basic analysis was done with the assumption that the inputs were driven by a differential voltage source. Here we shall analyze a current driven stage, which could suitably be used as an intermediate amplifier stage with the T-coil peaking at the input, such as the cascode of Fig. 3.6.1. The T-coil peaking has already been analyzed in detail, so we shall concentrate on the active part of the circuit and see how it can be improved.

In Fig. 3.8.1 we have the differential cascode circuit, driven differentially by the current source i_s. This current source is loaded by the resistance R_s which we have split in half and grounded its middle point. The output currents from the collectors of Q_3 and Q_4 drive the differential inputs of the next amplifying stage and we are not interested in it now. What we are interested in, is the current gain of this differential stage. The reader who has followed the previous analysis with some attention should by now be able to understand the following relation intuitively:

$$\frac{i_o}{i_s} \approx \frac{R_s}{R_e} \cdot \frac{1}{1 + s\left(R_s \frac{\tau_{T1}}{R_e} + \frac{R_s}{2} C_\mu\right)} \cdot \frac{1}{1 + s\,\tau_{T2}} \tag{3.8.1}$$

The last fraction is the frequency dependent part owed to $Q_{3,4}$.

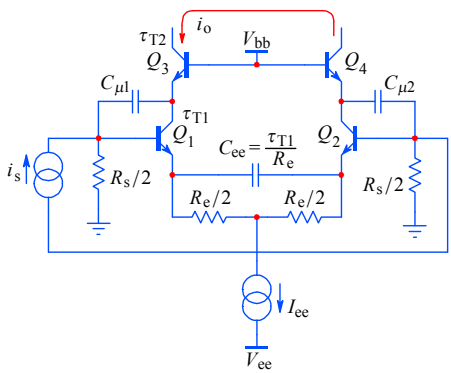


Fig. 3.8.1: The current driven differential cascode amplifier. We assume that the transistor pair $Q_{1,2}$, and the pair $Q_{3,4}$, respectively, have identical main parameters. The emitters of $Q_{1,2}$ see the R_eC_{ee} network, set to equal the time constant τ_{T1} of transistors $Q_{1,2}$.

The low frequency current gain is simply:

$$A_i = \frac{R_s}{R_e} \tag{3.8.2}$$

We can substitute this in Eq. 3.8.1 to highlight the bandwidth dependence on gain:

$$\frac{i_o}{i_s} \approx A_i \frac{1}{1 + s\left(A_i\,\tau_{T1} + A_i \frac{R_e\,C_\mu}{2}\right)} \cdot \frac{1}{1 + s\,\tau_{T2}} \tag{3.8.3}$$

This means that by reducing the gain A_i we can extend the bandwidth. On the basis of this there arises the idea that we can add another differential stage to double the current (and therefore also the current gain) and then optimize the stage, choosing between the doubling of gain with the same bandwidth and the doubling of bandwidth with the same gain, or any factor in between.

The basic f_T doubler circuit, developed by *C.R. Battjes*, [Ref. 3.1], is presented in Fig. 3.8.2. Each differential pair amplifies the voltage drop on its own $R_s/2$ and each pair sees its own R_e between the emitters, thus the current gain is simply:

$$\frac{i_o}{i_s} \approx \frac{R_s}{R_e} \cdot \frac{1}{1 + s\left(R_s \frac{\tau_{T1}}{2\,R_e} + \frac{R_s C_\mu}{2}\right)} \cdot \frac{1}{1 + s\,\tau_{T2}}$$

$$\approx A_i \frac{1}{1 + s\,A_i\left(\frac{\tau_{T1}}{2} + \frac{R_e\,C_\mu}{2}\right)} \cdot \frac{1}{1 + s\,\tau_{T2}} \qquad (3.8.4)$$

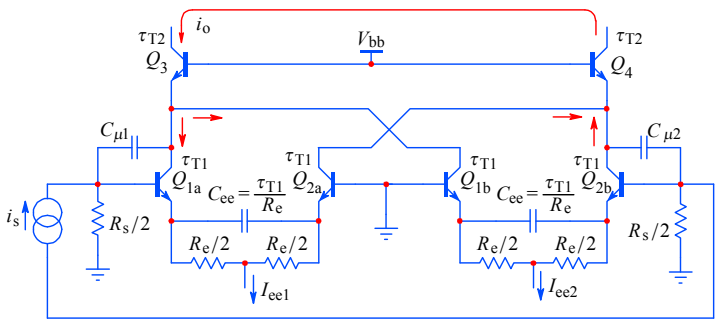


Fig. 3.8.2: The basic f_T doubler circuit. We assume equal transistors for $Q_{1a,2a}$ and $Q_{1b,2b}$. The low input impedance of $Q_{3,4}$ emitters allows summing the collector current of each pair, but cross-coupled for in phase signal summing.

Another advantage of this circuit is the reduced input capacitance:

$$C_i = \frac{2\,\tau_{T1}}{R_e} + C_\mu \qquad \text{for the circuit in Fig. 3.8.1} \qquad (3.8.5)$$

$$C_i = \frac{\tau_{T1}}{R_e} + C_\mu \qquad \text{for the circuit in Fig. 3.8.2} \qquad (3.8.6)$$

This will ease the application of T-coil peaking at the input.

But there are also limitations. As can be deduced from Eq. 3.8.4, the 'doubler' term in the circuit name is misleading, because the term with τ_{T2} is not influenced by the reduced gain. Therefore, the amount of bandwidth improvement depends on which time constant is larger. The part with $R_e C_\mu$ can be reduced by selecting transistors with low C_μ, but this we would do anyway. And we do not want to reduce R_e, because it would increase the gain (for same R_s).

Another problem is that, although the transfer function is of second order, there are two real poles, so we can not 'tune' the system for an efficient peaking. By forcing the system to have complex conjugate poles with emitter peaking, we would increase the emitter capacitance C_{ee}, which would be reflected into the base as an increased input capacitance; this would increase exactly that term which we have just halved.

A quick estimate will give us a little more feeling of the improvement achievable. Let us have a number of transistors, with $f_T = 3.5\,\text{GHz}$, $C_\mu = 1\,\text{pF}$, $R_s = 2 \times 50\,\Omega$, $Q_{3,4}$ collector load $R_L = 2 \times 50\,\Omega$, $C_L = 1\,\text{pF}$ and the total current gain $A_i = 3$. Assuming that the system's response is governed by a dominant pole, we can calculate the rise time of the conventional system as:

$$t_r = 2.2\sqrt{A_i^2\left(\frac{1}{2\pi f_T} + \frac{R_e C_\mu}{2}\right)^2 + \left(\frac{1}{2\pi f_T}\right)^2 + [R_L(C_L + C_\mu)]^2} \tag{3.8.7}$$

Then, for the ordinary differential cascode in Fig. 3.8.1:

$$t_{r1} = 476\,\text{ps} \tag{3.8.8}$$

whilst for the f_T doubler we have:

$$t_{r2} = 355\,\text{ps} \tag{3.8.9}$$

and the improvement factor is 1.34, much less than 2. Transistors with lower f_T might give an apparently greater improvement (about 1.7 could be expected) owing to the lower contribution of the source's impedance. However, it seems that a better idea would be to remain with the original bandwidth and use the gain doubling instead, which could lead to a system with a lower number of stages, which in turn could be optimized more easily.

On the other hand, the reduced input capacitance is really beneficial to the loading of the input T-coils. With the data from the example above, we can calculate the T-coils for the conventional and the doubler system and the resulting bandwidths. From Eq. 3.6.21 we can find that:

$$\omega_H = \sqrt{\frac{12}{(R_L C_i)^2}\left(1 + \frac{r_b}{R_L}\right)} \tag{3.8.10}$$

By assuming an $r_b = 15\,\Omega$ and C_i of 6.5 and 3.7 pF, respectively (Eq. 3.8.5 and 3.8.6), we can calculate an f_H of 1.9 and 3.4 GHz, a ratio of nearly 1.8, which is worth considering.

In principle one could use the same doubler implementation with 4, 6, or more transistor pairs; however, the input capacitance poses a practical limit. A system with 4 pairs is already slower than the system with two pairs.

3.9 JFET Source Follower

Wideband signals come usually from two source types: low impedance sources are usually those from the output of other wideband amplifiers, while medium and high impedance sources are usually those from sensors or other very low power sources.

In the first case, we employ standardized impedances, 50 Ω or 75 Ω, so that both the source and the load have the same impedance as the characteristic impedance of the cable that connects them. In this way we preserve the bandwidth and prevent reflections which would distort the signal, but we pay for this by the 50 % (−6 dB) attenuation of the signal's amplitude.

In the second case we also want a standardized value of the impedance, but this time the value is 1 MΩ, in parallel with some (inevitable) capacitance, usually 20 pF (but values from 10 to 25 pF can also be found). The standardized capacitance is helpful not only in determining the loading of the source at high frequencies, but also to allow the use of 'probes', which are actually special HF attenuators (÷10 or ÷100), so that the source can be loaded by a 10 MΩ or even a 100 MΩ resistance, while keeping the loading capacitance below some 12 pF.

With the improvement of semiconductor production processes, the so called 'active' probes have been developed, used mostly for extremely wideband signals, such as those found in modern communications and digital computers. Active probes usually have a 10 kΩ || 2 pF input impedances, with no reduction in amplitude.

The key component of both high input impedance amplifiers and active probes is the JFET (*junction field effect transistor*) source follower [Ref. 3.16].

The basic JFET source follower circuit configuration is shown in Fig. 3.9.1. In contrast to the BJT emitter follower (with an input resistance of about βR_e), the JFET source follower has a very high input resistance (between 10^9 and 10^{12} Ω), owed to the specific construction of the JFET. Its gate (a p–n junction with the drain–source channel) is reverse biased in normal operation, modulating the channel width by the electrical field only, so the input current is mainly owed to the reverse biased p–n junction leakage and the input capacitances, C_{gd} and C_{gs}.

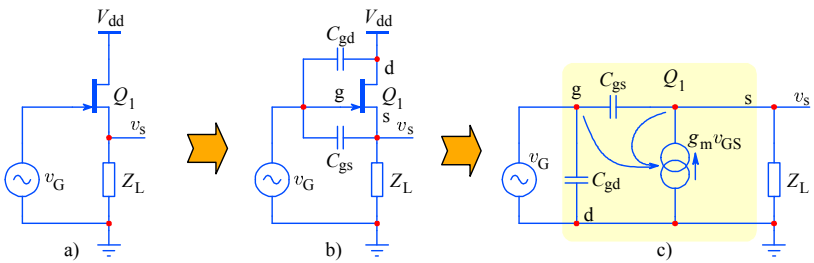


Fig. 3.9.1: The JFET source follower: a) circuit schematic; b) the same circuit, but with an ideal JFET and the inter-electrode capacitances drawn as external components; c) equivalent circuit.

A MOSFET (*metal oxide silicon field effect transistor*) has even greater input resistance (up to ~10^{15} Ω); however it also has a greater input capacitance (between 20

and 200 pF; it is also more noisy and more sensitive to damage by being overdriven), so it is not suitable for a wideband amplifier input stage.

In Fig. 3.9.1b we have drawn an ideal JFET device and its inter-electrode capacitances are modeled as external components. These capacitances determine the response at high frequencies [Ref. 3.8, 3.16, 3.20, 3.35]. Fig. 3.9.1c shows the equivalent circuit.

The source follower is actually the common drain circuit with a voltage gain of nearly unity, as the name 'follower' implies. The meaning of the circuit components is:

C_{gd} gate–drain capacitance; in most manufacturer data sheet it is labeled as C_{rss} (*common source circuit reverse capacitance*); values usually range between 1 and 5 pF;

C_{gs} gate–source capacitance; in their data sheet, manufacturers usually report the value of C_{iss}, the *common source total input capacitance*, therefore we obtain $C_{gs} \approx C_{iss} - C_{rss}$; values of C_{gs} usually range from 3 to 15 pF;

g_m JFET transconductance; usual values range between 1 000 and 15 000 µS (in some data-sheets, the symbol 'mho' is used to express that the unit *siemens* $[S] = [1/\Omega]$);

Z_L the loading impedance of the JFET source.

The JFET drain is connected to the power supply which must be a short circuit for the drain signal current, therefore we can connect C_{gd} to ground, in parallel with the signal source. We assume the signal source impedance to be zero; so we can forget about C_{gd} for a while.

From the equivalent circuit in Fig. 3.91c we find the currents for the node g:

$$i_g = v_G\, s\, C_{gd} + (v_G - v_s)\, s\, C_{gs} \tag{3.9.1}$$

and the currents for the node s:

$$(v_G - v_s)\, s\, C_{gs} + (v_G - v_s)\, g_m = \frac{v_s}{Z_L} \tag{3.9.2}$$

which can be rewritten as:

$$v_G\left(1 + \frac{g_m}{s\, C_{gs}}\right) = v_s\left(1 + \frac{g_m}{s\, C_{gs}} + \frac{1}{s\, C_{gs} Z_L}\right) \tag{3.9.3}$$

From this we obtain the system's voltage gain:

$$A_v = \frac{v_s}{v_G} = \frac{Z_L\left(1 + \dfrac{s\, C_{gs}}{g_m}\right)}{Z_L\left(1 + \dfrac{s\, C_{gs}}{g_m}\right) + \dfrac{1}{g_m}} \tag{3.9.4}$$

In practical follower circuits we want to make the output signal's dynamic range as high and the voltage gain as close to 1 as possible. The simplest way to achieve this is by replacing Z_L with a constant current generator, Fig. 3.9.2, much as we have done in

the differential amplifier. By doing this, we increase the real (resistive) part of the loading impedance, but we can do little to reduce the always present loading capacitance C_L.

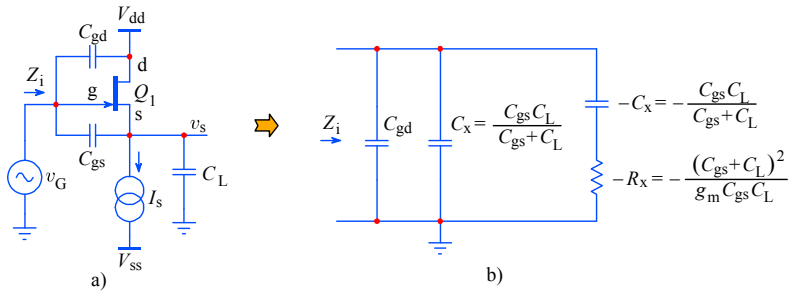


Fig. 3.9.2: The JFET source follower biased by a current generator and loaded only by the inevitable stray capacitance C_L: a) circuit schematic; b) the input impedance has two **negative** components, owed to C_{gs} and g_m (see Sec. 3.9.5).

In Eq. 3.9.4 the term C_{gs}/g_m is obviously the characteristic JFET time constant, τ_{FET}:

$$\frac{C_{gs}}{g_m} = \tau_{FET} = \frac{1}{\omega_{FET}} \tag{3.9.5}$$

Since we now have $Z_L = 1/j\omega C_L$ we can rewrite Eq. 3.9.4 as:

$$A_v = \frac{v_s}{v_G} = \frac{1 + \dfrac{j\omega}{\omega_{FET}}}{1 + j\omega\left(\dfrac{1}{\omega_{FET}} + \dfrac{C_L}{g_m}\right)} \tag{3.9.6.}$$

and by replacing g_m with $\omega_{FET} C_{gs}$ (Eq. 3.9.5) we obtain:

$$\frac{v_s}{v_G} = \frac{1 + \dfrac{j\omega}{\omega_{FET}}}{1 + \dfrac{j\omega}{\omega_{FET}} \cdot \dfrac{1}{D_c}} \tag{3.9.7.}$$

Here D_c is the input to output capacitive divider, which would set the output voltage if only the capacitances were in place:

$$D_c = \frac{C_{gs}}{C_{gs} + C_L} \tag{3.9.8.}$$

We would like to express Eq. 3.9.7 by its pole s_1 and zero s_2, so we need the normalized canonical form:

$$F(s) = A_0 \frac{-s_1}{s - s_1} \cdot \frac{s - s_2}{-s_2} \tag{3.9.9}$$

Therefore we replace $j\omega$ with s, multiply both the numerator and the denominator by $\omega_{\text{FET}}\, D_c$ and obtain:

$$F(s) = D_c \frac{s - (-\omega_{\text{FET}})}{s - (-\omega_{\text{FET}}\, D_c)} \tag{3.9.10}$$

At zero frequency, the transfer function gain is:

$$A_0 = F(0) = D_c \frac{\omega_{\text{FET}}}{\omega_{\text{FET}}\, D_c} = 1 \tag{3.9.11}$$

so the zero is:

$$s_2 = -\omega_{\text{FET}} \tag{3.9.12}$$

and the pole is:

$$s_1 = -\omega_{\text{FET}}\, D_c \tag{3.9.13}$$

These simple relations are the basis from which we shall calculate the frequency response magnitude, the phase, the group delay and the step response of the JFET source follower (simplified at first and including the neglected components later).

3.9.1 Frequency response magnitude

The frequency response magnitude is the normalized absolute value of $F(s)$ and we want to have the normalization in both gain and frequency. Eq. 3.9.7 is already normalized in frequency (to ω_{FET}) and the gain $A_0 = 1$. To get the magnitude, we must multiply $F(j\omega)$ by its complex conjugate, $F(-j\omega)$ and take the square root:

$$|F(\omega)| = \sqrt{F(j\omega)\, F(-j\omega)} = \sqrt{\frac{1 + \dfrac{j\omega}{\omega_{\text{FET}}}}{1 + \dfrac{j\omega}{\omega_{\text{FET}}} \cdot \dfrac{1}{D_c}} \cdot \frac{1 - \dfrac{j\omega}{\omega_{\text{FET}}}}{1 - \dfrac{j\omega}{\omega_{\text{FET}}} \cdot \dfrac{1}{D_c}}}$$

$$|F(\omega)| = \sqrt{\frac{1 + \left(\dfrac{\omega}{\omega_{\text{FET}}}\right)^2}{1 + \left(\dfrac{\omega}{\omega_{\text{FET}}\, D_c}\right)^2}} \tag{3.9.14}$$

Since we want to examine the influence of loading we shall plot the transfer function for three different values of the ratio C_L/C_{gs}: 0.5, 1.0, and 2.0 (the corresponding values of D_c being 0.67, 0.5, and 0.33, respectively). The plots, shown in Fig. 3.9.3, have three distinct frequency regions: in the lower one, the circuit behaves as a voltage follower, with the JFET playing an active role, so that $v_s = v_G$, whilst in the highest frequency region, only the capacitances are important; in between we have a transition between both operating modes.

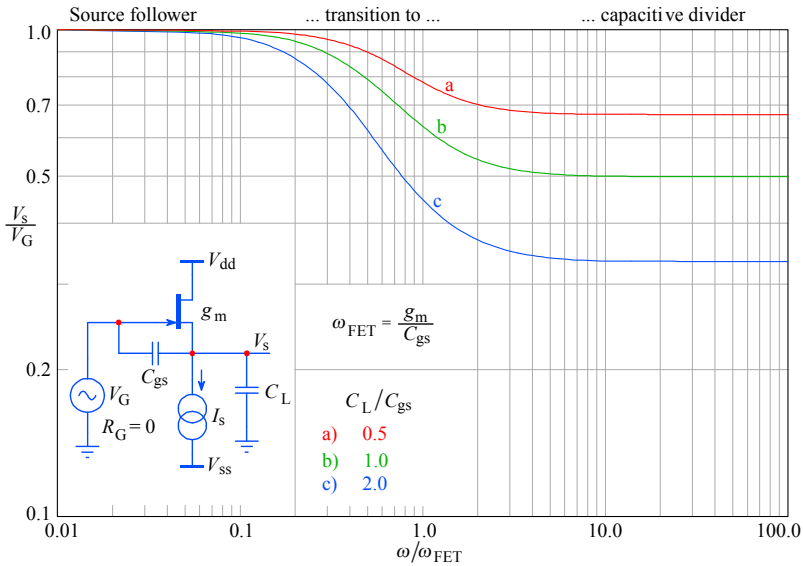


Fig. 3.9.3: Magnitude of the frequency response of the JFET source follower for three different capacitance ratios C_L/C_{gs}. The pole $s_3 = -1/R_G C_i$ has not been taken into account here (see Fig. 3.9.7 and Fig. 3.9.8).

The relation for the upper cutoff frequency is very interesting. If we set $|F(\omega)|$ to be equal to:

$$\sqrt{\frac{1+(\omega_h/\omega_{FET})^2}{1+(\omega_h/\omega_{FET}\,D_c)^2}} = \frac{1}{\sqrt{2}} \tag{3.9.15}$$

it follows that:

$$\omega_h = \omega_{FET}\,\frac{D_c}{\sqrt{1-2\,D_c^2}} \tag{3.9.16}$$

From Eq. 3.9.15 we can conclude that by putting $D_c = 1/\sqrt{2}$ the denominator is reduced to zero, thus $\omega_h = \infty$. This means that for such a capacitive ratio the magnitude never falls below $1/\sqrt{2}$. However attractive the possibility of achieving infinite bandwidth may seem, this can never be realized in practice, because any signal source will have some, although small, internal resistance R_G, resulting in an additional input pole $s_3 = -1/R_G C_i$, where C_i is the total input capacitance of the JFET. The complete transfer function will now be (see Fig. 3.9.7 and 3.9.8):

$$F(s) = \frac{s_1 s_3}{-s_2} \cdot \frac{s-s_2}{(s-s_1)(s-s_3)} \tag{3.9.17}$$

3.9.2 Phase

We obtain the phase response from Eq. 3.9.7 by taking the arctangent of the ratio of the imaginary to the real part of $F(j\omega)$:

$$\varphi(\omega) = \arctan \frac{\Im\{F(j\omega)\}}{\Re\{F(j\omega)\}} \tag{3.9.18}$$

Since in Eq. 3.9.7 we have a single real pole an a single real zero, the resulting phase angle is calculated as:

$$\varphi(\omega) = \arctan \frac{\omega}{\omega_{\mathrm{FET}}} - \arctan \frac{\omega}{\omega_{\mathrm{FET}}\, \mathrm{D_c}} \tag{3.9.19}$$

In Fig. 3.9.4 the three phase plots with same $C_{\mathrm{L}}/C_{\mathrm{gs}}$ ratios are shown. Because of the zero, the phase returns to the initial value at high frequencies.

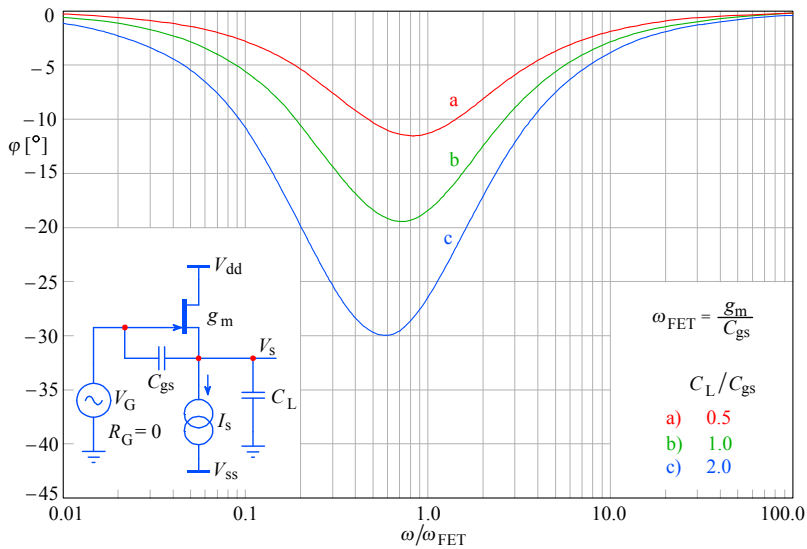


Fig. 3.9.4: Phase plots of the JFET source follower for the same three capacitance ratios.

3.9.3 Envelope delay

We obtain the envelope delay by taking the ω derivative of the phase:

$$\tau_{\mathrm{e}} = \frac{d\varphi}{d\omega} \tag{3.9.20}$$

but we usually prefer the normalized expression $\tau_{\mathrm{e}}\, \omega_{\mathrm{h}}$. In our case, however, the upper cut off frequency, ω_{h}, is changing with the capacitance divider D_{c}. So instead of ω_{h} we

shall, rather, normalize the envelope delay to the characteristic frequency of the JFET itself, ω_{FET}:

$$\tau_e\,\omega_{FET} = \frac{1}{1+(\omega/\omega_{FET})^2} - \frac{D_c}{D_c^2+(\omega/\omega_{FET})^2} \tag{3.5.21}$$

The envelope delay plots for the three capacitance ratios are shown in Fig. 3.9.5. Note that for all three ratios there is a frequency region in which the envelope delay becomes **positive**, implying that the output signal advance, in correlation with the phase plots, goes up with frequency. We have explained the physical background of this behavior in Part 2, Fig. 2.2.5, 2.2.6. The positive envelope delay influences the input impedance in a very unfavorable way, as we shall soon see.

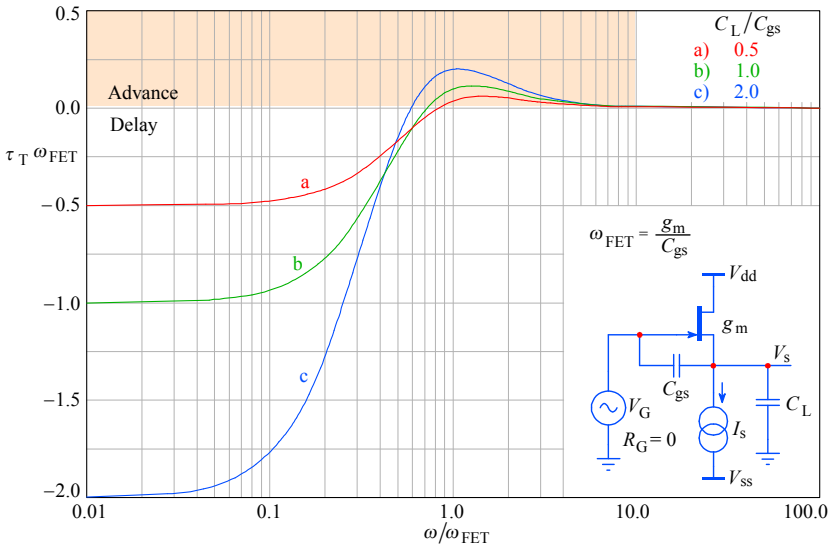


Fig. 3.9.5: The JFET envelope delay for the three capacitance ratios. Note the positive peak (phase advance region): trouble in sight!

3.9.4 Step response

We are going to use Eq. 3.9.9, which we multiply by the unit step operator $1/s$ to obtain the step response in the complex frequency domain; we then obtain the time domain response by applying the inverse Laplace transform:

$$G(s) = \frac{1}{s}\,F(s) = \frac{1}{s}\,D_c\,\frac{s-s_2}{s-s_1} \tag{3.9.22}$$

$$g(t) = \mathcal{L}^{-1}\{G(s)\} = D_c \sum_{s=0}^{s_1} \text{res}\,\frac{(s-s_2)\,e^{st}}{s\,(s-s_1)} \tag{3.9.23}$$

The residue at $s \to 0$ is:

$$\text{res}_0 = \lim_{s \to 0} s \, \frac{(s - s_2)\, \mathrm{e}^{st}}{s\,(s - s_1)} = \frac{s_2}{s_1} \tag{3.9.24}$$

and the residue at $s \to s_1$ is:

$$\text{res}_1 = \lim_{s \to s_1} (s - s_1) \, \frac{(s - s_2)\, \mathrm{e}^{st}}{s\,(s - s_1)} = \frac{s_1 - s_2}{s_1} \, \mathrm{e}^{s_1 t} \tag{3.9.25}$$

By entering both residues back into Eq. 3.9.23 we get:

$$g(t) = D_\mathrm{c} \left(\frac{s_2}{s_1} + \frac{s_1 - s_2}{s_1} \, \mathrm{e}^{s_1 t} \right) \tag{3.9.26}$$

and, by considering Eq. 3.9.12 and 3.9.13, as well as that $\omega_\mathrm{FET} = 1/\tau_\mathrm{FET}$ and using the normalized time t/τ_FET, we end up with:

$$g(t) = 1 - (1 - D_\mathrm{c})\, \mathrm{e}^{-D_\mathrm{c} t/\tau_\mathrm{FET}} \tag{3.9.27}$$

The plot of this relation is shown in Fig. 3.9.6, again for the same three capacitance ratios. The initial output signal jump at $t = 0$ is the input signal cross-talk (through C_gs) multiplied by the D_c factor:

$$g(0) = D_\mathrm{c} \tag{3.9.28}$$

Following the jump is the exponential relaxation towards the normal follower action at lower frequencies. If the input pole $s_3 = -1/R_\mathrm{G}C_\mathrm{i}$ is taken into account, the jump would be slowed down to an exponential rise with a time constant of $R_\mathrm{G}C_\mathrm{i}$.

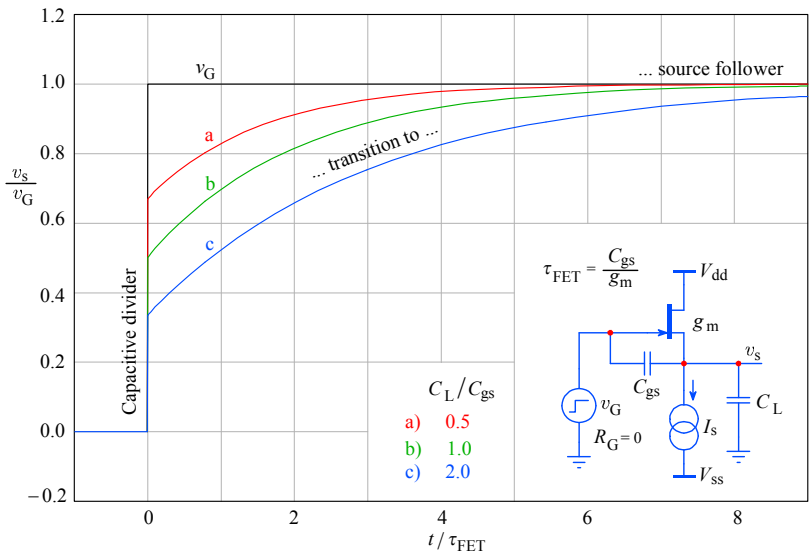


Fig. 3.9.6: The JFET source follower step response for the three capacitance ratios.

As mentioned earlier in connection with Eq. 3.9.17, when the value of R_G is comparable to $1/g_m$, an additional pole has to be to considered. We must derive the system transfer function again, from the following two equations:

For the currents into the node g:

$$\frac{v_G - v_g}{R_G} = \frac{v_g}{\frac{1}{s\,C_{gd}}} + \frac{v_g - v_s}{\frac{1}{s\,C_{gs}}} \tag{3.9.29}$$

and for the currents into the node s:

$$\frac{v_g - v_s}{\frac{1}{s\,C_{gs}}} + (v_g - v_s)\,g_m = \frac{v_s}{\frac{1}{s\,C_L}} \tag{3.9.30}$$

We first express v_g as a function of v_s from Eq. 3.9.30:

$$v_g = v_s\left(1 + \frac{s\,C_L}{s\,C_{gs} + g_m}\right) \tag{3.9.31}$$

Then we replace v_g in Eq. 3.9.29 by 3.9.31:

$$v_G = v_s\left(1 + \frac{s\,C_L}{s\,C_{gs} + g_m}\right)(1 + R_G\,s\,C_{gd} + R_G\,s\,C_{gs}) - v_s R_G\,s\,C_{gs} \tag{3.9.32}$$

After some further manipulation we arrive at:

$$\frac{v_s}{v_G} = \frac{1}{1 + R_G\,s\,C_{gd} + \frac{s\,C_L}{s\,C_{gs} + g_m}(1 + R_G\,s\,(C_{gd} + C_{gs}))} \tag{3.9.33}$$

Now we put this into the normalized canonical form and use Eq. 3.9.5 again to replace the term g_m/C_{gs} with ω_{FET}. Also, we express all the time constants as functions of ω_{FET} and the appropriate capacitance ratios. Finally, we want to see how the response depends on the product $g_m R_G$, so we multiply all the terms containing R_G with g_m and compensate each of them accordingly. The final expression is:

$$\frac{v_s}{v_G} = \frac{(s + \omega_{FET})\frac{1}{g_m R_G} \cdot \frac{\omega_{FET}\frac{C_{gs}}{C_{gd}}}{\left(1 + \frac{C_L}{C_{gs}} + \frac{C_L}{C_{gd}}\right)}}{s^2 + s\,\frac{\omega_{FET}\left[1 + \frac{1}{g_m R_G}\left(\frac{C_{gs}}{C_{gd}} + \frac{C_L}{C_{gd}}\right)\right]}{\left(1 + \frac{C_L}{C_{gs}} + \frac{C_L}{C_{gd}}\right)} + \frac{1}{g_m R_G} \cdot \frac{\omega_{FET}^2\frac{C_{gs}}{C_{gd}}}{\left(1 + \frac{C_L}{C_{gs}} + \frac{C_L}{C_{gd}}\right)}} \tag{3.9.34}$$

To plot the responses we shall set:

$$\frac{C_L}{C_{gs}} = 1, \qquad \frac{C_{gs}}{C_{gd}} = 5, \quad \omega_{FET} = 1, \quad s = j\omega, \quad \text{and} \quad g_m R_G = [0.3,\ 1,\ 3]$$

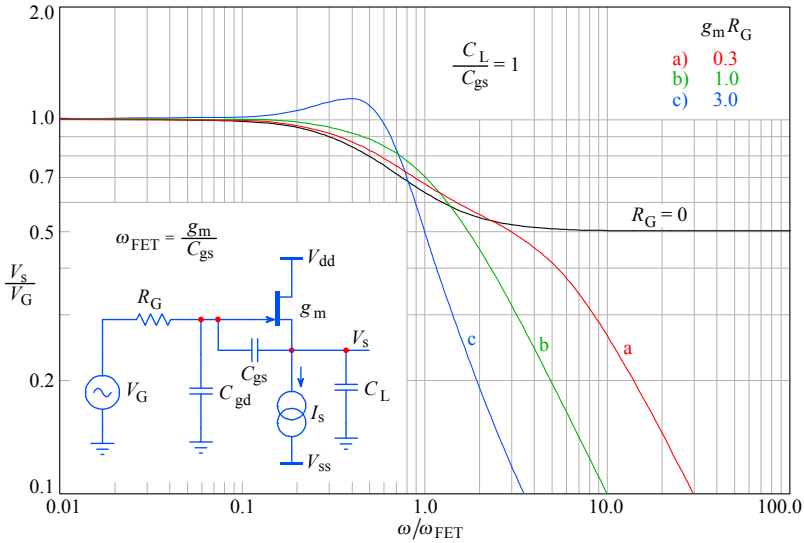


Fig. 3.9.7: The JFET source follower frequency response for a ratio $C_L/C_{gs} = 1$ and a variable signal source impedance, so that $R_G g_m$ is 0.3, 1, and 3, respectively. Note the response peaking for $g_m R_G = 3$.

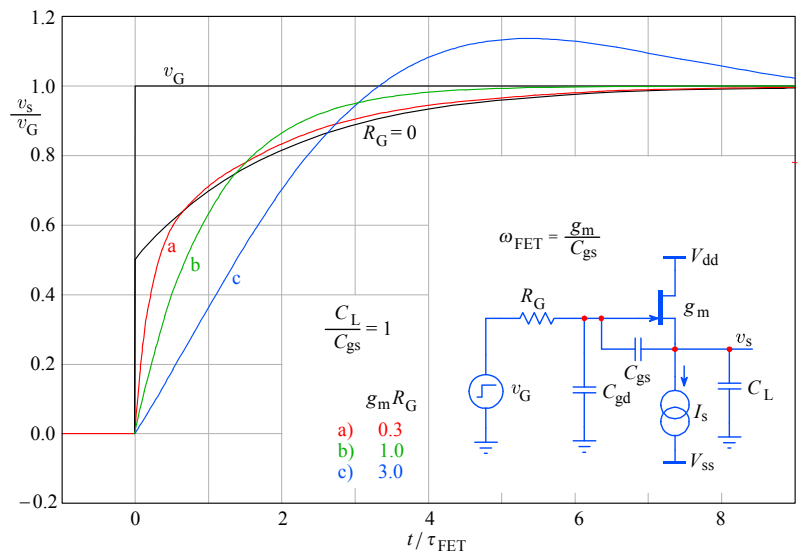


Fig. 3.9.8: The JFET source follower step response for the same conditions as in Fig. 3.9.7.

3.9.5 Input impedance

In Fig. 3.9.7 and 3.9.8 we have seen how the JFET source follower response is affected by its input impedance; this behavior becomes evident when the signal source has a non-zero resistance. Here, we are going to explore the circuit in more depth to examine the influence of a complex and, in particular, inductive signal source.

As we have done in the previous analysis, the gate–drain capacitance C_{gd} will appear in parallel with the input, so we can treat its admittance separately and concentrate on the remaining input components.

We start from Eq. 3.9.1 by solving it for v_s:

$$v_s = v_G - \frac{i_i}{s\,C_{gs}} \tag{3.9.35}$$

This we insert into Eq. 3.9.2:

$$-v_G\,(s\,C_{gs} + g_m) + \left(v_g - \frac{i_i}{s\,C_{gs}}\right)\left(s\,C_{gs} + g_m + \frac{1}{Z_L}\right) = 0 \tag{3.9.36}$$

Because the JFET source is biased from a constant current generator (whose impedance we assume to be infinite) the loading admittance is $1/Z_L = s\,C_L$. Let us put this back into Eq. 3.3.36 and rearrange it a little:

$$v_G\,s\,C_L = i_i\left(1 + \frac{g_m}{s\,C_{gs}} + \frac{C_L}{C_{gs}}\right) \tag{3.9.37}$$

Furthermore:

$$v_G = i_i\left(\frac{1}{s\,C_L} + \frac{g_m}{s^2\,C_{gs}\,C_L} + \frac{1}{s\,C_{gs}}\right)$$

$$= i_i\,\frac{s\,(C_{gs} + C_L) + g_m}{s^2\,C_{gs}\,C_L} \tag{3.9.38}$$

The input impedance (without C_{gd}, hence the prime [′]) is then:

$$Z_i' = \frac{v_g}{i_i} = \frac{s\,(C_{gs} + C_L) + g_m}{s^2\,C_{gs}\,C_L} \tag{3.9.39}$$

To see more clearly how this impedance is comprised, we invert it to find the admittance and apply the continuous fraction synthesis in order to identify the individual components.

$$Y_i' = \frac{s^2\,C_{gs}\,C_L}{s\,(C_{gs} + C_L) + g_m} = s\,\frac{C_{gs}\,C_L}{C_{gs} + C_L} - \frac{g_m\,s\,\dfrac{C_{gs}\,C_L}{C_{gs} + C_L}}{s\,(C_{gs} + C_L) + g_m} \tag{3.9.40}$$

The first fraction is the admittance of the capacitances C_{gs} and C_L connected in series. Let us name this combination C_x:

$$C_x = \frac{C_{gs}\,C_L}{C_{gs} + C_L} \tag{3.9.41}$$

The second fraction, which has a negative sign, must be further simplified. We invert it again, and after some simple rearrangement we obtain the impedance:

$$Z_x = -\frac{(C_{gs}+C_L)^2}{g_m C_{gs} C_L} - \frac{C_{gs}+C_L}{s C_{gs} C_L} \tag{3.9.42}$$

The first part is interpreted as a negative resistance, which we shall label $-R_x$ in order to follow the negative sign in the following analysis more clearly:

$$-R_x = -\frac{(C_{gs}+C_L)^2}{g_m C_{gs} C_L} \tag{3.9.43}$$

The second part as a negative capacitance, which we label $-C_x$ because it has the same absolute value as C_x from the Eq. 3.9.41:

$$-C_x = -\frac{C_{gs} C_L}{C_{gs}+C_L} \tag{3.9.44}$$

Now that we have all the components we can reintroduce the gate–drain capacitance C_{gd}, so that the final equivalent input impedance looks like Fig. 3.9.9. We can write the complete input admittance:

$$Y_i = j\omega (C_{gd}+C_x) - \frac{1}{R_x + \dfrac{1}{j\omega C_x}} \tag{3.9.45}$$

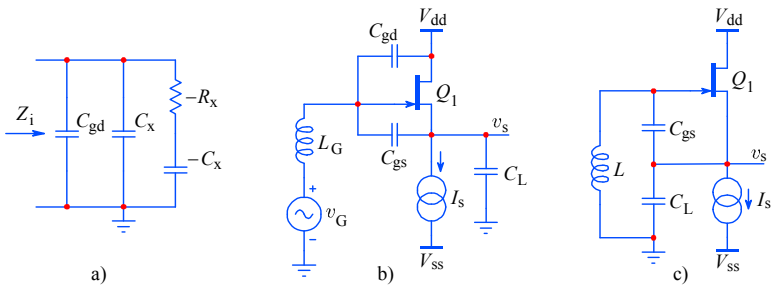


Fig. 3.9.9: a) The equivalent input impedance of the capacitively loaded JFET source follower has negative components which can be a nuisance if, as in b), the signal source has an inductive impedance, forming c) a familiar Colpitts oscillator. If C_{gd} is small, the circuit will oscillate for a broad range of inductance values.

We can separate the real and imaginary part of Y_i by putting Eq. 3.9.45 on a common denominator:

$$\begin{aligned} Y_i &= \Re\{Y_i\} + \Im\{Y_i\} \\ &= -\frac{\omega^2 C_x^2 R_x}{1+\omega^2 C_x^2 R_x^2} + j\omega \frac{C_{gd} + \omega^2 C_x^2 R_x^2 (C_{gd}+C_x)}{1+\omega^2 C_x^2 R_x^2} \end{aligned} \tag{3.9.46}$$

The negative real part can cause some serious trouble [Ref. 3.24]. Suppose we are troubleshooting a circuit with a switching power supply and we suspect it to be a cause of a strong electromagnetic interference (EMI); we want to use a coil with an appropriate inductance L (which, of course, has its own real and imaginary admittance) to inspect the various parts of the circuit for EMI intensity and field direction. If we connect this coil to the source follower and if the coil resistance is low, we would have:

$$\Re\{Y_L\} + \Re\{Y_\mathrm{i}\} \leq 0 \tag{3.9.47}$$

and the source follower becomes a familiar Colpitts oscillator, Fig. 3.9.9c [Ref. 3.25]. Indeed, some older oscilloscopes would burst into oscillation if connected to such a coil and with its input attenuator switched to maximum sensitivity (a few highly priced instruments built by respectable firms, back in early 1970's, were no exception).

By taking into account Eq. 3.9.42, 3.9.43 and 3.9.9 and substituting $\omega_\mathrm{FET} = g_\mathrm{m}/C_\mathrm{gs}$, the real part of the input impedance can be rewritten as:

$$\Re(Y_\mathrm{i}) = G_\mathrm{i} = -g_\mathrm{m}\,\frac{C_\mathrm{L}}{C_\mathrm{gs}} \cdot \frac{(\omega/\omega_\mathrm{FET})^2}{1 + (\omega/\omega_\mathrm{FET}\,D_\mathrm{c})^2} \tag{3.9.48}$$

The last fraction represents the normalized frequency dependence of this admittance:

$$G_\mathrm{iN} = \frac{(\omega/\omega_\mathrm{FET})^2}{1 + (\omega/\omega_\mathrm{FET}\,D_\mathrm{c})^2} \tag{3.9.49}$$

Fig. 3.9.10 shows the plots of G_iN for the same ratios of $C_\mathrm{L}/C_\mathrm{gs}$ as before. Note the quadratic dependence (of D_c) at high frequencies.

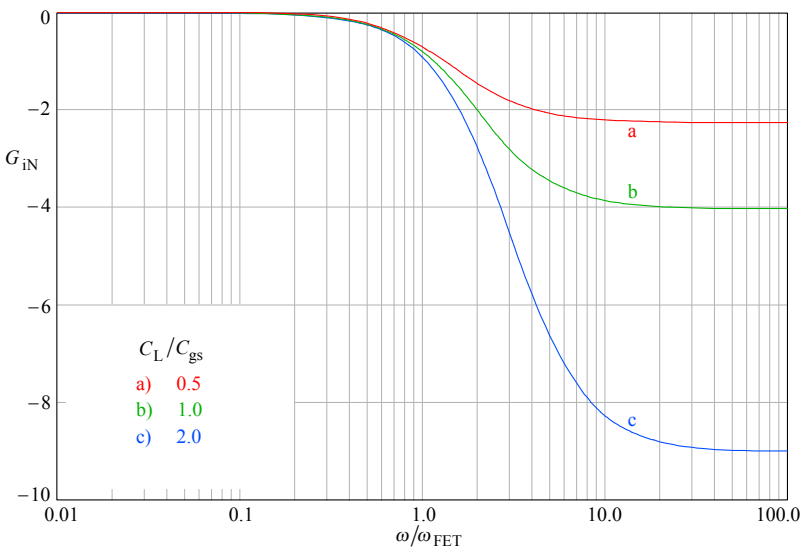


Fig. 3.9.10: Normalized negative input conductance G_iN vs. frequency.

A negative input impedance is always highly undesirable and we shall show a few possible solutions. The obvious way would be to introduce a resistor in series with

the JFET gate; since we should have some series resistance anyway in order to protect the sensitive input from static discharge or accidental overdrive, this will seem to be the preferred choice. However, after a closer look, this protection resistance is too small to prevent oscillations in case of an inductive signal source impedance. The required resistance value which will guarantee stability in all conditions will be so high that the bandwidth will be reduced by nearly an order of magnitude. Thus this method of compensation is used only if we do not care how much bandwidth we obtain.

A more elegant method of compensation is the one which we have used already in Fig. 3.5.3. If we introduce a serially connected $R_x C_x$ network in parallel with the JFET input, as shown in Fig. 3.9.11, we obtain $Y_x = 0$ and $Z_x = \infty$. Note the corresponding phasor diagram: we first draw the negative components, $-R_x$ and $-C_x$, find the impedance vector $-Z_x$ and invert it to find the negative admittance, $-Y_x$. We then compensate it by a positive admittance Y_x such that their sum $Y_{ic} = 0$. We finally invert Y_x to find Z_x and decompose it into its real and imaginary part, R_x and C_x:

$$Y_{ic} = \frac{1}{-R_x - \dfrac{1}{j\omega C_x}} + \frac{1}{R_x + \dfrac{1}{j\omega C_x}} = 0 \qquad (3.9.50)$$

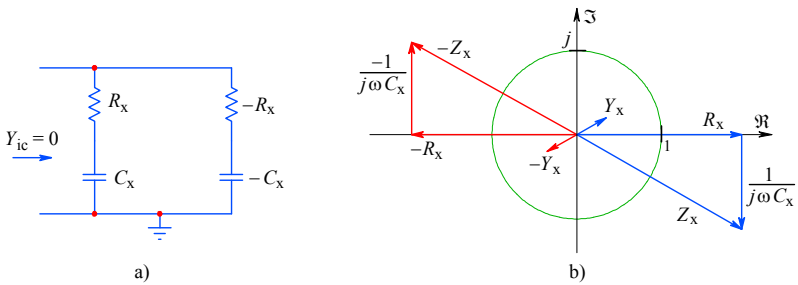


Fig. 3.9.11: a) The negative components of the input impedance can be compensated by an equal but positive network, connected in parallel, so that their admittances sum to zero (infinite impedance). In b) we see the corresponding phasor diagram.

With this compensation, the total input impedance is the one belonging to the parallel connection of C_{gd} with C_x and assuming a 1 MΩ gate bias resistor R_{in}:

$$Z_i = \frac{1}{\dfrac{1}{R_{in}} + j\omega(C_{gd} + C_x)} = \frac{R_{in}}{1 + j\omega\left(C_{gd} + \dfrac{C_{gs}\,C_L}{C_{gs} + C_L}\right)R_{in}} \qquad (3.9.51)$$

The analysis of the input impedance would be incomplete without Fig. 3.9.12, where the Nyquist diagrams of the impedance are shown revealing its frequency dependence, as well as the influence of different signal source impedances.

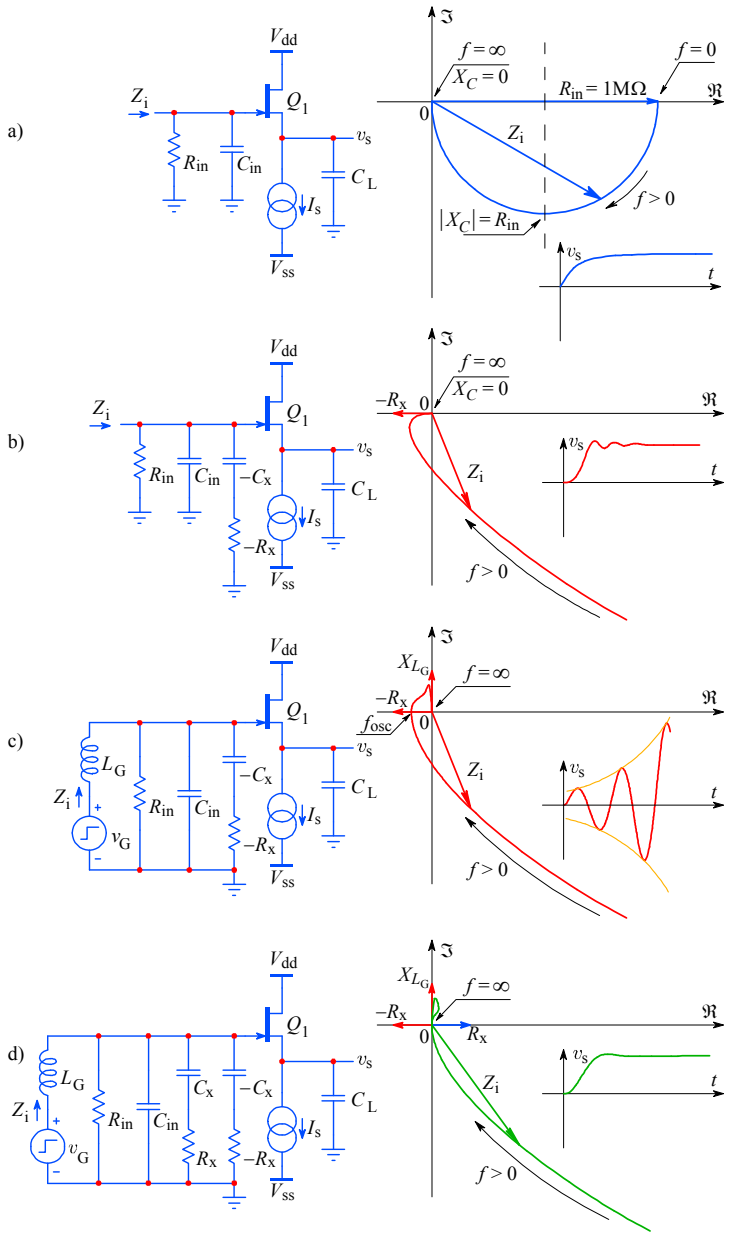


Fig. 3.9.12: **a)** The input impedance of the JFET source follower, assumed to be purely capacitive and in parallel with a 1 MΩ gate biasing resistor; thus at $f = 0$ we see only the resistor and at $f = \infty$ the reactance of the input capacitance is zero; **b)** the negative input impedance components affect the input impedance near the origin; **c)** with an inductive signal source, the point in which the impedance crosses the negative real axis corresponds to the system resonant frequency, provoking oscillations. **d)** The compensation removes the negative components.

In Fig. 3.9.12a the JFET gate is tied to ground by a 1 MΩ resistor, which, with a purely capacitive input impedance, would give a phasor diagram in the form of a half circle with frequency varying from DC to infinity.

In Fig. 3.9.12b we concentrate on the small area near the complex plane origin (high frequencies, close to f_{FET}), where we draw the influence of the negative input impedance components, assuming a resistive signal source.

In Fig. 3.9.12c an inductive signal source (with a small resistive component) will cause the impedance crossing the real axis in the negative region, therefore the circuit would oscillate at the frequency at which this crossing occurs.

Finally, in Fig. 3.9.12d we see the same situation but with the negative components compensated as in Fig. 3.9.11. Note the small loop in the first quadrant of the impedance plot — it is caused by the small resistance R_G of the coil L_G, the coil inductance, and the total input capacitance C_{in}.

In Fig. 3.9.13 we see yet another way of compensating the negative input impedance. Here the compensation is achieved by inserting a small resistance R_d in the drain, thus allowing the anti-phase signal at the drain to influence the gate via C_{gd} and cancel the in-phase signal from the JFET source via C_{gs}. This method is sometimes preferred over the former method, because the PCB pads, which are needed to accommodate the additional compensation components, also create some additional parasitic capacitance from the gate to ground.

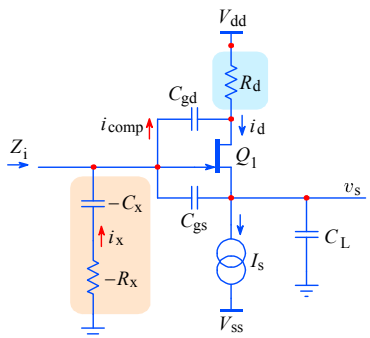


Fig. 3.9.13: Alternative compensation of the input impedance negative components, using negative feedback from the JFET drain.

It should be noted, however, that the negative input impedance compensation can be achieved for small signals only. Large signals vary the JFET gate's reverse bias voltage and the drain–source voltage considerably, therefore both C_{gs} and C_{gd}, as well as g_m, change with voltage. We therefore expect some nonlinear effects to appear with a large signal drive. We shall examine this and some other aspects of the source follower's performance in Part 5.

Résumé of Part 3

In this part we have analyzed some basic circuits for wideband amplification, examined their most important limitations, and explained several ways of improving their high frequency performance. The property which can cause most trouble, even for experienced designers, is the negative input impedance of some of the most useful wideband circuits, and we have shown a few possible solutions.

The reader must realize, however, that the analytical tools and solutions presented are by no means the ultimate design examples. For a final design, many other aspects of circuit performance must also be carefully considered, and, more often than not, these other factors will compromise the wideband performance severely.

As we have indicated at some points, there are ways of compensating certain unwanted circuit behavior by implementing the system in a differential configuration, but, on the negative side, this doubles the number of active components, increasing cost, power dissipation, circuit size, strays and parasitics and also the production and testing complexity. From the wideband design point of view, having many active components usually means many more poles and zeros that must be carefully analyzed and appropriately 'tuned'.

In Part 4 and Part 5 we shall explain some theoretical and practical techniques for an efficient design approach at the system level.

References:

[3.1] *C.R. Battjes*, Amplifier Frequency Response and Risetime, AFTR Class Notes (Amplifier Frequency and Transient Response), Tektronix, Inc., Beaverton, 1977

[3.2] *P. Starič*, Transistor as an Impedance Converter (in Slovenian with an English Abstract), Elektrotehniški Vestnik, 1989, pp. 17–22.

[3.3] *D.L. Feucht*, Handbook of Analog Circuit Design, Academic Press, Inc., San Diego, 1990

[3.4] *I.E. Getreu*, Modeling the Bipolar Transistor, Elsevier Scientific Publishing Company, Amsterdam, 1978

[3.5] *R.I. Ross*, T-coil Transistor Interstage Coupling, AFTR Class Notes, Tektronix, Inc., Beaverton, 1968

[3.6] *P. Starič*, Application of T-coil Transistor Interstage Coupling in Wideband Pulse Amplifiers, Elektrotehniški Vestnik, 1990, pp. 143–152.

[3.7] *P.R. Gray & R.G. Meyer*, Analysis and Design of Analog Integrated Circuits, John Wiley, New York, 1969

[3.8] *B.E. Hofer*, Amplifier Risetime and Frequency Response, AFTR Class Notes, Tektronix, Inc., Beaverton, Oregon, 1982

[3.9] *J. J. Ebers & J.L. Moll*, Large-Signal Behavior of Junction Transistors, Proceedings of IRE, Vol. 42, pp. 1761–1772, December 1954

[3.10] *H.K. Gummel & H.C. Poon*, An Integral Charge Control Model of Bipolar Transistors, Bell Systems Technical Journal, Vol. 49, pp. 827–852, May 1970

[3.11] *J.M. Early*, Effects of Space-Charge Layer Widening in Junction Transistors, Proceedings of IRE, Vol. 40, pp. 1401–1406, November 1952

[3.12] *P.E. Gray & C.L Searle*, Electronic Principles, Physics, Models and Circuits, John Wiley, New York, 1969

[3.13] *L.J. Giacoletto*, Electronics Designer's Handbook, Second Edition, McGraw-Hill, New York 1961.

[3.14] *P.M. Chirlian*, Electronic Circuits, Physical Principles and Design, McGraw-Hill, 1971.

[3.15] *J.E. Cathey*, Theory and Problems of Electronic Devices and Circuits, Schaum's Outline Series in Engineering, McGraw-Hill, New York, 1989

[3.16] *A.D. Evans*, Designing with Field-Effect Transistors, (Siliconix Inc.), McGraw-Hill, New York, 1981

[3.17] *J.M. Pettit & M. M. McWhorter*, Electronic Amplifier Circuits, Theory and Design, McGraw-Hill, New York, 1961

[3.18] *B. Orwiller*, Vertical Amplifier Circuits, Tektronix Inc., Beaverton, Oregon, 1969

[3.19] *C.R. Battjes*, Technical Notes on Bridged T-coil Peaking, Internal Publication, Tektronix, Inc., Beaverton, Oregon, 1969

[3.20] *P. Antognetti & G. Massobrio*, Semiconductor Device Modeling with SPICE, McGraw-Hill, New York, 1988

[3.21] *P. Horowitz & W. Hill*, The Art of Electronics, Cambridge University Press, Cambridge, 1987

[3.22] *G. Bruun*, Common-Emitter Transistor Video-Amplifiers, Proceedings of the IRE, Nov. 1956, pp. 1561–1572

[3.23] *F.W. Grover*, Inductance Calculation, (Reprint),
Instrument Society of America, Research Triangle Park, N. C. 27 709, 1973.

[3.24] *G.B. DeBella*, Stability of Capacitively-Loaded Emitter Follower,
Hewlett-Packard Journal, Vol. 17, No. 8, April 1966

[3.25] *L. Strauss*, Wave Generation and Shaping,
McGraw-Hill, New York, 1960

[3.26] *J.L. Addis*, private e-mail exchange with the authors (see wbapdx32.PDF on the web)

[3.27] *M.E. Van Valkenburg*, Introduction to Modern Network Synthesis,
John Wiley, New York, 1960

[3.28] SPICE - An Overview,
http://www.seas.upenn.edu/~jan/spice/spice.overview.html

[3.29] ORCAD PSpice page - Info and download of latest free evaluation version,
http://www.orcad.com/products/pspice/pspice.htm

[3.30] MicroCAP, Spectrum Software, Inc.,
http://www.spectrum-soft.com

[3.31] *R.J. Widlar*, Some Circuit Design Techniques for Linear Integrated Circuits,
IEEE Transactions on Circuit Theory, Vol. CT-12, Dec. 1965, pp. 586–590

[3.32] *G.R. Wilson*, A Monolithic Junction FET-NPN Operational Amplifier,
IEEE Journal of Solid-State Circuits, Vol. SC-3, Dec. 1968, pp. 341–348

[3.33] *A.P. Brokaw*, A Simple Three-Terminal IC Bandgap Reference,
IEEE Journal of Solid-State Circuits, Vol. SC-9, Dec. 1974, pp. 388–393

[3.34] *S. Roach*: Signal Conditioning in Oscilloscopes and the Spirit of Invention,
J. Williams (Editor), The Art and Science of Analog Circuit Design, Ch. 7,
Butterworth-Heinemann, 1995, ISBN 0-7506-9505-6

[3.35] *P. Starič*, Wideband JFET Source Follower,
Electronic Engineering, Aug. 1992, pp. 29–34

[3.36] *J.M. Miller*, <http://www.ieee.org/organizations/history_center/legacies/miller.html>

P. Starič, E. Margan:

Wideband Amplifiers

Part 4:

Cascading Amplifier Stages, Selection of Poles

For every circuit design there is an equivalent and opposite redesign!
A generalization of Newton's Law by
Derek F. Bowers, Analog Devices

To Calculate Or Not To Calculate — That Is Not A Question

In the fourth part of this book we discuss some basic system integration procedures and derive, based on two different optimization criteria, the two system families of poles, which we have already used extensively in previous parts: the Butterworth system family, i.e., the systems with a maximally flat amplitude response, and the Bessel system family, i.e., the systems with a maximally flat envelope delay.

Once we derive the relations from which the system poles are calculated, we shall present the resulting poles in table form, in the same way as is traditionally done in the literature (but, more often than not, without the corresponding derivation procedures).

Some readers might ask why on earth in the computer era do we bother to write tables full of numbers, which very probably no one will ever refer to? The answer is that many amplifier designers are 'analog by vocation', they use the computer only when they must and they like to do lots of paperwork before they finally sit by the breadboard. For them, without those tables a book like this would be incomplete (even if many of them first sit by the breadboard with the soldering iron in one hand and a 'scope probe in the other, and do the paperwork later!).

Anyway, for the younger generations we provide the necessary computer routines in Part 6.

The principles developed in this part will be used in Part 5, where some more sophisticated amplifier building blocks are discussed.

Contents:

List of Figures:

List of Tables:

4.0 Introduction

In a majority of cases the desired *gain* × *bandwidth* product is not achievable with a single transistor amplifier stage. Therefore more stages must be connected in cascade. But to do it correctly we must find the answer to several questions:

- Should all stages be equal or different?
- What is the optimal pole pattern for obtaining a desired response?
- Is it important which pole (pair) is assigned to which stage?
- Is it better to use many simple (first- and second-order) stages or is it worth the trouble to try more complex (third-, fourth- or higher order) stages?
- What is the optimum single stage gain to achieve the greatest possible gain × bandwidth product for a given number of stages?
- Is it possible to construct an ideal multi-stage amplifier with either a maximally flat amplitude (MFA) response or a maximally flat envelope delay (MFED), and if not, how close to the ideal response can we come?

These are the main questions which we shall try to answer in this part.

In Sec. 4.1 we discuss a cascade of identical DC coupled amplifier stages, with loads consisting of a parallel connection of a resistance and a (stray) capacitance. There we derive the formula for the calculation of an optimum number of amplifying stages to obtain the required gain with the smallest rise time possible for the complete amplifier.

Next we derive the expression for the optimum gain of an individual amplifying stage of a multi-stage amplifier in order to achieve the smallest possible rise time. We also discuss the effect of AC coupling between particular stages by means of a simple RC network.

Butterworth poles, which are needed to achieve an MFA response, are derived next. This leads to the discussion of the (im)possibility to design an ideal MFA amplifier.

Then we derive the Bessel poles which provide the MFED response. Since they are derived from the condition for a unit envelope delay, the upper cut off frequency increases with the number of poles. Therefore we also present the derivation of two different pole normalizations: to equal cut off frequency and to equal stop band asymptote. We discuss the (im)possibility of designing an amplifier with the frequency response approaching an ideal Gaussian curve. Further we discuss the interpolation between the Bessel and the Butterworth poles.

Finally, we explain the merit of using staggered Bessel poles versus repeated second-order Bessel pole pairs.

Wherever practical, we calculate and plot the frequency, phase, group delay and step response to allow a quick comparison of different concepts.

4.1 A Cascade of Identical, DC Coupled, *RC* Loaded Stages

A multi-stage amplifier with DC coupled, RC loaded stages is shown in Fig. 4.1.1. All stages are assumed to be identical. Junction field effect transistors (JFETs) are being used as active devices, since we want to focus on essential design problems; with bipolar junction transistors (BJTs), we would have to consider the loading effect of a relatively complicated base input impedance [Ref. 4.1].

At each stage load the capacitance C_k $(k = 1 \ldots n)$ represents the sum of all the stray capacitances at the k^{th} node, including C_{GS} (the gate–source capacitance) and the $C_{\text{GD}}(1 + A_k)$ equivalent capacitance (where C_{GD} is the gate–drain capacitance and A_k is the voltage gain of the individual stage). By doing so we obtain a simple parallel RC load in each stage. The input resistance of a JFET is many orders of magnitude higher than the loading resistor R, so we can neglect it. All loading resistors are of equal value, and so are the mutual conductances g_{m} of all the JFETs; therefore all individual gains A_k are equal as well. Consequently, all the half power frequencies $\omega_{\text{h}k}$ and all the rise times $\tau_{\text{r}k}$ of the individual stages are also equal. In order to simplify the circuit further, we have not drawn the bias voltages and the power supply (which must represent a short circuit for AC signals; obviously, a short for DC would not be very useful!).

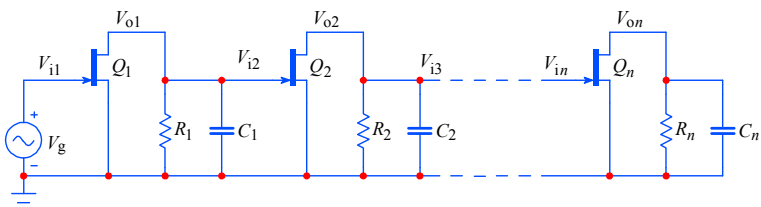


Fig. 4.1.1: A multi-stage amplifier as a cascade of identical, DC coupled, RC loaded stages.

The voltage gain of an individual stage is:

$$A_k = g_{\text{m}}\, R\, \frac{1}{1 + j\omega RC} \tag{4.1.1}$$

with the magnitude:

$$|A_k| = \frac{g_{\text{m}}\, R}{\sqrt{1 + (\omega/\omega_{\text{h}})^2}} \tag{4.1.2}$$

where:

g_{m} = mutual conductance of the JFET, [S] (siemens, [S] = [1/Ω]);

$\omega_{\text{h}} = 1/RC$ = upper half power frequency of an individual stage, [rad/s].

4.1.1 Frequency Response and the Upper Half Power Frequency

We have n equal stages with equal gains: $A_1 = A_2 = \cdots = A_n = A_k$. The gain of the complete amplifier is then:

$$A = A_1 \cdot A_2 \cdot A_3 \cdots A_n = A_k^n = \left[\frac{g_{\text{m}}\, R}{1 + j\omega\, RC}\right]^n \tag{4.1.3}$$

The magnitude is:

$$|A| = \left[\frac{g_m R}{\sqrt{1 + (\omega/\omega_h)^2}} \right]^n \tag{4.1.4}$$

To be able to compare the bandwidth of the multi-stage amplifier for different number of stages, we must normalize the magnitude by dividing Eq. 4.1.4 by the system DC gain $A_0 = (g_m R)^n$. The plots are shown in Fig. 4.1.2. It is evident that the system bandwidth, ω_H, shrinks with each additional amplifying stage.

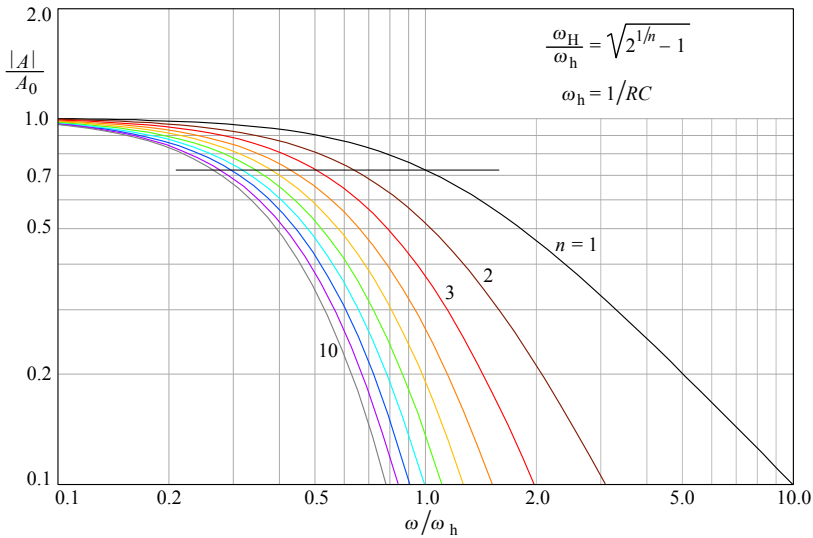


Fig. 4.1.2: Frequency response of an n-stage amplifier ($n = 1, 2, \ldots, 10$). To compare the bandwidth, the gain was normalized, i.e., divided by the system DC gain, $(g_m R)^n$. For each n, the bandwidth (the crossing of the 0.707 level) shrinks by $\sqrt{2^{1/n} - 1}$.

The upper half power frequency of the amplifier can be calculated by a simple relation:

$$\left[\frac{1}{\sqrt{1 + (\omega_H/\omega_h)^2}} \right]^n = \frac{1}{\sqrt{2}} \tag{4.1.5}$$

By squaring we obtain:

$$\left[1 + (\omega_H/\omega_h)^2\right]^n = 2 \quad \Rightarrow \quad \left(\frac{\omega_H}{\omega_h} \right)^2 = 2^{1/n} - 1 \tag{4.1.6}$$

The upper half power frequency of the complete n-stage amplifier is:

$$\boxed{\omega_H = \omega_h \sqrt{2^{1/n} - 1}} \tag{4.1.7}$$

At high frequencies, the first stage response slope approaches the $-6\,\mathrm{dB/octave}$ asymptote ($-20\,\mathrm{dB/decade}$). The meaning of this slope is explained in Fig. 4.1.3. For the second stage the slope is twice as steep, and for the n^{th} stage it is n times steeper.

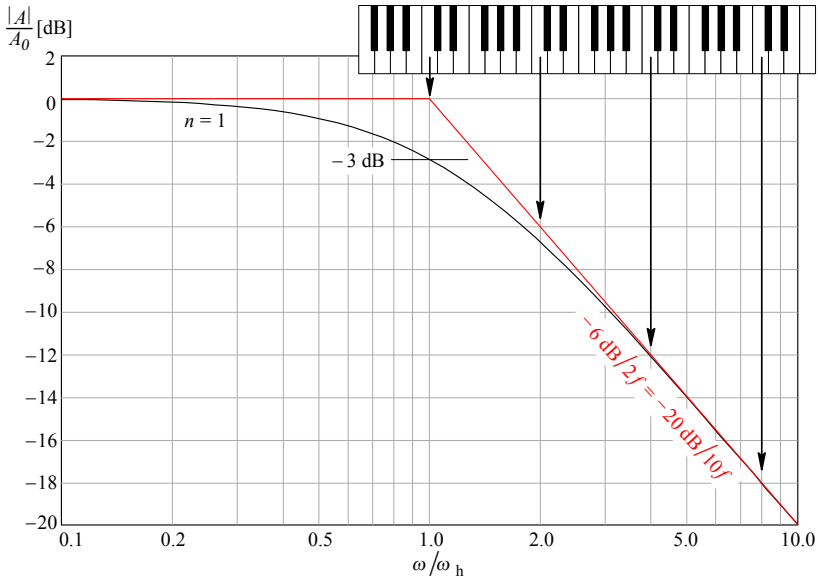


Fig. 4.1.3: The first-order system response and its asymptotes. Below the cut off, the asymptote is the level equal to the system gain at DC (normalized here to 0 dB). Above the cut off, the slope is $-6\,\mathrm{dB/octave}$ (an octave is a frequency span from f to $2f$), which is also equal to $-20\,\mathrm{dB/decade}$ (a frequency decade is a span from f to $10\,f$).

The values of ω_{H} for $n = 1-10$ are reported in Table 4.1.1.

Table 4.1.1

n	1	2	3	4	5	6	7	8	9	10
ω_{H}	1.000	0.644	0.510	0.435	0.386	0.350	0.323	0.301	0.283	0.269

With ten equal stages connected in cascade the bandwidth is reduced to a poor $0.269\,\omega_{\mathrm{h}}$; such an amplifier is definitely not very efficient for wideband amplification.

Alternatively, in order to preserve the bandwidth a n-stage amplifier should have all its capacitors reduced by the same factor, $\sqrt{2^{1/n}-1}$. But in wideband amplifiers we already strive to work with stray capacitances only, so this approach is not a solution.

Nevertheless, the amplifier in Fig. 4.1.1 is the basis for more efficient amplifier configurations, which we shall discuss later.

4.1.2 Phase Response

Each individual stage of the amplifier in Fig. 4.1.1 has a frequency dependent phase angle:

$$\varphi_k = \arctan \frac{\Im\{F(j\omega)\}}{\Re\{F(j\omega)\}} = \arctan(-\omega/\omega_h) \tag{4.1.8}$$

where $F(j\omega)$ is taken from Eq. 4.1.1. For n equal stages the total phase angle is simply n times as much:

$$\boxed{\varphi_n = n \arctan(-\omega/\omega_h)} \tag{4.1.9}$$

The phase responses are plotted in Fig. 4.1.4. Note the high frequency asymptotic phase shift increasing by $\pi/2$ (or 90°) for each n. Also note the shift at $\omega = \omega_h$ being exactly $n\,\pi/4$, in spite of a reduced ω_h for each n.

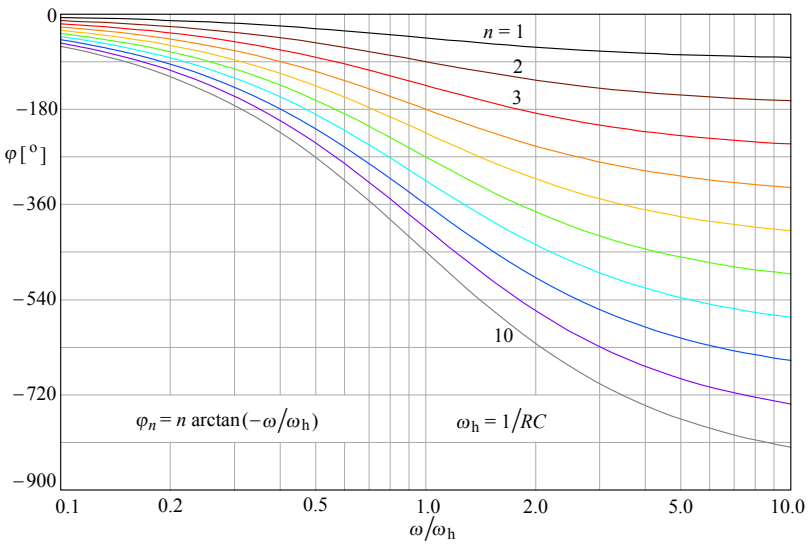


Fig. 4.1.4: Phase angle of the amplifier in Fig. 4.1.1, for $n = 1$–10 amplifying stages.

4.1.3 Envelope Delay

For a single amplifying stage ($n = 1$) the envelope delay is the frequency derivative of the phase, $\tau_{en} = d\varphi_n/d\omega$ (where φ_n is given by Eq. 4.1.9). The normalized single stage envelope delay is:

$$\tau_e\,\omega_h = \frac{-1}{1 + (\omega/\omega_h)^2} \tag{4.1.10}$$

and for n equal stages:

$$\tau_{en}\,\omega_h = \frac{-n}{1+(\omega/\omega_h)^2} \tag{4.1.11}$$

Fig. 4.1.5 shows the frequency dependent envelope delay for $n = 1$–10. Note the delay at $\omega = \omega_h$ being exactly $1/2$ of the low frequency asymptotic value.

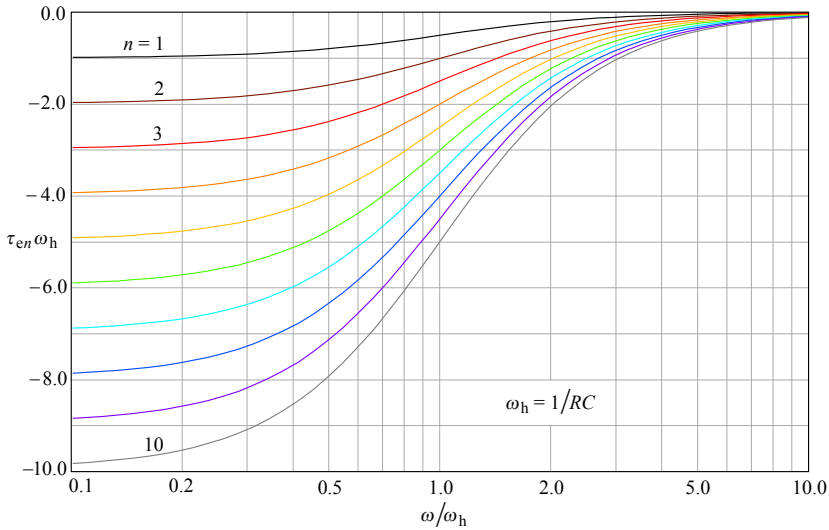


Fig. 4.1.5: Envelope delay of the amplifier in Fig. 4.1.1, for $n = 1$–10 amplifying stages. The delay at $\omega = \omega_h$ is $1/2$ of the low frequency asymptotic value. Note that if we were using f/f_h for the abscissa, we would have to divide the τ_e scale by 2π.

4.1.4 Step Response

To obtain the step response, the amplifier in Fig. 4.1.1 must be driven by the unit step function:

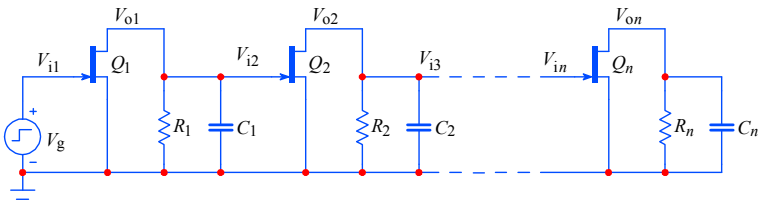


Fig. 4.1.6: Amplifier with n equal DC coupled stages, excited by the unit step

We can derive the step response expression from Eq. 4.1.1 and Eq. 4.1.3. In order to simplify and generalize the expression we shall normalize the magnitude by dividing the transfer function by the DC gain, $g_m R$, and normalize the frequency by setting $\omega_h = 1/RC = 1$. Since we shall use the $\mathcal{L}^{-1}$ transform we shall replace the variable $j\omega$ by the complex variable $s = \sigma + j\omega$.

With all these changes we obtain:

$$F(s) = \frac{1}{(1+s)^n} \tag{4.1.12}$$

The amplifier input is excited by the unit step, therefore we must multiply the above formula by the unit step operator $1/s$:

$$G(s) = \frac{1}{s\,(1+s)^n} \tag{4.1.13}$$

The corresponding function in the time-domain is:

$$g(t) = \mathcal{L}^{-1}\{G(s)\} = \sum \text{res}\, \frac{\mathrm{e}^{s\,t}}{s\,(1+s)^n} \tag{41.14}$$

We have two residues. The first one does not depend of n:

$$\text{res}_0 = \lim_{s\,\to\,0} s\left[\frac{\mathrm{e}^{s\,t}}{s\,(1+s)^n}\right] = 1$$

whilst the second does:

$$\text{res}_1 = \lim_{s\,\to\,1} \frac{1}{(n-1)!} \cdot \frac{d^{(n-1)}}{ds^{(n-1)}}\left[(1+s)^n\, \frac{\mathrm{e}^{s\,t}}{s\,(1+s)^n}\right] =$$

$$= \lim_{s\,\to\,1} \frac{1}{(n-1)!} \cdot \frac{d^{(n-1)}}{ds^{(n-1)}}\left(\frac{\mathrm{e}^{s\,t}}{s}\right) \tag{4.1.15}$$

Since res_1 depends on n, for $n = 1$ we obtain:

$$\text{res}_1\Big|_{n=1} = -\,\mathrm{e}^{-t} \tag{4.1.16}$$

for $n = 2$:

$$\text{res}_1\Big|_{n=2} = -\,\mathrm{e}^{-t}\,(1+t) \tag{4.1.17}$$

for $n = 3$:

$$\text{res}_1\Big|_{n=3} = -\,\mathrm{e}^{-t}\left(1 + t + \frac{t^2}{2}\right) \tag{4.1.18}$$

… etc.

The general expression for the step response for any n is:

$$\boxed{g_n(t) = \mathcal{L}^{-1}\{\mathrm{G}(s)\} = \text{res}_0 + \text{res}_1(n) = 1 - \mathrm{e}^{-t} \sum_{k=1}^{n} \frac{t^{k-1}}{(k-1)!}} \tag{4.1.19}$$

Here we must consider that $0! = 1$, by definition.

As an example, by inserting $n = 5$ into Eq. 4.1.19 we obtain:

$$g_5(t) = 1 - \mathrm{e}^{-t}\left(1 + \frac{t}{1!} + \frac{t^2}{2!} + \frac{t^3}{3!} + \frac{t^4}{4!}\right) \tag{4.1.20}$$

The step response plots for $n = 1$–10, calculated by Eq. 4.1.19, are shown in Fig. 4.1.7. Note that there is **no overshoot** in any of the curves. Unfortunately, the efficiency of this kind of amplifier in the sense of the 'bandwidth per number of stages' is poor, since it has no peaking networks which would prevent the decrease of bandwidth with n.

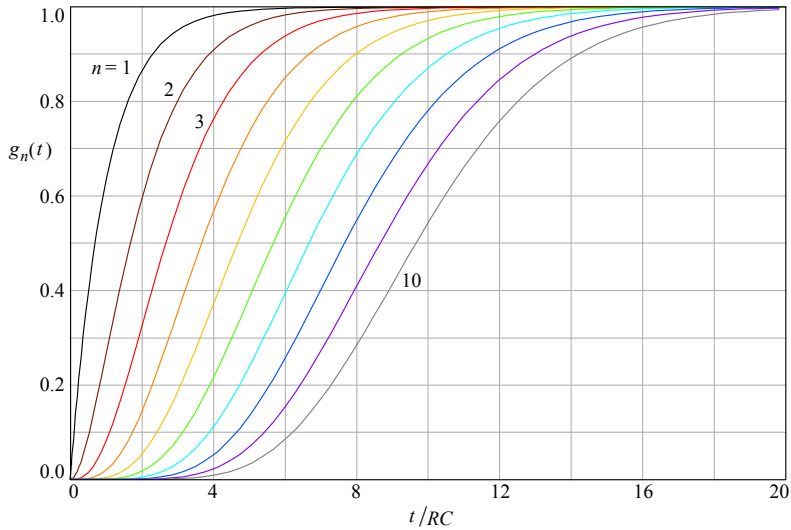


Fig. 4.1.7: Step response of the amplifier in Fig. 4.1.6, for $n = 1$–10 amplifying stages

4.1.5 Rise Time Calculation

In a multi-stage amplifier, where each particular stage has its respective rise time, $\tau_{r1}, \tau_{r2}, \ldots, \tau_{rn}$, we approximate the system's rise time [Ref. 4.2] as:

$$\tau_r \approx \sqrt{\tau_{r1}^2 + \tau_{r2}^2 + \tau_{r3}^2 + \cdots + \tau_{rn}^2} \tag{4.1.21}$$

In Part 2, Sec. 2.1.1, Eq. 2.1.1–4, we have calculated the rise time of an amplifier with a simple RC load to be $\tau_{r1} = 2.20\,RC$. Since here we have n equal stages the rise time of the complete amplifier is approximately:

$$\boxed{\tau_r \approx \tau_{r1}\sqrt{n} = 2.20\,RC\sqrt{n}} \tag{4.1.22}$$

Table 4.1.2 shows the actual rise time increasing with the number of stages. The numbers were calculated by using Eq. 4.1.19.

Table 4.1.2

n	1	2	3	4	5	6	7	8	9	10
τ_{rn}	2.20	3.36	4.22	4.94	5.56	6.12	6.64	7.11	7.56	7.98
τ_{rn}/τ_{r1}	1.00	1.53	1.92	2.25	2.53	1.79	3.02	3.24	3.44	3.63

4.1.6 Slew Rate Limit

The equations derived so far describe the small signal properties of an amplifier. If the signal amplitude is increased the maximum slope of the output signal dV_o/dt becomes limited by the maximum current available to charge any capacitance present at the particular node. The amplifier stage which must handle the largest signal is the first to run into the slew rate limiting; usually it is the output stage.

To find the slew rate limit we drive the amplifier by a sinusoidal signal and increase the input amplitude until the output amplitude just begins to saturate; then we increase the frequency until we notice that the middle part of the sinusoidal waveform becomes distorted (changing linearly with time) and then decrease the frequency until the distortion just disappears. That frequency is equal to the *full power bandwidth* ω_{FP} (in *radians per second*, [rad/s]). Generally, an amplifier need not have the positive and negative slope equally steep; then it is the less steep slope to set the limit.

$$\frac{dV_o}{dt} = \frac{I_{omax}}{C} \tag{4.1.23}$$

For a sinusoidal input signal of angular frequency ω_{FP} and amplitude V_{max}, the slope varies with time as:

$$\frac{d(V_{max}\,\sin\omega_{FP}t)}{dt} = V_{max}\,\omega_{FP}\cos\omega_{FP}t \tag{4.1.24}$$

and it has a maximum at $\cos\omega_{FP}t = 1$ (which is at $t = 0$; see Fig. 4.1.8). Therefore:

$$\boxed{\textit{slew rate}: \quad SR = V_{max}\,\omega_{FP}} \tag{4.1.25}$$

The slew rate is usually expressed in *volts per microsecond* [V/µs]; for contemporary amplifiers a more appropriate figure would be *volts per nanosecond* [V/ns].

The time elapsed from the negative to the positive peak of the sinusoid with the amplitude V_{max} and frequency ω_{FP} is equal to one half of the period, $T_{FP}/2$, of the full power bandwidth frequency, f_{FP}, in Hz:

$$\boxed{\frac{T_{FP}}{2} = \frac{\pi}{\omega_{FP}} = \frac{1}{2\,f_{FP}}} \tag{4.1.26}$$

If we increase the signal frequency beyond f_{FP} the waveform will eventually be distorted into a linear ramp shape, with a reduced amplitude. If we imagine a ramp with the amplitude equal to that of the sine wave the slewing time can be found from:

$$SR = \frac{\Delta V}{\Delta t} = \frac{2\,V_{max}}{t_{slew}} \quad \Rightarrow \quad t_{slew} = \frac{2\,V_{max}}{V_{max}\,\omega_{FP}} = \frac{2}{\omega_{FP}} \tag{4.1.27}$$

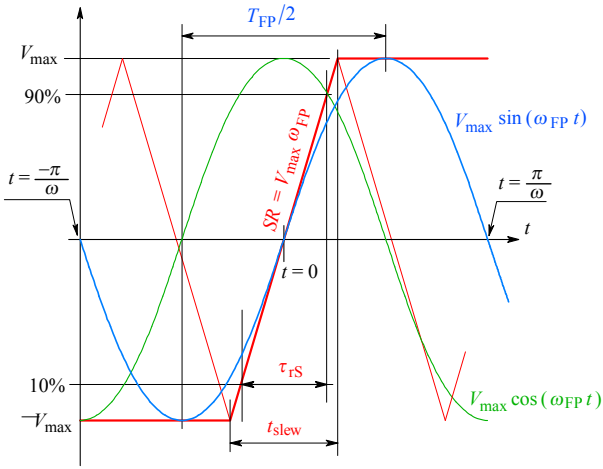


Fig. 4.1.8: Slew rate limiting: definitions of parameters.

All these calculations are valid for sinusoidal signals only, the slope of which decreases progressively and becomes zero at the peak voltage $\pm V_{max}$. These criteria can not be applied if the amplifier input is excited by a step pulse with the same output voltage range. In this case we, rather, speak of rise time. From 10 % to 90 % of $2V_{max}$, the output voltage rises by $0.8 \times 2V_{max}$. So the results of Eq. 4.1.27 must be multiplied by 0.8 to obtain:

$$\tau_{rS} = \frac{0.8 \times 2}{\omega_{FP}} = \frac{1.6}{2\,\pi\,f_{FP}} \approx \frac{0.2546}{f_{FP}} \tag{4.1.28}$$

Eq. 4.1.26 is generally used to characterize operational amplifiers, whilst for wideband and pulse amplifiers we prefer Eq. 4.1.28. It is very important to distinguish between the two definitions, since in most technical specifications they are tacitly assumed.

4.1.7 Optimum Single Stage Gain and Optimum Number of Stages

The first stage of the n-stage cascade amplifier of Fig. 4.1.6 has the voltage gain $A_1 = v_{o1}/v_{i1}$, the second stage gain is $A_2 = v_{o2}/v_{i2} = v_{o2}/v_{o1}$ and so on, up to $A_n = v_{on}/v_{in} = v_{on}/v_{o(n-1)}$ for the n^{th} stage. Then the total gain is the product of individual stage gains:

$$A = \frac{v_{on}}{v_{i1}} = A_1 \cdot A_2 \cdots A_n \tag{4.1.29}$$

If all the amplifying stages are identical, we denote the gain of each stage as A_k, each loading resistor as R, and each loading capacitor as C. Then the total gain is:

$$A = (A_k)^n \tag{4.1.30}$$

The rise time of the complete multi-stage amplifier is approximated as the square root of the sum of the individual rise times squared (Eq. 4.1.21), but since the amplitude after each gain stage is different, we must normalize the rise times by multiplying each with its own gain factor A_k:

$$\tau_r = \sqrt{\sum_{i=1}^{n} (A_k\, \tau_{rk})^2} \tag{4.1.31}$$

We have assumed that all stages have identical gain A_k and identical rise time τ_{rk}. Therefore:

$$\tau_r = \sqrt{n\,(A_k\, \tau_{rk})^2} = \sqrt{n}\; A_k\, \tau_{rk} = \sqrt{n}\; A_k\, 2.2\, RC \tag{4.1.32}$$

where τ_{rk} is the rise time of an individual stage, as calculated in Part 2, Eq. 2.1.4. By considering Eq. 4.1.30, we obtain the following relation:

$$\boxed{\frac{\tau_r}{\tau_{rk}} = \sqrt{n}\; A^{1/n}} \tag{4.1.33}$$

We have plotted this relation in Fig. 4.1.9, using the system total gain A as the parameter. Note that, in order to see the function $\tau_r(n)$ better, we have assumed a continuous n; of course, we can not have, say, 4.63 stages — n must be an integer.

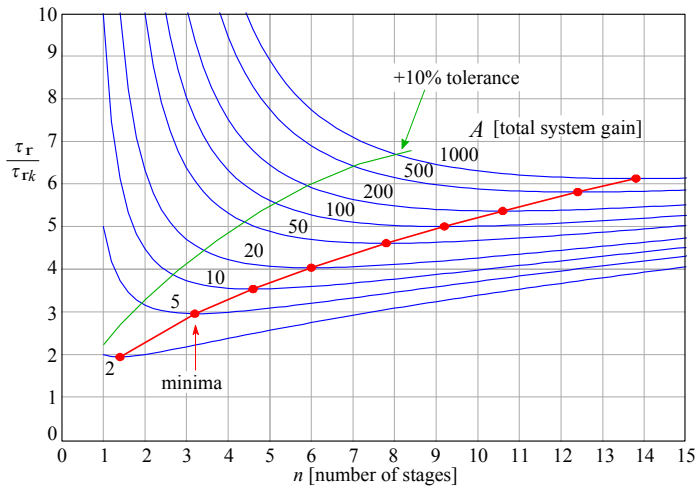


Fig. 4.1.9: Minimal relative rise time as a function of the number of stages n and the total system gain A. Close to the minima the curves are relatively flat, so in practice we can trade off, say, a 10% increase in the system rise time and reduce the required number of stages accordingly; i.e., to achieve the gain of 100, only 5 stages could be used instead of 9, with a slight rise time increase.

From this diagram we can find the optimum number of the amplifying stages n_{opt} if the total system gain A is known. These optima lie on the valleys of the curves and in the following discussion we will derive the necessary formulae.

The ratio A/τ_r characterizes the design efficiency of the amplifier. To design a multi-stage amplifier with the smallest possible rise time, we can differentiate τ_r from Eq. 4.1.33 with respect to n, and equate the result to zero to find the minimum:

$$\frac{\partial \tau_r}{\partial n} = \frac{\partial \left(\sqrt{n}\, A^{1/n}\, \tau_{rk}\right)}{\partial n} = 0 \tag{4.1.34}$$

The differentiation is solved as:

$$\frac{\tau_{rk}}{A}\left(\frac{1}{2\sqrt{n}}\, A^{1/n} - \sqrt{n}\, \frac{A^{1/n}}{n^2}\, \ln A\right) = 0 \tag{4.1.35}$$

Because neither τ_{rk} nor A are zero we can equate the expression in parentheses to zero:

$$\frac{1}{2\sqrt{n}} - \frac{\sqrt{n}}{n^2}\, \ln A = 0 \tag{4.1.36}$$

By multiplying this by the $2\sqrt{n}$, we obtain an important intermediate result:

$$1 - \frac{2}{n}\ln A = 0 \quad \Rightarrow \quad n = 2\ln A \tag{4.1.37}$$

For a given total gain A the optimum number of amplifying stages is approximately:

$$\boxed{1 \le n_{opt} \le \text{int}(2\ln A)} \tag{4.1.38}$$

and since we can not have, say, 3.47 amplifying stages, we round the result to the nearest integer, the smallest obviously being 1.

On the basis of this simple relation we can draw the line *a* in Fig. 4.1.10 for a quick estimation of the number of amplifying stages necessary to obtain the smallest rise time if the total system gain A is known. Again, the required number of amplifying stages can be reduced in practice, as indicated in Fig. 4.1.10 by the line *b*, without significantly increasing the rise time. Owing to reasons of economy, the most simple systems are often designed far from optimum, as indicated by the bars and the line *c*.

From Eq. 4.1.38 we can find the optimal gain value of the individual stage, independent of the actual number of stages in the system. For n equal stages it is:

$$A_k = A^{1/n} = A^{1/(2\ln A)} \tag{4.1.39}$$

By taking a logarithm of this expression, we obtain:

$$\ln A_k = \frac{1}{2\ln A} \cdot \ln A = \frac{1}{2} \tag{4.1.40}$$

The optimal individual stage gain for the total minimal rise time is then:

$$\boxed{A_{k\text{opt}} = \mathrm{e}^{1/2} = \sqrt{\mathrm{e}} \simeq 1.65} \tag{4.1.41}$$

Note that Eq. 4.1.41, as well as the approximations Eq. 4.1.33 and Eq. 4.1.38, can be applied also to amplifiers with peaking stages, although peaking stages are much more efficient in the gain-bandwidth product sense, as we shall see in Sec. 4.3 and 4.4.

This expression gives us the optimal value of the gain of the individual amplifying stage which minimizes the total rise time of a multi-stage amplifier. In practice we usually take a higher value, say, between 2 and 4, in order to decrease the cost and simplify the design. Eq. 4.1.41 can also be used for peaking stages.

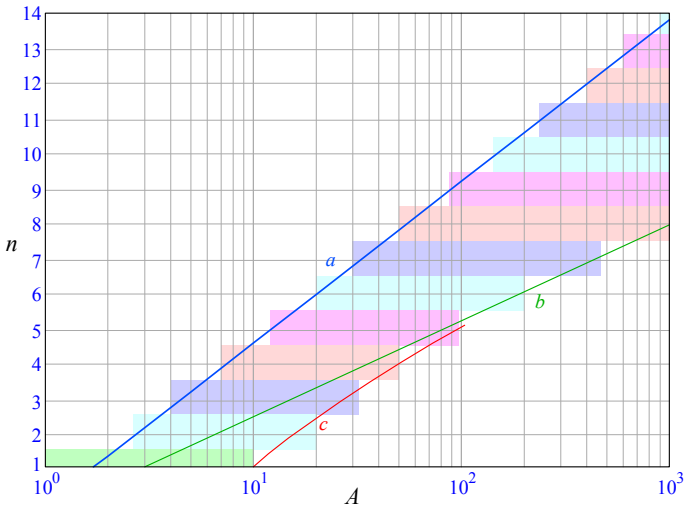


Fig. 4.1.10: The optimal number of stages, n, required to achieve the minimal rise time, given the total system gain A, as calculated by Eq. 4.1.38, is shown by line *a*. In practice, owing to economy reasons, we tend to use a lower number; the line *b* shows the same +10% rise time trade off as in Fig. 4.1.9. In low complexity systems we usually make even greater tradeoffs, as in *c*. The bars indicate the range of gain A which can be achieved with a particular number of stages n without departing much from the lowest rise time.

4.2 A Multi-stage Amplifier with Identical, AC Coupled Stages

In Fig. 4.1.1 all the amplifying stages were DC coupled. In times when the amplifiers were designed with electronic tubes the DC coupling was generally avoided to prevent drift owed to tube aging, unstable cathode heater voltages (changing of temperature), hum, and poor insulation between the heater and cathode. With the appearance of bipolar transistors and FETs only the temperature dependent drift remained on this nasty list. Because of it, many TV video and broadband RF amplifiers still use AC coupling between amplifying stages.

However, AC coupling introduces other undesirable properties, which in certain cases might not be acceptable. It is therefore interesting to investigate a multi-stage amplifier with equal stages, similar to that in Fig. 4.1.1, except that all the stages are AC coupled. In Fig. 4.2.1 we see a simplified circuit diagram of such an amplifier and here we are interested in its low-frequency performance.

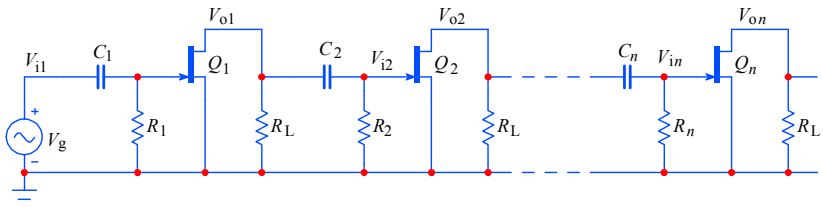


Fig. 4.2.1: Multi-stage amplifier with AC coupled stages. Again, to simplify the analysis, the power supply and the bias voltages are not shown.

Since we want to focus on essential problems only, here, too, we use FETs, instead of BJTs, in order to avoid the complicated expression for the base input impedance of each stage. Moreover, in a wideband amplifier we can assume $R_L \ll R_n$, so we shall neglect the effect of R_n on gain. On the other hand, R_n and C_n set the low frequency limit of each stage, which is $\omega_1 = 1/R_nC_n = 1/RC$, if all stages are identical. Usually, ω_1 is many orders of magnitude below ω_h, so we can neglect the stray and input capacitances (both effectively in parallel to the loading resistors R_L) as well. Thus, near ω_1, the voltage gain of each stage is:

$$A_n = g_m R_L \frac{j\omega/\omega_1}{1 + j\omega/\omega_1} \tag{4.2.1}$$

and the magnitude is:

$$|A_n| = g_m R_L \frac{\omega/\omega_1}{\sqrt{1 + (\omega/\omega_1)^2}} \tag{4.2.2}$$

With all input time constants equal to RC the system gain A is A_n to the n^{th} power:

$$A = \left[g_m R_L \frac{j\omega/\omega_1}{1 + j\omega/\omega_1} \right]^n \tag{4.2.3}$$

with the magnitude:

$$|A| = \left[g_m R_L \frac{\omega/\omega_1}{\sqrt{1 + (\omega/\omega_1)^2}} \right]^n \tag{4.2.4}$$

4.2.1 Frequency Response and Lower Half Power Frequency

In Fig. 4.2.2 we show the frequency response plots according to Eq. 4.2.4, normalized in amplitude by dividing it by $(g_m R_L)^n$.

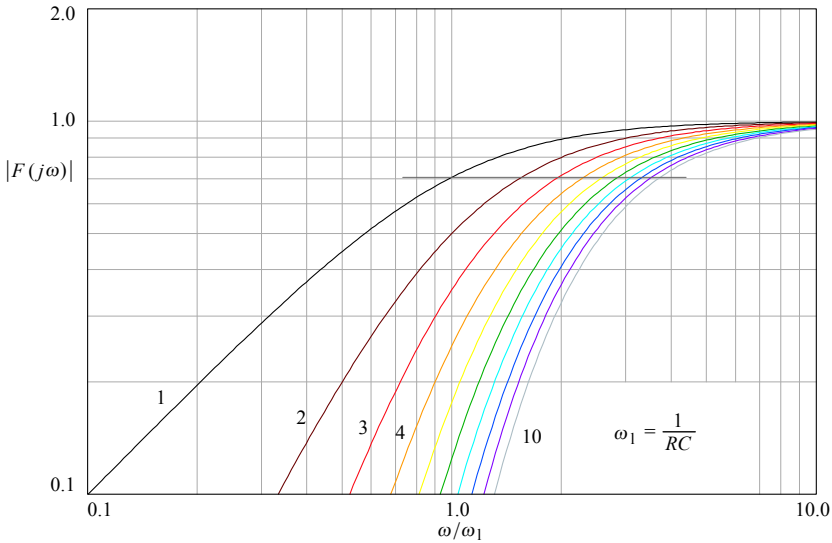


Fig. 4.2.2: Frequency response magnitude of the AC coupled amplifier for $n = 1$–10. The frequency scale is normalized to the lower cut off frequency ω_1 of the single stage.

It is evident that the lower half power frequency ω_L of the complete amplifier increases with the number of stages. We can express ω_L as a function of n from:

$$\left[\frac{\omega_L/\omega_1}{\sqrt{1+(\omega_L/\omega_1)^2}}\right]^n = \frac{1}{\sqrt{2}} \tag{4.2.5}$$

By eliminating the fractions:

$$[1+(\omega_L/\omega_1)^2]^n = 2\,(\omega_L/\omega_1)^{2n} \tag{4.2.6}$$

and taking the n^{th} root:

$$1+(\omega_L/\omega_1)^2 = 2^{1/n}(\omega_L/\omega_1)^2 \tag{4.2.7}$$

and rearranging a little:

$$(\omega_L/\omega_1)^2\,(2^{1/2}-1) = 1 \tag{4.2.8}$$

we obtain the lower half power frequency of the complete multi-stage amplifier:

$$\boxed{\omega_L = \omega_1 \frac{1}{\sqrt{2^{1/n}-1}}} \tag{4.2.9}$$

We normalize this equation in frequency by setting $\omega_1 = 1/RC = 1$. The normalized values of the lower half power frequency for n equal stages, $n = 1$–10, are shown in Table 4.2.1:

Table 4.2.1

n	1	2	3	4	5	6	7	8	9	10
ω_L/ω_1	1.000	1.554	1.961	2.299	2.593	2.858	3.100	3.334	3.534	3.733

4.2.2 Phase Response

The phase shift for a single stage is:

$$\varphi_i = \arctan(\omega_1/\omega) \tag{4.2.10}$$

The phase shift is positive and this means a phase advance. For n stages the total phase advance is simply n times as much:

$$\boxed{\varphi_n = n \arctan(\omega_1/\omega)} \tag{4.2.11}$$

The corresponding plots for $n = 1$–10 are shown in Fig. 4.2.3. Note the phase shift at $\omega = \omega_1$ being exactly $1/2$ the low frequency asymptotic value.

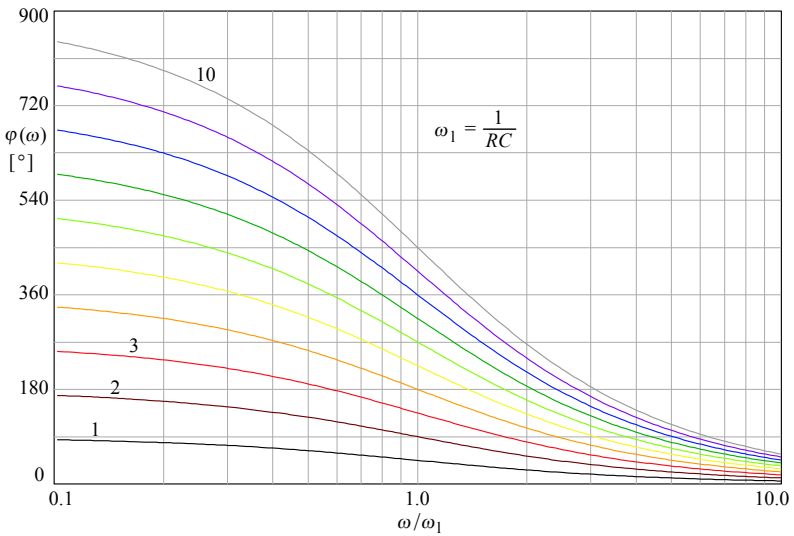


Fig. 4.2.3: Phase angle as a function of frequency for the AC coupled n-stage amplifier, $n = 1$–10. The frequency scale is normalized to the lower cutoff frequency of the single stage.

We will omit the calculation of envelope delay since in the low frequency region this aspect of amplifier performance is not very important.

4.2.3 Step Response

By replacing the sine wave generator in Fig. 4.2.1 with a unit step generator, we obtain the time domain step response of the AC coupled multi-stage amplifier. We want the plots to be normalized in amplitude, so we normalize Eq. 4.2.3 by dividing it by $(g_m R_L)^n$, the total amplifier gain at DC. We will use the $\mathcal{L}^{-1}$ transform, so we replace the normalized variable $j\omega/\omega_1$ by the complex variable $s = \sigma + j\omega$:

$$F_n(s) = \left(\frac{s}{1+s}\right)^n \tag{4.2.12}$$

The system's frequency response must be multiplied by the unit step operator $1/s$:

$$G_n(s) = \frac{1}{s}\left(\frac{s}{1+s}\right)^n \tag{4.2.13}$$

Now we apply the $\mathcal{L}^{-1}$ transformation and obtain the time domain step response:

$$g_n(t) = \mathcal{L}^{-1}\{G_n(s)\} = \operatorname{res} G_n(s)\,\mathrm{e}^{st} = \operatorname{res} \frac{s^{n-1}}{(1+s)^n}\,\mathrm{e}^{st} \tag{4.2.14}$$

Since we have here a single pole repeated n times we have only a single residue, but — as we will see — it is composed of n summands. A general expression for the residue for an arbitrary n is:

$$g_n(t) = \lim_{s \to -1} \frac{1}{(n-1)!} \cdot \frac{d^{\,n-1}}{d\,s^{\,n-1}} \left[(1+s)^n \frac{s^{n-1}}{(1+s)^n}\,\mathrm{e}^{st}\right] \tag{4.2.15}$$

or, simplified:

$$g_n(t) = \frac{1}{(n-1)!} \cdot \frac{d^{\,n-1}}{d\,s^{\,n-1}} \left[s^{n-1}\,\mathrm{e}^{st}\right]\Big|_{s \to -1} \tag{4.2.16}$$

A few examples:

$$\begin{aligned}
n = 1 \;&\Rightarrow\; g_1(t) = \frac{1}{0!}\,\mathrm{e}^{-t} = \mathrm{e}^{-t} \\
n = 2 \;&\Rightarrow\; g_2(t) = \frac{1}{1!}\left(\mathrm{e}^{-t} - t\,\mathrm{e}^{-t}\right) = \mathrm{e}^{-t}(1-t) \\
n = 3 \;&\Rightarrow\; g_3(t) = \mathrm{e}^{-t}\left(1 - 2\,t + 0.5\,t^2\right) \\
n = 4 \;&\Rightarrow\; g_4(t) = \mathrm{e}^{-t}\left(1 - 3\,t + 1.5\,t^2 - 0.1667\,t^3\right) \\
n = 5 \;&\Rightarrow\; g_5(t) = \mathrm{e}^{-t}\left(1 - 4\,t + 3\,t^2 - 0.6667\,t^3 + 0.0417\,t^4\right)
\end{aligned} \tag{4.2.17}$$

The coefficients decrease rapidly with increasing number of stages n; e.g., the last summand for $n = 10$ is $-2.775 \cdot 10^{-6}\,t^9$.

The corresponding plots are drawn in Fig. 4.2.4. The plots for $n = 6$–9 are not shown, since it would be very difficult to distinguish the individual curves. We note that the n^{th}-order response intersects the abscissa $(n-1)$ times.

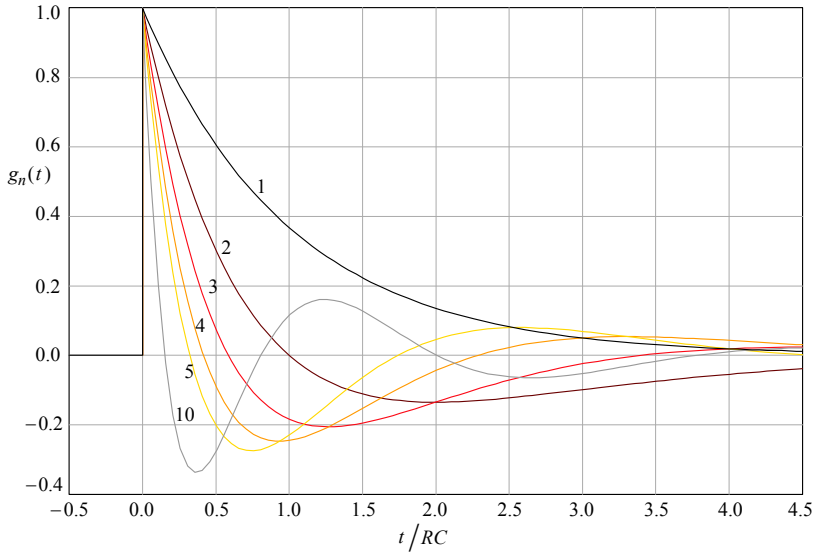


Fig. 4.2.4: Step response of the multi-stage AC coupled amplifier for $n = 1$–5 and $n = 10$.

For pulse amplification only the short starting portions of the curves come into consideration. An example for $n = 1$, 3, and 8 is shown in Fig. 4.2.5 for a pulse width $\Delta t = 0.1\,RC$.

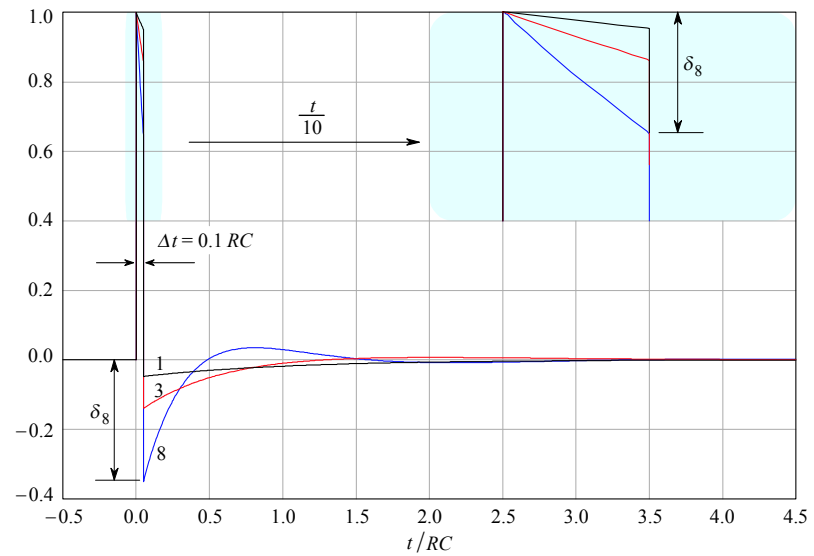


Fig. 4.2.5: Pulse response of the AC coupled multi-stage amplifier ($n = 1$, 3, and 8).

Note that the pulse in Fig. 4.2.5 sags, both on the leading and trailing edge, the sag increasing with the number of stages. We conclude that the AC coupled amplifier of Fig. 4.2.1 is not suitable for a faithful pulse amplification, except when the pulse duration is very short in comparison with the time constant RC of a single amplifying stage (say, $\Delta t \leq 0.001\,RC$).

Another undesirable property of the AC coupled amplifier is that the output voltage makes $n-1$ damped oscillations when the pulse ends, no matter how short its duration is. This is especially annoying because the input voltage is by now already zero. The undesirable result is that the effective output DC level will depend on the pulse repetition rate.

Since today the DC amplification technique has reached a very high quality level, we can consider the AC coupled amplifier an inheritance from the era of electronic tubes and thus almost obsolete. However, we still use AC coupled amplifiers to avoid the drift in those cases where the deficiencies described are not important.

4.3 A Multi-stage Amplifier with Butterworth Poles (MFA)

In multi-stage amplifiers, like the one in Fig. 4.1.1, we can apply inductive peaking at each stage, see Fig. 4.3.7. As we have seen in Part 2, Sec. 2.9, where we discussed the shunt–series peaking circuit, the equations became very complicated because we had to consider the mutual influence of the shunt and series peaking circuit. If both circuits are separated by a buffer amplifier, the analysis is simplified. Basically, this was considered by *S. Butterworth* in his article *On the Theory of Filter Amplifiers* in the review *Experimental Wireless & the Wireless Engineer* in 1930 [Ref. 4.6]. When writing the article, which Butterworth wrote when he was serving in the British Navy, he obviously did not expect that his technique might also be applied to wideband amplifiers. In general his article became the basis for filter design for generations of engineers up to the present time.

The basic Butterworth equation, which, besides to filters, can also be applied to wideband amplifiers, either with a single or many stages, is:

$$F(\omega) = \frac{1}{1 + j\,(\omega/\omega_{\mathrm{H}})^{n}} \tag{4.3.1}$$

where ω_{H} is the upper half power frequency of the (peaking) amplifier and n is an integer, representing the number of stages. A network corresponding to this equation has a maximally flat amplitude response (MFA). The magnitude of $F(\omega)$ is:

$$\boxed{|F(\omega)| = \frac{1}{\sqrt{1 + (\omega/\omega_{\mathrm{H}})^{2\,n}}}} \tag{4.3.2}$$

The magnitude derivative, $d\big|F(\omega)\big|/d\omega$ is zero at $\omega = 0$:

$$\frac{d\big|F\,(\omega)\big|}{d\,\omega} = \frac{-\,n\,(\omega/\omega_{\mathrm{H}})^{2n-1}}{[1 + (\omega/\omega_{\mathrm{H}})^{2n}]^{3/2}} \cdot \frac{1}{\omega_{\mathrm{H}}} = 0 \Bigg|_{\omega\,=\,0} \tag{4.3.3}$$

and not just the first derivative, but all the $n - 1$ derivatives of an n^{th}-order system are also zero at origin. This means that at very low frequencies ($\omega \ll \omega_{\mathrm{H}}$) the filter is essentially flat. The number of poles in Eq. 4.3.1 is equal to the parameter n and the flatness of the frequency response in the passband also increases with n. The parameter n is called the *system order*. To derive the expression for the poles we start with the denominator of Eq. 4.3.2, where the expression under the root can be simplified into:

$$1 + (\omega/\omega_{\mathrm{H}})^{2\,n} = 1 + s^{2\,n} \tag{4.3.4}$$

Whenever this expression is equal to zero, we have a pole, $F(j\omega) \rightarrow \pm\,\infty$. Thus:

$$1 + s^{2\,n} = 0 \qquad \text{or} \qquad s^{2\,n} = -1 \tag{4.3.5}$$

The roots of these equations are the poles of Eq. 4.3.1 and they can be calculated by the following general expression:

$$s = \sqrt[2n]{-1} \tag{4.3.6}$$

We solve this equation using *De Moivre*'s formula [Ref. 4.7]:

$$-1 = \cos(\pi + 2\pi q) + j\sin(\pi + 2\pi q) \tag{4.3.7}$$

where q is either zero or a positive integer. Consequently the poles are:

$$s_q = \sqrt[2n]{-1} = \sqrt[2n]{\cos(\pi + 2\,\pi\,q) + j\sin(\pi + 2\,\pi\,q)}$$

$$= \cos\left(\pi\,\frac{1+2\,q}{2\,n}\right) + j\sin\left(\pi\,\frac{1+2\,q}{2\,n}\right) \tag{4.3.8}$$

If we insert the value $0, 1, 2, \ldots, (2n-1)$ for q, we obtain $2n$ roots. The roots lie on a circle of radius $r = 1$, spaced by the angles $\pi/2n$. With this condition no pole is repeated and none of the poles lies on the imaginary axis. One half ($= n$) of the poles lie in the left side of the s-plane; these are the poles of Eq. 4.3.1. The other half of the poles lie in the positive half of the s-plane and they can be associated with the complex conjugate of $F(j\omega)$; as shown in Fig. 4.3.1, owing to the Hurwitz stability requirement, they are not useful for our purpose.

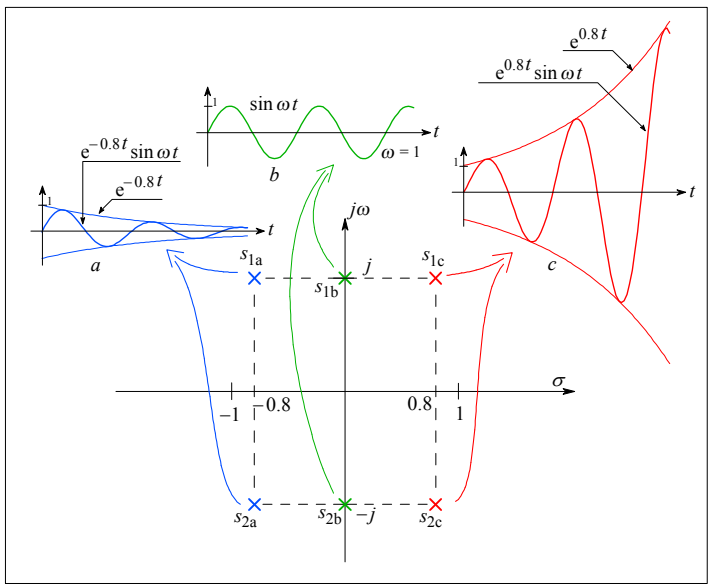


Fig. 4.3.1: Impulse response of three different complex conjugate pole pairs: The real part determines the system stability: s_{1a} and s_{2a} make the system unconditionally stable, since the negative exponent forces the response to decrease with time; s_{1b} and s_{2b} make the system conditionally stable, whilst s_{1c} and s_{2c} make it unstable.

This left- and right-half pole division is not arbitrary, but, as we have explained in Part 1, it reflects the direction of energy flow. If an unconditionally stable system is energized and then left alone, it will eventually dissipate all the energy into heath and RF radiation, so it is lost (from the system point of view) and therefore we agree to give it a negative sign. This is typical of a dominantly resistive systems. On the other hand, generators produce energy and we agree to give them a positive sign. In effect, generators can be treated as negative resistors. Inductances and capacitances can not dissipate energy, they can only store it in their associated electromagnetic fields (for a while). We therefore assign the resistive and generative action to the real axis, and the inductive and capacitive action to the imaginary axis.

For example, if we take a two–pole system with poles forming a complex conjugate pair, $s_1 = \sigma + j\omega$ and $s_2 = \sigma - j\omega$, the system impulse response function has the form:

$$f(t) = \mathrm{e}^{\sigma t} \sin \omega t \tag{4.3.9}$$

By referring to Fig. 4.3.1, let us first consider the poles $s_{1a} = -0.8 + j$ and $s_{2a} = -0.8 - j$, where $\omega = 1$. Their impulse function is a damped sinusoid:

$$f(t) = \mathrm{e}^{-0.8 t} \sin \omega t \tag{4.3.10}$$

This means that for any impulse disturbance the system reacts with a sinusoidal oscillation (governed by ω), exponentially damped (by the rate set by σ). Such behavior is typical for an unconditionally stable system. But if we move the poles to the imaginary axis ($\sigma = 0$) so that $s_{1b} = j$ and $s_{2b} = -j$ (again, $\omega = 1$), then, since there is no damping ($\mathrm{e}^0 = 1$), an impulse excites the system into a continuous sine wave:

$$f(t) = \sin \omega t \tag{4.3.11}$$

If we push the poles further to the right side of the s plane, so that $s_{1c} = 0.8 + j$ and $s_{2c} = 0.8 - j$, keeping $\omega = 1$, the slightest impulse disturbance, or even just the system's own noise, excites an exponentially rising sine wave:

$$f(t) = \mathrm{e}^{0.8 t} \sin \omega t \tag{4.3.12}$$

The poles on the imaginary axis are characteristic of a sine wave oscillator, in which we have the active components (amplifiers) set to make up for (and exactly match) any energy lost in resistive components. The poles on the right side of the s-plane also result in oscillations, but there the final amplitude is limited by the system power supply voltages. Because the active components provide much more energy than the system is capable of dissipating thermally, the top and bottom part of the waveform will be saturated, thus limiting the energy produced. Since we are interested in the design of amplifiers and not of oscillators, we shall not use the last two kinds of poles.

Let us return to the Butterworth poles. We want to find the general expression for n poles on the left side of the s-plane. A general expression for a pole s_q, derived from Eq. 4.3.8 is:

$$s_q = \cos \theta_q + \sin \theta_q \tag{4.3.13}$$

where:

$$\theta_q = \pi \frac{1 + 2q}{2n} \tag{4.3.14}$$

The poles with the angle:

$$\frac{\pi}{2} < \theta_q < \frac{3\,\pi}{2} \tag{4.3.15}$$

lie in the left side of the s-plane. If we multiply Eq. 4.3.8 by j, we rotate it by $+\pi/2$, achieving the condition expressed in Eq. 4.3.15 for the first n poles:

$$s_q = -\sin\pi\,\frac{1+2\,q}{2\,n} + j\cos\pi\,\frac{1+2\,q}{2\,n} \tag{4.3.16}$$

The parameter q is an integer from 0, 1, …, $(n-1)$. We would prefer to have the poles starting with 1 and ending with n. To do so, we introduce a new parameter $k = q+1$ and consequently $q = k-1$. With this, we arrive to the final expression for Butterworth poles:

$$\boxed{s_k = \sigma_k + j\,\omega_k = -\sin\pi\,\frac{2\,k-1}{2\,n} + j\cos\pi\,\frac{2\,k-1}{2\,n}} \tag{4.3.17}$$

where k is an integer from 1 to n. As shown in Fig. 4.3.2 (for n = 1–5), all these poles lie on a semicircle with the radius $r = 1$ in the left half of the s-plane:

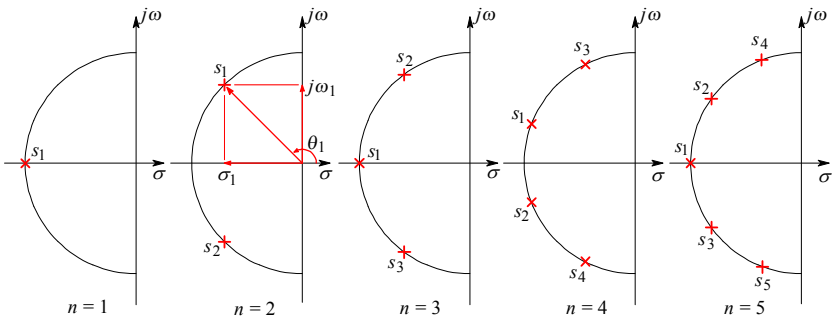


Fig. 4.3.2: Butterworth poles for the system order n = 1–5.

The numerical values of the poles for systems of order n = 1–10, together with the corresponding angle θ, are listed in Table 4.3.1. Obviously, if n is even the system has complex conjugate pole pairs only. If the n is odd, one of the poles is real, and in the normalized presentation its value is $s_1 = -\omega/\omega_H = -1$. In the non-normalized form, the value of the real pole is equal to $-\omega_H$. Since this is the radius of the circle on which all the poles lie, we can calculate the upper half power frequency also from any pole (for Butterworth poles only!):

$$\boxed{\omega_H = |s_i| = \sqrt{\sigma_i^2 + \omega_i^2}} \tag{4.3.18}$$

or, when one of the poles ($s_1 = \sigma_1$) is real:

$$\boxed{\omega_H = -\sigma_1} \tag{4.3.19}$$

4.3.1 Frequency Response

The normalized frequency response magnitude plots, expressed by Eq. 4.3.2, with $\omega_H = 1$ and for $n = 1$–10, are drawn in Fig. 4.3.3. Evidently the passband's flatness increases with increasing n.

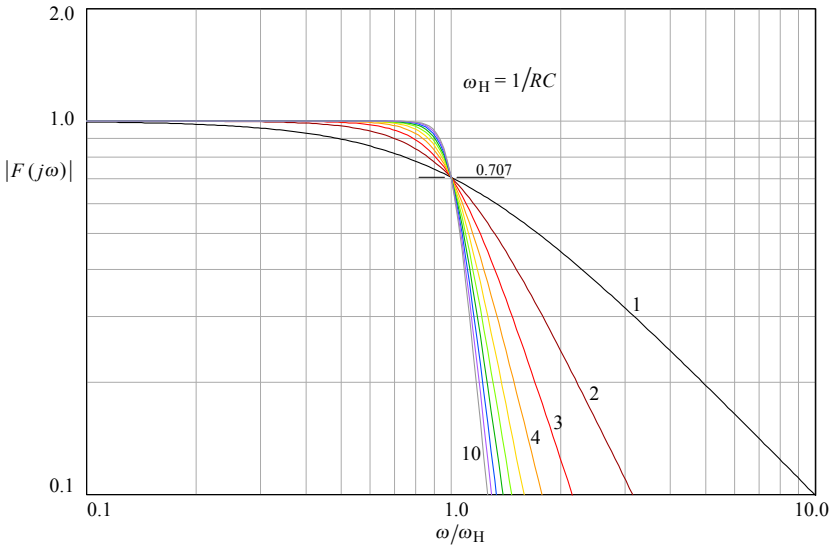


Fig. 4.3.3: Frequency response magnitude of n^{th}-order system with Butterworth poles, $n = 1$–10.

We can write the frequency response of an amplifier with Butterworth poles of order, say, $n = 5$, in three different ways. The general expression with poles is:

$$F_5(s) = \frac{(-1)^5\, s_1\, s_2\, s_3\, s_4\, s_5}{(s-s_1)(s-s_2)(s-s_3)(s-s_4)(s-s_5)} \tag{4.3.20}$$

with $s = j\omega/\omega_H$ and $s_i = \sigma_i + j\omega_i$ (the values of σ_i and ω_i are listed in Table 4.3.1). By multiplying all the expressions in parentheses, we obtain:

$$F_5(s) = \frac{a_0}{s^5 + a_4\, s^4 + a_3\, s^3 + a_2\, s^2 + a_1\, s + a_0} \tag{4.3.21}$$

where:

$$\begin{aligned}
a_4 &= -s_1 - s_2 - s_3 - s_4 - s_5 \\
a_3 &= s_1\, s_2 + s_1\, s_3 + s_1\, s_4 + s_1\, s_5 + s_2\, s_3 + s_2\, s_4 + s_2\, s_5 + s_3\, s_4 + s_3\, s_5 + s_4 s_5 \\
a_2 &= -s_1\, s_2\, s_3 - s_1\, s_2\, s_4 - s_1\, s_2\, s_5 - s_2\, s_3\, s_4 - s_2\, s_3\, s_5 - s_3\, s_4\, s_5 \\
a_1 &= s_2\, s_3\, s_4\, s_5 + s_1\, s_3\, s_4\, s_5 + s_1\, s_2\, s_4\, s_5 + s_1\, s_2\, s_3\, s_5 + s_1\, s_2\, s_3\, s_4 \\
a_0 &= -s_1\, s_2\, s_3\, s_4\, s_5
\end{aligned} \tag{4.3.22}$$

If we use the normalized poles with the numerical values listed in Table 4.3.1 to calculate the coefficients $a_0 \dots a_4$, we obtain:

$$F_5(s) = \frac{1}{s^5 + 3.2361\,s^4 + 5.2361\,s^3 + 5.2361\,s^2 + 3.2361\,s + 1} \tag{4.3.23}$$

For the magnitude only, by applying Eq. 4.3.2, we have:

$$|F_5(\omega)| = \frac{1}{\sqrt{1 + (\omega/\omega_H)^{10}}} \tag{4.3.24}$$

The reason why we took particular interest for the function with the normalized numerical values of the order $n = 5$ is that in Sec. 4.5 we shall compare it with the function having Bessel poles of the same order.

4.3.2 Phase response

The general expression for the phase angle is:

$$\varphi = \sum_{i=1}^{n} \arctan \frac{\dfrac{\omega}{\omega_H} + \omega_i}{\sigma_i} \tag{4.3.25}$$

For an odd number of poles the imaginary part of the middle pole $\omega_{n/2+1} = 0$. For the remaining poles or in the case of even n, we enter the complex conjugate pair components: $s_{i,n-i+1} = \sigma_i \pm j\,\omega_i$. The phase response plots are drawn in Fig. 4.3.4. By comparing it with Fig. 4.1.4 we note that Butterworth poles result in a much steeper phase slope near the system's cut off frequency at $\omega = \omega_H$ (which is even more evident in the envelope delay).

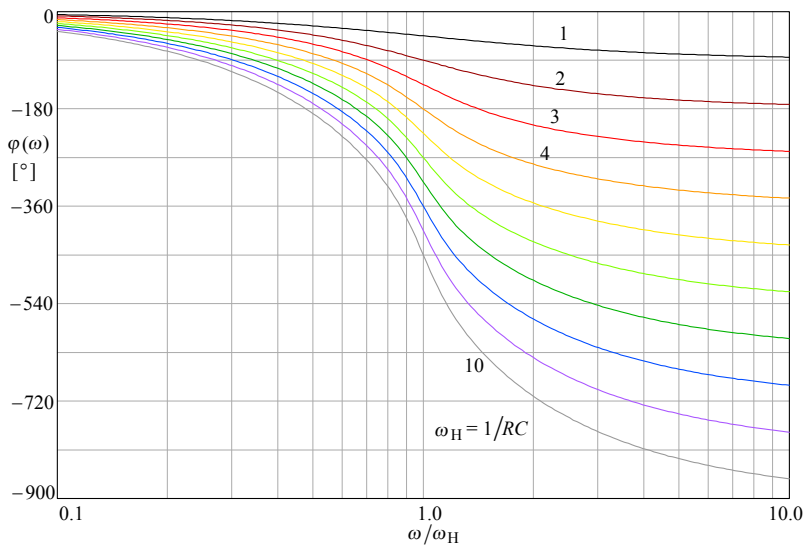


Fig. 4.3.4: Phase angle of n^{th}-order system with Butterworth poles, n = 1–10.

4.3.3 Envelope Delay

We obtain the expressions for envelope delay by making a frequency derivative of Eq. 4.3.25:

$$\tau_e\,\omega_H = \sum_{i=1}^{n} \frac{\sigma_i}{\sigma_i^2 + \left(\frac{\omega}{\omega_H} + \omega_i\right)^2} \tag{4.3.26}$$

The envelope delay plots for $n = 1$–10 are shown in Fig. 4.3.5. Owing to the ever steeper phase shift, the curves for $n > 1$ dip around the system cut off frequency. Those frequencies are delayed more than the rest of the spectrum, thus revealing the system resonance on transients. Therefore we expect that amplifiers with Butterworth poles will exhibit an increasing amount of ringing in the step response, a property not acceptable in pulse amplification.

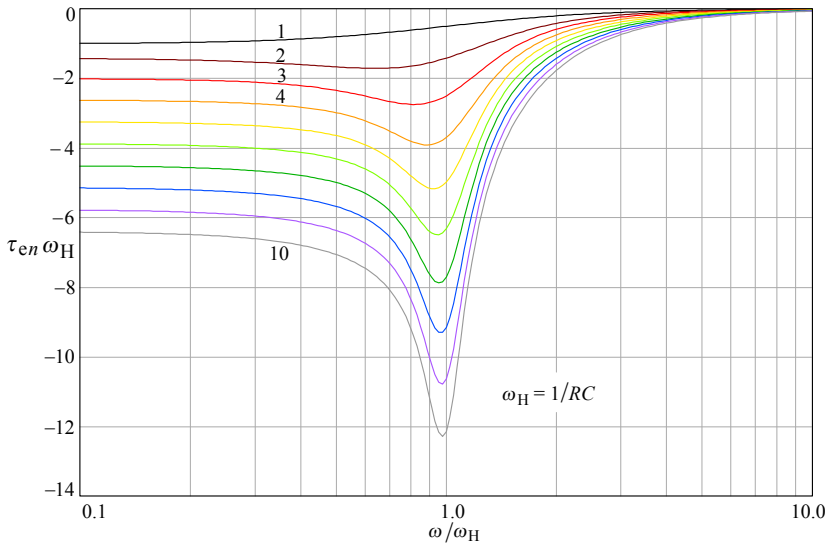


Fig. 4.3.5: Envelope delay of n^{th}-order system with Butterworth poles, $n = 1$–10.

4.3.4 Step Response

Since we have n non-repeating poles we start with the frequency function in the form which is suitable for the $\mathcal{L}^{-1}$ transform:

$$F(s) = \frac{(-1)^n s_1 s_2 \cdots s_n}{(s - s_1)(s - s_2) \cdots (s - s_n)} \tag{4.3.27}$$

We multiply this by the unit step operator $1/s$ and obtain:

$$G(s) = \frac{(-1)^n s_1 s_2 \cdots s_n}{s\,(s - s_1)(s - s_2)\,\cdots (s - s_n)} \tag{4.3.28}$$

To obtain the step response in the time domain we use the $\mathcal{L}^{-1}$ transform:

$$g(t) = \mathcal{L}^{-1}\{G(s)\} = \sum_{i=1}^{n} \text{res}_i \frac{(-1)^n s_1 s_2 \cdots s_n \, e^{st}}{s\,(s - s_1)(s - s_2) \cdots (s - s_n)} \qquad (4.3.29)$$

It would take too much space to list the complete analytical calculation for systems with 1 to 10 poles. Some examples can be found in the Appendix 2.3 (web only). Here we shall use the computer routines, which we develop and discuss in detail in Part 6. The plots for $n = 1$–10 are shown in Fig. 4.3.6.

These plots confirm our expectation that amplifiers with Butterworth poles are not suitable for pulse amplification. The main advantage of Butterworth poles is the flat frequency response (MFA) in the passband (evident from the plots in Fig. 4.3.3). Therefore for measuring sinusoidal signals in a wide range of frequencies, e.g., in an electronic voltmeter, Butterworth poles offer the best solution.

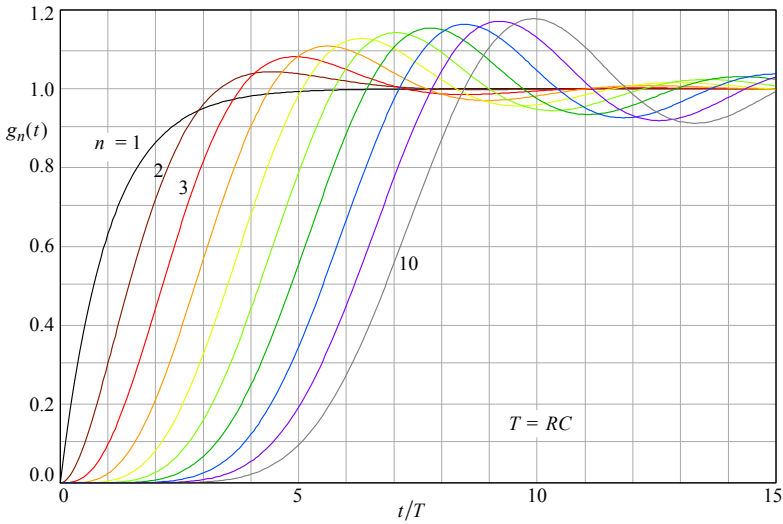


Fig. 4.3.6: Step response of n^{th}-order system with Butterworth poles, $n = 1$–10.

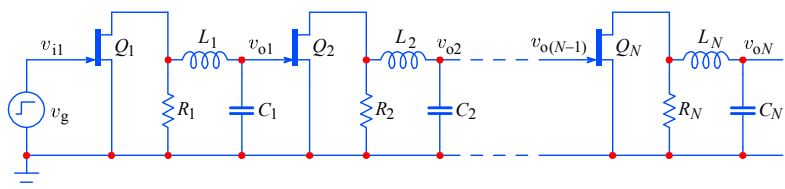


Fig. 4.3.7: An example of an amplifier with N shunt peaking stages. Since each stage has one pair of complex conjugate poles the number of stages is equal to one half of the number of poles, $N = n/2$. For odd-order systems one stage (usually the first one) is of a single pole configuration ($L_1 = 0$), and $N = 1 + (n - 1)/2$. Of course, instead of the shunt peaking, other peaking networks can be used.

Table 4.3.1: Butterworth Poles

Order n	σ [rad/s]	ω [rad/s]	θ [°]
1	− 1.0000	0.0000	180
2	− 0.7071	± 0.7071	180 ∓ 45.0000
3	− 1.0000	0.0000	180
	− 0.5000	± 0.8660	180 ∓ 60.0000
4	− 0.9239	± 0.3827	180 ∓ 21.5000
	− 0.3827	± 0.9239	180 ∓ 67.5000
5	− 1.0000	0.0000	180
	− 0.8090	± 0.5878	180 ∓ 36.0000
	− 0.3090	± 0.9511	180 ∓ 72.0000
6	− 0.9659	± 0.2588	180 ∓ 15.0000
	− 0.7071	± 0.7071	180 ∓ 45.0000
	− 0.2588	± 0.9659	180 ∓ 75.0000
7	− 1.0000	0.0000	180
	− 0.9010	± 0.4339	180 ∓ 25.7143
	− 0.6235	± 0.7818	180 ∓ 51.4286
	− 0.2225	± 0.9749	180 ∓ 77.1429
8	− 0.9808	± 0.1951	180 ∓ 11.2500
	− 0.8315	± 0.5556	180 ∓ 33.7500
	− 0.5556	± 0.8315	180 ∓ 56.2500
	− 0.1951	± 0.9808	180 ∓ 78.7500
9	− 1.0000	0.0000	180
	− 0.9397	± 0.3420	180 ∓ 20.0000
	− 0.7660	± 0.6428	180 ∓ 40.0000
	− 0.5000	± 0.8660	180 ∓ 60.0000
	− 0.1736	± 0.9848	180 ∓ 80.0000
10	− 0.9877	± 0.1564	180 ∓ 9.0000
	− 0.8910	± 0.4540	180 ∓ 27.0000
	− 0.7071	± 0.7071	180 ∓ 45.0000
	− 0.4540	± 0.8910	180 ∓ 63.0000
	− 0.1564	± 0.9877	180 ∓ 81.0000

Note: in this and all other tables we have arranged the poles in complex conjugate pairs, i.e., $s_{k,n-k+1} = \sigma_k \pm j\,\omega_k$, where $k = 1, 2, \ldots, n$. For n odd the first pole is real.

4.3.5 Ideal MFA Filter; Paley–Wiener Criterion

The following discussion will be given in an abridged form, since a complete derivation would detract us too much from the discussion of amplifiers. We are interested in designing an amplifier with the ideal frequency response, maximally flat in the passband and zero outside, as in Fig. 4.3.8 (shown also for negative frequencies), expressed as:

$$A(\omega) = \begin{cases} 1\big|_{|\omega/\omega_H|<1} \\ 0\big|_{|\omega/\omega_H|>1} \end{cases} \tag{4.3.30}$$

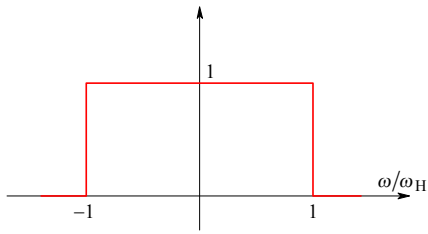


Fig. 4.3.8: Ideal MFA frequency response.

For the time being we assume that the function $A(\omega)$ is real, and consequently it has no phase shift. At the instant $t = 0$ we apply a unit step voltage to the input of the amplifier (multiply $A(s)$ by the unit step operator $1/s$). By applying the basic formula for the $\mathcal{L}^{-1}$ transform (Part 1, Eq. 1.4.4), the output function in the time domain is the integral of the $\sin(t)/t$ function [Ref. 4.2]:

$$g(t) = \frac{1}{\pi} \int_{-\infty}^{t} \frac{\sin t}{t}\, dt \tag{4.3.31}$$

The normalized plot of this integral is shown in Fig. 4.3.9. Here we have 50% of the signal amplitude at the instant $t = 0$. Also, there is some response for $t < 0$, before we applied any step voltage to the amplifier input, which is impossible. Any physically realizable amplifier would have some phase shift and an envelope delay, therefore the step response would be shifted rightwards from the origin. However, an infinite phase shift and delay would be needed in order to have no response for time $t < 0$.

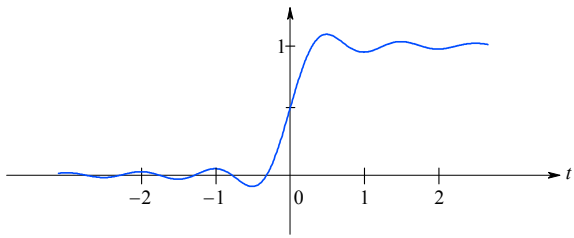


Fig. 4.3.9: Step response of a network having the ideal frequency response of Fig. 4.3.8.

What we would like to know is whether there is any phase response, linear or not, which the amplifier should have in order to suit Eq. 4.3.30 without any response for time $t < 0$. The answer is negative and it was proved by *R.E.A.C. Paley* and *N. Wiener*. Their criterion is given by an amplitude function [Ref. 4.2]:

$$\int_{-\infty}^{\infty} \frac{|\log A(\omega)|}{1+\omega^2}\, d\omega < \infty \tag{4.3.32}$$

Outside the range $-\omega_H < \omega < \omega_H$, $A(\omega) = 0$, as required by Eq. 4.3.30, but the magnitude in the numerator is infinite ($|\log 0| = \infty$); therefore the condition expressed by Eq. 4.3.32 is not met. Thus it is not possible to make an amplifier with a continuous infinite attenuation in a certain frequency band (it is, nevertheless, possible to have an infinite attenuation, but at **distinct** frequencies only). As we can derive from Eq. 4.3.32, the problem is not the steepness of the frequency response curve at ω_H in Fig. 4.3.8, but the requirement for an infinite attenuation everywhere outside the defined passband $-\omega_H < \omega < \omega_H$.

If we allow that outside the passband $A(\omega) = \epsilon$, no matter how small ϵ is, such a frequency response is possible to achieve. In this case the corresponding phase response must be [Ref. 4.2]:

$$\varphi(\omega) = \ln|\epsilon| \cdot \ln\left|\frac{1+\omega}{1-\omega}\right| \tag{4.3.33}$$

However, such an amplifier would still have a step response very similar to that in Fig. 4.3.9, except that it would be shifted rightwards and there would be no response for $t < 0$. This is because we have almost entirely (down to ϵ) and suddenly cut the signal spectrum above ω_H. The overshoot is approximately 9 %. We have met a similar situation in Part 1, Fig.1.2.7.a,b in connection with the square wave when we were discussing the *Gibbs*' phenomenon [Ref. 4.2].

Some readers may ask themselves why the step response overshoot of some systems with Butterworth poles in Fig. 4.3.6 exceeds 9%? The reason is the corresponding nonlinear phase response, resulting in a peak in the envelope delay, as shown in Fig. 4.3.5. This is a characteristic of not just the Butterworth poles, but also of any pole pattern, e.g., Chebyshev Type I and Elliptic (Cauer) systems, for which the magnitude and phase change more steeply near the cut off frequency.

We shall use Eq. 4.3.32 again when we shall discuss the possibility of obtaining the ideal Gaussian response of an amplifier.

4.4 Derivation of Bessel Poles for MFED Response

If we want to preserve the waveform shape, the amplifier must pass all frequency components with the same delay. From the requirement for a constant delay we can derive the system poles. The frequency response function having a constant delay T [Ref. 4.8, 4.9] is of the form:

$$F(s) = \mathrm{e}^{-sT} \tag{4.4.1}$$

Let us normalize this expression by choosing a unit delay, $T = 1$. It is possible to approximate e^{-s} by a rational function, where the denominator is a polynomial and all its roots (the poles of $F(s)$) lie in the left half of the s-plane. In this case the denominator is a so called *Hurwitz polynomial* [Ref. 4.10]. The approximation then fulfills the constant delay condition expressed by Eq. 4.4.1 to a certain accuracy only up to a certain frequency. The higher the polynomial degree, the higher is the accuracy.

We can write e^{-s} also by using the hyperbolic sine and cosine functions:

$$F(s) = \frac{1}{\sinh s + \cosh s} = \frac{\dfrac{1}{\sinh s}}{1 + \dfrac{\cosh s}{\sinh s}} \tag{4.4.2}$$

Both hyperbolic functions can be expressed with their corresponding series:

$$\cosh s = 1 + \frac{s^2}{2!} + \frac{s^4}{4!} + \frac{s^6}{6!} + \frac{s^8}{8!} + \cdots \tag{4.4.3}$$

$$\sinh s = s + \frac{s^3}{3!} + \frac{s^5}{5!} + \frac{s^7}{7!} + \frac{s^9}{9!} + \cdots \tag{4.4.4}$$

With these suppositions and using ‘long division’ we obtain:

$$\frac{\sinh s}{\cosh s} = \frac{1}{s} + \cfrac{1}{\dfrac{3}{s} + \cfrac{1}{\dfrac{5}{s} + \cfrac{1}{\dfrac{7}{s} + \cfrac{1}{\dfrac{9}{s} + \cdots}}}} \tag{4.4.5}$$

With a successive multiplication we can simplify this continuous fraction into a simple rational function. By truncating the fraction at $9/s$ we obtain the following approximation:

$$\frac{\sinh s}{\cosh s} \approx \frac{15\,s^4 + 420\,s^2 + 945}{s^5 + 105\,s^3 + 945\,s} \tag{4.4.6}$$

Now we put this and Eq. 4.4.4 into Eq. 4.4.2 and perform the suggested division by $\sinh s$. A normalized expression, where $F(s) = 1$ if $s = 0$ is obtained by multiplying the numerator by 945. With these operations we obtain:

$$F(s) = \mathrm{e}^{-s} \approx \frac{945}{s^5 + 15\,s^4 + 105\,s^3 + 420\,s^2 + 945\,s + 945} \tag{4.4.7}$$

The poles of this equation are the roots of the denominator:

$$\begin{aligned} s_{1,2} &= -3.3520 \pm j\,1.7427 \\ s_{3,4} &= -2.3247 \pm j\,3.5710 \\ s_5 &= -3.6467 \end{aligned}$$

A critical reader might ask why have we taken such a circuitous way to come to this result instead of deriving it straight from *McLaurin*'s series as:

$$\mathrm{e}^{-s} = \frac{1}{\mathrm{e}^{s}} \approx \frac{1}{1 + s + \frac{s^2}{2!} + \frac{s^3}{3!} + \frac{s^4}{4!} + \frac{s^5}{5!}} \tag{4.4.8}$$

In this case the roots are:

$$\begin{aligned} s_{1,2} &= -1.6496 \pm j\,1.6936 \\ s_{3,4} &= +0.2898 \pm j\,3.1283 \\ s_5 &= -2.1806 \end{aligned}$$

and the roots $s_{3,4}$ lie in the **right half** of the s-plane. Therefore the denominator of Eq. 4.4.8 is not a Hurwitz polynomial [Ref. 4.10] (a closer investigation would reveal that the denominator is not a Hurwitz polynomial if its degree exceeds 4, but even for low order systems it can be shown that the McLaurin's series results in an envelope delay which is far from being maximally flat). Thus Eq. 4.4.8 describes an unstable system or an unrealizable transfer function.

Let us return to Eq. 4.4.7, which we express in a general form:

$$\boxed{F(s) = \frac{a_0}{s^n + a_{n-1}s^{n-1} + \cdots + a_2\,s^2 + a_1\,s + a_0}} \tag{4.4.9}$$

where the numerical values for the coefficients can be calculated by the equation:

$$\boxed{a_i = \frac{(2\,n - i)!}{2^{n-i}\,i!\,(n-i)!}} \tag{4.4.10}$$

We can express the ratio of hyperbolic functions also as:

$$\frac{\cosh s}{\sinh s} = \frac{J_{-1/2}\,(-j\,s)}{j\,J_{1/2}\,(j\,s)} \tag{4.4.11}$$

where $J_{-1/2}\,(-j\,s)$ and $j\,J_{1/2}\,(j\,s)$ are the spherical Bessel functions [Ref. 4.10, 4.11]. Therefore we name the polynomials having their coefficients expressed by Eq. 4.4.10 *Bessel polynomials*. Their roots are the poles of Eq. 4.4.9 and we call them *Bessel poles*. We have listed the values of Bessel poles for polynomials of order $n = 1$–10 in Table 4.4.1, along with the corresponding pole angles θ_i.

We usually express Eq. 4.4.9 in another normalized form which is suitable for the $\mathcal{L}^{-1}$ transform:

$$F(s) = \frac{(-1)^n s_1\, s_2\, s_3 \cdots s_n}{(s - s_1)(s - s_2)(s - s_3) \cdots (s - s_n)} \tag{4.4.12}$$

where s_1, s_2, s_3, …, s_n are the poles of the function $F(s)$.

In Sec. 4.3 we saw that Butterworth poles lie in the left half of the s-plane on an origin centered unit circle. Since the denominator of Eq. 4.4.12 is also a Hurwitz polynomial [Ref. 4.10], all the poles of this equation must also lie in the left half of the s-plane. This is evident from Fig. 4.4.1, where the Bessel poles of the order $n = 1$–10 are drawn. However, Bessel poles lie on ellipses (not on circles). The characteristics of this family of ellipses is that they all have the near focus at the origin of the complex plane and the other focus on the positive real axis.

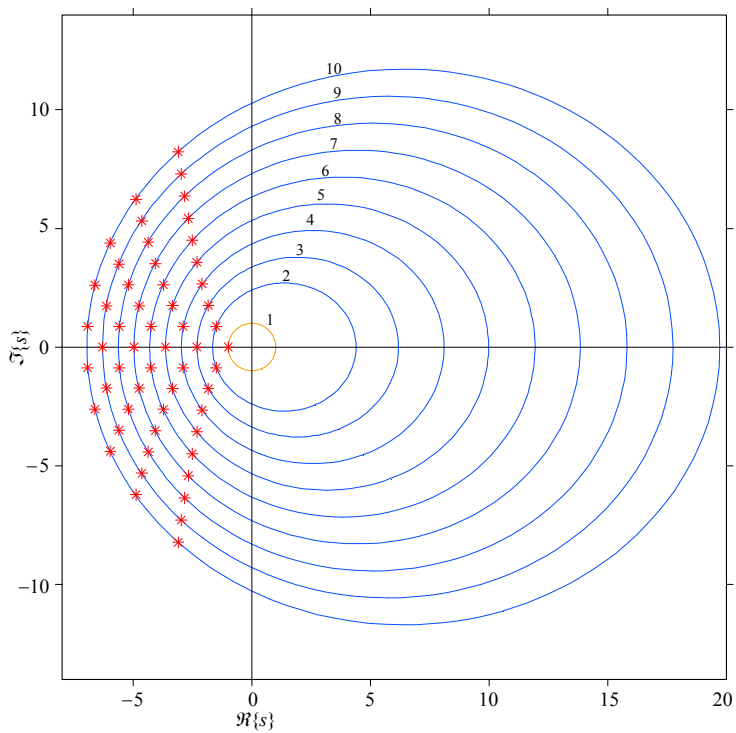


Fig. 4.4.1: Bessel poles for order $n = 1$–10. Bessel poles lie on a family of ellipses with one focus at the origin of the complex plane and the other focus on the positive real axis. The first-order pole is the same as for the Butterworth system and lies on the unit circle.

To implement these poles in a practical amplifier, the same circuit of Fig. 4.3.7, which we have used for Butterworth poles, can also be used for Bessel poles. Here too, instead of the shunt peaking shown, other, more efficient peaking networks can be used.

4.4.1 Frequency Response

A general normalized expression for the frequency response is the magnitude (absolute value) of Eq. 4.4.12:

$$|F(s)| = \left| \frac{s_1\, s_2\, s_3 \cdots s_n}{(s-s_1)(s-s_2)(s-s_3)\cdots(s-s_n)} \right|_{s\,=\,j\omega/\omega_\mathrm{h}} \tag{4.4.13}$$

where $\omega_\mathrm{h} = 1/RC$ is the non-peaking stage cut off frequency. If we put the numerical values for poles $s_\mathrm{i} = \sigma_\mathrm{i} \pm j\,\omega_\mathrm{i}$ and $s = j\,\omega/\omega_\mathrm{h}$ as suggested, then this formula obtains a form similar to Part 2, Eq. 2.6.10 (where we had 4 poles only).

The magnitude plots for $n = 1$–10 are shown in Fig. 4.4.2. By comparing this figure with Fig. 4.3.3, where the frequency response curves for Butterworth poles are displayed, we note an important difference: for Butterworth poles the upper half power frequency is always 1, regardless of the number of poles. In contrast, for Bessel poles the upper half power frequency increases with n.

The reason for the difference is that the derivation of n Butterworth poles was based on $\sqrt[2n]{-1}$ (for magnitude), whilst the Bessel poles were derived from the condition for a unit envelope delay. This difference prevents any direct comparison of the bandwidth extension and the rise time improvement between both kinds of poles. To be able to compare the two types of systems on a fair basis we must normalize the Bessel poles to the first-order cut off frequency. We do this by recursively multiplying the poles by a correction factor and calculate the cut off frequency, until a satisfactory approximation is reached (see Sec. 4.4.6). Also, a special set of Bessel poles is derived in Sec. 4.5, allowing us to interpolate between Bessel and Butterworth poles. The BESTAP algorithm (in Part 6) calculates the Bessel poles in any of the three options.

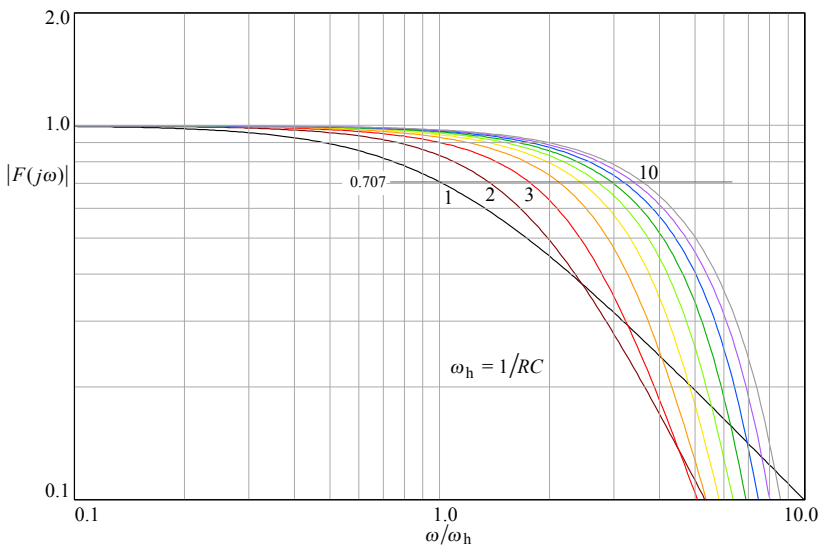


Fig. 4.4.2: Frequency-response magnitude of systems with Bessel poles for order $n = 1 \ldots 10$.

4.4.2 Upper Half Power Frequency

To find the cut off frequency and the bandwidth improvement offered by the Bessel poles an inversion formula should be developed from Eq. 4.4.13. This inversion formula would be different for each n, so there would not be a general solution. Instead, we shall use a computer program (see Part 6) to calculate the complete frequency response magnitude and find the cut off frequency from it. Calculated in this way, the bandwidth improvement factors $\eta_b = \omega_H/\omega_h$ for Bessel poles of the order $n = 1$–10 are listed in the following table:

Table 4.4.1: Relative Bandwidth Improvement with Bessel Poles

n	1	2	3	4	5	6	7	8	9	10
η_b	1.00	1.36	1.75	2.12	2.42	2.70	2.95	3.18	3.39	3.59

However, some special circuits can have a bandwidth improvement different from the one shown in the table for a corresponding order; e.g., T-coil circuits improve the bandwidth for $n = 2$ by exactly twice as much. We shall discuss this in Part 5, where we shall analyze a two-stage amplifier with 7 staggered Bessel poles, having one three-pole T-coil and one four-pole L+T circuit and obtain a total bandwidth improvement $\eta_b = 3.55$ for the complete amplifier.

4.4.3 Phase Response

The calculation is similar to the phase response for Butterworth poles; however there we had the normalized upper half power frequency $\omega_H = 1$, whilst for Bessel poles ω_H increases with the order n, so here we must use $\omega_h = 1/RC$ as the reference, where RC is the time constant of the non-peaking amplifier. Then for the phase response we simply repeat Eq. 4.3.25 and write ω_h (instead of ω_H):

$$\varphi(\omega) = \sum_{i=1}^{n} \arctan \frac{\dfrac{\omega}{\omega_h} + \omega_i}{\sigma_i} \tag{4.4.14}$$

Fig. 4.4.3 shows the phase plots of Eq. 4.4.14 for Bessel poles for the order $n = 1$–10 (owing to the cutoff frequency increasing with order n, the frequency scale had to be extended to see the asymptotic values at high frequencies).

So far we have used a logarithmic frequency scale for our phase response plots. However, by using a linear frequency scale, as in Fig. 4.4.4, the plots show that the phase response for Bessel poles is linear up to a certain frequency [Ref. 4.10], which increases with an increased order n.

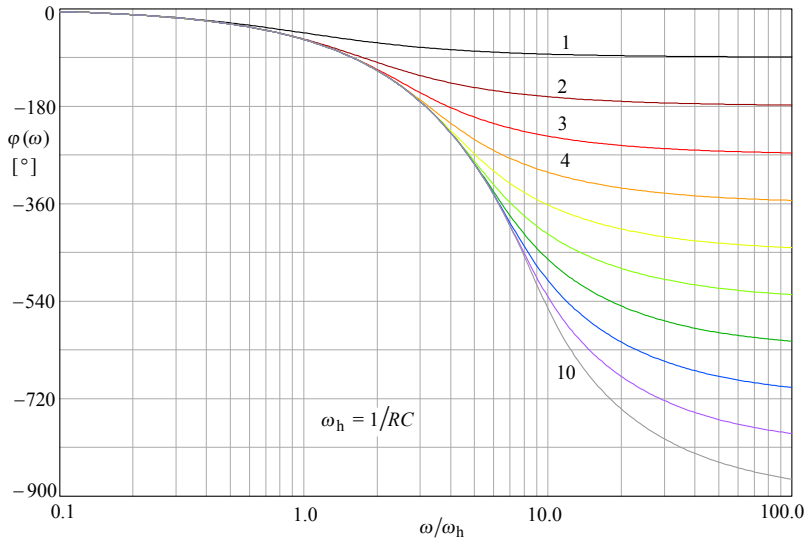


Fig. 4.4.3: Phase angle of the systems with Bessel poles of order $n = 1 \ldots 10$.

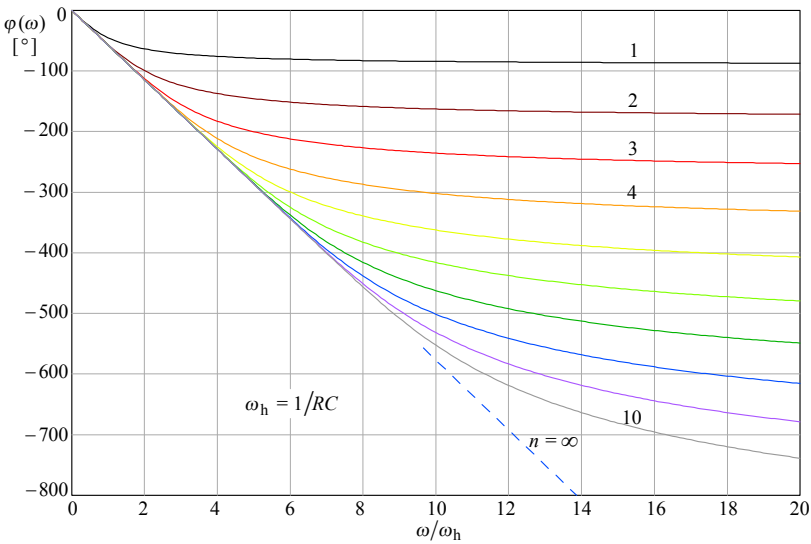


Fig. 4.4.4: Phase-angle as in Fig. 4.4.3, but in a linear frequency scale. Note the linear phase frequency-dependence extending from the origin to progressively higher frequencies.

4.4.4 Envelope-delay

Here, too, we take the corresponding formula from Butterworth poles and replace the frequency normalization ω_H by ω_h:

$$\tau_e\,\omega_h = \sum_{i=1}^{n} \frac{\sigma_i}{\sigma_i^2 + \left(\frac{\omega}{\omega_h} + \omega_i\right)^2} \tag{4.4.15}$$

The envelope delay plots are shown in Fig. 4.4.5. The delay is flat up to a certain frequency, which increases with increasing order n. This was our goal when we were deriving the Bessel poles, starting with Eq. 4.4.1. Therefore the name MFED (Maximally Flat Envelope Delay) is fully justified by this figure. This property is essential for pulse amplification. Because pulses contain a broad range of frequency components, all of them, (or in practice, the most significant ones, i.e., those which are not attenuated appreciably) should be subject to equal time delay when passing through the amplifier in order to preserve the pulse shape at the output as much as possible.

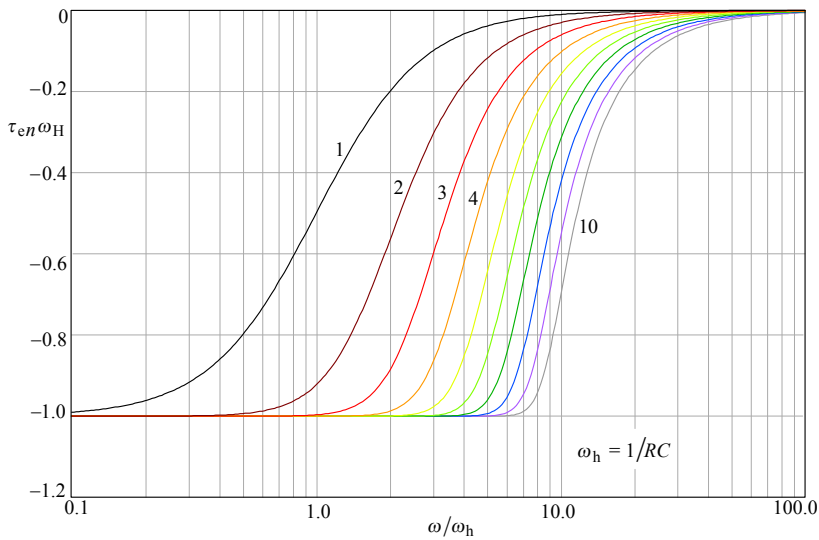


Fig. 4.4.5: Envelope delay of the systems with Bessel poles for order $n = 1$–10. Note the flat unit delay response increasing with system order. This figure demonstrates the fulfillment of the criterion from which we have started the derivation of MFED.

4.4.4 Step Response

We start with Eq. 4.4.12 and multiply it by the unit step operator $1/s$ to obtain:

$$G(s) = \frac{(-1)^n\, s_1\, s_2\, s_3 \cdots s_n}{s\,(s - s_1)(s - s_2)(s - s_3)\cdots(s - s_n)} \tag{4.4.16}$$

By applying the $\mathcal{L}^{-1}$-transform we obtain:

$$g(t) = \mathcal{L}^{-1}\{G(s)\} = \sum_{i=1}^{n} \text{res}_i \frac{(-1)^n\, s_1\, s_2\, s_3 \cdots s_n\, \mathrm{e}^{st}}{s\,(s-s_1)(s-s_2)(s-s_3)\cdots(s-s_n)} \qquad (4.4.17)$$

By inserting the numerical pole values from Table 4.4.3 for the systems of order $n = 1$–10 we can proceed in the same way as in the examples in Appendix 2.3, but it would take too much space. Instead, we shall use the routines developed in Part 6 to generate the plots of Fig. 4.4.6. This diagram is notably different from the step response plots of Butterworth poles in Fig. 4.3.6. Again, the reason is that for normalized Butterworth poles the upper half power frequency ω_H is always one, regardless of the order n, consequently the step response always has the same maximum slope, but a progressively larger delay. The Bessel poles, on the contrary, have progressively steeper slope, whilst the delay approaches unity. This is also reflected by the improvement in rise time, listed in Table 4.4.2. Of course, the improvement in rise time is even higher for peaking circuits using T-coils.

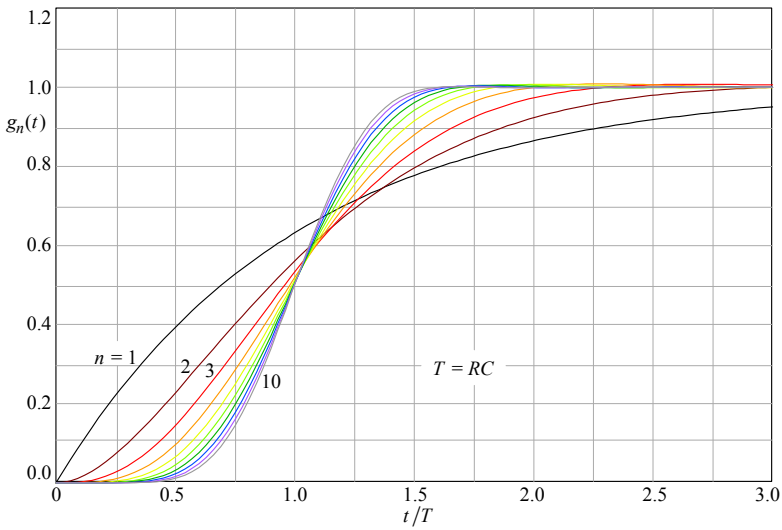


Fig. 4.4.6: Step response of systems with Bessel poles of order $n = 1$–10. Note the 50% amplitude delay approaching unity as the system order increases.

Table 4.4.2: Relative Rise time Improvement with Bessel Poles

n	1	2	3	4	5	6	7	8	9	10
η_r	1.00	1.38	1.75	2.09	2.39	2.68	2.93	3.18	3.41	3.60

Table 4.4.3: Bessel Poles (identical envelope delay)

Order n	σ [rad/s]	ω [rad/s]	θ [°]
1	− 1.0000	0.0000	180
2	− 1.5000	± 0.8660	180 ∓ 30.0000
3	− 2.3222	0.0000	180
	− 1.8389	± 1.7544	180 ∓ 43.6525
4	− 2.8962	± 0.8672	180 ∓ 16.6697
	− 2.1038	± 2.6574	180 ∓ 51.6325
5	− 3.6467	0.0000	180
	− 3.3520	± 1.7427	180 ∓ 27.4696
	− 2.3247	± 3.7510	180 ∓ 56.9366
6	− 4.2484	± 0.8675	180 ∓ 11.5411
	− 3.7357	± 2.6263	180 ∓ 35.1079
	− 2.5159	± 4.4927	180 ∓ 60.7508
7	− 4.9718	0.0000	180
	− 4.7583	± 1.7393	180 ∓ 20.0787
	− 4.0701	± 3.5172	180 ∓ 40.8316
	− 2.6857	± 5.4207	180 ∓ 63.6439
8	− 5.5879	± 0.8676	180 ∓ 8.8257
	− 5.2048	± 2.6162	180 ∓ 26.6861
	− 4.3683	± 4.4144	180 ∓ 45.3011
	− 2.8390	± 6.3539	180 ∓ 65.9245
9	− 6.2970	0.0000	180
	− 6.1294	± 1.7378	180 ∓ 15.8295
	− 5.6044	± 3.4982	180 ∓ 31.9715
	− 4.6384	± 5.3173	180 ∓ 48.9007
	− 2.9793	± 7.2915	180 ∓ 67.7753
10	− 6.9220	± 0.8677	180 ∓ 7.1447
	− 6.6153	± 2.6116	180 ∓ 21.5430
	− 5.9675	± 4.3849	180 ∓ 36.3085
	− 4.8862	± 6.2250	180 ∓ 51.8703
	− 3.1089	± 8.2327	180 ∓ 69.3119

Important:

From Fig 4.4.2 and 4.4.6 one could make a false conclusion that the upper half power frequency increases and the rise time decreases if more equal amplifier stages are cascaded. This is not true, because all the parameters of systems having Bessel poles are defined with respect to the **single stage non-peaking amplifier**, where $\omega_h = 1/RC$.

In the case of a system with n Bessel poles this would mean chopping the stray capacitance of a single amplifying stage into smaller capacitances and separating them by coils to create n poles.

Unfortunately there is a limit in practice because each individual amplifier stage input sees two capacitances: the output capacitance of the previous stage and its own input capacitance. Therefore, in a multi-stage amplifier an individual stage can have four poles at most (either Bessel, Butterworth, or of any other pole family). However, a single-stage amplifier can have up to 8 poles (if we apply a T-coil peaking to both input and output).

If we use more than one stage, we can assign a small group of staggered poles from the n^{th}-group (from either Table 4.3.1 or Table 4.4.3) to each stage, so that the system as a whole has the poles as specified by the n^{th}-group chosen. Then **no stage by itself will be optimized, but the amplifier as a whole will be**. More details of this technique are given in Sec. 4.6 and some examples can be found in Part 5 and Part 7.

4.4.5 Ideal Gaussian Frequency Response

Suppose that we have succeeded making an amplifier with zero phase shift and an ideal Gaussian frequency response:

$$G(\omega) = \mathrm{e}^{-\omega^2} \tag{4.4.18}$$

the plot of which is shown in Fig. 4.4.7 for both positive and negative frequencies (and, to acquire a feeling for Bessel systems, compared to the magnitude of a 5th-order system with modified Bessel poles).

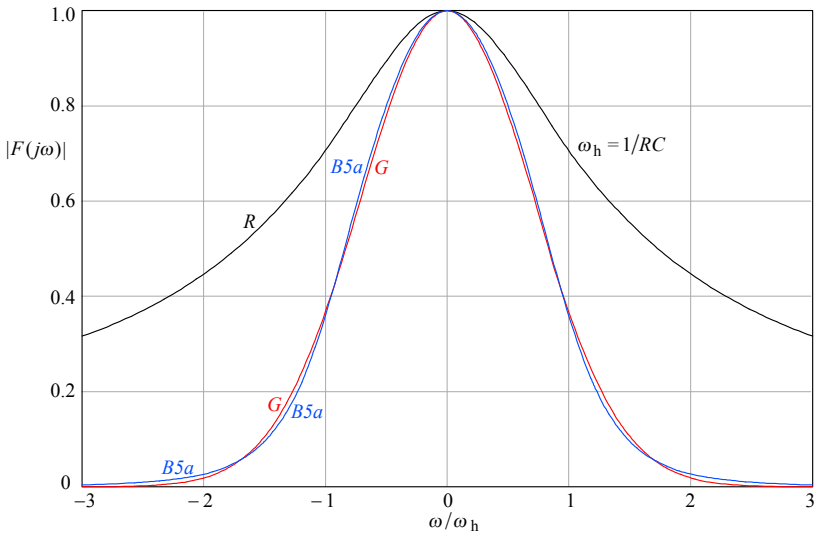


Fig. 4.4.7: Ideal Gaussian (MFED) frequency response G (real only, with no phase shift), compared to the magnitude of a 5th-order modified Bessel system $B5a$ (identical cutoff asymptote, Table 4.5.1). The frequency scale is two-sided, linear, and normalized to $\omega_\mathrm{h} = 1/RC$ of the first-order system, which is shown as the reference R.

By examining Eq. 4.4.9 and Eq. 4.4.18 we come to the conclusion that it is possible to **approximate** the Gaussian response with any desired accuracy up to a certain frequency. At higher frequencies, the Gaussian response falls faster than the approximated response. This is brought into evidence in Fig. 4.4.8 where the same responses are plotted in log-log scale.

By applying a unit step at the instant $t = 0$ to the input of the hypothetical amplifier having a Gaussian frequency response the resulting step response is equal to the so called *error-function*, which is defined as the time integral of the exponential function of time squared [Ref. 4.2]:

$$g_\mathrm{G}(t) = \mathrm{erf}(t) = \frac{1}{2\sqrt{\pi}} \int_{-\infty}^{t_1} \mathrm{e}^{-t^2/4}\, dt \tag{4.4.19}$$

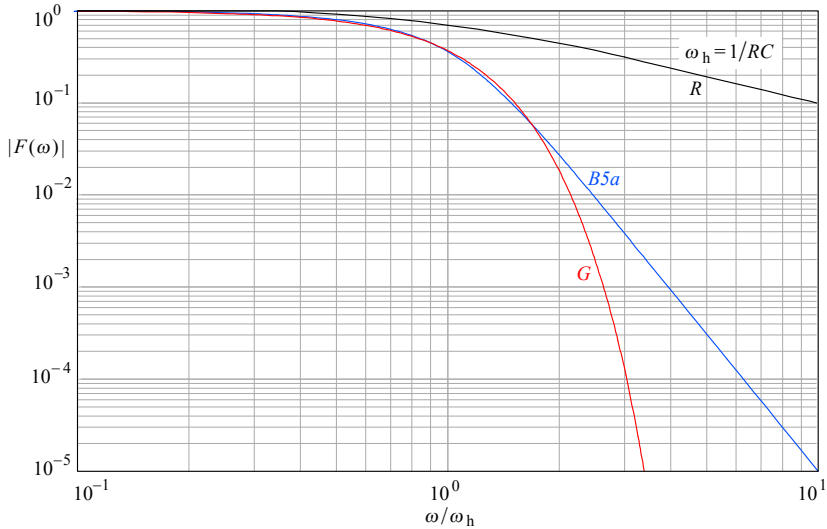


Fig. 4.4.8: Frequency response in log-log scale brings into evidence how the ideal Gaussian response G decreases much more steeply with frequency than the 5th-order Bessel response $B5a$. The Bessel system would have to be of infinitely high order to match the Gaussian response.

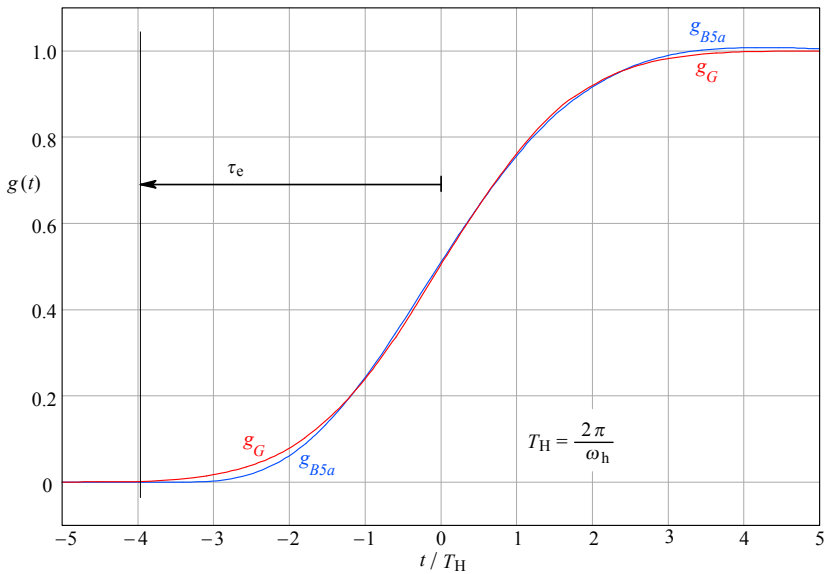


Fig. 4.4.9: Step response of a hypothetical system, g_G, having the ideal Gaussian frequency response with no phase shift, as the one in Fig. 4.4.7 and 4.4.8. Compare it with the step response of a 5th-order Bessel system, g_{B5a}, with modified Bessel poles, Table 4.5.1, and envelope delay compensated ($\tau_e = -3.94\,T_H$) for minimal difference in the half amplitude region.

The plot of Eq. 4.4.19 calculated by the *Simpson method* is shown in Fig. 4.4.9. The step response is symmetrical, without any overshoot. However, here too, we have a response for $t < 0$ as it was in the ideal MFA amplifier.

If our hypothetical amplifier were to have any linear phase delay the curve g_G in Fig. 4.4.9 would be shifted rightwards from the origin, but an infinite phase shift would be required in order to have no response for time $t < 0$ (the same as for Fig. 4.3.8).

By looking back at Eq. 4.4.2–7, we realize that we would need an infinite number of terms in the denominator (= an infinite number of poles) in order to justify an ' = ' sign instead of an approximation (≈). This would mean an infinite number of system components and amplifying stages, and therefore the conclusion is that we can not make an amplifier with an ideal Gaussian response (but we can come very close).

A proof, based on the *Paley–Wiener Criterion*, can be carried out in the following way: if we compare the step response in Fig. 4.4.9 with the step response of a non-peaking multi-stage amplifier in Fig. 4.1.7, for $n = 10$, there is a great similarity. Therefore we can ask ourselves if a Gaussian response could be achieved by increasing the number of stages to some arbitrarily large number ($n \to \infty$). By doing so, the phase response diverges when $n \to \infty$ and it becomes infinite if $\omega \to \infty$. Therefore for both reasons (infinite number of stages and divergent phase response) it is not possible to make an amplifier with an ideal Gaussian response.

4.4.6 Bessel Poles Normalized to Identical Cutoff Frequency

Because the Bessel poles are derived from the requirement for an identical envelope delay there is no simple way of renormalizing them back to the same cut off frequency. However, such a renormalization would be very useful, not only for comparing the systems with different pole families and equal order, but also for comparing systems of different order within the Bessel family itself.

What is difficult to do analytically is often easily done numerically, especially if the actual number crunching is executed by a machine. The normalization procedure goes by taking the original Bessel poles and finding the system magnitude by Eq. 4.4.13 at the unit frequency ($\omega/\omega_h = 1$). We obtain an attenuation value, say, $|F(1)| = a$ and we want $|F(1)|$ to be $1/\sqrt{2}$. The ratio $q = 1/a\sqrt{2}$ is the correction factor by which we multiply all the poles and insert the new poles again into Eq. 4.4.13. We keep repeating this procedure until $|q - 1/\sqrt{2}| < \varepsilon$, with ε being an arbitrarily small error; for practical purposes, a value of $\varepsilon = 0.001$ is adequate. In the algorithm presented in Part 6, this tolerance is reached in only 6 to 9 iterations, depending on system order.

The following graphs were made using the computer algorithms presented in Part 6 and show the performance of cut off frequency normalized Bessel systems of order $n = 1$–10, as in the previous figures.

Fig. 4.4.10 shows the frequency response magnitude; the plots for $n = 5$–9 are missing, since the difference is too small to identify them on such a vertical scale (the difference in high frequency attenuation becomes significant with higher magnitude resolution, say, down to 0.001 or less). Fig. 4.4.11 shows the phase, Fig. 4.4.12 the envelope delay and Fig. 4.4.13 shows the step response. Finally, in Table 4.4.4 we report

the values of Bessel poles and their respective angles for systems with equal cut off frequency.

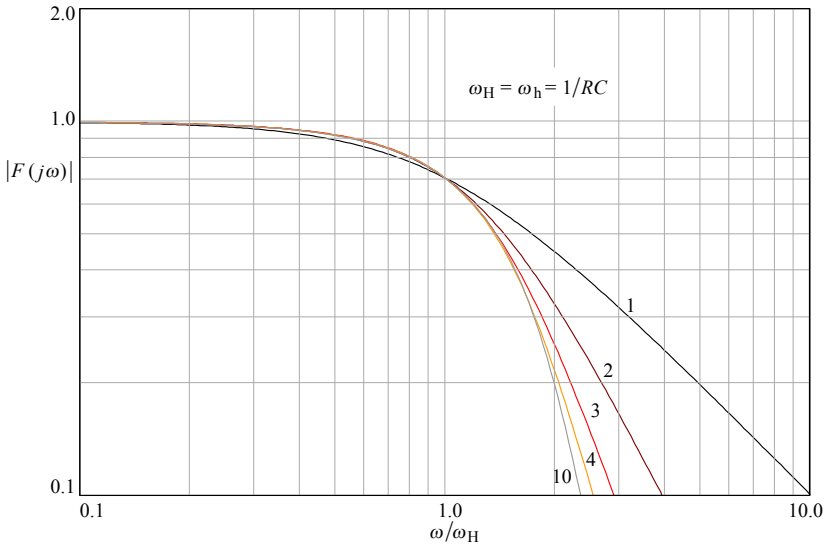


Fig. 4.4.10: Frequency response magnitude of systems with normalized Bessel poles of order $n = 1, 2, 3, 4,$ and 10. Note the nearly identical passband response — this is the reason why we can approximate the oscilloscope (multi-stage) amplifier rise time from the cut off frequency, using the relation for the first-order system: $\tau_r = 0.35/f_T$.

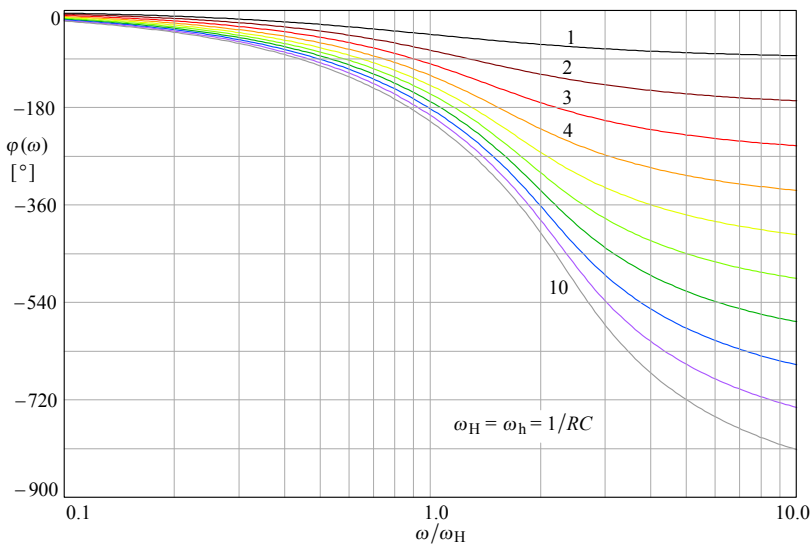


Fig. 4.4.11: Phase angle of systems with normalized Bessel poles of order $n = 1$–10.

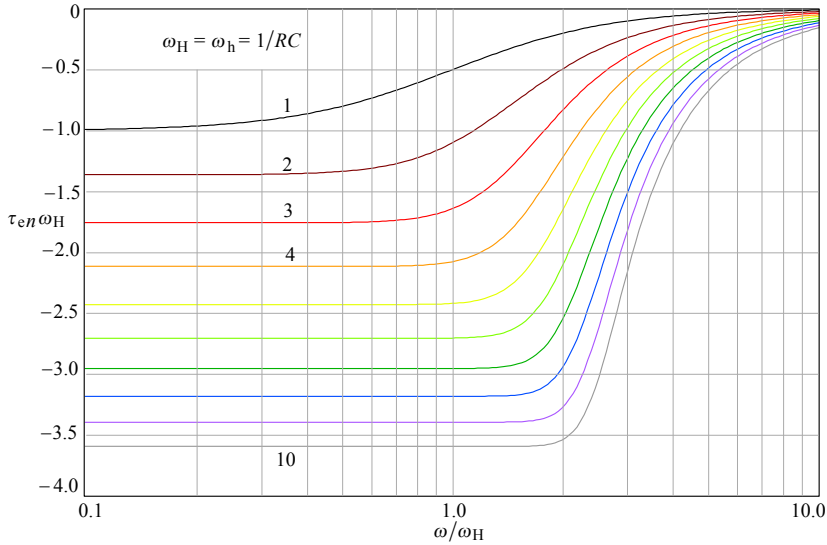


Fig. 4.4.12: Envelope delay of systems with normalized Bessel poles of the order $n = 1$–10. Although the bandwidth is the same, the delay flatness extends progressively with system order, already reaching beyond the system cut off frequency for $n = 5$.

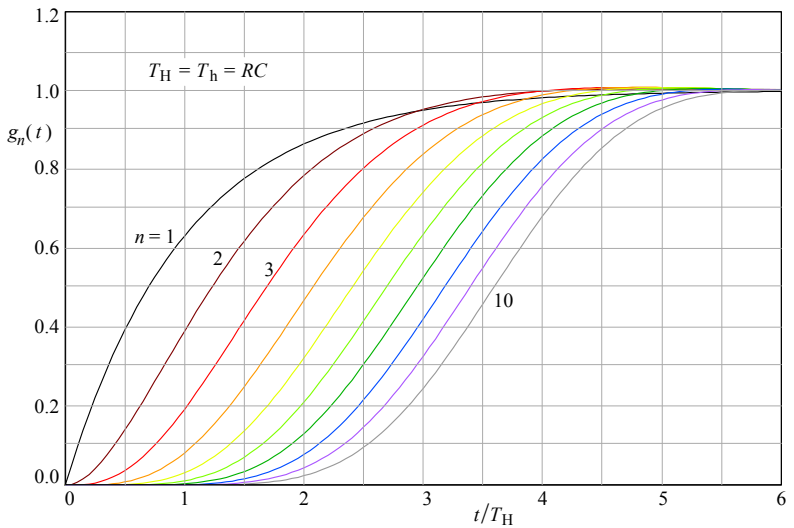


Fig. 4.4.13: Step response of systems with normalized Bessel poles of order $n = 1$–10. Note the half amplitude slope being almost equal for all systems, indicating an equal system cut off frequency.

Table 4.4.4: Bessel Poles (equal cut off frequency)

Order n	σ [rad/s]	ω [rad/s]	θ [°]
1	− 1.0000	0.0000	180
2	− 1.1017	± 0.6360	180 ∓ 30.0000
3	− 1.3227	0.0000	180
	− 1.0475	± 0.9993	180 ∓ 43.6525
4	− 1.3701	± 0.4103	180 ∓ 16.6697
	− 0.9952	± 1.2571	180 ∓ 51.6325
5	− 1.5024	0.0000	180
	− 1.3810	± 0.7180	180 ∓ 27.4696
	− 0.9578	± 1.4713	180 ∓ 56.9366
6	− 1.5716	± 0.3209	180 ∓ 11.5411
	− 1.3819	± 0.9715	180 ∓ 35.1079
	− 0.9307	± 1.6619	180 ∓ 60.7508
7	− 1.6845	0.0000	180
	− 1.6122	± 0.5893	180 ∓ 20.0787
	− 1.3790	± 1.1917	180 ∓ 40.8316
	− 0.9099	± 1.8366	180 ∓ 63.6439
8	− 1.7575	± 0.2729	180 ∓ 8.8257
	− 1.6370	± 0.8228	180 ∓ 26.6861
	− 1.3739	± 1.3884	180 ∓ 45.3011
	− 0.8929	± 1.9984	180 ∓ 65.9245
9	− 1.8567	0.0000	180
	− 1.8072	± 0.5124	180 ∓ 15.8295
	− 1.6525	± 1.0314	180 ∓ 31.9715
	− 1.3676	± 1.5678	180 ∓ 48.9007
	− 0.8784	± 2.1499	180 ∓ 67.7753
10	− 1.9277	± 0.2416	180 ∓ 7.1447
	− 1.8423	± 0.7273	180 ∓ 21.5430
	− 1.6619	± 1.2212	180 ∓ 36.3085
	− 1.3608	± 1.7336	180 ∓ 51.8703
	− 0.8658	± 2.2927	180 ∓ 69.3119

4.5 Pole Interpolation

Sometimes we desire to design an amplifier, or just a single stage, with a performance which is somewhere between the Butterworth and Bessel response. We shall derive the corresponding poles by the pole interpolation procedure which was described by *Y. Peless* and *T. Murakami* [Ref. 4.14].

4.5.1 Derivation of Modified Bessel poles

In order to be able to interpolate between Butterworth and Bessel poles, the later must be modified so that, for both systems of equal order, the frequency response magnitude will have the same asymptotes in both passband and stopband. This is achieved if the product of all the poles is equal to one, as in Butterworth systems:

$$(-1)^n \prod_{k=1}^{n} s_k = 1 \tag{4.5.1}$$

A general expression for the frequency response normalized in amplitude is:

$$F(s) = \frac{a_0}{s^n + a_{n-1}\, s^{n-1} + \cdots + a_2\, s^2 + a_1\, s + a_0} \tag{4.5.2}$$

where:

$$a_0 = (-1)^n \prod_{k=1}^{n} s_k \tag{4.5.3}$$

If we divide both the numerator and the denominator by a_0 we obtain:

$$F(s) = \frac{1}{\frac{1}{a_0}\, s^n + \frac{a_{n-1}}{a_o}\, s^{n-1} + \cdots + \frac{a_2}{a_0}\, s^2 + \frac{a_1}{a_0}\, s + 1} \tag{4.5.4}$$

Next we introduce another variable $\underline{s}$ such that:

$$\underline{s}^n = \frac{s^n}{a_0} \qquad \text{or} \qquad \underline{s} = \frac{s}{a_0^{1/n}} \tag{4.5.5}$$

Then Eq. 4.5.4 can be written as:

$$F(s) = \frac{1}{\underline{s}^n + b_{n-1}\, \underline{s}^{n-1} + \cdots + b_2\, \underline{s}^2 + b_1\, \underline{s} + 1} \tag{4.5.6}$$

where the coefficients b_i are:

$$\begin{aligned} b_{n-1} &= \frac{a_{n-1}}{a_0^{1/n}} \\ b_{n-2} &= \frac{a_{n-2}}{a_0^{2/n}} \\ &\cdots \\ b_1 &= \frac{a_1}{a_0^{1-1/n}} \end{aligned} \tag{4.5.7}$$

The coefficients a_k for the Bessel polynomials are calculated by Eq. 4.4.10. Then the coefficients b_k are those of the modified Bessel polynomial, from which we can calculate the modified Bessel poles of $F(s)$ for the order $n = 1$–10. These poles are listed together with the corresponding pole radii r and pole angles θ in Table 4.5.1.

4.5.2 Pole Interpolation Procedure

At the time of the German mathematician *Friedrich Wilhelm Bessel* (1784–1846), there were no electronic filters and no wideband amplifiers, to which the roots of his polynomials could be applied. *W.A. Thomson* [Ref. 4.9] was the first to use them and he also derived the expressions required for MFED network synthesis. Therefore some engineers use the name *Thomson poles* or, perhaps more correctly, *Bessel–Thomson poles*.

In the following discussion we shall interpolate between Butterworth and the modified Bessel poles. If we were to label the poles by initials only, a confusion would result in the graphs and formulae. Therefore, to label the modified Bessel poles, we shall use the subscript 'T' in honor of W.A. Thomson.

The procedure of pole interpolation can be explained with the aid of Fig. 4.5.1.

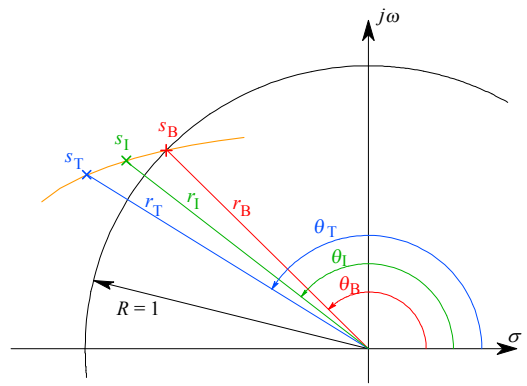


Fig. 4.5.1: Pole interpolation procedure. Butterworth (index B) and Bessel poles (index T) are expressed in polar coordinates, $s(r,\theta)$. The trajectory going through both poles is the interpolation path required to obtain the transitional pole s_1.

We first express the poles in polar coordinates with the well known conversion:

$$r_k = \sqrt{\sigma_k^2 + \omega_k^2} \tag{4.5.8}$$

and

$$\theta_k = \pi + \arctan\frac{\omega_k}{\sigma_k} \tag{4.5.9}$$

Here the π radians added are required because the arctangent function repeats with a period of π radians, so it does not distinguish between the poles in quadrant III form those in I, and the same is true for quadrants IV and II.

By using the polar coordinates a pole s_k is expressed as:

$$s_k = \sigma_k + j\omega_k = r_k\, \mathrm{e}^{j\theta_k} \tag{4.5.10}$$

In Eq. 4.5.3 the coefficient a_0 is equal to the product of all poles. Because we have divided the polynomial coefficients a_k by a_0 to obtain the coefficients b_k we have effectively normalized the product of all poles to one:

$$\left| \prod_{k=1}^{n} s_k \right| = 1 \tag{4.5.11}$$

and this is now true for both Butterworth and the modified Bessel poles. Therefore we can assume that there exists a trajectory going through the k^{th} Butterworth pole $s_{\text{B}k}$ and the k^{th} Bessel pole $s_{\text{T}k}$ and each point on this trajectory can represent a pole $s_{\text{I}k}$ which can be expressed as:

$$s_{\text{I}k} = r_{\text{I}k}\, \mathrm{e}^{j\theta_{\text{I}k}} \tag{4.5.12}$$

such that the absolute product of all interpolated poles s_{I} is kept equal to one. Then:

$$r_{\text{I}k} = r_{\text{T}k}^{m} \tag{4.5.13}$$

and:

$$\theta_{\text{I}k} = \theta_{\text{B}k} + m\,(\theta_{\text{T}k} - \theta_{\text{B}k}) \tag{4.5.14}$$

The parameter m can have any value between 0 and 1.

If $m = 0$ we have the Butterworth poles and if $m = 1$ we have the modified Bessel poles. By using, say, $m = 0.5$, the characteristics of such a network would be just half way between the Butterworth and modified Bessel poles.

It is obvious that we need to calculate only one half of the poles, say, those with the positive imaginary value, $\sigma_k + j\omega_k$, since the complex conjugate poles $\sigma_k - j\omega_k$ have the same magnitude, only the sign of their imaginary component is negative. In the cases with odd n the interpolated real pole remains on the real axis ($\theta_{\text{I}k} = \pi$) between the two real poles belonging to the Buttreworth and the modified Bessel system.

Table 4.5.1: Modified Bessel Poles
(with HF asymptote identical as Butterworth)

Order n	σ [rad/s]	ω [rad/s]	r	θ [°]
1	− 1.0000	0.0000	1.0000	180
2	− 0.8660	± 0.5000	1.0000	180 ∓ 30.0000
3	− 0.9416	0.0000	0.9416	180
	− 0.7456	± 0.7114	1.0305	180 ∓ 43.6525
4	− 0.9048	± 0.2709	0.9444	180 ∓ 16.6697
	− 0.6572	± 0.8302	1.0588	180 ∓ 51.6325
5	− 0.9264	0.0000	0.9246	180
	− 0.8516	± 0.4427	0.9598	180 ∓ 27.4696
	− 0.5906	± 0.9072	1.0825	180 ∓ 56.9366
6	− 0.9094	± 0.1857	0.9282	180 ∓ 11.5411
	− 0.7997	± 0.5622	0.9775	180 ∓ 35.1079
	− 0.5386	± 0.9617	1.1022	180 ∓ 60.7508
7	− 0.9195	0.0000	0.9195	180
	− 0.8800	± 0.3217	0.9369	180 ∓ 20.0787
	− 0.7527	± 0.6505	0.9948	180 ∓ 40.8316
	− 0.4967	± 1.0025	1.1188	180 ∓ 63.6439
8	− 0.9097	± 0.1412	0.9206	180 ∓ 8.8257
	− 0.8473	± 0.4259	0.9483	180 ∓ 26.6861
	− 0.7111	± 0.7187	1.0110	180 ∓ 45.3011
	− 0.4622	± 1.0344	1.1329	180 ∓ 65.9245
9	− 0.9155	0.0000	0.9155	180
	− 0.8911	± 0.2527	0.9262	180 ∓ 15.8295
	− 0.8148	± 0.5086	0.9605	180 ∓ 31.9715
	− 0.6744	± 0.7731	1.0259	180 ∓ 48.9007
	− 0.4331	± 1.0601	1.1451	180 ∓ 67.7753
10	− 0.9091	± 0.1140	0.9162	180 ∓ 7.1447
	− 0.8688	± 0.3430	0.9341	180 ∓ 21.5430
	− 0.7838	± 0.5759	0.9726	180 ∓ 36.3085
	− 0.6418	± 0.8176	1.0394	180 ∓ 51.8703
	− 0.4083	± 1.0813	1.1558	180 ∓ 69.3119

4.5.3 A Practical Example of Pole Interpolation

Let us calculate the frequency, phase, time delay and step response of a network with three poles. Three poles are just enough to demonstrate the procedure of pole interpolation completely. Let us select $m = 0.5$. From Table 4.3.1 we find the following values for Butterworth poles of order $n = 3$:

$$\begin{array}{lll} s_{1B}: & r_{1B} = 1 & \theta_{1B} = \ 180° \\ s_{2B}: & r_{2B} = 1 & \theta_{2B} = \ +120° \\ s_{3B}: & r_{3B} = 1 & \theta_{3B} = \ -120° \end{array} \tag{4.5.15}$$

and in Table 4.5.1, order $n = 3$, we find these values for modified Bessel poles:

$$\begin{array}{lll} s_{1T}: & r_{1T} = 0.9416 & \theta_{1T} = 180° \\ s_{2T}: & r_{2T} = 1.0305 & \theta_{2T} = \ +136.35° \\ s_{3T}: & r_{3T} = 1.0305 & \theta_{3T} = \ -136.35° \end{array} \tag{4.5.16}$$

Now we interpolate the radii and the pole angles:

$$\begin{aligned} r_1 &= {r_{1T}}^m = 0.9416^{0.5} = 0.9704 \\ r_{2,3} &= {r_{1T}}^m = 1.0305^{0.5} = 1.0151 \\ \theta_{2,3} &= \theta_B + m\,(\theta_T - \theta_B) = 120° + 0.5\,(136.35° - 120°) = 128.175° \end{aligned} \tag{4.5.17}$$

With these values we calculate the real and imaginary components of transitional Butterworth–Bessel poles (TBT):

$$\begin{aligned} s_1 &= \ -0.9704 = \sigma_1 \\ s_{2,3} &= \ -r_2 \cos\theta_2 \pm j\, r_2 \sin\theta_2 \\ &= \ -1.0151 \cos 128.175° \pm j\, 1.0151 \sin 128.175° \\ &= \ -0.6274 \pm j\, 0.7980 = \sigma_2 \pm j\,\omega_2 \end{aligned} \tag{4.5.18}$$

The relation for the normalized frequency response magnitude is:

$$\left|F(\omega)\right| = \frac{1}{\sqrt{(\sigma_1 \ +\omega)^2\left[\sigma_2^2 + (\omega+\omega_2)^2\right]\left[\sigma_2^2 + (\omega-\omega_2)^2\right]}} \tag{4.5.19}$$

$$= \frac{1}{\sqrt{(0.9704+\omega)^2\left[0.6274^2 + (\omega+0.7980)^2\right]\left[0.6274^2 + (\omega-0.7980)^2\right]}}$$

The magnitude plot is shown in Fig. 4.5.2 (TBT); for comparison, the magnitude plots with Butterworth poles and modified Bessel poles are also drawn.

The normalized phase response is calculated as:

$$\begin{aligned} \varphi &= \arctan\frac{\omega}{\sigma_1} + \arctan\frac{\omega+\omega_2}{\sigma_2} + \arctan\frac{\omega-\omega_2}{\sigma_2} = \\ &= \arctan\frac{\omega}{-0.9704} + \arctan\frac{\omega+0.7981}{-0.6275} + \arctan\frac{\omega-0.7981}{-0.6275} \end{aligned} \tag{4.5.20}$$

In Fig. 4.5.3 the phase plot of the transitional (TBT) system, together with the plots for Butterworth and modified Bessel systems are drawn.

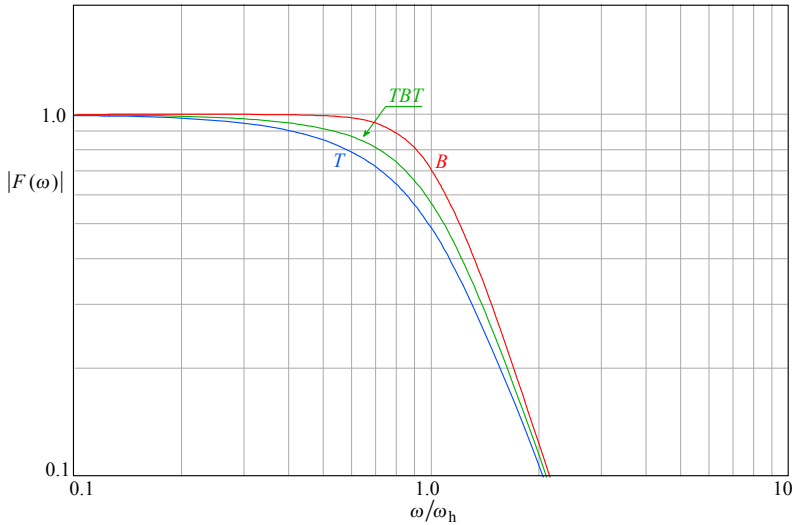


Fig. 4.5.2: Frequency response's magnitude of the Transitional Bessel–Butterworth three-pole system (TBT), along with the modified Bessel (T) and Butterworth (B) responses.

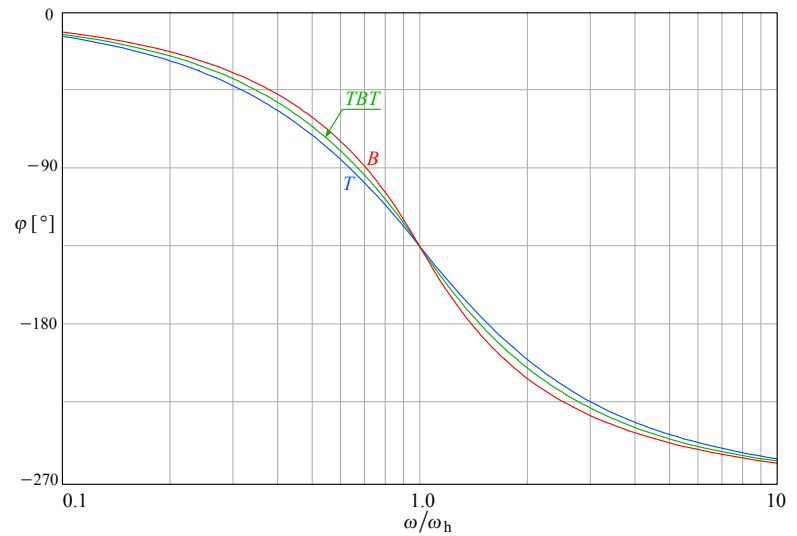


Fig. 4.5.3: Phase angle of the Transitional Bessel–Butterworth three-pole system (TBT), along with the Bessel (T) and Butterworth (B) phase.

The normalized envelope delay is calculated as:

$$\tau_e = \frac{\sigma_1}{\sigma_1^2 + \omega^2} + \frac{\sigma_2}{\sigma_2^2 + (\omega + \omega_2)^2} + \frac{\sigma_2}{\sigma_2^2 + (\omega - \omega_2)^2} \tag{4.5.21}$$

$$= -\frac{0.9704}{0.9704^2 + \omega^2} - \frac{0.6275}{0.6275^2 + (\omega + 0.7981)^2} - \frac{0.6275}{0.6275^2 + (\omega - 0.7981)^2}$$

The envelope delay plot is shown in Fig. 4.5.4 (TBT), along with the delays for the Butterworth and the modified Bessel system.

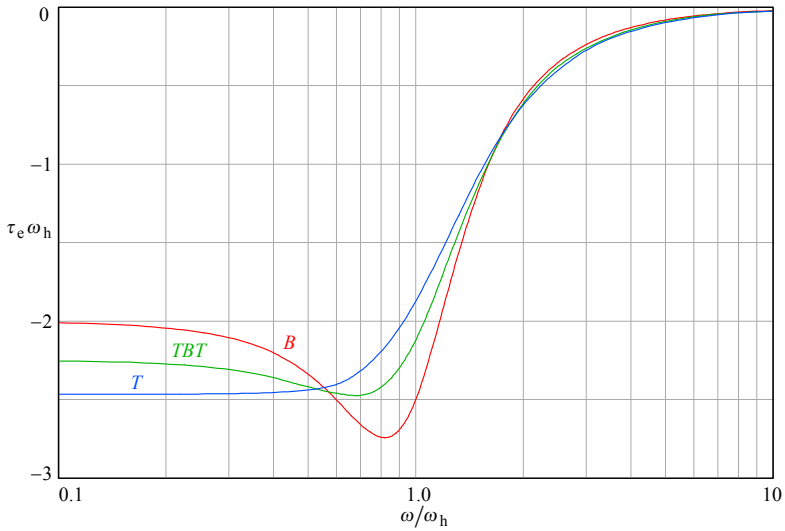


Fig. 4.5.4: Envelope delay of the Transitional Bessel–Butterworth three-pole system (TBT), along with the Bessel (T) and Butterworth (B) delays.

The starting point for the step response calculation is the general formula for a three pole function multiplied by the unit step operator $1/s$:

$$G(s) = \frac{-\,s_1\,s_2\,s_3}{s\,(s - s_1)(s - s_2)(s - s_3)} \tag{4.5.22}$$

We calculate the corresponding step response in the time domain by the $\mathcal{L}^{-1}$ transform:

$$g(t) = \mathcal{L}^{-1}\{G(s)\} = \sum_{i=1}^{n} \mathrm{res}_i\, G(s)\, \mathrm{e}^{st}$$

$$= \sum_{i=1}^{3} \mathrm{res}_i \frac{-\,s_1\,s_2\,s_3\,\mathrm{e}^{st}}{s\,(s - s_1)(s - s_2)(s - s_3)} \tag{4.5.23}$$

After the sum of the residues is calculated, we insert the poles $s_1 = \sigma_1$ and $s_{2,3} = \sigma_2 \pm j\,\omega_2$ to obtain (see Appendix 2.3):

$$g(t) = 1 - \frac{\sigma_1 \sqrt{\left[\sigma_2(\sigma_2 - \sigma_1) - \omega_2^2\right]^2 + \omega_2^2(2\sigma_2 - \sigma_1)^2}}{\omega_2\left[(\sigma_2 - \sigma_1)^2 + \omega_2^2\right]}\, e^{\sigma_2 t} \sin(\omega_2\, t + \theta)$$

$$- \frac{(\sigma_2^2 + \omega_2^2)}{(\sigma_2 - \sigma_1)^2 + \omega_2^2}\, e^{\sigma_1 t} \qquad (4.5.24)$$

where the angle θ is:

$$\theta = \pi + \arctan \frac{-\omega_2(2\sigma_2 - \sigma_1)}{\sigma_2(\sigma_2 - \sigma_1) - \omega_2^2} \qquad (4.5.25)$$

By inserting the numerical values for poles from Eq. 4.5.18, we arrive at the final relation:

$$g(t) = 1 + 1.4211\, e^{-6275\, t} \sin(0.7981\, t + 2.8811) - 1.3660\, e^{-0.9704\, t} \qquad (4.5.26)$$

The plot based on this formula is shown in Fig. 4.5.5, (TBT). By inserting the appropriate pole values in Eq. 4.5.24 we obtain the plots of Butterworth (B) and modified Bessel (T) system's step responses.

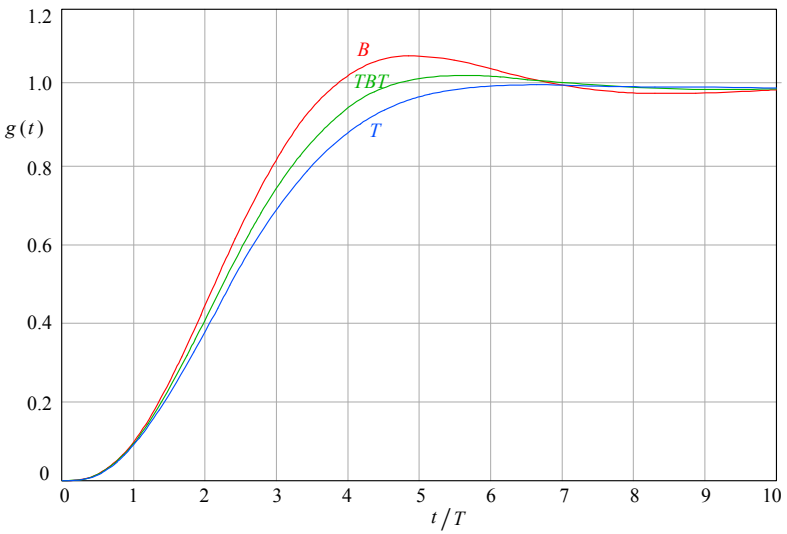


Fig. 4.5.5: The step response of the Transitional Bessel–Butterworth three-pole system (TBT), along with the Bessel (T) and Butterworth (B) responses.

4.6 Staggered vs. Repeated Bessel Pole Pairs

In order to compare the performance of an amplifier with staggered (index 's') vs. repeated Bessel pole pairs (index 'r') we need to compare the following two frequency response functions:

$$\left|F_{\mathrm{s}}(s)\right| = \left| \frac{s_1 s_2 \cdots s_n}{(s-s_1)(s-s_2)\cdots(s-s_n)} \right| \tag{4.6.1}$$

and:

$$\left|F_{\mathrm{r}}(s)\right| = \left| \frac{(s_1 s_2)^{n/2}}{(s-s_1)^{n/2}\,(s-s_2)^{n/2}} \right| \tag{4.6.2}$$

where n is an even integer (2, 4, 6, …). For a fair comparison we must use the poles from Table 4.4.4, the Bessel poles normalized to the same cutoff frequency.

The plots in Fig. 4.6.1 of these two functions were made by a computer, using the numerical methods described in Part 6. From this figure it is evident that an amplifier with staggered poles (as reported in the Table 4.4.4 for each n) preserves the intended bandwidth. On the other hand, the amplifier with the same total number of poles, but of second-order, repeated $(n/2)$-times, does not — its bandwidth shrinks with each additional second-order stage. Obviously, if $n = 2$ the systems are identical.

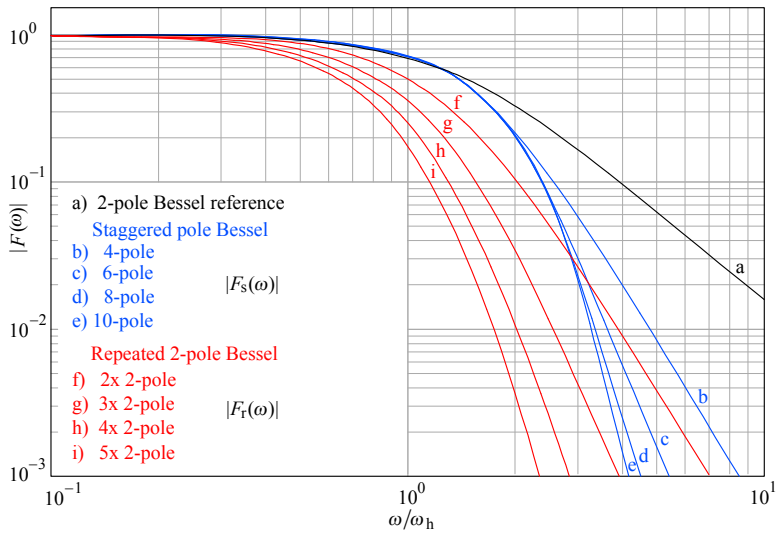


Fig. 4.6.1: Frequency response magnitude of systems with staggered poles, compared with systems with repeated second-order pole pairs. By using staggered poles the bandwidth is preserved, whilst for systems with repeated poles it decreases with each additional stage.

Even if the poles were of a different kind, e.g., Butterworth or Chebyshev poles, the staggered poles would also preserve the bandwidth, but the system with repeated second-order pole pairs will not. For the same total number of poles the curves tend to

the same cut off asymptote (from Fig. 4.6.1 this not evident, but it would have been clear if the graphs had been plotted with increased vertical scale, say, down to 10^{-6}).

In the time domain the decrease of rise times is even more evident. To compare the step responses we take Eq. 4.6.1 and 4.6.2 (without the magnitude sign) and multiply them by the unit-step operator $1/s$, obtaining:

$$G_{\mathrm{s}}(s) = \frac{s_1 s_2 \cdots s_n}{s\,(s-s_1)(s-s_2)\cdots(s-s_n)} \tag{4.6.3}$$

and:

$$G_{\mathrm{r}}(s) = \frac{(s_1 s_2)^{n/2}}{s\,(s-s_1)^{n/2}\,(s-s_2)^{n/2}} \tag{4.6.4}$$

with n being again an even integer.

By using the $\mathcal{L}^{-1}$ transform, we obtain the step responses in the time domain:

$$g_{\mathrm{s}}(t) = \mathcal{L}^{-1}\{G_{\mathrm{s}}(s)\} = \sum \mathrm{res}\, \frac{s_1 s_2 \cdots s_n\, \mathrm{e}^{st}}{s\,(s-s_1)(s-s_2)\cdots(s-s_n)} \tag{4.6.5}$$

and:

$$g_{\mathrm{r}}(t) = \mathcal{L}^{-1}\{G_{\mathrm{r}}(s)\} = \sum \mathrm{res}\, \frac{(s_1 s_2)^{n/2}\, \mathrm{e}^{st}}{s\,(s-s_1)^{n/2}\,(s-s_2)^{n/2}} \tag{4.6.6}$$

The analytical calculation of these two equations, for n equal to 2, 4, 6, 8 and 10, should be a pure routine by now (at least for readers who have followed the calculations in Part 2 and Appendix 2.3; for those who have skipped them, there will be a lot of opportunities to revisit them later, when the need will force them!). Anyway, this can be done more easily by using a computer, employing the numerical methods described in Part 6. The plots obtained are shown in Fig. 4.6.2. The figure is convincing enough and does not need any comment.

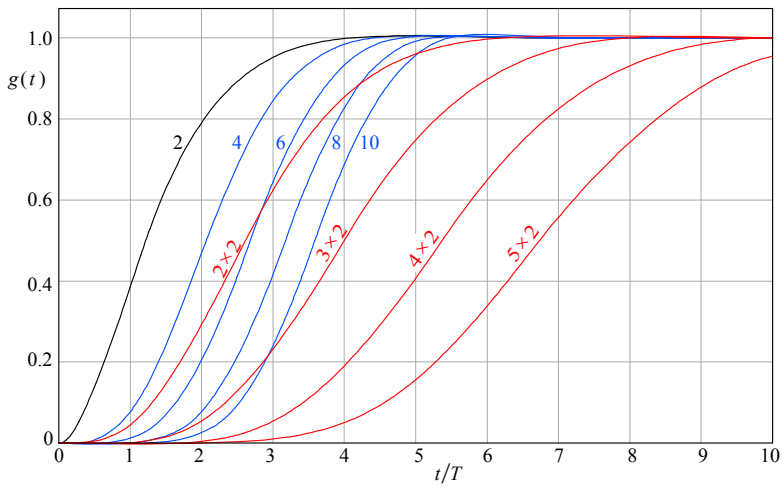


Fig. 4.6.2: Step response of systems with staggered poles, compared with systems with repeated second-order pole pairs. The rise times of systems with repeated poles increase with each additional stage, whilst for systems with staggered poles they even decrease slightly!

4.6.1 Assigning the Poles For Maximum Dynamic Range

Readers who usually pay attention to details may have noted that we have listed the poles in our tables (and also in figures) in a particular order. We have combined them in complex conjugate pairs and listed them by the increasing of their imaginary part. Yes, there is a reason for this, beyond pure aesthetics.

In general we can choose any number (m) of poles from the total number (n) of poles in the system, and assign them to any stage of the total number of stages (k). Sometimes, a particular choice would be limited by reasons other than gain and bandwidth, e.g., in oscilloscopes, the first stage is a JFET source follower, which provides the high input impedance required and it is usually difficult to design an effective peaking around a unity gain stage, so almost universally this stage has only a single real pole. In most other cases the choice is governed mainly by the gain × bandwidth product available for a given number of stages.

However, the main reason which dictates the particular pole ordering is the dynamic range. Remember that in wideband amplifiers we are, more often than not, at the limits of a system's realizability. If we want to extract the maximum performance from a system we should limit any overshoot at each stage to a minimum.

If we consider a rather extreme example, by putting the pole pair with the highest imaginary part in the first amplifier stage, the step response of this stage would exhibit a high overshoot. Consequently, the maximum amplitude which the system could handle linearly would be reduced by the amount of that overshoot.

In order to make the argument more clear, let us take a 3-stage 5-pole system with Bessel poles (Table 4.4.3, $n = 5$) and analyze the step response of each stage separately for two different assignments. In the first case we shall use a reversed pole assignment: the pair $s_{4,5}$ will be assigned to the first stage, the pair $s_{2,3}$ to the second stage and the real pole s_1 to the last stage. In the second case we shall assign the poles in the preferred order, the real pole first and the pair with the largest imaginary part last.

Our poles have the following numerical values:

$$\begin{aligned} s_1 &= -3.6467 \\ s_{2,3} &= -3.3520 \pm j\,1.7427 \\ s_{4,5} &= -2.3247 \pm j\,3.7510 \end{aligned}$$

We can model the actual amplifier by three voltage driven current generators loaded by appropriate CR or CLR networks. Since the passive components are isolated by the generators, each stage response can be calculated separately. We have thus one first-order function and two second-order functions:

$$g_1(t) = \sum \operatorname{res} \frac{s_1\, \mathrm{e}^{st}}{s\,(s - s_1)} = 1 + \frac{1}{s_1}\, \mathrm{e}^{s_1 t}$$

$$g_2(t) = \sum \operatorname{res} \frac{s_2 s_3\, \mathrm{e}^{st}}{s\,(s - s_2)(s - s_3)} = 1 + \frac{1}{|\sin\theta_2|}\, \mathrm{e}^{\sigma_2 t} \sin(\omega_2 t + \theta_2)$$

$$g_3(t) = \sum \operatorname{res} \frac{s_4 s_5\, \mathrm{e}^{st}}{s\,(s - s_4)(s - s_5)} = 1 + \frac{1}{|\sin\theta_4|}\, \mathrm{e}^{\sigma_4 t} \sin(\omega_4 t + \theta_4)$$

In the reverse pole order case the first stage, with the pole pair $s_{4,5}$, is excited by the unit step input signal. We know that the second-order system from Table 4.4.3 has an optimal step response; since the imaginary to real part ratio of $s_{4,5}$ is larger (greater $\tan\theta_4$) than it is with the poles of the optimal case, we thus expect that the stage with $s_{4,5}$ would exhibit a pronounced overshoot.

In Fig. 4.6.3 we have drawn the response of each stage when it is driven individually by the unit step input signal (the responses are gain-normalized to allow comparison). It is evident that the stage with poles $s_{4,5}$ has a 13% overshoot.

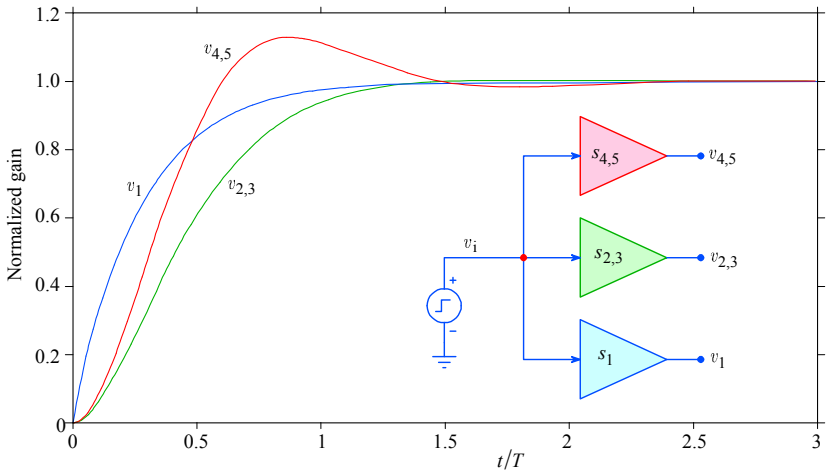


Fig. 4.6.3: Step response of each of the three stages taken individually.

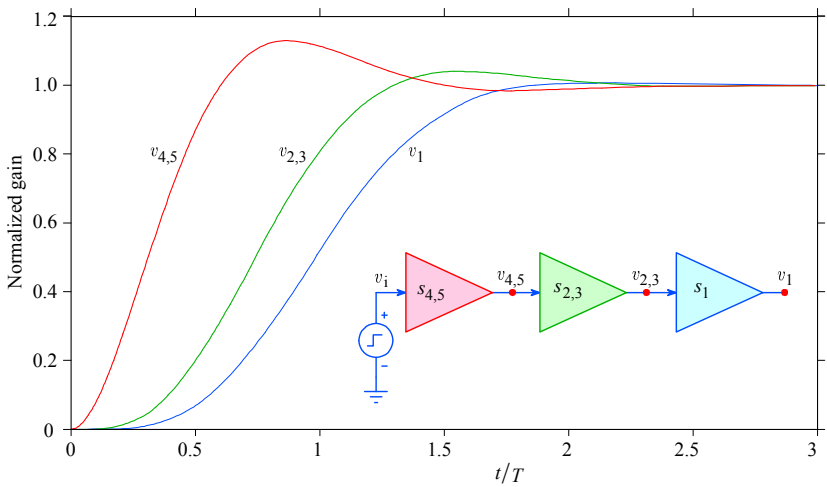


Fig. 4.6.4: Step response of the complete amplifier with reverse pole order at each stage. Although the second stage had no overshoot of its own, it overshoots by nearly 5% when processing the $v_{4,5}$ output.

In Fig. 4.6.4 the step response of the complete amplifier is drawn, showing the signal after each stage. Note that although the second stage exhibits no overshoot when driven by the unit step (Fig. 4.6.3), it will overshoot by nearly 5% when driven by the output of the first stage, $v_{4,5}$. And the overshoot would be even higher if the pole s_1 had been assigned to the middle stage.

The dynamic range of the input stage will therefore have to be larger by 13% and that of the second stage by 5% in order to handle the signal linearly. Fortunately the maximal input signal is equal to the maximal output, divided by the total gain. If we have followed the rule given by Eq. 4.1.33 and Fig. 4.1.9 the input stage will have only $1/3$ of the total system gain, so its output amplitude will be only a fraction of the supply voltage. On the other hand, the optimal stage gain is rather low, as given by Eq. 4.1.38, so the dynamic range may become a matter of concern after all.

The circuit configuration which is most critical in this respect is the cascode amplifier, since there are two transistors effectively in series with the power supply, so the biasing must be carefully chosen. In traditional discrete circuits, with relatively high supply voltages, the dynamic range was rarely a problem; the major concern was about poor linearity for large signals, since no feedback was used. In modern ICs, with lots of feedback and a supply of only 5V or just 3V, the usable dynamic range can be critical.

We can easily prevent this limitation if we use the correct pole ordering, so that the first stage has the real pole s_1 and the last stage the pole pair $s_{4,5}$. As we can see in Fig. 4.6.5, the situation improves considerably, since in this case the two front stages exhibit no overshoot, while the output overshoot is 0.4% only. Note that the final response in all cases is the same, though.

In a real amplifier, the pole assignment chosen can be affected by other factors, e.g., the stage with the largest capacitance will require the poles with the lowest imaginary part; alternatively a lower loading resistor and thus lower gain can be chosen for that stage.

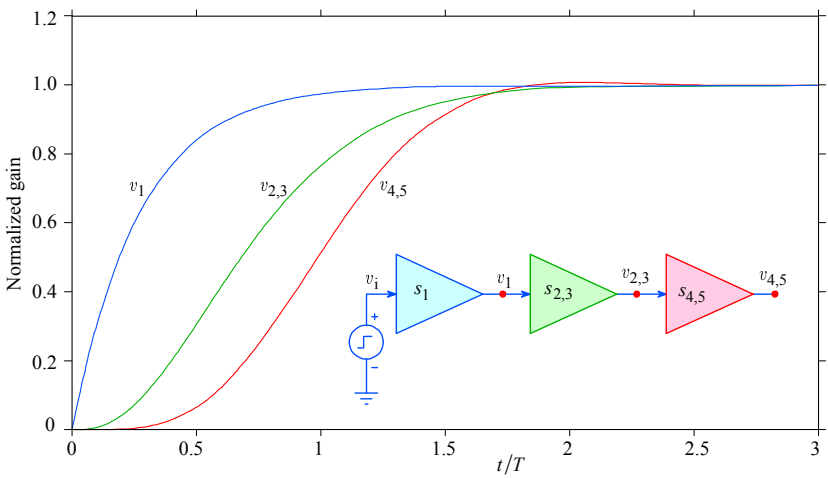


Fig. 4.6.5: Step response of the complete amplifier, but with the correct pole assignment.

Résumé of Part 4

The study of this part should have given the reader enough knowledge to acquire an idea of how multi-stage amplifiers could be optimized by applying inductive peaking circuits, discussed in Part 2, at each stage.

Also, the merit of using DC over AC coupled multi-stage amplifiers should be clearly understood.

A proper pole pattern selection is of fundamental importance for the amplifier's performance. In particular, for a smooth, low overshoot transient response the envelope delay extended flatness offered by the Bessel poles provides a nearly ideal solution, approaching the ideal Gaussian response very quickly: with only 5 poles, the system's frequency and step response conform exceptionally well to the ideal, with a difference of less than 1% throughout most of the transient.

Finally, the advantage of staggered vs. repeated pole pairs should be strictly considered in the design of multi-stage amplifiers when gain × bandwidth efficiency is the primary design goal. We hope that the reader has gained awareness of how the bandwidth, which has been achieved by hard work following the optimizations of each basic amplifying stage given in Part 3, can be easily lost by a large factor if the stages are not coupled correctly.

A few examples of how these principles are used in practice are given in Part 5.

References:

[4.1] *P.M. Chirlian*, Electronic Principles and Design,
McGraw-Hill, New York, 1971.

[4.2] *G.E. Valley & H. Wallman*, Vacuum Tube Amplifiers,
MIT Radiation Laboratory Series Vol.18, McGraw-Hill, New York, 1948.

[4.3] *J. Bednařík & J. Daněk*, Obrazové zesilovače pro televisi a měřicí techniku,
Statní nakladatelství technické literatury, Prague, 1957.

[4.4] *L.T. Bruton*, *RC*–Active Circuits, Theory and Design,
Prentice-Hall, 1980.

[4.5] *J.M. Pettit and M. McWhorter*, Electronic Amplifier Circuits,
McGraw-Hill, New York, 1961.

[4.6] *S. Butterworth*, On the Theory of Filter Amplifiers,
Experimental Wireless & Wireless Engineer, Vol.7, 1930, pp. 536–541.

[4.7] *W. Gellert, H. Küstner, M. Hellwich, H. Kästner*, The VNR Concise Encyclopedia of
Mathematics, second edition, Van Nostrand Reinhold Company, New York, 1992.

[4.8] *L. Storch*, Synthesis of Constant Time-Delay Ladder Networks Using Bessel Polynomials,
Proceedings of I.R.E., Vol. 7, 1954, pp. 536–541.

[4.9] *W.E. Thomson*, Networks with Maximally Flat Delay,
Wireless Engineer, Vol. 29, 1952, pp. 253–263.

[4.10] *M.E. Van Valkenburg*, Introduction to Modern Network Synthesis,
John Wiley, New York, 1960.

[4.11] *M. Abramowitz & I.A. Stegun*, Handbook of Mathematical Functions,
Dover Publication, New York, 1970.

[4.12] *J.N. Little & C.B. Moller*, *MATLAB* User's Guide,
The MathWorks, Inc., South Natick, USA, 1990.

[4.13] *P. Starič*, Interpolation Between Butterworth and Thomson Poles (in Slovenian),
Elektrotehniški vestnik, 1987, pp. 133–139.

[4.14] *Y. Peless & T. Murakami*, Analysis and Synthesis of Transitional Butterworth-Thomson Filters
and Bandpass Amplifiers, RCA Review, Vol.18, March, 1957, pp. 60–94.

[4.15] *D.L. Feucht*, Handbook of Analog Circuit Design,
Academic Press, Inc., San Diego, 1990.

[4.16] *A.B. Williams and F.J. Taylor*, Electronic Filter Design Handbook
(LC, Active and Digital Filters), Second Edition, McGraw-Hill, New York, 1988.

[4.17] *A.I. Zverev*, Handbook of Filter Synthesis,
John Wiley and Sons, New York, 1967.

P. Starič, E. Margan:

Wideband Amplifiers

Part 5:

System Synthesis And Integration

Any sufficiently advanced technology is indistinguishable from magic.
Arthur C. Clarke
(Profiles of the Future: An Inquiry
into the Limits of the Possible, 1973)

Contents:

List of Figures:

List of Tables:

5.0 'The Product Is Greater Than The Sum Of Its Parts'

... and that can be true in both the mathematical and technological sense! Well, in math, at least as long as we are dealing with numbers greater than two; but in technology the goal might not be so straightforward and neither are the means of achieving it.

Electronics engineering is a difficult and demanding job. Most electronics engineers pay attention to the input and output constraints imposed by the real world as a matter of course. Many will also silently nod their head when the components logistician tells them to use another part instead of the one originally specified, just because it's cheaper. A number of them will also agree to take it into account when the marketing manager tells them that the customer wants and expects from the product a feature which was not foreseen initially as one of the design goals. And almost all will shrug their shoulders when the chief director announces that the financial resources for their project have been cut low or even that the project has been canceled. But almost all will go mad when the mechanical engineer or the enclosure designer casually stops by and asks if a switch or a pot could be moved from the left side of the printed circuit board to the right. Fortunately for the electronics engineer, he has on his side the most powerful argument of all: "Yeah, maybe I could do that, but probably the final performance would suffer!" Electronics engineering is a difficult and demanding job, indeed.

In the past 50 years electronics engineers have been delivering miracles at an ever increasing rate. Not just the general public, but also other people involved within the electronics industry have become accustomed to this. In the mid 1980s no one ever asked you if something could be done; instead, the question was 'for how much and when?'. Today, no one asks 'how much' either — it has to be a bargain (between brothers) and it had better be ready yesterday!

How many of you could give a name or two of engineers who became rich or just famous in the last 50 or so years? Let us see: William R. Hewlett was famous, but that was probably due more to his name in the Hewlett–Packard firm's title and less to the actual recognition of his work by the general public. Then there is Ray Dolby, known for his noise reduction system. Gordon Moore of Intel is known for his 'law'. Errr ... who else? Oh, yes, Bill Gates is both rich and famous, but he is more famous because he is so rich and far less for his own work!

Do you see what we mean? No doubt, many engineers have started a very profitable business based on their own original circuit ideas. True, most commercially successful products are the result of a team effort; still, the key solution is often born in the mind of a gifted individual. But, frankly speaking, is there a single engineer or scientist who could attract 50,000 delirious spectators to a stadium every Sunday? OK, 5,000? Maybe Einstein could have been able to do so, but even he was considered to be 'a bit nuts'. Well, there you have it, the offer–demand economy.

In this book we have tried to pay a humble tribute to a small number of amplifier designers. But our readers are engineers themselves, so we are preaching to the already converted. Certainly we are not going to influence, let alone reverse, any of those trends.

So, if tomorrow your boss tells you that he will be paying you 5% less, shrug your shoulders and go back to your drawing board. And do not forget to stop by the front-panel designer to tell him that he is asking too much!

5.1 Geometrical Synthesis of Inductively Compensated Multi-Stage Amplifiers — A Simple Example

The reader who has patiently followed the discussion presented in previous chapters is probably eager to see all that theory being put into practice.

Before jumping to some more complex amplifier circuits we shall give a relatively simple example of a two-stage differential cascode amplifier, by which we shall illustrate the actual system optimization procedure in some detail, using the previously developed principles in their full potential.

Since we want to grasp the 'big picture' we shall have to leave out some less important topics, such as negative input impedance compensation, cascode damping, etc.; these are important for the optimization of each particular stage which, once optimized, can be idealized to some extent. We have covered that extensively enough in Part 3, so we shall not explicitly draw the associated components in the schematic diagram. But, at the end of our calculations, we shall briefly discuss the influence of those components to final circuit values.

A two-stage amplifier is a 'minimum complexity' system for which the multi-stage design principles still apply. To this we shall add a 3-pole T-coil and a 4-pole L+T-coil peaking networks, discussed in Part 2, as loads to each stage, making a total of 7 poles. There is, however, an additional real pole, owed to the Q_1 input capacitance and the total input and signal source resistance. As we shall see later, this pole can be neglected if its distance from the complex plane's origin is at least twice as large as that of the system real pole set by $-1/R_{\mathrm{a}}C_{\mathrm{a}}$.

Such an amplifier thus represents an elementary example in which everything that we have learned so far can be applied. The reader should, however, be aware that this is by no means the ideal or, worse still, the only possibility. At the end of our calculations, when we shall be able to assess the advantages and limitations offered by our initial choices at each stage, we shall examine a few possibilities of further improvement.

We shall start our calculations from the unavoidable stray capacitances and the desired total voltage gain. Then we shall apply an optimization process, which we like to refer to as the **geometrical synthesis**, by which we shall calculate all the remaining circuit components in such a way that the resulting system will conform to the 7-pole normalized Bessel–Thomson system. The only difference will be that the actual amplifier poles will be larger by a certain factor, proportional (but not equal) to the upper half power frequency ω_{H}. We have already met the geometrical synthesis in its basic form in Part 2, Fig. 2.5.3 when we were discussing the 3-pole T-coil circuit. The name springs from the ability to calculate all the peaking network components from simple geometrical relations which involve the pole's real and imaginary components, given, of course, the desired pole pattern and a few key component values which can either be chosen independently or set by other design requirements. Here we are going to see a generalization of those basic relations applied to the whole amplifier.

It must be admitted that the constant and real input impedance of the T-coil network is the main factor which allows us to assign so many poles to only two stages. A cascade of passive 2-pole sections could have been used, but those would load each

other and, as a result, the bandwidth extension factor would suffer. Another possibility would be to use an additional cascode stage to separate the last two peaking sections, but another active stage, whilst adding gain, also adds its own problems to be taken care of. It is, nevertheless, a perfectly valid option.

Let us now take a quick tour of the amplifier schematic, Fig. 5.1.1. We have two differential cascode stages and two current sources, which set both the transistor's transconductance and the maximum current available to the load resistors, R_a and R_b. This limits the voltage range available to the CRT. Since the circuit is differential the total gain is a double of each half. The total DC gain is (approximately):

$$A_0 = 2\,\frac{R_a}{R_{e1}} \cdot \frac{R_b}{R_{e3}} \tag{5.1.1}$$

The values of R_{e1} and R_{e3} set the required capacitive bypass, $C_{e1}/2$ and $C_{e3}/2$, to match the transistor's time constants. In turn, this sets the input capacitance at the base of Q_1 and Q_3, to which we must add the inevitable C_{cb} and some strays.

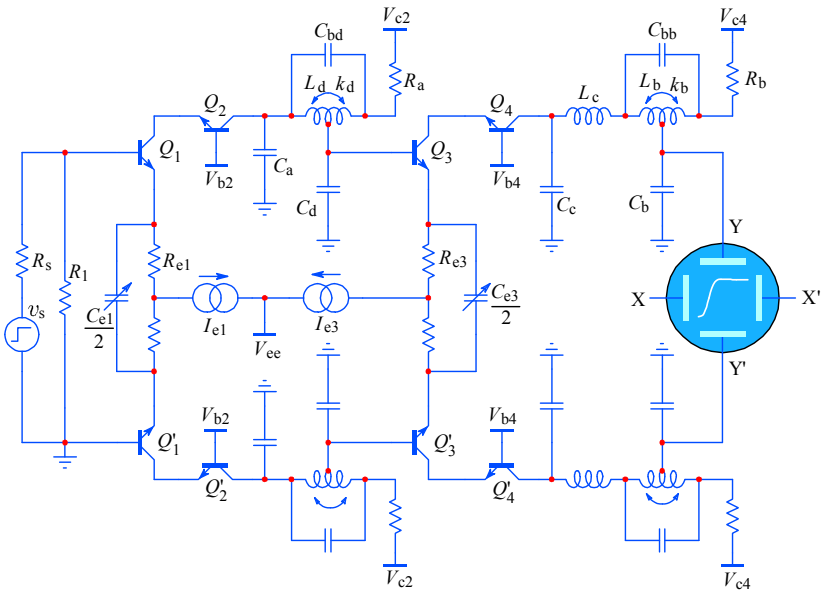


Fig. 5.1.1: A simple 2-stage differential cascode amplifier with a 7-pole peaking system: the 3-pole T-coil inter-stage peaking (between the Q_2 collector and the Q_3 base) and the 4-pole L+T-coil output peaking (between the Q_4 collector and the vertical plates of the cathode ray tube). The schematic was simplified to emphasize the important design aspects — see text.

The capacitance C_d should thus consists of, preferably, only the input capacitance at the base of Q_3. If required by the coil 'tuning', a small capacitance can be added in parallel, but that would also reduce the bandwidth. Note that the associated T-coil L_d will have to be designed as an inter-stage peaking, as we have discussed in Part 3, Sec. 3.6, but we can leave the necessary corrections for the end.

The capacitance C_b, owed almost entirely to the CRT vertical plates, is much larger than C_d, so we expect that R_a and R_b can not be equal. From this it follows that

it might be difficult to apply equal gain to each stage in accordance with the principle explained in Part 4, Eq. 4.1.39. Nevertheless, the difference in gain will not be too high, as we shall see.

Like any other engineering process, geometrical synthesis also starts from some external boundary conditions which set the main design goal. In this case it is the CRT's vertical sensitivity and the available input voltage, from which the total gain is defined. The next condition is the choice of transistors by which the available current is defined. Both the CRT and the transistors set the lower limit of the loading capacitances at various nodes. From these the first circuit component R_b is fixed.

With R_b fixed we arrive at the first 'free' parameter, which can be represented by several circuit components. However, since we would like to maximize the bandwidth this parameter should be attributed to one of the capacitances. By comparing the design equations for the 3-pole T-coil and the 4-pole L+T-coil peaking networks in Part 2, it can be deduced that C_a, the input capacitance of the 3-pole section, is the most critical component.

With these boundaries set let us assume the following component values:

$$\begin{aligned} C_b &= 11\,\text{pF} && (9\,\text{pF of the CRT vertical plates, } 2\,\text{pF stray}) \\ C_a &= 4\,\text{pF} && (3\,\text{pF from the } Q_2\ C_{cb},\ 1\,\text{pF stray}) \\ R_b &= 360\ \Omega && (\text{determined by the desired gain and the available current}) \end{aligned} \qquad (5.1.2)$$

The pole pattern is, in general, another 'free' parameter, but for a smooth, minimum overshoot transient we must apply the Bessel–Thomson arrangement. As can be seen in Fig. 5.1.2, each pole (pair) defines a circle going through the pole and the origin, with the center on the negative real axis.

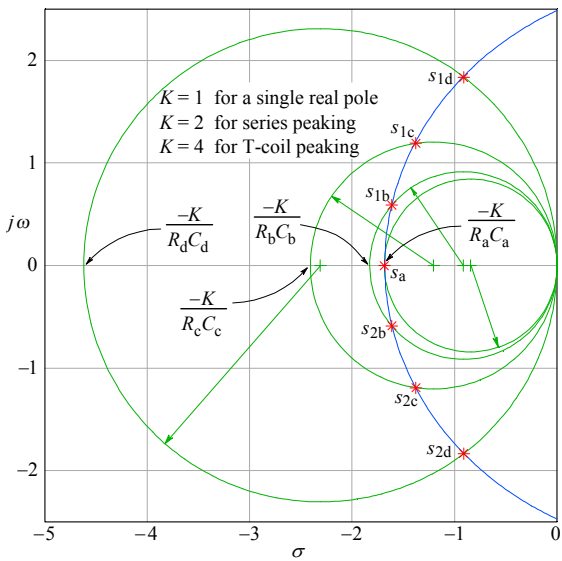


Fig. 5.1.2: The 7 normalized Bessel–Thomson poles. The characteristic circle of each pole (pair) has a diameter determined by the appropriate RC constant and the peaking factor K, which depends on the type of network chosen.

The poles in Fig. 5.1.2 bear the index of the associated circuit components and the reader might wonder why we have chosen precisely that assignment.

In a general case the assignment of a pole (pair) to a particular circuit section is yet another 'free' design parameter. If we were designing a low frequency filter we could indeed have chosen an arbitrary assignment (as long as each complex conjugate pole pair is assigned as a pair, a limitation owed to physics, instead of circuit theory).

If, however, the bandwidth is an issue then we must seek those nodes with the largest capacitances and apply the poles with the lowest imaginary part to those circuit sections. This is because the capacitor impedance (which is dominantly imaginary) is inversely proportional both to the capacitor value and the signal frequency.

In this light the largest capacitance is at the CRT, that is, C_b; thus the pole pair with the lowest imaginary part is assigned to the output T-coil section, formed by L_b and R_b, therefore acquiring the index 'b', $s_{1\mathrm{b}}$ and $s_{2\mathrm{b}}$.

The real pole is the one associated with the 3-pole stage and there it is set by the loading resistor R_a and the input capacitance C_a, becoming s_a.

The remaining two pole pairs should be assigned so that the pair with the larger imaginary part is applied to that peaking network which has a larger bandwidth improvement factor. Here we must consider that $K = 4$ for a T-coil, whilst $K = 2$ for the series peaking L-section (of the 4-pole L+T-scetion). Clearly the pole pair with the larger imaginary part should be assigned to the inter-stage T-coil, L_d, thus they are labeled $s_{1\mathrm{d}}$ and $s_{2\mathrm{d}}$. The L-section then receives the remaining pair, $s_{1\mathrm{c}}$ and $s_{2\mathrm{c}}$.

We have thus arrived at a solution which seems logical, but in order to be sure that we have made the right choice we should check other combinations as well. We are going to do so at the end of the design process.

The poles for the normalized 7th-order Bessel–Thomson system, as taken either from Part 4, Table 4.4.3, or by using the BESTAP (Part 6) routine, along with the associated angles, are:

$$\begin{aligned}
s_\mathrm{a} = \sigma_\mathrm{a} &= -4.9718 && \theta_\mathrm{a} = 180° \\
s_\mathrm{b} = \sigma_\mathrm{b} \pm j\,\omega_\mathrm{b} &= -4.7583 \pm j\,1.7393 && \theta_\mathrm{b} = 180° \mp 20.0787° \\
s_\mathrm{c} = \sigma_\mathrm{c} \pm j\,\omega_\mathrm{c} &= -4.0701 \pm j\,3.5172 && \theta_\mathrm{c} = 180° \mp 40.8316° \\
s_\mathrm{d} = \sigma_\mathrm{d} \pm j\,\omega_\mathrm{d} &= -2.6857 \pm j\,5.4207 && \theta_\mathrm{d} = 180° \mp 63.6439°
\end{aligned} \qquad (5.1.3)$$

So, let us now express the basic design equations by the assigned poles and the components of the two peaking networks.

For the real pole s_a we have the following familiar proportionality:

$$s_\mathrm{a} = \sigma_\mathrm{a} = D_\mathrm{a} = -4.9718 \;\propto\; \frac{-1}{R_\mathrm{a}\,C_\mathrm{a}} \qquad (5.1.4)$$

At the output T-coil section, according to Part 2, Fig. 2.5.3, we have:

$$D_\mathrm{b} = \frac{\sigma_\mathrm{b}}{\cos^2\theta_\mathrm{b}} = \frac{-4.7583}{0.8821} = -5.3941 \;\propto\; \frac{-4}{R_\mathrm{b}\,C_\mathrm{b}} \qquad (5.1.5)$$

For the L-section of the L+T output network, because the T-coil input impedance is equal to the loading resistor, we have:

$$D_c = \frac{\sigma_c}{\cos^2\theta_c} = \frac{-4.0701}{0.5725} = -7.1094 \quad \propto \quad \frac{-2}{R_b\,C_c} \tag{5.1.6}$$

And finally, for the inter-stage T-coil network:

$$D_d = \frac{\sigma_d}{\cos^2\theta_d} = \frac{-2.6857}{0.1917} = -13.6333 \quad \propto \quad \frac{-4}{R_a\,C_d} \tag{5.1.7}$$

From these relations we can calculate the required values of the remaining capacitances, C_c and C_d. If we divide Eq. 5.1.5 by Eq. 5.1.6, we have the ratio:

$$\frac{D_b}{D_c} = \frac{-\dfrac{4}{R_b\,C_b}}{-\dfrac{2}{R_b\,C_c}} = \frac{2\,C_c}{C_b} \tag{5.1.8}$$

It follows that the capacitance C_c should be:

$$C_c = \frac{C_b}{2}\cdot\frac{D_b}{D_c} = \frac{11}{2}\cdot\frac{-5.3941}{-7.1094} = 4.1730\,\text{pF} \tag{5.1.9}$$

Likewise, if we divide Eq. 5.1.4 by Eq. 5.1.7, we obtain:

$$\frac{D_a}{D_d} = \frac{-\dfrac{1}{R_a\,C_a}}{-\dfrac{4}{R_a\,C_d}} = \frac{C_d}{4\,C_a} \tag{5.1.10}$$

Thus C_d will be:

$$C_d = 4\,C_a\,\frac{D_a}{D_d} = 4\cdot 4\cdot\frac{-4.9718}{-13.6333} = 5.8349\,\text{pF} \tag{5.1.11}$$

Of course, for most practical purposes, the capacitances do not need to be calculated to such precision, a resolution of 0.1 pF should be more than enough. But we would like to check our procedure by recalculating the actual poles from circuit components and for that purpose we shall need this precision.

Now we need to know the value of R_a. This can be readily calculated from the ratio D_a/D_b:

$$\frac{D_a}{D_b} = \frac{-\dfrac{1}{R_a\,C_a}}{-\dfrac{4}{R_b\,C_b}} = \frac{R_b\,C_b}{4\,R_a\,C_a} \tag{5.1.12}$$

resulting in:

$$R_a = \frac{R_b}{4}\cdot\frac{C_b}{C_a}\cdot\frac{D_b}{D_a} = \frac{360}{4}\cdot\frac{11}{4}\cdot\frac{-5.3941}{-4.9718} = 268.5\,\Omega \tag{5.1.13}$$

We are now ready to calculate the inductances L_b, L_c and L_d. For the two T-coils we can use the Eq. 2.4.19:

$$L_b = R_b^2\, C_b = 360^2 \cdot 11 \cdot 10^{-12} = 1.4256\,\mu H \tag{5.1.14}$$

and

$$L_d = R_a^2\, C_d = 268.5^2 \cdot 5.8349 \cdot 10^{-12} = 0.4206\,\mu H \tag{5.1.15}$$

For L_c we use Eq. 2.2.26 to obtain the proportionality factor of the RC constant:

$$L_c = \frac{1+\tan^2\theta_b}{4}\, R_b^2\, C_c = \frac{360^2 \cdot 4.1730 \cdot 10^{-12}}{4 \cdot 0.8821} = 0.1533\,\mu H \tag{5.1.16}$$

The magnetic coupling factors for the two T-coils are calculated by Eq. 2.4.36:

$$k_b = \frac{3-\tan^2\theta_b}{5+\tan^2\theta_b} = \frac{3-0.1336}{5+0.1336} = 0.5584 \tag{5.1.17}$$

and likewise:

$$k_d = \frac{3-\tan^2\theta_d}{5+\tan^2\theta_d} = \frac{3-4.0738}{5+4.738} = -0.1183 \tag{5.1.18}$$

Note that k_d is negative. This means that, instead of the usually negative mutual inductance, we need a positive inductance at the T-coil tap. This can be achieved by simply mounting the two halves of L_d perpendicular to each other, in order to have zero magnetic coupling and then introduce an additional coil, L_e (again perpendicular to both halves of L_d), with a value of the required positive mutual inductance, as can be seen in Fig. 5.1.3. Another possibility would be to wind the two halves of L_d in opposite direction, but then the bridge capacitance C_{bd} might be difficult to realize correctly.

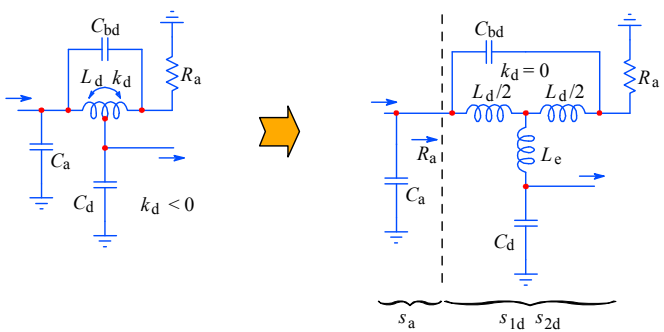


Fig. 5.1.3: With the assigned poles and the resulting particular component values the 3-pole stage magnetic coupling k_d needs to be negative, which forces us to use non-coupled coils and add a positive mutual inductance L_e. Even with a negative k_d the T-coil reflects its resistive load to the network input, greatly simplifying the calculations of component values.

The additional inductance L_e can be calculated from the required mutual inductance given by the negative value of k_d. In Part 2, Eq. 2.4.1–2.4.5 we have defined the T-coil inductance, its two halves, and its mutual inductance by the relations repeated in Eq. 5.1.19 for convenience:

$$L = L_1 + L_2 + 2\,L_M$$

$$L_1 = L_2 = \frac{L}{2\,(1+k)} \qquad (5.1.19)$$

$$L_M = -\,k\,\sqrt{L_1\,L_2}$$

Thus, if $k = 0$ we have:

$$L_{1d} = L_{2d} = \frac{L_d}{2} = \frac{0.4206}{2} = 0.2103\,\mu H \qquad (5.1.20)$$

and:

$$L_e = -\,k_d\,\sqrt{\frac{L_d}{2}\cdot\frac{L_d}{2}} = -\,k_d\,\frac{L_d}{2} = 0.1183\,\frac{0.4206}{2} = 0.025\,\mu H \qquad (5.1.21)$$

If we were to account for the Q_3 base resistance (discussed in Part 3, Sec. 3.6) we would get k_d even more negative and also $L_{1d} \neq L_{2d}$.

The coupling factor k_b, although positive, also poses a problem: since it is greater than 0.5 it might be difficult to realize. As can be noted from the above equations, the value of k depends only on the pole's angle θ. In fact, the 2nd-order Bessel system has the pole angles of $\pm 150°$, resulting in a $k = 0.5$, representing the limiting case of realizability with conventionally wounded coils. Special shapes, coil overlapping, or other exotic techniques may solve the coupling problem, but, more often than not, they will also impair the bridge capacitance. The other limiting case, when $k = 0$, is reached by the ratio $\Im\{s\}/\Re\{s\} = \sqrt{3}$, a situation occurring when the pole's angle $\theta = 120°$.

In accordance with previous equations we also calculate the value of the two halves of L_b:

$$L_{1b} = L_{2b} = \frac{L_b}{2\,(1+k_b)} = \frac{1.4256}{2\,(1+0.5584)} = 0.4574\,\mu H \qquad (5.1.22)$$

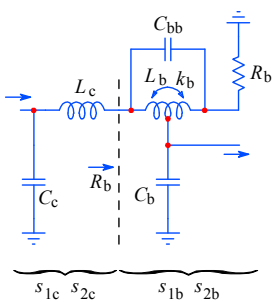


Fig. 5.1.4: The 4-pole output L+T-coil stage and its pole assignment.

The last components to be calculated are the bridge capacitances, C_{bb} and C_{bd}. The relation between the T-coil loading capacitance and the bridge capacitance has been given already in Part 2, Eq. 2.4.31, from which we obtain the following expressions for C_{bb} and C_{bd}:

$$C_{bb} = C_b \frac{1+\tan^2\theta_b}{16} = 11\,\frac{1+0.1336}{16} = 0.7793\,\text{pF} \qquad (5.1.23)$$

and:

$$C_{bd} = C_d \frac{1+\tan^2\theta_d}{16} = 5.8349\,\frac{1+4.0738}{16} = 1.8503\,\text{pF} \qquad (5.1.24)$$

This completes the calculation of amplifier components necessary for the inductive peaking compensation and thus achieving the Bessel–Thomson system response. We would now like to verify the design by recalculating the actual pole values. To do this we return to the relations which we have started from, Eq. 5.1.3 to Eq. 5.1.7 and for the imaginary part using the relations in Part 2, Fig. 2.5.3 . In order not to confuse the actual pole values with the normalized values, from which we started, we add an index 'A' to the actual poles:

$$\sigma_{aA} = -\frac{1}{R_a C_a} = -\frac{1}{268.5\cdot 4\cdot 10^{-12}} = -931.1\cdot 10^6\,\text{rad/s} \qquad (5.1.25)$$

$$\sigma_{bA} = -\frac{4\cos^2\theta_b}{R_b C_b} = -\frac{4\cdot 0.8821}{360\cdot 11\cdot 10^{-12}} = -891.0\cdot 10^6\,\text{rad/s}$$

$$\omega_{bA} = \pm\frac{4\cos\theta_b\sin\theta_b}{R_b C_b} = \pm\frac{4\cdot 0.9392\cdot 0.3433}{360\cdot 11\cdot 10^{-12}} = \pm 325.7\cdot 10^6\,\text{rad/s}$$

$$\sigma_{cA} = -\frac{2\cos^2\theta_c}{R_b C_c} = -\frac{2\cdot 0.5725}{360\cdot 4.1730\cdot 10^{-12}} = -762.2\cdot 10^6\,\text{rad/s}$$

$$\omega_{cA} = \pm\frac{2\cos\theta_c\sin\theta_c}{R_b C_c} = \pm\frac{2\cdot 0.7566\cdot 0.6538}{360\cdot 4.1730\cdot 10^{-12}} = \pm 658.5\cdot 10^6\,\text{rad/s}$$

$$\sigma_{dA} = -\frac{4\cos^2\theta_d}{R_a C_d} = -\frac{4\cdot 0.1917}{268.5\cdot 5.8349\cdot 10^{-12}} = -489.5\cdot 10^6\,\text{rad/s}$$

$$\omega_{dA} = \pm\frac{4\cos\theta_d\sin\theta_d}{R_a C_d} = \pm\frac{4\cdot 0.4439\cdot 0.8961}{268.5\cdot 5.8349\cdot 10^{-12}} = \pm 1015.6\cdot 10^6\,\text{rad/s}$$

If we divide the real amplifier pole by the real normalized pole, we get:

$$\frac{\sigma_{bA}}{\sigma_b} = \frac{-931.1\cdot 10^6}{-4.9718} = 187.3\cdot 10^6 \qquad (5.1.26)$$

and this factor is equal for all other pole components. Unfortunately, from this we cannot calculate the upper half power frequency of the amplifier. The only way to do that (for a Bessel system) is to calculate the response for a range of frequencies around the cut off and then iterate it using the bisection method, until a satisfactory tolerance has been achieved.

Instead of doing it for only a small range of frequencies we shall, rather, do it for a three decade range and compare the resulting response with the one we would get from a non-compensated amplifier (in which all the inductances are zero). Since to this point we were not interested in the actual value of the voltage gain, we shall make the comparison using amplitude normalized responses.

The non-compensated amplifier has two real poles, which are:

$$s_{1N} = -\frac{1}{R_a\,(C_a + C_d)} \quad \text{and} \quad s_{2N} = -\frac{1}{R_b\,(C_b + C_c)} \tag{5.1.27}$$

Consequently, its complex frequency response would then be:

$$F_N(s) = \frac{s_{1N}\,s_{2N}}{(s - s_{1N})(s - s_{2N})} \tag{5.1.28}$$

with the magnitude:

$$\left|F_N(\omega)\right| = \frac{\sqrt{s_{1N}\,s_{2N}}}{\sqrt{(\omega^2 - s_{1N}^2)(\omega^2 - s_{2N}^2)}} \tag{5.1.29}$$

and the step response:

$$\begin{aligned} g(t) &= \mathcal{L}^{-1}\left\{\frac{s_{1N}\,s_{2N}}{s\,(s - s_{1N})(s - s_{2N})}\right\} \\ &= 1 + \frac{s_{2N}}{s_{1N} - s_{2N}}\,e^{s_{1N}t} - \frac{s_{1N}}{s_{1N} - s_{2N}}\,e^{s_{2N}t} \end{aligned} \tag{5.1.30}$$

The rise time is:

$$\tau_r = 2.2\sqrt{\frac{1}{s_{1N}^2} + \frac{1}{s_{2N}^2}} \tag{5.1.31}$$

and the half power frequency:

$$f_h = \frac{\sqrt{s_{1N}\,s_{2N}}}{2\,\pi} \tag{5.1.32}$$

In contrast, the complex frequency response of the 7-pole amplifier is:

$$F_A(s) = A_0\,\frac{-\,s_{aA}\,s_{1bA}\,s_{2bA}\,s_{1cA}\,s_{2cA}\,s_{1dA}\,s_{2cA}}{(s - s_{aA})(s - s_{1bA})(s - s_{2bA})(s - s_{1cA})(s - s_{2cA})(s - s_{1dA})(s - s_{2dA})} \tag{5.1.33}$$

and the step response is the inverse Laplace transform of the product of $F_A(s)$ with the unit step operator $1/s$:

$$g(t) = \mathcal{L}^{-1}\left\{\frac{1}{s}\,F_A(s)\right\} = \sum \mathrm{res}\left(\frac{1}{s}\,F_A(s)\,e^{st}\right) \tag{5.1.34}$$

We shall not attempt to solve either of these functions analytically, since it would take too much space, and, anyway, we have solved them separately for its two parts (3rd- and 4th-order) in Part 2. Because the systems are separated by an amplifier (Q_3, Q_4), the frequency response would be a simple multiplication of the two responses. For the step response we now have 8 residues to sum (7 of the system poles, in addition to the one from the unit step operator). Although lengthy, it is a relatively simple operation and we leave it as an exercise to the reader. Instead we are going to use the computer routines, the development of which can be found in Part 6.

In Fig. 5.1.5 we have made a polar plot of the poles for the inductively compensated 7-pole system and the non-compensated 2-pole system. As we have learned in Part 1 and Part 2, the farther from origin the smaller is the pole's influence on the system response. It is therefore obvious that the 2-pole system's response will be dominated by the pole closer to the origin and that is the pole of the output stage, s_{2N}. The bandwidth of the 7-pole system is, obviously, much larger.

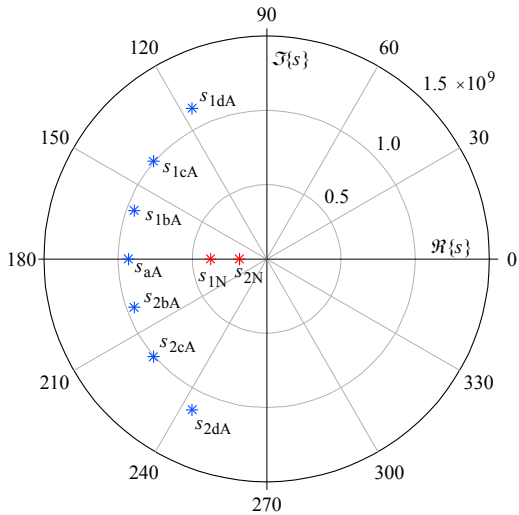


Fig. 5.1.5: The polar plot of the 7-pole compensated system (poles with index 'A') and the 2-pole non-compensated system (index 'N'). The radial scale is $\times 10^9$ rad/s. The angle is in degrees.

The pole layout gives us a convenient indication of the system's performance, but it is the magnitude vs. frequency response that reveals it clearly. As can be seen in Fig. 5.1.6, the non-compensated system has a bandwidth of less than 25 MHz. The compensated amplifier bandwidth is close to 88 MHz, more than 3.5 times larger.

The comparison of step responses in Fig. 5.1.7 reveals the difference in performance even more dramatically. The rise time of the non-compensated system is about 14 ns, whilst for the compensated system it is only 3.8 ns, also a factor of 3.5 times better; in addition, the overshoot is only 0.48 %.

Both comparisons show an impressive improvement in performance. But is it the best that could be obtained from this circuit configuration? After all, in Part 2 we have seen a similar improvement from just the 4-pole L+T-coil section and we expect that the addition of the 3-pole section should yield a slightly better result at least.

One obvious way of extending the bandwidth would be to lower the value of R_b, increase the bias currents, and scale the remaining components accordingly. Then we should increase the input signal amplitude to get the same output. But this is the 'trivial' solution (mathematically, at least; not so when building an actual circuit).

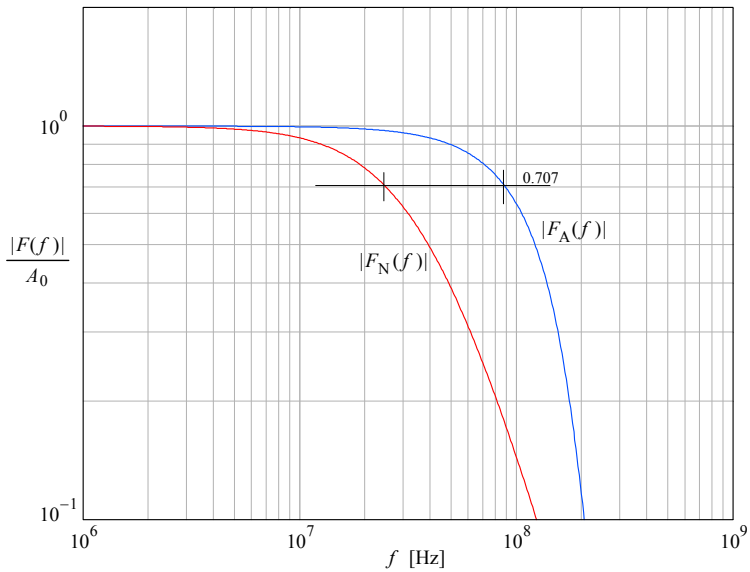


Fig. 5.1.6: The gain normalized magnitude vs. frequency of the 7-pole compensated system $|F_A(f)|$ and the 2-pole non-compensated system, $|F_N(f)|$. The bandwidth of F_N is about 25 MHz and the bandwidth of F_A is about 88 MHz, more than 3.5 times larger.

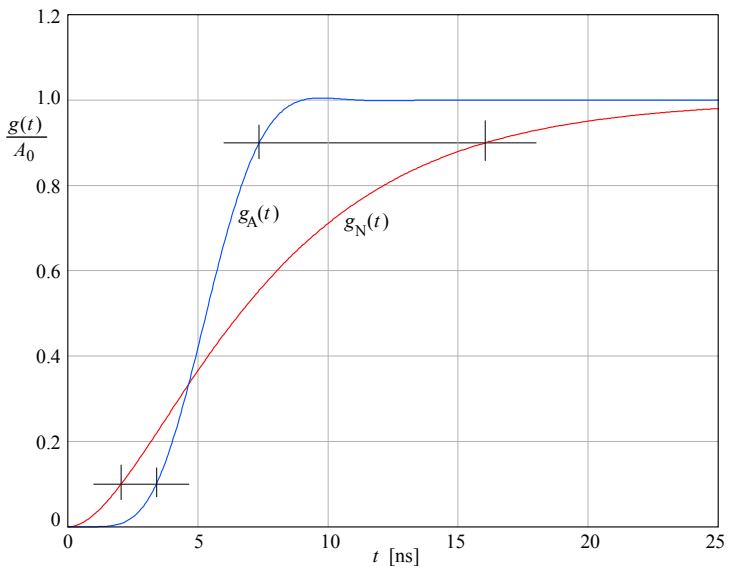


Fig. 5.1.7: The gain normalized step responses of the 7-pole compensated system $g_A(t)$ and the 2-pole non-compensated system $g_N(t)$. The rise time is 14 ns for the $g_N(t)$, but only 3.8 ns for $g_A(t)$. The overshoot of $g_A(t)$ is only 0.48 %.

By a careful inspection of the amplifier design equations and comparing them with the analysis of the two sections in Part 2, we come to the conclusion that the most serious bandwidth drawback factor is the high value of the CRT capacitance, which is much higher than C_a or C_d. But if so, did we limit the possible improvement by assigning the poles with the lowest imaginary part to the output? Should not we obtain a better performance if we add more peaking to the output stage?

Since we have put the design equations and the response analysis into a computer routine, we can now investigate the effect of different pole assignments. To do so we simply reorder the poles and run the routine again. Besides the pole order that we have described, let us indicate it by the pole order: abcd (Eq. 5.1.3), we have five additional permutations: abdc, acbd, adbc, acdb, adcb. The last two permutations result in a rather slow system, requiring a large inductance for L_b and large capacitances C_c and C_d. But the remaining ones deserve a look.

In Fig. 5.1.8 we have plotted the four normalized step responses and, because there are two identical pairs of responses, we have displaced them vertically by a small offset in order to distinguish them more clearly.

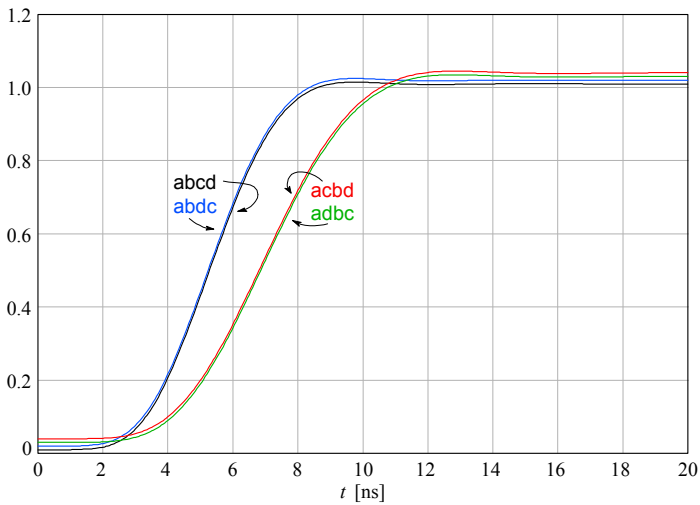


Fig. 5.1.8: The normalized step responses of the four possible combinations of pole assignments. There are two pairs of responses, here spaced vertically by a small offset to allow easier identification. One of the two faster responses (labeled 'abcd') is the one for which the detailed analysis has been given in the text.

If the pole pairs s_c and s_d are mutually exchanged the result is the same as our original analysis. But by exchanging s_b with either s_c or s_d the result is sub-optimal.

A closer look at Table 5.1.1 reveals that both of the two slower responses have $R_a = 354\,\Omega$ instead of $268\,\Omega$. The higher value of R_a means actually a higher gain, as can be seen in Fig. 5.1.9, where the original system was set for a gain of $A_0 \approx 10$, in contrast with the higher value, $A_0 \approx 13$. The higher gain results from a different 'tuning' of the 3-pole T-coil stage, in accordance with the different pole assignment.

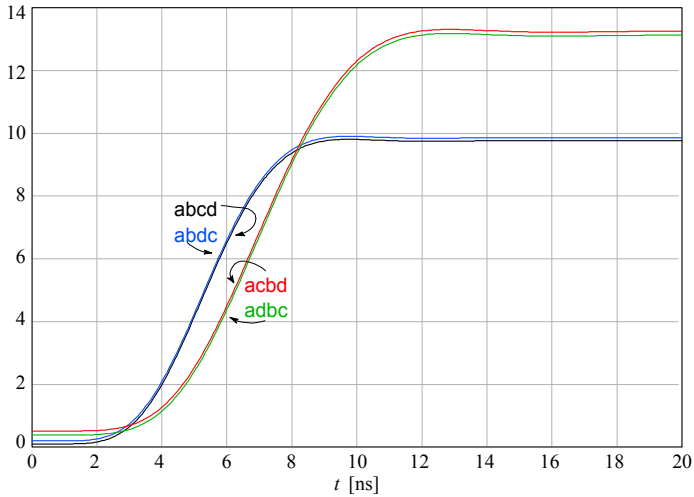


Fig. 5.1.9: The slower responses of Fig. 5.1.8, when plotted with the actual gain, are actually those with a higher value of R_a and therefore a higher gain.

Since our primary design goal is to maximize the bandwidth with a given gain, let us recalculate the slower system for a lower value of R_b. If $R_b = 316\,\Omega$ (from the E96 series of standard values, 0.5 % tolerance), the gain is restored. Fig. 5.1.10 shows the recalculated responses, labeled 'acbd' and 'abdc', whilst the 'abcd' and 'abdc' responses are the same as before.

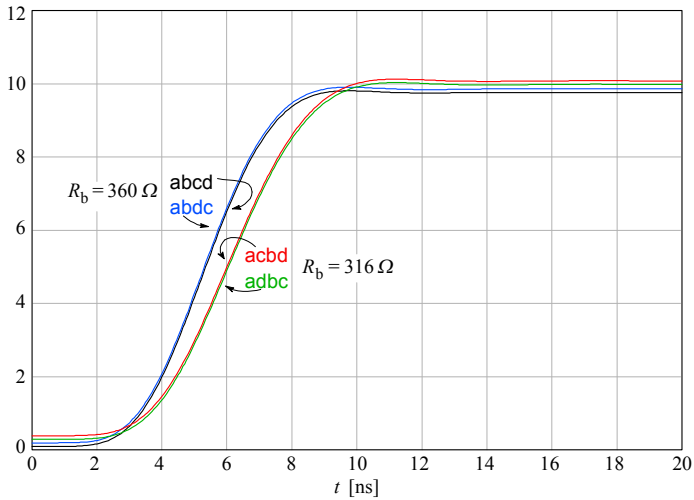


Fig. 5.1.10: If the high gain responses are recalculated by reducing R_b from the original 360 Ω to 316 Ω, the gain is nearly equal in all four cases, However, those pole assignments which put the poles with the higher imaginary part at the output stage still result in a slightly slower system.

The difference in rise time between the two pairs is much smaller now; however, the recalculated pair is still slightly slower. This shows that our initial assumptions of how to achieve maximum bandwidth (within a given configuration) were not guessed by sheer luck.

In Table 5.1.1 we have collected all the design parameters for the four out of six possible pole assignments. The systems in the last two columns have the same pole assignments as in the middle two, but have been recalculated from a lower R_b value, in order to obtain the total voltage gain nearly equal to the first system. From a practical point of view the first and the last column are the most interesting: the system represented by the first column is the fastest (as the second one, but the latter is difficult to realize, mainly owing to low C_c value), whilst the last one is only slightly slower but much easier to realize, mainly owing to a lower magnetic coupling k_b and the non-problematic values of C_c and C_d.

Table 5.1.1

R_b [Ω]	360	360	360	360	316	316
pole order:	abcd	abdc	acbd	adbc	acbd	adbc
A_0	9.667	9.667	12.74	12.74	9.817	9.817
R_a [Ω]	268.5	286.5	353.9	353.9	310.7	310.7
C_c [pF]	4.173	2.177	2.870	7.249	2.870	7.249
C_d [pF]	5.838	11.19	14.75	5.838	14.75	5.838
C_{bb} [pF]	0.779	0.779	1.201	1.201	1.201	1.201
C_{bd} [pF]	1.851	1.222	1.045	1.851	1.045	1.851
L_b [μH]	1.426	1.426	1.426	1.426	1.098	1.098
L_c [μH]	0.153	0.080	0.162	0.410	0.125	0.316
L_d [μH]	0.421	0.807	1.847	0.731	1.423	0.563
k_b	0.558	0.558	0.392	0.392	0.392	0.392
k_d	−0.118	0.392	0.558	−0.118	0.558	−0.118
η_b	3.57	3.57	2.35	2.35	2.84	2.84
η_r	3.55	3.55	2.33	2.33	2.81	2.81

Table 5.1.1: Circuit components for 4 of the 6 possible pole assignments. The last two columns represent the same pole assignment as the middle two, but have been recalculated for $R_b = 316\,\Omega$ and nearly equal gain. The first column is the example calculated in the text and its response is one of the two fastest. The other fast system (second column) is probably non realizable (in discrete form), because $C_c \approx 2$ pF. The last column (adbc) is, on the other hand, only slightly slower, but probably much easier to realize (T-coil coupling and the capacitance values). The bandwidth and rise time improvement factors η_b and η_r were calculated by taking the non-compensated amplifier responses as the reference.

The main problem encountered in the realization of our original 'abcd' system is the relatively high magnetic coupling factor of the output T-coil, k_b. A possible way of improving this could be by applying a certain amount of emitter peaking to either the Q_1 or Q_3 emitter circuit. Then we would have a 9-pole system and we would have to recalculate everything. However, the use of emitter peaking results in a negative input impedance which has to be compensated (see Part 3, Sec. 3.5), and the compensating network adds more stray capacitance.

A 9-pole system might be more easily implemented if, instead of the 3-pole section, we were to use another L+T-coil 4-pole network. The real pole could then be provided by the signal source resistance and the Q_1 input capacitance, which we have chosen to neglect so far. With 9 poles both T-coils can be made to accommodate those two pole pairs with moderate imaginary part values (because the T-coil coupling factor depends only on the pole angle θ), so that the system bandwidth could be more easily maximized. A problem could arise with a low value of some capacitances, which might become difficult to achieve. But, as is evident from Table 5.1.1, there are many possible variations (their number increases as the factorial of the number of poles), so a clever compromise can always be made. Of course, with a known signal source an additional inductive peaking could be applied at the input, resulting in a total of 11 or perhaps even 13 poles, but then the component tolerances and the adjustment precision would set the limits of realizability.

Finally, we would like to verify the initial claim that the input real pole s_1, owed to the signal source resistance and base spread resistance and the total input capacitance, can be neglected if it is larger than the system real pole s_a. Since the input pole is separated from the rest of the system by the first cascode stage, it can be accounted for by simply multiplying the system transfer function by it. In the frequency response its influence is barely noticeable. In the step response, Fig. 5.1.11, it affects mostly the envelope delay and the overshoot, whilst the rise time (in accordance with the frequency response) remains nearly the same.

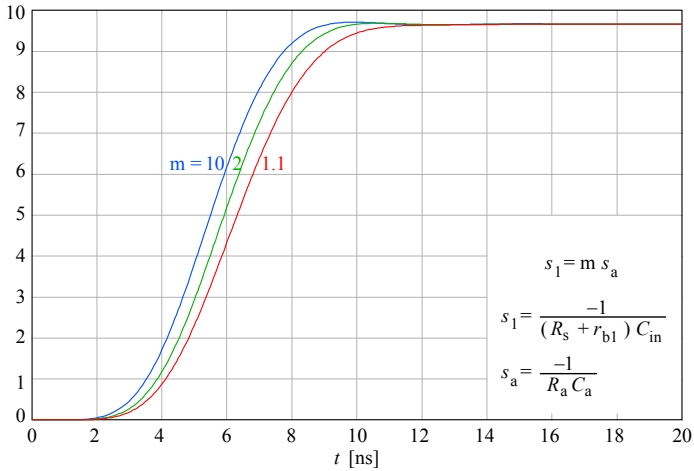


Fig. 5.1.11: If the real input pole s_1 is at least twice as large as the system's real pole s_a, its influence on the step response can be seen merely as an increased envelope delay and a reduced overshoot, whilst the rise time remains nearly identical.

Usually the signal source's impedance is 50 Ω or less; an input capacitance of several pF would still ensure that the input pole is high above the system real pole. However, an oscilloscope needs an input buffer stage and a preamplifier, with variable gain and attenuation, to adapt the signal amplitude to the required level. This preamplifier's output impedance, driving the input capacitance of our amplifier, can be high enough, forcing us to account for it. In such cases, as already stated before, it

might become feasible to replace the 3-pole peaking network by another 4-pole L+T-coil network and make the input pole the main system real pole.

As already mentioned, the relatively high capacitance of the CRT vertical deflection plates is the dominant cause for the amplifier bandwidth limitation.

To avoid this problem the most advanced CRTs from the analog 'scope era have had their deflection plates made in a number of sections, connected externally by a series of T-coils (see Fig. 5.1.12), thus reducing the capacitance seen by the amplifier to just a fraction of the original value. At the same time, the T-coils have provided a delay required to match the signal propagation to the electron velocity in the writing beam (compensating for the electron's finite travel time by the deflection plates, as well as some non negligible relativistic effects! — see Appendix 5.1), thus aiding to a better beam control.

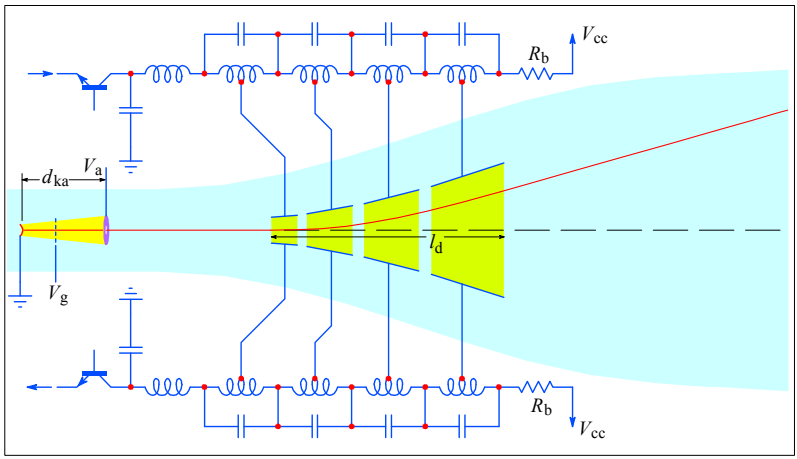


Fig. 5.1.12: If the CRT deflection plates are made in a number of sections (usually between 4 and 8), connected by a series of T-coil peaking circuits, the amplifier would effectively be loaded by a much smaller capacitance, allowing the system cut off frequency to be several times higher. The T-coils also provide the time delay necessary to keep the deflecting voltage (as seen by the electrons in the writing beam) almost constant throughout the electron's travel time across the deflecting field. For simplicity, only the vertical deflection system is shown, but a similar circuit could be used for the horizontal deflection, too (such an example can be found in the 1 GHz Tektronix 7104 model; see Appendix 5.1 for further details). Note that, owing to the increasing distance between the plates their length should also vary accordingly, in order to compensate for the reduced capacitance. Fortunately, the capacitance is also a function of the plate's width, not just length and distance, so a well balanced compromise can always be found.

5.2 High Input Impedance Selectable Attenuator with a JFET Source Follower

A typical oscilloscope vertical input must incorporate a number of passive signal conditioning functions:

1) a selectable 50 Ω/1 MΩ input resistance, with low reflection coefficient on the 50 Ω setting;
2) a selectable DC–GND–AC coupling;
3) a selectable 1:1/10:1/100:1 or similar attenuation with a 1 MΩ || 10 pF input impedance (independent of the selected attenuation; resistance tolerance 0.1 %, capacitance between 10–20 pF, since the external probes are adjustable, but its dielectric properties must be constant with frequency and temperature);
4) a 2 kV electrostatic discharge spark gap, able to protect the input from a human body model discharge (200 pF, 15 kV);
5) a 400 V (DC + peak AC) continuous protection of the delicate input amplifier at no input attenuation setting (except for the 50 Ω setting).

To this we must add the following requirements:

6) the upper cut off frequency at least twice higher than the system's bandwidth;
7) the upper cut off frequency should be independent of any of the above settings;
8) the gain flatness must be kept within 0.5 % from DC to 1/5 of the bandwidth;
9) the protection diodes must survive repeating 1–2 A surge currents with < 1 ns rise and 50 µs decay, their leakage must be < 100 pA and capacitance < 1 pF;

To preserve a high degree of signal integrity the stray capacitances and inductances must be kept low throughout the input stage, which means small components with a small size of soldering pads and the traces as short as possible.

In addition, the unity gain JFET buffer stage performance should include:

10) a > 100 MΩ input resistance;
11) a < 1 pF input capacitance;
12) a 50 Ω output resistance or the ability to drive such loads;
13) the bandwidth at least twice higher than the rest of the system;
14) $< 5\,\mathrm{nV}/\sqrt{\mathrm{Hz}}$ input noise density (see [Ref. 5.66] for low noise design);
15) $< 0.5\,\mathrm{mV_{pp}}$ wideband noise (at 5 mV/div.; important for digitizing 'scopes);
16) gain close to unity, flat within 0.5 %, up to 1/5 of the system bandwidth;
17) low overshoot and undershoot, any resonance up to 10× the system's bandwidth should be well damped to reduce ringing;
18) fast step settling time to within 0.1% of the final signal value;
19) recovery from input overdrive as short as possible;
20) ability to handle signals in a wide range, from $8\,\mathrm{mV_{pp}}$ (1 mV/div) up to $40\,\mathrm{V_{pp}}$ (5V/div, with input attenuation), with a DC offset of ± 1/2 screen at least;

This is an impressive list, indeed. Especially if we consider that for a 500 MHz system bandwidth the above requirements should be fulfilled for a 1 GHz bandwidth.

A typical input stage block diagram is shown in Fig. 5.2.1. The attenuator and the unity gain buffer stage will be analyzed in the following sections.

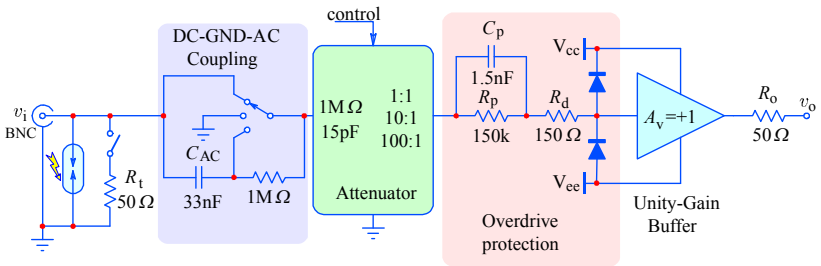


Fig. 5.2.1: A typical conventional oscilloscope input section. All the switches must be high voltage types, controlled either mechanically or as electromagnetic relays (but other solutions are also possible, as in [Ref. 5.2]). The spark gap protects against electrostatic discharge. The R_t 50 Ω resistor is the optional transmission line termination. The 1 MΩ resistor in the DC–GND–AC selector charges the AC coupling capacitor in the GND position, reducing the overdrive shock through C_{ac} in presence of a large DC signal component. The attenuator is analyzed in detail in Sec. 5.2.1–3. The overdrive protection limits the input current in case of an accidental connection to the 240 V_{ac} with the attenuator set to the highest sensitivity. The unity gain buffer/impedance transformer is a >100 MΩ R_{in}, 50 Ω R_o JFET or MOSFET source follower, analyzed in Sec. 5.2.4 and 5.2.6.

5.2.1 Attenuator High Frequency Compensation

A simple resistive attenuator, like the one in Fig. 5.2.2a is too sensitive to any capacitive loading by the following amplifier stage. For oscilloscopes the standard input impedance $R_a = R_1 + R_2$ is 1 MΩ, so for a 10:1 attenuation the resistance values must be $R_1 = 900\,\text{k}\Omega$ and $R_2 = 100\,\text{k}\Omega$. With such values the output impedance, which equals the parallel connection of both resistances, would be about 90 kΩ. Assuming an amplifier input capacitance of only 1 pF, the resulting system bandwidth would be only 1.77 MHz [$f_h = (R_1 + R_2)/(2\pi C_i R_1 R_2)$].

Therefore high frequency compensation, as shown in Fig. 5.2.2b, is necessary if we want to obtain higher bandwidth. The frequency compensation, however, lowers the input impedance at high frequencies.

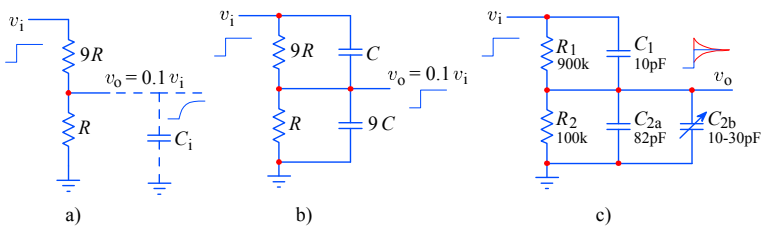


Fig. 5.2.2: The 10:1 attenuator; **a) resistive:** with $R = 100\,\text{k}\Omega$, the following stage input capacitance of just 1 pF would limit the bandwidth to only 1.77 MHz; **b) compensated:** the capacitive divider takes over at high frequencies but the input capacitance of the following stage of 1 pF would spoil the division by 1%; **c) adjustable:** in practice, the capacitive divider is trimmed for a perfect step response.

In general, at DC the signal source is loaded by the total attenuation resistance R_a; for an attenuation factor A, the values of resistor R_1 and R_2 must satisfy the following equations:

$$R_1 + R_2 = R_\mathrm{a} \tag{5.2.1}$$

The current through the resistive path is:

$$i_\mathrm{i} = \frac{v_\mathrm{i}}{R_1 + R_2} = \frac{v_\mathrm{o}}{R_2} \tag{5.2.2}$$

so, from the last two expressions, the attenuation is:

$$\frac{v_\mathrm{o}}{v_\mathrm{i}} = \frac{1}{A} = \frac{R_2}{R_1 + R_2} \tag{5.2.3}$$

and the required resistance relation is:

$$R_1 = (A - 1)R_2 \tag{5.2.4}$$

Thus, for an $R_\mathrm{a} = 1\,\mathrm{M\Omega}$ and $A = 10$:

$$R_1 = 900\,\mathrm{k\Omega} \qquad \text{and} \qquad R_2 = 100\,\mathrm{k\Omega} \tag{5.2.5}$$

The high frequency compensation consists of a capacitive divider having the same attenuation factor as the high impedance resistive divider in parallel with it, as in Fig. 5.2.2b. In order to achieve a precise attenuation, resistors with 0.1% tolerance are used, giving a maximum error of 0.2%. However, capacitors with a comparably tight tolerance are not readily available, and, even if they were, the layout strays would dominate. So in practice the capacitive divider is made adjustable, Fig. 5.2.2c. Trimming of C_1 should be avoided, in order to reduce the circuit size (and thus stray inductances and capacitances); a much better choice is to trim only some 20–30% of C_2. Care should be taken to connect the variable plate to the ground, otherwise the metal tip of the adjusting screwdriver would modify the capacitance by contact alone.

For a well trimmed attenuator, the capacitive reactance ratio at high frequencies must match the resistor ratio at DC and low frequencies:

$$\frac{R_1}{R_2} = \frac{X_{C1}}{X_{C2}} = \frac{\dfrac{1}{j\omega C_1}}{\dfrac{1}{j\omega C_2}} = \frac{C_2}{C_1} \tag{5.2.6}$$

which also implies that the two RC constants must be equal:

$$R_1 C_1 = R_2 C_2 = \tau_\mathrm{a} \tag{5.2.7}$$

The input impedance now becomes:

$$\begin{aligned} Z_\mathrm{a} = Z_1 + Z_2 &= \frac{1}{\dfrac{1}{R_1} + j\omega C_1} + \frac{1}{\dfrac{1}{R_2} + j\omega C_2} \\ &= \frac{R_1}{1 + j\omega C_1 R_1} + \frac{R_2}{1 + j\omega C_2 R_2} = (R_1 + R_1)\frac{1}{1 + j\omega\,\tau_\mathrm{a}} \end{aligned} \tag{5.2.8}$$

In the latter expression we have taken into account Eq. 5.2.7. This is the same as if we would have a single parallel $R_a C_a$ network:

$$Z_a = R_a \frac{1}{1 + j\omega C_a R_a} \tag{5.2.9}$$

where:

$$R_a = R_1 + R_2 \qquad \text{and} \qquad C_a = \frac{1}{\frac{1}{C_1} + \frac{1}{C_2}} \tag{5.2.10}$$

By substituting $C_2 = (A-1)C_1$ the input capacitance C_a relates to C_1 as:

$$C_a = \frac{1}{\frac{1}{C_1} + \frac{1}{(A-1)C_1}} = C_1 \frac{A-1}{A} \tag{5.2.11}$$

The transfer function can then be calculated from the attenuation:

$$F(j\omega) = \frac{1}{A} = \frac{v_{out}}{v_{in}} = \frac{Z_2}{Z_1 + Z_2} = \frac{1}{1 + \frac{R_1}{R_2} \cdot \frac{1 + j\omega C_2 R_2}{1 + j\omega C_1 R_1}} \tag{5.2.12}$$

Obviously, the frequency dependence will vanish if the condition of Eq. 5.2.7 is met. However, the transfer function will be independent of frequency only if the signal's source impedance is zero (we are going to see the effects of the signal source impedance a little later).

The transfer function of an unadjusted attenuator ($R_1 C_1 \neq R_2 C_2$) has a simple pole and a simple zero, as can be deduced from Eq. 5.2.12. If we rewrite the impedances as:

$$Z_1 = \frac{1}{\frac{1}{R_1} + sC_1} = R_1 \frac{\frac{1}{R_1 C_1}}{s - \left(-\frac{1}{R_1 C_1}\right)} = R_1 \frac{-s_1}{s - s_1} \tag{5.2.13}$$

and

$$Z_2 = \frac{1}{\frac{1}{R_2} + sC_2} = R_2 \frac{\frac{1}{R_2 C_2}}{s - \left(-\frac{1}{R_2 C_2}\right)} = R_2 \frac{-s_2}{s - s_2} \tag{5.2.14}$$

where s_1 and s_2 represent the poles in each impedance arm, explicitly:

$$s_1 = -\frac{1}{R_1 C_1} \qquad \text{and} \qquad s_2 = -\frac{1}{R_2 C_2} \tag{5.2.15}$$

The transfer function is then:

$$\frac{v_{out}}{v_{in}} = \frac{Z_2}{Z_1 + Z_2} = \frac{R_2 \frac{-s_2}{s - s_2}}{R_1 \frac{-s_1}{s - s_1} + R_2 \frac{-s_2}{s - s_2}} \tag{5.2.16}$$

By solving the double divisions, the transfer function can be rewritten as:

$$\frac{v_{\text{out}}}{v_{\text{in}}} = \frac{-\,s_2R_2(s-s_1)}{-\,s_1R_1(s-s_2)-s_2R_2(s-s_1)} \tag{5.2.17}$$

We can replace the products $s_{\text{i}}R_{\text{i}}$ by $1/C_{\text{i}}$:

$$\begin{aligned}\frac{v_{\text{out}}}{v_{\text{in}}} &= \frac{\frac{1}{C_2}(s-s_1)}{\frac{1}{C_1}(s-s_2)+\frac{1}{C_2}(s-s_1)} = \frac{C_1(s-s_1)}{C_2(s-s_2)+C_1(s-s_1)} = \\ &= \frac{C_1(s-s_1)}{s(C_1+C_2)-s_2C_2-s_1C_1} = \frac{C_1}{C_1+C_2}\cdot\frac{(s-s_1)}{s-\frac{s_2C_2+s_1C_1}{C_1+C_2}}\end{aligned} \tag{5.2.18}$$

This can be simplified by defining a few useful substitutions: the capacitive divider attenuation:

$$A_C = \frac{C_1}{C_1+C_2} \tag{5.2.19}$$

the system zero:

$$s_{\text{z}} = s_1 = -\,\frac{1}{R_1C_1} \tag{5.2.20}$$

and the system pole:

$$s_{\text{p}} = \frac{s_2C_2+s_1C_1}{C_1+C_2} \tag{5.2.21}$$

Further, the system pole can be rewritten as:

$$\begin{aligned}s_{\text{p}} &= \frac{s_2C_2+s_1C_1}{C_1+C_2} = \frac{-\,C_2\frac{1}{R_2C_2}-C_1\frac{1}{R_1C_1}}{C_1+C_2} = -\,\frac{\frac{1}{R_2}+\frac{1}{R_1}}{C_1+C_2} \\ &= -\,\frac{R_1+R_2}{R_1R_2}\cdot\frac{1}{C_1+C_2}\end{aligned} \tag{5.2.22}$$

From the system pole we note that the system time constant is equal to the parallel connection of all four components.

We will also define the resistance attenuation as:

$$A_R = \frac{R_2}{R_1+R_2} \tag{5.2.23}$$

and we then rewrite the system pole as:

$$s_{\text{p}} = -\,\frac{R_1+R_2}{R_2}\cdot\frac{1}{R_1(C_1+C_2)} = -\,\frac{1}{A_R}\cdot\frac{1}{R_1(C_1+C_2)} \tag{5.2.24}$$

With all these substitutions the complex frequency response is:

$$\boxed{F(s) = \frac{v_{\text{out}}}{v_{\text{in}}} = A_C\,\frac{s-s_{\text{z}}}{s-s_{\text{p}}}} \tag{5.2.25}$$

Again, it is obvious that the frequency dependence vanishes if $s_{\text{p}} = s_{\text{z}}$.

From $F(s)$ we derive the magnitude:

$$M(\omega) = \Big|F(j\omega)\Big| = \sqrt{F(j\omega)\cdot F(-j\omega)} = A_C\sqrt{\frac{j\omega-\sigma_z}{j\omega-\sigma_p}\cdot\frac{-j\omega-\sigma_z}{-j\omega-\sigma_p}} \quad (5.2.26)$$

which results in:

$$\boxed{M(\omega) = A_C\sqrt{\frac{\omega^2+\sigma_z^2}{\omega^2+\sigma_p^2}}} \quad (5.2.27)$$

The phase angle is the arctangent of the imaginary to real component ratio of the frequency response $F(j\omega)$:

$$\varphi(\omega) = \arctan\frac{\Im\{F(j\omega)\}}{\Re\{F(j\omega)\}} = \arctan\frac{\Im\left\{\dfrac{j\omega-\sigma_z}{j\omega-\sigma_p}\right\}}{\Re\left\{\dfrac{j\omega-\sigma_z}{j\omega-\sigma_p}\right\}} \quad (5.2.28)$$

First we must rationalize $F(j\omega)$ by multiplying both the numerator and the denominator by the complex conjugate of the denominator $(-j\omega-\sigma_p)$:

$$\frac{j\omega-\sigma_z}{j\omega-\sigma_p} = \frac{(j\omega-\sigma_z)(-j\omega-\sigma_p)}{(j\omega-\sigma_p)(-j\omega-\sigma_p)} = \frac{\omega^2+j\omega\,\sigma_z-j\omega\,\sigma_p+\sigma_z\sigma_p}{\omega^2+\sigma_p^2}$$

and then we separate the real and imaginary part:

$$\frac{\omega^2+j\omega\,\sigma_z-j\omega\,\sigma_p+\sigma_z\sigma_p}{\omega^2+\sigma_p^2} = \frac{\omega^2+\sigma_z\sigma_p}{\omega^2+\sigma_p^2} + j\,\frac{\omega(\sigma_z-\sigma_p)}{\omega^2+\sigma_p}$$

The phase angle is then:

$$\varphi(\omega) = \arctan\frac{\omega(\sigma_z-\sigma_p)}{\omega^2+\sigma_z\sigma_p} = \arctan\frac{\dfrac{\sigma_z-\sigma_p}{\omega}}{1+\dfrac{\sigma_z\sigma_p}{\omega^2}} \quad (5.2.29)$$

By using the identity:

$$\arctan x - \arctan y = \arctan\frac{x-y}{1+xy}$$

we can write:

$$\boxed{\varphi(\omega) = \arctan\frac{\sigma_z}{\omega} - \arctan\frac{\sigma_p}{\omega}} \quad (5.2.30)$$

With $\sigma_p = \sigma_z$ the phase angle is zero for any ω.

The envelope delay is the frequency derivative of the phase:

$$\tau_d = \frac{d\varphi}{d\omega} = \frac{d}{d\omega}\left(\arctan\frac{\sigma_z}{\omega} - \arctan\frac{\sigma_p}{\omega}\right)$$

$$= \frac{1}{1+\left(\dfrac{\sigma_z}{\omega}\right)^2}\left(-\frac{\sigma_z}{\omega^2}\right) - \frac{1}{1+\left(\dfrac{\sigma_p}{\omega}\right)^2}\left(-\frac{\sigma_p}{\omega^2}\right) \quad (5.2.31)$$

So the result is:

$$\tau_\mathrm{d} = \frac{-\sigma_\mathrm{z}}{\omega^2 + \sigma_\mathrm{z}^2} - \frac{-\sigma_\mathrm{p}}{\omega^2 + \sigma_\mathrm{p}^2} \tag{5.2.32}$$

Again, note that for $\sigma_\mathrm{p} = \sigma_\mathrm{z}$ the envelope delay is zero.

We have plotted the magnitude, phase and envelope delay in Fig. 5.2.3. The plots are made for the matched and two unmatched cases in order to show the influence of trimming the attenuator by C_2 (±10 pF).

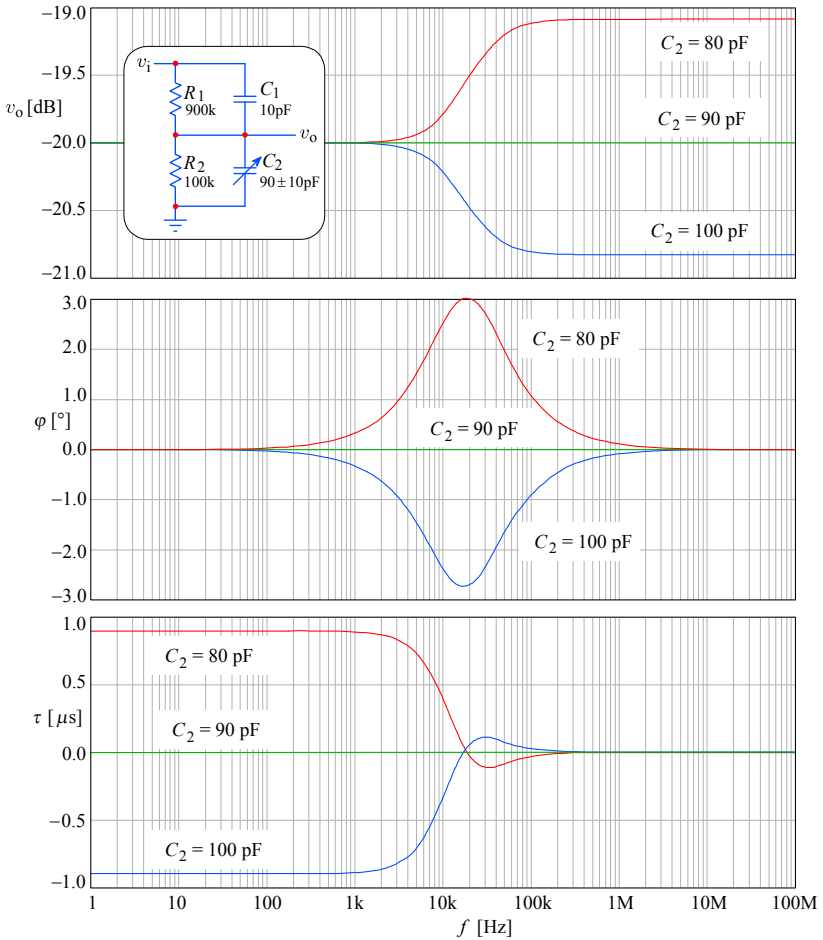


Fig. 5.2.3: The attenuator magnitude, phase, and envelope delay responses for the correctly compensated case (flat lines), along with the under- and over-compensated cases (C_2 is trimmed by ±10 pF). Note that these same figures apply also to oscilloscope passive probe compensation, demonstrating the importance of correct compensation when making single channel pulse measurements and two channel differential measurements.

The step response is obtained from $F(s)$ by the inverse Laplace transform, using the theory of residues:

$$\mathcal{L}^{-1}\left\{\frac{1}{s}\,F(s)\right\} = \frac{A_C}{2\pi j}\oint_C \frac{s-s_\mathrm{z}}{s(s-s_\mathrm{p})}\,\mathrm{e}^{st}\,ds = A_C\sum_{s=0}^{s=s_\mathrm{p}}\mathrm{res}\left\{\frac{s-s_\mathrm{z}}{s(s-s_\mathrm{p})}\,\mathrm{e}^{st}\right\}$$

We have two residues. One is owed to the unit step operator, $1/s$:

$$\begin{aligned}
\mathrm{res}_1 &= A_C \lim_{s\to 0} s\left\{\frac{s-s_\mathrm{z}}{s(s-s_\mathrm{p})}\,\mathrm{e}^{st}\right\} = A_C \lim_{s\to 0}\left\{\frac{s-s_\mathrm{z}}{s-s_\mathrm{p}}\,\mathrm{e}^{st}\right\} \\
&= A_C\,\frac{-s_\mathrm{z}}{-s_\mathrm{p}} = A_C\,\frac{-\dfrac{1}{R_1C_1}}{-\dfrac{1}{A_R}\cdot\dfrac{1}{R_1(C_1+C_2)}} \\
&= A_C\,A_R\frac{R_1(C_1+C_2)}{R_1C_1} = A_C\,A_R\frac{C_1+C_2}{C_1} = A_C\,A_R\,\frac{1}{A_C} \\
&= A_R = \frac{R_2}{R_1+R_2}
\end{aligned} \tag{5.2.33}$$

As expected, the residue for zero frequency (DC) is set by the resistance ratio.

The other residue is due to the system pole, s_p:

$$\begin{aligned}
\mathrm{res}_2 &= A_C \lim_{s\to s_\mathrm{p}} (s-s_\mathrm{p})\left\{\frac{s-s_\mathrm{z}}{s(s-s_\mathrm{p})}\,\mathrm{e}^{st}\right\} = A_C \lim_{s\to s_\mathrm{p}}\left\{\frac{s-s_\mathrm{z}}{s}\,\mathrm{e}^{st}\right\} = \\
&= A_C\,\frac{s_\mathrm{p}-s_\mathrm{z}}{s_\mathrm{p}}\,\mathrm{e}^{s_\mathrm{p}t} = A_C\,\frac{-\dfrac{1}{A_R}\cdot\dfrac{1}{R_1(C_1+C_2)} - \left(-\dfrac{1}{R_1C_1}\right)}{-\dfrac{1}{A_R}\cdot\dfrac{1}{R_1(C_1+C_2)}}\mathrm{e}^{s_\mathrm{p}t} \\
&= A_C\,\frac{\dfrac{1}{A_R}\cdot\dfrac{1}{R_1(C_1+C_2)} - \dfrac{1}{R_1C_1}}{\dfrac{1}{A_R}\cdot\dfrac{1}{R_1(C_1+C_2)}}\mathrm{e}^{s_\mathrm{p}t} \\
&= A_C\left(1 - A_R\frac{C_1+C_2}{C_1}\right)\mathrm{e}^{s_\mathrm{p}t} = A_C\left(1-\frac{A_R}{A_C}\right)\mathrm{e}^{s_\mathrm{p}t} \\
&= (A_C - A_R)\,\mathrm{e}^{s_\mathrm{p}t}
\end{aligned} \tag{5.2.34}$$

The result is a time decaying exponential, with the time constant set by the system pole, s_p, and the amplitude set by the difference between the capacitive and resistive divider.

The step response is the sum of both residues:

$$\begin{aligned}
f(t) &= \sum \mathrm{res} = A_R + (A_C - A_R)\,\mathrm{e}^{s_\mathrm{p}t} \\
&= A_R + (A_C - A_R)\,\mathrm{e}^{-\frac{1}{A_R}\cdot\frac{1}{R_1(C_1+C_2)}t}
\end{aligned} \tag{5.2.35}$$

So the explicit result is:

$$f(t) = \frac{R_2}{R_1 + R_2} + \left(\frac{C_1}{C_1 + C_2} - \frac{R_2}{R_1 + R_2} \right) \mathrm{e}^{-\frac{R_1+R_2}{R_1 R_2} \cdot \frac{1}{C_1+C_2} t} \tag{5.2.36}$$

When $A_C = A_R$, the exponential function coefficient is zero, thus:

$$f(t) = \frac{R_2}{R_1 + R_2} \tag{5.2.37}$$

The system's time constant, as we have already seen in Eq. 5.2.24, is:

$$\tau_\mathrm{a} = -\frac{1}{s_\mathrm{p}} = \frac{R_1 R_2}{R_1 + R_2}(C_1 + C_2) \tag{5.2.38}$$

For a well compensated attenuator, the following is true:

$$\tau_\mathrm{a} = \frac{R_1 R_2}{R_1 + R_2}(C_1 + C_2) = R_1 C_1 = R_2 C_2 \tag{5.2.39}$$

We have plotted the step response in Fig. 5.2.4. The plots are made for the matched and two unmatched cases in order to show the influence of trimming the attenuator by C_2 (±10 pF), as in the frequency domain plots.

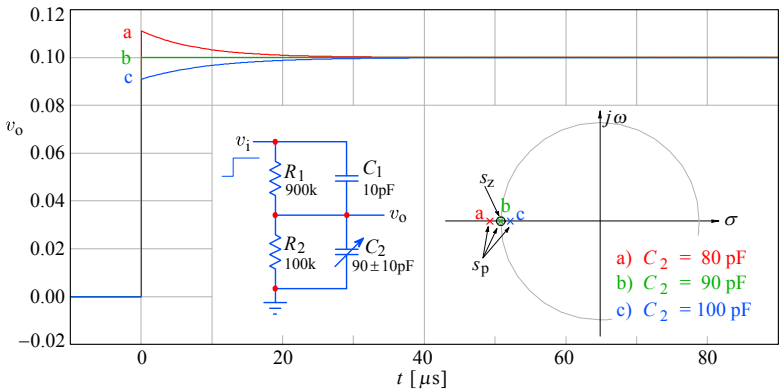


Fig. 5.2.4: The attenuator's step response for the correctly compensated case, along with the under- and over-compensated cases (C_2 is trimmed by ±10 pF). Note that by changing C_2 the system pole also changes but the system zero remains the same.

Now we are going to analyze the influence of the non-zero source impedance on the transfer function. Since we can reuse some of the results we shall not need to recalculate everything.

The capacitive divider presents a relatively high output capacitance to the following amplifier, and the amplifier input capacitance appears in parallel with C_2, changing the division slightly, but that is compensated by trimming.

However, the attenuator input capacitance C_a (Eq. 5.2.10) is smaller than C_1 (for an attenuation of $A = 10$, the input capacitance is $C_\mathrm{a} = \frac{9}{10} C_1$). The actual values

of C_1 and C_2 are dictated mainly by the need to provide a standard value for various probes (compensated attenuators themselves, too). Historically, values between 10 and 20 pF have been used for C_a. Although small, this load is still significant if the signal source internal impedance is considered.

High frequency signal sources are designed to have a standardized impedance of $R_g = 50\,\Omega$ (75 Ω for video systems). The cable connecting any two instruments must then have a characteristic impedance of $Z_0 = 50\,\Omega$, and it must always be terminated at its end by an equal impedance in order to prevent signal reflections. As shown in Fig. 5.2.5a and 5.2.5b, the internal source resistance R_g and the termination resistance R_t form a ÷2 attenuator (neglecting the 1 MΩ of the 10:1 attenuator):

$$\frac{v_o}{v_g} = \frac{R_t}{R_g + R_t} = \frac{50\,\Omega}{50\,\Omega + 50\,\Omega} = \frac{1}{2} \tag{5.2.40}$$

Therefore the effective signal source impedance seen by the attenuator is:

$$R_{ge} = \frac{R_g R_t}{R_g + R_t} = \frac{2500}{100} = 25\,\Omega \tag{5.2.41}$$

With a 9 pF equivalent attenuator input capacitance ($C_a = 0.9\,C_1$) a pole at $s_h = -1/R_{ge}C_a$ is formed, resulting in an $f_h = 1/(2\,\pi\,R_{ge}C_a) = 707$ MHz cut off.

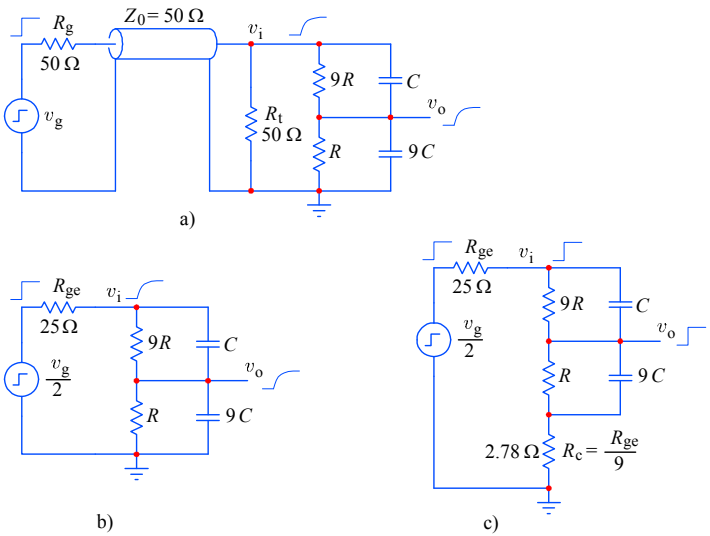


Fig. 5.2.5: **a)** When working with 50 Ω impedance the terminating resistance must match the generator internal resistance, forming a ÷2 attenuator with an effective output impedance of 25 Ω; **b)** With a 9 pF attenuator input capacitance, a HF cut off at 707 MHz results; **c)** The cut off for the ÷10 attenuator can be compensated by a 25/9 Ω resistor between the lower end of the attenuator and the ground.

Owing to the high resistances involved, we can neglect the attenuator's resistive arms and consider only the equivalent signal source impedance and the capacitive divider (assuming the attenuator is correctly compensated). If $F_0(s)$ is the

attenuator transfer function for a zero signal source impedance (Eq. 5.2.25), the transfer function for the impedance $R_{ge} = 25\,\Omega$ is:

$$F_1(s) = \frac{1}{2} R_{ge} \cdot \frac{-\dfrac{1}{R_{ge}C_a}}{s - \left(-\dfrac{1}{R_{ge}C_a}\right)} \cdot F_0(s) \qquad (5.2.42)$$

The resulting pole s_h can be compensated by inserting a resistor R_c of $25/9 = 2.78\,\Omega$ between the lower attenuator end and the ground, as in Fig. 5.2.5c; with the equivalent input resistance of $R_g = 25\,\Omega$, the R_c provides the 10:1 division at highest frequencies, as required. The effective output resistance at these frequencies is only $2.5\,\Omega$, so the pole formed by it and the unity gain buffer total input capacitance of, say, 2 pF would be well beyond the bandwidth of interest (~32 GHz). However, the bandwidth of the passive part of the input circuitry is impaired by other effects, as we shall soon see.

A typical attenuator consists of three switch selectable sections, 1:1, 10:1, and 100:1, as shown in Fig. 5.2.6. This allows us to cover the required input amplitude range from millivolts to tens of volts (intermediate 2:1 and 5:1 attention settings are usually implemented after the input buffer stage with low value resistors). Because the attenuator's compensation is adjustable, the input capacitance changes, so it has to be 'standardized' by an additional compensating capacitor to make the input capacitance of all sections equal.

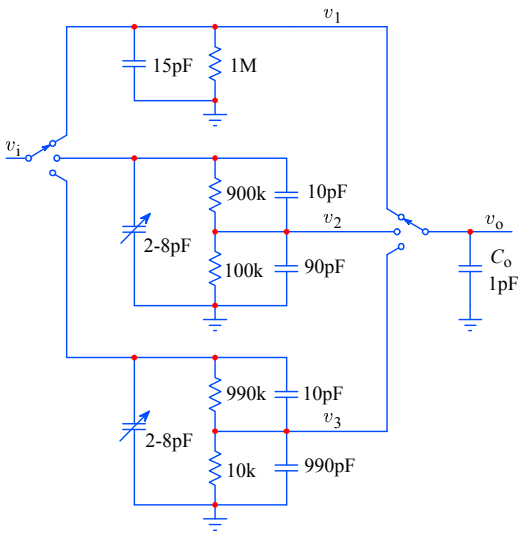


Fig. 5.2.6: The direct and the two attenuation paths are switched at both input and output, in order to reduce the input capacitance. For low cross-talk, the input and output of each unused section should be grounded (not shown here). The variable capacitors in parallel with the two attenuation sections are adjusted so that the input capacitance is equal for all settings. Of course, other values are possible, e.g., ÷1, ÷20, ÷400 (as in Tek 7A11), with the highest attenuation achieved by cascading two ÷20 sections. The advantage is that the parasitic serial inductance of the largest capacitance in the highest attenuation section is avoided; a disadvantage is that it is very difficult to trim correctly.

Unfortunately, for attenuator settings other than 10:1 the resistive compensation shown in Fig. 5.2.5 can be very difficult to achieve. For a 100:1 attenuation the resistance required between the attenuator and the ground would be only 0.2525 Ω and such a low value might be present in the circuit already (in the form of a ground return trace if it is not wide enough). Even if we are satisfied by a 1% tolerance we still need a well controlled design, taking care of every 10 mΩ.

The most critical is the direct (1:1) path. In order to present the same input load to an external 10:1 probe this path must have the same input resistance and capacitance as the higher attenuation paths. However, the direct path has no attenuation, so **the direct path can not be compensated** by a low value resistor.

It is for this reason that many designers avoid using the direct path altogether and opt for a 2:1 and 20:1 combination instead. Such a circuit, showing also the resistive compensation, is drawn in Fig. 5.2.7. As a bonus, the amplifier input current limiting and input overdrive protection circuitry is easier to realize (no 1:1 path), needing lower serial impedance and thus allowing higher bandwidths.

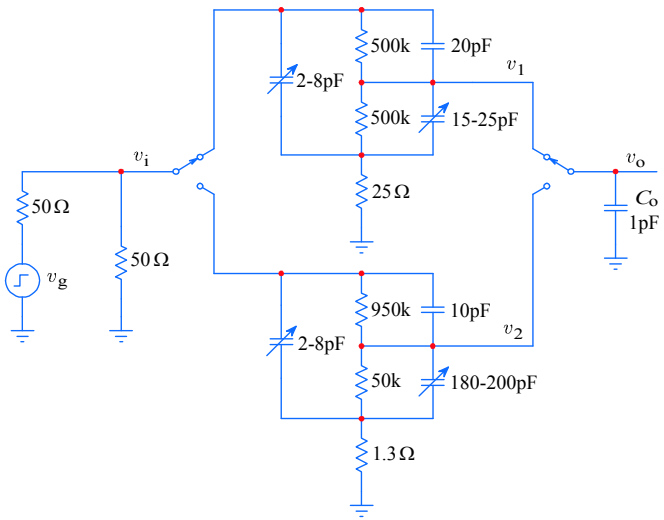


Fig. 5.2.7: The attenuator with no direct path, in which the 25 Ω effective source impedance compensation can be used for both settings. Low ground return path impedance is necessary.

On the negative side, by using the ÷2 and ÷20 attenuation, the amplifier must provide for another gain of two, making the system optimization more difficult. Fortunately, for modern amplifiers, driving an AD converter, the gain requirement is low, since the converter requires only a volt or two for a full range display; in contrast, a conventional 'scope CRT requires tens of volts on the vertical deflecting plates. Thus a factor of at least 10 in gain reduction (and a similar bandwidth increase!) is in favor to modern circuits.

Whilst the gain requirements are relaxed, modern sensitive circuits require a higher attenuation to cover the desired signal range. But obtaining a 200:1 attenuation can be difficult, because of capacitive feed through: even a 0.1 pF from the input to the

buffer output, together with a non-zero output impedance, can be enough to spoil the response. If we can tolerate a feed through error of one least significant bit of an 8 bit analog to digital converter, the 200:1 attenuator would need an effective isolation of $20\log_{10}(200\times 2^8) = 94\,\text{dB}$, which is sometimes hard to achieve even at audio frequencies, let alone GHz. A cascade of two sections could be the solution.

5.2.2 Attenuator Inductance Loops

Designer's life would be easy with only resistances and capacitances to deal with. But every circuit also has an inductance, whether we intentionally put it in or desperately try to avoid it. As we have learned in Part 2, in wideband amplifiers, instead of trying to avoid the unavoidable, we rather try to put the inductance to use by means of fine tuning and adequate damping.

In Fig. 5.2.8 we have indicated the two inductances associated with the attenuator circuit. Because of the high voltages involved, the attenuator circuit can not use arbitrarily small components, packed arbitrarily close together. As a consequence, the circuit will have loop dimensions which can not be neglected and, since the inductance value is proportional to the loop area, the inductance values can be relatively large (for wideband amplifiers).

As for stray capacitance, the value of stray inductance can not be readily predicted, at least not to the precision required. Each component in Fig. 5.2.8 will have its own stray inductances, one associated with the internal component structure and the other associated with the component leads, the soldering pads, and PCB traces. These will be added to the loop inductance.

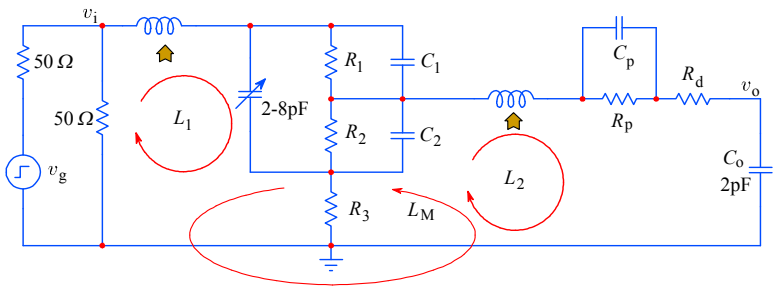


Fig. 5.2.8: Inductances owed to circuit loops can be modeled as inductors in series with the signal path. Note that in addition to the two self inductances there is also a mutual inductance between the two. The actual values depend on the loop's size, which in turn depends on the size of the components and the circuit's layout. Smaller loops have less inductance. Mutual inductance can be reduced by shielding, although this can increase the stray capacitances.

Nevertheless, it is relatively easy to estimate both loop inductances, at least to an order of magnitude. Basically, a single loop current I causes a magnetic flux Φ with a density B within the loop area S, so the self inductance is:

$$L = \frac{\Phi}{I} = \frac{BS}{I} = \frac{\mu HS}{I} = \frac{\mu_0\,\mu_r\,HS}{I} \qquad (5.2.43)$$

The current I and the magnetic field strength H are proportional: $H = I/2r$ for a single loop, where r is the loop radius. In a linear non-magnetic environment (with the relative permeability $\mu_r = 1$) I and B are also proportional because $B = \mu H$. Furthermore, μ_0 is the free space magnetic permeability, also known as the 'induction constant', the value of which has been set by the SI agreement about the Ampere: $\mu_0 = 4\pi \cdot 10^{-7}$ [V s A^{-1} m^{-1}]. This means that a current of 1 A encircling once a loop area of 1 m^2 causes a magnetic field strength of 1 Vs. Because for a circular loop $S = \pi r^2$, our loop inductance equation can be reduced to:

$$L = \frac{\mu_0 A}{2r} = \frac{\mu_0 \pi r^2}{2r} = \frac{\mu_0 \pi r}{2} = \mathrm{k}\, r \qquad (5.2.44)$$

where $\mathrm{k} = 2\pi^2 \cdot 10^{-7}$ H/m. The inductance of a 1 m^2 loop ($r = 0.5642$ m) is then $\approx 1.14 \cdot 10^{-6}$ H (the unit of inductance is 'henry', after *Joseph Henry*, 1791–1878; [H] = [V s A^{-1}]).

As a more practical figure, a loop of 10 cm^2 (≈ 0.0178 m radius circle) has an inductance of ≈ 35 nH. This does not look much, but remember that in our circuit the loop inductance L_1 is effectively in series with the signal source and is loaded by the attenuator's input capacitance, forming a 2nd-order low pass filter with a cut off frequency $f_h = 1/\left(2\pi\sqrt{L_1 C_a}\right) \approx 268$ MHz, assuming $L_1 = 35$ nH and $C_a = 10$ pF. With such values the step response rings long, since the equivalent signal source resistance ($R_{ge} = 25\,\Omega$) is not high enough to damp the resonance (such damping would be adequate for $L_1 < 5$ nH).

The above inductance estimation is based on a circular loop model, whilst our loops will usually be of a square form (thus increasing L), with additional stray inductances owing to the internal geometry of the components (capacitors) and their leads, or just the PCB traces if surface mounted components are used.

The loop inductances can, of course, be measured. If we replace C_1, C_2 and R_3 (Fig. 5.2.8) by a wire of the same total length, the input resistance and L_1 form a high pass filter, whose cut off frequency can be measured. Next, by removing the wire and also R_3 and replacing R_p, R_d and C_o by another wire, we can measure $L_1 + L_2$. Obtaining L_2 is then a matter of simple subtraction. Finally, by applying a signal to the input, shorting R_3, and measuring the signal induced in L_2, we can calculate the mutual inductance L_M. Note that a thin wire will have a somewhat larger inductance than a wide PCB trace.

The best way to reduce the loop area (and consequently L) is to use a 3-layer PCB, and make the middle layer a 'ground plane'. In addition, using surface mounted components reduces the circuit size and also allows us to place them on both sides of the board. However, this technique also increases the stray capacitances and can also cause reflections if the 'microstrip' trace impedances are not well matched to the circuit. Therefore, a careful PCB design is needed, with wider ground clearance around sensitive pads and using a material with low ε_r. The most sensitive node in this respect is the attenuator output.

Another way of reducing the effect of stray inductance is to employ the same technique as we did for the low value resistors. This means that the inductance in the ground path (the signal return path) should not be too small, as it would be in the ground plane case; rather, the return path inductance should be kept in the same ratio

to the L_1 as the attenuation ratio. Precision in this respect is difficult, but not impossible to achieve. Our inductance expression Eq. 5.2.44 does not show it, but inductance is also inversely proportional to trace width. Powerful finite element numerical simulation routines will be required for the job.

However, the same trick can not be used for L_2 (no attenuation in this loop!). Fortunately, as will become clear from the analysis below, the input inductance L_1 is more critical than L_2, since the latter is loaded by a much smaller capacitance (C_o) and can be suitably damped with a larger resistance (R_d, which is already in the circuit because it is required for the FET gate protection).

We shall analyze the attenuator loops by assuming perfectly matched time constants, $R_1C_1 = R_2C_2$, matched also to the other attenuator paths, so that the variable capacitor in parallel is not needed. Also, we shall replace the two 50 Ω resistors with a single 25 Ω one, representing the effective signal source resistance R_s in series with the input, with $v_i = v_g/2$. The loop inductances are represented by discrete components, $L_1$and L_2 in the forward signal paths, as drawn in Fig. 5.2.8.

In the second loop the first thing to note is that C_2 is many times larger than C_o (10–500 ×, depending on the attenuation setting) and the same is true for C_p, which means that their reactance will be comparably low and can thus be neglected. Likewise, the resistances R_2 and R_p in parallel with these capacitances are large in comparison with their reactances. What remains is the loop inductance L_2 in series with $R_d + R_3$, driving the amplifier input capacitance C_o. If the attenuated input voltage is v_i/A, the output voltage will be:

$$v_o = \frac{v_i}{A} \cdot \frac{1}{sC_o} \cdot \frac{1}{sL_2 + R_d + R_3 + \dfrac{1}{sC_o}} \tag{5.2.45}$$

So we have a 2nd-order transfer function:

$$F_2(s) = \frac{v_o}{v_i} = \frac{1}{A} \cdot \frac{\dfrac{1}{L_2C_o}}{s^2 + s\dfrac{R_d + R_3}{L_2} + \dfrac{1}{L_2C_o}} \tag{5.2.46}$$

Since R_3 is fixed and of quite low value, R_d is used to provide the desired damping.

The input loop analysis is similar. Here we have the equivalent source resistance $R_s + R_3$ in series with L_1, driving the equivalent input attenuator capacitance C_a (Eq. 5.2.9; the attenuator resistance $R_1 + R_2$ can be neglected at high frequencies). At the top of the attenuator we have:

$$v_i = \frac{v_g}{2}\left(R_3 + \frac{1}{sC_a}\right)\frac{1}{sL_1 + R_s + R_3 + \dfrac{1}{sC_a}} \tag{5.2.47}$$

which results in the following second-order transfer function:

$$F_1(s) = \frac{2\,v_i}{v_g} = \frac{\dfrac{1}{L_1C_a} + s\dfrac{R_3}{L_1}}{s^2 + s\dfrac{R_s + R_3}{L_1} + \dfrac{1}{L_1C_a}} \tag{5.2.48}$$

The numerator can be written as:

$$\frac{1}{L_1 C_a} + s\,\frac{R_3}{L_1} = \frac{1}{L_1 C_a}(1 + sC_a R_3) \tag{5.2.49}$$

It is clear that the frequency of the zero, $1/C_a R_3$, is much higher than the frequency of the pole pair, $1/\sqrt{L_1 C_a}$. Also, if $L_1 \approx L_2$ and C_a is at least 5 to 10 times larger than C_o, then C_a will dominate the response. Fortunately, as discussed above, with a clever layout of components and a suitable ground plane, L_1 can be broken into L_{1a} and L_{1b}, so that L_{1b} is in the ground return path. If we can make $L_{1a} = 9L_{1b}$ we would achieve an effective inductance compensation in this loop.

We are thus left with the L_2 loop and its transfer function, Eq. 5.2.49. However, this 2nd-order function will be transformed by the pole of the JFET source follower into a 3rd-order function, owing to its capacitive loading.

Although the inductance is always caused by a current loop, the inductance of a straight PCB trace can be estimated as some 7–10 nH/cm (length), depending on the trace width. In [Ref. 5.16] a good empirical approximation is offered:

$$L = 0.2\,l\left[\ln\left(\frac{2\,l}{w+h}\right) + 0.2235\left(\frac{w+h}{l}\right) - 0.5\right] \tag{5.2.50}$$

where the trace length l, width w and thickness h are all in mm, resulting in the inductance in nH (no ground plane in this case!). With surface mounted components, by using capacitors with low serial inductance, and using miniature relay switches in the attenuator, the inductance L_2 can be reduced to less than 10 nH, making the pole (pair) at $1/\sqrt{L_2 C_o}$ high, compared to the source follower real pole (set by the damping resistance R_d and the source follower loading capacitance C_L (see the JFET source follower discussion in Part 3, Sec. 3.9). However, by making L_2 somewhat larger, say, 30–50 nH, we can achieve a 3rd-order Bessel pole pattern, improving the bandwidth and reducing the rise time. In Fig. 5.2.9 we see the attenuator circuit of the $A = 10$ section, followed by a JFET source follower.

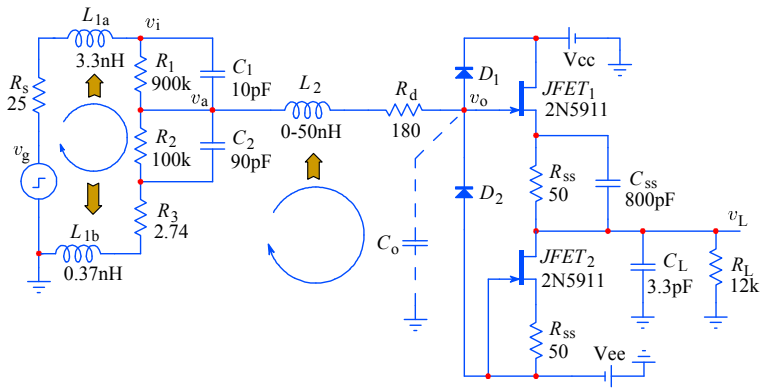


Fig. 5.2.9: The attenuator and the source follower JFET$_1$ (JFET$_2$ acts as a constant current source bias for JFET$_1$). The input loop inductance L_1 should be low, but the attenuation can be compensated by L_{1b}. The inductance L_2 of the second loop can be 'tuned' and damped by an appropriate value of R_d to provide a Bessel step response, as seen in Fig. 5.2.10.

Note that here we have not drawn the protecting components C_p and R_p, but since a 325 V (peak value of the 230 V AC-mains) at the input results in a 32.5 V at the attenuator output, these components are absolutely necessary. Also, C_p should be a high voltage type (500 V), in order to survive the 325 V in the direct path (and still 163 V for a ÷2 attenuator); therefore, it will be of larger dimensions, so its internal serial inductance will have to be taken into account.

Note also that for high bandwidth a low value of C_o must be ensured. Since the negative input impedance compensation network (as in Part 3, Sec. 3.9), as well as R_d, D_1, D_2, C_{GD}, and C_L are present at the v_o node, C_o will tend to be high.

We have analyzed the step response in Fig. 5.2.10 for two values of L_2 (10 and 50 nH; R_d has been chosen for a correct Bessel damping).

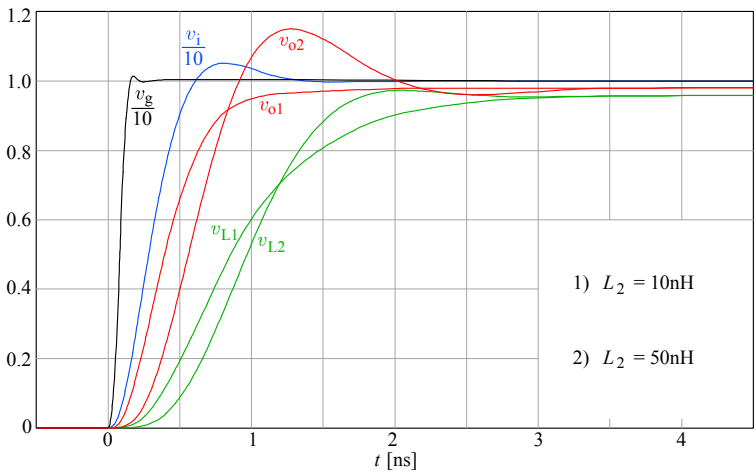


Fig. 5.2.10: Step response of the circuit in Fig. 5.2.9. With a low L_1, a correctly damped L_2, and a good JFET, a 350 MHz bandwidth (v_{L2} rise time $\approx$ 1 ns), can be easily achieved. The source follower gain is a little less than one. v_o and v_L are drawn for the two L_2 cases.

5.2.3 The 'Hook–Effect'

The discussion about high impedance attenuators would not be complete without mentioning the so called 'hook–effect'. The name springs from the shape of the step response signal, which, owing to a sag in the 10–300 kHz region, resembles a hook at slower time base values (Fig. 5.2.11). The effect is caused by the frequency dependent relative permittivity, ε_r, of the PCB material (the standard glass epoxy FR4, FR stands for 'flame resistant', has an average $\varepsilon_r = 4.5$, but it changes with frequency and temperature considerably).

The capacitance in farads of a parallel plates capacitor is expressed as:

$$C = \varepsilon_0 \, \varepsilon_r \, \frac{S}{d} \qquad [\mathrm{F}] = [\mathrm{As/V}] \qquad (5.2.51)$$

where S is the plate area [m^2], d is their distance [m], $\varepsilon_0 = 8.85 \times 10^{-12}$ [As/Vm] is the permittivity of the free space (vacuum) and ε_r is the relative permittivity of the

dielectric between the plates. A pad on the PCB thus has some small stray capacitance towards the ground (large if a ground plane is used). This capacitance changes with frequency proportionally with ε_r. Also, the material is porous and the fibres are long, extending to the edge of the board, allowing moisture in (water $\varepsilon_r = 80$), which causes long term changes. The problem is not specific to this material only, it is encountered with all traditional PCB materials (as well as many other insulators).

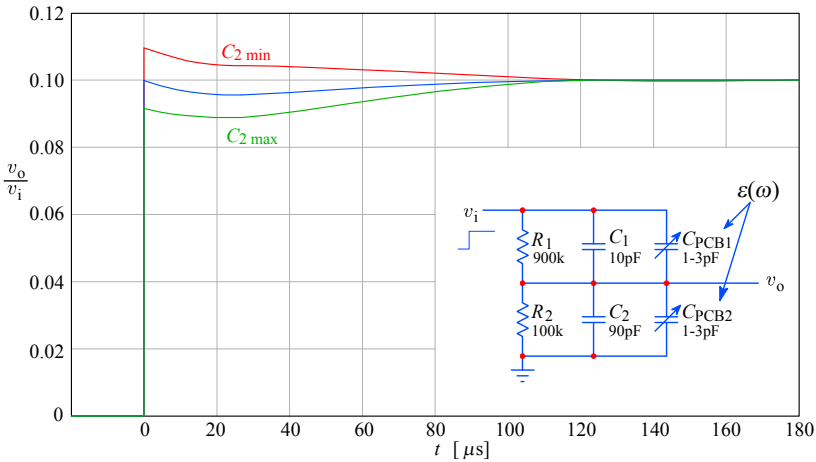


Fig. 5.2.11: The 'hook–effect' is most noticeable in the frequency range 10–300 kHz. Because the relative permittivity, ε_r, of a common PCB material is not exactly constant with frequency, the high impedance attenuator will exhibit a hook in its step response, which can not be trimmed out by the usual adjustment of C_2. The C_{PCB} stray capacitance can vary by some 10–30 %, depending on the actual topology involved. Since C_1 is small, it is affected by a few percent. The lower attenuator leg is less affected, due to a larger value of C_2.

To solve this problem, special Teflon® based material is used for instrument front end, but it is expensive and not readily available. If it can not be obtained, one possible solution could be to implement two large pads on a two sided PCB, in parallel to C_1 and C_2, with their areas in the same ratio as the attenuation factor required [Ref. 5.68]. Then, the effect would be equally present in both legs, canceling out the hook, Fig. 5.2.12. Even some trimming can be done by drilling small holes in the larger pad pair (in contrast to cutting a pad corner, drilling removes the dielectric, thus lowering both A and ε).

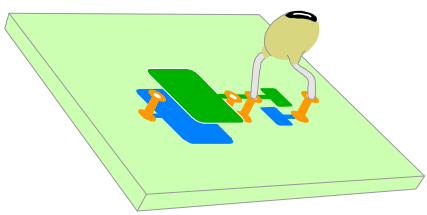


Fig. 5.2.12: Canceling the hook–effect in the common PCB material is achieved by intentionally adding two capacitances in form of large PCB pads, with areas in the same ratio as required by the attenuation (since the area is proportional to the square of the linear dimensions, for a 9:1 capacitance ratio, a 3:1 dimension ratio is needed). Trimming is possible by drilling small holes in the larger pad.

The main problem with this solution is that the use of external probes will expose the hook again, although to a lesser extent (owing to the large capacitance of the probe compensation).

5.2.4 Improving the JFET Source Follower DC Stability

The DC performance of a JFET source follower is far from perfect. Even if we use a dual JFET in the same case and on the same substrate, i.e., the Siliconix 2N5911 as in Fig. 5.2.9, their characteristics will not match perfectly. The 2N5911 data sheet state a V_{GS} mismatch of 10 mV maximum and a temperature drift of 20 µV/K. The circuit in Fig. 5.2.13 offers moderate DC stability; the resistor R_T is trimmed for a zero v_L to v_{in} DC offset.

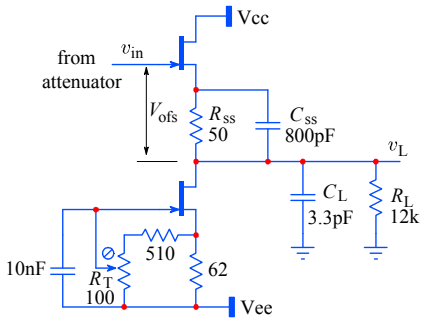


Fig. 5.2.13: Simple offset trimming of a JFET source follower.

Traditionally, oscilloscopes have a 'vertical position' control on the front panel (one for each channel), which is adjusted by the user in accordance with the particular measurement conditions, which differ from one situation to another, so the offset and drift (if not too high) are not of particular concern.

However, in modern instrumentation some additional features are becoming important, such as automated measurement, where we can not rely on the presence of a human operator to make adjustments every so often. Also it is not uncommon to find digital oscilloscopes with an 8 bit resolution (1:256) at high sampling rates, but capable of 12 bit (1:4096) or even 16 bit (1:65536) resolution at low speed, and in digital equipment it is expected that DC errors are of the order of ±1 LSB.

By trimming the current source, we reduce the DC offset, but the temperature drift will remain. The gate current of a JFET, although normally in the < 100 pA range, is also temperature dependent and approximately doubles with every 10 K. The source follower input sees an attenuation dependent source resistance (from 1 MΩ to 10 kΩ), so an additional offset component will be present owing to the gate current and the attenuator output impedance. A typical oscilloscope input has a maximum sensitivity of 5 mV/div., or a 40 mV full screen range; the 1 LSB resolution for an 8 bit sampling is $40/256 = 0.15$ mV, therefore the simple trimming circuit is inadequate for digital equipment, and an active offset compensation technique is required to keep the DC error below some 200 µV.

Basically, there are three ways of achieving a low DC error, each having its own advantages and drawbacks. While DC performance is not of primary interest in this book, it should be implemented so that high frequency performance is preserved.

The first technique is suitable for microprocessor controlled equipment, where the input can be temporarily switched to ground, the offset measured, and the error either adjusted by a digital to analog converter or subtracted from the sampled signal in memory. But this operation should not be repeated too often or take a considerable amount of time, otherwise the equipment would be missing valid trigger events or, worse still, introduce errors by loading and unloading the signal source with the instrument's input impedance. This is a rather inelegant solution and it should be taken as the last resort only.

A better way, shown in Fig. 5.2.14, is to use a good differential amplifier to monitor the difference between the Q_1 gate and the output, integrate it and modify the Q_2 current to minimize the offset. Note that this technique works well only while the input is within the linear range of the JFET; when in the non-linear range or when overdriven, the integrator will develop a high error voltage, which will be seen as a long 'tail' after the signal returns within the linear range. Also, owing to the presence of R_1 and R_2, the attenuator lower arm resistors will need to be readjusted.

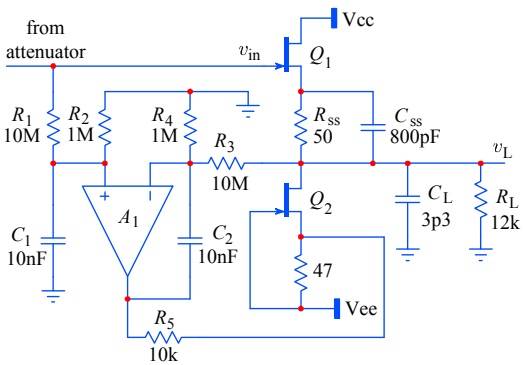


Fig. 5.2.14: Active DC correction loop. The amplifier A_1 amplifies and integrates the difference between the Q_1 gate and the output, driving through R_5 the source of Q_2 and modifying its current to minimize the offset. The resulting offset is equal to the offset of A_1, multiplied by the loop gain $(1 + R_3/R_4)$. The differential amplifier with a very low offset will usually have its input bias current much larger than the JFET input current, therefore resistors R_2 and R_4 provide a lower impedance to ground. C_2 is the integration capacitor, whilst C_1 provides an equal time constant to the non-inverting input. The feedback divider, R_3 and R_4 should be altered to compensate for the system gain slightly lower than one (this is achieved by adding a suitably low value resistor in series with R_4). For a low error the amplifier A_1 must have a high common mode rejection up to the frequency set by C_2 and $R_3||R_4$.

But the most serious problem is owed to the amplifier A_1: in order to minimize the system offset it should have both low offset and low input bias current itself. Although A_1 can be a low bandwidth device, the low input error requirements can easily put us back to where we started from.

The example in Fig. 5.2.14 is relatively simple to implement. However, for a low error we must keep an eye on several key parameters. Ideally we would like to get

rid of the resistor R_2 (and R_4) to avoid the DC path to ground, because it alters the attenuator balance.

Unfortunately, the input common mode range of the error amplifier is limited and, more importantly, amplifiers with a low DC offset are usually made with bipolar transistors at the input, so their input bias current can be in the nA range, much higher than the JFET gate's leakage (< 20 pA). The bias current would then introduce a high DC offset over R_1 (and R_3). Here R_2 and R_4 come to the rescue, by conducting the large part of the bias current to ground over their lower resistance. On the other hand, the amplifier input offset voltage is then effectively amplified by the DC loop gain, $1 + R_3/R_4$. The amplifier is selected so that the total offset error is minimized:

$$V_{\text{ofs}} = \left(1 + \frac{4\,\Delta R}{R}\right)\left(V_{\text{A}_1\text{ofs}}\,\frac{R_3 + R_4}{R_4} + I_{\text{A}_1\text{ofs}}\,\frac{R_1 R_2}{R_1 + R_2}\right) \qquad (5.2.52)$$

where $\Delta R/R$ is the nominal resistor tolerance and $V_{\text{A}_1\text{ofs}}$ and $I_{\text{A}_1\text{ofs}}$ are the amplifier's voltage and current input offset, respectively.

An industry standard amplifier, the OP-07, has $V_{\text{ofs}} = 30\,\mu\text{V}$ and $I_{\text{ofs}} = 0.4\,\text{nA}$ typical, so by taking the resistor values as in Fig. 5.2.14 (with a 1 % tolerance), we can estimate the typical total system offset to be within $\pm 728\,\mu\text{V}$, which is slightly larger than we would like. The offset can be reduced using a chopper stabilized amplifier, such as the Intersil's ICL-7650 or the LTC-1052 from Linear Technology, which have a very low voltage offset (< 5 µV) and low current offset (< 50 pA), but their switching noise must be filtered at the output; also their input switches are very delicate and must be well protected from over-voltage. Therefore, we can not do without R_2 and R_4 and consequently the attenuator must be corrected by increasing the lower resistance appropriately. See [Ref. 5.2] for more examples of such solutions.

The third technique involves separate low pass and high pass amplifier paths and summing their outputs.

The example in Fig. 5.2.15 is made on the assumption that the sum of the two outputs restores the original signal in both phase and amplitude. As the readers who have tried to build loudspeaker crossover networks will know from experience, this can be done correctly only for simple, first-order RC filters (with just two paths; for higher order filters a third, band pass path is necessary).

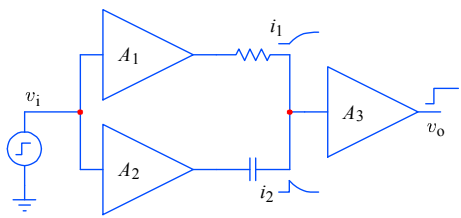


Fig. 5.2.15: The principle of separate low pass and high pass amplifiers.

Here the main problem is with the input of the low pass amplifier A_1, which must have an equally low input bias current as the high pass A_2, but should also have a very low voltage offset. Although in A_1 we don't need to worry about the high

frequency response, we are essentially again at the start, since JFETs and MOSFETs, which have low input current, have a high offset voltage and vice versa for the BJTs.

But we can combine Fig. 5.2.14 and 5.2.15, and, by putting the RC network in front of the source follower, we can eliminate the amplifier A_2. Fig. 5.2.16 shows a possible implementation.

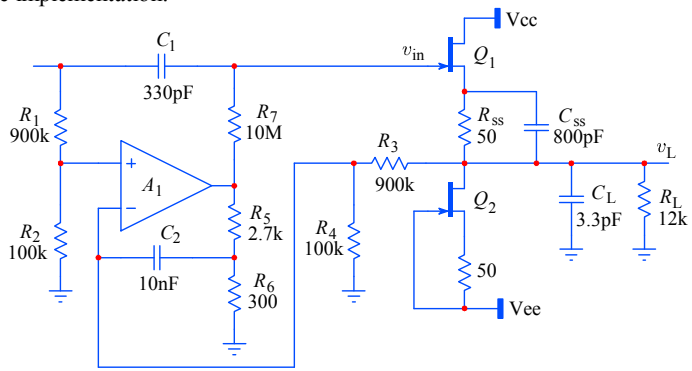


Fig. 5.2.16: With this configuration we can eliminate the need for a separate high pass amplifier. The DC correction is now applied to the Q_1 gate through R_7. The error integrating amplifier A_1 must have a gain of 10 in order to compensate for the $1 + R_1/R_2$ and $1 + R_3/R_4$ attenuation. Resistors R_1 and R_2 now provide the 1 MΩ input impedance for all attenuation settings, and this requires the compensated attenuators in front to be corrected accordingly.

Furthermore, instead of using a single differential amplifier we can invert the output by another low offset amplifier and rearrange the error amplifier into an inverting integrator, as in Fig. 5.2.17. We can also self bias Q_1 by bootstrapping the resistor R_7. This increases the input impedance by a very large factor, allowing us to reduce C_1 and thus further limit the current under overload conditions. However, be aware of the possibility of leakage currents from the protection diodes and the JFET gate itself, now that its DC input impedance has been increased.

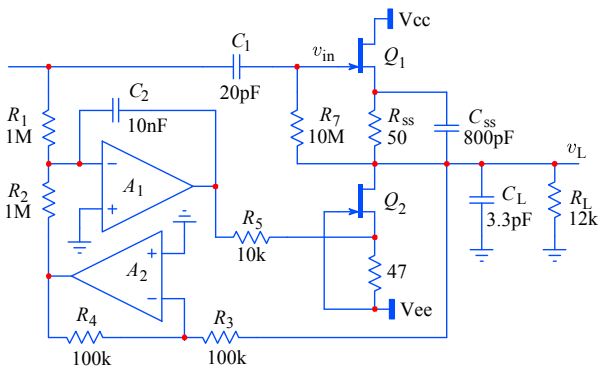


Fig. 5.2.17: By inverting the output the error amplifier becomes an inverting integrator and the offset correction is independent from the attenuator settings. The bootstrapping of R_7 produces an effective input resistance of about 2.4 GΩ.

Of course, now the DC error correction path must be returned to the current source Q_2. The input resistor R_1 must be increased to 1 MΩ, since now the input of A_1 is at the virtual ground; likewise R_2 must be equal to R_1. Note that both A_1 and A_2 offsets add to the final DC error.

Further evolution of this circuit is possible by combining a DC gain switching (R_2 or R_4 adjusting) with input attenuation. A very interesting result has been described in [Ref. 5.2], where also all input relays have been eliminated (using 3 source followers with the switching at their supply voltages by PIN diodes).

5.2.5 Overdrive Recovery

The integration loop will reduce the DC error only if the output follows the input. However, under a hard overdrive the JFET will saturate and the integrator will build up a charge proportional to the input overdrive amplitude and duration. When the overdrive is removed, the loop will reestablish the original DC conditions, but with the integration time constant, so the follower will exhibit a very long 'tail'.

This is one of the most annoying properties of modern instrumentation, because we often want to measure the settling time of an amplifier and a convenient specification is the time from start of the transient to within 0.1 % of the final value. With a good old analog 'scope we would simply increase the vertical sensitivity and adjust the vertical position so that the final signal level is within the screen range. But with modern DC compensated circuits this is not possible, and in order to avoid the post-overdrive tail we must use a specially built external limiter, [Ref. 5.6], to keep the input signal within the linear range of the 'scope. The quality and speed of this limiter will also influence the measurement.

Note that simple follower circuits, like the one in Fig. 5.2.13, would also exhibit a small but noticeable post-overdrive tail, mainly owed to thermal effects. Also, high amplitude step response can be nonlinear, as shown in Fig. 5.2.18, owing to the variation of the JFET gate to channel capacitance with voltage (Eq. 5.2.18–19), but the time constant involved here is relatively small.

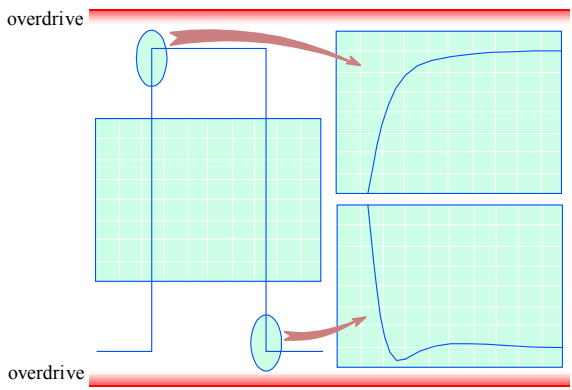


Fig. 5.2.18: Step response for large signals (but still below overdrive) is nevertheless nonlinear, caused by the variation of the JFET gate–drain capacitance with voltage.

5.2.6 Source Follower with MOSFETs

For a very long time, ever since semiconductors replaced electronic tubes in instrumentation, JFETs were the only components used for the source follower input section. Even today, JFETs outshine all other components in all performance aspects but one — shear speed. Unfortunately, BJT input impedance is much too low for the 1 MΩ required. And MOSFETs, although having higher DC input resistance than JFETs, can have (depending on their internal geometry) higher input leakage current, are notoriously noisy, and their gate is easily damaged by overdrive.

If, however, we are ready to accept the design challenge to help the MOSFET by external circuitry, we might be rewarded with a faster follower. Also MOSFETs lend themselves nicely to integration, and this is where the experience gained from the design of high speed digital circuits can help. Circuit area reduction minimizes the stray capacitance and inductance, and new IC processing and semiconductor materials (e.g., GaAs, SiGe) increase charge mobility.

Note that for source follower applications a depletion type MOSFET is needed in order to achieve the required drain–source conductance with zero gate–source voltage. With appropriate doping, the supply voltage can be reduced to only 2 or 3 V (in contrast to several tens of volts required by conventional circuits), whilst retaining good high frequency operation. This also reduces the power dissipation and, more importantly, with low system supply voltage, the need for voltage gain is lower.

As with BJTs and JFETs, the parasitic capacitances of MOSFETs are also voltage dependent, but only partially, as will become evident from the following comparison with JFETs.

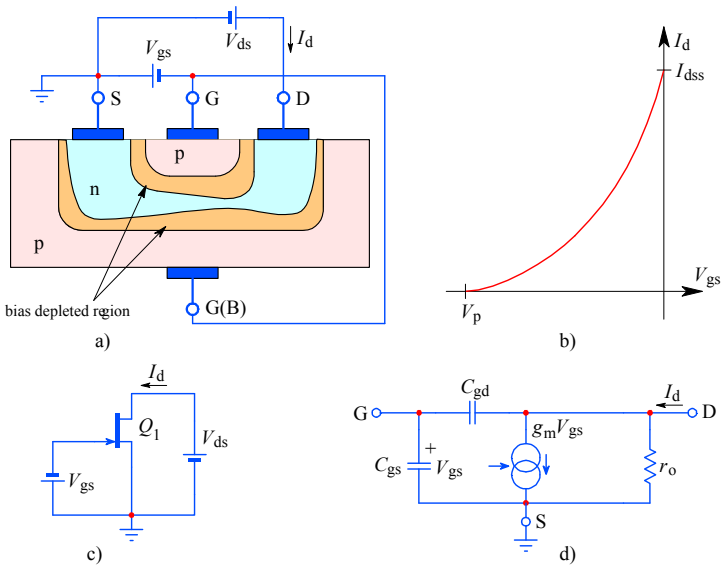


Fig. 5.2.19: a) A typical n-channel JFET structure cross-section under the bias condition. The p-type substrate is in contact with the p-type gate. The n-type channel is formed between the source and the drain. The bias voltage depletes the channel. **b)** The V_{gs}–I_d characteristic. **c)** The symbolic circuit. **d)** The equivalent circuit model.

The JFET capacitances C_{gd} and C_{gs} are voltage dependent:

$$C_{gd} = C_{gd0}\left(1 + \frac{V_{gd}}{V_{bi}}\right)^{-n_J} \tag{5.2.53}$$

$$C_{gs} = C_{gs0}\left(1 + \frac{V_{gs}}{V_{bi}}\right)^{-n_J} \tag{5.2.54}$$

where:

n_J is the junction grading coefficient ($1/2$ for abrupt and $1/3$ for graded junctions; most JFETs are built with a graded junction);

$V_{bi} = \frac{k_B T}{q_e} \ln \frac{N_A\, N_D}{n_i^2}$ is the intrinsic zero bias built in potential;

N_A is the acceptor doping density in p-type material ($\approx 10^{21}$ atoms/m^3);

N_D is the donor doping density in n-type material ($\approx 10^{22}$ atoms/m^3);

n_i is the intrinsic Si charge density (1.5×10^{16} electrons/m^3 at 300 K).

The built in potential V_{bi} relates to the JFET pinch off voltage parameter V_P as:

$$V_P = a^2\, \frac{q_e\, N_A}{2\, \varepsilon_{Si}} \left(1 + \frac{N_A}{N_D}\right) - V_{bi} \tag{5.2.55}$$

where:

a is the channel thickness ($\approx 2 \times 10^{-6}$ m);

ε_{Si} is the silicon dielectric permeability ($= 1.04 \times 10^{-10}$ F/m).

With the typical values above, $V_{bi} \approx 0.64$ V and $V_P \approx 3.4$ V.

In integrated circuits, a JFET would also have a gate to substrate capacitance C_{gss}, which, accounting for an abrupt junction, can be expressed as:

$$C_{gss} = C_{gss0}\left(1 + \frac{V_{gss}}{V_{bi}}\right)^{-\frac{1}{2}} \tag{5.2.56}$$

A typical zero bias range of values for these capacitances is:

$$C_{gd0} = 0.3\text{–}1 \text{ pF}$$

$$C_{gs0} = 1\text{–}4 \text{ pF}$$

$$C_{gss0} = 4\text{–}8 \text{ pF}$$

So a JFET with a transconductance $g_m = 2.5 \times 10^{-3}$ A/V and a total gate capacitance $C_T = C_{gd} + C_{gs} + C_{gss} = 4$ pF (under appropriate bias) would yield a cut off frequency:

$$f_T = \frac{1}{2\,\pi} \cdot \frac{g_m}{C_T} = \frac{2.5 \times 10^{-3}}{6.28 \times 4 \times 10^{-12}} \approx 100 \text{ MHz} \tag{5.2.57}$$

For MOSFETs, the situation is slightly different. Fig. 5.2.20 shows a typical n-channel MOS transistor cross-section and the equivalent circuit model.

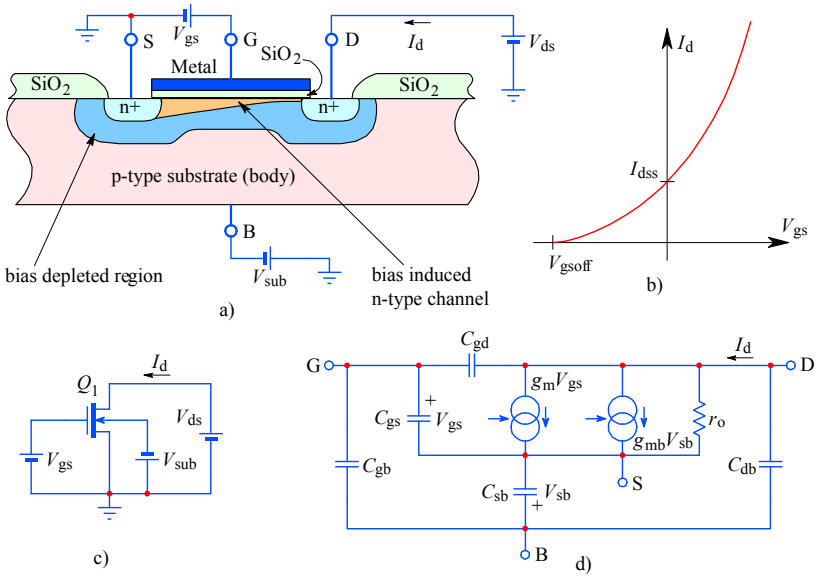


Fig. 5.2.20: **a)** A typical n-channel MOSFET structure's cross-section under bias condition. Two heavily doped n+ regions (source and drain) are manufactured on a p-type substrate and a metal gate covers a thin insulation layer, slightly overlapping the n+ regions. The bias voltage depletes a thick region in the substrate, within which an n-type channel is induced between the source and the drain. **b)** The V_{gs}–I_d characteristic. **c)** The symbolic circuit. **d)** The equivalent circuit model has two current sources, one owed to the usual mutual transconductance g_m and the gate–source voltage V_{gs}; the other is owed to the so called 'body effect' transconductance g_{mb} and the associated source–body voltage V_{sb}. The g_{mb} is typically an order of magnitude lower than g_m.

From the MOSFET structure cross-section it can be deduced that C_{gb} is small, owing to the relatively large depleted region. Ordinarily its value is about 0.1 pF and it is relatively constant. Likewise the depletion region capacitances C_{sb} and C_{db} are also small (they are proportional to the gate and source area), but they are voltage dependent:

$$C_{sb} = C_{sb0}\left(1 + \frac{V_{sb}}{V_{bi}}\right)^{-\frac{1}{2}} \tag{5.2.58}$$

$$C_{db} = C_{db0}\left(1 + \frac{V_{db}}{V_{bi}}\right)^{-\frac{1}{2}} \tag{5.2.59}$$

The capacitances C_{gs} and C_{gd} are owed to the SiO_2 insulation layer between the gate and the channel. If S_g is the gate area and C_x is the unit area capacitance of the oxide layer under the gate, then the total capacitance is:

$$C_{gs0} + C_{gd0} = S_g\, C_x \tag{5.2.60}$$

Most MOSFETs are built with symmetrical geometry, thus the total zero bias capacitance is simply split in half. But in the saturation region the channel narrows, so

the drain voltage influence is small, resulting in a nearly constant C_{gd} whose value is essentially proportional to the small gate–drain overlapping area. Thus typical C_{gd} values range between 0.002 and 0.020 pF.

C_{gs} is larger, typically some $2/3$ of the $S_g\, C_x$ value, or about 1–2 pF.

Although MOSFETs' g_m is typically lower than in JFETs, it is the very small capacitances, in particular C_{gd} and C_{gb}, which are responsible for the wider bandwidth of a MOSFET source follower. Cut off frequencies of many GHz are easily achieved.

5.2.7 Input Protection Network

The input protection network is needed for two distinct real life situations. The first one is the (occasional) electrostatic discharge, the second one is a long term overdrive.

Imagine a technician sitting on a well insulated chair, wearing woolen or synthetic clothes, and rubber plated shoes, repairing a circuit on his bench. For a while he rubs his clothes on the chair by reaching for the schematic, the spare parts, some tools, etc., thus quickly charging himself up to an average 500 V. Suddenly, he needs to put a 1:1 'scope probe somewhere on the rear panel and he stands up, touching the probe to identify its trace by the characteristic capacitive AC mains pickup. By standing up, he has increased his average distance from the chair by a large factor, say 30, but the charge on the chair and his clothes remains unchanged. This is equivalent to charging a parallel plates capacitor and then increasing the plates distance, so that the capacitance drops inversely to the distance (Eq. 5.2.51). Because $V = Q/C$, his effective voltage would increase 30 times, reaching some 15 kV!

The average capacitance of the human body towards the surroundings of an average room is about 200 pF. So, when our technician touches the probe tip, he will discharge the 15 kV of his 200 pF right into the input of the poor 'scope. And such a barbaric act can be repeated hundreds of times during an average repairing session.

At the instant the probe tip is touched the effective input voltage falls for the first 5 ns (the propagation delay of the 1 m long probe cable) to a level set by the resulting capacitive divider $a = 1/(1 + C_{cable}/C_{body})$, so if $C_{cable} = 100\,\text{pF/m}$, $V = a\, V_{body} \approx 10\,\text{kV}$. Here we assume a signal propagation velocity of 0.2 m/ns (about $2/3$ of the speed of light). Also, note that the probe cable is made as a lossy transmission line (the inner conductor is made of a thin resistive wire, about $50\,\Omega/\text{m}$).

After 5 ns the cable capacitance is fully charged and the signal reaches the spark gap. The spark gap fires, limiting the input voltage to its own breaking voltage (1500–2000 V), providing a low impedance path to ground and discharging $C_{cable} + C_{body}$. Some 25 ns later the voltage falls below the spark threshold.

Now the total capacitance $C_{cable} + C_{body} + C_{in}$ is discharged into the remaining input resistance. With the attenuator set to highest sensitivity (1:1), the input resistance is equal to the $1\,\text{M}\Omega$ of R_{in}, in parallel to the series connection of the damping resistor R_d and one of the protection diodes (depending on the voltage polarity). The diode must withstand a peak current $I_{dpk} = V_{spark}/R_d$; if $R_d = 150\,\Omega$, then $I_{dpk} = 2000/150 = 13.3\,\text{A}$! Fortunately, the peak current is lowered also by the loop inductance. The spark discharges the capacitance in less than 30 ns and then the

current falls exponentially as the total capacitance is discharged through R_d, which lasts another 250 ns. At this time the voltage is lower than $V_{cc} + V_{D1}$ and the capacitance is discharged through R_{in}.

In Fig. 5.2.21 we have plotted the first 250 ns of the discharge event, along with the schematic, showing only the most important circuit components.

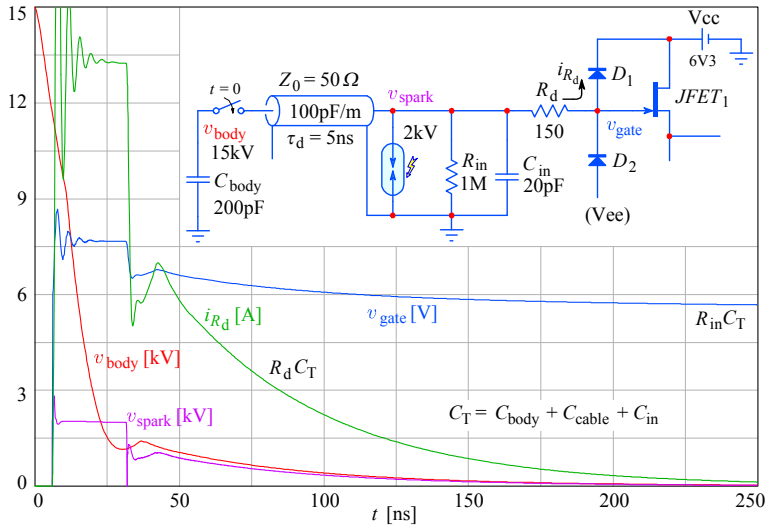


Fig. 5.2.21: A human body model of electrostatic discharge into the oscilloscope input. About 5 ns after touching the probe tip the probe cable is charged and the voltage reaches the spark gap. The spark gap fires and limits the voltage to its firing threshold. The arc provides a low impedance path discharging the body and cable capacitance until the voltage falls below the firing threshold (~25 ns). The remaining charge is fed through R_d and one of the protection diodes, until the voltage falls below $V_{cc} + V_{D1}$ (~250 ns). Finally, the capacitance is discharged through R_{in}.

A different situation occurs in case of a long term overdrive. Fig. 5.2.22 shows the protection network.

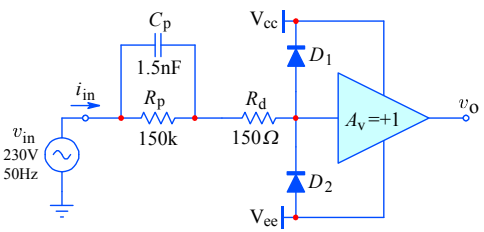


Fig. 5.2.22: Input protection network for long term overdrive: R_p limits the DC and LF current; at HF it is bypassed by C_p, leaving R_d to limit the input current.

The most severe long term input overdrive occurs when the oscilloscope input is on its highest sensitivity setting (no attenuation) and the user inadvertently connects a 1:1 probe to a high voltage DC or AC power supply. A typical highest sensitivity setting of 2 mV/div or ±8 mV range is brutally exceeded by the 230 V_{eff}, 650 V_{pp} AC mains voltage. Since with a well designed instrument nothing dramatic would happen

(no flash, no bang, no smoke), the user might realize his error only after a while (a few seconds at best and several minutes in the evening at the end of a long working day). The instrument must be designed to withstand such a condition for indefinitely long.

With component values as in Fig. 5.2.22 the peak current through R_p is:

$$I_{Rpk} = (V_{inpk} - V_{cc})/R_p = (325 - 10)/150 \times 10^3 = 2.1\,\text{mA}$$

and the peak current through C_p:

$$I_{Cpk} = (V_{inpk} - V_{cc})\omega C_p = (325 - 10)(2\pi \times 50 \times 1.5 \times 10^{-9}) = 0.15\,\text{mA}$$

Of course, I_C leads I_R in phase by $\pi/2$, so the total current through R_p is the vector sum, $\sqrt{I_C^2 + I_R^2}$, and its value is essentially that of I_R, since the mains frequency (50–60 Hz) is much lower than the network cutoff, $1/(2\pi C_p R_p)$ or 707 Hz.

One could easily be unimpressed by such low current values, however we must not forget the transient conditions. With abundant help from Mr Murphy, we shall make the connection at the instant when the mains voltage is at its peak. And Mr Gauss will ensure a 50% probability that the instantaneous voltage will be above the effective value. Then the input current is limited by R_d only (2.1 A peak!). Fortunately the current falls exponentially with the $R_d C_p$ time constant (225 ns), so the transient is over in about 1 µs. The value of C_p should not be too low, either; note that it forms a capacitive divider with the JFET input capacitance and ground strays. If these are about 1.5 pF, the high frequency gain will be lower than at DC by about 0.1 %. All these components must be specified to survive voltage transients of at least 500 V, so their larger physical dimensions will increase the circuit size, and as a consequence the parasitic loop inductance and stray capacitances. We should also not forget to account for the JFET's negative input impedance compensation componets, as discussed already in Part 3 (see also [Ref. 5.69]).

Note also that by using a ÷2 basic high-Z input attenuation, the 150 kΩ and 1.5 nF can be safely omitted because the attenuator takes over their function.

5.2.8 Driving the Low Impedance Attenuator

The high impedance attenuator, discussed in Sec. 5.2.1, is almost exclusively implemented as a two or three decade switch. The intermediate attenuation and gain settings of the 1–2–5 sequence vertical sensitivity selector are usually realized in the stages following the FET source follower. For highest bandwidth the 1–2–5 attenuator is designed as a 50 Ω resistive divider and there are some advantages (regarding the linear signal handling range) if this attenuator is put immediately after the FET source follower. However, the FET by itself can not drive such a low impedance load and additional circuitry is required to help it to do so.

An interesting solution is shown in Fig. 5.2.23, patented by *John L. Addis* in 1983 [Ref. 5.21].

The input FET Q_1 is biased by the constant current source Q_2, as we have seen in Fig. 5.2.13. It is also actively compensated for large signal transient nonlinearity (by C_2 and Q_4) and bootstrapped by Q_3, which reduces the input capacitive loading by

keeping the voltage at C_{gd} nearly constant. The output complementary emitter follower is driven from the Q_1 source and R_3 (R_3 should be equal to R_4 to reduce the DC offset). The Q_3 bootstrap is provided by DZ_2, which, together with DZ_3, also bootstrap the bias circuit ($R_{9,10}$ and $D_{3,4}$) for the bases of Q_5 and Q_6, lowering in this way the load to Q_1 source.

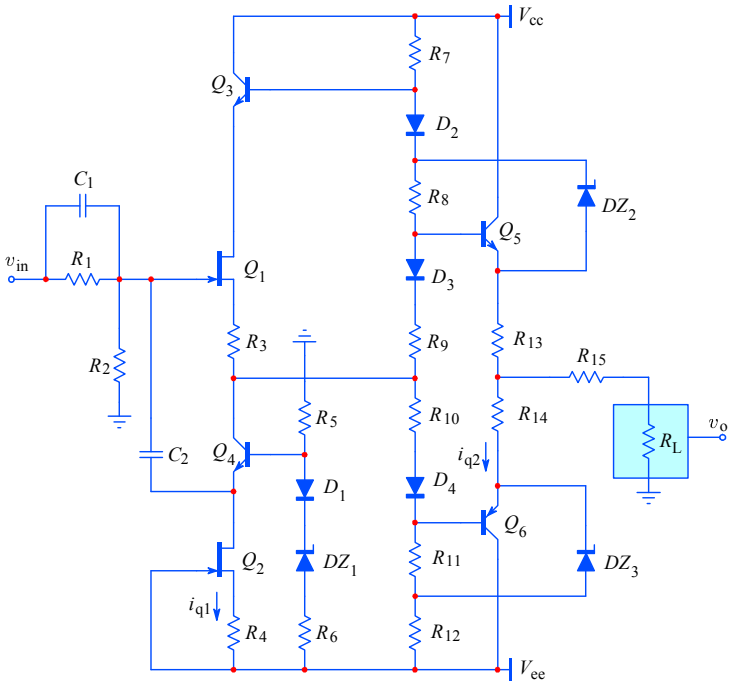


Fig. 5.2.23: The FET source follower with a buffer (patented by J.Addis in 1983) has a dynamic nonlinearity compensation and is capable of driving a low impedance load.

Bootstrapping increases the DC and low frequency impedance seen by Q_1, but note that its use will make sense at high frequencies only if the $Q_{5,6}$ and Q_3 are substantially faster than the FET itself. Otherwise, the bootstrap circuitry would only increase the parasitic capacitances and thus increase the rise time.

With enough driving capability made available by Q_5 and Q_6 the load resistor R_L can now be realized as the low impedance three step attenuator with a direct path and two attenuated paths, ÷2 and ÷4. Besides the usual maximum input sensitivity of 5 mV/div., this attenuator will provide the next two settings of 10 and 20 mV/div. The following lower sensitivity settings (50 mV/div., etc.) are achieved by switching in the ÷10 and ÷100 sections of the high impedance input attenuator, achieving the lowest sensitivity of 2 V/div. An external ÷10 probe will decrease this to 20 V/div.

Fig. 5.2.24 shows two possible implementations of the low impedance divider, having 50 Ω impedance at both input and output. The first attenuator design is based

on the τ-type network and the other on the π-type network. If the input signal is a current, the series 50 Ω in the ÷1 branch can be omitted.

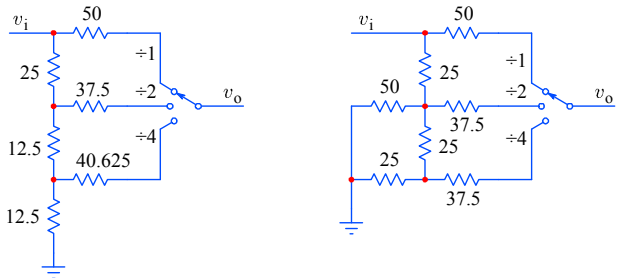


Fig. 5.2.24: The low impedance attenuator (50 Ω input and output) can be built as a straight τ-type ladder or a π-type ladder. If driven by a current source the series 50 Ω in the ÷1 branch can be omitted.

Here we assume that the input impedance of the following amplifier stage is high enough and its input capacitance is low enough for the division factor to remain correct at each setting and the bandwidth does not change. It is also important to keep the switch capacitive cross-talk low and preserve the nominal impedance by designing it as a microstrip transmission line.

Unfortunately, placing the low impedance attenuator immediately after the unity gain buffer would seriously degrade the noise performance at some attenuator settings. It is therefore better to move the attenuator after a further amplifying stage. But the ability to drive a 50 Ω load is nevertheless useful, e.g., in Tektronix 11A32 and 11A34 the input buffer had to drive a 25 cm long coaxial cable to the main board where the next stages were located (the M377 chip, described in Sec. 5.4.6, where we shall also see other ways of implementing the gain/sensitivity switching).

In addition to the discrete step attenuation, oscilloscopes, as well as other high speed instruments, often need a continuously variable attenuation (or gain), although within a restricted range (a range of 0.3 to 1 is often enough). A passive potentiometer is, of course, an obvious solution and it was used extensively in early days. However, this potentiometer is usually placed somewhere in the middle of the amplifier and its control shaft has to be brought to the instrument front panel, which can often be a mechanical nightmare. Also, its variable impedance causes the bandwidth to vary and this is a very undesirable property. An electronically controlled amplifier gain with constant bandwidth would therefore be welcome. We shall examine such circuits at the end of Sec. 5.4.

5.3 High Speed Operational Amplifiers

From about 1980 we have been witnessing both the development of a radically different operational amplifier topology and a major improvement in complementary semiconductor devices' technology, resulting in a steep rise in performance. At the same time, the market's hunger for higher bandwidth has been met by a massive production increase, so that the prices have remained fairly low. With the accompanying development in digital technology, both in terms of switching speed and circuit complexity, the techniques which have previously been forbiddingly expensive and too demanding to realize suddenly became feasible and within reach.

At the turn of the century opamps with the unity gain × bandwidth product of about 1 GHz or more (such as the Burr–Brown OPA-640) became available at a price comparable to that of a couple of good discrete high frequency transistors. Add to this a relatively good DC performance and, using surface mounting devices, a circuit area of $\sim 1\,\mathrm{cm}^2$, a low power consumption, and a noise level comparable to the thermal noise of a $100\,\Omega$ resistor, the advantages are obvious.

Clearly, we have come a long way from the ubiquitous μA741.

However, in order to better evaluate the performance and the design requirements of the new devices we shall first expose the weak points of the classical configuration.

5.3.1 The Classical Opamp

The name 'operational amplifier' springs from the analog computer era, in which amplifier blocks could be combined with passive components to perform various mathematical operations, from simple signal addition and subtraction to integration and differentiation.

The main performance limitation of early IC opamps was imposed by the integration technology itself: whilst fairly good NPN transistors could be easily produced, PNP transistors could be made along with NPN ones only as very slow, 'lateral' structures. This has restricted their use to only those parts of the opamp, where the needed bandwidth, gain, and load could be low. In practical terms, the only such place in a typical opamp is the so called 'middle stage level translator', as shown in Fig. 5.3.1. Even so, the opamp open loop bandwidth was almost always below 100 Hz, mainly owing to the Miller effect and the need to provide enough phase margin at low closed loop gain to ensure unconditional system stability.

On the other hand, for general purpose applications, the important parameter was a high negative feedback factor, used to minimize the circuit performance variations owed to the transistor parameters (which in the early days were difficult to control) and instead rely on passive components which could be easily produced with a relatively tight tolerance. It was this ability to deliver predictable performance by a simple choice of two feedback resistors which made the opamp a popular and widely used circuit component. And not only a well defined gain, but also a broad range of other signal conditioning functions is made possible by combining various passive and active components in the feedback loop.

The feedback concept is so simple and works so well that too many people take it for granted; and equally many are surprised to discover that it can cause as much trouble as the solutions it offers.

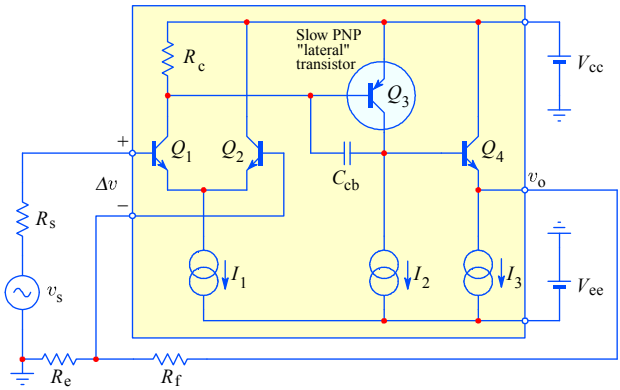


Fig. 5.3.1: The classical opamp, simplified. The input differential pair Q_1 and Q_2 subtract the feedback from the input signal, driving with the difference the 'level translator' stage Q_3, which in turn drives the output emitter follower Q_4, which provides a low output impedance. The feedback voltage is derived from the output voltage v_o by dividing it as $\beta = R_e/(R_e + R_f)$. If the opamp open loop gain A_o is much higher than $1/\beta$, the closed loop gain is $A_{cl} \approx 1/\beta$ (see text).

In the (highly simplified) opamp circuit in Fig. 5.3.1 the open loop gain is equal to the gain of the differential pair Q_1 and Q_2, multiplied by the gain of the level translator Q_3. The output emitter follower Q_4 has a unit voltage gain. However, all the three stages have their gain frequency dependent, as was explained in Part 3. Fortunately, the three poles are far apart (all are real) and the poles of the first and the third stage can be (and usually are) easily set high enough for the amplifier open loop frequency response to be dominated by the second stage pole (which was in turn named 'the dominant pole').

The dominant pole of the circuit in Fig. 5.3.1 is set by the Q_1 collector resistor R_c and the Miller capacitance C_M:

$$s_0 = -\frac{1}{R_c C_M} \tag{5.3.1}$$

where we have neglected the input resistance of Q_3 in parallel with R_c, which we can do if R_c is small.

C_M appears effectively in parallel with R_c and its value is equal to the Q_3 collector to base capacitance C_{cb}, multiplied by the Q_3 gain:

$$C_M = C_{cb}(1 + A_3) \tag{5.3.2}$$

The gain A_3 is set by the Q_3 transconductance and the loading resistance, in the form of the equivalent (shunt) input resistance at the base of Q_4:

$$A_3 \approx g_{m3} R_{b4} = \frac{q_e I_2}{k_B T} R_{b4} \tag{5.3.3}$$

The equivalent input resistance of Q_4 is approximately equal to the amplifier loading resistance reflected into the Q_4 base:

$$R_{\mathrm{b4}} = R_{\mathrm{L}}(1 + \beta_4) \tag{5.3.4}$$

where β_4 is the Q_4 current gain.

The gain of the input differential pair is set by the Q_1 collector load resistor R_{c} and the transconductance of both Q_1 and Q_2 (again neglecting the Q_3 input resistance):

$$A_1 \approx (g_{\mathrm{m1}} + g_{\mathrm{m2}})\, R_{\mathrm{c}} = \frac{q_{\mathrm{e}} I_1}{k_{\mathrm{B}} T}\, R_{\mathrm{c}} \tag{5.3.5}$$

where we have assumed both transconductances to be equal, owing to the current I_1 being equally divided into I_{c1} and I_{c2}.

As a result the open loop transfer function can be written as:

$$A(s) = A_0\, \frac{-\, s_0}{s - s_0} = A_1 A_3\, \frac{-\, s_0}{s - s_0} \tag{5.3.6}$$

Now we can derive the closed loop transfer function. By considering that the feedback factor β is set by the feedback resistive divider:

$$\beta = \frac{R_{\mathrm{e}}}{R_{\mathrm{e}} + R_{\mathrm{f}}} \tag{5.3.7}$$

the voltage at the inverting input is equal to the output voltage multiplied by the feedback factor:

$$v_{\mathrm{i}} = v_{\mathrm{o}}\, \beta \tag{5.3.8}$$

The signal being amplified is the difference between the source voltage and the voltage provided by feedback:

$$\Delta v = v_{\mathrm{s}} - v_{\mathrm{i}} \tag{5.3.9}$$

This voltage is amplified by the amplifier open loop transfer function, $A(s)$, to give the output voltage:

$$v_{\mathrm{o}} = A(s)\, \Delta v \tag{5.3.10}$$

By considering Eq. 5.3.6 and 5.3.9, we can write:

$$v_{\mathrm{o}} = A_0\, \frac{-\, s_0}{s - s_0}\, (v_{\mathrm{s}} - v_{\mathrm{i}}) \tag{5.3.11}$$

and, since v_{i} is a feedback scaled v_{o}:

$$v_{\mathrm{o}} = A_0\, \frac{-\, s_0}{s - s_0}\, (v_{\mathrm{s}} - v_{\mathrm{o}}\, \beta) \tag{5.3.12}$$

By rearranging this into:

$$v_{\mathrm{o}} \left(1 + \beta A_0\, \frac{-\, s_0}{s - s_0} \right) = A_0\, \frac{-\, s_0}{s - s_0}\, v_{\mathrm{s}} \tag{5.3.13}$$

we can obtain the explicit expression for v_o:

$$v_o = \frac{A_0 \dfrac{-s_0}{s-s_0}}{1+\beta A_0 \dfrac{-s_0}{s-s_0}} v_s \tag{5.3.14}$$

or:

$$v_o = \frac{1}{\dfrac{1}{A_0 \dfrac{-s_0}{s-s_0}} + \beta} v_s \tag{5.3.15}$$

From this last expression it is obvious that if the open loop gain A_0 is very high the amplifier gain v_o/v_s is reduced to the familiar $1/\beta$, or $(R_f + R_e)/R_e$.

Likewise, for a finite value of A_0, the frequency dependent part increases, thus lowering the closed loop gain at higher frequencies.

Take, for example, the µA741 opamp, Fig 5.3.2: owing to its dominant pole, the open loop cut off frequency is at about 10 Hz, whilst the open loop gain at DC is about 10^5. The unity gain crossover frequency f_1 is therefore about 1 MHz.

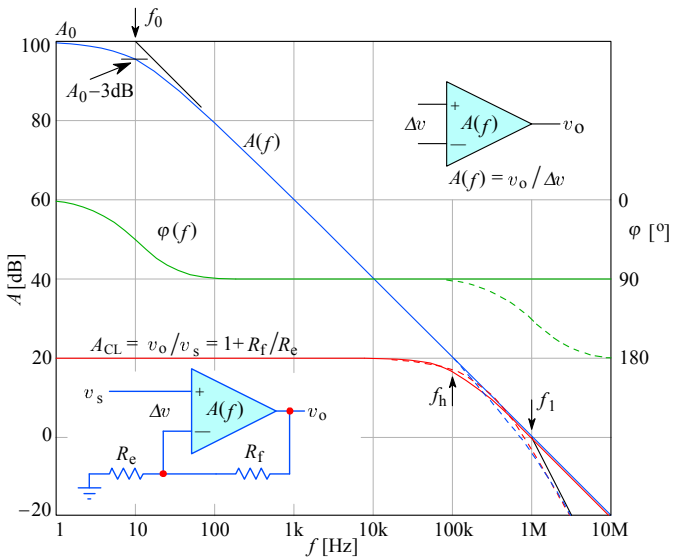


Fig. 5.3.2: A typical opamp open loop gain and phase compared to the closed loop gain. The dashed lines show the influence of a secondary pole (usually the input differential stage pole), which, for stability requirements, must be set at or above the unity gain transition frequency, $f_1 = 1$ MHz. f_h is the closed loop cutoff frequency.

For a closed loop gain of 10, $\beta = 0.1$; since the frequency dependence term is a ratio, the factor 2π can be extracted and canceled, leaving $f_0/(jf + f_0)$, where f_0 is the open loop cutoff frequency. By putting this into Eq. 5.3.15, we see that the amplifier will be making corrections to its own non-linearity by a factor 10^4 (80 dB) at 1 Hz, but only by a factor of 10^2 (40 dB) at 1 kHz; and at 100 kHz there would be only

3 dB of feedback, resulting in a 50% gain error. This means that for a source signal of 0.1 V there would be a Δv of 0.05 V, resulting in an output voltage of $v_o = 0.5$ V (instead of the 1 V as at low frequencies). Above the closed loop cutoff frequency the amplifier has practically no feedback at all.

An additional error is owed to the phase shift: at 100 kHz a single pole amplifier would have the output at 90° phase lag against the input. An amplifier with an additional input differential stage pole at 1 MHz would shift the phase by 135°, so there would be only a 45° phase margin at this frequency and the circuit would be practically at the edge of closed loop stability. If we were to need this amplifier to drive a 2 m long coaxial cable (capacitance 200 pF), by considering the amplifier output impedance of about 75 Ω the additional phase shift of 5° would be enough to turn the amplifier into a high frequency oscillator.

5.3.2 Slew Rate Limiting

The discussion so far is valid for the small signal amplification. For large signals the bandwidth would be much lower than the small signal one. This is owed to the Miller capacitance causing Q_3 to act as an integrator. For a positive input step larger than 2 $k_B T/q_e$ ($+I_1 R_{e1}$ if the input differential pair has emitter degeneration resistors), the transistor Q_1 will be fully open, while Q_2 will be fully closed. Therefore, the maximum current available to charge C_M will be equal to the tail current I_1. The voltage across C_M will increase linearly until the input differential stage will be out of saturation. Consequently, the slew rate limit is:

$$\mathrm{SR} = \frac{dv}{dt} = \frac{I_1}{C_M} \tag{5.3.16}$$

Usually I_1 is of the order of 100 µA (or even lower if low noise is the main design goal). Also, owed to the gain of Q_3 the Miller capacitance C_M can be large; say, with $C_{cb} = 4$ pF and $A_3 = 50$, C_M will be about 200 pF, giving a slew rate SR = 0.5 V/µs. We know that for a sine wave the maximum slope occurs at zero crossing, where the derivative is $dv/dt = d(V_p \sin\omega t)/dt = \omega V_p \cos\omega t$; at zero crossing $t = 0$ and $\cos(0) = 1$, so the slew rate equation can be written as:

$$\mathrm{SR} = \omega V_p = \frac{I_1}{C_M} \tag{5.3.17}$$

For a supply voltage of ±15 V, the signal amplitude just before clipping would probably be around 12 V, so the maximal full power sine wave frequency would be $f_{max} = I_1/2\pi C_M V_p$, or approximately 6.5 kHz only!

The frequency at which the sine wave becomes a linear ramp, with a nearly equal peak amplitude, is slightly higher: $f_r = 1/4t_r = \mathrm{SR}/4V_p = 10$ kHz (note that the SR of the circuit in Fig. 5.3.1 is not symmetrical, since C_M is charged by I_1 and discharged through R_e; in an actual opamp circuit, such as in Fig. 5.3.3, R_e is replaced by a current mirror, driven by the collector current of Q_2, giving a symmetrical slew rate).

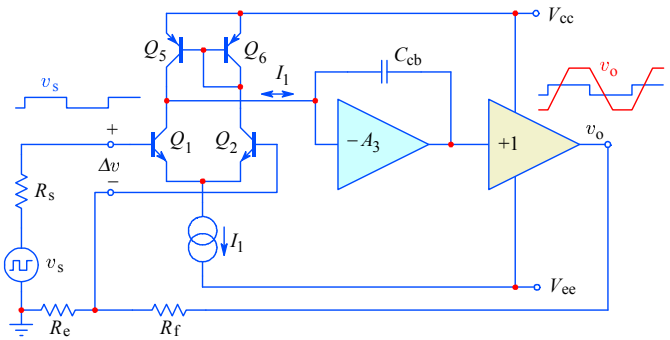


Fig. 5.3.3: Simplified conventional opamp circuit with the current mirror as the active load to Q_1. The second stage is modeled as a Miller integrator with large gain. This circuit exhibits symmetrical slew rate limiting. The dominant pole $s_0 = -g_m/C_M$, where g_m is the differential amplifier's transconductance and $C_M = C_{cb}(1 + A_3)$.

5.3.3 Current Feedback Amplifiers

The circuit in Fig. 5.3.1 could be characterized as a 'voltage feedback' amplifier and in the previous analysis we have shown its most important performance limitations. Instead the circuit in Fig. 5.3.4 is characterized as a 'current feedback' amplifier, since the feedback signal is in the form of a current, which, as will become evident soon, offers several distinct advantages.

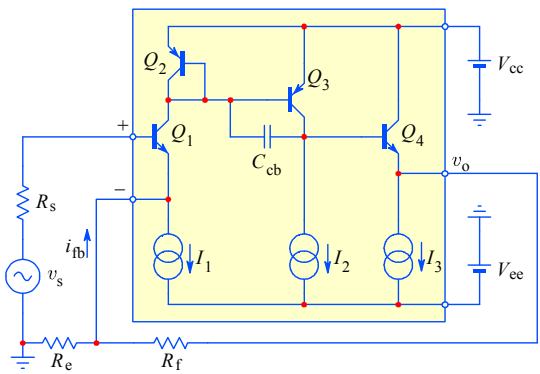


Fig. 5.3.4: Current feedback opamp, derived from the voltage feedback opamp (Fig. 5.3.1): we first eliminate Q_2 from the input differential amplifier and introduce the feedback into the Q_1 emitter (low impedance!). Next, we load the Q_1 collector by a diode connected Q_2, forming a current mirror with Q_3. Finally, we use very low values for R_f and R_e. The improvements in terms of speed are two: first, for large signals, the current available for charging C_{cb} is almost equal to the feedback current i_{fb}, eliminating slew rate limiting; second, C_{cb} is effectively grounded by the low impedance of Q_2, thus avoiding the Miller effect. A disadvantage is that the voltage gain is provided by Q_3 alone, so the loop gain is lower. Nevertheless, high frequency distortion can be lower than in classical opamps, because, for the equivalent semiconductor technology, the dominant pole is at least two decades higher, providing more loop gain for error correction at high frequencies.

The amplifier in Fig. 5.3.4 would still run into slew rate limiting for high amplitude signals, owing to the fixed bias of the first stage current source I_1. This is avoided by using a complementary symmetry configuration, as shown in Fig. 5.3.5. Of course, the complementary symmetry can be used throughout the amplifier, not just in the first stage.

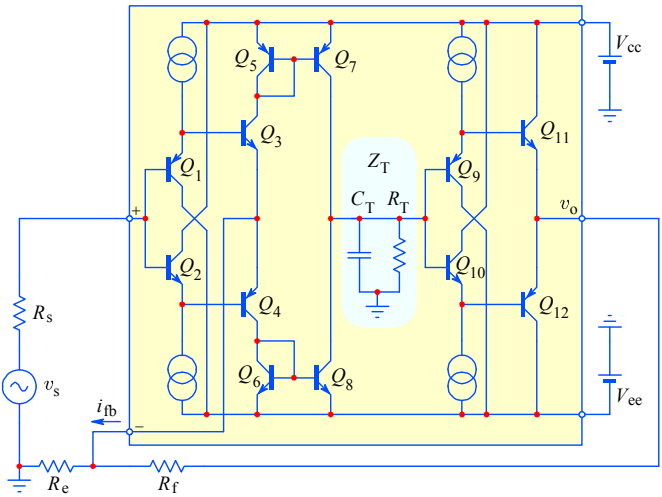


Fig. 5.3.5: A fully complementary current feedback amplifier model. It consists of four parts: transistors $Q_{1–4}$ form a unity gain buffer, the same as $Q_{9–12}$, with the four current sources providing bias; $Q_{5,7}$ and $Q_{6,8}$ form two current mirrors. In contrast to the voltage feedback circuit, both of whose inputs are of high impedance, the inverting input of the current feedback amplifier is a low impedance output of the first buffer. The current flowing in or out of the emitters of $Q_{3,4}$ is (nearly) equal to the current at the $Q_{3,4}$ collectors. This current is reflected by the mirrors and converted into a voltage at the $Q_{7,8}$ collectors, driving the output unity gain buffer. The circuit stability is ensured by the transimpedance Z_T, which can be modeled as a parallel connection of a capacitor C_T and resistor R_T. The closed loop bandwidth is set by R_f and the gain by R_e (the analysis is presented later in the text). One of the first amplifiers of this kind was the Comlinear CLC400.

Perhaps, it would be more correct to label the structure in Fig. 5.3.5 as a 'current on demand' type of amplifier, owing to the fact that the feedback current, which is proportional to the input–output error, feeds the dominant pole capacitance. The larger the error, the larger the current, which practically eliminates the slew rate limiting. The slew rate will be limited nevertheless, due to secondary effects, which will be discussed later, but the maximum current charging C_T is usually much greater than in conventional amplifiers. Also, C_T is small (not affected by the Miller effect).

Another name often found in the literature is the 'transimpedance amplifier', after the transimpedance equation (Z_T, see the analysis below). Owing to historical reasons, we shall keep the name used in the section title.

The complementary symmetry nature of the circuit in Fig. 5.3.5 would have been difficult to realize with the available opamp integration technology of the late 1960s and early 1970s, owing to the different characteristics of PNP and NPN transistors. A major technological breakthrough, made between 1980 and 1985 at

Comlinear Corporation and Elantec, later followed by Analog Devices, Burr–Brown and others, enabled PNP transistors to have their f_T almost as high as the NPN had. Fig. 5.3.6 shows a typical chip cross-section and Table 5.3.1 presents the typical values of the most important production parameters improving over the years.

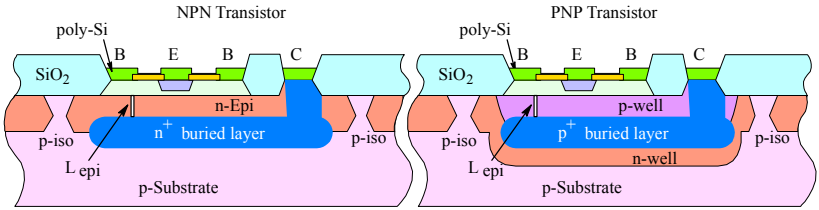


Fig. 5.3.6: Cross-section of the Complementary-Bipolar process.

Process (yr)	VIP1 (1986)		VIP2 (1988)		VIP3 (1994)		VIP10 (2000)		Units
Parameter	NPN	PNP	NPN	PNP	NPN	PNP	NPN	PNP	—
i_c/i_b	250	150	250	80	150	60	100	55	—
Early V_A	200	60	150	40	150	50	120	40	V
f_T	0.4	0.2	0.8	0.5	3	2.5	9	8	GHz
C_{js}	2.0	2.2	1.5	1.8	0.5	0.8	0.005	0.007	pF
E width	15		11		2		1		µm
Area	20000		18000		2400		300		µm²
$V_{ce\ max}$	36		36		32		12		V

Table 5.3.1: Typical production parameters of the Complementary Bipolar process [Ref. 5.33]

Although the same technology is now used also for conventional voltage feedback amplifiers, the current feedback structure offers further advantages which result in improved circuit bandwidth, as put in evidence by the following discussion.

The stability of the amplifier in Fig. 5.3.5 is ensured by the so called transimpedance network, Z_T, which can be modeled as a parallel $R_T\,C_T$ network. Note that the compensating capacitor, C_T (consisting of 4 parts, C_{T1}–C_{T4}), is effectively grounded, as can be seen in Fig. 5.3.7, since C_{cb} of $Q_{9,10}$ are tied to the supply voltages directly, whilst the C_{cb} of $Q_{7,8}$ are tied to the supply by the low impedance CE path of $Q_{5,6}$.

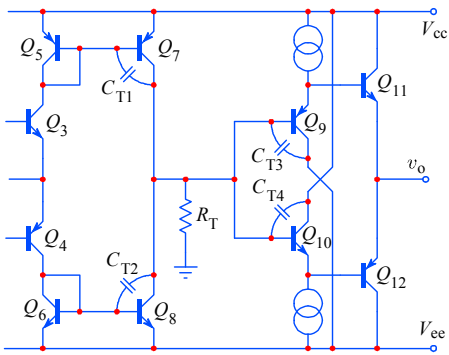


Fig. 5.3.7: The capacitance C_T consists of four components, all effectively grounded.

Therefore in this configuration the Miller effect is substantially eliminated. This means that, for the same driving current, this circuit is capable of much higher bandwidth, compared to the conventional opamp.

Also, owing to the two current mirrors, the current which charges and discharges C_T is equal to the current injected by the feedback network into the inverting input (the first buffer output). Since this current is feedback derived, it is proportional to the input–output error; thus for a fast input voltage step there would initially be a large input–output error, causing a large current into the inverting input and an equally large current will charge C_T, so its voltage, and consequently the output voltage, will increase fast. As the output voltage increases, the error is reduced, reducing the error and lowering the current.

To analyze the circuit operation we shall at first assume ideal buffers and current mirrors, as modeled in Fig. 5.3.8. Later, we shall see how the real circuit parameters limit the performance.

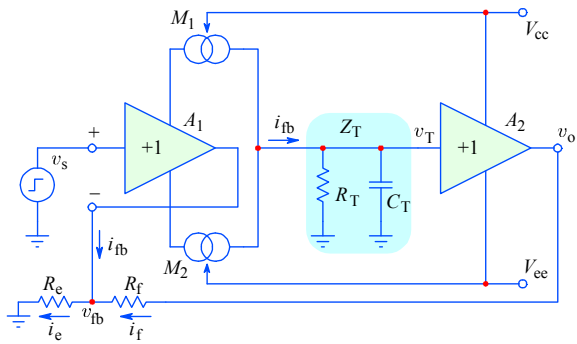


Fig. 5.3.8: Current feedback amplifier model used for the analysis.

Imagine for a moment that R_f is taken out of the circuit. Essentially this would be an open loop configuration, the gain of which can be expressed by the ratio of two resistors, R_T/R_e. The gain is provided by the current mirrors L_M; their output currents are summed, so the two mirrors behave like a single stage; consequently the gain value, compared with that of a conventional opamp, is relatively low (in practice, maximum gains between 60 and 80 dB are common). It is important to note, however, that the open loop (voltage) gain does not play such an important role in current feedback amplifiers. As the name implies, it is more important how the feedback current is processed.

Let us now examine a different situation: we put back R_f and disconnect R_e. If there were to be any voltage difference between the outputs of the two buffers, a current would be forced through R_f, increasing the output current of the first buffer, A_1. The two current mirrors would reflect this onto the input of the second buffer, A_2, in order to null the output voltage difference. In other words, **the output of the first buffer A_1 represents an inverting current mode input** of the whole system.

If we now reconnect R_e it is clear that the A_1 output must now deliver an additional current, i_e, flowing to the ground. The current increase is reflected by the mirrors into a higher v_T, so the output voltage v_o would increase, forcing the current

i_f (through R_f) into the A_1 output. By looking from the A_1 output, i_e flows in the direction opposite to i_f, so the total current i_fb of the A_1 output will be equal to their difference. Thus with R_e and R_f a classical feedback divider network is formed, but the feedback signal is a current. As expected, the output of A_2 must now become $(R_\mathrm{f} + R_\mathrm{e})/R_\mathrm{e}$ times higher than the output of A_1 to balance the feedback loop.

Deriving the circuit equations is simple. The transimpedance equation (assuming an ideal unity gain buffer A_2, thus $v_\mathrm{T} = v_\mathrm{o}$) is:

$$v_\mathrm{o} = Z_\mathrm{T}\, i_\mathrm{fb} \tag{5.3.18}$$

The feedback current (assuming an ideal unity gain buffer A_1, thus $v_\mathrm{fb} = v_\mathrm{s}$) is:

$$i_\mathrm{fb} = \frac{v_\mathrm{s}}{R_\mathrm{e}} - \frac{v_\mathrm{o} - v_\mathrm{s}}{R_\mathrm{f}} = v_\mathrm{s}\left(\frac{1}{R_\mathrm{e}} + \frac{1}{R_\mathrm{f}}\right) - \frac{v_\mathrm{o}}{R_\mathrm{f}} \tag{5.3.19}$$

The closed loop gain (from both equations above) is:

$$\frac{v_\mathrm{o}}{v_\mathrm{s}} = \left(1 + \frac{R_\mathrm{f}}{R_\mathrm{e}}\right) \frac{1}{1 + \dfrac{R_\mathrm{f}}{Z_\mathrm{T}}} \tag{5.3.20}$$

We see that the equation for the closed loop gain has two terms, the first one resulting from the feedback network divider and the second one containing the transimpedance Z_T and R_f, but not R_e! This is in contrast to what we are used to in conventional opamps.

If we now replace Z_T by its equivalent network, $1/(s\,C_\mathrm{T} + 1/R_\mathrm{T})$, then the closed loop gain Eq. 5.3.20 can be rewritten as:

$$\frac{v_\mathrm{o}}{v_\mathrm{s}} = \left(1 + \frac{R_\mathrm{f}}{R_\mathrm{e}}\right) \frac{1}{1 + \dfrac{R_\mathrm{f}}{R_\mathrm{T}} + s\,C_\mathrm{T}\,R_\mathrm{f}} \tag{5.3.21}$$

We can rewrite this to reveal the system's pole, in the way we are used to:

$$\frac{v_\mathrm{o}}{v_\mathrm{s}} = \frac{1 + \dfrac{R_\mathrm{f}}{R_\mathrm{e}}}{1 + \dfrac{R_\mathrm{f}}{R_\mathrm{T}}} \cdot \frac{\left(1 + \dfrac{R_\mathrm{f}}{R_\mathrm{T}}\right)\dfrac{1}{C_\mathrm{T}\,R_\mathrm{f}}}{s - \left[-\left(1 + \dfrac{R_\mathrm{f}}{R_\mathrm{T}}\right)\dfrac{1}{C_\mathrm{T}\,R_\mathrm{f}}\right]} \tag{5.3.22}$$

By comparing this with the general single pole system transfer function:

$$F(s) = A_0\,\frac{-\,s_1}{s - s_1} \tag{5.3.23}$$

we note that the term:

$$s_1 = -\left(1 + \frac{R_\mathrm{f}}{R_\mathrm{T}}\right)\frac{1}{C_\mathrm{T}\,R_\mathrm{f}} \tag{5.3.24}$$

is the closed loop pole, which sets the closed loop cutoff frequency: $f_\mathrm{h} = |s_1|/2\pi$.

We also note that the system closed loop gain is:

$$A_0 = \frac{1 + \dfrac{R_\mathrm{f}}{R_\mathrm{e}}}{1 + \dfrac{R_\mathrm{f}}{R_\mathrm{T}}} \tag{5.3.25}$$

Since R_f is normally much smaller than R_T, the term $R_\mathrm{T}/R_\mathrm{f}$ represents the open loop gain, so $R_\mathrm{f}/R_\mathrm{T}$ represents the closed loop gain error (in analogy with the finite open loop gain error at DC in classical amplifiers).

If, for example, we have an amplifier with $R_\mathrm{T} = 300\,\mathrm{k\Omega}$ and $C_\mathrm{T} = 2\,\mathrm{pF}$ and we form the feedback with $R_\mathrm{f} = 330\,\Omega$ and $R_\mathrm{e} = 110\,\Omega$, the amplifier would have a closed loop bandwidth of about 240 MHz and a gain of 4. Moreover, its loop gain would be $R_\mathrm{T}(R_\mathrm{f} + R_\mathrm{e})/R_\mathrm{f}R_\mathrm{e} \approx 3000$ and flat up to some 265 kHz, more than three orders of magnitude higher than the usual 100 Hz in conventional voltage feedback operational amplifiers.

Thus, we find that the closed loop bandwidth depends mainly on R_f (but not on R_e), whilst the gain, once R_f has been chosen, can be set by R_e alone. With current feedback the amplifier designer has independent control over the two most important circuit parameters and must only watch for possible second-order effects.

The benefits of the current feedback amplifier are all due to two main points:

a) since there is only one voltage gain stage (the two current mirrors, working effectively in parallel) and only one internal high impedance node (Z_T), this structure is inherently a single pole system (since the integration technology allows the poles of both buffers to be much higher). As a consequence, the system may always be made unconditionally stable, whilst not compromising the available system bandwidth;

b) since the feedback is entered in the form of a current the system bandwidth depends on R_f but not on R_e, so that if R_f is held constant while R_e is varied, the bandwidth remains (almost) independent of gain. As a bonus there is practically no slew rate limiting mechanism, because the feedback current drives a grounded C_T and the larger the input voltage step, the larger will be the current charging C_T. So the amplifier's step response will always look like that of a low pass $R_\mathrm{f}C_\mathrm{T}$ network for any signal amplitude up to the clipping level.

It might be interesting to return to Eq. 5.3.19 with the result of Eq. 5.3.21 and express the current which charges C_T as a function of input voltage:

$$i_{C_\mathrm{T}} \approx i_\mathrm{fb} = v_\mathrm{s}\left(\frac{1}{R_\mathrm{f}} + \frac{1}{R_\mathrm{e}}\right)\left(1 - \frac{1}{1 + R_\mathrm{f}/R_\mathrm{T} + s\,C_\mathrm{T}\,R_\mathrm{f}}\right) \tag{5.3.26}$$

Therefore we can express i_{C_T} (see the transient response in Fig. 5.3.9) as:

$$i_{C_\mathrm{T}} \approx i_\mathrm{fb} = v_\mathrm{s}\left(\frac{1}{R_\mathrm{f}} + \frac{1}{R_\mathrm{e}}\right)\frac{s}{s + \left(1 + \dfrac{R_\mathrm{f}}{R_\mathrm{T}}\right)\dfrac{1}{C_\mathrm{T}\,R_\mathrm{f}}} \tag{5.3.27}$$

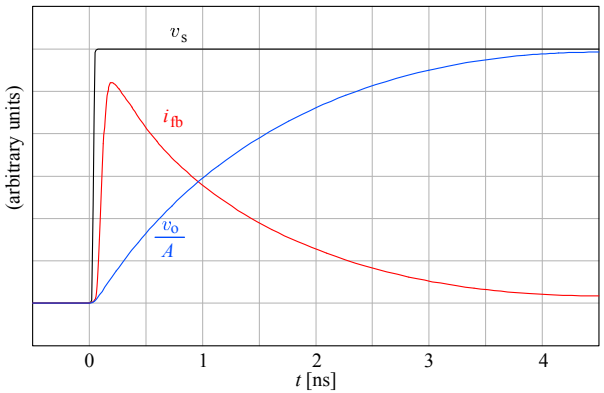


Fig. 5.3.9: 'Current on demand': The step response reveals that the feedback current is proportional to the difference between the input and output voltage, essentially a high pass version of the output voltage, as shown by Eq. 5.3.27.

In Fig. 5.3.10 we compare the cut off frequency vs. gain of a voltage feedback and a current feedback amplifier. The voltage feedback amplifier bandwidth is inversely proportional to gain; in contrast, the current feedback amplifier bandwidth is, in principle, independent of gain. This property makes current feedback amplifiers ideal for wideband programmable gain applications.

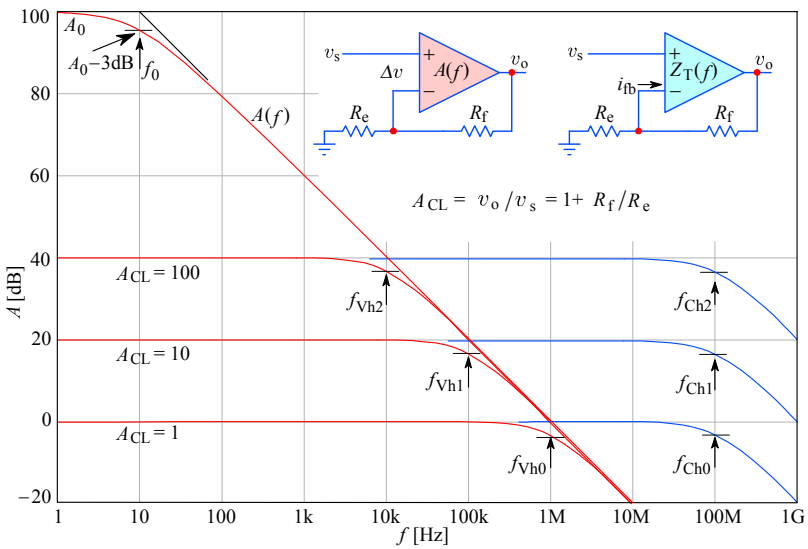


Fig. 5.3.10: Comparison of closed loop cut off frequency vs. gain of a conventional and a current feedback amplifier. The conventional amplifier has a constant GBW product (higher gain, lower cut off). But the current feedback cut off frequency is (almost) independent of gain.

Of course, a real amplifier will have some second-order effects, which we have not discussed so far, so its performance will be somewhat less than ideal.

The main causes for non-ideal performance are:

- the small but finite inverting input resistance (non-zero output impedance of A_1 in Fig. 5.3.8), which causes a voltage error between the non-inverting and the inverting input and, consequently, a lower feedback current through R_f;
- the non-zero output impedance of A_2, which, combined with the output load forms an additional feedback divider;
- the asymmetry and non-linearity of the two current mirrors, which directly influences the transfer function;
- the finite current gain of the output stage A_2, which, if too low, would allow the amplifier load to be reflected at the input of A_2 and influence Z_T;
- the secondary poles at high frequencies, owed to the finite bandwidth of the transistors within the amplifier.

The last four points are equally well known from conventional amplifiers and their influence is straightforward and easy to understand, so we shall not discuss them any further. The first point, however, deserves some attention.

5.3.4 Influence of a Finite Inverting Input Resistance

The current feedback amplifier requires a low (ideally zero) impedance at the inverting input in order to sense the feedback current correctly. This, in addition to the manufacturing imperfections between the transistors Q_{1-4} (Fig. 5.3.11), results in a relatively high input offset, owed to both DC voltage errors and current errors.

The offset is reduced by using the current mirror technique for the biasing current sources (Q_{a-d}), making the currents of $Q_{1,2}$ equal. Further reduction is achieved by adding low value resistors, R_{1-4} (a value of about $10\,r_e$ is usually sufficient) to Q_{1-4} emitters. This, however, increases the inverting input resistance.

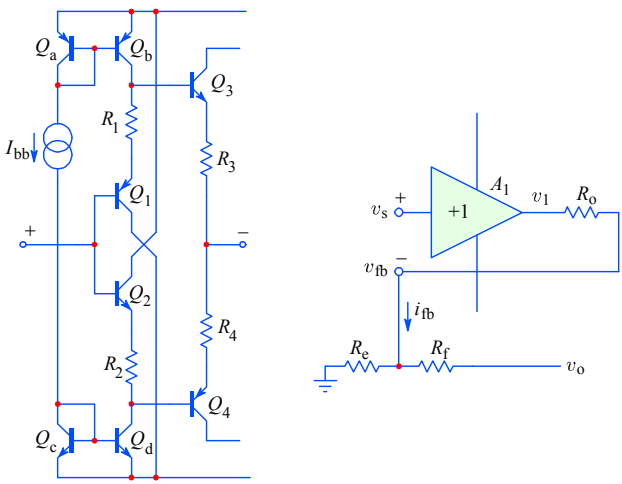


Fig. 5.3.11: The resistors R_{1-4} used to balance the inputs are the cause for the non-zero inverting input resistance, $R_3 \| R_4$, modeled by R_o; this causes an additional voltage drop, reducing the feedback current (see analysis).

A typical circuit of the first buffer implementing DC offset reduction by current mirror biasing and using emitter degeneration resistors is shown in Fig. 5.3.11; the equivalent inverting input resistance $R_3||R_4$ is modeled by R_o. It causes an additional voltage drop which reduces the feedback current.

For the analysis we assume that the buffer has a unity gain, thus $v_{o1} \approx v_s$. Since the feedback current flows through R_o, the voltage at the inverting input, v_{fb}, will be lower than v_s by $i_{fb}R_o$:

$$v_{fb} = v_s - i_{fb}R_o \tag{5.3.28}$$

By summing the currents at the v_{fb} node we have:

$$\frac{v_o - v_{fb}}{R_f} = \frac{v_{fb}}{R_e} - \frac{v_s - v_{fb}}{R_o} \tag{5.3.29}$$

Note that the last term in this equation is the feedback current from Eq. 5.3.28:

$$i_{fb} = \frac{v_s - v_{fb}}{R_o} \tag{5.3.30}$$

and from the transimpedance equation Eq. 5.3.18 the feedback current required to produce the output voltage v_o is:

$$i_{fb} = \frac{v_o}{Z_T} \tag{5.3.31}$$

By substituting v_{fb} and i_{fb} in Eq. 5.3.28, we obtain the transfer function:

$$\frac{v_o}{v_s} = \frac{\left(1 + \frac{R_f}{R_e}\right)}{1 + \left(1 + \frac{R_f}{R_e}\right)\frac{R_o}{Z_T} + \frac{R_f}{Z_T}} \tag{5.3.32}$$

If we express Z_T by its components:

$$Z_T = \frac{1}{\frac{1}{R_T} + sC_T} \tag{5.3.33}$$

we obtain:

$$\frac{v_o}{v_s} = \frac{\left(1 + \frac{R_f}{R_e}\right)}{1 + \left(1 + \frac{R_f}{R_e}\right)\frac{R_o}{R_T} + \frac{R_f}{R_T} + sC_T\left[R_f + R_o\left(1 + \frac{R_f}{R_e}\right)\right]} \tag{5.3.34}$$

which we reorder in the usual way to separate the DC gain from the frequency dependent part:

$$\frac{v_\mathrm{o}}{v_\mathrm{s}} = \frac{1+\dfrac{R_\mathrm{f}}{R_\mathrm{e}}}{1+\left(1+\dfrac{R_\mathrm{f}}{R_\mathrm{e}}\right)\dfrac{R_\mathrm{o}}{R_\mathrm{T}}+\dfrac{R_\mathrm{f}}{R_\mathrm{T}}} \cdot \frac{\dfrac{1+\left(1+\dfrac{R_\mathrm{f}}{R_\mathrm{e}}\right)\dfrac{R_\mathrm{o}}{R_\mathrm{T}}+\dfrac{R_\mathrm{f}}{R_\mathrm{T}}}{C_\mathrm{T}\left[R_\mathrm{f}+R_\mathrm{o}\left(1+\dfrac{R_\mathrm{f}}{R_\mathrm{e}}\right)\right]}}{s+\dfrac{1+\left(1+\dfrac{R_\mathrm{f}}{R_\mathrm{e}}\right)\dfrac{R_\mathrm{o}}{R_\mathrm{T}}+\dfrac{R_\mathrm{f}}{R_\mathrm{T}}}{C_\mathrm{T}\left[R_\mathrm{f}+R_\mathrm{o}\left(1+\dfrac{R_\mathrm{f}}{R_\mathrm{e}}\right)\right]}} \tag{5.3.35}$$

Again, a comparison with the general normalized first-order transfer function:

$$F(s) = A_0 \frac{-s_1}{s-s_1} \tag{5.3.36}$$

reveals the DC gain:

$$A_0 = \frac{1+\dfrac{R_\mathrm{f}}{R_\mathrm{e}}}{1+\left(1+\dfrac{R_\mathrm{f}}{R_\mathrm{e}}\right)\dfrac{R_\mathrm{o}}{R_\mathrm{T}}+\dfrac{R_\mathrm{f}}{R_\mathrm{T}}} \tag{5.3.37}$$

and the pole:

$$s_1 = -\frac{1+\left(1+\dfrac{R_\mathrm{f}}{R_\mathrm{e}}\right)\dfrac{R_\mathrm{o}}{R_\mathrm{T}}+\dfrac{R_\mathrm{f}}{R_\mathrm{T}}}{C_\mathrm{T}\left[R_\mathrm{f}+R_\mathrm{o}\left(1+\dfrac{R_\mathrm{f}}{R_\mathrm{e}}\right)\right]} \tag{5.3.38}$$

The bandwidth is calculated from the pole s_1:

$$f_\mathrm{h} = \frac{1}{2\pi}\,|s_1| \tag{5.3.39}$$

The DC closed loop gain A_0 contains the desired gain A_cl:

$$A_\mathrm{cl} = 1+\frac{R_\mathrm{f}}{R_\mathrm{e}} \tag{5.3.40}$$

and an error term ε:

$$\varepsilon = A_\mathrm{cl}\frac{R_\mathrm{o}}{R_\mathrm{T}}+\frac{R_\mathrm{f}}{R_\mathrm{T}} \tag{5.3.41}$$

which is small since R_T is usually 100 kΩ or higher, whilst R_o is between 5 and 50 Ω and R_f is between 100 Ω and 1 kΩ; thus, even if A_cl is 100 or more, ε rarely exceeds 10^{-3}. The transfer function can therefore be expressed as:

$$\frac{v_\mathrm{o}}{v_\mathrm{s}} = \frac{A_\mathrm{cl}}{1+\varepsilon} \cdot \frac{\dfrac{1+\varepsilon}{C_\mathrm{T}(R_\mathrm{f}+R_\mathrm{o}A_\mathrm{cl})}}{s+\dfrac{1+\varepsilon}{C_\mathrm{T}(R_\mathrm{f}+R_\mathrm{o}A_\mathrm{cl})}} \tag{5.3.42}$$

Whilst the gain error ε is small, the bandwidth error can be rather high at high gain; e.g., if $R_f = 1\,k\Omega$ and $R_o = 10\,\Omega$, with $A_{cl} = 100$, the time constant would double and the bandwidth would be halved. In Fig. 5.3.12 we have plotted the bandwidth reduction as a function of gain for a typical current feedback amplifier. Although not constant as in theory, the bandwidth is reduced far less than in voltage feedback opamps (about 50× less for a gain of 100).

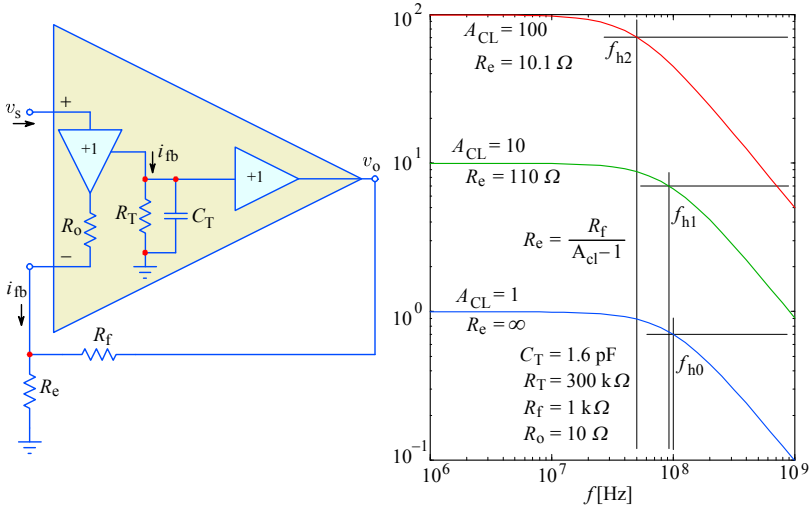


Fig. 5.3.12: The bandwidth of an actual current feedback amplifier is gain dependent, owing to a small but finite inverting input resistance R_o. The nominal bandwidth of 100 MHz at unity gain is reduced to 90 MHz at the gain of 10 and to only 50 MHz at the gain of 100 if R_o is 10 Ω. Nevertheless, this is still much better than in voltage feedback amplifiers.

From these relations we conclude two things: first, both the actual closed loop gain and bandwidth are affected by the desired closed loop gain $A_{cl} = 1 + R_f/R_e$; second and more important, for a given R_o, we can reduce R_f by $R_o A_{cl}$ and recalculate the required R_e to arrive at slightly modified values which **preserve both the desired gain and bandwidth!**

Note that in the above analysis we have assumed a purely resistive feedback path; additional influence of R_o will show up when we shall consider the effect of stray capacitances in the following section.

5.3.5 Noise Gain and Amplifier Stability Analysis

A classical voltage feedback unity gain compensated amplifier (for which any secondary pole lies above the open loop unity gain crossover) usually remains stable if a capacitor C_f is added in parallel to R_f, as in Fig. 5.3.13a. Because C_f lowers the bandwidth, it is often used to prevent problems at and above the closed loop cutoff. In contrast, a capacitor C_e in parallel with R_e, as shown in Fig. 5.3.13b, would reduce the feedback at high frequencies, leading to instability.

A different situation is encountered with current feedback amplifiers, which become unstable in the presence of any capacitance in parallel with either resistor of the feedback loop. Therefore the behavior of the circuit in Fig. 5.3.13a in the case of a current feedback amplifier is at odds with what we were used to.

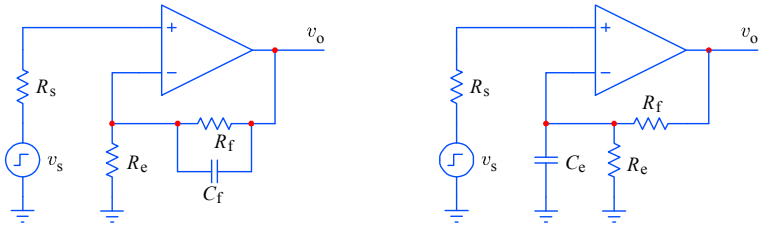


Fig. 5.3.13: **a)** A unity gain compensated voltage feedback amplifier remains stable with capacitive feedback, whilst in **b)** it is unstable. In contrast, the stability of a current feedback amplifier is upset by either **a)** C_f in parallel with R_f, or **b)** C_e in parallel with R_e.

Before we attempt to explain the unusual sensitivity of current feedback amplifiers to capacitively affected feedback, let us introduce an extremely useful concept, called **noise gain**. In contrast with the name, the noise gain is not used just to evaluate the circuit noise, but the circuit stability as well. It can be applied to all kinds of amplifiers, not just current feedback ones.

The noise, generated by the amplifier input stage, undergoes the full open loop amplification, so the input stage noise dominates over the noise of other stages. Therefore a noisy amplifier can be modeled as a noise generator in series with the input of a noiseless amplifier, as in Fig. 5.3.14, regardless of the actual signal amplification topology, be it inverting or non-inverting.

Any signal within the amplifier feedback loop is processed in the same way as the input stage noise. Thus by grounding the signal input and by analyzing the noise gain within the amplifier and its attenuation in the feedback network, we shall be able to predict the amplifier behavior.

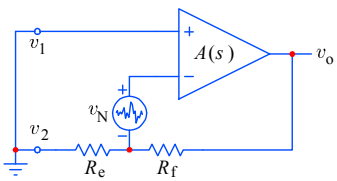


Fig. 5.3.14: Noise gain definition: A noisy amplifier is modeled as a noise generator v_N in series with the input of a noiseless amplifier. The inverting signal gain is $v_o/v_2 = -R_f/R_e$; the non-inverting signal gain is $v_o/v_1 = 1 + R_f/R_e$; the noise gain is $v_o/v_N = -(1 + R_f/R_e)$. The noise generator polarity is indicated only as a reference for the noise gain polarity inversion.

The noise gain is inverting in phase, but equal in value to the non-inverting signal gain:

$$A_N = \frac{v_o}{v_N} = -\left(1 + \frac{R_f}{R_e}\right) \tag{5.3.43}$$

For the voltage feedback amplifier the closed loop bandwidth is equal to the unity gain bandwidth frequency and the noise gain:

$$f_{\mathrm{cl}} = \frac{f_1}{|A_{\mathrm{N}}|} \tag{5.3.44}$$

If the feedback network is purely resistive the noise gain is independent of frequency; reactive components (usually capacitances) within the feedback loop will cause the noise gain to change with frequency.

To see this, we usually draw the asymptotes of the transfer function magnitude (absolute value) in a log-log *Bode* plot, with the breakpoints representing the poles and zeros, each pole adding a slope of $-20\,\mathrm{db}/10f$ and each zero a $+20\,\mathrm{dB}/10f$. We then approximate the phase angle at each breakpoint and in the middle of the linear section (this is a simple and straightforward process if the breakpoints are far apart — see an arbitrary example in Fig. 5.3.15 illustrating some most common possibilities). From this we can evaluate the feedback loop phase margin and, consequently, the amplifier stability.

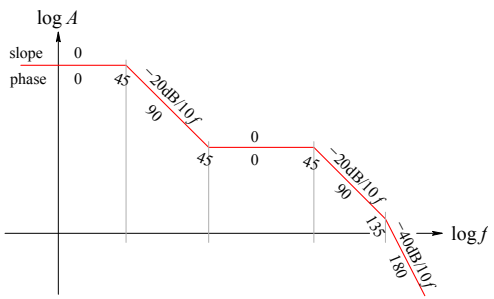


Fig. 5.3.15: An arbitrary example of the phase angle estimated from the gain magnitude, its slopes and various breakpoints. If two breakpoints are relatively close the phase would not actually reach the value predicted from the slope value, but an intermediate value instead.

In the same manner, along the amplifier open loop gain asymptotes, we draw the noise gain, as in Fig. 5.3.16, and we look at the crossover point of these two characteristics. Two important parameters can be derived from this plot: the first is the amount of gain at the crossover frequency f_{c}; the second is the relative slope difference between the two lines at f_{c}, which also serves as an indication of their phase difference.

If the available gain at f_{c} is greater than unity the phase difference determines the amplifier stability. The feedback can be considered 'negative' and the amplifier operation stable if the loop phase margin is at least 45°; this means that, if a 360° phase shift is 'positive', the maximum phase shift within the feedback loop must always be less than 315° (if $A(f_{\mathrm{c}}) > 1$). Since the inverting input provides 180°, the total phase shift of the remaining amplifier stages (secondary poles) and the feedback network should never exceed 135°. Note also that a phase margin of 90° or more results in a smooth transient response; for a phase margin between 90° and 45° an increasing amount of peaking would result.

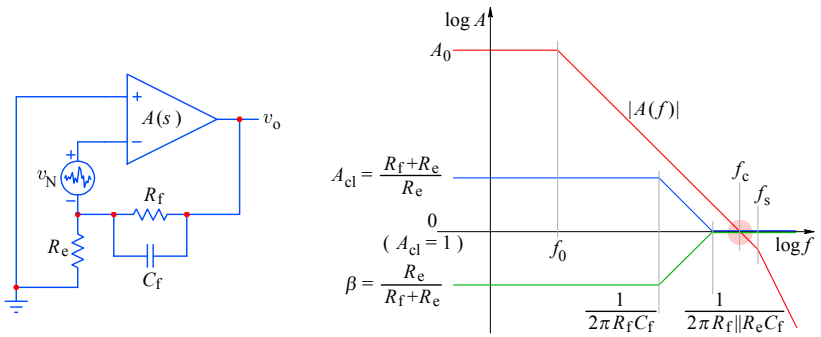


Fig. 5.3.16: Voltage feedback amplifier noise gain is derived from the equivalent circuit noise model (a noise generator in series with the input of a noiseless amplifier). The Bode plot shows the relationships between the most important parameters. Here $|A(f)|$ is the open loop gain magnitude, f_0 is the dominant pole, and f_s is the secondary pole, owed to the slowest internal stage. The inverse of the feedback attenuation β is the noise gain and it is equivalent to the amplifier closed loop gain A_{cl}. Note that the noise gain is flat up and beyond the open loop crossover f_c, owed to the zero at $1/(2\pi R_f||R_eC_f)$. The amplifier is stable since the noise gain crosses the open loop gain at a point where their slope difference is $20\,\mathrm{dB}/10f$. If the amplifier open loop gain was higher, the gain at the secondary pole (at f_s) would be higher than unity and the slope difference would be $40\,\mathrm{dB}/10f$. Then, the increased phase (135° at f_s and approaching 180° above), along with the 180° of the amplifier inverting input, would make the feedback positive (→360°) and the amplifier would oscillate.

Now, let us find the noise gain of the voltage feedback amplifier in Fig. 5.3.16.

Note that while there is some feedback available the amplifier tries to keep the difference between the inverting and non-inverting input as small as its open loop gain allows; so, with a high open loop gain, the input voltage difference tends to be zero (plus the DC voltage offset).

Note also that, in order to keep track of the phase inversion by the amplifier, we have added polarity indicators to the noise generator. If the '+' side of the noise generator v_N tries to push the inverting input positive, the output voltage must go negative to compensate it.

With C_f in parallel with R_f we shall have:

$$\frac{v_o}{v_N} = -\left(1 + \frac{R_f}{R_e} \cdot \frac{\dfrac{1}{C_f R_f}}{s + \dfrac{1}{C_f R_f}}\right) \tag{5.3.45}$$

which we can also rewrite as:

$$\frac{v_o}{v_N} = -\left(\frac{\dfrac{1}{C_f R_f}}{s + \dfrac{1}{C_f R_f}} + \frac{s + \dfrac{1}{C_f R_e}}{s + \dfrac{1}{C_f R_f}}\right) \tag{5.3.46}$$

Here we have a pole at $1/C_f R_f$ and a zero at $1/C_f R_e$. Eq. 5.3.46 is the noise gain (and also the closed loop gain A_{cl}; see the two distinct breakpoint frequencies in Fig. 5.3.16). The inverse of this is the feedback attenuation β.

With the open loop gain as shown in Fig. 5.3.16 the amplifier is stable, since the noise gain crosses over the open loop gain at f_c, where the open loop and closed loop slope difference is 20 dB/decade, and the associated phase shift is (nearly) 90°. In addition to the 180° of the amplifier inverting input, the total phase angle is then 270°. The minimum phase margin for a stable amplifier would be 45° and here we have 90° (360 − 270), so the feedback can still be considered "negative".

However, if the open loop gain was higher (and if the poles remain at the same frequencies) the gain at f_s (the frequency of the secondary pole) can be greater than unity. In this case at the crossover of the noise gain and open loop gain the slope difference will be 40 dB/10f, with the associated phase of 135° at f_s and approaching 180° above. The feedback will become 'positive' and, with the gain at f_s greater than unity, the amplifier would oscillate.

In the case of the current feedback amplifier in Fig. 3.5.17 we first note that instead of gain our Bode plot shows the feedback **impedances** and the amplifier **transimpedance**, all as functions of frequency.

Intuitively speaking, a capacitance C_f in parallel to R_f would reduce the impedance of the feedback network at high frequencies, thus also reducing the closed loop gain. However, intuition is misleading us: since the current feedback system bandwidth is inversely proportional to the feedback impedance in the f-branch (as demonstrated by Eq. 5.3.24), the addition of C_f increases the bandwidth. By itself this would be welcome, but note that at the crossover frequency f_c the slope difference between the transimpedance Z_T and the 'noise transimpedance' (in analogy with noise gain) is 40 dB/10f, causing a phase shift of 180°. This means that at f_c the feedback current becomes positive and the amplifier will oscillate.

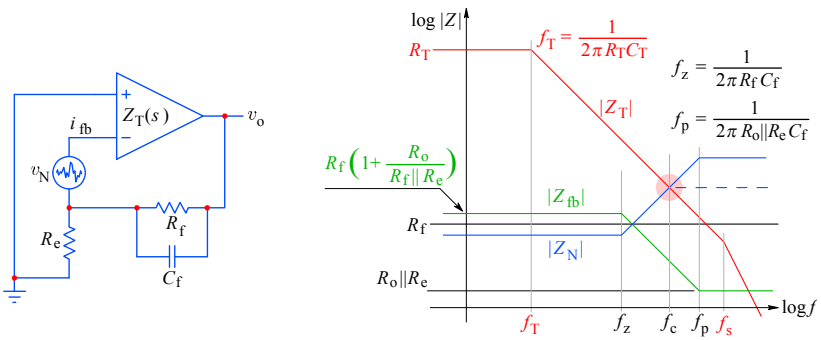


Fig. 5.3.17: For the current feedback amplifier we draw the impedances, not gain. The feedback network impedance $|Z_{fb}|$, as seen from v_o, is slightly higher than R_f at DC (owing to R_o, the inverting input resistance) and falls to $R_o||R_e$ at high frequencies; its inverse (about R_f) is the amplifier noise transimpedance, $|Z_N|$. The feedback network pole becomes the zero of the noise transimpedance: $s_z = -1/C_f R_f$ ($f_z = |s_z|/2\pi$); likewise, the feedback zero becomes the noise transimpedance pole $s_p = -1/C_f(R_o||R_e)$, ($f_p = |s_p|/2\pi$). At f_c the crossover with $|Z_T|$, the slope difference is 40 dB/10f and the relative phase angle is 180°; the amplifier will inevitably oscillate, even if the secondary pole is far away and its Z_T breakpoint is well below R_f. The dashed line is the transimpedance required for stability, realized by an R in series with C_f.

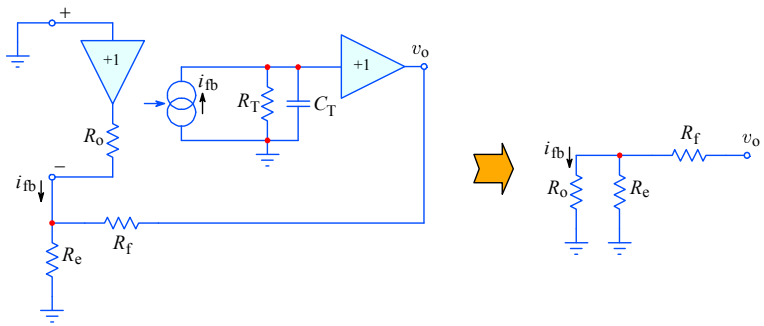


Fig. 5.3.18: The current feedback amplifier and its 'noise transimpedance' equivalent, v_o/i_{fb}.

We can find the noise transimpedance as simply as we found the noise gain for voltage feedback amplifiers. By assuming that the feedback current is noise generated, from the equivalent circuit in Fig. 5.3.18 we calculate the ratio of the output voltage and the feedback current:

$$\frac{v_o}{i_{fb}} = R_o + R_f\left(1 + \frac{R_o}{R_e}\right) \tag{5.3.47}$$

By adding a capacitance C_f in parallel with R_f the noise transimpedance becomes:

$$\frac{v_o}{i_{fb}} = R_o + R_f \frac{\dfrac{1}{C_f R_f}}{s + \dfrac{1}{C_f R_f}}\left(1 + \frac{R_o}{R_e}\right) \tag{5.3.48}$$

and this equation is represented by $|Z_N|$ in Fig. 5.3.17.

In most practical cases there will be stray capacitances in parallel to both R_f and R_e, and in addition between both inputs, as well as from the non-inverting input to ground, and also from output to ground. A real world situation can be rather complicated.

As we have shown, current feedback amplifiers are extremely sensitive to any capacitances within the negative feedback loop. This means that whole families of circuits (such as integrators, differentiators, some filter topologies, current amplifiers, I to V converters, logarithmic amplifiers, etc.) can not be realized in the same way as with conventional amplifiers. Fortunately, there are alternative ways of performing the same functions and some of the most common ones are shown in Fig. 5.3.19 for a quick comparison:

- an inverting integrator can be implemented using two current feedback (CFB) opamps, with the bonus of providing both the inverting and non-inverting configuration within the same circuit;
- a single pole inverting filter amplifier can be implemented by exploiting the internal capacitance C_T and the external feedback resistor R_f (useful for high frequency cut off; for lower frequencies a high value of R_f is impractical since it would cause a large voltage offset, owing to a large input bias current);

- for filters the Sallen–Key non-inverting configuration is recommended for use with CFB amplifiers. This configuration can be easily cascaded (using second- and third-order sections) to realize multi-pole high order filters, in the same way as the 'multiple feedback' inverting configuration can be cascaded;
- current sources, such as some digital to analog converters, photo-diodes, etc., have a relatively large capacitance in parallel. If a CFB amplifier is used as an inverting current to voltage converter, the source capacitance must be compensated. The gain setting resistors R_f and R_e are in series with the compensation capacitor, thus preventing instability at high frequencies. Also, these resistors can be used to scale the compensation capacitor value to more practical values of 10–20 pF, instead of 1 pF or less which would normally be needed for high speed response.

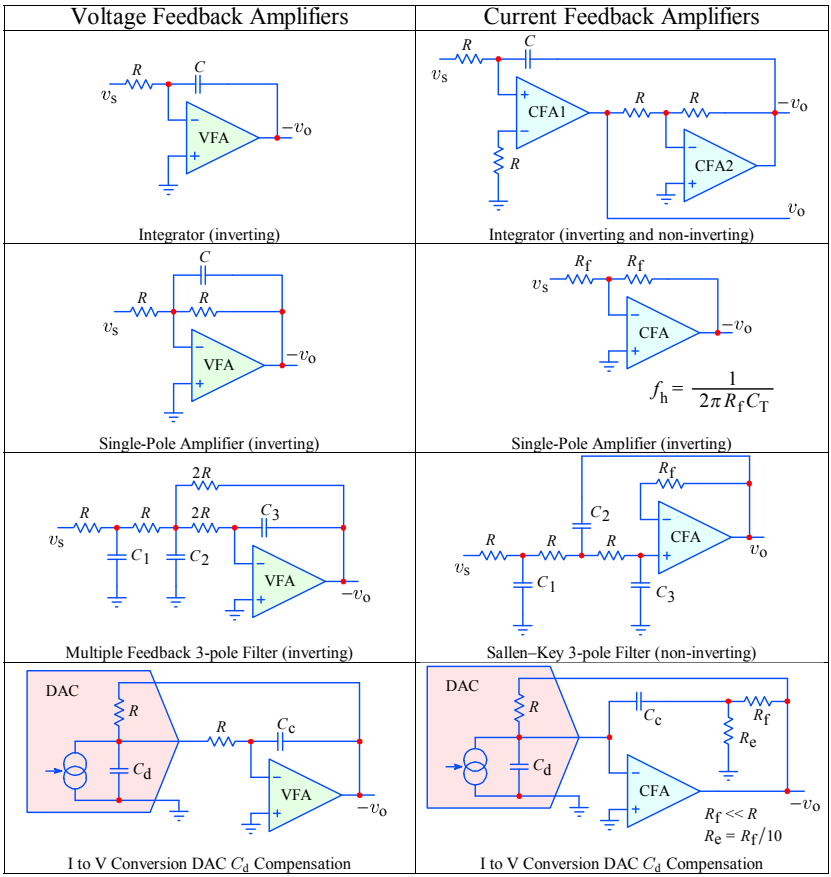


Fig. 5.3.19: Functionally equivalent circuits with conventional and current feedback amplifiers. Integrators, filters and current to voltage converters in inverting configurations cannot be achieved using a single CF amplifier. However, two-amplifier circuits can provide inherent amplifier pole compensation, which is very important at high frequencies. Filters can be realized in the non-inverting configuration. And feedback capacitance can be isolated by a resistive divider.

5.3.6 Feedback Controlled Gain Peaking

As we have just seen, the current feedback amplifier is sensitive to capacitive loading of its inverting input. But stray capacitances are unavoidable and they can cause significant gain peaking and even oscillations in high speed low gain designs. Fortunately, the current feedback topology offers a simple way of controlling this by choosing such values of R_f and R_e which would set the bandwidth and the gain to the optimum. Although CFB amplifier performance is usually optimized for a specific value of the bandwidth defining resistance R_f, ample trade-off range is often possible.

However, as we have learned in Part 4, in multi-stage amplifiers it is necessary to optimize the system as a whole and not just each stage individually. The possibility of controlling the gain peaking with feedback resistors lends itself nicely to our purpose. In practice we would have to iteratively adjust both R_f and R_e to obtain the desired gain, bandwidth, and peaking. What we would like is to be able to adjust the bandwidth and peaking by a single resistor, without affecting the gain. The circuit in Fig. 5.3.20 does just that.

The feedback resistors R_f and R_e should be chosen for the gain required, but with the lowest possible values, which would not overload the output stage. Then the resistor between the feedback divider and the inverting input, R_b, should be adjusted for the required response, assuming a fixed value of the stray capacitance.

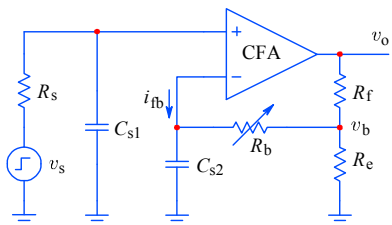


Fig. 5.3.20: This circuit exploits the ability of current feedback amplifiers to adjust the bandwidth and gain peaking independently of the gain. The price to pay is the lower slew rate limit. See the frequency and the step response below.

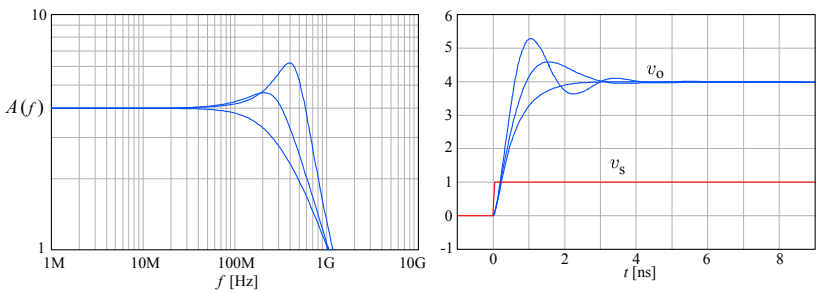


Fig. 5.3.21: **a)** Frequency response; **b)** Step response of the amplifier in Fig. 5.3.20. The closed loop gain $A_{cl} = 1 + R_f/R_e = 4$, $R_f = 150\,\Omega$, $R_e = 50\,\Omega$, the source resistance $R_s = 50\,\Omega$, the stray capacitances, $C_{s1,2} = 1\,\text{pF}$, the amplifier transcapacitance $C_T = 1\,\text{pF}$, while R_b is varied from $150\,\Omega$ for highest peaking to $750\,\Omega$ for lowest peaking.

Note, however, that in this way we lose the current on demand property of the CFB amplifier, since R_b will reduce the slew rate.

In a similar manner as was done for passive circuits in Part 2 and in Sec. 5.1, the resulting gain peaking can be used to improve the step response of a multi-stage system. As shown in Fig. 5.3.21, the gain peaking reveals the amplifier resonance, which decreases with increasing R_b, while the DC gain remains almost unchanged.

5.3.7 Improved Voltage Feedback Amplifiers

The lessons learned from the current feedback technology can be used to improve conventional voltage feedback amplifiers.

Besides the improved semiconductor manufacturing technology, basically there are two approaches: one is to take the voltage feedback amplifier and modify it using the techniques of current feedback to avoid its week points. One such example is shown in Fig. 5.3.22. The other way, like the circuit in Fig. 5.3.23, is to take the current feedback amplifier and modify it to make it appear to the outside world as a voltage feedback amplifier.

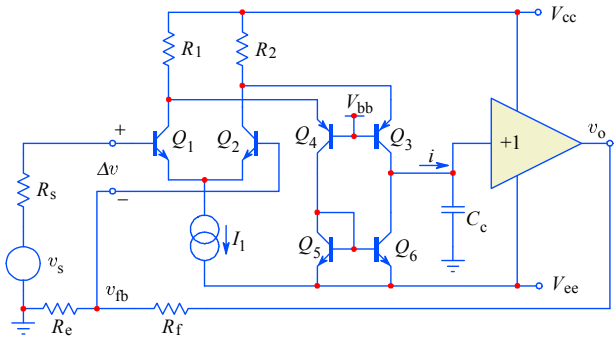


Fig. 5.3.22: The voltage feedback amplifier, improved. The transistors Q_{1-4} form a differential 'folded' cascode, which drives the current mirror $Q_{5,6}$. In this way the input is a conventional high impedance differential, but the dominant pole compensation capacitor C_c is grounded, eliminating the Miller effect. This circuit still exhibits slew rate limiting, although at much higher frequencies. Typical examples of this configuration are Analog Devices' AD-817 and Burr–Brown's OPA-640.

The differential folded cascode Q_{1-4} and the current mirror $Q_{5,6}$ of Fig. 5.3.22 can be modeled by a transconductance, g_m, driven by the input voltage difference, $\Delta v = v_s - v_{fb}$. Here v_s is the signal source voltage and v_{fb} is the feedback voltage, derived from the output voltage v_o and the feedback network divider, $R_e/(R_f + R_e)$. The current $i = \Delta v\, g_m$ drives the output buffer and the capacitance C_c:

$$v_o = i\,\frac{1}{s\,C_c} = g_m(v_s - v_{fb})\,\frac{1}{s\,C_c} = g_m\left(v_s - v_o\,\frac{R_e}{R_f + R_e}\right)\frac{1}{s\,C_c} \qquad (5.3.49)$$

From this we obtain:

$$v_o\left(1 + \frac{R_e}{R_f + R_e}\cdot\frac{g_m}{s\,C_c}\right) = v_s\,\frac{g_m}{s\,C_c} \qquad (5.3.50)$$

and, finally, the transfer function:

$$\frac{v_o}{v_s} = \left(1 + \frac{R_f}{R_e}\right) \frac{\dfrac{R_e}{R_f + R_e} \cdot \dfrac{g_m}{C_c}}{s + \dfrac{R_e}{R_f + R_e} \cdot \dfrac{g_m}{C_c}} \tag{5.3.51}$$

If we compare this with a general first-order amplifier transfer function:

$$\frac{v_o}{v_s} = A_{cl} \frac{-s_1}{s - s_1} \tag{5.3.52}$$

we see that the closed loop gain is, as usual, $A_{cl} = 1 + R_f/R_e$, whilst the amplifier closed loop pole is $s_1 = -R_e\, g_m/C_c(R_f + R_e)$, and therefore the cut off frequency is an inverse function of the closed loop gain, just as in voltage feedback amplifiers.

A similar situation is encountered in Fig. 5.3.23, where the current i_b charging C_T is set by the input voltage difference and R_b: $i_b = \Delta v/R_b$.

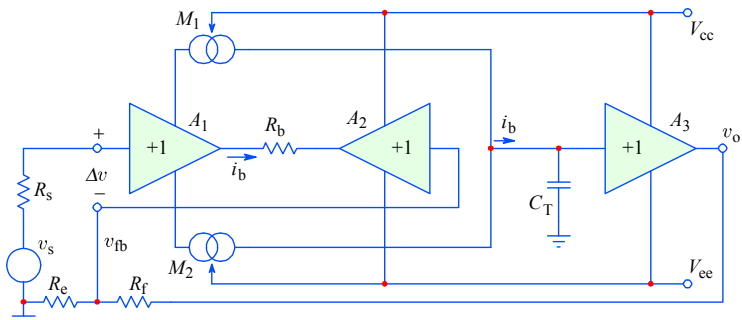


Fig. 5.3.23: The basic current feedback amplifier (A_1, A_3, M_1, M_2) is improved by adding another buffer, A_2, which presents a high impedance to the feedback divider, R_f and R_e; an additional resistor, R_b, now takes the role of converting the voltage feedback into current and provide bandwidth setting. Like the original current feedback amplifier, this circuit is also (almost) free from slew rate limiting. However, the closed loop bandwidth is, as in voltage feedback amplifiers, gain dependent. A typical representative of this configuration is Analog Devices' OP-467.

The output voltage is:

$$v_o = i_b \frac{1}{s\,C_T} = \Delta v \frac{1}{s\,C_T R_b} = \left(v_s - v_o \frac{R_e}{R_f + R_e}\right) \frac{1}{s\,C_T R_b} \tag{5.3.53}$$

so the transfer function is:

$$\frac{v_o}{v_s} = \left(1 + \frac{R_f}{R_e}\right) \frac{\dfrac{R_e}{R_f + R_e} \cdot \dfrac{1}{C_T R_b}}{s + \dfrac{R_e}{R_f + R_e} \cdot \dfrac{1}{C_T R_b}} \tag{5.3.54}$$

The closed loop gain is the same as in the previous case, whilst the pole is $s_1 = -R_e/C_T R_b(R_f + R_e)$, so the closed loop cutoff frequency is again an inverse function of the closed loop gain.

One of the important parameters in integrated circuit design is the available bandwidth vs. quiescent current. The average technology achievement around the year 1990 was about 10 MHz/mA and around the year 2000 it was already about 100 MHz/mA; the figure is steadily rising. With the ever increasing number of transistors on a silicon chip it is important to keep this value high. The implementation of structures which convey the supply current efficiently to the signal helps to reduce the waste of power.

An example, named the 'Quad Core' [Ref. 5.32], is shown in Fig. 5.3.24. This is an interesting combination of circuits in Fig. 5.3.22 and 5.3.23, where the two input buffers, formed by $Q_{1,2}$ and $Q_{3,4}$, combine their currents by the current mirrors, $Q_{5–8}$, driving the following gain stage $Q_{9,10}$ in a differential push–pull mode.

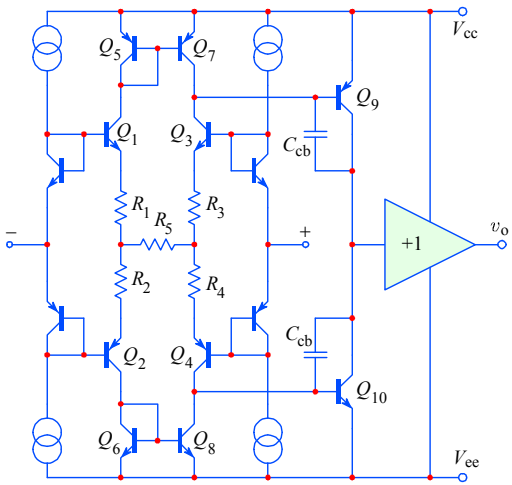


Fig. 5.3.24: An interesting combination of circuits in Fig. 5.3.22 and 5.3.23 is the so called 'Quad Core' structure, [Ref. 5.32]. Here both the inverting and the non-inverting input buffer currents are combined by the current mirrors to drive the C_{cb} of $Q_{9,10}$. The non-labeled transistors provide the V_{be} compensation for $Q_{1–4}$. Typical representatives are Analog Devices' AD-8047, AD-9631, AD-8041 and others.

The current available to charge the C_{cb} capacitances of $Q_{9,10}$ is set by the input voltage difference and R_5. This current is effectively doubled by the input structure, thus increasing the bandwidth, the loop gain, and linearity. A further bandwidth improvement is achieved by the low impedance of $Q_{7,8}$, which are practically fully open and so provide a tight control of the $Q_{9,10}$ base voltages, reducing the Miller effect considerably. The circuit behaves as a voltage feedback amplifier with the advantages of low offset and high loop gain and with a bandwidth and slew rate limiting close to that of current feedback amplifiers.

The output buffer stage can also be improved for greater current handling efficiency. An example is shown in Fig. 5.3.25.

Here the collectors of $Q_{2,3}$ and $Q_{1,4}$ are summed and mirrored by $Q_{5,7}$ and $Q_{6,8}$, respectively, and finally added to the output load current. With appropriate bias this scheme allows a reduction of the quiescent power supply current to just one third

of the conventional buffer, whilst not compromising the full power bandwidth. At the same time, the circuit has a comparable loading capability and offers better linearity.

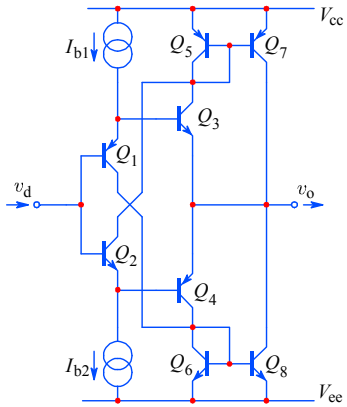


Fig. 5.3.25: Output buffer stage with improved current handling.

5.3.8 Compensating Capacitive Loads

Another very important property of high speed opamps is their ability to drive capacitive loads. To the amplifier in Fig. 5.3.26, because of its non-zero output resistance R_o, the capacitive load C_L adds a high frequency pole within the feedback loop. The feedback becomes frequency dependent and the phase margin is lowered, thus compromising the stability.

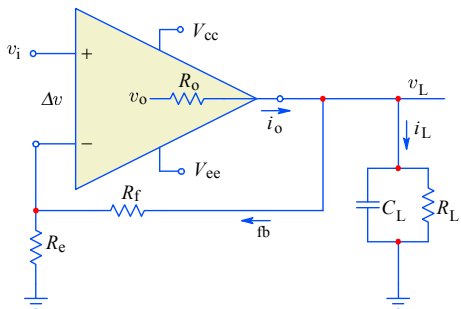


Fig. 5.3.26: Owing to the non-zero output impedance a capacitive load adds another pole within the feedback loop. If the closed loop gain is too low the resulting increase in phase can make the feedback positive at high frequencies, instead of negative, destabilizing the amplifier.

The load voltage v_L can be expressed as a function of the internal voltage v_o (seen when no load is present), the factor D_R, and the frequency dependence:

$$v_L = v_o \, D_R \, \frac{-s_L}{s - s_L} \tag{5.3.55}$$

Here D_R is the resistive divider formed by the output resistance R_o, the load resistance R_L and the total resistance of the feedback divider $R_\mathrm{f} + R_\mathrm{e}$:

$$D_R = \frac{\dfrac{R_\mathrm{L}(R_\mathrm{f} + R_\mathrm{e})}{R_\mathrm{L} + R_\mathrm{f} + R_\mathrm{e}}}{R_\mathrm{o} + \dfrac{R_\mathrm{L}(R_\mathrm{f} + R_\mathrm{e})}{R_\mathrm{L} + R_\mathrm{f} + R_\mathrm{e}}} \tag{5.3.56}$$

The pole s_L is formed by the load capacitance C_L and the equivalent resistance seen by it, R_q, whilst ω_L is the appropriate cut off frequency:

$$s_\mathrm{L} = -\frac{1}{R_\mathrm{q} C_\mathrm{L}}\,; \qquad \omega_\mathrm{L} = |s_\mathrm{L}| \tag{5.3.57}$$

R_q is simply the parallel combination of all the resistances at the output node:

$$R_\mathrm{q} = \frac{1}{\dfrac{1}{R_\mathrm{o}} + \dfrac{1}{R_\mathrm{L}} + \dfrac{1}{R_\mathrm{f} + R_\mathrm{e}}} \tag{5.3.58}$$

The internal output voltage, v_o, is a function of the input voltage difference, Δv, and the amplifier open loop gain A, which, in turn, is also a function of frequency, $A(s)$:

$$v_\mathrm{o} = A(s)\,\Delta v \tag{5.3.59}$$

The input voltage difference is, of course, the difference between the signal source voltage and the output (load) voltage, attenuated by the feedback resistors:

$$\Delta v = v_\mathrm{s} - v_\mathrm{L}\,\frac{R_\mathrm{e}}{R_\mathrm{f} + R_\mathrm{e}} \tag{5.3.60}$$

The open loop gain $A(s)$ is defined by the DC open loop gain A_0 and the frequency dependent term owed to the amplifier dominant pole at the frequency ω_0:

$$A(s) = A_0\,\frac{\omega_0}{s + \omega_0} \tag{5.3.61}$$

With this in mind, we can express the internal output voltage:

$$v_\mathrm{o} = A_0\,\frac{\omega_0}{s + \omega_0}\left(v_\mathrm{s} - v_\mathrm{L}\,\frac{R_\mathrm{e}}{R_\mathrm{f} + R_\mathrm{e}}\right) \tag{5.3.62}$$

and by inserting this back in Eq. 5.3.53 we have the load voltage:

$$v_\mathrm{L} = A_0\,\frac{\omega_0}{s + \omega_0}\left(v_\mathrm{s} - v_\mathrm{L}\,\frac{R_\mathrm{e}}{R_\mathrm{f} + R_\mathrm{e}}\right) D_R\,\frac{\omega_\mathrm{L}}{s + \omega_\mathrm{L}} \tag{5.3.63}$$

which we solve for v_L explicitly:

$$v_\mathrm{L}\left(1 + \frac{R_\mathrm{e}}{R_\mathrm{f} + R_\mathrm{e}}\,A_0\,\frac{\omega_0}{s + \omega_0}\,D_R\,\frac{\omega_\mathrm{L}}{s + \omega_\mathrm{L}}\right) = v_\mathrm{s}\,A_0\,\frac{\omega_0}{s + \omega_0}\,D_R\,\frac{\omega_\mathrm{L}}{s + \omega_\mathrm{L}} \tag{5.3.64}$$

Now we can write the transfer function:

$$\frac{v_\mathrm{L}}{v_\mathrm{s}} = \frac{A_0\, D_R \dfrac{\omega_0\,\omega_\mathrm{L}}{(s+\omega_0)(s+\omega_\mathrm{L})}}{1+\dfrac{R_\mathrm{e}}{R_\mathrm{f}+R_\mathrm{e}} A_0\, D_R \dfrac{\omega_0\,\omega_\mathrm{L}}{(s+\omega_0)(s+\omega_\mathrm{L})}} \tag{5.3.65}$$

where we separate the closed loop gain term:

$$\frac{v_\mathrm{L}}{v_\mathrm{s}} = \frac{R_\mathrm{f}+R_\mathrm{e}}{R_\mathrm{e}} \cdot \frac{\dfrac{R_\mathrm{e}}{R_\mathrm{f}+R_\mathrm{e}} A_0\, D_R \dfrac{\omega_0\,\omega_\mathrm{L}}{(s+\omega_0)(s+\omega_\mathrm{L})}}{1+\dfrac{R_\mathrm{e}}{R_\mathrm{f}+R_\mathrm{e}} A_0\, D_R \dfrac{\omega_0\,\omega_\mathrm{L}}{(s+\omega_0)(s+\omega_\mathrm{L})}} \tag{5.3.66}$$

and, by multiplying the numerator and the denominator by $(s+\omega_0)(s+\omega_\mathrm{L})$, which we expand into a polynomial, we obtain the expression for the transfer function. Clearly it is a second-order function of frequency:

$$\frac{v_\mathrm{L}}{v_\mathrm{s}} = \frac{R_\mathrm{f}+R_\mathrm{e}}{R_\mathrm{e}} \cdot \frac{\dfrac{R_\mathrm{e}}{R_\mathrm{f}+R_\mathrm{e}} A_0\, D_R\, \omega_0\,\omega_\mathrm{L}}{s^2 + s(\omega_0+\omega_\mathrm{L}) + \omega_0\omega_\mathrm{L}\left(1+\dfrac{R_\mathrm{e}}{R_\mathrm{f}+R_\mathrm{e}} A_0\, D_R\right)} \tag{5.3.67}$$

The product of the poles, $s_1 s_2$, is a function of not just ω_0 and ω_L, but also of the open loop gain A_0 and the closed loop feedback dividers, D_R and $R_\mathrm{e}/(R_\mathrm{f}+R_\mathrm{e})$ (refer to Appendix 2.1 to find the system poles of a 2nd-order function). Since the output resistance, R_o, is usually much lower than both the load resistance R_L and the feedback resistances $R_\mathrm{f}+R_\mathrm{e}$, the output divider D_R is usually between 0.9 and 1. The system's stability is therefore dictated by the amount of loop gain when $s \to \omega_\mathrm{L}$. Thus close to ω_L the loop gain will be higher than 1 either if A_0 is very high, or if ω_L is relatively low and $R_\mathrm{f} \to 0$, that is, if the closed loop gain approaches unity!

This is often counter-intuitive, not just to beginners, but sometimes even to experienced engineers. Usually, if we want to enhance the amplifier's stability, we increase the feedback at high frequencies by placing a capacitor in parallel to R_f. However, in the case of a capacitive load the amplifier will be turned into an oscillator by that procedure. In contrast, the stability will improve if we increase the closed loop gain (increase R_f or decrease R_e). This is illustrated in Fig. 5.3.27, where the gain and the phase are plotted for the three values of R_f (∞, $4R_\mathrm{e}$ and 0) and the capacitive load is such, that the loop gain at $f_\mathrm{L} \approx 2 \cdot 10^4\, f_0$ is about 3.

Conventional opamp compensation schemes (consisting usually of a resistor or a series RC network, connected between both inputs, thus increasing the noise gain, without affecting the signal gain) increase the stability at the expense of either the gain, or the bandwidth, or both! Conventional compensation should be used as the last resort only, when the circuit must meet an unknown load capacitance, which can vary considerably.

In fixed applications, when the capacitive load is known, or is within a narrow range, it is much better to compensate the load by inductive peaking, as we have seen in Part 2. The simplest approach is shown in Fig. 5.3.28, where the load impedance appears real to the amplifier output, so that the feedback is not frequency dependent.

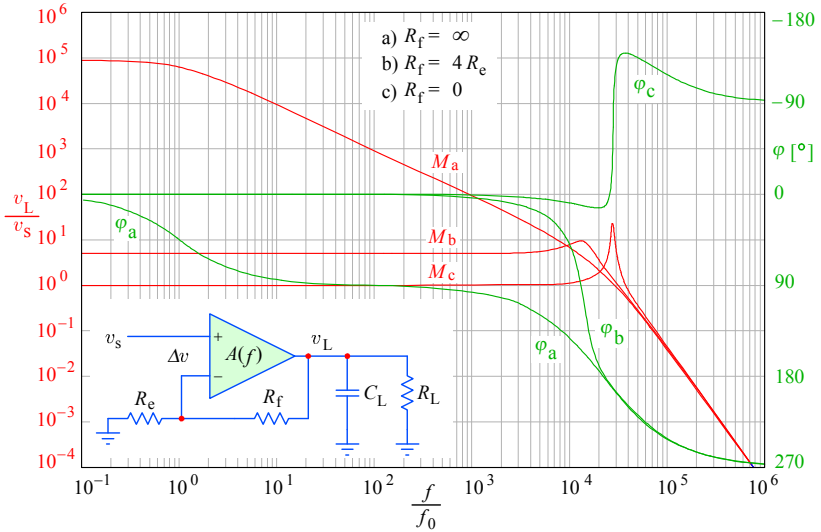


Fig. 5.3.27: An amplifier driving a capacitive load can become unstable if its close loop gain is decreased too much: a) with no feedback and b) with a gain of 5, the amplifier is stable, although in the latter case there is already a pronounced peaking; whilst in c) with the closed loop gain reduced to 1 the peaking is very high and the phase goes over 360° and oscillation is excited at the frequency at which the phase equals 360°.

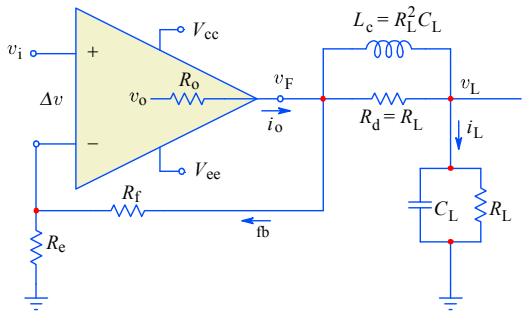


Fig. 5.3.28: Capacitive load compensation which makes the load to appear real and equal to R_L to the opamp. This works for fixed load impedances.

Here the inductance L_c and its parallel damping resistance, R_d, are chosen so that $R_d = R_L$ and $L_c = R_L^2\,C_L$, and the amplifier sees an impedance equal to R_L from DC up to the frequency at which the coil stray capacitance starts to influence the compensation. With a careful inductance design the frequency at which this happens can be much higher than the critical amplifier frequency.

However, even with such compensation, the bandwidth can be lower than desired, since the compensation network forms a low pass filter with the load, and the value of the inductance is influenced by both the load resistance R_L and the load capacitance C_L.

The filter transfer function is:

$$\frac{v_L}{v_F} = \frac{\dfrac{1}{L_c C_L} + s\dfrac{1}{R_L C_L}}{s^2 + s\dfrac{2}{R_L C_L} + \dfrac{1}{L_c C_L}} \tag{5.3.68}$$

The cut off frequency is $\omega_h = 1/\sqrt{L_c C_L} = 1/R_L C_L$ and that is much lower than ω_L of Eq. 5.3.57, which would apply if the amplifier could be made stable by some other means. If R_L and C_L can be separated, it is possible to build a 2-pole series peaking or a T-coil peaking, tuned to form a 3-pole system together with the pole associated with the amplifier closed loop cut off. This procedure is similar to the one described in Part 2 and also in Sec. 5.1, so we leave it as an exercise to the reader.

Another compensation method, shown in Fig. 5.3.29, is to separate the AC and DC feedback path by a small resistance R_c in series with the load and a feedback bridging capacitance C_c:

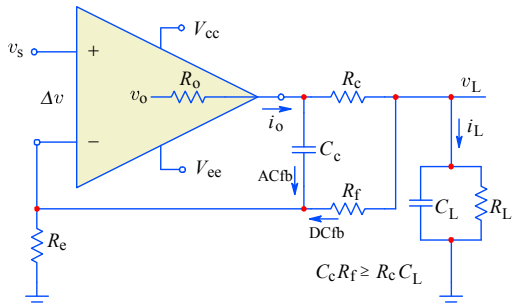


Fig. 5.3.29: The capacitive load is separated from the AC feedback path by a small resistor R_c in series with the output; owing to the capacitance C_c this type of compensation can be applied only to voltage feedback unity gain compensated amplifiers.

This type of compensation can be very effective, since very small values of R_c can be used (5–15 Ω or so), lowering the bandwidth only slightly; however, due to the feedback bridging capacitance C_c, it can be applied only to voltage feedback unity gain compensated amplifiers; it can not be used for current feedback amplifiers.

A more serious problem is the fact that, in some applications, the load capacitance would vary considerably; for example, some types of fast AD converters have their input capacitance code dependent (thus also signal level dependent!). It is therefore desirable to design the amplifier output stage with the lowest possible output resistance and employ a compensation scheme which would work for a range of capacitance values.

Fig. 5.3.30 shows the implementation found in some CFB opamps, where the compensation network, formed by C_c and R_c, is in parallel with the output buffer. With a high impedance load the voltage drop on the output resistance R_o is small and the current through the compensation network is low; but with a capacitive load or other low impedance load the output current causes a high voltage drop on R_o, and consequently the current through C_c increases. Effectively the series combination of

C_c and C_L is added in parallel to C_T, thus lowering the system open loop bandwidth in proportion with the load and keeping the loop stable.

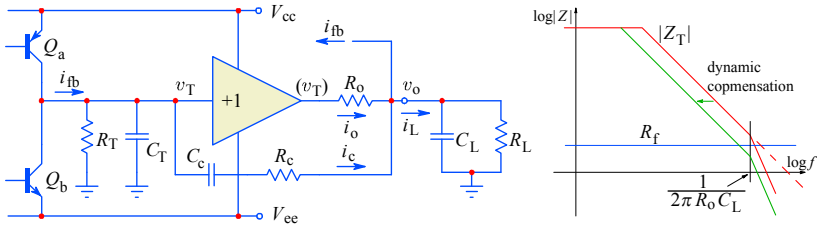


Fig. 5.3.30: The output buffer with a finite output resistance R_o would, when driving a capacitive load C_L, present an additional pole within the feedback loop (taken from v_o), which could condition the amplifier stability. The compensation network, formed by a serial connection of C_c and R_c, draws part of the feedback current to the output node, effectively increasing C_T in proportion to the load, reducing the transimpedance and preserving the closed loop stability.

In Fig. 5.3.30 we have three RC impedances: Z_T is the CFB amplifier transimpedance consisting of R_T and C_T in parallel; Z_c is the compensation impedance consisting of R_c and C_c in series; and Z_L is the load impedance consisting of R_L and C_L in parallel. The output unity gain buffer is assumed to be ideal and the real circuit is modeled by its output resistance R_o.

If v_T is its input voltage, the output voltage v_o will be lower by $i_o R_o$, where i_o is the output current. Since Z_c is connected between the buffer input and the load, the voltage over Z_c is equal to $i_o R_o$, so:

$$v_T - v_o = i_o R_o \tag{5.3.69}$$

$$i_c = \frac{v_T - v_o}{Z_c} = \frac{i_o R_o}{Z_c} \tag{5.3.70}$$

The transimpedance Z_T is driven by the feedback current i_{fb}; the voltage v_T, which in an ideal case would be equal to $i_{fb} Z_T$, is now lower, because part of the current is stolen by the compensation network Z_c:

$$v_T = (i_{fb} - i_c) Z_T \tag{5.3.71}$$

With a few simple substitutions we obtain:

$$v_o = i_{fb} Z_T - i_o R_o \left(\frac{Z_T}{Z_c} + 1 \right) \tag{5.3.72}$$

This equation shows that the original transimpedance equation (Eq. 5.3.18) is modified by the output current and the Z_T/Z_c ratio. Thus a capacitive load, which would draw high currents at high frequencies (or at the step transition), will automatically lower the system open loop cut off frequency. Consequently the gain at high frequencies is reduced so that the closed loop crossover remains well above the secondary pole (created by R_o and C_L).

Note that the distortion at high frequencies of the compensated amplifier will be worse than that of a non-compensated one.

5.3.9 Fast Overdrive Recovery

The ability to resume linear signal amplification after a prolonged large overdrive is one of the most important oscilloscope design considerations. The average oscilloscope user is often tempted to zoom in on a small detail of a relatively large signal to inspect or debug the performance of electronic equipment.

With the input sensitivity turned high, depending on the attenuation and gain settings, various stages of a multi-stage amplifier can be overdriven into their non-linear region or even saturated, whilst others can remain within their linear region; at some settings it is the output stage that is overdriven first, at others it will be so with the input stage (this is because the input attenuator is always varied in steps of ×10, the following attenuation and gain steps are usually lower, ×2 or ×2.5). When overdriven, different amplifier stages will have different electrical and thermal histories, so that when the signal falls back within the linear range the circuit will not re-balance immediately, but will take some time before it regains its original accuracy, often with many different time constants, characterized by relatively long rising or falling 'tails'.

For decades, Tektronix oscilloscopes excelled in fast recovery, far above all its competitors. Although the problem of overdrive recovery has never been easy to solve, in the old days of electronic tubes and discrete transistors the power supply voltage was always very high, allowing ample signal range. With modern ICs, having several transistors one on top of the other and with ever decreasing power supply voltages, the useful linear range is often only a volt or so. Therefore, special local limiting circuitry must be added to smoothly switch in and out with overdrive.

The overdrive problem is even more pronounced in many modern fast ADCs (e.g., the 'flash type', especially those with a two-stage pipeline architecture) whose input voltage range is only 1–2 V. When overdriven, they become slower; or they can generate large error codes, even if the overdrive level is just a few least significant bits (LSBs) above maximum. Such ADCs require a well controlled clipping amplifier to drive their input. With a voltage feedback opamp, the solution could be simple, by adding diodes shunting the feedback loop. Precise levels with sharp knees are needed, so a Schottky diode bridge with a biased Zener diode is often used. Fig. 5.3.31 shows one such possible circuit.

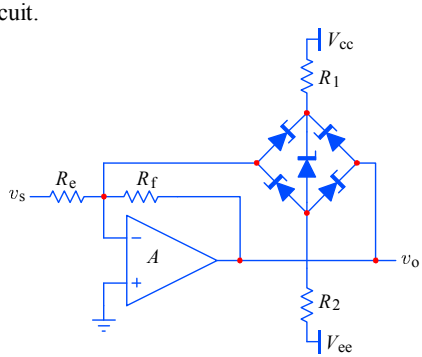


Fig. 5.3.31: The output level clipping is more precise if a biased Zener diode in a Schottky diode bridge is controlling the feedback. However, this circuit can be used only with voltage feedback unity gain compensated amplifiers.

Current feedback opamps do not like changes in feedback impedance, therefore a different output clipping scheme is used, having separate supply voltages for the output stage, as in Fig. 5.3.32. However, when the output reaches the level of its supply voltage the input stage loses feedback, so the input signal can overdrive the input differential stage. Usually we can prevent this by adding two anti-parallel Schottky diodes between the two inputs. Unfortunately, this would also increase the input capacitance.

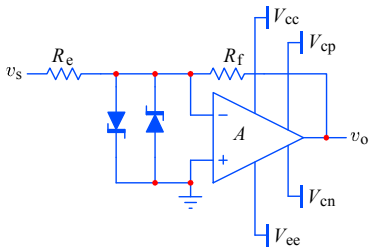


Fig. 5.3.32: The output buffer with a separate lower supply voltage can be used for output signal clipping with current feedback amplifiers. Since feedback is lost during clipping, the input stage must be protected from overdrive by anti-parallel Schottky diodes, which, inevitably, increase the input capacitance.

Another solution, often used with current feedback topology, is realized by limiting the voltage at the C_T internal node, using two normally closed voltage followers, as shown in Fig. 5.3.33. The addition of the voltage limiters increases the total capacitance at the C_T node, so, all other things being equal, limiting amplifiers tend to be slower. But with a careful design the bandwidth can still be quite high.

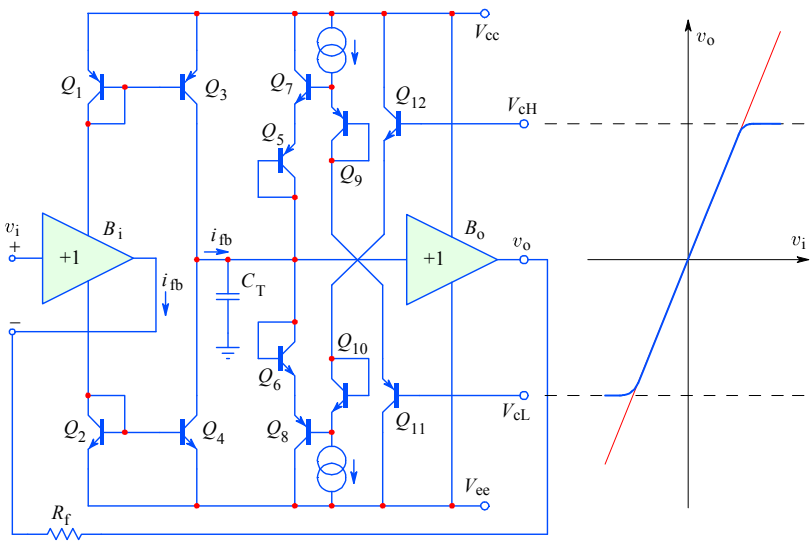


Fig. 5.3.33: Output signal clipping by limiting the internal C_T node of a current feedback amplifier. The transistors $Q_{5\text{–}8}$ are normally reverse biased. For positive voltages $Q_{6,8}$ ($Q_{5,7}$ for negative) start conducting only when the voltage at C_T reaches V_{cH} (V_{cL}).

Transistors $Q_{1–4}$ form the two current mirrors, which reflect the feedback current i_{fb} from the inverting input (the first buffer output) into the transimpedance node (at C_T). Transistors Q_5 and Q_7 are normally reverse biased and so are the B–E junctions of Q_6 and Q_8. The transistors $Q_{5,7}$ start to conduct when the C_T voltage (and consequently the output voltage v_o) falls below V_{cL}. Likewise, the transistors $Q_{6,8}$ conduct when the C_T voltage exceeds V_{cH}. When either of these transistors conduct, they diverge the mirrored current i_{fb} to one of the supplies. Note that the voltages which set the clipping levels can be as high as the supply voltage. Also, they can be adjusted independently, as long as $V_{cH} > V_{cL}$. Since only two transistors at a time are required to switch on or off, the limiting action, as well as the recovery from limiting, can be very fast.

The most common misconception of overdrive recovery, even amongst more experienced engineers, is the belief that short switching times can be achieved only if the transistors are prevented from being saturated by artificially keeping them within a linear signal range. It often comes as a surprise if this does not solve the problem or, in some cases, can make it even worse.

It is true that adding Schottky diodes to a TTL gate makes it faster than ordinary TTL, and the inherently non-saturating ECL is even faster. Fast recovery is ultimately limited by the so called *minority carrier storage time* within the semiconductor material, and it depends on the type and concentration of dopants which determine the mobility of minority charge carriers. However, in analog circuits the problem is radically different from digital circuits, since we are interested in not just how quickly the output will start to catch up with the input, but rather how quickly it will follow the input to within 0.1 %, or even 0.01 %. In current state of the art circuitry, the best recovery times are < 5 ns for 0.1 % error and < 25 ns for 0.01 %.

In this respect it is the thermal ‘tails’ that are causing trouble. Wideband circuits need more power than conventional circuits, so good thermal balance is critical. Careful circuit design is required in order to keep temperature differences low, both during linear and non-linear modes of operation.

To some extent, we have been dealing with thermal problems in Part 3. There we were discussing two-transistor circuits, such as the cascode and the differential amplifier. But the problem in multi-transistor circuits is that, even if it is differentially or complementary symmetrical, only one or two transistors will be saturated during overdrive, the rest of the circuit will either remain linear or cut off, which in this last case means cooling down. In saturation there is a low voltage across the device (usually a few tens of mV), so, even if the current through it is large, the power dissipation is low. Inevitably this results in different thermal histories of different parts of the circuit.

In integrated circuits the temperature gradients across the die are considerably lower than those between transistors in discrete circuits, but complex circuits can be large and heat conduction can be limited, so designers tend to reduce the power of auxiliary circuitry which is not essential for high speed signal handling. Therefore, hot spots can still exist and can cause trouble. Another important factor is gain switching and DC level adjustment, which must not affect the thermal balance, either because of amplifier working point changes or because of the control circuitry.

Circuits which rely on feedback for error correction can be inherently less sensitive to thermal effects (except for the input differential transistor pair!). However, the feedback stability, or, more precisely, the no feedback stability during overdrive, can cause identical or even worse problems. If there is insufficient damping during the transition from saturation back to the linear range, long ringing can result, impairing the recovery time. Such problems are usually solved by adding normally reverse biased diodes, which conduct during the saturation of a nearby transistor, allowing the remaining part of the circuit some control over critical nodes.

We will review a few possible solutions in the following section.

5.4 Improving Amplifier Linearity

The discussion of modern wideband amplifiers would not be complete without considering their linearity. In older wideband instrumentation a non-linearity of 1% was considered adequately low. In oscilloscopes this figure was comparable to the width of the luminous trace on the CRT screen.

In modern digitizing oscilloscopes the vertical resolution is limited by the resolution of the analog to digital converter (ADC), which for high sampling rate systems is rarely better than 8 bit ($1{:}2^8$, or ~0.4 %). At lower sampling rates the resolution can be 10 bit ($1{:}2^{10}$, or ~0.1 %) or 12 bit ($1{:}2^{12}$, or ~0.025 %). However, in such cases the digital display's resolution limits the readout accuracy (some new digital oscilloscopes have LCD screens with 1024 horizontal by 768 vertical pixels).

Nevertheless, the linearity issue is still important, because digital systems are often used when additional signal processing is required, either by the digital signal processor (DSP) within the 'scope itself or by an external computer to which the sampled signal is transferred. This processing can range from simple two-channel signal subtraction, addition, multiplication, etc., to averaging, convolution, or spectral analysis by fast Fourier transform (FFT). Note that spectral analysis, even if performed on 8 bit data, can offer an apparent 80 dB range (0.01 % resolution) if the signal in memory is long (1 MB signal acquisition length is not uncommon!) and if the FFT is performed on a large number of waveforms. It is therefore important to preserve all the linearity which the amplifier and the digitizer can offer.

At low frequencies the simplest way of improving linearity is by applying some sort of local corrective (negative) feedback at each amplifying stage, as we have seen in Sec. 5.2. But at high frequencies the feedback can give more trouble than it can solve, owing to multiple high frequency poles and the total phase and time delay within the loop. Pole–zero compensation and phase correction can be used to a certain extent, but ultimately the amplifier's time delay sets the limit. With feedback the error can be reduced only and never eliminated, since the error reduction is proportional to the loop gain, which can not be infinite, and it also falls with frequency. So the error is small at low frequencies, increasing to its full value at the closed loop cut off.

At the highest frequencies the only error correction technique which can be made to work is the so called error 'feedforward'. This technique involves taking the driving signal error from an earlier circuit stage and then subtracting it from the output signal, so that the error is effectively canceled.

For the younger generation of engineers it is perhaps interesting to mention the historical perspective of feedback and feedforward error correction. Feedback is an omnipresent concept today, but it was not always so! In fact, both the feedback and the feedforward concepts were invented by *Harold S. Black* [Ref. 5.39 and 5.40]. However, feedforward was invented in 1923 and feedback in 1927, some 4 years later. While the patent for feedforward was granted in 1928, the feedback was such a "strange and counterintuitive" concept that the patent was granted almost 10 years after the application, in 1937! In spite of the fact that the feedback concept has been invented later and has been slow to catch up, once it happened it soon became the preferred method of error reduction, mostly owing to the work of *H. Nyquist* (1889–1976) and *H.W. Bode* (1905–1982), [Ref. 5.67]. On the other hand, feedforward was

almost forgotten, then 'reinvented' from time to time, only to be rediscovered by the broader engineering community in 1975, when the Quad 405 audio power amplifier came onto the market [Ref. 5.42 and 5.43]. Following the presentation article by the 405's designer, *Peter J. Walker*, the idea was refined and generalized by a number of authors, among the first by *J. Vanderkooy* and *S.P. Lipshitz* [Ref. 5.44].

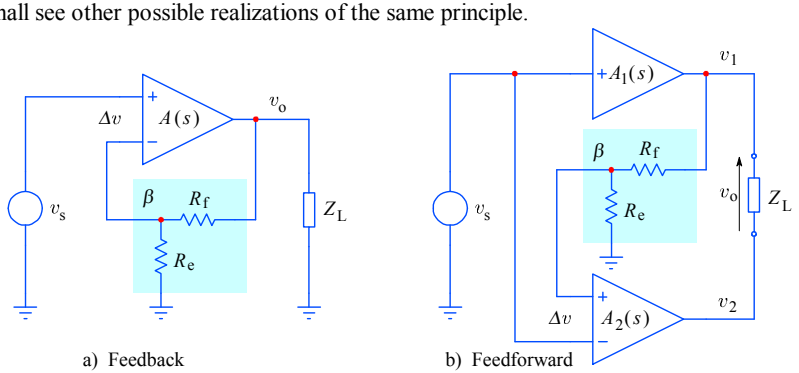
Later *M.J. Hawksford* [Ref. 5.46] showed that between the two extremes (pure feedback on one end and pure feedforward on the other) there is a whole spectrum of solutions combining both concepts. Moreover, such solutions can be applied both at system level (like the Quad 405 itself or similar solutions, as in [Ref. 5.48]) or down to the transistor level (as in [Ref. 5.47]).

It is interesting that while there were several examples of feedforward application in the field of RF communications (some even before 1975), most of the theoretical work was done with audio power amplification in mind. Apparently it took some time before the designers of high speed circuits grasped the full potential and the inherent advantages of the feedforward error correction techniques. In a way, this situation has not changed much, for even today we meet feedforward error correction mostly in RF communications systems and top level instrumentation. At the IC level we find feedforward only as a method of phase compensation (bypassing a slow inter-stage, not error correction), mostly in older opamps. From 1985 the advance in semiconductor processing has been extremely fast, discouraging amplifier designers from seeking more 'exotic' circuit topologies.

Before we examine the benefits of the feedforward technique for wideband amplification we shall first compare the feedback and feedforward principles from the point of view of error correction.

5.4.1 Feedback and Feedforward Error Correction

Fig. 5.4.1 shows the comparison of amplifiers with feedback and feedforward error correction. The feedforward amplifier is shown in its simplest form — later we shall see other possible realizations of the same principle.



a) Feedback b) Feedforward

Fig. 5.4.1: Comparison of amplifiers with feedback and feedforward error correction. The feedback amplifier has excess gain, $A(s)$, which is reduced to the required level by taking the output voltage, suitably attenuated (β), to the inverting input of the differential amplifier (hence the name 'negative feedback'). The feedforward case, in its most simple form, has two amplifiers: the main amplifier $A_1(s)$ is assisted by the auxiliary amplifier $A_2(s)$, which takes the difference between the input voltage and the attenuated main amplifier output forward to the load.

The analysis of the **feedback amplifier** has already been presented in Sec. 5.3, but we shall repeat some expressions in order to correlate them with the feedforward amplifier. From Fig. 5.4.1a we can write:

$$v_o = (v_s - \beta\, v_o) \cdot A(s) \tag{5.4.1}$$

where $\beta = R_e/(R_f + R_e)$. By solving for v_o we have:

$$v_o = v_s \frac{A(s)}{1 + \beta A(s)} \tag{5.4.2}$$

The fraction on the right hand side is the amplifier closed loop gain G_{fb}; it can be rewritten in such a form, from which it is easily seen that the gain expression can be approximated as $1/\beta$ if $A(s)$ is large:

$$G_{fb} = \frac{A(s)}{1 + \beta A(s)} = \frac{1}{\frac{1}{A(s)} + \beta} \approx \frac{1}{\beta}\bigg|_{A(s)\to\infty} = 1 + \frac{R_f}{R_e} \tag{5.4.3}$$

Of course, this final simplification is valid at low frequencies only. Since $A(s)$ is finite and falls with frequency owing to the amplifier dominant pole s_0:

$$A(s) = A_0 \frac{-s_0}{s - s_0} \tag{5.4.4}$$

the closed loop transfer function must also have a pole, but at a frequency s_h at which $A(s_h) = 1/\beta$. Since $f_h = |s_h|/2\pi$ and $f_0 = |s_0|/2\pi$ we can write:

$$\frac{1}{\beta} = A_0 \frac{f_0}{f_h + f_0} \tag{5.4.5}$$

and, considering that $\beta\, A_0 \gg 1$, the closed loop cut off frequency is:

$$f_h = f_0(\beta\, A_0 - 1) \approx f_0\, \beta\, A_0 = f_0 \frac{A_0}{G_{fb}} \tag{5.4.6}$$

So the closed loop cut off frequency of a voltage feedback amplifier is inversely proportional to the closed loop gain.

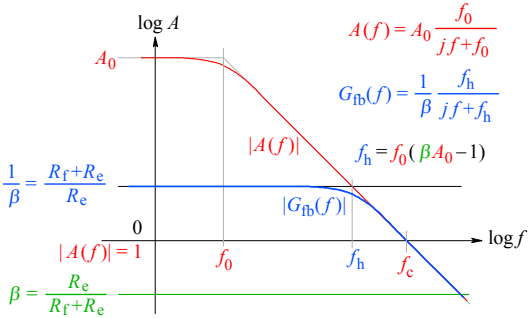


Fig. 5.4.2: For a voltage feedback amplifier the closed loop frequency response $G_{fb}(f)$ depends on the open loop gain A_0, its dominant pole f_0 and the feedback attenuation factor β. The transition frequency f_c is equal to the amplifier gain bandwidth product (but only if the amplifier does not have a secondary pole close to or lower than f_c).

For the **feedforward amplifier**, Fig. 5.4.1b, we must first realize that the load voltage, v_o, is the difference of the output voltages of individual amplifiers:

$$v_o = v_1 - v_2 \tag{5.4.7}$$

Since the main amplifier output is:

$$v_1 = v_s\, A_1(s) \tag{5.4.8}$$

and the auxiliary amplifier output is:

$$v_2 = (\beta\, v_1 - v_s) \cdot A_2(s) \tag{5.4.9}$$

it follows that the system output is:

$$\begin{aligned} v_o &= v_s\, A_1(s) - (\beta\, v_1 - v_s) \cdot A_2(s) \\ &= v_s\, [A_1(s) - \beta\, A_1(s)\, A_2(s) + A_2(s)] \end{aligned} \tag{5.4.10}$$

So by temporarily neglecting the frequency dependence, the closed loop gain of the feedforward amplifier is:

$$G_{ff} = A_1 + A_2 - \beta\, A_1\, A_2 \tag{5.4.11}$$

The following reasoning is the most important point in feedforward amplifier analysis and it is probably difficult to foresee, but once we do, it becomes all too obvious. Let us say that, within the frequency range of interest, we can achieve:

$$\beta\, A_1 = 1 \tag{5.4.12}$$

Then:

$$G_{ff} = \frac{1}{\beta} + A_2 - \beta\, \frac{1}{\beta}\, A_2 = \frac{1}{\beta} \tag{5.4.13}$$

So, whatever the actual value of the auxiliary amplifier gain A_2, the system's gain G_{ff} will be equal to $1/\beta$ if we can make $A_1 = 1/\beta$. Note that we have not requested any of the two gains to be very high, as we were forced to for feedback amplifiers, therefore this result is achieved without any approximation! True, if A_1 is frequency dependent and β is not, at high frequencies Eq. 5.4.12 would not hold and, consequently, Eq. 5.4.13 would not be so simple.

However, when A_1 starts to fall with frequency the factor $-\beta\, A_1 A_2$ is also reduced by the same amount, and the appropriate part of A_2 compensates the loss. This would be so as long as the gain A_2 remains constant with frequency (and even beyond its own cut off, provided that there still is enough gain for correction!).

Thus feedforward (in principle) achieves the dream goal:
a zero error, high cut off frequency gain, using non-ideal amplifiers!

We shall, of course, still have to deal with component tolerances, temperature dependencies, uncontrollable strays, time delays, etc., but with a manageable effort all these factors can be minimized, or, at least, kept below some predictable limit.

If you now think that there are no more surprises with feedforward amplifiers, consider the following points:

Eq. 5.4.11 is, in a sense, symmetrical, thus $G_{\text{ff}} = 1/\beta$ (as in Eq. 5.4.13) can be achieved also if we decide to make:

$$\beta A_2 = 1 \tag{5.4.14}$$

However, the advantage of making $\beta A_1 = 1$ is that the input signal is canceled at the auxiliary amplifier differential input (remaining as a common mode signal only). In contrast with the 'output balance' condition represented by Eq. 5.4.14, Eq. 5.4.12 represents the so called 'input balance' condition, in which the auxiliary amplifier needs only a very low level amplitude swing at low frequencies (but rising to $1/2$ amplitude and higher at A_1 cut off and beyond, respectively). It therefore processes only the errors of the main amplifier, canceling them at the load and leaving only those of the auxiliary amplifier; as errors in processing the main amplifier error, those are secondary errors only.

It is also possible to make both gains equal to $1/\beta$:

$$A_1 = A_2 = 1/\beta \tag{5.4.15}$$

achieving in this way both input and output node balance.

In some instances there are good reasons to make A_1 larger than $1/\beta$, achieving also gain correction (in this case, the auxiliary amplifier will have to partly handle the input signal, too). Of course, in all these cases the gain matching has to be very precise, since in feedforward amplifiers the error cancellation is additive, as can be deduced from Eq. 5.4.11, not multiplicative as in feedback amplifiers (Eq. 5.4.2).

Another nice property of the feedforward circuit is that it is feedback-free, so there are no loops causing potential instability, and consequently no stability criterion to satisfy. As a disadvantage, the output impedance is not lowered by the feedforward action (as it is in feedback amplifiers), so for high amplifier power efficiency it has to be made as low as possible within the frequency range of interest.

Let us return to the frequency dependence of the feedforward system gain. By defining:

$$A_1(s) = A_{01}\,\frac{-s_1}{s-s_1} \qquad \text{and} \qquad A_2(s) = A_{02}\,\frac{-s_2}{s-s_2} \tag{5.4.16}$$

the system gain is:

$$G_{\text{ff}} = A_{01}\,\frac{-s_1}{s-s_1} + A_{02}\,\frac{-s_2}{s-s_2} - \beta\,A_{02}A_{01}\,\frac{s_1 s_2}{(s-s_1)(s-s_2)} \tag{5.4.17}$$

where A_{01} and A_{02} are the main and auxiliary amplifier DC gain, respectively. By choosing $A_{01} = 1/\beta$ we get:

$$\begin{aligned} G_{\text{ff}} &= \frac{1}{\beta}\cdot\frac{-s_1}{s-s_1} + A_{02}\,\frac{-s_2}{s-s_2}\left(1 - \frac{-s_1}{s-s_1}\right) \\ &= \frac{1}{\beta}\cdot\frac{-s_1}{s-s_1} + A_{02}\,\frac{-s_2}{s-s_2}\cdot\frac{s}{s-s_1} \end{aligned} \tag{5.4.18}$$

Note that the auxiliary amplifier gain is effectively multiplied by the high pass version of the main amplifier's frequency dependence. If we now decide to make $A_{02} = 1/\beta$ also, we obtain:

$$G_{\text{ff}} = \frac{1}{\beta}\left(\frac{-s_1}{s-s_1} + \frac{-s_2}{s-s_2}\cdot\frac{s}{s-s_1}\right)$$

$$= \frac{1}{\beta}\left[\frac{-s_1}{s-s_1}\left(1+\frac{s_2}{s_1}\cdot\frac{s}{s-s_2}\right)\right] \qquad (5.4.19)$$

The question is: is it desirable to make $s_1 = s_2$? Let us see:

$$G_{\text{ff}} = \frac{1}{\beta}\left(\frac{-s_1}{s-s_1} + \frac{-s_1 s}{(s-s_1)^2}\right) \qquad (5.4.20)$$

The second fraction represents a second-order band pass response, which will add some gain peaking and extend the bandwidth by a factor of almost 3:

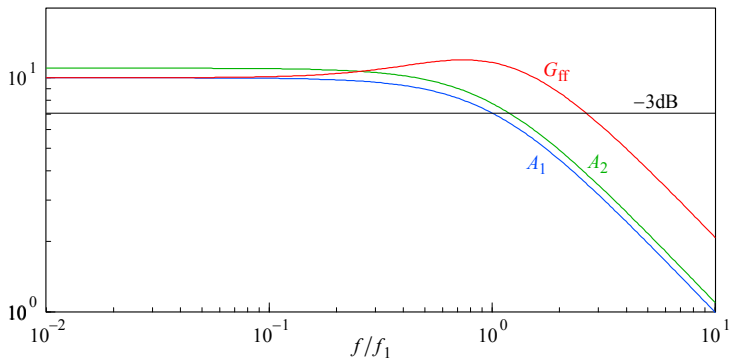


Fig. 5.4.3: The feedforward amplifier bandwidth G_{ff} is highest (and optimized in the sense of lowest gain × bandwidth requirement of the auxiliary amplifier) if both amplifiers have the same bandwidth and the gains equal to $1/\beta$. In this figure $1/\beta = 10$, $A_{01} = 10$, and $A_{02} = 11$ (in order to distinguish A_2 from A_1 more easily). If $s_1 \neq s_2$ the G_{ff} bandwidth will be lower.

5.4.2 Error Reduction Analysis

In order to analyze the error reduction by both feedback and feedforward action we have to determine the system gain **sensitivity** on the amplifier gain variations. Generally, the concept of sensitivity of some system property, let us call it P, to variations of some subsystem parameter x is expressed as the x fraction of P, multiplied by the partial derivative of P on x:

$$S_x^P = \frac{x}{P}\cdot\frac{\partial P}{\partial x} \qquad (5.4.21)$$

and it represents the amount of change in P for a unit change in x.

For the feedback amplifier we want to know the influence of variations in the amplifier open loop gain A to the closed loop gain G_{fb}, as defined by Eq. 5.4.3:

$$S_A^{G_{\text{fb}}} = \frac{A}{G_{\text{fb}}} \cdot \frac{\partial G_{\text{fb}}}{\partial A} = \frac{A}{\dfrac{A}{1+\beta A}} \cdot \frac{\partial \dfrac{A}{1+\beta A}}{\partial A}$$

$$= (1+\beta A) \cdot \left[\frac{1}{1+\beta A} - \frac{\beta A}{(1+\beta A)^2} \right] = 1 - \frac{\beta A}{1+\beta A}$$

$$= \frac{1}{1+\beta A} \longrightarrow 0 \Big|_{A \to \infty} \tag{5.4.22}$$

This means that the gain sensitivity is low only if A is very high. We also want to know the influence of variations of the feedback attenuation β:

$$S_\beta^{G_{\text{fb}}} = \frac{\beta}{G_{\text{fb}}} \cdot \frac{\partial G_{\text{fb}}}{\partial \beta} = \frac{\beta}{\dfrac{A}{1+\beta A}} \cdot \frac{\partial \dfrac{A}{1+\beta A}}{\partial \beta}$$

$$= \frac{\beta(1+\beta A)}{A} \cdot \left[- \frac{A^2}{(1+\beta A)^2} \right]$$

$$= - \frac{\beta A}{1+\beta A} \longrightarrow -1 \Big|_{A \to \infty} \tag{5.4.23}$$

In the case of the feedforward amplifier, the influence of A_1 and A_2 on the system gain, as well as the influence of β, using Eq. 5.4.11 for G_{ff}, is:

$$S_{A_1}^{G_{\text{ff}}} = \frac{A_1}{G_{\text{ff}}} \cdot \frac{\partial G_{\text{ff}}}{\partial A_1} = \frac{A_1(1-\beta A_2)}{A_1(1-\beta A_2)+A_2} = \begin{cases} 1-\beta A_2 & \text{if } \beta A_1 = 1 \\ 0 & \text{if } \beta A_2 = 1 \end{cases} \tag{5.4.24}$$

$$S_{A_2}^{G_{\text{ff}}} = \frac{A_2}{G_{\text{ff}}} \cdot \frac{\partial G_{\text{ff}}}{\partial A_2} = \frac{A_2(1-\beta A_1)}{A_2(1-\beta A_1)+A_1} = \begin{cases} 0 & \text{if } \beta A_1 = 1 \\ 1-\beta A_1 & \text{if } \beta A_2 = 1 \end{cases} \tag{5.4.25}$$

$$S_\beta^{G_{\text{ff}}} = \frac{\beta}{G_{\text{ff}}} \cdot \frac{\partial G_{\text{ff}}}{\partial \beta} = \frac{-\beta A_1 A_2}{A_1 + A_2 - \beta A_1 A_2} = \begin{cases} -\beta A_2 & \text{if } \beta A_1 = 1 \\ -\beta A_1 & \text{if } \beta A_2 = 1 \\ -1 & \text{if } A_1 = A_2 = 1/\beta \end{cases} \tag{5.4.26}$$

It is evident that the second condition in Eq. 5.4.24 and the first in Eq. 5.4.25, as well as the third condition in Eq. 5.4.26, are the same as for the feedback amplifier. However, remember that for the feedback amplifier the results belong to the ideal case for which $A \to \infty$, so in practice they can be approximated only, whilst for the

feedforward amplifier they can be realized without any approximation (but within the limits of the specified component tolerances).

In a similar way we can determine the error reduction. For the feedback amplifier we have:

$$\frac{\varepsilon_{\text{fb}}}{\varepsilon_A} = \frac{1}{1+\beta A} \longrightarrow 0\Big|_{A\to\infty} \tag{5.4.27}$$

and again, zero distortion is achievable only in the idealized case of infinite gain.

In contrast, for the feedforward amplifier we have:

$$\frac{\varepsilon_{\text{ff}}}{\varepsilon_{A_1}} = 1-\beta A_2 = 0\Big|_{\beta A_2=1} \tag{5.4.28}$$

and this extraordinary result can be realized (not only approximated!) to whatever degree of precision we are satisfied with (accounting also for the technology cost).

Also, we must not forget that the open loop gain of the feedback amplifier decreases with frequency, so, for a given A_0, the theoretically achievable maximum error reduction, $1/(1+\beta A_0)$, is obtained only from DC up to the frequency of the dominant pole, f_1; beyond f_1 the error increases proportionally with frequency.

In contrast, feedforward amplifiers offer the same level of error reduction from DC up to the full feedforward system bandwidth and even beyond!

5.4.3 Alternative Feedforward Configurations

The main drawback of the feedforward amplifier in Fig. 5.4.1b is the 'floating' load (between the outputs of both amplifiers); in most cases we would prefer a ground referenced load, instead.

We have already noted that the output impedance of the main amplifier is not reduced by feedforward action; in fact, it does not need to be low in order to achieve effective error canceling. This leads to the idea of summing passively the two outputs, with that of the auxiliary amplifier inverted in phase:

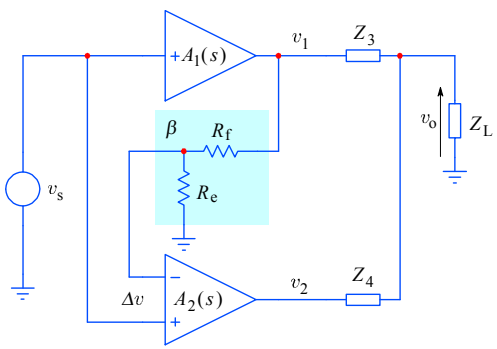


Fig. 5.4.4: Grounded load feedforward amplifier. Note the inverted input polarity of the auxiliary amplifier, compared with the circuit in Fig. 5.4.1b. This allows passive signal summing over the output impedances Z_3 and Z_4.

The impedances Z_3 and Z_4 should be lower than Z_L, but this condition is dictated mainly by amplifier power efficiency, not by the summing process. We have labeled Z_3 and Z_4 in accordance with tradition (emerging from the Quad 405 circuit, see Fig. 5.4.6) and also in accordance with a general form, in which Z_1 and Z_2 appear as feedback components to A_1 and A_2. As implied by the Z symbol, these impedances can also be complex.

The output voltage v_o can be calculated by summing the two output currents:

$$i_1 = \frac{v_1 - v_o}{Z_3} \qquad \text{and} \qquad i_2 = \frac{v_2 - v_o}{Z_4} \tag{5.4.29}$$

so that:

$$v_o = Z_L(i_1 + i_2) \tag{5.4.30}$$

which results in:

$$v_o = \frac{Z_L}{Z_L + \dfrac{Z_3 Z_4}{Z_3 + Z_4}} \left(\frac{Z_3}{Z_3 + Z_4} v_1 + \frac{Z_4}{Z_3 + Z_4} v_2 \right) \tag{5.4.31}$$

By extracting the common attenuation factor, a:

$$a = \frac{Z_L}{Z_L + \dfrac{Z_3 Z_4}{Z_3 + Z_4}} \cdot \frac{Z_3}{Z_3 + Z_4} \tag{5.4.32}$$

the system's gain is:

$$G_{ff} = \frac{v_o}{v_s} = a \left[A_1 + \frac{Z_4}{Z_3}(A_2 - \beta A_1 A_2) \right] = \frac{a}{\beta} \Bigg|_{\substack{\text{when } \beta A_1 = 1 \\ \text{or } \beta A_2 = Z_3/Z_4}} \tag{5.4.33}$$

Because of the passive summing, the correct balance condition and error canceling for this circuit is achieved when the two output voltages are in the inverse ratio as the impedances:

$$\frac{v_2}{v_1} = \frac{Z_3}{Z_4} \tag{5.4.34}$$

If we assume that the output balance has been achieved, then:

$$v_2 = v_1 \beta A_2 \tag{5.4.35}$$

But the output balance condition is $\beta A_2 = 1$, therefore:

$$\frac{v_1 \beta A_2}{v_1} = \frac{Z_3}{Z_4} \qquad \Rightarrow \qquad \beta A_2 = \frac{Z_3}{Z_4} \tag{5.4.36}$$

On the other hand, the input balance condition, $\beta A_1 = 1$, because of Eq. 5.4.34 and 5.4.36, results in the gain ratio equal to the impedance ratio:

$$\frac{A_2}{A_1} = \frac{Z_3}{Z_4} \tag{5.4.37}$$

The auxiliary amplifier will, under this condition, draw considerable current, even without the load. To reduce the current demand, we have to give up the input balance condition. If we set $v_2 = v_1$ then there would be no current if $Z_L = \infty$, and by choosing $Z_3||Z_4 \ll Z_L$ the current demand is reduced for the nominal load:

$$v_2 = A_2\left(\frac{v_1}{A_1} - \beta v_1\right) = v_1 \tag{5.4.38}$$

and this means that:

$$\frac{A_2}{A_1} = \beta A_2 + 1 \tag{5.4.39}$$

which, considering Eq. 5.4.35, results in:

$$\frac{A_2}{A_1} = \frac{Z_3}{Z_4} + 1 \tag{5.4.40}$$

and this should be compared with the 'simple' balance condition in Eq. 5.4.37.

Another configuration, known as the 'error take off', by *Sandman* [Ref. 5.49], is shown in Fig. 5.4.5. Here both the main and the auxiliary amplifier are of the negative feedback type; however, the auxiliary amplifier senses both the distortion and the gain error from the main amplifier feedback input and delivers it to the load in the same feedforward passive summing manner.

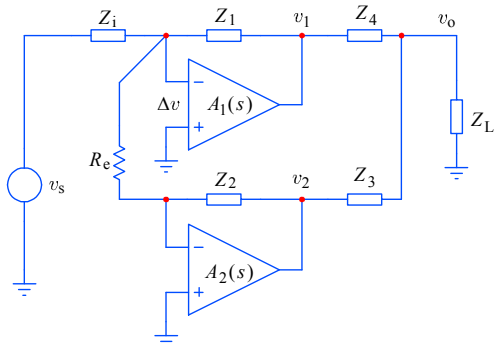


Fig. 5.4.5: 'Error take off' principle. The error Δv of the main amplifier, left by feedback, is taken by the auxiliary amplifier and fed forward to the load, where it is passively summed to the main output. The impedances Z_1 to Z_4 form the balancing bridge.

With an ideal main amplifier the voltage at its inverting input would be at a (virtual) ground potential; any signal Δv present at this point represents an attenuated version of the main amplifier error (gain error and distortion). If adequately amplified, it can be added to the main output to cancel the error at the load:

$$\Delta v = \beta\varepsilon = \frac{Z_i}{Z_i + Z_1}\varepsilon \tag{5.4.41}$$

To achieve effective error cancellation, we must set:

$$\frac{Z_2}{R_e} \cdot \frac{Z_i}{Z_i + Z_1} = \frac{Z_3}{Z_4} \tag{5.4.42}$$

A variation of this circuit is shown in Fig. 5.4.6, which actually represents the original Quad 405 'current dumping' amplifier configuration, and we now see how it follows from both the 'error take off' circuit and the pure feedforward circuit. If the balance condition is achieved A_2 must compensate whatever the amount of error at the A_1 output. It is important to realize that the input signal of A_1 can be taken from any suitable point within the circuit (the only condition is that it should, preferably, not be out of phase); the A_2 output represents just such a convenient point. The balance condition for the 405 is:

$$\frac{Z_3}{Z_4} = \frac{Z_2}{Z_1} \tag{5.4.43}$$

and the main amplifier error is canceled. Again, the impedances Z_n can be real or complex, whichever combination satisfies this equation.

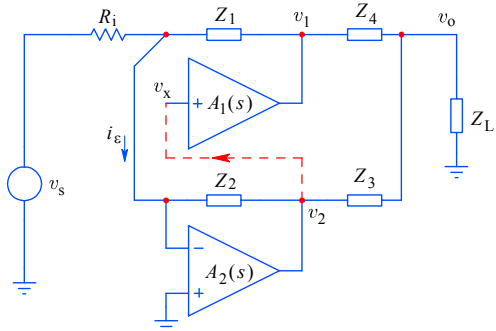


Fig. 5.4.6: Current dumping: by taking the main amplifier input signal (v_x) from the auxiliary amplifier (v_2), we obtain the original Quad 405 amplifier configuration. Although it is effective in error cancellation, its main disadvantages are the requirement for large voltage at the auxiliary amplifier output and a relatively low cut off frequency. In the Quad 405, Z_1 and Z_3 are resistors, Z_2 is a capacitor and Z_4 is an inductor.

A disadvantage of the current dumping scheme is that the auxiliary amplifier, although supplying relatively low current, must supply all the output voltage plus the error term; in such a condition the auxiliary amplifier's error can be rather high, and although it is a second-order error it can be significant. Also, in the classical 405 realization Z_2 (the feedback impedance of the auxiliary amplifier) is a capacitance (compensating the inductance Z_4), which results in a relatively slow system (high speed is not an issue at audio frequencies).

5.4.4 Time Delay Compensation

Let us return to the basic feedforward amplifier of Fig. 5.4.1b. It is clear that if both the main and the auxiliary amplifier have limited bandwidth, time delays are inevitable. In order to work properly the feedforward scheme must include some time delay compensation, otherwise error cancellation close to and beyond the system cut off frequency would not occur.

For two sinusoidal signals a small time delay between them is transformed into a small amplitude error when summed and it might even be possible to compensate it

by altering the balance condition. However, for a square wave or a step function, large spikes, equal to full amplitude difference, would result and these can not be corrected by the balance setting components. Moreover, these spikes can overdrive the input of the auxiliary amplifier and saturate its output; error correction in such conditions is non-operating. Thus for high speed amplification some form of time delay compensation is mandatory.

In Fig. 5.4.7 we see a general principle of time delay compensation. Since each amplifier has its own time delay acting on a different summing node, at least two separate time delay circuits are needed. Here, τ_1 compensates the delay of the main amplifier, allowing the auxiliary amplifier input to perform the difference between the input and β–attenuated signal with the correct phase. Likewise, τ_2 does so for the auxiliary amplifier delay for correct summing at the output.

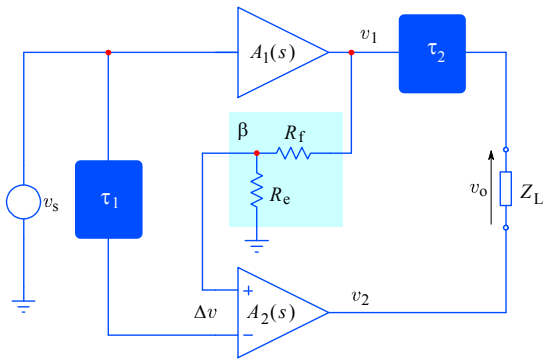


Fig. 5.4.7: Time delay compensation principle for the basic feedforward amplifier: τ_1 compensates the delay introduced by the main amplifier and τ_2 compensates the delay introduced by the auxiliary amplifier.

5.4.5 Circuits With Local Error Correction

In this section we shall briefly discuss a few simple circuits in which we shall employ our knowledge of feedback and feedforward error correction. We shall demonstrate how the same technique, which was developed at the system level, can also be applied at the local (subsystem) level. Local error correction often gives better results, since the linearization of individual amplifier stages lowers the requirements or indeed completely eliminates the need for system level correction. In many applications, such as oscilloscopes and adaptable data acquisition instrumentation, system level correction can be difficult to implement, owing to variable inter-stage conditions (variable attenuation, gain, DC level setting, trigger pick up delays, etc.).

The most interesting circuits to which the error correction is applied are the differential amplifier and the differential cascode amplifier. We have discussed them briefly in Part 3, Sec. 3.7. Here we shall review the analysis with the emphasis on their non-linearity. The dominant non-linearity mechanism is the familiar exponential dependence of I_e to V_{be} of a single transistor:

$$I_e = I_s\left(e^{\frac{q_e}{k_B T}V_{be}} - 1\right) \qquad (5.4.44)$$

Here I_s is the saturation current (about 10^{-14} A in silicon transistors, depending on the dopant concentrations in the p–n junctions); the remaining symbols have the usual meaning (see Part 3, Sec. 3.1). Under normal operating conditions the DC emitter bias current I_{e0} exceeds I_s by at least 10^{11}, simplifying Eq. 5.4.44 to:

$$I_e \approx I_s \, e^{\frac{q_e}{k_B T} V_{be}} \Big|_{I_{e0} \gg I_s} \tag{5.4.45}$$

For small signals, not altering the junction temperature considerably, we can assume $V_T = k_B T / q_e$ to be constant. So if we also neglect the dependence of the current gain and I_s on temperature and biasing, as well as the internal resistance and capacitance variations with the signal, we can express the non-linearity in form of the internal emitter resistance:

$$r_e = \frac{\partial V_{be}}{\partial I_e} \tag{5.4.46}$$

For the differential pair of Fig. 5.4.8 the effective resistance seen by their base–emitter junctions is the sum:

$$r_{ed} = \frac{\partial V_{be1}}{\partial I_{e1}} + \frac{\partial V_{be2}}{\partial I_{e2}} \tag{5.4.47}$$

which, since one increases and the other decreases with the signal, varies much less over the much larger input signal range than in the single transistor case.

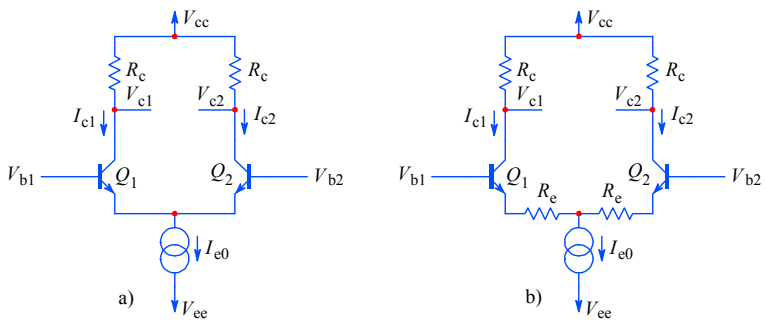


Fig. 5.4.8: **a)** A simple differential amplifier, showing the voltages and currents used in the analysis. **b)** The same, but with the emitter degeneration resistors R_e.

For the amplifier in Fig. 5.4.8a we must first realize that the differential input voltage is equal to the difference of the two V_{be} junction voltages:

$$V_{di} = V_{b1} - V_{b2} = V_{be1} - V_{be2} \tag{5.4.48}$$

We calculate the V_{be} junction voltages from Eq. 5.4.45:

$$V_{be} = V_T \ln \frac{I_e}{I_s} \tag{5.4.49}$$

In an integrated circuit, $I_{s1} \approx I_{s2}$, so the ratio of emitter currents is:

$$\frac{I_{e1}}{I_{e2}} = \frac{e^{V_{be1}/V_T}}{e^{V_{be2}/V_T}} = e^{\frac{V_{be1}-V_{be2}}{V_T}} = e^{\frac{V_{di}}{V_T}} \tag{5.4.50}$$

The collector current $I_c = \alpha_F I_e$; also, the sum of emitter currents must be equal to the bias provided by the constant current source, I_{e0}. Thus:

$$I_{c1} + I_{c2} = \alpha_F (I_{e1} + I_{e2}) = \alpha_F I_{e0} \tag{5.4.51}$$

From the last two equations we obtain:

$$I_{c1} = \frac{\alpha_F I_{e0}}{1 + e^{-V_{di}/V_T}} \qquad \text{and} \qquad I_{c2} = \frac{\alpha_F I_{e0}}{1 + e^{+V_{di}/V_T}} \tag{5.4.52}$$

The collector voltage is equal to the potential drop $R_c I_c$ from the supply voltage:

$$V_{o1} = V_{cc} - R_c I_{c1} \qquad \text{and} \qquad V_{o2} = V_{cc} - R_c I_{c2} \tag{5.4.53}$$

Therefore the differential output voltage will follow a hyperbolic tangent function of the input differential voltage:

$$\begin{aligned} V_{do} &= V_{o1} - V_{o2} = R_c (I_{c2} - I_{c1}) \\ &= \alpha_F R_c I_{e0} \left(\frac{1}{1 + e^{+V_{di}/V_T}} - \frac{1}{1 + e^{-V_{di}/V_T}} \right) \\ &= \alpha_F R_c I_{e0} \tanh \frac{-V_{di}}{V_T} \end{aligned} \tag{5.4.54}$$

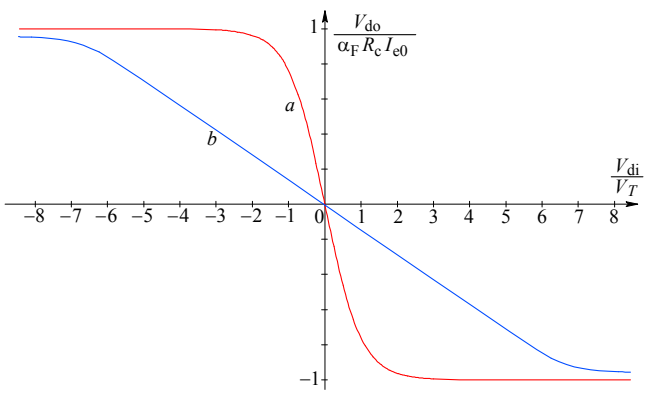


Fig. 5.4.9: a) The DC transfer function of the differential amplifier in Fig. 5.4.8a: the input differential voltage is normalized to V_T and the output is normalized to $\alpha_F R_c I_{e0}$. **b)** With the emitter degeneration, as in Fig. 5.4.8b, the transfer function is more linear, but at the expense of the gain (lower slope).

The system gain is represented by the slope of the plot, which, for $V_{di} \approx 0$, is:

$$\frac{V_{do}}{V_{di}} = \frac{\alpha_F R_c I_{e0}}{V_T} = \frac{\alpha_F R_c I_{e0}}{r_e I_{e0}} = \alpha_F \frac{R_c}{r_e} \tag{5.4.55}$$

Now, in wideband amplifier applications, the emitters are usually 'degenerated' by the addition of external resistors R_e. The degeneration resistor acts as a local current feedback, extending the linear part of the DC transfer function by $I_{e0}R_e + V_T$, instead of just V_T. Of course, this reduces the gain to $R_c/(R_e + r_e)$.

By considering one half of the differential pair and accounting for the bias current I_{e0} and signal current i flowing into R_e, the input voltage can be expressed from Eq. 5.4.44 as:

$$V_{in} = (i - I_{e0})R_e + V_T \ln\left(\frac{i + I_{e0}}{I_s}\right) \tag{5.4.56}$$

and by differentiating this we obtain:

$$\partial V_{in} = R_e\,\partial i + \frac{V_T}{i + I_{e0}}\,\partial i \tag{5.4.57}$$

We separate the linear and non-linear components:

$$\partial V_{in} = \underbrace{\left(R_e + \frac{V_T}{I_{e0}}\right)\partial i}_{\text{linear}} - \underbrace{\frac{V_T\, i}{I_{e0}(i + I_{e0})}\,\partial i}_{\text{non-linear}} \tag{5.4.58}$$

We define the loading factor x as the ratio of the signal current to the bias current:

$$x = \frac{i}{I_{e0}} \tag{5.4.59}$$

So we can express the incremental non-linearity (INL) factor N as the ratio of the non-linear gain component to the linear one:

$$N(x) = \frac{-x}{1 + x} \cdot \frac{V_T}{V_T + I_{e0}R_e} \tag{5.4.60}$$

Generally N can be (and usually is) a function of many variables, not just one.

In a similar way we can derive INL for the differential pair, where the differential input voltage is:

$$V_{in} = iR_e + (V_{be1} - V_{be2}) = iR_e + V_T \ln\left(\frac{I_{e0} + i}{I_{e0} - i}\right) \tag{5.4.61}$$

The linear and non-linear components are:

$$\partial V_{in} = \left(R_e + \frac{2V_T}{I_{e0}}\right)\partial i + \frac{2V_T\, i^2}{I_{e0}(i^2 - I^2)}\,\partial i \tag{5.4.62}$$

and the INL:

$$N(x) = \frac{x^2}{1 - x^2} \cdot \frac{2V_T}{2V_T + I_{e0}R_e} \tag{5.4.63}$$

This expression can be used to estimate the amount of error for a given signal and bias current, which an error correction scheme attempts to suppress.

In the following pages we are going to show a collection of differential amplifier circuits, employing some form of error correction, either feedback, feedforward or both. We are also going to present their frequency and time domain

performance to compare how the bandwidth has been affected as a result of increased circuit complexity (against a simple cascode amplifier).

For a fair comparison all circuits have been arranged to suit the test setup shown in Fig. 5.4.10; the amplifiers were set for a gain of 2, using the same type of transistors (BF959) and biased by a 10 mA current source. The input signal was modeled by a 10 mA step driven current source, loaded by two 50 Ω || 1 pF networks. An equal network was used as the output load. Finally, all circuits were 'tuned' for a Bessel system response (MFED), using only capacitive emitter peaking (of course, inductive peaking can be used in the final design). Note that this setup offers only a relative indication of what can be achieved, not a final optimized design.

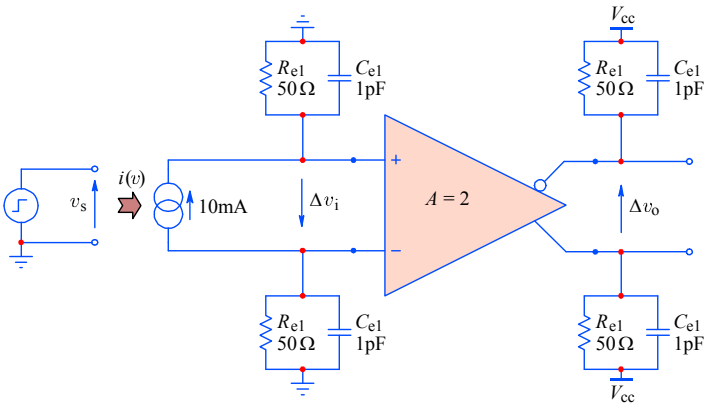


Fig. 5.4.10: Test set up used to compare different amplifier configurations.

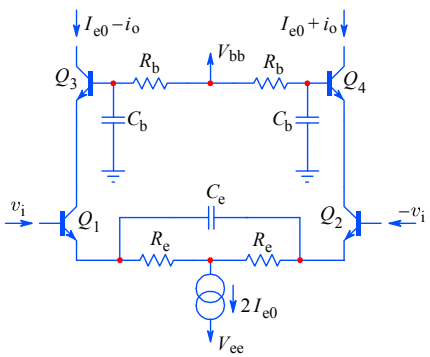


Fig. 5.4.11: This simple differential cascode amplifier, employing no error correction, is used in the test set up circuit of Fig. 5.4.10, representing the reference against which all other amplifiers are compared. The emitter peaking and base impedance are adjusted for a MFED response.

The simple differential cascode amplifier of Fig. 5.4.11, employing no error correction, represents a reference against which all other amplifiers will be compared. The $Q_{1,2}$ emitter peaking capacitor C_e and the $Q_{3,4}$ base network R_bC_b are adjusted

for a MFED response. Fig. 5.4.12 and 5.4.13 show the frequency domain and time domain responses, respectively.

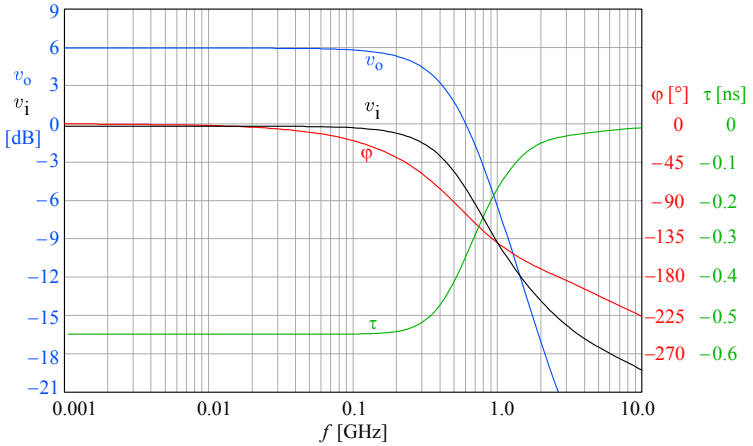


Fig. 5.4.12: Frequency domain performance of a simple differential cascode amplifier of Fig. 5.4.11 (no error correction) used in the test set up circuit of Fig. 5.4.10. This will be used as the reference for all other circuits. The bandwidth achieved is a little over 400 MHz.

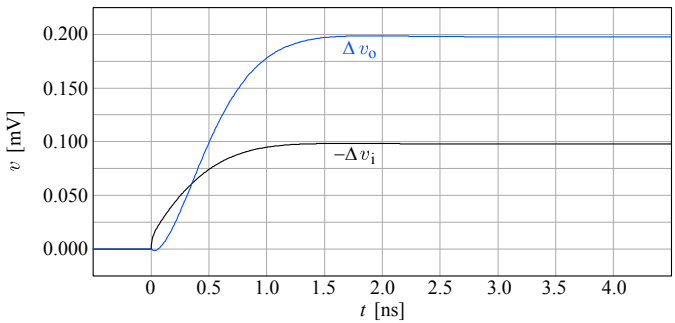


Fig. 5.4.13: Time domain performance of a simple differential cascode amplifier of Fig. 5.4.11 (no error correction) used in the test set up circuit of Fig. 5.4.10. This will be used as the reference for all other circuits. The input voltage indicates the input impedance dynamics in the first 100 ps and up to 1.5 ns. Note also the small undershoot at the output, owed to the cross-talk via C_{bc} of $Q_{1,2}$. The output rise time is less than 1 ns.

The first circuit to be compared with the reference is shown in Fig. 5.4.14. The circuit is owed to *C.R. Battjes* [Ref. 5.18] and is functionally a Darlington connection (Q_1, Q_2), improved by the addition of Q_3. Used as the differential input stage of a cascode amplifier, it enhances the input characteristics and increases both the output current handling and the bandwidth. At a first glance it may seem that the diode connected Q_3 (shorted collector and base) can not do much. However, it allows Q_1 to carry a current much larger than the Q_2 base current, delivering it to the resistance R_e and lowering the impedance seen by the base of Q_2, thus extending the bandwidth. The compound device has about twice the current gain of a single transistor.

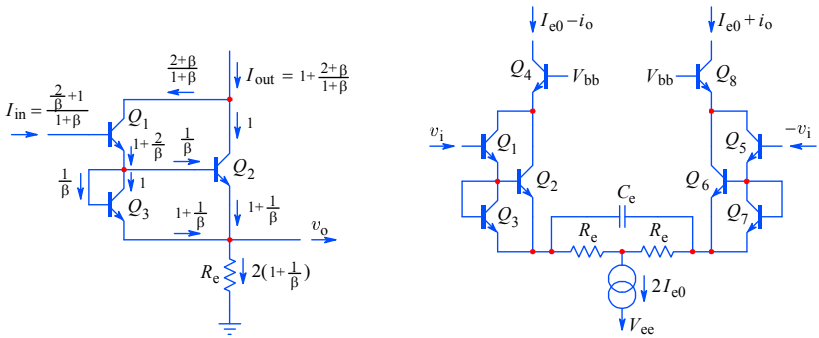


Fig. 5.4.14: a) Improved Darlington. b) Used as the input differential stage of the cascode amplifier — see the performance in the following figures.

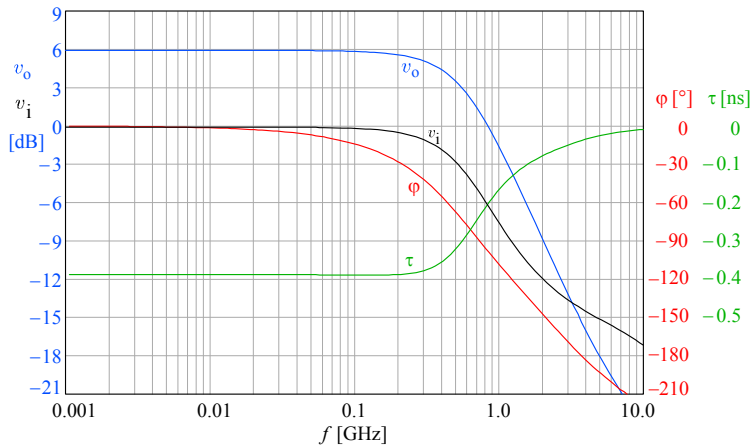


Fig. 5.4.15: Frequency domain performance of the differential cascode amplifier using the circuit of Fig. 5.4.14b. The bandwidth is about 560 MHz. Note the input voltage changing slope above 2 GHz.

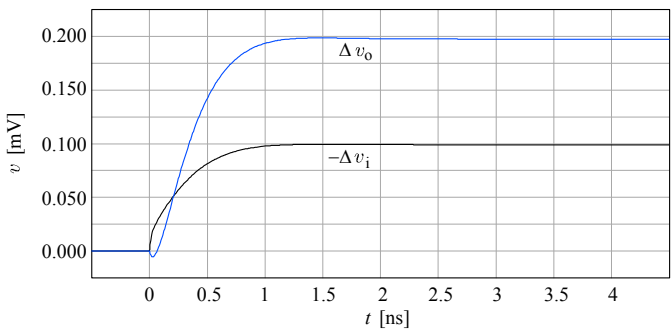


Fig. 5.4.16: Time domain performance of the differential cascode amplifier of Fig. 5.4.14b. The undershoot has increased, but the rise time is less than 0.7 ns.

In Fig. 5.4.17 Q_1 and Q_2 form the differential amplifier, whose error current i_1 is sensed at the resistor R_1 and is available at the collector of Q_3 for summing with the output current i_2 (error feedforward) further in the following circuit.

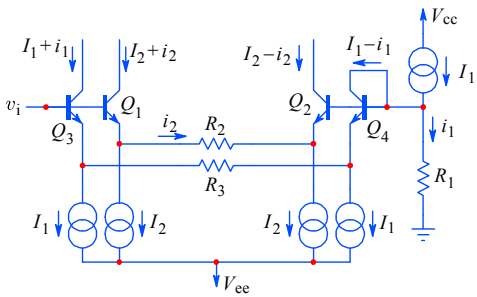


Fig. 5.4.17: A simple differential amplifier with feedforward error correction. Accurate matching of transistors is required only for DC error reduction, not for the feedforward linearization. Here i_2 is the differential current, whilst the error current, i_1, sensed at R_1, is available at the Q_3 collector to be added to the output current further in the circuit.

Two such circuits can form a differential amplifier, employing a double error feedforward correction, as shown in Fig. 5.4.18. The error currents can now be summed directly with output currents, without further processing.

However, the main problem with this linearization technique is that R_1 must be relatively high for a suitable error sensing, so it can reduce the bandwidth considerably. In part the bandwidth can be improved by adding precisely matched capacitors in parallel to both R_2 and R_3 (emitter peaking), but then the input impedance can become negative and should be compensated accordingly. This negative input impedance compensation is easily implemented at $\pm v_i$ inputs, but, by adding it to the inputs connected to R_1, the error sensing will be affected, since a part of i_1 would flow into the compensating networks, reducing error correction at high frequencies.

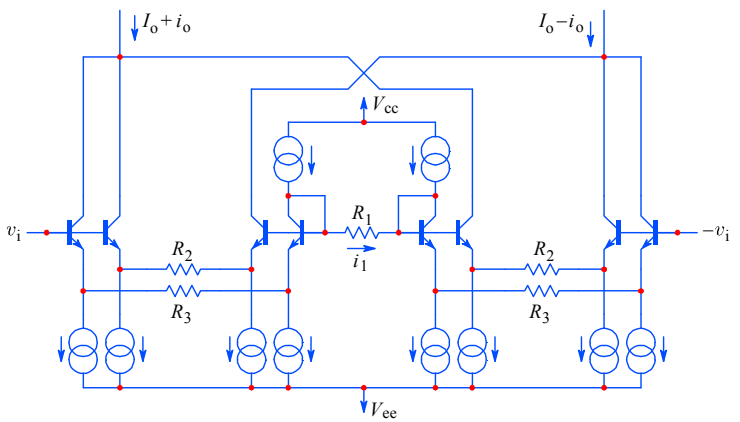


Fig. 5.4.18: Two circuits from Fig. 5.4.17 can form a differential amplifier with a double error feedforward directly summed with the output.

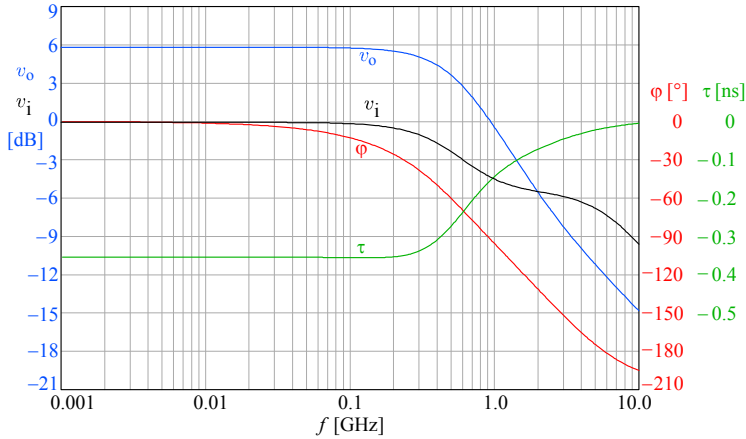


Fig. 5.4.19: Frequency domain performance of the circuit from Fig. 5.4.18. The bandwidth can be high (560 MHz), but for a suitable error sensing the required high value of R_1 would compromise it. The plot of v_i indicates the negative input impedance at high frequencies, which would need additional compensation networks at both inputs and at R_1.

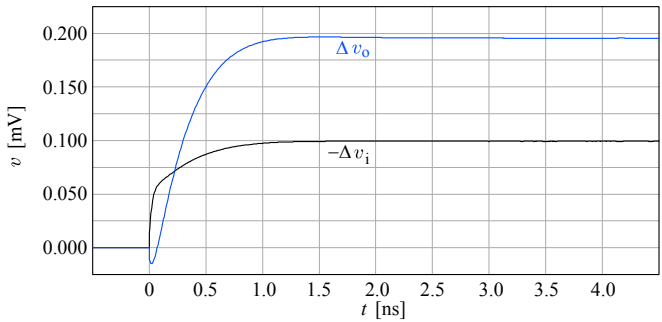


Fig. 5.4.20: Time domain performance of the circuit in Fig. 5.4.18. Note the jump of v_i in the first 100 ps and a pronounced undershoot in Δv_o.

A very interesting circuit, known as the 'Cascomp' (compensated cascode), shown in Fig. 5.4.21, was invented by *Patrick Quinn* [Ref. 5.53–54]. Here the usual differential cascode amplifier, $Q_{1–4}$, is enhanced by indirect error sensing and feedforward error correction. The error, generated at the emitters of $Q_{1,2}$, is also available at the emitters of $Q_{3,4}$, where it appears almost identical (owed to the same bias and signal currents and thus similar emitter resistances). Sensed by $Q_{5,6}$ and amplified adequately, the error is subtracted from the $Q_{3,4}$ collector currents. The addition of a further common base stage, $Q_{7,8}$, lowers the summing node impedance, which increases the summing precision, and at the same time improves the bandwidth of the error sensing amplifier, $Q_{5,6}$. Since the error signal voltage at the $Q_{3,4}$ emitters is low (several mV, or so), the emitter resistances R_{e2} of the error amplifier $Q_{5,6}$ can also be very low, or even eliminated completely, without degrading the linearity.

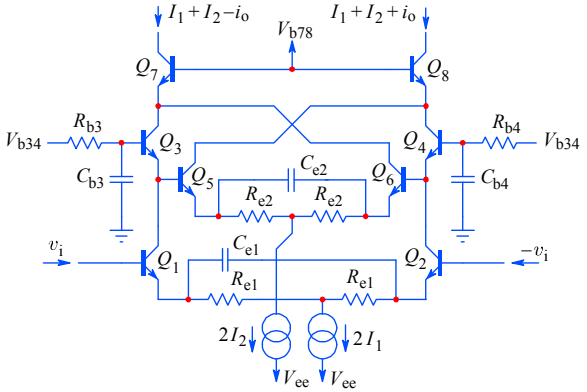


Fig. 5.4.21: The Quinn's 'Cascomp' employs indirect error sensing and error feedforward.

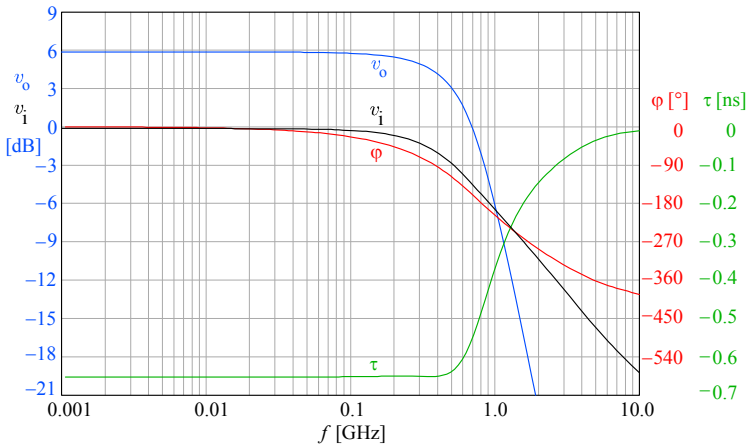


Fig. 5.4.22: Frequency domain performance of the 'Cascomp' amplifier. The bandwidth is about 500 MHz, but note also the high flatness of the envelope delay, right up to the bandwidth limit.

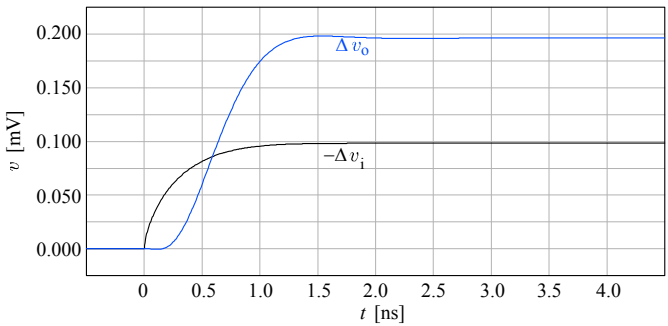


Fig. 5.4.23: Time domain performance of the 'Cascomp' amplifier. The rise time is about 0.7 ns and the initial undershoot is very low.

Fig. 5.4.24 shows a similar circuit, but with a feedback error correction. The error signal is taken at the same point as before, but its amplified version is applied to the emitters of the input differential pair $Q_{1,2}$. Unfortunately, in spite of its attractive concept this configuration is not suitable for high frequencies, since capacitive emitter peaking can not be used (a capacitance in parallel to R_1 would short the auxiliary amplifier outputs, reducing the amount of error correction at high frequencies), thus the bandwidth is about 180 MHz only. But when bandwidth is not the primary design requirement this amplifier can be a valid choice. We shall not plot its performance.

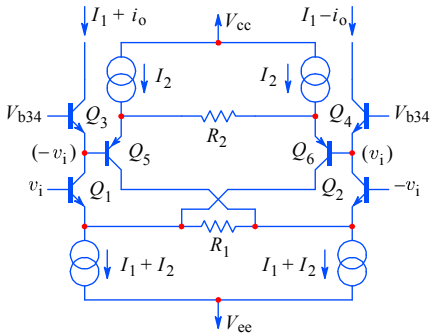


Fig. 5.4.24: Differential cascode amplifier with indirect error sensing and error feedback. Unfortunately this configuration tends to be rather slow.

The circuit in Fig. 5.4.25 represents a modification of the feedforward error correction in which the correction current is summed at the same point where the error was generated. This configuration requires reasonably well matched devices with a high current gain β. Transistors $Q_{1,2}$ form the differential pair, $Q_{5,6}$ form the error sensing amplifier and $Q_{3,4}$ provide an additional V_{be} voltage to enhance the error amplifier dynamic range.

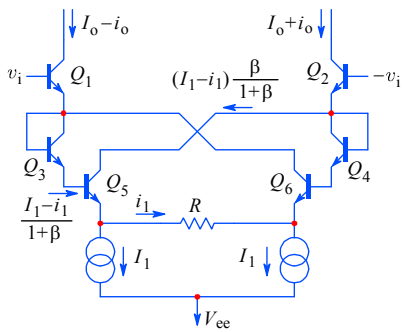


Fig. 5.4.25: A possible modification of error feedforward, employing direct error sensing and direct feedforward error correction.

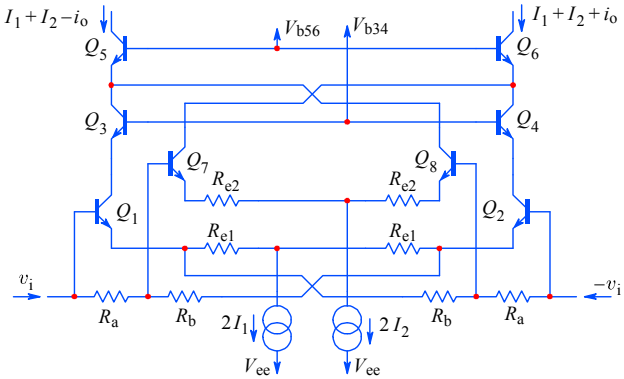


Fig. 5.4.26: Another evolution of the Cascomp is realized by feedback derived error sensing and feedforward error correction. The junction of R_a and R_b is at the 'virtual ground' at which the error of the $Q_{1,2}$ pair is sensed and amplified by the auxiliary amplifier $Q_{7,8}$. The error is subtracted from the output current at the emitters of $Q_{5,6}$.

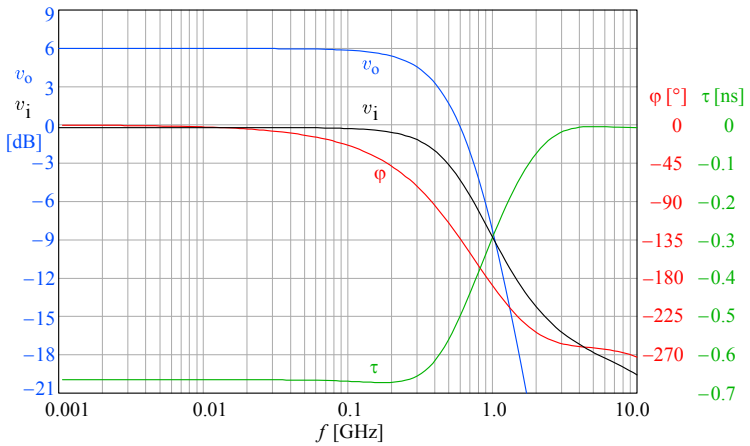


Fig. 5.4.27: Frequency domain performance of the Cascomp evolution amplifier. The bandwidth is a little over 400 MHz.

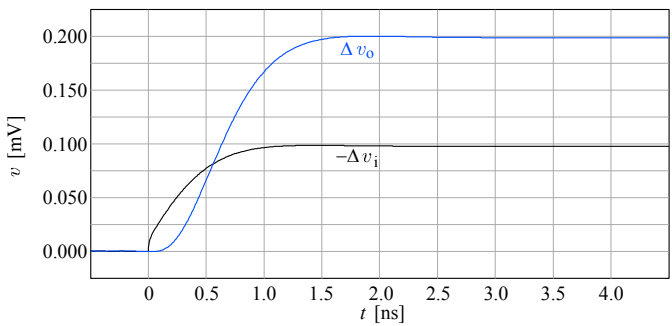


Fig. 5.4.28: Time domain performance of the Cascomp evolution amplifier.

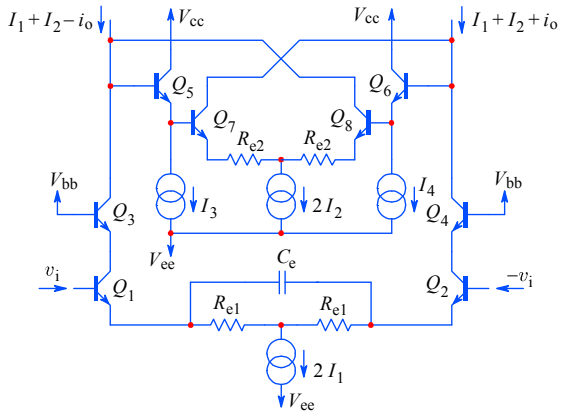


Fig. 5.4.29: This 'output impedance compensation', also patented by Pat Quinn, has direct error sensing and direct feedforward error correction, performed by Q_{5-8}.

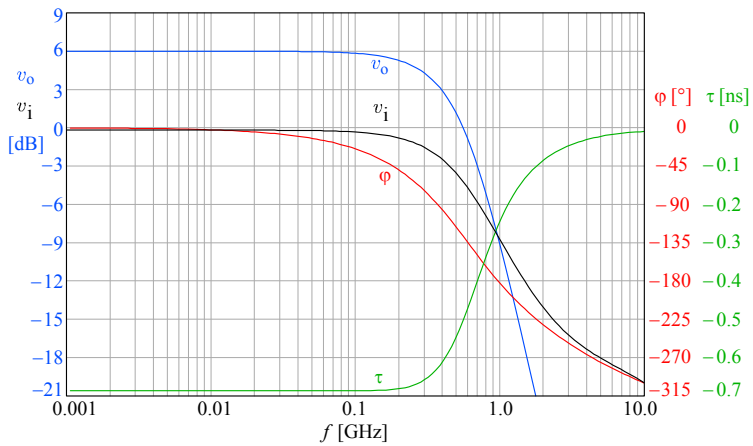


Fig. 5.4.30: Frequency domain performance of the amplifier in Fig. 5.4.29. The bandwidth is about 400 MHz.

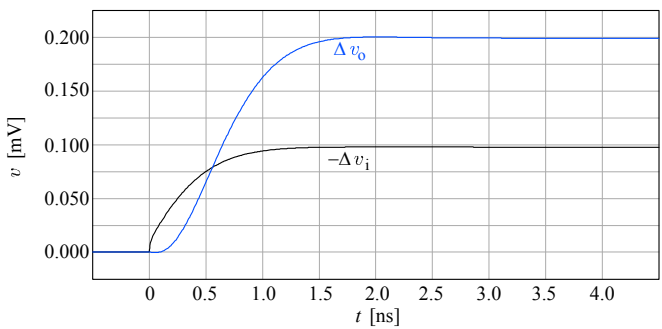


Fig. 5.4.31: Time domain performance of the amplifier in Fig. 5.4.29.

5.4.6 The Tektronix M377 IC

In this section we shall have a brief discussion of the M377 IC, made by Tektronix for their 11 000 series oscilloscope, later also employed in many other models. The M377 was described in [Ref. 5.50 and 5.51] by its designer *John Addis*.

When it was designed, back in mid 1980s, this integrated circuit started a revolutionary trend in oscilloscope design which is still evolving today in the first decade of the XXI century. One of the important design goals was to eliminate the need for manual adjustments as much as possible. Classical oscilloscopes, made with discrete components, required a lot of manual adjustment; for example, the Tektronix 7104 oscilloscope with its two single-channel plug-ins required 32 manual adjustments for correcting the thermal effects only, many of which needed iterative corrections in combination with another one or several other settings. This caused long calibration periods and a lot of bench testing by experienced personnel, increasing the production cost considerably. In contrast, the M377 needs only one manual adjustment (for optimizing the transient response), all other calibration procedures are performed at power–up by a microprocessor, which varies the settings using several DC voltage control inputs. Of course, the calibration can also be done upon the user's request by pressing a push button on the front panel. Some settings, such as the high impedance attenuator calibration, is done in production by laser trimming of the components.

The elimination of almost all electromechanical adjustments and their replacement by electronic controls resulted in circuit miniaturization, reducing strays and parasitics, thus improving bandwidth. But it also increased the circuit complexity (the number of active devices) and density, resulting in higher power dissipation, and consequently higher temperatures, requiring careful thermal compensation.

The first step in this direction was the 'Cascomp', [Ref. 5.53–5.54], which we have met already in Fig. 5.4.21. The first instrument to use the Cascomp was the Tektronix 2465 oscilloscope. Although it represented a significant improvement in precision over a simple cascode differential pair, it also had some limitations. First, the addition of the error amplifier, $Q_{5,6}$, helps to reduce both the nonlinearity and, if the operating points are chosen carefully, also the thermally induced tails. But since both the main and the error amplifier have nearly equal gain, the system gain can not be high, otherwise the error amplifier's non-linearity would show up.

Another problem is the stack of three transistor pairs in the signal path, which results in an increased gain sensitivity to α variations. The gain, as a function of temperature, increases as α^3, about 225 ppm/K for a Cascomp using transistors with $\beta = 80$, compared to some 150 ppm/K of the standard cascode. But a standard cascode also has a counteracting temperature dependent gain term, ≈ -185 ppm/K (at $I_e = 20$ mA, $T_j = 60$°C and $R_e = 40\,\Omega$) owed to the dynamic emitter resistance r_e, which in the Cascomp is compensated. To some degree α effects can be compensated by adding resistors $R_{b3,4}$ to the bases of $Q_{3,4}$, but these resistors can never be made large enough, owing to the transient response requirement for low base resistance (see Part 3, Fig. 3.4.6 and the associated analysis).

The three transistor pair stack (and not forgetting the current source, I_1) further disadvantage the IC against the discrete design, since for a given supply voltage the output linear voltage range is severely reduced. Also, any level shifting back down

to the negative supply voltage, needed by an eventual subsequent stage, requires a greater level shift than in a conventional cascode.

Finally, the Cascomp has a limited ability to handle overdrive signals. The emitters of $Q_{3,4}$ do not 'see' the whole signal during overdrive, thus the error amplifier signal is clipped off at peaks, and as a result the main and the error amplifier experience different thermal histories. Additional circuitry is needed to ensure correct input signal clipping and acceptable thermal behavior.

All these limitations dictated a different approach in the M377 IC. The basic amplifier block (shown in Fig. 5.4.32 and a differential version in Fig. 5.4.33) is already inherently insensitive to thermal influence, and uses local feedback, resulting in high linearity. It can be viewed (simplified) as a compound transistor, with the base represented by the Q_1 base, the emitter by the Q_3 emitter, and the collector by the Q_3 collector, respectively. Since Q_3 is current driven, its V_{be} variation with temperature is not important. Compared with a single transistor, operating at the same current, such a compound transistor has greater g_m and β, an excellent linearity, and no thermals.

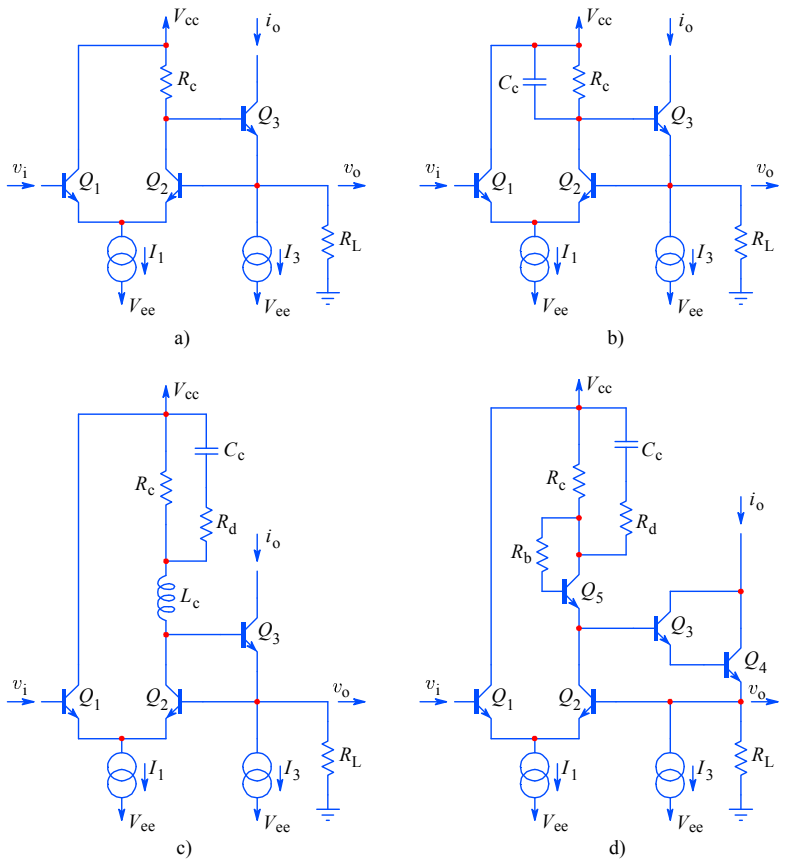


Fig. 5.4.32: **a)** The M377 IC main amplifier block, basic scheme. **b)** Dominant pole compensation. **c)** Inductive peaking. **d)** Inductance created by the Q_5 emitter impedance.

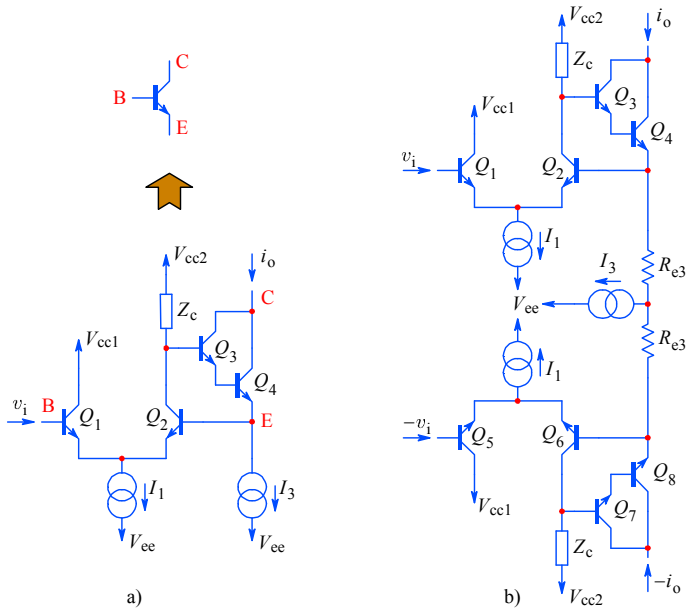


Fig. 5.4.33: a) The basic wideband amplifier block of the M377 IC can be viewed as a compound transistor, in which the Q_1 base, the Q_3 emitter, and the Q_3 collector represent the base, emitter and collector of the compound device. **b)** Two such blocks form a differential amplifier. An improved design results if Q_3 is of a Darlington configuration, Q_3+Q_4. To a high precision the output current is $i_o = v_i/R_{e3}$. The impedance Z_c is the compensation shown in Fig. 5.4.32d.

However, the α of Q_3 is not improved, neither by Q_1, nor Q_2. This could be corrected if, for example, the Q_1 collector were to be connected to the Q_3 emitter, which, besides improving α, would also bootstrap the collector of Q_1, lowering the input capacitance. Unfortunately, the low collector voltage and, consequently the operating point of Q_1, would reduce its cut off frequency. Also, owing to feedback through the capacitance C_{cb} of Q_1, such a circuit can have a negative input impedance at very high frequencies, compromising stability. The best solution for a high f_T ($\approx$ 8.5 GHz in the M377) is to use a Darlington for Q_3 (Q_3+Q_4 in Fig 5.4.33).

One of the most important parameters of a high speed amplifier is the overdrive recovery time. Following the ever increasing requirements for speed and precision we usually specify the time (in ns) needed to settle to, say, 0.1% or even 0.01% of the final level, following an overdrive level of many times the maximum level over a relatively long overdrive period. Fig. 5.4.34 shows the configuration used in the M377, which recovers to 0.04% in just 6 ns after a 2 V overdrive; recovery to 0.01% is about 25 ns. The current sources and Schottky diodes, added to the original circuit of Fig. 5.4.33b, allow separate feedback paths under overdrive. It is important to realize that in this circuit, if not compensated, only the half with the negative input voltage would be overdriven. The positive part then takes all the current provided by the I_3 source. Under this condition D_5 and D_6 cut off, while D_4 conducts owing to I_2 and I_4, allowing the feedback of the lower circuit half to remain operative, preserving the delicate thermal balance.

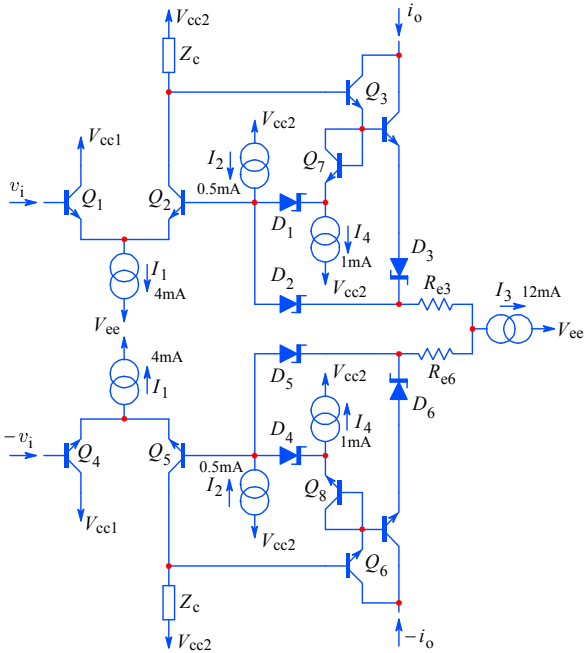


Fig. 5.4.34: The M377 amplifier with the components for high speed overdrive recovery.

The frequency domain and time domain responses of the circuit in Fig. 5.4.34 are shown in Fig. 5.4.35 and 5.4.36, respectively. Note that the compensating impedance Z_c was adjusted to the needs of the transistors used for circuit simulation (BF959, as in all previous circuits, thus allowing comparison), therefore the graphs do not represent the true M377 performance capabilities.

Although it can be argued that the simulation has been performed using a simplified basic version of the circuit, there are a few points to note, which are nevertheless valid. First, there is the potential instability problem (owed to feedback), indicated by the phase plot turning upward and the envelope delay going positive above some 4 GHz. If proper care is not taken, especially compensating the parasitics and strays in an IC environment, the step response might display some initial waveform aberrations, or even ringing and oscillations.

Another point of special attention is the parasitic capacitance to the substrate of the Schottky diodes. Being within the feedback loop, these capacitances can be troublesome. Proper forward bias for low impedance is needed to move those unwanted poles (transfer function zeros) well above the cutoff frequency. High bias would result in high temperatures, which, in a densely packed IC such as this one, can be problematic. Also, since noise increases with temperature, the bias can not be as high as one would like.

As an advantage, judging by the constant slope of the input voltage plot, the circuit input impedance is well behaved, thus the loading of a previous stage should not be critical. Likewise, the active inductive peaking, realized by the base resistance

R_b and transformed as inductance at the Q_5 emitter (as shown in Fig. 5.4.32d), offers a simple way of adjusting the frequency compensation network.

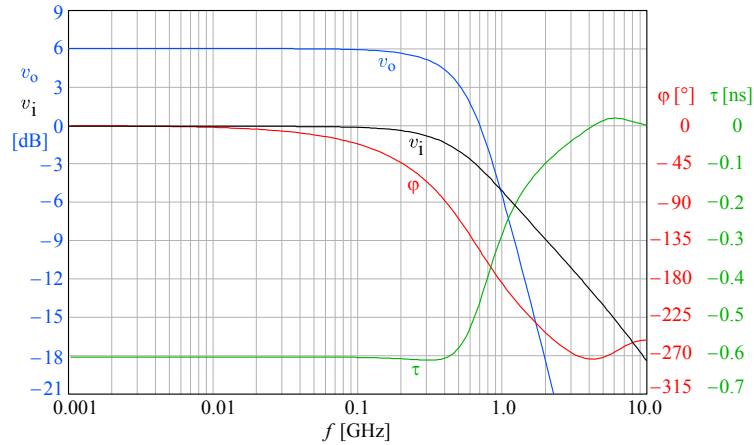


Fig. 5.4.35: Frequency domain performance of the amplifier in Fig. 5.4.34. The bandwidth of the simulated circuit is about 500 MHz, but this is owed to the transistor used (BF959) and the frequency compensation network adjusted in accordance, therefore the graph is not representative of the actual M377 IC performance.

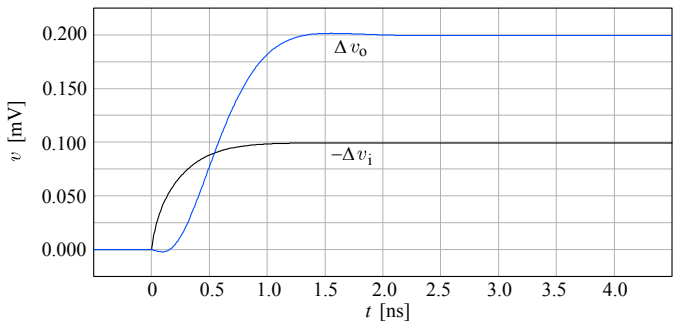


Fig. 5.4.36: Time domain performance of the amplifier in Fig. 5.4.34. The comment in the caption of Fig. 5.4.35 also holds here.

Now take a close look at the circuit in Fig. 5.4.34; in particular, the diode pairs $D_{2,3}$ and $D_{5,6}$, the resistors $R_{e3,6}$ and the current source I_3; if another such block is added in parallel (with different values of resistors $R_{e3,6}$), and if the current sources are switched on one at a time, a gain switching in steps can be achieved. The bandwidth would change only slightly with switching. Fig. 5.4.37 shows such a circuit with two gain values, but several more can easily be added.

Another way of changing the vertical sensitivity is to use a fixed gain amplifier and switch the attenuation at its output, as shown in Fig. 5.4.38. In this way the bandwidth is preserved, but attenuation switching has its own weak points, such as a reduced signal range and higher noise at higher attenuation.

As a point of principle, switching the amplifier gain is preferred to fixed gain with switched attenuation. Although it alters the bandwidth, gain switching preserves the signal's dynamic range at all settings, whilst the system with fixed gain and an attenuator will have a comparable dynamic range only with no attenuation; at any other attenuation setting the dynamic range would be proportionally reduced. Also, gain switching systems will preserve the same noise level, whilst the fixed gain systems will have the lowest signal to noise ratio at maximum attenuation.

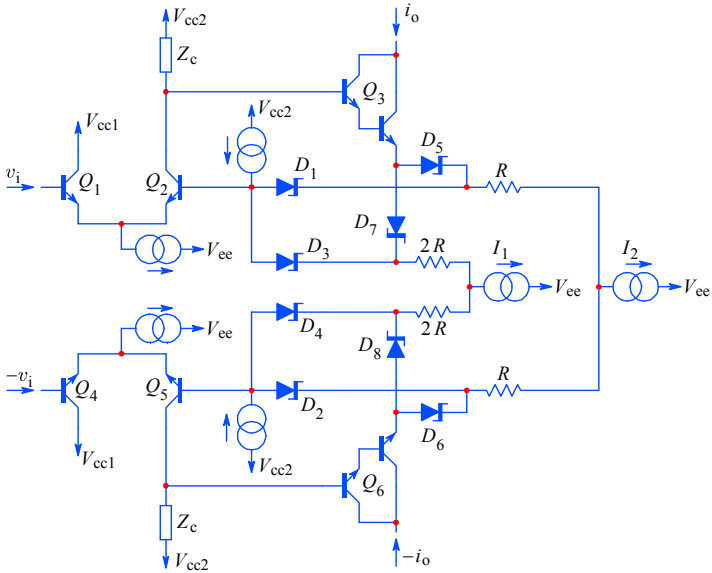


Fig. 5.4.37: Gain switching in steps was achieved by adding one or more emitter current sources (I_1, I_2, …) with appropriate resistor values and Schottky diode pairs and switching on one current source at a time.

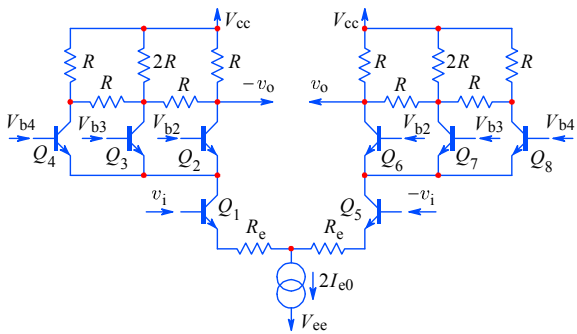


Fig. 5.4.38: With an R–$2R$ network the attenuation can be switched in steps by applying a positive DC voltage $V_{b2,3,4}$ to $Q_{2,6}$, $Q_{3,7}$, $Q_{4,8}$, one pair at a time; at each collector the load is $R||2R = 2R/3$. With $Q_{2,6}$ on the gain is $A = 2R/3R_e$, with $Q_{3,7}$ on $A = R/3R_e$, and with $Q_{4,8}$ on $A = R/6R_e$, effectively halved by each step. A similar circuit was used in the Tek 2465.

For small gain or attenuation changes of, say, 1:4, as commonly found in oscilloscope amplifiers, these differences can be small. However, in M377, the gain

switching had a 50:1 range. With such a high gain range the frequency compensation needed to be readjusted (for the highest gain no compensation was needed).

5.4.7 The Gilbert Multiplier

A similar problem of compensation readjustment as a function of gain is encountered with a continuously variable gain. Although used only occasionally, a continuously variable gain is a standard feature of almost all oscilloscopes and no manufacturer dares to exclude it, even in digital instruments (although there it is done in very small steps).

In older analog instruments a simple potentiometer was used. This worked well up to some 20 MHz. For higher frequencies the 'pot' size and the variable impedance at the slider represented major difficulties, even if the required gain change was within a relatively small range, ordinarily about 3:1. At Tektronix, an ingenious wire pot was used, having a bifiliar winding to cancel the inductance, but its parasitic capacitance, which also varied with the setting, was causing too much cross-talk at higher frequencies. Finally, there was also the mechanical problem of placing the pot at the correct point in the circuit and still being able to bring its axis on the front panel, aligned with the main attenuator switch.

A much more elegant choice is to use some sort of electronic gain control, by using either a voltage or a current controlled amplifier (VCA or ICA). Such an amplifier modulates (ideally) only the signal amplitude, a process which can be mathematically described as multiplication of a signal by a DC voltage; thus we often refer to those amplifiers as multipliers or modulators. Of course, electronic gain control has its own problems and great care is needed to make it linear enough and fast enough, as well as not too noisy. But it solves the problem of mechanical pot placement, since it now has to handle only a DC control signal, so the pot can be placed at any convenient place. In digital systems, the pot is replaced by a digital to analog converter (DAC; in lower speed instruments, a multiplying DAC can be used to attenuate the signal directly, replacing the VCA altogether).

Oscilloscopes do not need to exploit the full modulation range, as RF modulators normally do. In contrast to RF modulators, which are four–quadrant devices (both the carrier and the modulation are AC signals), the gain control in oscilloscopes needs to work in two quadrants only (AC signal and DC control); four quadrants would allow simple gain inversion, but this is more accurately done by a switch. Therefore the modulation cross-talk or the common mode rejection at HF is not an issue. On the other hand, DC stability is important since it directly affects measurement accuracy. Whilst RF modulators operate over a limited frequency range, for oscilloscopes the wideband gain flatness at all gain settings is also very important.

The simple differential amplifier in Fig. 5.4.8a can perform the variable gain control by varying the emitter current. If we assume that the modulation voltage is $v_M = V_M - V_{be} - V_{ee}$, the modulation current is:

$$2I_e = \frac{v_M}{R_e} \tag{5.4.64}$$

By inserting this into the gain equation of the differential amplifier the multiplication function results:

$$v_o = v_{bb}\frac{2R}{2\,r_e} = v_{bb}\frac{R\,I_e}{V_T} = v_{bb}\cdot v_M\,\frac{R}{2\,R_e V_T} \tag{5.4.65}$$

Unfortunately the bandwidth is also proportional to the emitter current and with the usual values of R and stray capacitances, the dominant pole at low currents can be very low. In addition the output common mode level also changes with current.

Almost all wideband multipliers are based on one of the variations of the basic circuit, now known as the Gilbert multiplier, after its inventor *Barrie Gilbert* (see [Ref. 5.56–5.62]). The circuit development can be followed from Fig. 5.4.39a by noting that if the output is to be a linear function the input has to be nonlinear.

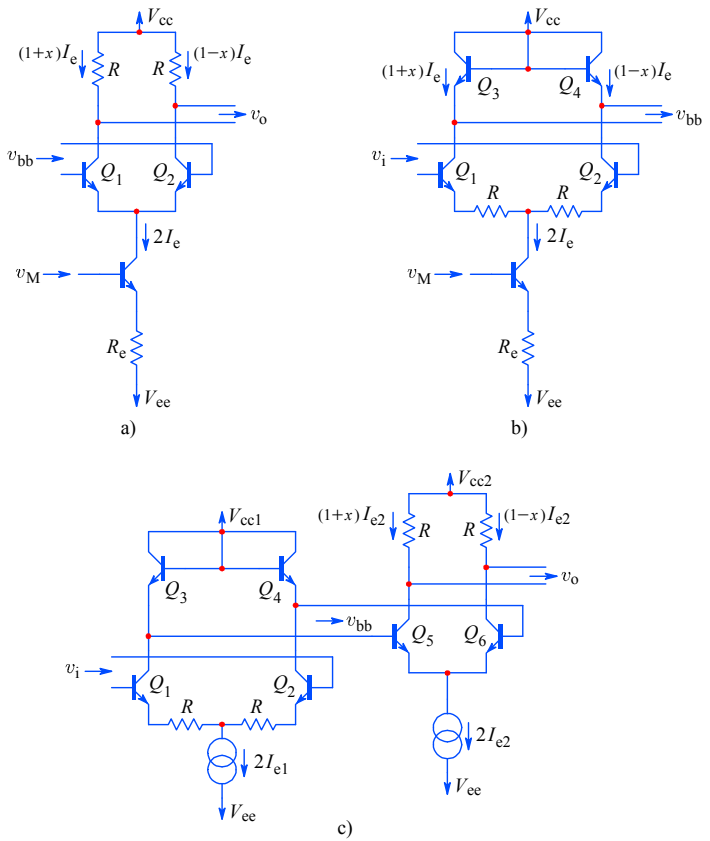


Fig. 5.4.39: The Gilbert multiplier development. **a)** The gain of an ordinary differential amplifier is proportional to the emitter current, but so is the bandwidth. Also, for a linear output a nonlinear input is required. **b)** Inverting the function by linearizing the g_m stage and loading it by simpler p–n junctions gives a nonlinear output, such as required by the circuit in a) to give a linear output. **c)** The simple combination of b) and a) gives the so called 'translinear' stage.

Starting from the exponential I_e–V_{be} relationship:

$$I_e = I_s\left(e^{V_{be}/V_T} - 1\right) \tag{5.4.66}$$

and considering that the intrinsic current $I_s \approx 10^{-14}$ A, then even for currents as low as 1 nA we can say that $I_e \gg I_s$; so we are not making a big mistake if we use:

$$I_e \approx I_s\, e^{V_{be}/V_T} \tag{5.4.67}$$

Since the differential pair was made by the same IC process, we can expect that the devices will be reasonably well matched, so $I_{s1} \approx I_{s2}$ and their temperature dependence will also be well matched if the transistors are at the same temperature:

$$\frac{I_{e1}}{I_{e2}} = \frac{I_{s1}}{I_{s2}}\, e^{(V_{be1}-V_{be2})/V_T} \approx e^{(V_{be1}-V_{be2})/V_T} \tag{5.4.68}$$

In order to achieve a linear output current the expected input voltage should follow the logarithmic function:

$$v_{bb} = \Delta V_{be} = V_T \ln\frac{I_{e1}}{I_{e2}} \tag{5.4.69}$$

If i_e is the modulation component superimposed on the DC current I_{e0}, we can write:

$$I_{e1} = I_{e0} + i_e \qquad \text{and} \qquad I_{e2} = I_{e0} - i_e \tag{5.4.70}$$

Let x be the AC to DC component ratio:

$$x = \frac{i_e}{I_{e0}} \tag{5.4.71}$$

Then the currents are:

$$I_{e1} = I_{e0}(1 + x) \qquad \text{and} \qquad I_{e2} = I_{e0}(1 - x) \tag{5.4.72}$$

and, returning to Eq. 5.4.69, we obtain the required nonlinear input function:

$$v_{bb} = V_T \ln\frac{I_{e0}(1 + x)}{I_{e0}(1 - x)} = V_T \ln\frac{1 + x}{1 - x} \tag{5.4.73}$$

How can this nonlinear relationship at the differential amplifier's input be realized? By inverting the Fig. 5.4.39a circuit's function, such as in Fig. 5.4.39b; that is, by using relatively large emitter degeneration resistors, resulting in a linear voltage to current conversion and loading the collectors by similar p–n junctions as V_{be}, would produce exactly such a nonlinear relationship as the original circuit needs at its input to produce a linear output.

Fig. 5.4.39c is thus a simple combination of the b and a circuits, but with some very interesting properties. First, it is compensated and thus quite linear within the entire input range ($-1 \le x \le +1$). Also the v_{bb} swing is small (less than V_T) owing to the very low impedance ($\approx V_T/I_e$) of $Q_{3,4}$, which means that charging and discharging of stray capacitances is minimal, so the bandwidth limit is owed to the collector impedances of $Q_{5,6}$ and the transistors' f_T. Finally, the gain is entirely 'current mode', with low temperature dependence (both V_T and I_s are canceled in the expression for v_o) and the gain control is set by the current ratio I_{e2}/I_{e1}.

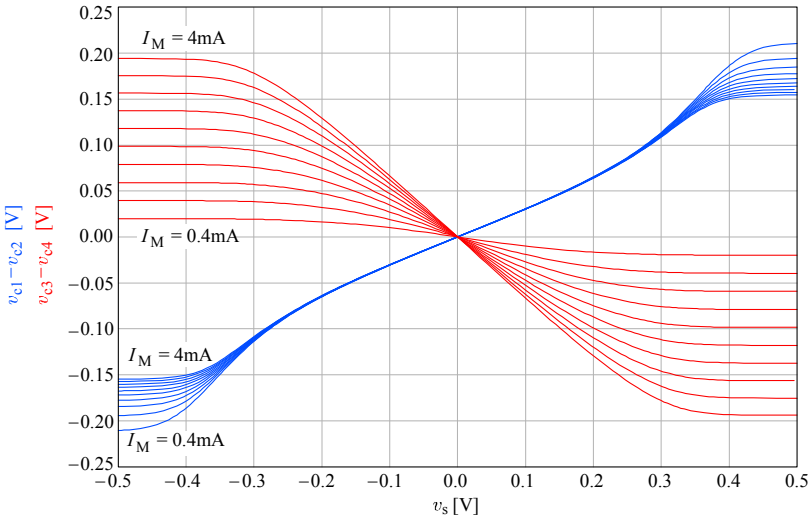


Fig. 5.4.40: DC transfer function of the Gilbert multiplier of Fig. 5.4.39, for $2I_{e1} = 4$ mA and modulation current $2I_{e2} = I_M = 0.4$–4 mA.. The signal source (v_s) range is ±0.5V.

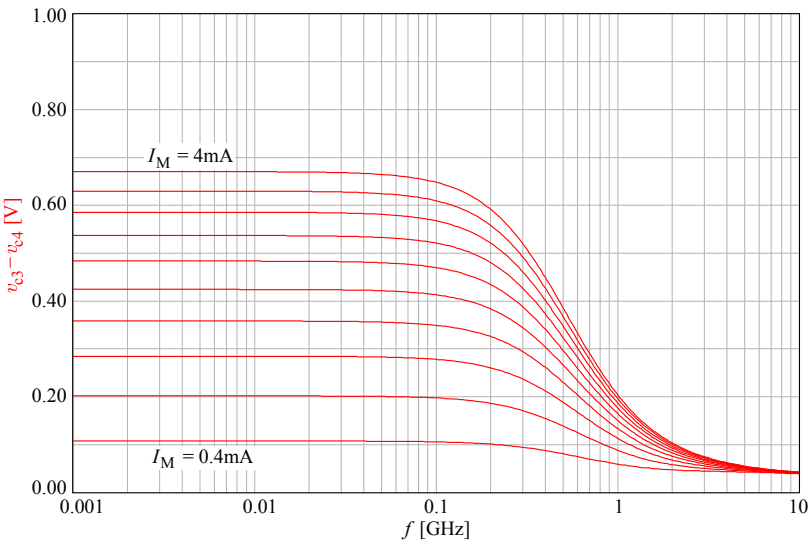


Fig. 5.4.41: The Gilbert multiplier bandwidth is almost constant over the 10:1 modulation current range.

Another way of developing this 'translinear gain cell' can be followed by observing Fig. 5.4.42. Two current mirrors, with the current gain proportional to the emitter area A, can be interconnected by breaking the two emitters carrying the mirrored currents and biasing them by a current source, thus forming a differential

amplifier, whose input nonlinearity is compensated by the nonlinearity of the remaining two transistors.

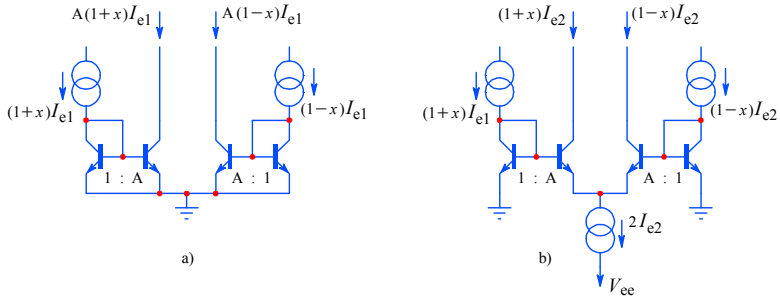


Fig. 5.4.42: Another way of developing the Gilbert multiplier is by interconnecting two current mirrors into a differential amplifier, whose input nonlinearity is compensated by the nonlinearity of the two grounded transistors.

Once the basic multiplier is developed it is relatively easy to construct a four quadrant multiplier, Fig. 5.4.43, by adding another differential pair with inputs in parallel and the collectors cross-coupled.

A further differential amplifier can be used to perform the voltage to current conversion and splitting the current source ($4I_e$) into the required signal currents. The gain is now controlled by biasing the compensation transistors with current.

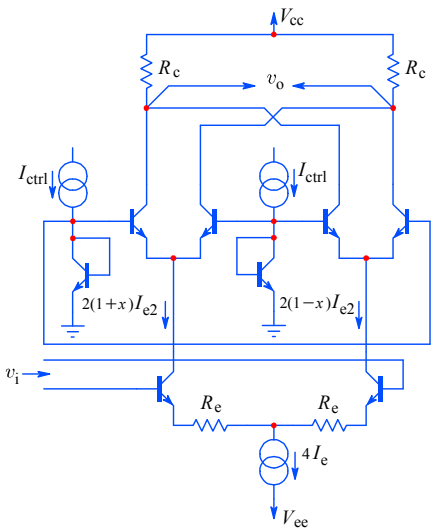


Fig. 5.4.43: A four quadrant multiplier developed from the previous circuit. The gain is controlled by current biasing the compensation transistors. Four quadrant operation is obtained from the fact that the cross-coupled collectors cancel out if the two tail currents are equal and the third differential amplifier allows it to distribute the tail currents in a symmetrical manner about the mid–bias value. Thus both the input and the control can be AC signals and can also be mutually exchanged, not compromising the bandwidth or the linearity.

By cross-coupling the collectors of the two differential pairs we have achieved effective output current cancellation if the two control currents are equal and the two emitter currents are equal. Varying any pair of currents about this mid–point changes the polarity as well as the gain of the multiplier for the other input. Thus, a nice byproduct of four–quadrant configuration is that the 'signal' and the 'control' inputs can be mutually exchanged, without compromising the bandwidth or the linearity.

Returning to the M377 design, where only a two–quadrant multiplication is needed, the multiplier in Fig. 5.4.44 represents the circuit used. It is almost identical to the four–quadrant multiplier, except that the collectors are not cross-coupled and the output is taken from a single pair. The differential circuit symmetry was retained merely because of its good thermal balance and DC stability. For the same reason, all four collectors must be equally loaded.

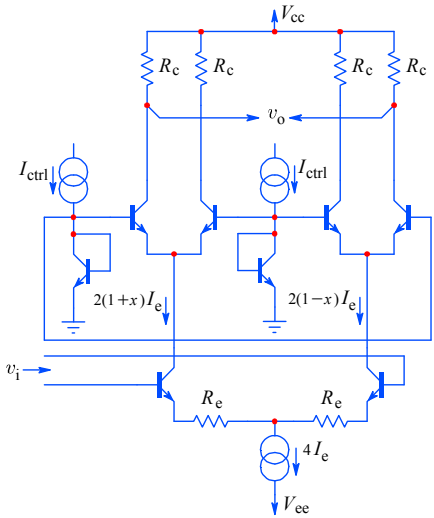


Fig. 5.4.44: Two quadrant multiplication is sufficient for the oscilloscope continuously variable gain control; however, the same differential symmetry of the four quadrant multiplier has been retained for the M377 because of good thermal balance and DC stability.

The multiplier circuits shown represent the basic linearization principle. In actual implementations, a number of additional linearization techniques are used, most of them also patented by Barrie Gilbert while he was at Tektronix[1], and some later when at Analog Devices, [Ref. 5.56–5.61].

[1] It might be interesting to note that Barrie Gilbert published an article [Ref. 5.56] describing his multiplier before Tektronix applied for a patent. Motorola quickly seized the opportunity and started producing it (as the MC1495). Tektronix claimed the primarity and Motorola admitted it, but nevertheless continued the production, since once published the circuit was 'public domain'. Barrie's misfortune gave the opportunity to many generations of electronics enthusiasts (including the authors of this book) to play with this little jewel and use it in many interesting applications. Thanks, Barrie!

Résumé of Part 5

In this part we have briefly analyzed some of the most important aspects of system integration and system level performance optimization, with a special emphasis on system bandwidth.

We have described the transient response optimization by a pole assignment process called 'geometrical synthesis' and showed how it can be applied using inductive peaking. We have discussed the problems of input signal conditioning, the linearization and error reduction and correction techniques, employing the feedback and feedforward topologies at either the system level or at the local, subsystem level. We have also revealed and compared some aspects of designing wideband amplifiers using discrete components and modern IC technology.

On the other hand, we have said very little about other important topics in wideband instrumentation design, such as adequate power supply decoupling, grounding and shielding, signal and supply path impedance control by strip line and microstrip transmission line techniques, noise analysis and low noise design, and the parasitic impedances of passive components. But we believe that those subjects are extensively covered in the literature, some of it also cited in the references, so we have tried to concentrate on the bandwidth and transient performance issues.

We have also said nothing about high sampling rate analog to digital conversion techniques, now already established as the essential ingredient of modern instrumentation. While there are many books discussing AD conversion, most of them are limited to descriptions of applying a particular AD converter, or, at most, to compare the merits of one conversion method against others. Only a few of them discuss ADC circuit design in detail, and even fewer the problems of achieving top sampling rates for a given resolution, either by an equivalent time or a real time sampling process, time interleaving of multiple converters, combining analog and digital signal processing and other techniques, which today (first decade of the XXI century) allow the best systems to reach sampling rates of up to 20 GSps (Giga Samples per second) and bandwidths of up to 6 GHz.

Just like many other books, this one, too, ends just as it has become most interesting (the reader might ask himself whether there is really nothing more to say or whether the authors simply ran out of ideas? — since most of the circuits presented are not of our origin, and electronics certainly is an art of infinite variations, we the authors can be, one hopes, spared the blame). As already said in the Foreword, the most difficult thing when writing about an interesting subject is not what to include, but what to leave out. Although we discuss the effects of signal sampling in Part 6 and a few aspects of efficient system design combining analog and digital technology in Part 7, this book is about amplifier design, so we leave the fast ADC circuit design discussion for another opportunity.

References:

[5.1] *D.L. Feucht*, Handbook of Analog Circuit Design,
Academic Press, Inc. San Diego, 1990
See also the latest compact disk version at <http://www.innovatia.com/>

[5.2] *S. Roach*, Signal Conditioning in Oscilloscopes and the Spirit of Invention,
(*J. Williams*, [Editor], The Art and Science of Analog Circuit Design, Part 2)
Butterworth–Heinemann, Boston, 1995

[5.3] *P.R. Gray & R.G. Meyer*, Analysis and Design of Analog Integrated Circuits,
John Wiley, New York, 1969

[5.4] *J. Williams*, Composite Amplifiers,
Linear Technology Application Note AN-21, July 1986

[5.5] *P. Starič*, Wideband JFET Source Follower,
Electronic Engineering, August 1992

[5.6] *J. Williams*, Measuring 16–bit Settling Times: the Art of Timely Accuracy,
EDN Magazine, Nov. 19, 1998, <http://www.ednmag.com/>

[5.7] *W. Kester*, High Speed Design Techniques, Analog Devices, 1996,
<http://www.analog.com/>

[5.8] *A.D. Evans* (Editor), Designing with Field Effect Transistors, Siliconix, McGraw-Hill, 1981

[5.9] *J.E. Lilienfeld*, Method and Apparatus for Controlling Electric Current,
US Patent 1 745 175, Jan. 28, 1930
(the FET patent, apparently predating the BJT patent by about 23 years)

[5.10] FET Design Catalog, Siliconix, 1982, <http://www.siliconix.com/>

[5.11] IC Applications Handbook,
Burr-Brown, 1994, <http://www.burr-brown.com/>, <http://www.ti.com/>

[5.12] A New Approach to OpAmp Design, Application Note AN 300-1,
Comlinear Corporation, March, 1985

[5.13] *S. Franco*, Design with Operational Amplifiers and Analog ICs, McGraw-Hill, 1988

[5.14] *F.A. Muller*, High Frequency Compensation of RC Amplifiers,
Proceedings of the I.R.E., August, 1954, pp. 1271–1276.

[5.15] *B. Orwiller*, Vertical Amplifier Circuits, Tektronix, Inc., Beaverton, Oregon, 1969.

[5.16] *F.W. Grover*, Inductance Calculation, (Reprint)
Instrument Society of America, Research Triangle Park, N. C. 27 709, 1973.

[5.17] *J. Williams*, High Speed Amplifier Techniques, Application Note AN-47,
Linear Technology, March, 1985, <http://www.lt.com/>

[5.18] *C.R. Battjes*, Monolithic Wideband Amplifier,
US Patent 4 236 119, Nov. 25, 1980[2]

[2] Note : For US Patents go to <http://www.uspto.com/> and type the patent number in the Search pad. Patent figures are in TIFF graphics format, so a TIFF Viewer software is recommended (links for downloading and installation are provided within the USPTO web pages).

[5.19] *J.L. Addis*, *P.A. Quinn*, Broadband DC Level Shift Circuit With Feedback,
US Patent 4 725 790, Feb. 16, 1988

[5.20] *J.L. Addis*, Precision Differential Amplifier Having Fast Overdrive Recovery,
US Patent 4 714 896, Dec. 22, 1987

[5.21] *J.L. Addis*, Buffer Amplifier,
US Patent 4 390 852, Jun. 28, 1983

[5.22] *J.L. Addis*, Feedbeside Correction Circuit for an Amplifier,
US Patent 4 132 958, Jan. 2, 1979

[5.23] *T.C. Hill*, *III*, Differential Amplifier with Dynamic Thermal Balancing,
US Patent 4 528 516, July 9, 1985

[5.24] *J. Woo*, Wideband Amplifier with Active High Frequency Compensation,
US Patent 4 703 285, Oct. 27, 1987

[5.25] *Wakimoto*, *Tsutomu*, *Azakawa*, *Yukio*, Wideband Amplifier,
US Patent 4 885 548, Dec. 5, 1989

[5.26] *H. Weber*, A Method of Predicting Thermal Stability,
Motorola, Application Note AN-128

[5.27] *R. Ivins*, Measurement of Thermal Properties of Semiconductor Devices,
Motorola, Application Note AN-226

[5.28] *B. Botos*, Nanosecond Pulse Handling Techniques in IC Interconnections,
Motorola, Application Note AN-270

[5.29] Field Effect Transistors In Theory and Practice,
Motorola, Application Note AN-211A

[5.30] *N. Freyling*, FET Differential Amplifier,
Motorola, Application Note AN-231

[5.31] *R.W. Anderson*, s-Parameter Techniques for Faster, More Accurate Network Designs,
Hewlett-Packard, Application Note AN-95-1

[5.32] *R. Gosser*, Wideband Transconductance Generator (Quad–Core Amplifier),
US Patent 5 150 074, Sep. 22, 1992

[5.33] New Complementary Bipolar Process, National Semiconductor, Application Note AN-860

[5.34] *J. Bales*, A Low Power, High Speed, Current Feedback OpAmp with a Novel Class AB High Current Output Stage, IEEE Journal of Solid-State Circuits, Vol. 32, No. 9, Sept. 1997

[5.35] Development of an Extensive SPICE Macromodel for Current Feedback Amplifiers,
National Semiconductor Application Note AN-840, July, 1992

[5.36] Topics on Using the LM6181 – A New Current Feedback Amplifiers,
National Semiconductor Application Note AN-813, March, 1992

[5.37] *T.T. Regan*, Designing with a New Super Fast Dual Norton Amplifier,
National Semiconductor Application Note AN-278, Sept., 1981

[5.38] Simulation SPICE Models for Comlinear OpAmps,
National Semiconductor Application Note OA-18, May, 2000

[5.39] *H.S. Black*, Translating System,
U.S. Patent 1 686 792, Oct. 9, 1928

[5.40] *H.S. Black*, Wave Translation System,
U.S. Patent 2 102 671, Dec. 21, 1937

[5.41] *H.S. Black*, Inventing the Negative Feedback Amplifier,
IEEE Spectrum, vol. 14, pp. 55–60, Dec., 1977

[5.42] *P.J. Walker*, Current Dumping Audio Amplifier,
Wireless World, vol. 81, pp. 560–562, Dec., 1975

[5.43] *P.J. Walker*, *M.P. Albinson*, Distortion–Free Amplifiers,
U.S. Patent 3 970 953, Jul. 20, 1976

[5.44] *J. Vanderkooy*, *S.P. Lipshitz*, Feedforward Error Correction in Power Amplifiers,
Journal of the Audio Engineering Society, vol. 28, No. 1/2, pp. 2–16, Jan./Feb., 1980

[5.45] *M.J. Hawksford*, Distortion Correction in Audio Power Amplifiers,
Journal of the Audio Engineering Society, vol. 29, No. 1/2, pp. 27–30, Jan./Feb., 1981

[5.46] *M.J. Hawksford*, Distortion Correction Circuits for Audio Amplifiers,
Journal of the Audio Engineering Society, vol. 29, No. 7/8, pp. 503–510, Jul./Aug., 1981

[5.47] *M.J. Hawksford*, Low Distortion Programmable Gain Cell Using Current Steering Cascode Topology, Journal of the Audio Engineering Society, vol. 30, No. 11, pp. 795–799, Nov., 1982

[5.48] *S. Takahashi*, *S. Tanaka*, Design and Construction of a Feedforward Error Correction Amplifier, Journal of the Audio Engineering Society, vol. 29, pp. 31–37, Jan./Feb., 1981

[5.49] *A.M. Sandman*, Reducing amplifier distortion by error add-on,
Wireless World, vol. 79, pp. 32, Jan., 1973

[5.50] *J.L. Addis*, Versatile Analogue Chip for Oscilloscope Plug-Ins,
Electronic Engineering, Aug., 1988 (Part I), Sept., 1988 (Part II)

[5.51] *J.L. Addis*, Good Engineering and Fast Vertical Amplifiers,
(*J. Williams*, [Editor], Analog Circuit Design, Art, Science and Personalities, Part 1)
Butterworth-Heinemann, Boston, 1991

[5.52] *C. R. Battjes*, Who Wakes the Buglar?,
(*J. Williams*, [Editor], The Art and Science of Analog Circuit Design, Part 2)
Butterworth-Heinemann, Boston, 1995

[5.53] *P.A. Quinn*, Feed-Forward Amplifier,
US Patent No. 4 146 844, Mar. 27, 1979

[5.54] *P.A. Quinn*, Feed-Forward Amplifier,
US Patent No. 4 146 844 (Reissue 31 545), Mar. 27, 1984

[5.55] *P.A. Quinn*, Differential Impedance Neutralization Circuit,
US Patent No. 4 692 712, Sep. 8, 1987

[5.56] *B. Gilbert*, A Precise Four–quadrant Multiplier with Subnanosecond Response,
IEEE Journal of Solid-State Circuits, pp. 365–373, Dec. 1968

[5.57] *B. Gilbert*, Multiplier Circuit,
US Patent No. 4 156 283, May 22, 1979

[5.58] *B. Gilbert*, High accuracy four–quadrant multiplier also capable of four–quadrant division,
US Patent No. 4 586 155, April 29, 1986

[5.59] *B. Gilbert*, Synchronous logarithmic amplifier,
US Patent No. 5 298 811, March 29, 1994

[5.60] *B. Gilbert*, Single supply analog multiplier,
US Patent No. 6 074 082, June 13, 2000

[5.61] *B. Gilbert*, Circuit having dual feedback multipliers,
US Patent No. 6 456 142, September 24, 2002

[5.62] *B. Gilbert*, Where Do Little Circuits Come From?,
(*J. Williams* [Editor], Analog Circuit Design, Part 1)
Butterworth-Heinemann, Boston, 1991

[5.63] *P. Gamand*, *C. Caux*, Semiconductor device comprising a broadband and high gain monolithic integrated circuit for a distributed amplifier,
US Patent No. 5,386,130, January 31, 1995

[5.64] *K. Nakahara*, *Y. Sasaki*, Distributed amplifier and bidirectional amplifier,
US Patent No. 5,414,387, May 9, 1995

[5.65] *R.W. Chick*, Non-uniformly distributed power amplifier,
US Patent No. 5,485,118, January 16, 1996

[5.66] *C.D. Motchenbacher*, *F.C. Fitchen*, Low Noise Electronic Design,
John Wiley, New York, 1973, ISBN 0-471-61950-7

[5.67] *H.W. Bode*, Network Analysis And Feedback Amplifier Design,
Van Nostrand, New York, 1949

[5.68] *E. Margan*, *RC* Attenuator Distortion,
Electronics World + Wireless World, Sept., 1992, Circuit Ideas, pp. 765

[5.69] *E. Margan*, Amplifier Instability,
Electronics World + Wireless World, Apr., 1998, pp. 311–312

P. Starič, E. Margan:

Wideband Amplifiers

Part 6:

Computer Algorithms for Analysis and Synthesis of Amplifier–Filter Systems

If you search for something long enough,
you will certainly discover something else.

(Erik's First Amendment to Murphy's Law
applied to scientific research)

How It All Began

Ever since I heard of 'electronic brains', back in the early 1960s, I was asking all sorts of people how those things work, but I never received any answer, with the exception of one, coming from a medical profession, saying that no one understood the biological brains either, so how could I expect to understand the electronic ones? Much later I discovered that an awful lot of people did not like using their brains at all, and they were doing just fine without it, thank you for asking!

In the autumn of 1974, as a student, I had a limited access to an IBM-1130 machine while attending a course on Fortran. It was not exactly a top model of punched card technology, but it could be used for many purposes, other than calculating the monthly wages of the University personnel. As a beginner, it took me nearly three months to program and run a simple $1 - e^{-t/RC}$ response. I had just heard of Moore's law, but I guessed he was exaggerating, since I could do the same job with a slide rule in little less than four hours, so I expected that within my professional lifetime I would never need a computing power that sophisticated. How wrong I was!

In the spring of 1975 my father bought me an HP-29-C, a programmable pocket calculator with 'scientific' functions, sines, cosines, logs, exps, and all that jazz. And in addition to the four stack registers it had some 96 program registers of 'continuous' memory (CMOS, thus the '-C' suffix). Many of my colleagues had similar toys, too, but while most of them were playing the then very popular 'Moon Landing Simulator' (in which you typed in the amount of fuel to burn at each step and in response it displayed your speed and altitude — and a flashing display on crashing), I was busy programming the 0.5 dB tolerance frequency response of an RIAA phonograph correction network, optimized to standard E-12 R and C values. Next year in summer, I made a preamp based on those calculations and inserted it between a simple five transistor power amp and my new Transcriptor's 'Skeleton' turntable with a 'Vestigal' tonearm and a Sonus 'Blue Label' pickup. It worked perfectly and sounded beautiful.

In the late 1970s I was totally devoted to audio; however, I had good relations with the local Motorola representative, who was supplying me with the latest data sheets and application notes, so I simply could not have missed the microprocessor revolution. But I was still using digital chips in the same way as analog ones — by building the hardware, its function was determined once and for all; I never thought of it as something programmable or adaptable to different tasks.

It was only in the early 1980s, with the first PCs, that I really began to devote a substantial part of my working time to programming, and even that was more out of necessity than desire. I was working on the signal sampling section for a digital oscilloscope project, so I had to know all the interactions with the microprocessor. I was also busy drawing printed circuit boards with one of the first CAD programs, the Wintek's 'smArtwork'. Some time later I received a first demo version of the Spectrum-Software's 'MicroCAP' circuit simulator and then the 'PCMatlab' by The MathWorks, Inc.

Then, one day Peter came to my lab and asked me if I could do a little circuit simulation for him. This turned out to be the start of a long friendship and one of the results of it is now in front of you.

Contents:

List of Figures:

List of routines:

6.0 Aim and Motivation

The analysis and synthesis of electronic amplifier–filter systems using symbolic calculus is clearly very labor intensive, as we have seen in the previous parts of the book. The introduction of computers opened up the possibility of solving such problems numerically. Then, of course, the solutions can be approximated only, but it is always possible to trade calculation time for accuracy.

Unfortunately, commercially available computer programs are not well suited to the job from the system designer's point of view. Most of the so called CAD/CAE programs require actual specifications of active devices and passive components before the circuit's simulation can start. The better the program, the more complex are the device models used. Also, these programs will perform circuit analysis only, so the user is left on his own for circuit synthesis and configuration selection, as well as for modeling the influence of parasitic components, which in return depend on the particular circuit topology and devices used.

When designing a system, we usually like to start from some ideal performance goals, so we would prefer to use small and flexible algorithms which would allow us to quickly calculate and compare the required responses of several systems, before we decide what kind of system to use and the technology to realize it. As this book deals with amplifiers, the emphasis will be on the behavior of **linear, time invariant, causal systems (LTIC)**, especially in the time domain.

The algorithms developed are based mostly on analytically derived formulae and procedures which can be easily implemented as numerical routines in any computer language. Some of these formulae have already been presented and discussed in previous parts, but are repeated here in order to put the computer algorithms into an adequate perspective. The algorithms have been written for a special maths oriented programming environment, named Matlab™ [Ref. 6.1, 6.2, 6.3]. Matlab has a large set of pre-programmed mathematical, logical, and graphical functions and allows the user to create macro functions (saved as the '*.M' files) of his own. Over the years Matlab has grown in popularity and has become one of the five most favorite problem solving environments, setting standards in data analysis and control software. Also, its syntax, similar to that of the 'C' programming language, allows easy implementation in embedded systems, so we feel its use here is justified. For the sake of readability we have tried to write the routines and the examples with mainly the basic Matlab instructions, adding as much comment as possible.

6.1 LTIC System Description — A Short Overview

Assume $x(t)$ to be the signal presented at the input of a **linear, time invariant, causal system** (LTIC) which has n dominant energy storing (reactive) components (or briefly a n^{th}-order system). The system output in the time domain, $y(t)$, may be expressed by a **linear differential equation** with constant coefficients:

$$\sum_{i=0}^{n} b_i\, y^{(i)}(t) = \sum_{j=0}^{m} a_j\, x^{(j)}(t) \tag{6.1.1}$$

where the coefficients $\boldsymbol{a_j}$ **and** $\boldsymbol{b_i}$ **are derived from the system's time constants**, whilst $y^{(i)}$ and $x^{(j)}$ are the i^{th} and j^{th} derivatives of the output and input signal, as required by the system's order. From the theory of differential equations we know that the solution of Eq. 6.1.1, given the initial conditions $y(0)$, $y'(0)$, $y''(0)$, …, $y^{(n-1)}(0)$ is of the form:

$$y(t) = \; y_{\text{h}}(t) + y_{\text{f}}(t) \tag{6.1.2}$$

Here $y_{\text{h}}(t)$ is the solution of the **homogeneous** differential equation (in which $x(t)$ and all its derivatives are zero), whilst $y_{\text{f}}(t)$ is the **particular** solution for $x(t)$, which means that $y_{\text{f}}(t) = y_{\text{f}}\{x(t)\}$. From circuit theory we know that $\boldsymbol{y_{\text{h}}(t)}$ **represents the natural (also free, impulse, transient, or relaxation from the initially energized state) response** and $\boldsymbol{y_{\text{f}}(t)}$ **represents the forced (also particular, final, steady state) response.**

Knowing that $y_{\text{f}}(t)$ is a description of the output signal in time very distant from the initial disturbance, when the system has regained a new state of (static or dynamic) balance, we can define the **system transfer function** $\boldsymbol{F(s)}$ from $y_{\text{f}}(t)$ and $x(t)$:

$$F(s) = \frac{y_{\text{f}}(t)}{x(t)} \qquad \text{where} \qquad x(t) = \mathrm{e}^{st} \tag{6.1.3}$$

Such an input signal has been chosen merely because it still retains its exponential form when differentiated, i.e.:

$$x^{(n)}(t) = s^n \mathrm{e}^{st} \tag{6.1.4}$$

For the same reason $y_{\text{f}}(t)$ is expected to be of the form:

$$y_{\text{f}}(t) = A\, \mathrm{e}^{st} \tag{6.1.5}$$

Eq. 6.1.1 may now be rewritten as:

$$(b_n s^n + b_{n-1} s^{n-1} + \cdots + b_1 s + b_0) A \mathrm{e}^{st} = (a_m s^m + a_{m-1} s^{m-1} + \cdots + a_1 s + a_0) \mathrm{e}^{st} \tag{6.1.6}$$

Then A must be:

$$A = \frac{a_m s^m + a_{m-1} s^{m-1} + \cdots + a_1 s + a_0}{b_n s^n + b_{n-1} s^{n-1} + \cdots + b_1 s + b_0} \tag{6.1.7}$$

Returning to Eq. 6.1.3 we can now define the system transfer function as a rational function of s:

$$F(s) = \frac{y_f(t)}{x(t)} = \frac{A\,e^{st}}{e^{st}} = \frac{a_m s^m + a_{m-1} s^{m-1} + \cdots + a_1 s + a_0}{b_n s^n + b_{n-1} s^{n-1} + \cdots + b_1 s + b_0} \tag{6.1.8}$$

Instead of $y_f(t)$, it is much easier in practice to find $F(s)$ first, since the coefficients a_i and b_i are derived from the system's time constants. The system's time domain response is then found from $F(s)$. From algebra we know that ***F(s)* can be expressed also as a function of its poles and zeros.** A n^{th}-order polynomial can be expressed as a product of terms containing its roots r_k:

$$P_n(s) = \sum_{i=0}^{n} a_i s^i = \prod_{k=1}^{n} (s - r_k) \tag{6.1.9}$$

The value of this product is zero whenever s assumes a value of a root r_k. Therefore we can rewrite Eq. 6.1.8 as:

$$F(s) = \frac{(s - z_1)(s - z_2)\ldots(s - z_{m-1})(s - z_m)}{(s - p_1)(s - p_2)\ldots(s - p_{n-1})(s - p_n)} \tag{6.1.10}$$

Here the roots of the polynomial in the numerator are the system's zeros, z_j, and the roots of the polynomial in the denominator are the system's poles, p_i.

We shall have this form in mind whenever a system is specified, because we shall always **start the design by specifying some optimum pole–zero pattern as the design goal** and then work towards the required system's time constants.

The system's time domain equivalent of $F(s)$, labeled $f(t)$, is the system's impulse response:

$$y_h(t) = f(t)\Big|_{x(t)=\delta(t)} \tag{6.1.11}$$

where $\delta(t)$ is the *Dirac*'s function (the infinitesimal time limit of the unit area impulse).

The response to an arbitrary input signal may then be found by convolving the input signal with the system's impulse response (for convolution see Part 1, Sec. 1.15; see also the VCON routine in Part 7, Sec. 7.2).

It is owed to *Oliver Heaviside* (1850–1925, [Ref. 6.4–6.8]), who pioneered the transform theory, that we solve differential equations through the use of the **Laplace transform**[1]**.**

The transform is applied to the time variable t through a single time domain integration, producing a new variable s, whose dimension is t^{-1} (frequency). In the frequency domain the n^{th}-order differential equation is reduced to an n^{th}-order polynomial, whilst the convolution is reduced to simple multiplication. Once solved (using simple algebra), the result is transformed back to the time domain.

[1] Apparently, Heaviside developed his 'operational calculus' in the 1890s independently of Laplace. Although useful and giving results in accordance with practice, his method was considered unorthodox and suspicious for quite a while and only in the mid 1930s it was realized that the theoretical basis for his work could be traced back to Laplace. Interestingly, he also develpoed the method of compensating the dominantly capacitive telegraph lines by inductive peaking, amongst many other things.

Let $F(s)$ represent the Laplace transform of $f(t)$. Then:

$$F(s) = \mathcal{L}\{f(t)\} = \int_{-\infty}^{+\infty} f(t)\,\mathrm{e}^{-st}\,dt \tag{6.1.12}$$

where $\mathcal{L}\{\}$ denotes the Laplace operator as defined by the integral (see Part 1, Sec.1.4).

Actually, the integration is usually made from $t = 0$ and not from $-\infty$, in order to preserve the response's causality (i.e., something happens only after closing the switch). This limitation is caused by the term e^{-st} which for $t < 0$ would not integrate to a finite value unless $f(t) = 0$ for $t \leq 0$. Such restriction is readily accomplished if we modulate the input signal by closing a switch at $t = 0$. Mathematically, this can be expressed by multiplying $f(t)$ by $h(t)$ — the Heaviside's unit step function. In our case this is not necessary, since for calculation of the transient response we consider such input signals which satisfy the convergence condition by definition. Also we shall always assume that the system under investigation was powered up for a time long enough to settle down, so we can safely say that all initial conditions are zero (or an additive constant at worst).

Physically, by multiplying the time domain function by e^{-st} in Eq. 6.1.12 we have canceled the rotation of the phasor e^{st} at that particular frequency (s), allowing the function to integrate to some finite value (see Part 1, Sec.1.2). At other frequencies the phasors will continue to rotate, integrating eventually to zero. By doing so for all frequencies we produce the frequency domain equivalent of $f(t)$. This same process is going on in a sweeping filter spectrum analyzer; the only difference is that in our case an infinitely narrow filter bandwidth is considered. Indeed, such bandwidth takes an infinitely long energy build up time, thus the integration must also last infinitely long and be performed in infinitely small steps.

The **inverse transform** process is defined as:

$$f(t) = \mathcal{L}^{-1}\{F(s)\} = \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} F(s)\,\mathrm{e}^{st}\,ds \tag{6.1.13}$$

where σ is an arbitrarily chosen real valued positive constant for which the inversion solution exists (this restriction is required for functions which do not decay to zero in some finite time and therefore the integral would not converge, e.g., the unit step).

Eq. 6.1.1 can now be written as:

$$y(t) = \mathcal{L}^{-1}\Big\{F(s)\cdot\mathcal{L}\{x(t)\}\Big\} \tag{6.1.14}$$

Note that for transient response calculation, $x(t)$ (the time domain input function), is either the Dirac function **(or $\delta(t)$ — the unity area impulse)** or the Heaviside function **(or $h(t)$ — the unity amplitude step).** In these two cases $\mathcal{L}\{x(t)\} = X(s)$ is either 1 (the transform of the unity area impulse), or $1/s$ (the transform of the unity amplitude step), as we have already seen in Part 1.

Eq. 6.1.14 has been used extensively in previous parts to calculate the transient responses analytically. However, for calculation of the frequency response, we are interested only in that part of the transformed function which is a function of a purely

imaginary s and therefore a special case of $F(s)$, that is $F(j\omega)$. It is thus interesting to examine the possibility of calculating the transient response using the **inverse Fourier transform** (a special case of the inverse Laplace transform) of the system frequency response. We have already seen in Part 1 that the only difference between the Laplace and Fourier transforms is that s is replaced by $j\omega$, which is the same as making σ, the real part of s, equal to zero.

Eq. 6.1.11 shows that the time domain equivalent of the system's frequency response $F(s)$ is $f(t)$ resulting from the excitation by $\delta(t)$, or, in words, the system impulse response. Since the impulse response of any system (except the conditionally stable systems, as well as the oscillating or regenerating systems) decays to zero after some finite time, we do not have to make those special precautions (as in the inverse Laplace transform) to allow the integral to converge, but instead we can use the inverse Fourier transform of $F(j\omega)$ to calculate the impulse response:

$$f(t) = \mathcal{F}^{-1}\{F(j\omega)\} \tag{6.1.15}$$

However, **the Fourier transform of the unity amplitude step does not converge**, so we shall have to use an additional procedure to calculate the step response.

It is possible to put Eq. 6.1.1 into numerical form [Ref. 6.20, 6.21]. Whilst there are ways of using Eq. 6.1.14 in numerical form [Ref. 6.22, 6.23], we shall rather concentrate on Eq. 6.1.15, since the Fast Fourier Transform algorithm (FFT, [Ref. 6.16–6.19]), which we are going to use, offers some very distinct advantages. In addition, we shall develop an algorithm based on the residue theory (Part 1, Sec. 1.9); the details are given in Sec. 6.6.

Another point to consider, known from modern filter theory, is that optimized high order systems are difficult to realize in direct form, because the ratio of the smallest to the largest time constant quickly falls below component tolerances as the system's order is increased. *Butterworth* [Ref. 6.11] has shown that optimum system performance is more easily met by a cascade of low order systems (several of second and only one third, if n is odd) separated by amplifiers. As a bonus such structures will satisfy the gain–bandwidth product requirement more easily. So in practice we shall rarely need to solve high order system equations, usually only at the system integration level.

The formulae presented above will be used as the starting point in algorithm development. We shall develop the algorithms for calculating the system poles for a desired system order, the complex frequency response, the magnitude and phase response, the group (envelope) time delay, the impulse response, the step response, and the numerical convolution. Those algorithms can, of course, be written to solve only our particular class of problems. It is wise, however, to write them to be as universally applicable as possible, in spite of losing some algorithm efficiency, to suit eventual future needs.

6.2 Algorithm Syntax And Terminology

Readers who have not used Matlab or other similar software before will probably have some difficulties in understanding the algorithm syntax and the operations implied. Here is some of the syntax and terminology used throughout Part 6:

`matrix` — An array of data, organized in m rows and n columns.
The operations involving matrices follow the standard matrix calculation rules.

`vector` — A single row or single column matrix; either m or n is equal to 1.

`scalar` — A single element matrix; both m and n are equal to 1.

`size` `length` — Matrix dimension; if `M` is an m by n matrix then `[m,n]=size(M)` returns the number of rows in `m` and the number of columns in `n`. Likewise, `max(size(M))` returns m or n, whichever is greater. For vectors `max(size(V))` is the `length`, or the total number of elements in `V`.

`submatrix` — A smaller matrix contained inside a larger one.
`A=V(k)` is the k^{th} element of the vector `V` (`k` is the index).
`A=V(k)` is the same as `A=V(round(k))` if `k` is non-integer.
`B=M(:,k)` is the k^{th} column of M.
`C=M(j:k,h:i)` is the matrix of h^{th} to i^{th} elements from the j^{th} to k^{th} row.

`+` `-` `*` `/` — Operations involving matrices of 'compatible' dimensions. A scalar can be added to, subtracted from, can multiply or divide a matrix of any dimension. Two matrices can be added or subtracted if they have the same dimensions. Multiplication `A*B` requires that the number of rows in `B` is equal to the number of columns in `A`. Division `A/B` requires an equal number of columns in `A` and `B`.

`.*` `./` — The dot before the operation specifies element by element multiplication and division. The matrix containing the inverse values of the elements in matrix `A` can be calculated as: `1 ./A` (note the **space** between 1 and the dot).

`^` `.^` — Powers: `A.^n` is a matrix with each element of `A` raised to the power of `n`;
`n.^A` is a matrix of `n` to the power of each element of `A`;
`A^n` is possible if `A` is a square matrix (equal number of rows and columns);
`exp(A)` is `e.^A`, where `e=2.71828...`;
`A^B`, if both `A` and `B` are matrices, is an error.

`:` — Colon. Indicates range.
`(1:5)` is a vector `[1,2,3,4,5]`; `V(1:2:N-1)` denotes all odd elements of V.
`A(:)` is all elements of `A` in a single column.

`=` — Equality (right to left assignment of the function or operation result); examples:
`[output arguments]=function(input arguments);` or:
`[out]=[in1]operation[in2];` i.e.: `y=sin(w*t+phi); c=a*b/(a+b);`

`==` — Identically equal (relation); `if x==0, do something; end`

`>` `>=` `<` `<=` — greater, greater or equal, smaller, smaller or equal;

`&` `|` `~` `~=` — logical operators: 'and', 'or', 'not', 'not equal'.

`;` — Semicolon, logical end of command line. For matrices it indicates the end of a row.

`2+3j` — Complex numbers: `2+3j` or `2+j*3` or `2+3*sqrt(-1);`
Most Matlab operations can deal with matrices containing complex elements.

`%` — Characters following `%` are ignored by Matlab. Used for comments.

6.3 Poles and Zeros

It is beyond the scope of this text to cover all the background of system optimization theory. Let us just mention the most important optimization criteria for each major system family. For the same bandwidth and system order n:

- the Butterworth family

  is optimal in the sense of having a maximally flat pass band magnitude;

- the Bessel–Thomson family

  is optimal in the sense of having a maximally flat group (envelope) delay and, consequently, a maximally steep step response with minimum overshoot;

- the Chebyshev family

  is optimal in the sense of a having maximally steep pass band to stop band transition, at the expense of some specified pass band ripple;

- the Inverse Chebyshev family

  is optimal in the sense of having a maximally steep stop band to pass band transition, at the expense of some specified stop band ripple;

- the Elliptic (Cauer) family

  is optimal in the sense of having a maximally steep pass band to stop band transition, at the expense of some specified pass band and stop band ripple.

It must be pointed out, however, that some system families, the Bessel–Thomson family in particular, can be realized in practice more easily than others, owing to the lower ratio of the largest to the smallest system time constant for a given system order, the maximum usable ratio being limited by component tolerances. Also, low order systems can be realized more easily than high order systems. We shall have to keep these things in mind when denormalizing the system to the actual upper frequency limit and deciding the number of stages used to achieve the total amplification factor.

The trouble is that during the design process we select the system poles and zeros in accordance with certain circuit simplifications, useful for speeding up the analysis. But the implementation of the poles and zeros in an actual amplifier is more a matter of practical know–how, instead of a rigorous theory. This is particularly true if we are pushing the performance to the limits of realizability, since in these conditions the component stray reactances must be taken into account when specifying the system time constants. Luckily for the amplifier designer, for the same component and layout strays the Bessel–Thomson system yields the highest system bandwidth, with the bonus of an optimal transient response. This is also true for 'feedback stabilized' systems, since the large phase margin offered by this system family aids system stability; also, feedback induced Q-factor enhancement at high frequencies lowers the required imaginary versus real part ratio of the response shaping component impedances even further.

In this text we shall only deal with the Butterworth [Ref. 6.11] and Bessel–Thomson [Ref. 6.12, 6.13] low pass systems for calculation of poles, since these are

required in wideband amplifier design. If needed, Chebyshev, inverse Chebyshev and elliptic (Cauer) functions are provided in the Matlab Signal Processing Toolbox, as well as the low pass to high pass, band pass and band stop transform algorithms. The toolbox also contains many other useful algorithms, such as RESIDUE, ROOTS, etc., which will not be considered here (see [Ref. 6.1, 6.2, 6.19]).

In order to be able to compare the performance of different systems on a fair basis we must specify some form of **system standardization:**

a) **all systems will have the pole values normalized for an upper half power angular frequency ω_h of 1 radian per second (equivalent to the cycle frequency of $f_h = 1/2\pi$ [Hz]). This leads to the use of a normalized frequency vector, implying that whenever we write either f or ω, we shall actually mean f/f_h or ω/ω_h, respectively.**

Please note that this can sometimes cause a bit of confusion, since f/f_h is the same as ω/ω_h; but $\omega = 2\pi f$, so we should keep an eye on the factor 2π, especially when denormalizing the poles to the actual system upper half power frequency.

The frequency response is calculated as a function of ω, since the poles and zeros are mapped in terms of $s = \sigma + j\omega$, where both the real and imaginary part are measured in [rad/s], but we usually plot it as a function of f (in [Hz]). If the values of the poles are normalized we can use the same normalized frequency vector to calculate and plot the frequency domain functions. Therefore to plot the magnitude and phase responses vs. frequency we shall not have to divide the frequency vector (of a length usually between 100 and 1000 elements) by 2π.

Since Matlab will not accept the symbol ω as a valid name for a variable we shall replace it by `w=2*pi*f` in our routines.

b) **all systems will have their DC gain (at $\omega = 0$) normalized to $A_0 = 1$** (throughout this text we shall consider low pass systems only). Nevertheless, we shall try to provide the correct gain treatment in the general case, in order to broaden the applicability of our algorithms.

Of course, to extract the actual system component values, as well as to scale the various frequency and time domain responses to comply with the desired upper frequency f_h, the poles and zeros will have to be **denormalized** (multiplied) by $2\pi f_h$. Also, each response will have to be scaled by the required gain factor.

I.e., for a simple current driven shunt RC system, the normalized pole, $s_{1n} = 1/(R_{1n}C_{1n}) = 1$, is first denormalized to the value of the desired bandwidth, $s_1 = s_{1n} \cdot 2\pi f_h$. From s_1 we get the new component values, $R_1C_1 = 1/(2\pi f_h)$. Finally, we multiply R_1 by the gain factor, $R = AR_1$, to obtain the desired output voltage from the available input current, that is $v_o = Ri_i$, and then reduce C_1 by the same amount, $C = C_1/A$. If C is a stray capacitance it cannot be reduced below the limit imposed by the circuit topology. Then we must work backwards by first finding the R which would give the desired bandwidth, and then determine the input current which will give the required output voltage.

6.3.1 Butterworth Systems

Butterworth systems ([Ref. 6.11], see also Part 4, Sec. 4.3) are optimal in the sense that all the derivatives of the frequency response are zero at the complex plane origin, resulting in a maximally flat magnitude response. The normalized squared magnitude response of an n^{th}-order Butterworth system is:

$$F^2(\omega) = \frac{1}{1 + (\omega^2)^n} \tag{6.3.1}$$

This can be rewritten as:

$$F(s)\,F(-s) = \frac{1}{1 + (-s^2)^n} \tag{6.3.2}$$

This is an all-pole system, since $F^2(\omega) \to \infty$ whenever:

$$1 + (\omega^2)^n = 0 \tag{6.3.3}$$

The roots of Eq. 6.3.3 are:

$$\omega = (-1)^{1/2n} \tag{6.3.4}$$

This can be solved using DeMoivre form:

$$(-1)^{1/2n} = \cos\frac{\pi + k\,2\pi}{2n} + j\sin\frac{\pi + k\,2\pi}{2n} \tag{6.3.5}$$

where $k = 0, 1, 2, \cdots, n-1$.

If, owing to the ***Hurwitz* stability requirement**, we associate the poles in the left half of the complex plane with $F(s)$, then:

$$F(s) = \frac{k_0}{\prod_{k=1}^{n}(s - s_k)} \tag{6.3.6}$$

where s_k are found from the expression of Eq. 6.3.5 in the exponential form:

$$s_k = \mathrm{e}^{j\pi\frac{1+\frac{2k-1}{n}}{2}} \qquad \text{for } k = 1, 2, 3, \ldots, n \tag{6.3.7}$$

and:

$$k_0 = \prod_{k=1}^{n}(-s_k) \tag{6.3.8}$$

In the general (non-normalized) case, $\omega_{\text{h}} = \sqrt[n]{k_0}$.

A Butterworth system is completely specified by the system order n. It is normalized so that it has its half power bandwidth limit at the unit frequency:

$$F(j)\,F(-j) = F^2(1) = \frac{1}{2} \quad \Rightarrow \quad \left|F(1)\right| = \frac{1}{\sqrt{2}} \tag{6.3.9}$$

Eq.6.3.7 and Eq.6.3.8 are implemented in the Matlab Signal Processing Toolbox function called BUTTAP (an acronym for BUTTerworth Analog Prototype):

```
function [z,p,k] = buttap(n)
%BUTTAP     Butterworth analog low pass filter prototype.
%      [z,p,k] = buttap(n) returns the zeros, poles, and gain
%      for the n-th order normalized prototype Butterworth analog
%      low pass filter. The resulting filter has n poles on the
%      unit circle in the left half plane, and no zeros.
%
%      See also BUTTER, CHEB1AP, and CHEB2AP.

%      J.N. Little and J.O. Smith 1-14-87
%      Revised 1-13-88 LS
%      (c) Copyright 1987, 1988,  by The MathWorks, Inc.

% Poles are on the unit circle in the left-half plane.
z = [];
p = exp(sqrt(-1)*(pi*(1:2:2*n-1)/(2*n) + pi/2)).';
k = real(prod(-p));
```

As an example see the complex plane layout of the poles of a 5th-order Butterworth system in Fig. 6.3.1.

For a desired attenuation $a = 1/A$ at some chosen $\omega_a > \omega_h$ we can calculate the required system order:

$$n = \frac{\log_{10}(A^2 - 1)}{2 \log_{10} \frac{\omega_a}{\omega_h}} \tag{6.3.10}$$

and round it to the first higher integer.

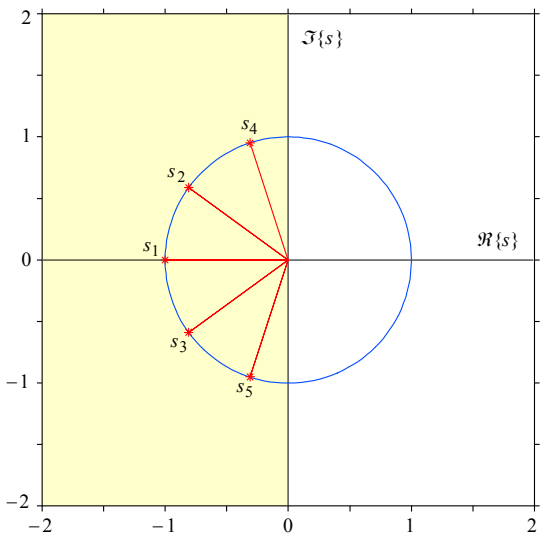


Fig. 6.3.1: The 5th-order Butterworth system poles in Cartesian coordinates of the complex plane. The left half of the figure is the s domain over which the magnitude in Fig. 6.4.1 is plotted.

6.3.2 Bessel–Thomson Systems

Bessel–Thomson systems (see [Ref. 6.12, 6.13]; see also Part 4, Sec. 4.4) are optimal in the sense that all the derivatives of the group (envelope) time delay response are zero at origin, which results in a maximally flat group delay. This means that all the relevant frequencies pass through the system with equal time delay, resulting in a transient response with a minimal overshoot. In the complex frequency plane a system with pure time delay is represented by:

$$F(s) = \mathrm{e}^{-sT} \tag{6.3.11}$$

We first normalize this by making $T = 1$. Then we expand e^{-s} as a polynomial. However, if this is done using the *Taylor* series expression for e^{x} (and if the polynomial degree exceeds 4), the resulting polynomial would not meet the ***Hurwitz* stability criterion,** because some of the poles would be in the right half of the complex plane. But there is another expression for e^{-s} which we can use:

$$\mathrm{e}^{-s} = \frac{1}{\sinh s + \cosh s} = \frac{\dfrac{1}{\sinh s}}{1 + \dfrac{\cosh s}{\sinh s}} \tag{6.3.12}$$

The Taylor series for hyperbolic sine function has even powers of s and the hyperbolic cosine has odd powers of s. When we divide these polynomials (using long division) the poles of the resulting polynomial meet the stability criterion. If we express this as a partial fraction expansion, truncated at the n^{th} fraction, an n^{th}-order Bessel–Thomson system results. This can be expressed as:

$$F(s) = \frac{c_0}{B_n(s)} \tag{6.3.13}$$

where $B_n(s)$ is an n^{th}-order Bessel polynomial:

$$B_n(s) = \sum_{k=0}^{n} c_k \, s^k \tag{6.3.14}$$

and each $B_n(s)$ satisfies one of the following relations:

$$\begin{aligned} B_0(s) &= 1 \\ B_1(s) &= s + 1 \\ B_n(s) &= (2n-1)\, B_{n-1}(s) + s^2\, B_{n-2}(s) \end{aligned} \tag{6.3.15}$$

The coefficients c_k of the resulting polynomial can be calculated as:

$$c_k = \frac{(2n-k)!}{2^{(n-k)}\; k!\; (n-k)!} \qquad \forall\ k = 0, 1, 2, \ldots, n-1, n \tag{6.3.16}$$

The function, which will calculate the Bessel polynomial coefficients using Eq. 6.3.16 will be called BESTAP (this stands for BESsel–Thomson Analog Prototype, but the name is also in good agreement with the best time domain response of this system family). Within this function the system poles are extracted using the ROOTS

function in Matlab. This works well up to $n = 24$; for higher orders the ratio of c_n to c_0 is so high that the computer numerical resolution ('double precision' or 16 significant digits) is exceeded, but this is not a severe limitation because in most circuit configurations the 1% component tolerances will limit system realizability to about $n = 13$ (assuming a 6-stage system, for which the highest reactive component value ratio is about 12:1). But if needed, we can always calculate the frequency response from the polynomial expression, using the coefficients c_k directly in Eq. 6.1.8, instead of using Eq. 6.1.10, as in the Matlab POLYVAL and FREQS routines.

Bessel–Thomson system poles are found in the left half of the complex plane on a family of ellipses, having a nearer focus at the complex plane origin and the other focus on the positive part of the real axis (see Fig. 6.3.2**)**.

The poles calculated in this way define a family of systems with equal envelope delay (normalized to 1 s). This results in a progressively larger bandwidth and smaller rise time for each higher n (see Fig. 6.5.11). In addition, two other normalizations of the Bessel–Thomson system are possible.

One is to make the asymptote of the magnitude roll off slope the same as it is for the Butterworth system of equal order (this is useful for calculating transitional Bessel to Butterworth systems, as we have seen in Part 4, Sec. 4.5.3). If ω_a is to become the half power cut off frequency of the new system:

$$\left|F(\omega_a)\right| = \frac{c_0}{2\,\omega_a^n} = \frac{1}{2} \qquad \Rightarrow \qquad \omega_a = c_0^{1/n} \tag{6.3.17}$$

In this case, with the roots of $B_n(s)$ divided by $c_0^{1/n}$, the envelope delay will be equal to $c_0^{1/n}$, instead of 1, and the system bandwidth will be smaller for each higher n.

The other is to have equal bandwidth for any n, possibly normalized to 1 rad/s, as is the Butterworth family; in this way we would be able to compare different systems on a fair basis. Unfortunately there is no simple way of matching the Bessel–Thomson system bandwidth to that of a Butterworth system of the same order. To achieve this we have to recursively multiply the poles by a correction factor proportional to the bandwidth ratio, until a satisfying approximation is reached (the values of poles modified in such a way for systems of order 2 to 10 are shown in Part 4, Table 4.5.1). The `while` loop at the end of the BESTAP routine has a tolerance of 0.0001 and it was experimentally found to match in only 8 to 12 loop iterations, depending on n; this tolerance is satisfactory for most practical purposes, but the reader can easily change it to suit his needs.

All these three normalization options (group delay, asymptote, and bandwidth) are being provided for by the BESTAP routine by entering, besides the system order n, an additional input argument in the form of a single character string:

`'n'` for 1 rad/s cutoff frequency normalization (the default),

`'t'` for unit time delay and

`'a'` for the same attenuation slope asymptote as Butterworth system of equal order.

As in the BUTTAP routine, three output variables are returned. But the number of arguments returned by BESTAP can be either 3, 2, or just 1. If all three output arguments are requested, the zeros are returned in `z`, the poles in `p`, and the non-

normalized system gain is returned in the output variable `k`. Since there are no zeros in this family of systems, an empty matrix is returned in `z`.

With just two output arguments, only `z` and `p` are returned.

When only one output argument is specified, instead of having an empty matrix returned in `z`, which would not be very useful, we have decided to return the Bessel polynomial coefficients c_k. Note that for the n^{th}-order system there are $n+1$ coefficients, from c_n to c_0. The system gain normalization is achieved by dividing the each coefficient by c_0, that is, the last one in the vector `c`, i.e., `c=c/c(n+1)`. The coefficients are scaled as for the `'t'` option (equal envelope delay); other options are then ignored. But, if necessary, we can always calculate the polynomial coefficients for those cases from the poles, by invoking the POLY routine, i.e., `c=poly(p)`.

```
function [z,p,k]=bestap(n,x)
%BESTAP  BESsel-Thomson Analog Prototype.
%        Returns the zeros z, poles p and gain k of the n-th order
%        Bessel-Thomson system. This is an all-pole system, so an
%        empty matrix is returned in z. The poles are calculated for
%        a maximally flat envelope (group) delay.
%
%        Call  : [z,p,k]=bestap(n,x);
%        where :
%                n is the system order
%                x is a single-character string, making the poles:
%                  'n' - normalized to a cutoff of 1 rad/s (default);
%                  'a' - normalized to have the same attenuation
%                        asymptote as a Butterworth system of same n;
%                  't' - scaled for a group-delay of ls.
%                k is the non-normalized system DC gain.
%                p are the poles (length-n column vector)
%                z are the zeros (no zeros, empty matrix returned)
%        With only one output argument :
%        c=bestap(n);
%        the n+1 coefficients of the system polynomial are returned,
%        scaled as in the 't' option, ignoring other options.

% Author :  Erik Margan, 881012,    Free of copyright !

if nargin == 1          % nargin is the number of input arguments
      x='n';            % by default, normalize to 1 rad/s cutoff
end
z=[];                   % no zeros
if n == 1
      if nargout == 1
            c=[1, 1];   % first-order system coefficients
      else
            p=-1;       % first-order pole
            k=1;        % gain
            return      % end execution of this routine
      end
else
                        % find the Bessel polynomial coefficients
                        % from factorials :
                        % 0!=1 by definition, the rest is calculated
                        % by CUMPROD (CUMulative PRODuct)
      fact=[1, cumprod(1:1:2*n)];
      binp=2 .^(0:1:n));      % a vector of binary powers
      c=fact(n+1:1:2*n+1)./(binp.*fact(1:1:n+1).*fact(n+1:-1:1));
                        % c is a vector of polynomial coefficients,
                        % c(1) is at s^n, c(n+1) is at s^0
```

```
end
if nargout == 1          % nargout is the number of output arguments
      z=c;               % the coefficients of the Bessel polynomial
      return             % end execution of this routine
end

c=c/c(n+1);              % Normalize system gain to 1 at DC

if x == 'a' | x == 'A'         % | means logical OR
                               % Normalize to Butterworth asymptote
      g=c(1) .^((n:-1:0)/n);   % c(1) is the coefficient at s^n
      c=c./g;                  % Normalize gain
end

p=roots(c);                    % ROOTS extracts poles from coefficients

if x == 'n' | x == 'N'
            % Bandwidth normalization to 1 rad/s results in
            % progressively greater envelope delay for increasing n
      P=p;  % copy the poles to P
            % Reference (-3 dB point)
      y3=1/sqrt(2);
      y=abs(freqw(P,1));       % attenuation at 1 rad/s (see FREQW)
      while abs( 1 - y3/y ) > 0.0001
            P=P*(y3/y);        % Make iterative corrections
            y=abs(freqw(P,1));
      end
      p=P;  % copy P back to p
end

k=real(prod(-p));              % non-normalized system gain
```

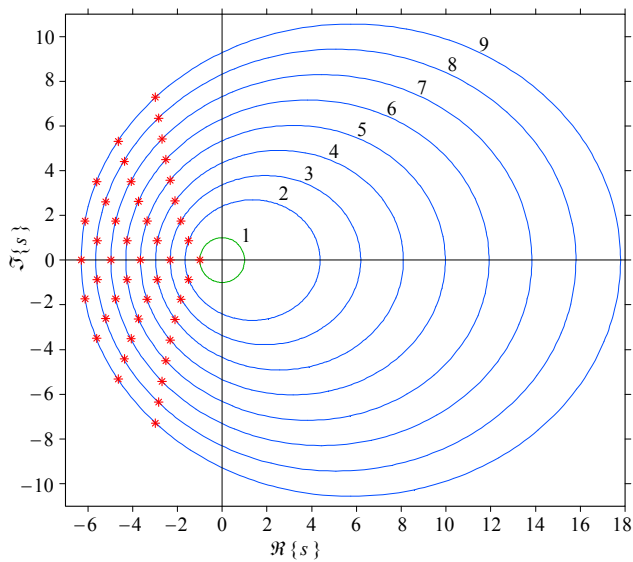


Fig. 6.3.2: Complex plane pole map for unit group delay Bessel–Thomson systems of order 2–9 (with the first-order reference).

6.4 Complex Frequency Response

In Matlab the frequency response is calculated by two routines, called POLY and FREQS, which require the polynomial coefficients as the input argument. We shall use two different routines: PATS, which calculates the product of terms containing polynomial roots at each value of s, according to Eq. 6.1.9 and FREQW, which returns the complex frequency response $F(s)$, after Eq. 6.1.10, given the zeros, poles and the normalized frequency vector. FREQW calls PATS as a subroutine.

```
function P=pats(R,s)
% PATS   Polynomial (product form) value AT S.
%        P=pats(R,s) returns the values of the product of terms,
%        containing n-th order polynomial roots R=[r1, r2,.., rn],
%        for each element of s.
%        Values are calculated according to the formula :
%
%     P(s)=(s-r1)*(s-r2)*...*(s-rn) / (((-1)^n)*(r1*r2*...*rn))
%
%        and are normalized so that P(0)=1.
%     PATS is used by FREQW. See also POLY and FREQS.

% Author :  Erik Margan, 881110, Free of copyright !

[m,n]=size(s);
P=ones(m,n);            % A matrix of all ones, same dimension as s
nr=max(size(R));        % number of elements in R
for k=1:nr
      if R(k) == 0
            P=P.*s;     % Multiply, but prevent from dividing by 0.
      else
            P=P.*(s-R(k))/(-R(k));
      end
end
```

```
function F=freqw(z,p,w)
% FREQW  returns the complex frequency response F(jw) of the system
%        described by the zeros (vector z=[z1,z2,...,zm]) and the
%        poles (vector p=[p1,p2,...,pn]).
%        Call : F=freqw(z,p,w);
%        w is the frequency vector; can be real, imaginary or complex.
%        F=freqw(p,w) assumes a system with poles only.
%        FREQW uses PATS. See also FREQS and FREQZ.

% Author :  Erik Margan, 881110, Free of copyright !

if nargin == 2          % nargin returns the number of input arguments
      w=p; p=z; z=[];   % assume a system with poles only
end
for k=1:max(size(p))
      if real(p(k)) >= 0
            disp('WARNING : This is not a Hurwitz-type system!')
      end
end
if ~any(imag(w))
      w=sqrt(-1)*w;     % if w is real, assume it to be imaginary
end
if isempty(z)
      F=1 ./pats( p, w ) ;
else
      F=pats( z, w )./pats( p, w ) ;
end
```

6.4.1 Frequency Dependent Response Magnitude

The absolute value of the complex frequency response is called **magnitude**; it is calculated as a square root of the product of $F(s)$ with its own complex conjugate:

$$|F(s)| = \sqrt{F(s)\,F^*(s)} = \sqrt{\Big(\Re\{F(s)\} + j\Im\{F(s)\}\Big)\Big(\Re\{F(s)\} - j\Im\{F(s)\}\Big)}$$

$$= \sqrt{\Big(\Re\{F(s)\}\Big)^2 + \Big(\Im\{F(s)\}\Big)^2} = M(s) \qquad (6.4.1)$$

Assuming a sinusoidal input signal, the magnitude represents the output to input ratio of the peak signal value at that particular frequency. In practice, when we talk about the system's 'frequency response', we usually mean 'the frequency dependent magnitude', $M(\omega)$. The magnitude contains no phase information.

We can calculate the magnitude by any of the following Matlab basic functions:

```
M=sqrt( (real(F)).^2 + (imag(F)).^2 );   % or :
M=sqrt( F .* conj(F) );                  % or :
M=sqrt( F .* ( F' ) );                   % or :
M=abs(F);                                % abs --> absolute value
```

We shall use the ABS command, not just because it is easy to type in, but because it executes much faster when there is a large amount of data to process.

In order to acquire a better understanding of what we are doing, let us write an example for a 5th-order Butterworth system. In the Matlab command window we write:

```
[z,p]=buttap(5);↵      % Note: a command line is executed by "ENTER"
```

If we now type:

```
z ↵        answer:          []

p ↵        answer:          -0.3090 + 0.9511i
                            -0.3090 - 0.9511i
                            -0.8090 + 0.5878i
                            -0.8090 - 0.5878i
                            -1.0000 + 0.0000i
```

Since there are no zeros an empty matrix (shown by square brackets) is returned in `z`. A 5-element column vector with complex conjugate pole values is returned in `p`. Let us plot these poles in the complex plane using Cartesian coordinates:

```
plot( real(p), imag(p), '*' ), axis([-2,2,-2,2]); ↵
                                % see the result in Fig.6.3.1.
```

and the result would look as in Fig. 6.3.1 (for clarity, the distance from the origin and the unit circle are also shown there, both needing extra 'plot' operations, not written in the example above). From now on we shall not write the ENTER character explicitly.

The system magnitude as a function of the complex frequency s has a very interesting 3D shape and it is instructive to have a closer look at it:

```
[z,p]=buttap(5);                % 5th-order Butterworth poles
r=(-2:1/20:0);                  % real frequency vector
w=(-2:1/20:2);                  % imaginary frequency vector
[x,y]=meshgrid(r,w);            % make the complex domain grid
M=abs(1 ./pats(p,x+j*y));       % magnitude due to poles in x+j*y domain
for m=1:max(size(M))
     n=find(M(m,:)>12 );        % find magnitude > 12
     M(m,n)=12;                 % limit magnitude to 12 for plot
end
waterfall(x',y',M')             % waterfall plot of magnitude in 3-D
                                % prime(') aligns plots along jw-axis
axis([-2,0,-2,2,0,12])          % set axes limits
view(50,25)                     % view(azimuth,elevation) set view angle
                                % add axes labels (Matlab-V format):
xlabel( '{\it\sigma} = \Re\{{\its}\}', 'FontSize', 10 )
ylabel( '{\itj\omega} = {\itj}\Im\{{\its}\}', 'FontSize', 10 )
zlabel( '{\itM}({\its}) = |{\itF}({\its})|', 'FontSize', 10 )
                                % see the result in Fig.6.4.1.
```

Fig. 6.4.1 has been created using the Matlab WATERFALL function and shows the 3D magnitude of the 5th-order Butterworth system over a limited s domain in the complex plane. The s domain here is the same as the left half of Fig. 6.3.1. Over the poles, the magnitude would extend to infinity, so we have had to limit the height of the plot in order to show the low level features in more detail.

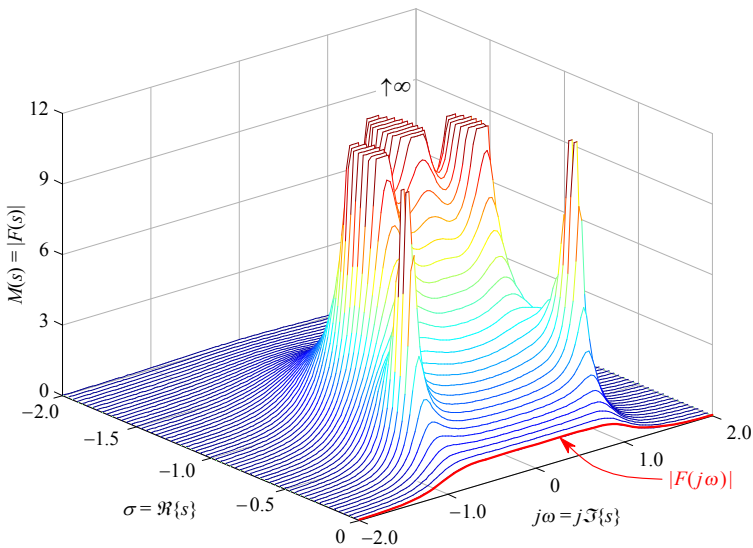


Fig. 6.4.1: The 5th-order Butterworth system magnitude, plotted over the same s-domain as the shaded left half of Fig. 6.3.1. The surface represents $|F(s)|$, but limited in height in order to reveal the low level details. Its shape above the $j\omega$ axis is $|F(j\omega)| = M(\omega)$.

Now, we have intentionally limited the s domain to just the left half of the complex plane (where the real part is either zero or negative). This highlights the shape of the plot along the imaginary axis, which is — guess what ? — $M(\omega)$.

Looking at those lines parallel to the imaginary axis we can see what would happen if the poles were moved closer to that axis: the magnitude would exhibit a progressively pronounced peak. Such is the consequence of lowering the real part of the poles. Since the negative real part is associated with energy dissipative (resistive) components, it is clear that its role is to suppress resonance. But when we design an oscillator we need to compensate any energy lost in the parasitic resistances of the reactive components by an active regeneration ('negative resistance' or a positive real part) in order to set the system poles (usually just one pair for oscillators) exactly on the imaginary axis.

What is interesting to note is the mirror like symmetry about the real axis, owed to the complex conjugate nature of the Laplace space. Here we see at work the concept of 'negative frequency', which will be discussed latter in Sec. 6.5, dealing with the Fourier transform inversion. This symmetry property will allow us to greatly improve the inverse transform algorithm efficiency.

It is also instructive to see the complex frequency response $F(j\omega)$ in 3D:

```
w=(-3:0.01:3);          % 601 frequencies, -3 to 3, in 0.01 increment
F=freqw(z,p,w);         % 601 points of complex frequency response
plot3(w,real(F),imag(F))% 3D plot of the Im and Re part of F(jw)
view(65,15);            % view angle, azimuth 65deg., elevation 15deg.
                        % see the result in Fig.6.4.2.
```

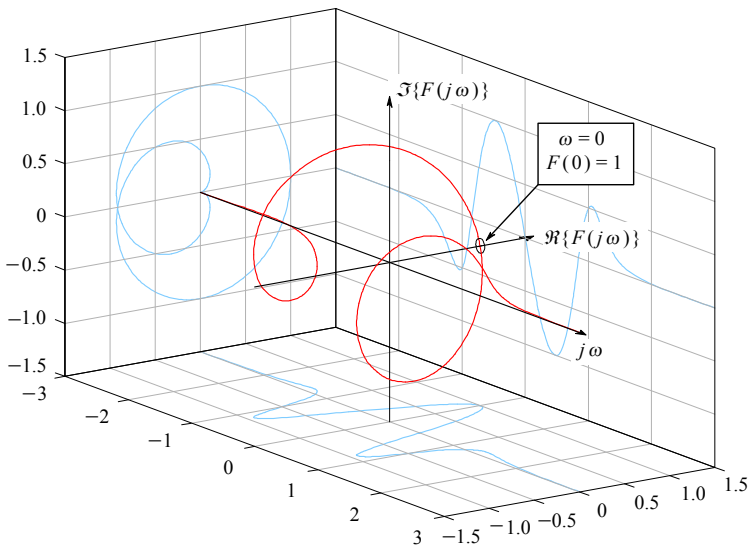


Fig. 6.4.2: The complex 3D plot of $F(j\omega)$. The response phasor rotates clockwise, going from negative to positive frequency. The distance from the frequency axis is the magnitude. The circle on the real axis marks the DC response point. The Nyquist plot (see Fig. 6.4.3) usually shows only the $\omega > 0$ part, viewing in the $-j\omega$ direction. The three projections are plotted to help those readers who do not have access to Matlab to visualize the shape.

Fig. 6.4.2, which has been created using the Matlab PLOT3 function, shows $F(j\omega)$ with the phase angle twisting about the $j\omega$ axis and the magnitude as the distance from the $j\omega$ axis. The circle marker denotes the point where $F(j\omega)$ crosses the real axis at zero frequency — the DC system gain normalized to 1.

Whilst the Fig. 6.4.1 waterfall plot shape was relatively easy to interpret and 'feel', the 3D curve shape is somewhat less clear. In Matlab one can use the `view(azimuth,elevation)` command to see the graph from different viewing angles. In Fig. 6.4.2, `view(65,15)` was used. In more recent versions of Matlab the user can even select the viewing point by the 'mouse'. To help the imagination of readers without access to Matlab we have also plotted the three 'shadows'.

Regarding the symmetry, $F(j\omega)$ is not a mirror image of $F(-j\omega)$ — unlike $M(\omega)$ — because the phasor preserves its sense of rotation (clockwise, negative by definition for any system with poles on the left) throughout the $j\omega$ axis. But if folded about the real axis the shape would match.

As a result of such symmetry $F(j\omega)$ can be plotted using only the $\omega \geq 0$ part of the axis, without any loss of information. The *Nyquist* plot [Ref. 6.9] shows both the magnitude and the phase angle on the same graph:

```
w=(0:0.01:3);           % 301 frequencies, 0 to 3, increment 0.01
F=freqw(z,p,w);         % 301 points of complex frequency response
axis('square');         % the plot axes will have a 1:1 aspect ratio
plot(real(F),imag(F))   % plot the imaginary part versus real part
```

The result should look like Fig. 6.4.3. The view is as if we look at the Fig. 6.4.2 in the opposite direction of the $j\omega$ axis (from $+\infty$ towards the origin).

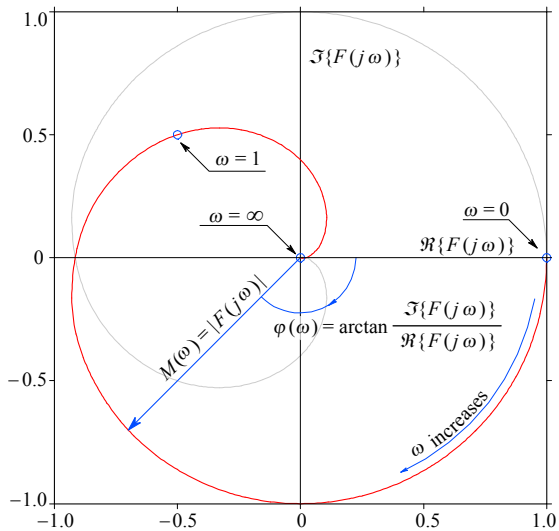


Fig. 6.4.3: The Nyquist plot of the 5th-order Butterworth system frequency response. The frequency axis is reduced to a single point projection at the origin and is parametrically incremented with the phase angle, from the DC point on the real axis, to the half power bandwidth point at `[-0.5, 0.5*j]` and to infinity at the origin.

However, in a Nyquist plot it is difficult to see the frequency dependence of the magnitude and phase, unless we intentionally mark a few chosen frequencies on the plot, as was done in Fig. 6.4.3 in order to make the orientation easier. Thus it has become a standard practice to make separate plots of magnitude and phase as functions of frequency, as introduced by *Bode* [Ref. 6.10]. Again, exploiting the symmetry, we can plot only the $\omega \geq 0$ part without loosing any information:

```
                        % following the previous example:
M=abs(F);               % 301 points of magnitude
plot(w,M)               % display M versus w, see Fig.6.4.4.
```

This should look like Fig. 6.4.4. The special point on the graph is the magnitude at the unit frequency — its value is $1/\sqrt{2}$, or 0.707, and since power is proportional to the magnitude squared this is the system's half power cut off frequency.

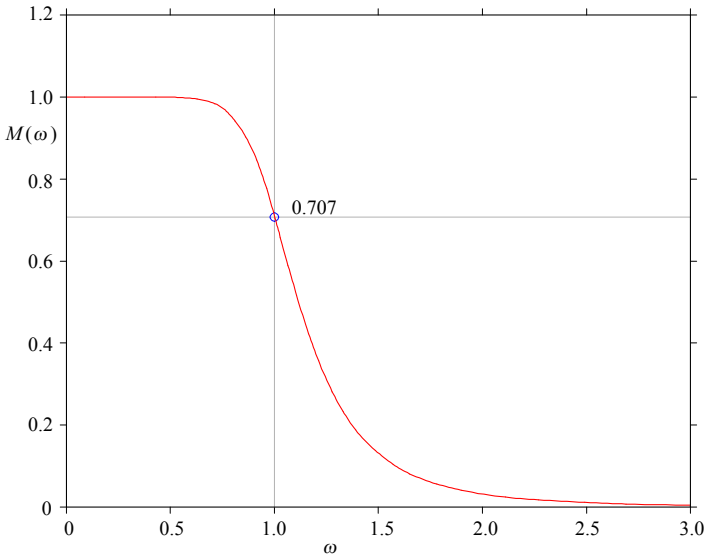


Fig. 6.4.4: The magnitude vs. frequency plot in a linear scale. The characteristic point is the half power bandwidth.

It has also become a standard practice to enhance the stop band detail by using either the $\log M$ vs. $\log \omega$ or the semilog dB(M) vs. $\log \omega$ plot scale (Fig. 6.4.5):

```
w=logspace(-1,1,301);   % frequency, 301 points, equally spaced in
                        % log-scale from 0.1 to 10
F=freqw(z,p,w);         % 301 points of complex frequency response
M=abs(F);               % 301 points of magnitude
semilogx(w,20*log10(M))% display M in dB versus log-scaled w
                        % see the result in Fig.6.4.5.
```

By using the log-log scale or a linear dB vs. log frequency scale we can quickly estimate the system order, since the slope is simply (for all pole systems) n times the first-order system slope (n×20 dB per frequency decade).

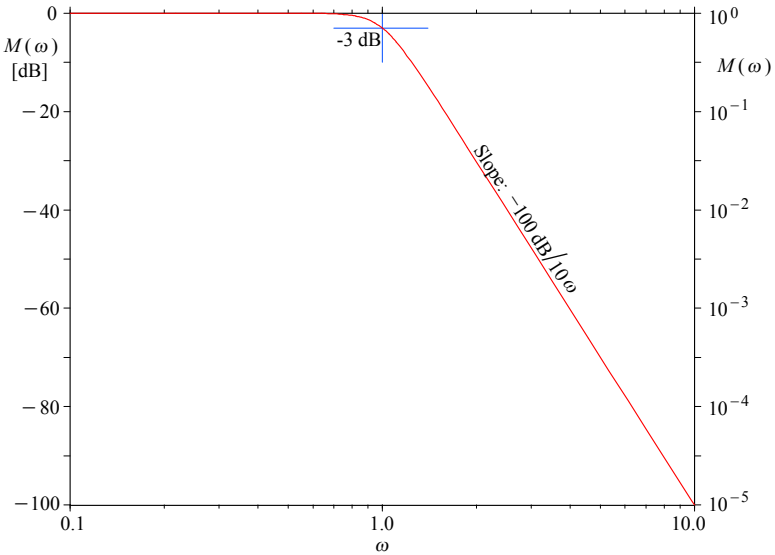


Fig. 6.4.5: Bode plot of the 5th-order Butterworth system magnitude, as in Fig. 6.4.4, but in a linear dB vs. log frequency scale. In such a scale, all-pole systems have an asymptotically linear attenuation slope, proportional to the system order (a factor of 10^n or n×20 dB/10ω). The marked −3dB reference point is the same half power cut off frequency point as in Fig. 6.4.4.

6.4.2 Frequency Dependent Phase Shift

The **phase** response is calculated from the complex frequency response as the arctangent of the ratio of the imaginary versus real part of $F(s)$:

$$\varphi(\omega) = \arctan \frac{\Im\{F(s)\}}{\Re\{F(s)\}} \tag{6.4.2}$$

Using the previously calculated $F(j\omega)$ we can write this as:

```
phi=atan(imag(F)./real(F));          % 301 samples of phase of F(jw)
```

Note that the Matlab arctangent function is called `atan`. However, Matlab also has a built in command named ANGLE, using the same Eq. 6.4.2, so:

```
phi=angle(F);                 % phase response, modulo 2pi ;
semilogx(w,phi);              % show phi in radians vs. log-scaled w ;
```

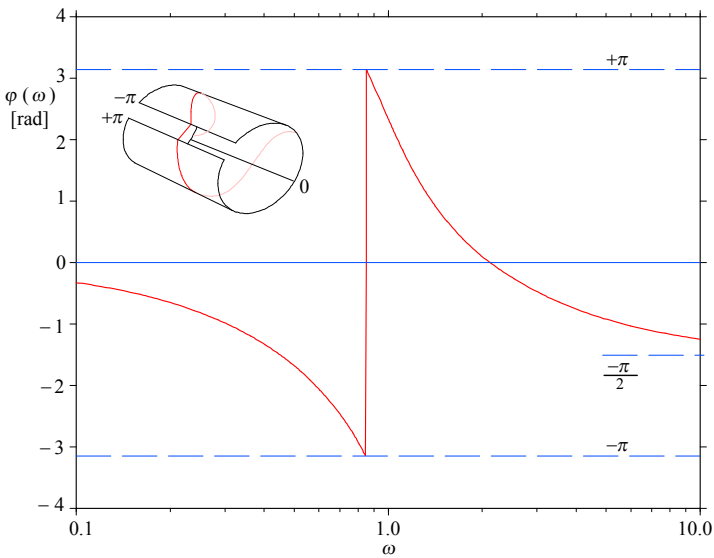


Fig. 6.4.6: The phase angle vs. frequency plot of the 5th-order Butterworth system. The circularity of trigonometric functions, defined within the range $\pm\pi$ radians, is the cause of the discontinuous phase vs. frequency relationship.

Clearly this is a circular function of modulo 2π radians. For systems of 3rd or greater order the phase will rotate by more than 2π, so there will be jumps from $-\pi$ to $+\pi$ in the phase graph, as in Fig. 6.4.6, and we must 'unwrap' it (roll the 'cylinder' along the φ axis) in order to get a continuous function. Matlab has the ADDTWOPI (older) and UNWRAP (newer) routine, but both perform irregularly for $\varphi > 4\pi$, especially for systems with zeros. The following EPHD routine works correctly.

```
function q=ephd(phi)
% EPHD      Eliminate PHase Discontinuities.
%     Outperforms UNWRAP and ADDTWOPI for systems with zeros.
%     Use :       q=ephd(phi);
%     where :
%     phi --> input phase vector in radians ( range: -pi>=phi>=pi );
%     q   --> output phase vector, "unwrapped";
%     If phi is a matrix, unwrapping is performed down each column.

% Author :  Erik Margan, 890505, Free of copyright !

[r,c]=size(phi);
if min(r,c) == 1
      phi=phi(:);       % column-wise orientation
      c=1;
end
q=diff(phi);            % differentiate to detect discontinuities
      % compensate for one element lost in diff and round the steps :
q=[zeros(1:c); pi*round(q/pi)];
q=cumsum(q);            % integrate back by cumulatively summing
q=phi-q;                % subtract the correcting values
if r == 1
      q=q.';            % restore orientation
end
```

The 'trick' used in the EPHD routine is to first differentiate the phase, in order to find where the discontinuities are and determine how large they are, then normalize them by dividing by π, round this to integers, multiply back by π, integrate back to obtain the corrections, and subtract the corrections from the original phase vector.

Following our 5th-order Butterworth example, we can now write:

```
alpha=ephd(phi);            % unwrapped ;
semilogx(w,180*alpha/pi)    % show alpha in degrees vs. log-scaled w;
```

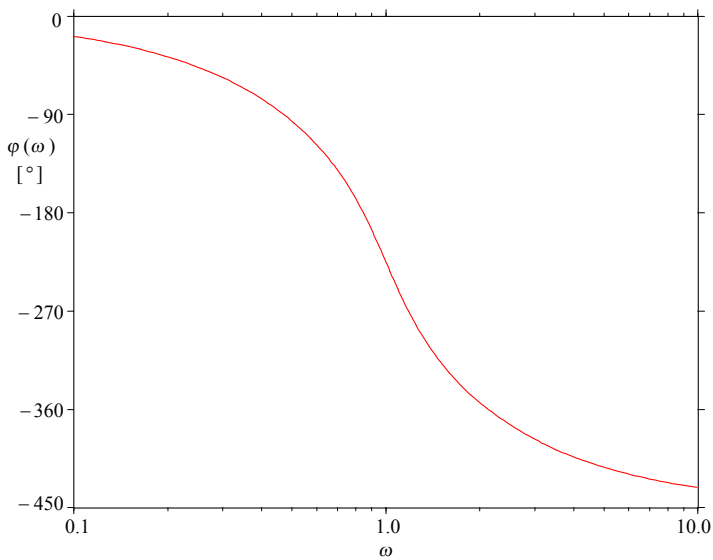


Fig. 6.4.7: Bode plot of 'unwrapped' phase, in a linear degree vs. log frequency scale.

Plotting the phase in linear degrees vs. log scaled frequency reveals an interesting fact: the system exhibits a 90° phase shift for each pole, 450° total phase shift for the 5^{th}-order system. Also, the phase shift at the cut off frequency ω_h is exactly one half of the total phase shift (at $\omega \gg \omega_h$).

Another important fact is that stable systems (those with poles on the left half of the complex plane) will always exhibit a **negative** phase shift, whatever the system configuration (low pass or high pass, inverting or non-inverting). If you ever see a phase graph with a positive slope, first inspect what is the system gain in that frequency region. If it is 0.1 or higher that is a cause for major concern (that is, if your intention was not to build an oscillator!).

6.4.3 Frequency Dependent Envelope Delay

The **envelope delay** is defined as the phase versus frequency derivative:

$$\tau_{\mathrm{d}}(\omega) = \frac{d\varphi(\omega)}{d\omega} \tag{6.4.3}$$

(note: φ must be in radians!). Now it becomes evident why we have had to 'unwrap' the circular phase function: each 2π discontinuity would, when differentiated, produce a very high, sharp spike in the envelope delay.

Numerical differentiation can be performed by simply taking the difference of each pair of adjacent elements for both the phase and the frequency vector:

```
dphi=phi(2:1:300)-phi(1:1:299);
dw=w(2:1:300)-w(1:1:299);
```

But Matlab has a built in command called DIFF, so let us use it:

```
tau=diff(phi)./(diff(w));
```

By doing so we run into an additional problem. Numerical differentiation assigns a value to each difference of two adjacent elements, so if we started from N elements the differentiation will return $N-1$ differences. Since each difference is assigned to the interval between two samples, instead of the samples themselves, this results in a half interval delay when the result is displayed against the original frequency vector `w`.

If we have low density data we should compensate this by redefining `w`. For a linearly scaled frequency vector we would simply take the algebraic mean, `w=(w(2:1:N)-w(1:1:N-1))/2`. But for a log-scaled frequency vector, a geometric mean (the square root of a product) is needed, as in the example below:

```
w1=sqrt(w(1:1:299).*w(2:1:300));
semilogx(w1,tau)                    % see the result in Fig.6.4.8.
```

Note that the values in variable `tau` are negative, reflecting the fact that the system output is delayed in time. Since we call this response a 'delay' by definition, we could use the absolute value. However, we prefer to keep the negative sign, because it also reflects the sense of the phase rotation (see Fig. 6.4.3 and 6.4.7). An upward rotating phase (or a counter-clockwise rotation in the Bode plot of the complex frequency response) would imply a positive time delay or output before input and, consequently, an unstable or oscillatory system.

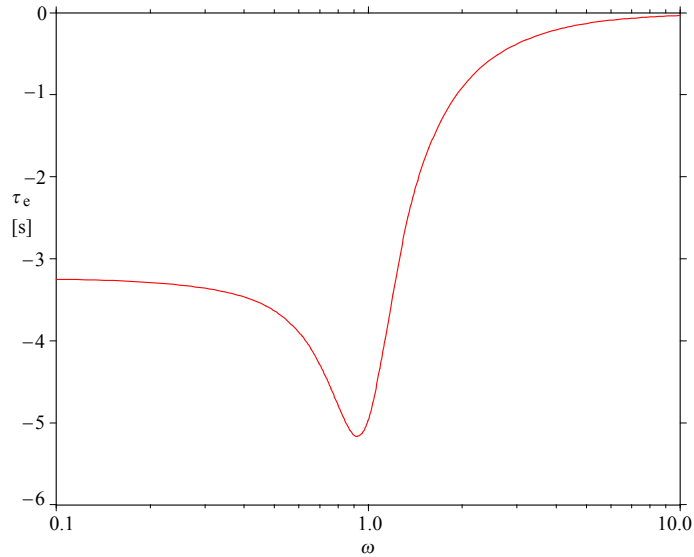


Fig. 6.4.8: The envelope (group) delay vs. frequency of the 5th-order Butterworth system. The delay is the largest for frequencies where the phase has the greatest slope.

So far we have derived the phase and group delay functions from the complex response to imaginary frequency. There are times, however, when we would like to save either processing time or memory requirement (as in embedded instrumentation applications). It is then advantageous to calculate the phase or the group delay directly from the system poles (and zeros, if any) and the frequency vector.

The phase influence of a single pole p_k can be calculated as:

$$\varphi_k(\omega, p_k) = \arctan \frac{\omega - \Im\{p_k\}}{\Re\{p_k\}} \qquad (6.4.4.)$$

The influence of a zero is calculated in the same way, but with a negative sign. The total system phase shift is equal to the sum of all particular phase shifts of poles and zeros:

$$\varphi(\omega) = \sum_{k=1}^{n} \varphi_k(\omega, p_k) - \sum_{i=1}^{m} \varphi_i(\omega, z_i) \qquad (6.4.5.)$$

But owing to the inherent complex conjugate symmetry of poles and zeros, only half of them need to be calculated and the result is then doubled. If the system order is odd the real pole is summed just once, the same is true for any real zero. This, of course, requires some sorting procedure of the system poles and zeros, but sorting is performed much quicker than multiplication with ω, which is usually a lengthy vector. If we are interested in getting data for a single frequency, or just two or three characteristic points, then it might be faster to skip sorting and calculate with all poles and zeros. In Matlab poles and zeros are already returned sorted. See the PHASE routine, in which Eq. 6.4.4 and 6.4.5 were implemented.

Note also that with the PHASE routine we obtain the 'unwrapped' phase directly and we do not have to recourse to the EPHD routine.

```
function phi=phase(z,p,w)
% PHASE  returns the phase angle of the system specified by the zeros
%        z and poles p for the frequencies in vector w :
%
%        Call :   phi=phase(z,p,w);
%
%     Instead of using angle(freqw(z,p,w)) which returns the phase
%     in the range +/-pi, this routine returns the "unwrapped" result.
%     See also FREQW, ANGLE, EPHD and GDLY.

% Author: Erik Margan, 890327, Last rev.: 980925, Free of copyright !

if nargin == 2
      w = p ;
      p = z ;
      z = [];           % A system with poles only.
end
if any( real( p ) > 0 )
      disp('WARNING :  This is not a Hurwitz-type system !' )
end
n = max( size( p ) ) ;
m = max( size( z ) ) ;
% find w orientation to return the result in the same form.
[ r, c ] = size( w ) ;
if c == 1
      w = w(:).' ;      % make it a row vector.
end
% calculate phase angle for each pole and zero and sum it columnwise.
phi(1,:) = atan( ( w - imag( p(1) ) ) / real( p(1) ) ) ;
for k = 2 : n
      phi(2,:) = atan( ( w - imag( p(k) ) ) / real( p(k) ) ) ;
      phi(1,:) = sum( phi ) ;
end
if m > 0
      for k = 1 : m
            phi(2,:) = atan( ( imag( z(k) ) - w ) / real( z(k) ) ) ;
            phi(1,:) = sum( phi ) ;
      end
end
phi( 2, : ) = [] ;      % result is in phi(1,:)
if c == 1
      phi = phi(:) ;    % restore the form same as w.
end
```

A similar procedure can be applied to the group delay. The influence of a single pole p_k is calculated as:

$$\tau_k(\omega, p_k) = \frac{\Re\{p_k\}}{\Re\{p_k\}^2 + \left(\Im\{p_k\} - \omega\right)^2} \qquad (6.4.6.)$$

As for the phase, the total system group delay is a sum of all delays for each pole and zero:

$$\tau_{\mathrm{d}}(\omega) = \sum_{k=1}^{n} \tau_k(\omega, p_k) - \sum_{i=1}^{m} \tau_i(\omega, z_i) \qquad (6.4.7.)$$

Again, owing to the complex conjugate symmetry, only half of the complex poles and zeros need to be taken into account and the result doubled, and any delay of an eventual real pole or zero is then added to it. The GDLY (Group DeLaY) routine implements Eq. 6.4.6 and 6.4.7.

```
function tau=gdly(z,p,w)
% GDLY returns the group (envelope) time delay for a system defined
%      by zeros z and poles z, at the chosen frequencies w.
%
%      Call :    tau=gdly(Z,P,w);
%
%      Although the group delay is defined as a positive time lag,
%      by which the system response lags the input, this routine
%      returns a negative value, since this reflects the sense of
%      phase rotation with frequency.
%
%      See also FREQW, PATS, ABS, ANGLE, PHASE.

% Author: Erik Margan, 890414, Last rev.: 980925, Free of copyright !

if nargin == 2
      w=p;
      p=z;
      z=[]; % system has poles only.
end
if any( real( p ) > 0 )
      disp( 'WARNING :  This is not a Hurwitz type system !' )
end
n=max(size(p));
m=max(size(z));
[r,c]=size(w);
if c == 1
      w=w(:).' ;  % make it a row vector.
end
tau(1,:) = real(p(1)) ./(real(p(1))^2 + (w-imag(p(1))).^2);
for k = 2 : n
      tau(2,:) = real(p(k)) ./(real(p(k))^2 + (w-imag(p(k))).^2);
      tau(1,:) = sum( tau ) ;
end
if m > 0
      for k = 1 : m
            tau(2,:)=-real(Z(k)) ./(real(Z(k))^2 + (w-imag(Z(k))).^2);
            tau(1,:) = sum( tau ) ;
      end
end
tau(2,:) = [] ;
if c == 1
      tau = tau(:) ;
end
```

6.5 Transient Response by Fourier Transform

There are several methods for time domain response calculation. Three of these that are interesting from the system designer's point of view, including the FFT method, were compared for efficiency and accuracy in [Ref. 6.23]. Besides the high execution speed, the main advantage of the FFT method is that we do not even have to know the exact mathematical expression for the system frequency response, but only the graph data (i.e. if we have measured the frequency and phase response of a system). Although the method was described in detail in [Ref. 6.23] we shall repeat here the most important steps, to allow the reader to follow the algorithm development.

There are **five difficulties** associated with the discrete Fourier transform that we shall have to solve:

a) the inability to transform some interesting functions (e.g., the unit step);

b) the correct treatment of the DC level in low pass systems;

c) preserving accuracy with as little spectral information input as possible;

d) find to what extent our result is an approximation owed to finite spectral density;

e) equally important, estimate the error owed to finite spectral length.

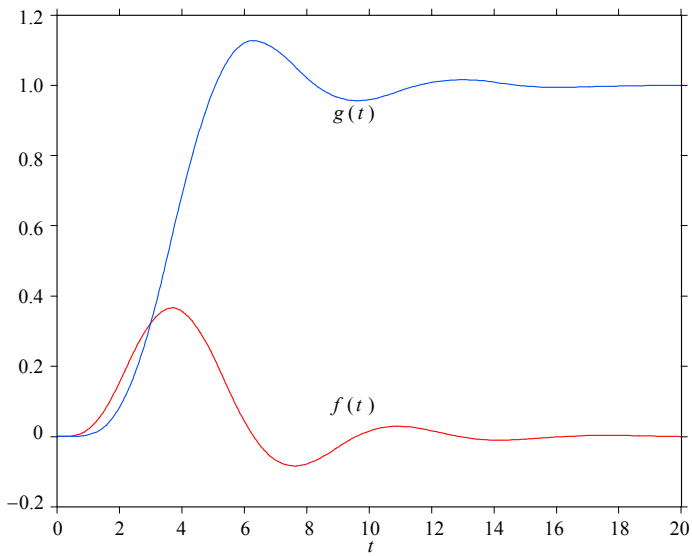


Fig. 6.5.1: The impulse and step response of the 5th-order Butterworth system. The impulse amplitude has been normalized to represent the response to an ideal, infinitely narrow, infinite amplitude input impulse. The impulse response reaches the peak value at the time equal to the envelope delay value at DC; this delay is also the half amplitude delay of the step response. The step response first crosses the final value at the time equal to the envelope delay maximum. Also the step response peak value is reached when the impulse response crosses the zero level for the first time. If the impulse response is normalized to have the area (the sum of all samples) equal to the system DC gain, the step response would be simply a time integral of it.

6.5.1 Impulse Response, Using FFT

The basic idea behind this method is that the Fourier transform is a special case of the more general Laplace transform and the Dirac impulse function is a special type of signal for which the Fourier transform solution always exists. Comparing Eq. 1.3.8 and Eq. 1.4.3 and taking in account that $s = \sigma + j\omega$, we see:

$$\mathcal{L}\{f(t)\} = \mathcal{F}\{f(t)\,\mathrm{e}^{-\sigma t}\} \tag{6.5.1}$$

Since the complex plane variable s is composed of two independent parts (real and imaginary), then $F(s)$ may be treated as a function of two variables, σ and ω. This can be most easily understood by looking at Fig. 6.4.1, in which the complex frequency response (magnitude) of a 5-pole Butterworth function is plotted as a 3D function over the Laplace plane.

In that particular case we had:

$$F(s) = \frac{-s_1 s_2 s_3 s_4 s_5}{(s-s_1)(s-s_2)(s-s_3)(s-s_4)(s-s_5)} \tag{6.5.2}$$

where $s_{1\text{–}5}$ have the same values as in the example at the beginning of Sec. 6.4.1.

When the value of s in Eq. 6.5.2 becomes close to the value of one of the poles, s_i, the magnitude $|F(s)|$ then increases until becoming infinitely large for $s = s_\mathrm{i}$.

Let us now introduce a new variable p such that:

$$p = s\,\Big|_{\sigma\,=\,0} \qquad \text{or:} \qquad p = j\omega \tag{6.5.3}$$

This has the effect of slicing the $|F(s)|$ surface along the imaginary axis, as we did in Fig. 6.4.1, revealing the curve on the surface along the cut, which is $|F(j\omega)|$, or in words: the magnitude $M(\omega)$ of the complex frequency response. As we have indicated in Fig. 6.4.5, we usually show it in a log-log scaled plot. However, for transient response calculation a linear frequency scale is appropriate (as in Fig. 6.4.2), since we need the result of the inverse transform in linear time scale increments.

Now that we have established the connection between the Laplace transformed transfer function and its frequency response we have another point to consider: conventionally, the Fourier transform is used to calculate waveform spectra, so we need to establish the relationship between a frequency response and a spectrum. Also we must explore the effect of taking **discrete values (sampling)** of the time domain and frequency domain functions, and see to what extent we **approximate** our **results** by taking **finite length vectors of finite density sampled data.** Those readers who would like to embed the inverse transform in a microprocessor controlled instrument will have to pay attention to **amplitude quantization (finite word length)** as well, but in Matlab this is not an issue.

We have examined the Dirac function $\delta(t)$ and its spectrum in Part 1, Sec. 1.6.6. Note that the spectral components are separated by $\Delta\omega = 2\pi/T$, where T is the impulse repetition period. If we allow $T \to \infty$ then $\Delta\omega \to 0$. Under these conditions we

can hardly speak of discrete spectral components because the spectrum has become very dense; we rather speak of **spectral density**. Also, instead of individual components' magnitude we speak of **spectral envelope** which for $\delta(t)$ is essentially flat.

However, if we do not have an infinitely dense spectrum, then $\Delta\omega$ is small but not 0, and this merely means that the impulse repeats after a finite period $T = 2\pi/\Delta\omega$ (this is the mathematical equivalent of testing a system by an impulse of a duration much shorter than the smallest system time constant and of a repetition period much larger than the largest system time constant).

Now let us take such an impulse and present it to a system having a selective frequency response. Fig. 6.5.2 shows the results both in the time domain and the frequency domain (magnitude). The time domain response is obviously the system impulse response, and its equivalent in the frequency domain is a spectrum, whose density is equal to the input spectral density, but with the spectral envelope shaped by the system frequency response. The conclusion is that we only have to sample the frequency response at some finite number of frequencies and perform a discrete Fourier transform inversion to obtain the impulse response.

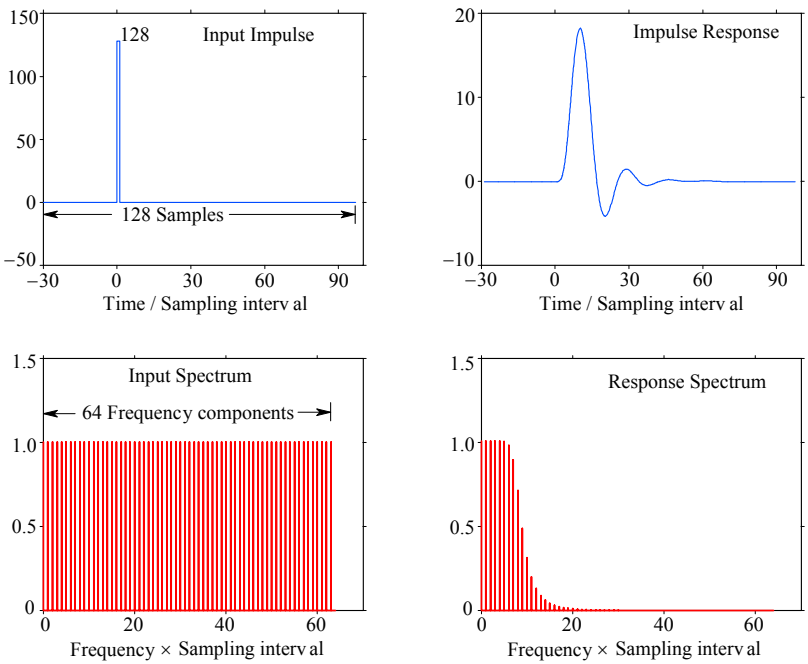


Fig. 6.5.2: Time domain and frequency domain representation of a 5-pole Butterworth system impulse response. The spectral envelope (only the magnitude is shown here) of the output is shaped by the system frequency response, whilst the spectral density remains unchanged. From this fact we conclude that the time domain response can be found from a system frequency response using inverse Fourier transform. The horizontal scale is the number of samples (128 in the time domain and 64 in the frequency domain — see the text for the explanation).

If we know the magnitude and phase response of a system at some finite number of equally spaced frequency points, then each point represents:

$$F_i = M_i \cos(\omega_i t - \varphi_i) \tag{6.5.4}$$

As the contribution of frequencies components which are attenuated by more than, say, 60 dB can be neglected, we do not have to take into account an infinitely large number of frequencies, and the fact that we do not have an infinitely dense spectrum merely means that the input impulse repeats in time. By applying the superposition theorem, the output is then equal to the sum of all the separate frequency components.

Thus for each time point the computer must perform the addition:

$$f(t_k) = \sum_{i(\omega_{\min})}^{i(\omega_{\max})} M_i \cos(\omega_i t_k - \varphi_i) \tag{6.5.5}$$

Eq. 6.5.5 is the **discrete Fourier transform**, with the exponential part expressed in trigonometric form. However, if we were to plot the response calculated after Eq. 6.5.5, we could see that the time axis is reversed, and from the theory of Fourier transform properties (symmetry property, [Ref. 6.14, 6.15, 6.18]), we know that the application of two successive Fourier transforms returns the original function but with the sign of the independent variable reversed:

$$\mathcal{F}\Big\{\mathcal{F}\{f(t)\}\Big\} = \mathcal{F}\{F(j\omega)\} = f(-t) \tag{6.5.6}$$

or more generally:

$$f(t) \underset{\mathcal{F}^{-1}}{\overset{\mathcal{F}}{\rightleftarrows}} F(j\omega) \underset{\mathcal{F}^{-1}}{\overset{\mathcal{F}}{\rightleftarrows}} f(-t) \underset{\mathcal{F}^{-1}}{\overset{\mathcal{F}}{\rightleftarrows}} F(-j\omega) \underset{\mathcal{F}^{-1}}{\overset{\mathcal{F}}{\rightleftarrows}} f(t) \tag{6.5.7}$$

The main drawback in using Eq. 6.5.5 is the high total number of operations, because there are three input data vectors of equal length (ω, M, φ) and each contributes to every time point result. It seems that greater efficiency might be obtained by using the input frequency response data in the complex form, with the frequency vector represented by the index of the $F(j\omega)$ vector.

Now $F(j\omega)$ in its complex form is a two sided spectrum, as was shown in Fig. 6.4.3, and we are often faced with only a single sided spectrum. It can be shown that a real valued $f(t)$ will always have $F(j\omega)$ symmetrical about the real axis σ. Thus:

$$F_N(j\omega) = F_P^*(-j\omega) \tag{6.5.8}$$

F_N and F_P are the $\omega < 0$ and $\omega > 0$ parts of $F(j\omega)$, with their inverse transforms labeled $f_N(t)$ and $f_P(t)$. Note that F_P^* is the complex conjugate of F_P.

This symmetry property follows from the definition of the **negative frequency concept**: instead of having a single phasor rotating counter-clockwise (positive by definition) in the complex plane, we can always have two half amplitude phasors rotating in opposite directions at the same frequency (as we have already seen drawn in Part 1, Fig. 1.1.1; for vector analysis see Fig. 6.5.3**)**. We can therefore conclude that the

inherent conjugate symmetry of the complex plane allows us to define 'negative frequency' as a clockwise rotating, half amplitude phasor, being the complex conjugate of the usual counter-clockwise (positive by definition) rotating (but now also half amplitude) phasor. And this is not just a fancy way of making simple things complex, but is rather a direct consequence of our dislike of sine–cosine representation and the preference for the complex exponential form, which is much simpler to handle analytically.

One interesting aspect of the negative frequency concept is the ***Shannon* sampling theorem**: for a continuous signal, sampled with a frequency f_s, all the information is contained within the frequency range between 0 and $f_s/2$, because the spectrum from $f_s/2$ to f_s is a mirror image, so the spectrum is symmetrical about $f_s/2$, the *Nyquist* frequency. Therefore a frequency equal to f_s can not be distinguished from a DC level, and any frequency from $f_s/2$ to f_s can not be distinguished from $f_s - f$.

But please, also note that this 'negative frequency' does not necessarily imply 'negative time', since the negative time is defined as the time before some arbitrary instant $t = 0$ at which the signal was applied. In contrast, the negative frequency response is just one half of the full description of the $t \geq 0$ signal.

However, those readers who are going explore the properties of the *Hilbert transform* will learn that this same concept can be extended to the $t < 0$ signal region, but this is beyond the scope of this text.

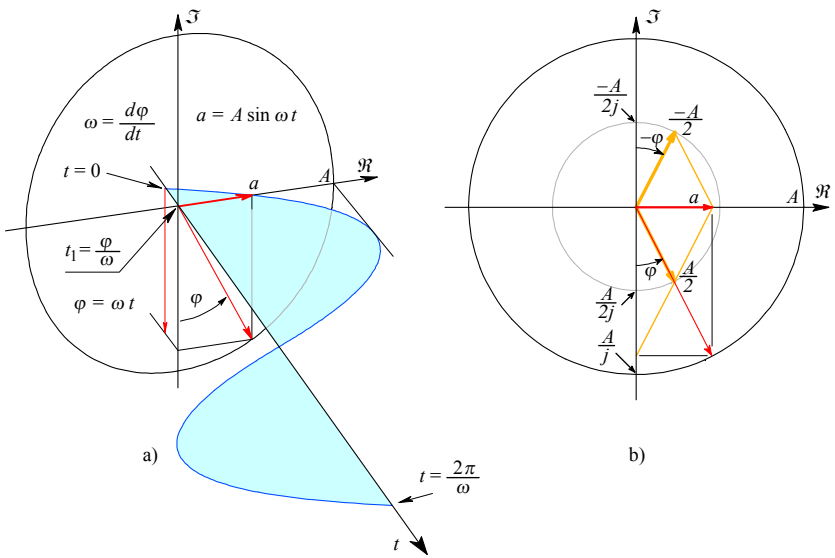


Fig. 6.5.3: As in Part 1, Fig. 1.1.1, but from a slightly different perspective: **a)** the real signal instantaneous amplitude $a(t) = A \sin\varphi$, where $\varphi = \omega t$; **b)** the real part of the instantaneous signal phasor, $\Re\{\vec{0A}\} = \vec{0a} = A \sin\varphi$, can be decomposed into two half amplitude, oppositely rotating, complex conjugate phasors, $-(A/2j)\sin\varphi + (A/2j)\sin(-\varphi)$. The second term has rotated by $-\varphi = -\omega t$ and, since t is obviously positive (see the **a)** graph), the negative sign is attributed to ω; thus, **clockwise rotation is interpreted as a 'negative frequency'**.

Eq. 6.5.8 can thus be used to give:

$$F(j\omega) = F_P(j\omega) + F_P^*(-j\omega) \tag{6.5.9}$$

but from Eq. 6.5.7 we have:

$$\mathcal{F}^{-1}\Big\{F_P^*(-j\omega)\Big\} = f_P^*(t) \tag{6.5.10}$$

hence using Eq. 6.5.9 and Eq. 6.5.10 and taking into account the cancellation of imaginary parts, we obtain:

$$f(t) = f_P(t) + f_P^*(t) = 2\,\Re\Big\{f_P(t)\Big\} \tag{6.5.11}$$

Eq. 6.5.11 means that if $F_P(j\omega)$ is the $\omega \geq 0$ part of the Fourier transformed real valued function $f(t)$, its Fourier transform inversion $f_P(t)$ will be a complex function whose real part is equal to $f(t)/2$. Summing the complex conjugate pair results in a doubled real valued $f(t)$. So by Eq. 6.5.10 and Eq. 6.5.11 **we can calculate the system impulse response from just one half of its complex frequency response using the forward Fourier transform** (not inverse!):

$$f(t) = 2\,\Re\left\{\left[\mathcal{F}\Big\{F_P^*(j\omega)\Big\}\right]^*\right\} \tag{6.5.12}$$

Note that the second (outer) complex conjugate is here only to satisfy the mathematical consistency — in the actual algorithm it can be safely omitted, since only the real part is required.

As the operator $\mathcal{F}\{\}$ in Eq. 6.5.12 implies integration we must use the **discrete Fourier transform (DFT)** for computation. The DFT can be defined by decomposing the Fourier transform integral into a finite sum of N elements:

$$F(k) = \frac{1}{N}\sum_{i=0}^{N-1} f(i)\,\mathrm{e}^{-j\frac{2\pi k}{N}} \tag{6.5.13}$$

That means going again through a large number of operations, comparable to Eq. 6.5.5. Instead we can apply the **Fast Fourier Transform (FFT)** algorithm and, owing to its excellent efficiency, save the computer a great deal of work.

Cooley and *Tukey* [Ref. 6.16] have shown that if $N = 2^B$ and B integer, there is a smart way to use Eq. 6.5.13, owing to the periodical nature of the Fourier transform.

If Eq. 6.5.13 is expressed in a matrix form then the matrix which represents its exponential part can be divided into its even and odd part, and the even part can be assigned to $N/2$ elements. The remaining part can then also be divided as before and the same process can then be repeated over and over again, so that we end up with a

number ($\log_2 N$) of individual sub-matrices. Furthermore, it can be shown that these sub-matrices contain only two non-zero elements, one of which is always unity (1 or j).

Therefore multiplying by each of the factor matrices requires only N complex multiplications.

Finally (or firstly, depending on whether we are using the 'decimation in frequency' or the 'decimation in time' algorithm), we rearrange the data, by writing the position of each matrix element in a binary form, and reordering the matrix it by reversing the binary digits (this operation is often referred to as the 'reshuffling').

The total number of multiplications is thus $N \log_2 N$ instead of the N^2 required to multiply in one step. Other operations are simple and fast to execute (addition, change of sign, change of order). Thus in the case of $B = 10$, $N = 1024$, and $N^2 = 1048576$ whilst $N \log_2 N = 10240$, so a reduction of the required number of multiplications by **two orders of magnitude** has been achieved.

Matlab has a command named 'FFT' which uses the 'radix-2' type of algorithm and we shall use it as it is. Those readers who would like to implement the FFT algorithm for themselves can find the detailed treatment in [Ref. 6.16, 6.17 and 6.19].

A property of the FFT algorithm is that it returns the spectrum of a real valued signal as folded about the Nyquist frequency (one half of the frequency at which the signal was sampled). As we have seen in Fig. 6.5.2, if we have taken 128 signal samples, the FFT returns the first 64 spectral components from $\omega = 0, 1, 2, \ldots, 63$ but then the remaining 64 components, which are the complex conjugate of the first ones, are in the reversed order.

This is in contrast to what we were used to in the analytical work, as we expect the complex conjugate part to be on the $\omega < 0$ side of the spectrum. On the other hand, this is equivalent, since the 128 samples taken in the signal time domain window are implicitly assumed to repeat, and consequently the spectrum is also repeating every 128 samples. So if we use the standard inverse FFT procedure we must take into account all 128 spectral components to obtain only 64 samples of the signal back. However, note that the Eq. 6.5.12 **requires only a single–sided spectrum of N points to return N points of the impulse response**. This is, clearly, an **additional two–fold improvement** in algorithm efficiency.

6.5.2 Windowing

For calculating the transient response a further reduction in the number of operations is possible through the use of **'windowing'**. Windowing means multiplying the system response by a suitable window function. We shall use the windowing in the frequency domain, the reason for this we have already discussed when we were considering to what extent DFT is an approximation. Like many others before us, in particular the authors of various window functions, we, too, have found out that the accuracy improves if the influence of higher frequencies is reduced.

Since the frequency response of a high order system (third-order or greater) falls off quickly above the cut off frequency, we can take just $N = 256$ frequency samples, and after inverse FFT we still obtain a time domain response with an accuracy equal to or better than the vertical resolution of the VGA type of graphics (1/400 or 0.25%). And as the sample density (number of points) of the transient wavefront increases with the number of stop band frequency samples, it is clear that the smaller the contribution of higher frequencies, the greater is the accuracy.

But, in order to achieve a comparable accuracy with 1st- and 2nd-order systems we would have to use $N_1 = 4096$ and $N_2 = 1024$ frequency samples, respectively. Thus since we would like to minimize the length of the frequency vector, low order systems need to be artificially rolled off at the high end. This can be done by multiplying the frequency response by a suitable window function, element by element, as shown in Fig. 6.5.4. The window function used in Fig. 6.5.4 (and also in the TRESP routine) is a real valued Hamming type of window (note that we need only its right hand half, since we use a single-sided spectrum; the other half is implicitly used owing to Eq. 6.5.12).

```
W=0.54-0.46*cos(2*pi*(N+1:1:2*N)/(2*N)); % right half Hamming window
```

The physical effect of applying the Hamming window is similar to the implementation of an additional non-dominant pole of high order at about $12\,\omega_h$ and a zero at about $25\,\omega_h$. Multiplication in frequency domain is equivalent to convolution in time domain, and vice-versa; however, note that the window function is **real**, thus no phase distortion occurs!

After extensive experimentation with different types of windows, it was found that the Hamming window yields the lowest error, most notably at the first few points of the impulse response. This is understandable, since the FFT of this window has lowest spectral side lobes and the same is true for the inverse transform. For the same reason the final value error of the 1st- and 2nd-order system will also be reduced.

Note that the first point of the 1st-order impulse response will be in error anyway, because we have started from a spectrum of finite density, in which the distance between adjacent spectral components was equal to $1/N$. The time domain equivalent of this is an impulse whose amplitude is finite (N) and its width is $1/N$ (so that the impulse area is equal to 1). This means that the response rise time is not infinitely short, so its first point will be smaller than expected, the error being proportional to $1/N$.

If the system order and type could be precisely identified, this error might be corrected by forcing the first point of the 1st-order impulse response to a value equal to

the sampling time period, multiplied by the system DC gain, as has been done in the TRESP routine.

In Sec. 6.5.6 we give a more detailed error analysis for the 1st-, 2nd- and 3rd-order system for both windowed and non-windowed spectrum.

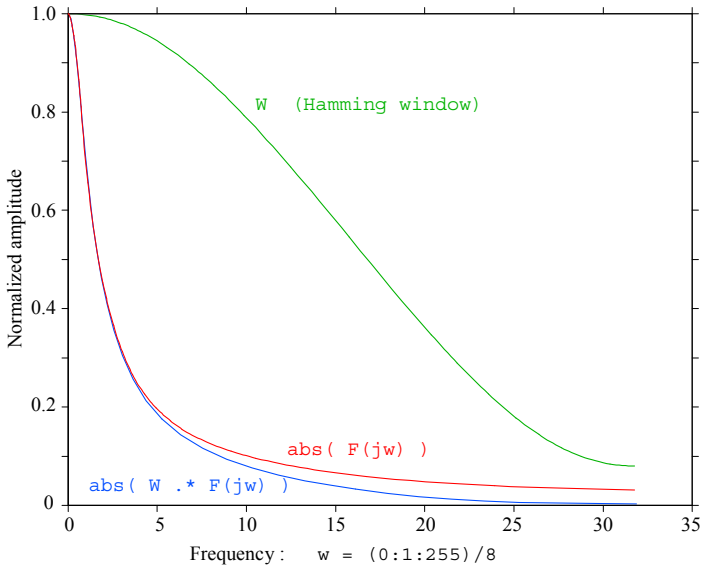


Fig. 6.5.4: Windowing example. The 1st-order frequency response (only the magnitude is shown on plot) is multiplied element by element by the Hamming type of window function in order to reduce the influence of high frequencies and improve the impulse response calculation accuracy. Note that the window function is real only, affecting equally the system real and imaginary part, thus the phase information is preserved and only the magnitude is corrected.

6.5.3 Amplitude Normalization

The impulse response, as returned by Eq. 6.5.12, has the amplitude such that the sum of all the N values is equal to N times the system gain. Also, if we are dealing with a low pass system and the first point of its frequency response (the DC component) is $F(0)$, then the impulse response will be shifted vertically by $F(0)$ (the DC component is added). Thus we must first cancel this 'DC offset' by subtracting $F(0)$ and then normalize the amplitude by dividing by N:

$$f_{\mathrm{n}}(t) = \frac{f(t) - F(0)}{N} \tag{6.5.14}$$

By default, the TRESP routine (see below) returns the impulse amplitude in the same way, representing a unity gain system's response. Optionally, we can denormalize it as if the response was caused by an ideal, infinitely high impulse; then the 1st-order response starts the exponential decay from a value very close to one, as it should. If the system's half power bandwidth, ω_{h}, is found at the `m+1` element of the frequency response vector, the amplitude denormalization factor will be:

$$A = \frac{N}{2\pi\,\mathrm{m}} \tag{6.5.15}$$

The 2π factor comes as a bit of surprise here. See Sec. 6.5.5 about time scale normalization for an explanation.

The term `m` can be entered explicitly, as a parameter. But it can also be derived from the frequency vector by finding the index at which it is equal to 1, or it can be found by examining the magnitude and finding the index of the point, nearest to the half power bandwidth value (in both of cases the index must be decremented by 1).

Another problem can be encountered with high order systems, which exhibit a high degree of ringing, e.g., Chebyshev systems of order 8 or greater. If `m<8`, some additional ringing is introduced into the time domain response. This ringing results from the time frame implicitly repeating with a period $T = 2\pi/\Delta\omega$, where $\Delta\omega$ describes the finite spectral density of input data. If we have specified the system cut off frequency too near to the origin of the frequency vector it would cause a time scale expansion. Thus overlapping of adjacent responses will introduce distortion if the impulse response has not decayed to zero by the end of the period T. Therefore the choice of placing the cut off frequency relative to the frequency vector is a compromise between the pass band and stop band description. In Matlab, the frequency vector of `N` linearly spaced frequencies, normalized to 1 at its `m+1` element, can be written as:

```
N=256;      m=8;          w=(0:1:N-1)/m;
```

The variable `m` specifies the normalized frequency unit. The transient response of both Butterworth and Bessel systems can be calculated with good accuracy by using a frequency vector normalized to 1 at its 5th sample (`m=4`). But by placing the cutoff frequency at the 9th sample (`m=8`) of a frequency vector of length 256, an acceptably low error will be achieved even for a 10th-order Chebyshev system. For higher order, high ringing systems one will probably need to increase the frequency vector to 512 or 1024 elements in order to prevent time window overlapping.

6.5.4 Step Response

Up to this point we have obtained the impulse response. The step response is not available straight from the Fourier transform (if the unit step is integrated, the integral will diverge to $+\infty$; this is why we prefer the more general Laplace transform for analytical work). However, from signal analysis theory we know that the response to an arbitrary input signal can be found by convolving it with the system's impulse response (for convolution, see Part 1, Sec. 1.15 and Fig. 1.15.1; see Part 7, Sec. 7.2 for the numerical algorithm). With the unit step as the input signal the convolution is reduced to a simple time domain integration of the system impulse response.

But this integration must give us the final step response value equal, or at least very close, to the system DC gain. So we must use the impulse response normalized to obtain the sum of all its elements equal to the system's gain. Numerical integration can be done by cumulative summing. This means that the first element is transferred unchanged, the second element is the sum of the first two, the third of the first three, and so on, up to the last element which is the sum of all elements. The CUMSUM command in Matlab will return in `Y` a cumulative sum of vector `X` like this:

```
                              % the CUMulative-SUM example :
Y(1)=X(1);                    % transfer the first sample unchanged
for k=2:max(size(X))
     Y(:,k)=Y(k-1)+X(k);      % next sample is the sum of all previous
end
```

However, inside the Matlab command interpreter, this `for`-loop executes considerably slower, owing to vector remapping, thus we shall use the built in CUMSUM command in the TRESP routine.

An interesting problem arises with the first-order system, as well as any system with zeros, since the first sample of the impulse response will be non-zero, therefore the step response will also start from above zero, which is not so in the real world. We can solve this problem by artificially adding a zero valued first element to the impulse response vector, but we then have to be careful not to increment the total number of elements by one. That is because we want to keep the sum of the original impulse response vector, divided by the number of elements, equal to the system gain (the increased number of elements affects the normalization).

Also, as we have seen in the envelope delay derived from the phase response, the numerical differentiation, owing to a finite number of elements taken into account, results in a one half sample shift. The numerical integration, being an inverse process, does the same thing, even if does not increase the number of elements by one. The integration shift is also in the opposite direction.

Both these problems will be treated in the next section dealing with time scale normalization.

6.5.5 Time Scale Normalization

When we plot a time domain response we want to be able to correlate it to the system's envelope delay, so we need to specify the time scale normalization factor. This factor depends on how many samples of the frequency response were used in the FFT and which sample was at the system's cut off frequency.

We have already seen that a Fourier transformed signal has a spectrum periodic in frequency, the period being $\omega_s = 2\pi/\Delta t$, thus inversely proportional to the sampling period. Also the spectral density $\Delta\omega$ reflects the overall time domain repetition period, $T_R = 2\pi/\Delta\omega$, which represents the size of the time domain window. Note that when we specify the single-sided spectrum, we usually do it from 0 up to the Nyquist frequency $\omega_N = \omega_s/2 = N\Delta\omega$.

Now our frequency vector has been defined as: `w=(0:1:N-1)/m`. Obviously, at its `m+1` element `w(m+1)=1`, which is the numerical equivalent of the normalized cut off frequency $\omega_h = 1$. Also, $\Delta\omega$ can be found as `w(2)-w(1)`, but `w(1)=0` and `w(2)=1/m`, so we can say that $\Delta\omega = \omega_h/m$. Since the Nyquist frequency is $\omega_N = N\Delta\omega$, and the sampling frequency is twice that, $\omega_s = 2N\Delta\omega$, then the sampling time interval is $\Delta t = 1/\omega_s = 1/2N\Delta\omega = m/2N\omega_h$.

But remember that Eq. 6.5.12 allows us to improve our algorithm, obtaining N time samples from N frequency samples, a result which we would otherwise get from $2N$ frequency samples. So, we have $\Delta t = m/N\omega_h$. Also remember that we have calculated the frequency response using a normalized frequency vector, ω/ω_h, and the term $\omega_h = 1$ was effectively replaced by $f_h = 1$, loosing 2π. Our sampling interval should therefore be:

$$\Delta t = \frac{2\pi m}{N} \tag{6.5.16}$$

You may have noted that this is exactly the inverse of the amplitude denormalization factor, Eq. 6.5.15, $\Delta t = 1/A$. This is not just a strange coincidence! Remember Fig. 6.5.2: the amplitude and the width of the input impulse were set so that $A\Delta t = N$, with N being also the time domain vector's length, so if $\Delta t = 1$ then $A = N$. For the unity gain system the response must contain the same amount of energy as the input, thus the sum of all the response values must also be equal to N. Of course, N is a matter of choice, but once its value has been chosen it is a constant. So it is only the system bandwidth, set by the factor `m`, that will determine the relationship between the response amplitude and its time scale.

Therefore, after Eq. 6.5.16:

```
dt=2*pi*m/N;            % delta-t
t=dt*(0:1:N-1);         % normalized length-N time vector
```

We may check the time normalization easily by calculating the impulse response of a first-order system and comparing it to the well known reference, the analytical first-order low pass RC system impulse response, which is just:

$$f_r(t) = e^{-t/RC} \tag{6.5.17}$$

Now, normalizing the time scale means showing it in increments of the system time constant $(RC, 2RC, 3RC, \ldots)$. Thus we simply set $RC = 1$. For the starting sample at $t = 0$, the response $f(0) = 1$, so in order to obtain the response of a unity gain system excited by a finite amplitude impulse we must denormalize the amplitude (see Eq. 6.5.15) by $1/A$:

$$f_{\mathrm{r}}(t) = \frac{1}{A}\mathrm{e}^{-t} \tag{6.5.18}$$

```
z=[];                                % no zeros,
p=-1;                                % just a single real pole
N=256;                               % total number of samples
m=8;                                 % samples in the frequency unit
w=(0:1:N-1)/m;                       % the frequency vector
dt=2*pi*m/N;                         % sampling time interval = 1/A
t=dt*(0:1:N-1);                      % the time vector
F=freqw(z,p,w);                      % the frequency response
In=(2*real(fft(conj(F)))-1)/N;       % the impulse response
Ir=dt*exp(-t);                       % 1st-order ref., denormalized
plot( t, Ir, t, In )
title('Ideal vs. windowed response'), xlabel('Time')
plot( t(1:30), Ir(1:30), t(1:30), In(1:30) )
title('Zoom first 30 samples')
```

In the above example (see the plot in Fig. 6.5.5), we see that the final values of normalized impulse response `In` are not approaching zero, and by zooming on the first 30 points we can also see that the first point is too low and the rest somewhat lower than the reference. Windowing can correct this:

```
W=0.54-0.46*cos(2*pi*(N+1:2*N)/(2*N));   % right half Hamming window
Iw=(2*real(fft(conj(F.*W)))-1)/N;        % impulse, windowed fr.resp.
plot( t, Ir, t, Iw ), xlabel('Time')
title('Ideal vs. windowed response')
plot( t(1:30), Ir(1:30), t(1:30), Iw(1:30) )
title('Zoom first 30 samples')
```

This plot fits the reference much better. But the first point is still far too low. From the amplitude denormalization factor, by which the reference was multiplied, we know that the correct value of the first point should be $1/A = \Delta t$. So we may force the first point to this value, but, by doing so, we would alter the sum of all values by `N*(dt-I(1))`. In order to obtain the correct final value of the step response, the impulse response requires the correction of all points by `1/(1+(dt-I(1))/N)`, as in the following example:

```
% the following correction is valid for 1st-order system only !!!
er1=dt-I(1);                         % the first point error
Iw(1)=dt;                            % correct first-point amplitude
% note that with this we have altered the sum of all values by er1,
% so we should modify all the values by :
Iw=Iw*(1/(1+er1/N));
Ir=Ir*(1/(1+er1/N));
% the same could also be achieved by :  Ix=Ix/sum(Ix);
plot( t(1:30), Ir(1:30), t(1:30), Iw(1:30) ), title('Zoom first 30')
plot( t, (Iw-Ir) ), title('Impulse response error plot')
```

Likewise we can compare the calculated step response. Our reference is then:

$$f_r(t) = 1 - e^{-t} \tag{6.5.19}$$

But if the first-order impulse response is numerically integrated the value of the first sample of the step response will be equal to the value of the first sample of the impulse response instead of zero, as it should be in the case of a low pass LTIC system.

Also, there is an additional problem resulting from numerical integration, which manifests itself as a one half sample time delay. Remember what we have observed when we derived the envelope delay from the phase: numerical differentiation has assigned each result point to each difference pair of the original data, so the resulting vector was effectively shifted left in (log-scaled) frequency by a (geometrical) mean of two adjacent frequency points, $\sqrt{\omega_n\,\omega_{n+1}}$. Because we work in linear scale, the shift is the arithmetic mean, $\Delta\omega/2$. Since the numerical integration is the inverse process of differentiation, the signal is shifted right. However, whilst the differentiation vector had one sample less, the numerical integration returns the same number of samples, not one more.

So, in order to see the actual shape of the error, we have to compensate for this shift of one half sample. We can do this by artificially adding a leading zero to the impulse response vector, then cumulatively sum the resulting $N + 1$ elements and finally take the mean value of this and the version shifted by one sample, as in the example below which is uses the vectors from above (see the result in Fig. 6.5.6):

```
Sr=1-exp(-t);                   % 1st-order step response reference
Sw=cumsum([0, Iw]);             % step resp. from impulse + leading zero
     % compensate the one half sample delay by taking the mean :
Sw=(Sw(1:N)+Sw(2:N+1))/2;
Sw(1)=0;                        % correct the first value
plot( t, Sr, t, Sw ), title('Ideal vs. windowed step response')
plot( t(1:50), Sr(1:50), t(1:50), Sw(1:50) )
title('Zoom first 50 samples')
plot( t, (Sw-Sr) ), title('Step response error plot')
```

This compensation will lower the algorithm's efficiency somewhat, but considering the embedded applications numerical addition is fast and division by 2 can be done by shifting bits one binary place to the right, so the operation can still be performed solely in integer arithmetics.

For the second-order system we can use the reference response which was calculated in Part 2, Eq. 2.2.37 (see the error plots in Fig. 6.5.7 and 6.5.8.):

```
[z,p]=buttap(2);                    % 2nd-order Butterworth system
% the following variables are the same as before:
N=256; m=8; w=(0:1:N-1)/m; dt=2*pi*m/N; t=dt*(0:1:N-1);
F=freqw(z,p,w);                     % complex frequency response
W=0.54-0.46*cos(2*pi*(N+1:2*N)/(2*N));  % a right-half Hamming window
Iw=(2*real(fft(conj(F.*W)))-1)/N;   % impulse resp., windowed fr.resp.
Sw=cumsum([0,Iw]);                  % numerical integration, step r.
Sw=(Sw(1:N)+Sw(2:N+1))/2;           % compensate half sample t-delay
T=sqrt(2);                          % 2nd-order response constants,
theta=pi/4;                         % see example: Part 2, Eq.2.2.39.
Sr=1-T*exp(-t/T).*sin(t/T+theta);   % the 2nd-order response reference
plot(t,Sr,t,Sw), title('Ideal vs. windowed step response')
plot(t(1:60),Sr(1:60),t(1:60),Sw(1:60)), title('Zoom samples 1-60')
plot( t, (Sw-Sr) ), title('Step response error plot')
```

Note that the TRESP routine allows us to enter the actual denormalized frequency vector, in which case all (but the first one) of its elements might be greater than 1. The normalized frequency unit `m` is then found from the frequency response, by checking which sample is closest to `abs(F(1))/sqrt(2)`, and then decremented by 1 to compensate for the frequency vector starting from 0. But in the case of a denormalized frequency vector we should also denormalize the time scale, by dividing the sampling interval by the actual upper cut off frequency, which is `w(m+1)`.

To continue with our 5th-order Butterworth example, we can now calculate the impulse and step response by using the TRESP routine in which we have included all the above corrections:

```
[z,p]=buttap(5);        % the 5th-order Butterworth system poles
w=(0:1:255)/8;          % form a linearly spaced frequency vector
F=freqw(z,p,w);         % the frequency response at w
[I,t]=tresp(F,w,'i');   % I : ideal impulse, t : normalized time
S=tresp(F,w,'s');       % S : step response ( time same as for I )
plot(t(1:100),I(1:100),t(1:100),S(1:100))
                        % plot 100 points of I and S vs. t
```

The results should look just like Fig. 6.5.1. Here is the TRESP routine:

```
function [y,t]=tresp(F,w,r,g)
%TRESP   Transient RESPonse, using Fast Fourier Transform algorithm.
%    Call : [y,t]=tresp(F,w,r,g);
%    where:
%     F --> complex-frequency response, length-N vector, N=2^B, B=int.
%     w --> can be the related frequency vector of F, or it
%           can be the normalized frequency unit index, or it
%           can be zero and the n.f.u. index is found from F.
%     r --> a character, selects the response type returned in y:
%           - 'u' is the unity area impulse response (the default)
%           - 'i' is the ideal impulse response
%           - 's' is the step response
%     g --> an optional input argument: plot the response graph.
%     y --> the selected system response.
%     t --> the normalized time scale vector.

% Author :  Erik Margan, 880414, Last rev. 000310, Free of copyright!

% ----------- Preparation and checking the input data ------------
if nargin < 3
      r='u';            % select the default response if not specified
end
G=abs(F(1));            % find system DC gain
N=length(F);            % find number of input frequency samples
v=length(w);            % get the length of w
if v == 1
      m=w;              % w is the normalized frequency unit or zero
elseif v == N           % find the normalized frequency unit
      m=find(abs(w-1)==min(abs(w-1)))-1;
      if isempty(m)
            m=0;        % not found, try from the half power bandwidth
      end
else
      error('F and w are expected to be of equal length !');
end
if m == 0               % find the normalized frequency unit index
      m=max(find(abs(F)>=G/sqrt(2)))-1;
```

```
end
% check magnitude slope between the 2nd and 3rd octave above cutoff
M=abs(diff(20*log10(abs(F(1+4*m*[1,2])))));
x=3;                    % system is 3rd-order or higher (>=18dB/2f)
if M < 15
      x=2;              % probably a 2nd-order system (12dB/2f)
elseif M < 9
      x=1;              % probably a 1st-order system (6dB/2f)
end

% ----------- Form the window function --------------------------
if x < 3
      W=0.54-0.46*cos(2*pi*(N+1:2*N)/(2*N)); % right half Hamming
      F=W.*F;           % frequency response windowed
end

% ----------- Normalize the time-scale --------------------------
A=2*pi*m;               % amplitude denormalization factor
dt=A/N;                 % calculate delta-t
if v == N
      dt=dt/w(m+1);     % correct for the actual frequency unit
end
t=dt*(0:1:N-1);         % form the normalized time scale

% ----------- Calculate the impulse response --------------------
y=2*real(fft(conj(F)))-G;     % calculate iFFT and null DC offset
if x == 1
      er1=A*G-y(1);     % fix the 1st-point error for 1st-order system
      y(1)=A*G;
end
if r == 'u' | r == 'U' | r == 's' | r == 'S'
      y=y/N;            % normalize area to G
end

% ----------- Calculate the step response -----------------------
if r == 's' | r == 'S'
      if x == 1
            y=y*(1/(1+er1/N)); % correct 1st-point error
      end
      y=cumsum([0, y]); % integrate to get the step response
      if x > 1
            y=(y(1:N)+y(2:N+1))/2;  % compensate half sample t-delay
            y(1)=0;
      else
            y=y(1:N);
      end
end

% ----------- Normalize the amplitude to ideal ------------------
if r == 'i' | r == 'I'
      y=y/A;            % denormalize impulse amplitude
end

% ----------- Plot the graph ------------------------------------
if nargin == 4
      plot( t, y, '-r' ), xlabel('Time')
      if r == 'i' | r == 'I' | r == 'u' | r == 'U'
            title('Impulse response')
      else
            title('Step response')
      end
end
```

6.5.6 Calculation Errors

Whilst the amount of error in low order system impulse responses might seem small, it would integrate to an unacceptably high level in the step responses if the input data were not windowed. In Fig. 6.5.5 to 6.5.10 we have computed the difference between analytically and numerically calculated impulse and step responses of Butterworth systems, using both the normal and windowed frequency response for the numerical method. Fig. 6.5.5 and 6.5.6 show the impulse and the step response of the 1st-order system, Fig. 6.5.7 and 6.5.8 show the 2nd-order system and Fig. 6.5.9 and 6.5.10 the 3rd-order system. The error plots, shown within each response plot, were magnified 10 or 100 times to reveal the details.

The initial value of the 1st-order system's impulse response, calculated from a non-windowed frequency response, is about 0.08% and falls off quickly with time. Nevertheless, it is about 10× higher than for the impulse response calculated from a windowed response. If the frequency response is not windowed, the impulse responses error eventually integrates to almost 4% of the final value in the step response.

The error plots of the second and higher order systems are much smaller, but they exhibit some decaying oscillations, independently of windowing. This oscillating error is inherent in the FFT method and it is owed to the Gibbs' phenomenon (see Part 1, Sec. 1.2). It can be easily shown that the frequency response of a rectangular time domain window follows a $(\sin x)/x$ curve, and the equivalent holds for the time response of the frequency domain rectangular window (remember Eq. 6.5.7). Since we have deliberately truncated the system's frequency response at the N^{th} sample, we can think of it as being a product of an infinitely long (but finite density) spectrum with a rectangular frequency window. This results in a convolution of the system impulse response with the $(\sin x)/x$ function in the time domain, hence the form of the error in Fig. 6.5.7 to 6.5.10.

This error can be lowered by taking the transform of a longer frequency response vector, but can never be eliminated. For example, if `N=2048` and we specify the frequency vector as: `w=(0:1:N-1)/8`, the error will be 8 times lower than in the case of a frequency vector with `N=256`, but the calculation will last more than 23 times longer (the number of multiplications is proportional to $N\log_2 N$ or 11 times, in addition to 8 times the number of sums and other operations).

As we have stated at the beginning, Sec. 6.0, our aim is to make quick comparisons of the performance of several systems, and on the basis of those decide which system suits us best. Since the resolution of the computer VGA type of graphics is more than adequate for this purpose, and the response error in the case of `N=256` can be seen only if compared with the exact response (and even so as only a one pixel difference), the extra time and memory requirements do not justify the improvement.

A better way of calculating the response more accurately and also directly from system poles and zeros is described in the next section.

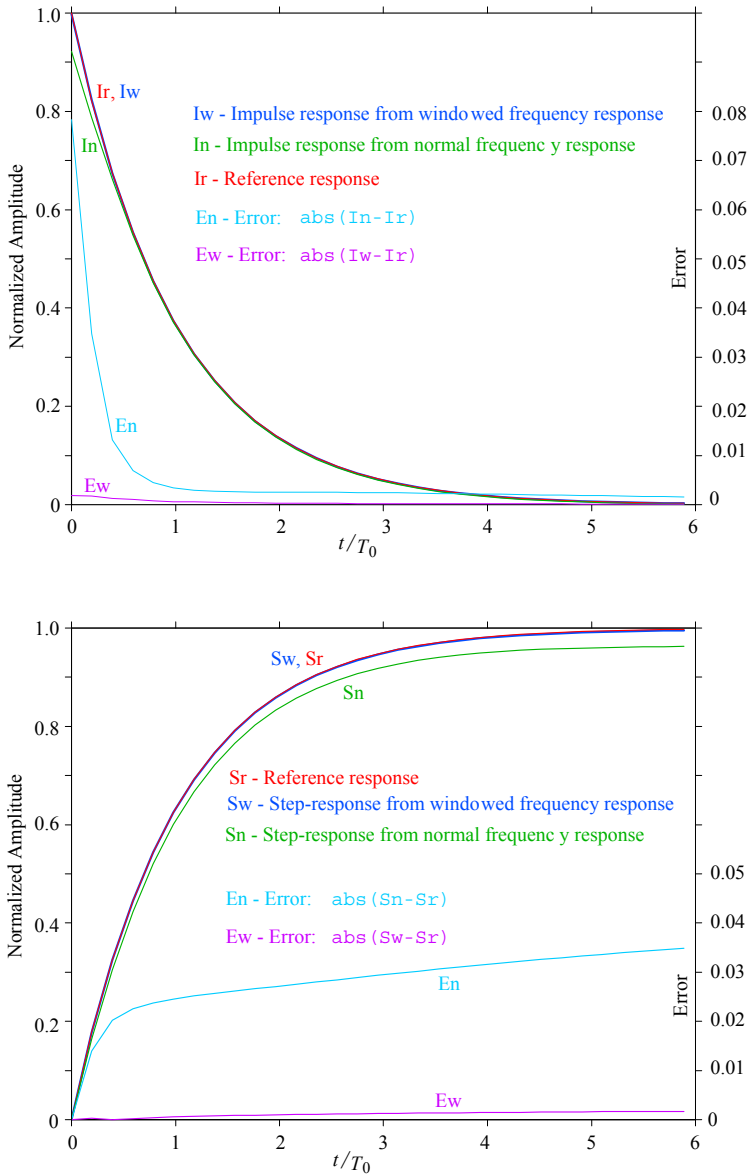


Fig. 6.5.5 and 6.5.6: The first 30 points of a 256 sample long 1st-order impulse and step response vs. the analytically calculated references. The error plots En and Ew are enlarged 10×. Although the impulse response, calculated from the normal frequency response, has a relatively small error, it integrates to an unacceptably high value (4%) in the step response. In contrast, by windowing the frequency response, both time domain errors are much lower, the step response final value is in error by less than 0.2%.

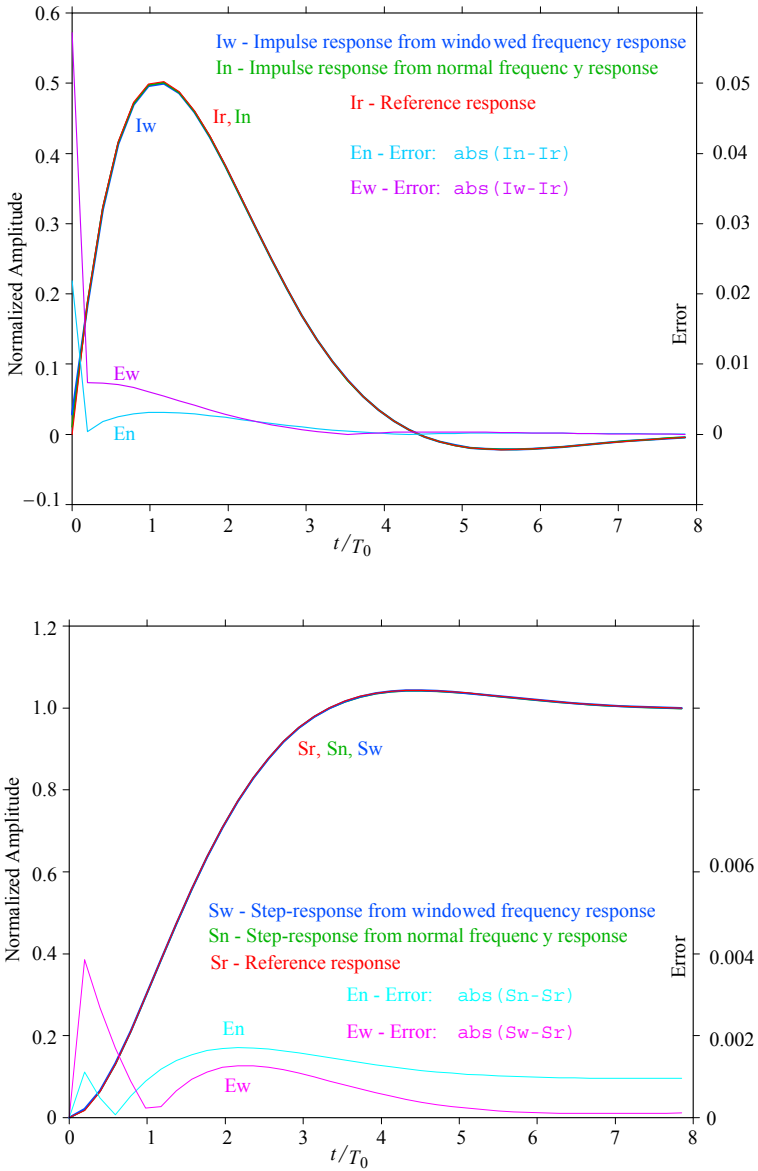


Fig. 6.5.7 and 6.5.8: As in Fig. 6.5.5 and 6.5.6, but with 40 samples of a 2nd-order Butterworth system. The impulse response error for the windowing procedure is higher at the beginning, but falls off more quickly, therefore the step response final value error is still much lower (note that the step response error plots are enlarged 100×). The oscillations in error plots, owed to the Gibbs' effect, also begin to show.

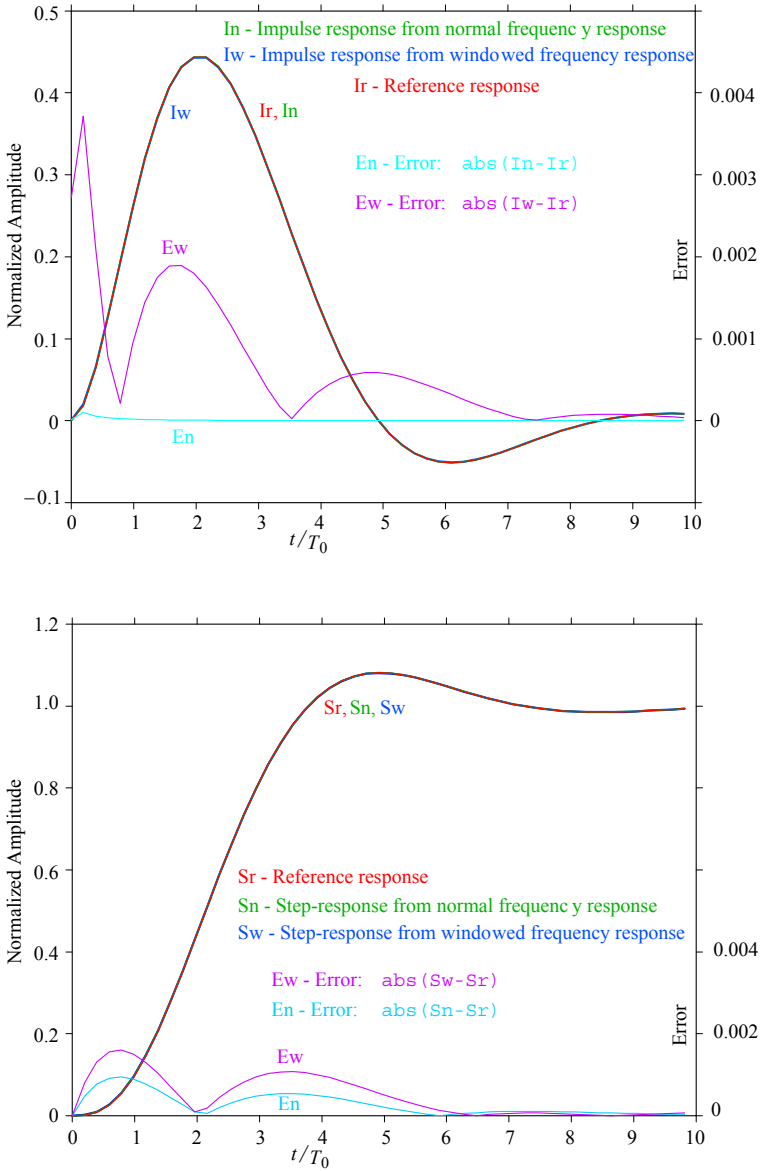


Fig. 6.5.9 and 6.5.10: As in Fig. 6.5.5–8, but with 50 samples of a 3rd-order Butterworth system. Windowing does not help any longer and produces even greater error. The dominant error is now owed to the Gibbs' effect.

6.5.7 Code Execution Efficiency

The TRESP routine executes surprisingly fast. Back in 1987, when these Matlab routines were developed and the first version of this text was written, I was using a 12 MHz PC with an i286-type processor, an i287 math coprocessor, and EGA type of graphics (640×400 resolution, 16 colors). To produce the 10 responses of Fig.6.5.11, starting from the system order, finding the system coefficients, extracting the poles, calculating the complex frequency response, running FFT to obtain the impulse response, integrating for the step response and finally plotting it all on a screen, that old PC worked less than 12 seconds. Today (March 2000), a 500 MHz Pentium-III based processor does it in a few tens of milliseconds (before you can release the ENTER key, once you have pressed it; although, it takes a lot more time for Matlab working under Windows to open the graph window). And note that we are talking about floating point, 'double precision' arithmetic! Nevertheless, being able to make fast calculations has become even more important for embedded instrumentation applications, which require real time data processing and adaptive algorithms.

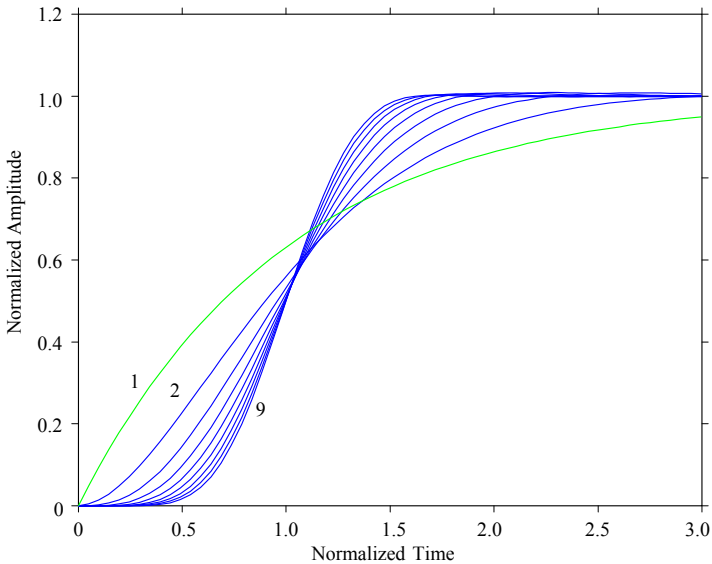


Fig. 6.5.11: Step responses of Bessel–Thomson systems (normalized to equal envelope delay), of order 2 to 9, including the 1st-order step response for comparison. Note the half amplitude delay approaching 1 and the bandwidth improving (shorter rise time) as the system order increases. The TRESP algorithm execution speed was tested by creating this figure.

6.6 Transient Response From Residues

The method of calculating the transient response by FFT, presented in Sec. 6.5, has several advantages over other algorithms. The most important ones are high execution speed, the possibility of computing from either a calculated complex frequency response or from a measured magnitude–phase relationship, and the use of the same FFT algorithm to work both time–to–frequency and frequency–to–time.

Its main disadvantage is the error resulting from the Gibbs' effect, which distorts the most interesting part of the time domain response. This error, although small, can sometimes prevent the system designer from resolving or identifying the cause of possible second-order effects that are spoiling the measured or simulated system performance to which the desired ideal response is being compared. In such a case the designer must have a firm reference, which should not be an approximation in any sense.

The algorithm, presented in this section, with the name ATDR (an acronym of 'Analytical Time Domain Response'), calculates the impulse and step responses by following the same analytical method, that has been used extensively in the previous parts of this book. The routine calculates the residues at each system transfer function pole, and then calculates the final response at specified time instants. However, the residues are not calculated by an actual infinitesimally limiting process, so it is not possible to apply this routine in the most general case (e.g., it fails for systems with coincident poles), but this restriction is not severe, since all of the optimized system families are covered properly. Readers who would like to implement a rigorously universal procedure can obtain the residues calculated by the somewhat longer RESIDUE routine in Matlab.

In contrast to the FFT method, whose execution time is independent of system complexity, this method works more slowly for each additional pole or zero.

A nice feature of this method is that the user has a direct control over the time vector: the response is calculated at exactly those time instants which were specified by the user. This may be important when making comparison with a measured response of an actual system prototype.

As we have seen in numerous examples, solved in the previous parts, a general expression for a residue at a pole p_k of an n^{th}-order system specified by Eq. 6.1.10 can be written like this:

$$r_k = \lim_{p \to p_k} (p - p_k) \, \frac{\prod_{i=1}^{n} (-p_i)}{\prod_{i=1}^{n} (p - p_i)} \cdot \frac{\prod_{j=1}^{m} (p - z_j)}{\prod_{j=1}^{m} (-z_j)} \, \mathrm{e}^{p_k t} \qquad (6.6.1)$$

Here n is the number of poles, m is the number of zeros, p is a vector whose elements are the system's poles p_i, p_k is the k^{th} pole for which the residue r_k is calculated, z_j are the zeros, whilst i and j are the indices.

The terms $(p - p_k)$ cancel for each $i = k$ before limiting. If we now make $p = p_k$, without using the limiting process, the general applicability of Eq. 6.6.1 is lost, but for all optimized system families (no coincident poles!) this will still be valid.

In the ATDR routine we form a vector `Z` of length n, containing the products over the index j of $(p_k - z_j)$ — one (k^{th}) element of `Z` for each residue — and divide these by the product of all transfer function zeros (if there are any).

Next, we form a matrix `D` of n by n elements, each element being a product of $(p - p_i)$. The elements on the diagonal of `D` will all have zero value (on the diagonal $p = p_i$), and since we need the products of the remaining terms, we must eliminate them to avoid multiplying by zero. This results in a `D` of $(n - 1)$ rows by n columns matrix.

In Matlab most matrix operations are designed to perform on columns, producing a single-row vector of results. The same is true for the PROD ('product') command, so `prod(D)` returns a row of n products performed over each column of `D`. This row is returned in `D`, then `Z` is divided by `D`, element by element. The resultant vector is multiplied by the product of all poles to produce the correct scaling factors of the residues, returned in the vector `P`.

If a step response is required, `P` must be divided by the poles p, element by element. Finally, each element of `P` multiplies a vector of (complex) exponential functions of the time vector multiplied by the k^{th} pole.

All the residue values at the same time point are summed (in rows) and each sum is then an element of the real valued result vector (the complex parts cancel to better than 10^{-14} and are neglected). In the case of the step response all values must be increased by 1, the value of the residue of the additional pole at the complex plane's origin resulting from the input step function transform operator $1/s$.

For the impulse response case there are two options: either the result is left as it is, representing a response (implicitly) normalized in amplitude to the response of the same system excited by the ideal (infinite amplitude, infinitely narrow) input impulse, or it can be normalized to represent a unity gain system by dividing each response value by the sum of all values. This is desirable when calculating convolution, etc., but then we have to specify the time vector as sufficiently long (in comparison with the dominant system time constants), to allow the impulse response to decay to a value close enough to zero, thus avoiding a system gain error.

Here is how the now familiar 5$^{\text{th}}$-order Butterworth system responses can be calculated using the ATDR routine:

```
[z,p]=buttap(5);        % 5th-order Butterworth zeros and poles
t=(0:1:300)/15;         % 301-point t-vector, 15 samples in one t-unit
I=atdr(z,p,t,'i');      % ideal impulse response
S=atdr(z,p,t,'s');      % step response
plot(t,I,t,S)           % plot I and S against t
```

The resulting plot should look the same as in Fig. 6.5.1 (but with a much better accuracy!).

```
function y=atdr(z,p,t,q)
%ATDR Analytical Time Domain Response by simplified residue calculus
%      (does not work for systems with multiple poles).
%      y=atdr(z,p,t) or
%      y=atdr(z,p,t,'n') returns the normalized impulse response of a
%                        unity gain system, specified by zeros z and
%                        poles p in time t.
%      y=atdr(z,p,t,'i') returns the impulse response, denormalized to
%                        the ideal impulse input.
%      y=atdr(z,p,t,'s') returns the step response of the system.
%
%      Specify the time as :  t=(0:1:N-1)/T, where N is the number of
%      desired time domain samples and T is the number of samples in
%      the time scale unit, i.e.:  t=(0:1:200)/10

% Author :  Erik Margan, 891008, Free of copyright !

if nargin==3
      q='n' ;      % by default, return the unity gain impulse response
end
n=max(size(p));    % find the number of poles
for k=1:n          % test for repeating poles
      P=p;
      P(k)=[];     % exclude the pole currently tested
      if all(abs(P-p(k)))==0  % is there another such pole ?
            error('ATDR cannot handle systems with repeating poles!')
      end
end
dc=1;              % set low pass system flag
if isempty(z)
      Z=1;         % no zeros
else
      % zeros
      if any( abs(z) <1e-6 )
            dc=0; % HP or BP system, clear dc flag
      end
      if all( abs( real( z ) ) < 1e-6 )
            z = j * imag( z ) ;      % all zeros on imaginary axis
      end
      Z=ones(size(p)) ;
      if dc
            for k=1:n
                  Z(:,k)=prod(p(k)-z)/prod(-z);
            end
      else
            for k = 1:np
                  for h = 1:nz
                        if z(h) == 0
                              Z(k,:) = Z(k,:)*p(k) ;
                        else
                              Z(k,:) = Z(k,:)*(p(k)+z(h))/z(h) ;
                        end
                  end
            end
      end
      Z=Z(:);      % column-wise orientation
end
if n == 1
      D=1;         % single pole case
else
      for k = 1:n
            d=p(k)-p;    % difference, column orientation
            d(k)=[];     % k-th element = 0, eliminate it
```

```
            D(:,k)=d;   % k-th column of D
      end
      if n > 2
            D=(prod(D)); % make column-wise product if D is a matrix
      end
      D=D.';      % column-wise orientation
end
P=prod(-p)*Z./D;  % impulse residues
if q == 's'
      P=P./p;     % if step response is required, divide by p
end
t=t(:).';         % time vector, row orientation
y=P(1)*exp(p(1)*t);     % response, first row
for k = 2:n
      y=[y; P(k)*exp(p(k)*t)];      % next row
      y=sum(y);   % sum column-wise, return a single row
end
y=real(y);        % result is real only (imaginary parts cancel)
if (q == 's') & ( isempty(z) | dc == 1 )
      y=y+1;      % if step resp., add 1 for the pole at 0+j0
end
if ( q == 'i' | q == 'n' ) & ( dc == 0 )
      y=-diff([0, y]);  % impulse response of a high pass system
end
if q == 'n'
      y=y/abs(sum(y));  % normalize impulse resp. to unity gain
end
```

6.7 A Simple Application Example

The algorithms which we have developed allow us now to make a quick comparison of the performance of two equal bandwidth 5th-order systems, a Butterworth and a Bessel–Thomson system. We shall compare the pole loci, the magnitude and the step response. Let us first calculate and plot the poles:

```
[z1,p1]=buttap(5);      % a 5-th order Butterworth zeros and poles
[z2,p2]=bestap(5,'n'); % a 5-th order Bessel system zeros and poles
p1=2*pi*1000*p1;        % denormalize the poles to 1kHz
p2=2*pi*1000*p2;
                        % plot the imag-vs.-real part of poles
plot( real(p1),imag(p1),'*r', real(p2),imag(p2),'*b' )
axis equal square ;     % set axes aspect ratio 1:1
```

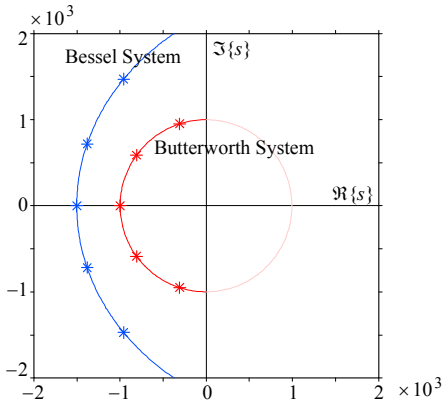


Fig. 6.7.1: The Butterworth poles (on the unit cycle) and the Bessel–Thomson poles (on the fitted ellipse). Note that for the same bandwidth (1 kHz) the values of Bessel–Thomson poles are much larger, but with a lower ratio of the imaginary to the real part.

Let us calculate and plot the frequency responses:

```
f=logspace(2,4,401);    % a log-spaced frequency vector 10^2 - 10^4 Hz
F1=freqw(z1,p1,2*pi*f);% Butterworth frequency response
F2=freqw(z2,p2,2*pi*f);% Bessel-Thomson frequency response
                        % plot the dB magnitude vs. log-frequency :
semilogx(f/1000,20*log10(abs(F1)),'-r',f/1000,20*log10(abs(F2)),'-b')
ylabel('Magnitude'), xlabel('f [kHz]')
```

The frequency response plots are shown in Fig. 6.7.2. Note the equal pass band (−3 dB point) and equal slope at high frequencies. However, the Butterworth system attenuation is an order of magnitude (20 dB) better.

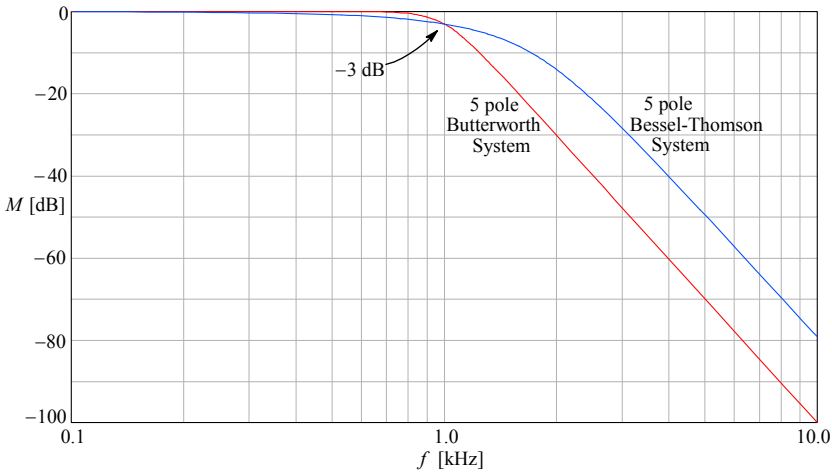


Fig. 6.7.2: Frequency responses of the Butterworth and Bessel–Thomson system. For an equal cut off frequency ($f_h = 1$ kHz), the Butterworth system stop band attenuation is about an order of magnitude (10× or 20 dB) better than that of the Bessel–Thomson.

Using the same poles and the ATDR routine, we compare the step responses:

```
t=(0:1e-5:3);           % the 3ms time vector, 100 samples/ms.
y1=atdr(z1,p1,t,'s');   % Butterworth step response
y2=atdr(z2,p2,t,'s');   % Bessel-Thomson step response
plot(t*1000,y1,'-r',t*1000,y2,'-b'), xlabel('t [ms]') % see Fig.6.7.3
```

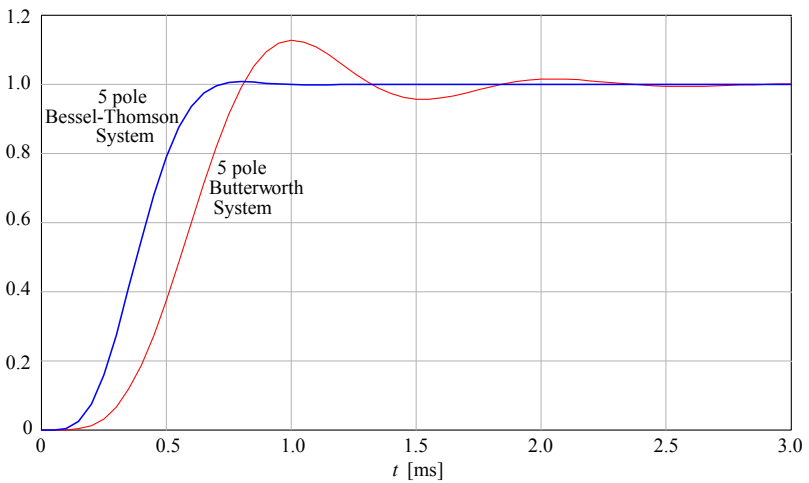


Fig. 6.7.3: Step responses of the Butterworth and Bessel–Thomson system. For the same cut off frequency (1 kHz) the Bessel–Thomson system's delay is smaller; the overshoot is only 0.4% and there is no ringing, so settling down to 0.1% occurs within the first 1 ms. Although the rise times are nearly equal, the Butterworth system is a poor choice if time domain performance is required, since it settles down to 0.1% only after some 5 ms (but Chebyshev and Elliptic filter systems are even much worse in this respect).

Résumé of Part 6

The algorithms shown are small, simple, easy to use, and fast in execution. They are ideal for starting the system's design from scratch, to specify the design goals, as well as to provide a reference with which a realized prototype can be compared.

We have shown how the system performance can easily be evaluated by using the routines developed for Matlab, the prediction of system time domain response in particular. We also hope that the development and application examples of these routines offer a deeper insight on how the system should be designed as a whole.

Still, the reader as the future system's designer is being let down at the most demanding task of finding the circuitry and hardware that will perform as required, and engineering experience is the only help here. This book should help to understand how it might be possible to push the bandwidth up, smooth the transient, and reduce the settling time. But there are also many other important parameters which must be carefully considered when designing an amplifier, such as noise, linearity, electrical and thermal stability, output power, slew rate limiting, the time it takes to recover from overdrive, etc.

However, these parameters (with the exception of electrical stability) are in most cases independent of the system pole and zero locations, but are strongly influenced by the circuit's topology and by the type of active devices used for the realization.

Once the design goals have been set and the circuit configuration selected, performance verification and iterative finalization can then be done using one of the many CAD/CAE programs available on the market.

To see the numerical convolution routine and calculation examples and an actual amplifier–filter system design example calculated using the algorithms developed so far, please turn to Part 7.

References:

[6.1] *J.N. Little*, *C.B. Moller*, PC–MATLAB For Students (containing disks with Matlab program), Prentice–Hall, 1989

[6.2] MATLAB–V For Students (containing a disk with Matlab program), Prentice–Hall, 1998

[6.3] The MathWorks, Inc., <http://www.mathworks.com/>

[6.4] *Oliver Heaviside*, Electromagnetic Theory, Chelsea Pub. Co., 3rd edition, 1971, ISBN: 082840237X

[6.5] *Oliver Heaviside*, Electrical Papers Edition Volume 1, American Mathematical Society; January 1970, ASIN: 0821828401

[6.6] *W. A. Atherton*, Pioneers: Oliver Heaviside — Champion of inductance, Wireless World, August 1987, pp. 789–790

[6.7] *Paul J. Nahin*, Oliver Heaviside: The Life, Work, and Times of an Electrical Genius of the Victorian Age, Johns Hopkins Univ. Pr., October 2002, ISBN: 0801869099

[6.8] *Douglas H. Moore*, Heaviside Operational Calculus; An Elementary Foundation, American Elsevier Pub. Co., ASIN: 0444000909

[6.9] *H. Nyquist*, <http://www.ieee.org/organizations/history_center/legacies/nyquist.html>

[6.10] *H.W. Bode*, <http://www.ieee.org/organizations/history_center/legacies/bode.html>

[6.11] *S. Butterworth*, On the Theory of Filter–Amplifiers, Experimental Wireless & The Wireless Engineer, Vol. 7, 1930, pp.536–541

[6.12] *W. E. Thomson*, Networks With Maximally Flat Delay, Wireless Engineer, Vol. 29, October 1952, pp.256–263

[6.13] *L. Storch*, Synthesis of Constant–Time–Delay Ladder Networks Using Bessel Polynomials, Proceedings of the I.R.E., Vol. 42, 1954, pp.1666–1675

[6.14] *O. Föllinger*, Laplace– und Fourier–Transformation, AEG Telefunken (Abt. Verlag), 1982

[6.15] *M. O'Flynn*, *E. Moriarty*, Linear Systems: Time–Domain and Transform Analysis, J. Wiley & Sons, 1987

[6.16] *J.W. Cooley & J.W. Tukey*, An Algorithm for the Machine Calculation of Complex Fourier Series, Math. of Computation, Vol. 19, No. 90, April 1965, pp. 297–301

[6.17] Special Issue on the Fast Fourier Transform, IEEE Transactions on Audio & Electroacoustics, Vol. AU-15, June 1967

[6.18] *E.O. Brigham*, The Fast Fourier Transform, Prentice–Hall, 1974

[6.19] *A.V. Oppenheim*, *R.W. Schafer*, Digital Signal Processing, Prentice–Hall, 1975.

[6.20] *R.I. Ross*, Evaluating the Transient Response of a Network Function, Proceedings of the IEEE, May 1967, pp.693–694.

[6.21] *R.I. Ross*, Iterative Transient Response Calculation Procedures that have Low Storage Requirement, presented at the Second International Symposium on Network Theory in Herceg-Novi, 1972

[6.22] *J. Vlach*, *K. Singhal*, Computer Methods for Circuit Analysis and Design, Van Nostrand Reinhold, 1983

[6.23] *E. Margan*, Calculating the Transient Response, Elekrotehniški vestnik, Ljubljana, ELVEA2 58, 1991, 1, pp.11–22

P. Starič, E. Margan:

Wideband Amplifiers

Part 7:

Algorithm Application Examples

Any computer program can be reduced by at least one command line.

Any computer program has at least one command line with an error.

...

Any computer program can eventually be reduced
to a single command line, having at least one error.

(Conservative extrapolation of Murphy's Law to computer programming)

Contents:

List of Figures:

List of Routines:

7.0 Introduction

In Part 6 we have developed a few numerical algorithms that will serve us as the basis of system analysis and synthesis. We have shown how simple it is to implement the analytical expressions related to the various aspects of system performance into compact, fast executing computer code which reduces the tedious mathematics to pure routine. Of course, a major contribution to this easiness was provided by the programming environment, a high level, maths–oriented language called Matlab™ (Ref. [7.1]).

As wideband amplifier designers, we want to be able to accurately predict amplifier performance, particularly in the time-domain. With the algorithms developed, we now have the essential tools to revisit some of the circuits presented in previous parts, possibly gaining a better insight into how to put them to use in our new designs eventually.

But the main purpose of Part 7 is to put the algorithms in a wider perspective. Here we intentionally use the term 'system', in order to emphasize the high degree of integration present in modern electronics design, which forces us to abandon the old paradigm of adding up separately optimized subsystems into the final product; instead, the design process should be conceived to optimize the total system performance from the start. As more and more digital processing power is being built in to modern products, the analog interface with the real world needs to be given adequate treatment on the system level, so that the final product eventually becomes a successful integration of both the analog and the digital world.

Now, we hear some of you analog circuit designers asking in a low voice "why do we need to learn any of this digital stuff?" The answer is that digital engineers would have a hard time learning the analog stuff, so there would be no one to understand the requirements and implications of a decent AD or DA interface. On the other hand, for an analog engineer learning the digital stuff is so simple, almost trivial, and it pays back well with better designs, and it acquires for you the respect due from fellow digital engineers.

7.1 Using Convolution: Response to Arbitrary Input Waveforms

7.1.1 From Infinitesimal to Discrete Time Integration

The time-domain algorithms that we have developed in Part 6 gave us the system response to two special cases of input signal: the unit-area impulse and the unit-amplitude step. Here we will consider the response to any type of input signal, provided that its application will not exceed neither the input nor the output system capabilities. In technical literature this is known as the BIBO-condition[1]. And, of course, we are still within the constraints of our initial LTIC-conditions[2].

As we have seen in Part 1, Sec. 1.14, the system's time domain response to an arbitrary input signal can be calculated in two ways:

a) by transforming the input signal to complex frequency domain, multiplying it by the system transfer function and transforming the result back to the time domain;

b) directly in time domain by the convolution integral.

A short reminder of the convolution integral definition and the transcription from differential to difference form is in order here. Let $x(t)$ be the time domain signal, presented to the input of a system being characterized by its impulse response $f(t)$. The system output can then be calculated by convolving $x(t)$ with $f(t)$:

$$y(t) = \int_{t_0}^{t_1} f(\tau - t)\, x(t)\, dt \tag{7.1.1}$$

where τ is a fixed time constant, its value chosen so that $f(t)$ is time reversed. Usually, it is sufficient to make τ large enough to allow the system's impulse response $f(t)$ to completely relax and reach the steady state again (not just the first zero-crossing point!) with a tolerance of some 0.01 % or less.

If $x(t)$ was applied to the system at t_0, then this can be the lower limit of integration. Of course, the time scale can always be renormalized so that $t_0 = 0$. The upper integration limit, labeled t_1, can be wherever needed, depending on how much of the input and output signal we are interested in.

Now, in Eq. 7.1.1 dt is implicitly approaching zero, so there would be an infinite number of samples between t_0 and t_1. Since our computers have a limited amount of memory (and we have a limited amount of time!) we must make a compromise between the sampling rate and the available memory length and adjust them so that we cover the signal of interest with enough resolution in both time and

[1] Bounded input → bounded output. This property is a consequence of our choice of basic mathematical assumptions; since our math tools were designed to handle an infinite amount of infinitesimal quantities, BIBO is the necessary condition for convergence. However, in the real analog world, we are often faced with UBIBO requirements (unbounded input), i.e., our instrumentation inputs must be protected from overdrive. Interestingly, the inverse of BIBO is in widespread use in the computer world, in fact, any digital computer is a GIGO type of device (garbage in → garbage out; unbounded!).

[2] Linearity, Time Invariance, Causality. Although some engineers consider oscillators to be 'acausal', there is always a perfectly reasonable cause why an amplifier oscillates, even if we fail to recognize it at first.

amplitude. So if M is the number of memory bytes reserved for $x(t)$, the required sampling time interval is:

$$\Delta t = \frac{t_1 - t_0}{M} \tag{7.1.2}$$

Then, if Δt replaces dt, the integral in Eq. 7.1.1 transforms into a sum of M elements, $x(t)$ and $y(t)$ become vectors `x(n)` and `y(n)`, where `n` is the index of a signal sample location in memory, and $f(\tau - t)$ becomes `f(m-n)`, with `m=length(f)`, resulting in:

$$\texttt{y(n)} = \sum_{\texttt{n=1}}^{\texttt{M}} \texttt{f(m-n)*x(n)} \tag{7.1.3}$$

Here Δt is implicitly set to 1, since the difference between two adjacent memory locations is a unit integer. Good book-keeping practice, however, recommends the construction of a separate time scale vector, with values from t_0 to t_1, in increments of Δt between adjacent values. All other vectors are then plotted against it, as we have seen it done in Part 6.

7.1.2 Numerical Convolution Algorithm

In Part 1 we have seen that solving the convolution integral analytically can be a time consuming task, even for a skilled mathematician. Sometimes, even if $x(t)$ and $f(t)$ are analytic functions, their product need not be elementarily integrable in the general case. In such cases we prefer to take the $\mathcal{L}$ transform route; but this route can sometimes be equally difficult. Fortunately numerical computation of the convolution integral, following Eq. 7.1.3 , can be programmed easily:

```
function y=vcon(f,x)
%VCON   Convolution, step-by-step example. See also CONV and FILTER.
%
%     Call :      y=vcon(f,x);
%     where:      x(t) --> the input signal
%                 f(t) --> the system impulse response
%                 y(t) --> the system response to x(t) by convolving
%                          f(t) with x(t).
%     If  length(x)=nx  and  length(f)=nf, then length(y)=nx+nf-1.

%     Erik Margan, 861019, Last editing 890416; Free of copyright!

                        % force f to be the shorter vector :
if length(f) > length(x)
      xx=x; x=f; f=xx;  % exchange x and f via xx
      clear xx
end
nf=length(f);           % get the number of elements in x and f
nx=length(x);
f=f(:).';               % organize x and f as single-row vectors
x=x(:).';
y=zeros(2,nx+nf-1);     % form a (2)-by-(nx+nf-1) matrix y, all zeros
y(1,1:nx)=f(1)*x;       % first row of y: multiply x by f(1)
for k=2:nf              % second row: multiply and shift (insert 0)
      y(2, k-1:nx+k-1)=[0, f(k)*x];
                        % sum the two rows column-wise and
      y(1,:)=sum(y);    % put result back into first row
end                     % repeat for all remaining elements of f;
y=y(1,:);               % the result is the first row only.
```

To get a clearer view of what the VCON routine is doing, let us write a short numerical example, using a 6-sample input signal and a 3-sample system impulse response, and display every intermediate result of the matrix y in VCON:

```
x=[0 1 3 5 6 6];     f=[1 3 -1];     y=vcon(x,f);

% initialization - all zeros, 2 rows, 6+3-1 columns:
    0     0     0     0     0     0     0     0
    0     0     0     0     0     0     0     0
% step 1: multiply x by the first sample of f, f(1)=1 and
% insert it into the first row:
    0     1     3     5     6     6     0     0
    0     0     0     0     0     0     0     0
% step 2: multiply x by the second sample of f, f(2)=3,
% shift it one place to the right by adding a leading zero and
% insert it into the second row:
    0     1     3     5     6     6     0     0
    0     0     3     9    15    18    18     0
% step 3: sum both rows vertically, put the result in the first row
    0     1     6    14    21    24    18     0
    0     0     3     9    15    18    18     0
% iterate steps 2 and 3, each iteration using the next sample of f:
    0     1     6    14    21    24    18     0
    0     0     0    -1    -3    -5    -6    -6

    0     1     6    13    18    19    12    -6
    0     0     0    -1    -3    -5    -6    -6
% after 2 iterations (because f is only 3 samples long)
% the result is the first row of y:
    0     1     6    13    18    19    12    -6
% actually, the result is only the first 6 elements:
    0     1     6    13    18    19
% since there are only 6 elements in x, the process assumes the rest
% to be zeros. So the remaining two elements of the result represent
% the relaxation from the last value (19) to zero by the integration
% of the system impulse response f.

% Basically, the process above does the following:
% (note the reversed sequence of f)

                0  1  3  5  6  6
          -1  3  1
(↓*) ==>        0                 (→+) ==> 0  --> y(1)

                0  1  3  5  6  6
             -1  3  1
(↓*) ==>        0  1              (→+) ==> 1  --> y(2)

                0  1  3  5  6  6
               -1  3  1
(↓*) ==>        0  3  3           (→+) ==> 6  --> y(3)

                0  1  3  5  6  6
                  -1  3  1
(↓*) ==>        0 -1  9  5        (→+) ==> 13  --> y(4)

% ......... etc.
```

For convolution Matlab has a function named CONV, which uses a built in FILTER command to run substantially faster, but then the process remains hidden from the user; however, the final result is the same as with VCON. Another property of Matlab is the matrix indexing, which starts with 1 (see the lower limit of the sum

symbol in Eq. 7.1.3), in contrast to most programming languages which use memory 'pointers' (base address + offset, the offset of the array's first element being 0).

7.1.3 Numerical Convolution Examples

Let us now use the VCON routine in a real life example. Suppose we have a gated sine wave generator connected to the same 5th-order Butterworth system which we inspected in detail in Part 6. Also, let the Butterworth system's half power bandwidth be 1 kHz, the generator frequency 1.5 kHz, and we turn on the gate in the instant the signal crosses the zero level. From the frequency response calculations, we know that the forced response amplitude (long after the transient) will be:

```
[z,p]=buttap(5);                      % Butterworth 5-pole system
Aout=Ain*abs(freqw(z,p,1500/1000));   % output steady-state amplitude
```

Here `p` are the poles of the normalized 5th-order Butterworth system, `z` is an empty matrix (no zeros); the signal frequency is normalized to the system's cut off.

But how will the system respond to the signal's 'turn on' transient? We can simulate this using the algorithms we have developed in Part 6 and VCON:

```
fh=1000;                % system half-power bandwidth, 1kHz
fs=1500;                % input signal frequency, 1.5kHz
t=(0:1:300)/(50*fh);    % time vector, 20us delta-t, 6ms range
nt=length(t);

[z,p]=buttap(5);        % 5th-order Butterworth system
p=2*pi*fh*p;            % denormalized system poles
Ir=atdr(z,p,t,'n');     % system impulse-response

d=25;                   % switch-on delay, 25 time-samples
                        % make the input signal :
x=[zeros(1,d), sin(2*pi*fs*t(1:nt-d))];

y=vcon(Ir,x);           % convolve x with Ir ;

A=nt/(2*pi*fh*max(t)); % denormalize amplitude of Ir for plot

                        % plot the input, the system impulse response
                        % and the convolution result :
plot( t*fh, x, '-g', ...
      t*fh, [zeros(1,d), Ir(1:nt-d)*A],'-r', ...
      t*fh, y(1:nt), '-b')
xlabel('t [ms]')
```

The convolution result, compared to the input signal and the system impulse response, is shown in Fig. 7.1.1.

Note that we have plotted only the first `nt` samples of the convolution result; however, the total length of `y` is `length(x)+length(Ir)-1`, or one sample less that the sum of the input signal and the system response lengths. The first `length(x)=nt` samples of `y` represent the system's response to `x`, whilst the remaining `length(Ir)-1` samples are the consequence of the system relaxation: since there are no more signal samples in `x` after the last point `x(nt)`, the convolution assumes that the input signal is

zero and calculates the system relaxation from the last signal value. This is equivalent to a response caused by an input step from `x(nt)` to zero. So if we are interested only in the system response to the input signal, we simply limit the response vector to the same length as was the input signal. Also, in the general case the length of the system's impulse response vector, `nr=length(Ir)`, does not have to be equal to the input signal vector length, `nx=nt`. In practice, we often make `nr<<nx`, but `Ir` should be made long enough to include the system relaxation to a level very close to zero, as only then will the sum of all elements of `Ir` not differ much from the system gain.

There is, however, an important difference in the plot and the calculation, that must be explained. The impulse response which we obtained from Butterworth system poles was normalized to represent a unity gain system, since we want to see the frequency dependence on the output amplitude by comparing the input and output signals. Thus our system should either have a gain of unity, or the output should be normalized to the input in some other way (e.g., if the gain is known, we could either divide the output signal by the gain, or multiply the input signal). But the unity gain normalized impulse response would be too small in amplitude, compared to the input signal, so we have plotted the ideal impulse response.

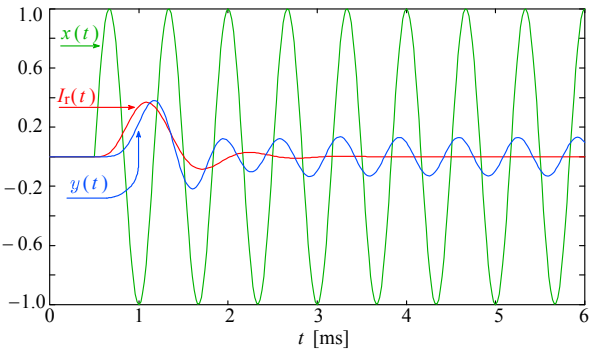


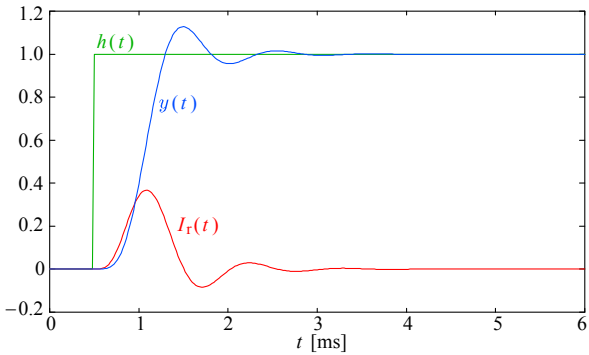
Fig. 7.1.1: Convolution example: response $y(t)$ to a sine wave $x(t)$ switched-on into a 5th-order Butterworth system, whose impulse-response is $I_r(t)$, shown here as the ideal response (instead of unity gain); both are delayed by the same switch–on time (0.5 ms). The system responds by phase shifting and amplitude modulating the first few wave periods, reaching finally the forced ('steady state') response.

Can we check whether our routine works correctly?

Apart from entering some simple number sequences as before, we can do this by entering an input signal for which we have already calculated the result in a different way, say, the unit step (see Fig. 6.5.1, Part 6). By convolving the impulse response with the unit step, instead of the sine wave, we should obtain the now known step response:

```
% continuing from above:
h=[zeros(1:d), ones(1:nt-d)];      % h(t) is the unit step function
y=vcon(Ir,h);                      % convolve h with Ir
plot( t*fh, h, '-g', ...
      t*fh, [zeros(1,d), Ir(1:nt-d)*A],'-r', ...
      t*fh, y(1:nt), '-b')
xlabel('t [ms]')
```

The resulting step response, shown in Fig. 7.1.2, should be identical to that of Fig. 6.5.1, Part 6, neglecting the initial 0.5 ms (25 samples) time delay and the different time scale:



sFig. 7.1.2: Checking convolution: response $y(t)$ of the 5th-order Butterworth system to the unit step $h(t)$. The system's impulse response $I_r(t)$ is also shown, but with its ideal size (not unity gain). Apart from the 0.5 ms (25-samples) time delay and the time scale, the step response is identical to the one shown in Part 6, Fig.6.5.1.

We can now revisit the convolution integral example of Part 1, Sec. 1.15, where we had a unit-step input signal, fed to a two-pole Bessel-Thomson system, whose output was in turn fed to a two-pole Butterworth system. The commands in the following window simulate the process and the final result of Fig. 1.15.1. But this time, let us use the TRESP (Part 6) routine for the frequency to time domain transform. See the result in Fig. 7.1.3 and compare it to Fig. 1.15.1g.

```
[z1,p1]=bestap(2,'t'); % Bessel-Thomson 2nd-order system poles
[z2,p2]=buttap(2);     % Butterworth 2nd-order system poles
N=256;                 % number of samples
m=4;                   % set the bandwidth factor
w=(0:1:N-1)/m;         % frequency vector, w(m+1)=1 ;
F1=freqw(p1,w);        % Bessel-Thomson system frequency response
F2=freqw(p2,w);        % Butterworth system frequency response

[S1,t]=tresp(F1,w,'s');% step-response of the Bessel-Thomson system
I2=tresp(F2,w,'u');    % unity-gain Butterworth impulse response ;
                       % both have the same normalized time vector
d=max(find(t<=15));    % limit the plot to first 15 time units
I2=I2(1:d);            % limit the I2 vector length to d

     % convolution of Bessel-Thomson system step-response with
     % the first d points of the Butterworth impulse response :
y=vcon(I2,S1);

A=N/(2*pi*m*max(t));   % amplitude denormalization for I2
     % plot first d points of all three responses vs. time :
plot( t(1:d), S1(1:d), '-r',...
      t(1:d), I2(1:d)*A, '-g',...
      t(1:d),  y(1:d), '-b' )
xlabel('Time [s]')
```

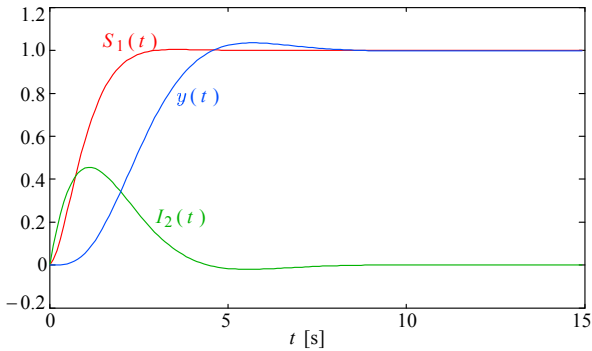


Fig. 7.1.3: Convolution example of Part 1, Sec. 1.15. A Bessel–Thomson 2-pole system step response $S_1(t)$ has been fed to the 2-pole Butterworth system and convolved with its impulse response $I_2(t)$, resulting in the output step response $y(t)$. Compare it with Fig. 1.15.1g.

The VCON function is a lengthy process. On a 12 MHz AT-286 PC, which was used for the first experiments back in 1986–7, it took more than 40 s to complete the example shown above, but even with today's fast computers there is still a noticeable delay. The Matlab CONV routine is much faster.

The reader might question the relevance of the total calculation time, since, for the purpose of a circuit design aid, anything below 1 s should be acceptable (this is comparable with the user's reaction time, when making a simple go/no go assessment of the result). However, imagine an automated optimization program loop, adjusting the values of some 10–20 circuit components. Such a loop might take hundreds or even thousands of executions before reaching satisfactory performance criteria, so a low routine time would be welcome. Moreover, if the routine will eventually be implemented in hardware to acquire and process the signal in real time, a low routine time is of vital importance. For example, in order to continuously process a 16 bit stereo audio stream, divided into 1024 sample chunks, using a 32 word long filter impulse response, the total routine calculation time must be less than 250 µs.

In some cases, particularly with long signal sequences (`N > 1000`), it could be interesting to take the Fourier transform route, numerically.

Here is an example using a signal recorded by a nuclear magnetic resonance imaging (MRI) system. The MRI RF signal is very weak (< 1 mV), so the detection is noisy and there is some interference from another source. We shall try to clean it by using a 5th-order Bessel-Thomson digital filter with a unity gain and a 1 MHz cut off:

```
load R.dat % load the recorded signal from a file "R.dat"
N=length(R);            % total vector length, N=2048 samples
Tr=102.4e-6;            % total record time 102.4 us
dt=Tr/N;                % sampling time interval, 50 ns
t=dt*(0:1:N-1);         % time vector reconstruction

% plot the first 1200 samples of the recorded signal
plot(t(1:1200,R(1:1200),'-g')
xlabel('Time [\mus]')  % input signal R, fist 60us, see Fig.7.1.4
G=fft(R);              % G is the FFT spectrum of R
G=G(1:N/2);            % use only up to the Nyquist freq. ( 10 MHz )
```

```
f=(1:1:N/2)/dt;          % frequency vector reconstructed

[z,p]=bestap(5,'n');     % 5th-order Bessel filter poles
p=p*2*pi*1e+6;           % half-power bandwidth is 1 MHz
F=freqw(z,p,2*pi*f);     % filter frequency response

% multiplication in frequency is equal to convolution in time:
Y=F.*G;                  % output spectrum

x=max(find(f<=8e+6));    % plot spectrum up to 8 MHz
M=max(abs(G));           % normalize the spectrum to its peak value
plot( f(1:x), abs(F(1:x)),    '-r', ...
      f(1:x), abs(G(1:x))/M, '-g', ...
      f(1:x), abs(Y(1:x))/M, '-b' )
xlabel('Frequency [MHz]'), ylabel('Normalized Magnitude')
                         % see Fig.7.1.5

y=2*(real(fft(conj(Y)))-1)/(N/2);  % return to time domain

a=max(find(t<=5e-5));
b=min(find(t>=20e-6));
plot( t(a:b), g(a:b), '-g', t(a:b), y(a:b), '-b' )
xlabel('Time [\mus]')    % see Fig.7.1.6
```

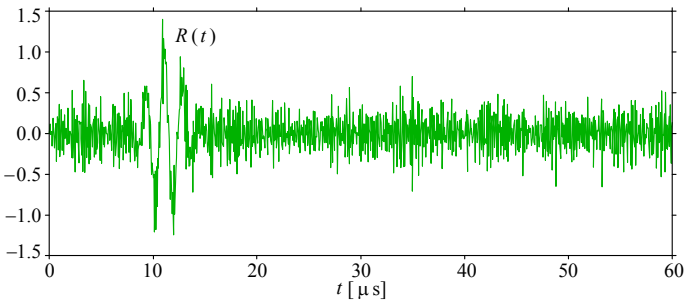


Fig. 7.1.4: Input signal example used for the spectral-domain convolution example (first 1200 samples of the 2048 total record length)

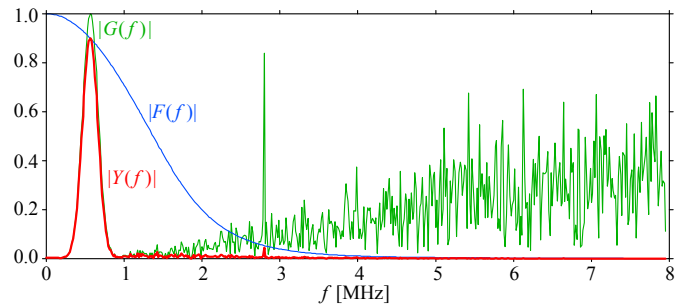


Fig. 7.1.5: The spectrum $G(f)$ of the signal in Fig. 7.1.4a is multiplied by the system's frequency response $F(f)$ to produce the output spectrum $Y(f)$. Along with the modulated signal centered at 560 kHz, there is a strong 2.8 MHz interference from another source and a high level of white noise (rising with frequency), both being substantially reduced by the filter.

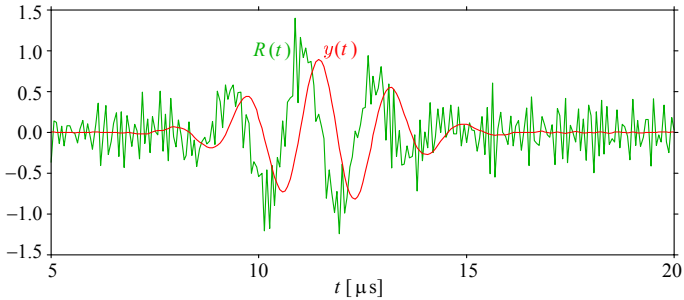


Fig. 7.1.6: The output spectrum is returned to time domain as $y(t)$ and is compared with the input signal $R(t)$, in expanded time scale. Note the small change in amplitude, the reduced noise level and the envelope delay (approx. $1/4$ period time shift), with little change in phase. The time shift is equal to $1/2$ the number of samples of the filter impulse response.

Fig. 7.1.6 illustrates the dramatic improvement in signal quality that can be achieved by using Bessel filters.

In MRI systems the test object is put in a strong static magnetic field. This causes the nucleons of the atoms in the test object to align their magnetic spin to the external field. Then a short RF burst, having a well defined frequency and duration, is applied, tilting the nucleon spin orientation perpendicular to the static field (this happens only to those nucleons whose resonant frequency coincides with that of the RF burst).

After the RF burst has ceased, the nucleons gradually regain their original spin orientation in a top-like precession motion, radiating away the excess electromagnetic energy. This EM radiation is picked up by the sensing coils and detected by an RF receiver; the detected signal has the same frequency as the excitation frequency, both being the function of the static magnetic field and the type of nucleons. Obviously the intensity of the detected radiation is proportional to the number of nucleons having the same resonant frequency[3].

In addition, since the frequency is field dependent a small field gradient can be added to the static magnetic field, in order to split the response into a broad spectrum. The shape of the response spectral envelope then represents the spatial density of the specific nucleons in the test object. By rotating the gradient around the object the recorded spectra would represent the 'sliced view' of the object from different angles. A computer can be used to reconstruct the volumetric distribution of particular atoms through a process called 'back-projection' (in effect, a type of spatial convolution).

From this short description of the MRI technique it is clear that the most vital parameter of the filter, applied to smooth the recorded signal, is its group delay flatness. Only a filter with a group delay being flat well into the stop band will be able to faithfully deliver the filtered signal, preserving its shape both in the time and the frequency domain, and Bessel–Thomson filters are ideal in this sense. Consequently a sharper image of the test object is obtained.

[3]The 1952 Nobel prize for physics was awarded to *Felix Bloch* and *Edward Mills Purcell* for their work on nuclear magnetic resonance; more info at <http://nobelprize.org/physics/laureates/1952/>.

7.2 System Front–End Design Consideration

7.2.1 General Remarks

The trend in modern instrumentation has been definitely going digital from the 1970s, benefiting from cheap microprocessor technology, being implemented as early as possible in the signal processing chain. Likewise, reverting back to analog occurs only if absolutely necessary and as late as possible. The key features are precision and repeatability of measurements, and those properties are often called upon to justify the sacrificing of other system properties (of those, the system bandwidth is no exception, in fact, it is the first victim in most cases!). In spite of this digital 'tyranny', analog engineers were (for now, at least) able to cope quite successfully with it. Actually, over the years they have managed to stay well in front of both demands and expectations.

Another key word of modern technology is miniaturization, and that in connection with ever-lowering power consumption. So the current technological front is concentrated on the integration of analog and digital functions on the same IC chip, using the lowest possible supply voltage and applying clever power management schemes in both hardware and software.

Of course, digitalization is causing many restrictions as well. In contrast with analog continuous time systems, digital systems operate in discrete time, on the transitions of the system clock. And, since there is only a limited amount of memory, which also has a finite access time, the sampling window is ever shrinking. As if this were not enough (and in spite of promoting precision!), digital systems are not very flexible to upgrade; for example, to change from an 8 bit to a 12 bit system, an analog circuit would have to be improved by 2^4 or 16 times (if it was not that good already!); in contrast, the digital part would have to be redesigned completely, not just changed. This is because an increased number of gates and flip-flops change state at each clock transition, increasing the supply current peaks; also the circuit area is increased, increasing the length of the interconnections. Both facts increase the RF interference and the possibility of noise injection back into the delicate analog input stage.

In Part 6, Sec. 6.5 we have already learned a few basic facts about the effects of signal sampling. We know that a finite sampling density means only that the signal repeats in time, so the length of the sampling window should be chosen in accordance with the signal length and the sampling frequency. More difficult to handle is the problem of finite word length, since it sets the effective system resolution and the conversion noise level.

7.2.2 Aliasing Phenomena in Sampling Systems

To illustrate further the use of the algorithms developed in Part 6, let us consider the design requirements of a front–end amplifier driving an analog to digital converter (ADC). In theory the minimum required amplifier bandwidth should be equal to the Nyquist frequency, which is one half of the ADC's sampling clock frequency, $f_N = f_c/2$. In practice, however, undistorted reconstruction of a periodic waveform can be achieved only if the signal content above the Nyquist frequency has been attenuated to levels lower than the ADC resolution. This is known in literature as the ***Shannon*'s sampling theorem** [Ref. 7.2].

The purpose of filtering the signal above the Nyquist frequency is to avoid **'aliasing'**. Fig. 7.2.1 shows a typical situation resulting in a signal frequency alias in relation to the sampling clock frequency.

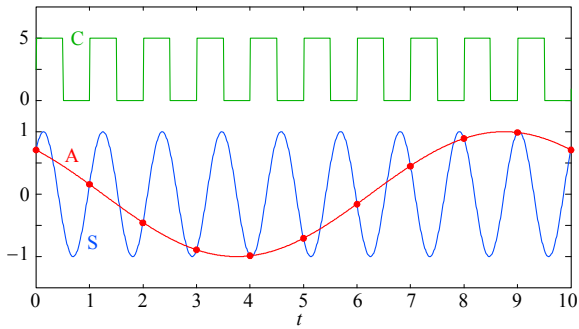


Fig. 7.2.1: Aliasing (frequency mirroring). A high frequency signal S, sampled by an ADC at each rising edge of the clock C of a comparably high frequency, can not be distinguished from its low frequency alias A, which is equal to the difference between the clock and signal frequency, $f_a = f_s - f_c$. In this figure, $f_s = (9/10)\, f_c$, therefore $f_a = -(1/10)\, f_c$ (Yes, a **negative** frequency! This can be verified by increasing the clock frequency very slightly and watch the aliased signal apparently moving backwards in time).

The alias frequency f_a is simply a difference between the signal frequency f_s and the sampling clock's frequency f_c:

$$f_a = f_s - f_c \tag{7.2.1}$$

Aliasing can be best understood if we recall a common scene in Western movies, where the wheels of the stage coach seem to be rotating backwards, while the horses are being whipped to run wild to escape from the desperados behind. The perceived frequency of rotation of the wheel is equal to the **difference** between the actual rotation frequency and the frequency at which the pictures were taken.

A wheel, rotating at the cycle frequency f_w equal to the picture rate f_p (or its integer multiple or sub-multiple, $f_w = n\, f_p/m$, where m is the number of wheel arms), would be perceived as stationary. Likewise, if an ADC's sampling frequency is equal to the signal frequency (see Fig. 7.2.2), the apparent result is a DC level.

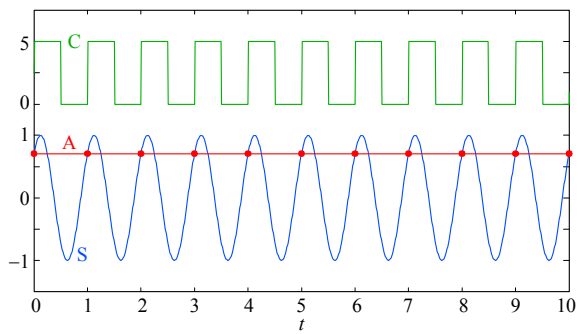


Fig. 7.2.2: Alias of a signal equal in frequency to the sampling clock looks like a DC.

Furthermore, a signal with a frequency slightly higher than the sampling frequency could not be distinguished from a low frequency equal to the difference of the two, as in Fig. 7.2.3. Experienced Hi-Fi enthusiasts and car mechanics will surely remember seeing this, if we remind them of the 'stroboscope effect'.

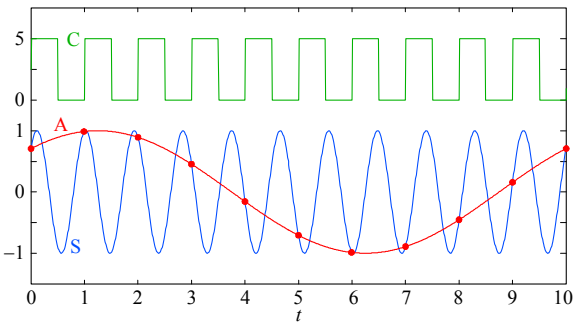


Fig. 7.2.3: A signal frequency slightly higher than the sampling frequency aliases into a low frequency, equal to the difference of the two (but now positive).

Here is the ALIAS routine for Matlab, by which we can calculate the aliasing for any clock and signal frequency desired (assuming infinitely short aperture time).

```
function fa=alias(fs,fc,phi)
% ALIAS calculates the alias frequency of a sampled sinewave signal.
%       Call :   fa = alias( fs, fc, phi ) ;
%       where:   fs is the signal frequency
%                fc is the sampling clock frequency
%                phi is the initial signal phase shift

% Erik Margan, 920807, Free of copyright!

if nargin < 3
     phi = pi/3 ;   % signal phase shift re. clk, arbitrary value
end
ofs = 2 ;           % clock offset
A = 1/ofs ;         % clock amplitude
m = 100 ;           % signal reconstruction factor is equal to
                    % the number of dots within a clock period
N = 1 + 10 * m ;    % total number of dots
dt = 1 / ( m * fc ) ;   % delta-t for time reconstruction
t = dt * ( 0 : 1 : N ) ;    % time vector

fa = fs - fc ;      % alias frequency (can be negative!)
clk = ofs + A * sign( sin( 2 * pi * fc *t ) ) ; % clock
sig = sin( 2 * pi * fs * t + phi ) ;     % sampled signal
sal = sin( 2 * pi * fa * t + phi ) ;     % alias signal

plot( t, clk, '-g',...
      t, sig, '-b',...
      t, sal, '-r',...
      t(1:m:N), sig(1:m:N), 'or')
xlabel( 't' )
```

Of course, the sampled signal is more often than not a spectrum, either discrete or continuous, and aliasing applies to a spectrum in the same way as to discrete frequency signals. In fact, the superposition theorem applies here, too.

We have noted that a sampled spectrum is symmetrical about the sampling frequency, because a signal, sampled by a clock with exactly the same frequency, aliases as a DC level, which in turn depends on the initial signal phase shift relative to the clock. However, something odd happens already at the Nyquist frequency, as can be seen in Fig. 7.2.4 and Fig. 7.2.5. In both figures the signal frequency is equal to the Nyquist frequency ($1/2$ the sampling frequency), but differs in phase. Although the correct alias signal is equal in amplitude to the original signal, we perceive an amplitude which varies with phase.

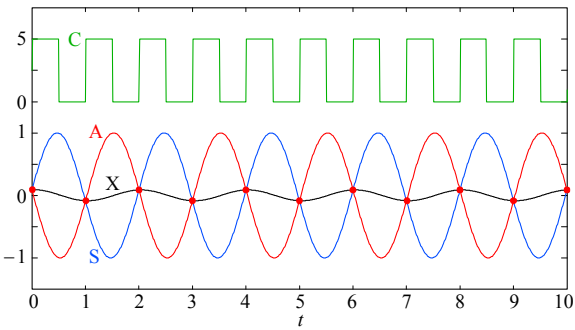


Fig. 7.2.4: When the signal frequency is equal to the Nyquist frequency, there are two samples per period and the correct alias signal is of the same amplitude as the original signal. However, the perceived alias amplitude is a function of the phase difference between the signal and the clock. A 10° phase shift results in a low apparent amplitude, as shown by the X waveform.

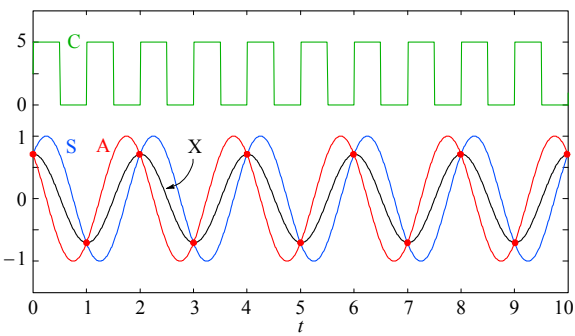


Fig. 7.2.5: Same as in Fig. 7.2.4, but with a 45° phase shift. The apparent amplitude of X is now higher.

In fact, if our ADC were to be sampling a slowly sweeping sinusoidal signal, its spectral envelope would follow the $(\sin \omega T_s)/\omega T_s$ function, shown in Fig. 7.2.6, with the first zero at the Nyquist frequency, the second zero at the sampling frequency and so on, a zero at every harmonic of the Nyquist frequency.

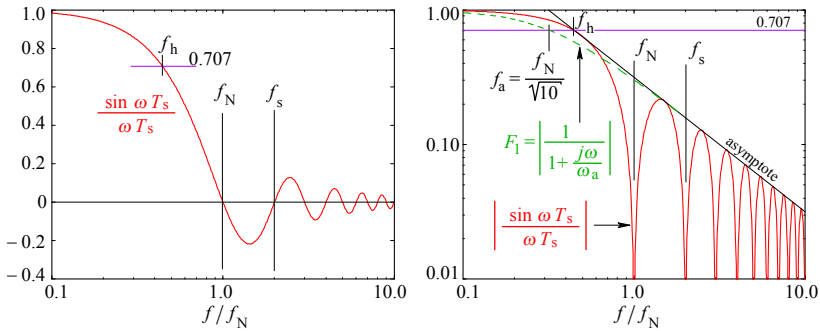


Fig. 7.2.6: The spectrum resulting from sampling a constant amplitude sinusoidal signal varying in frequency from $0.1 f_N$ to $10 f_N$ follows the $(\sin \omega T_s)/\omega T_s$ function, where $T_s = 1/f_s$. The function is shown in the linear vertical scale on the left and in the log of the absolute value on the right. The first zero occurs at the Nyquist frequency, the second at the sampling frequency and so on, at every Nyquist harmonic. Note that the effective sampling bandwidth f_h is reduced to about $0.43\, f_N$. The asymptote is the same as for a simple RC low pass filter, $-20\,\text{dB}/10f$ with a cut off at $f_a = f_N/\sqrt{10}$.

The aliasing amplitude follows this same $(\sin \omega T_s)/\omega T_s$ function, from the Nyquist frequency up. An important side effect is that the bandwidth is reduced to about $0.43\, f_N$. This can be taken into account when designing an input anti-aliasing filter and partially compensate the function pattern below the Nyquist frequency by an adequate peaking.

7.2.3 Better Anti-Aliasing With Mixed Mode Filters

By the term 'mixed mode filter' we mean a combination of analog and digital filtering which gives the same result as a single filter having the same total number of poles. The simplest way to understand the design requirements and optimization, as well as the advantages of such an approach, is by following an example.

Let us imagine a sampling system using an ADC with a 12 bit amplitude resolution and a 50 ns time resolution (sampling frequency $f_s = 20\,\text{MHz}$). The number of discrete levels resolved by 12 bits is $A = 2^{12} = 4096$; the ADC relative resolution level is simply $1/A$, or in dB, $a = 20 \log_{10}(1/A) = -72\,\text{dB}$. According to the Shannon sampling theorem the frequencies above the Nyquist frequency ($f_N = f_s/2 = 10\,\text{MHz}$) must be attenuated by at least 2^{-12} to reduce the alias of the high frequency spectral content (signal or noise) below the ADC resolution.

As we have just learned, the $(\sin \omega T_s)/\omega T_s$ function of the alias spectrum allows us to relax the filter requirements by some 4 bits (a factor of 4.5 or 13 dB) at the frequency $0.7\, f_s$; for a while, we are going to neglect this, leaving it for the end of our analysis.

Let us also assume a 4 V peak to peak ADC input signal range and let the maximum required vertical amplifier sensitivity be 5 mV/division. Since oscilloscope displays usually have 8 vertical divisions, this means 40 mV of input for a full scale display, or a gain of 100. We would like to achieve the required gain–bandwidth product with either a two- or a three-stage amplifier. We shall assume a 5-pole filter

for the two-stage amplifier (a 3-pole and a 2-pole stage), and a 7-pole filter for the three-stage amplifier (one 3-pole stage and two 2-pole stages). We shall also inspect the performance of a 9-pole (four-stage) filter to see if the higher bandwidth (achieved as a result of a steeper cut off) justifies the cost and circuit complexity of one additional amplifier stage.

Now, if our input signal was of a square wave or pulse form, our main requirement would be a shortest possible ADC 'aperture' time and an analog bandwidth as high as possible. Then we would be able to recognize the sampled waveform shape even with only 5 samples per period. But suppose we would like to record a transient event having the form of an exponentially decaying oscillating wave, along with lots of broad band noise, something like the signal in Fig. 7.1.4. To do this properly we require both aliasing suppression of the spectrum beyond the Nyquist frequency and preserving the waveform shape; the later requirement limits our choice of filter systems to the Bessel–Thomson family.

Finally, we shall investigate the possibility of improving the system bandwidth by filtering the recorded data digitally.

We start our calculations from the requirement that any anti-alias filter must have the attenuation at the Nyquist frequency f_N equal to the ADC resolution level. Since we know that the asymptote attenuation slope depends on the system order n (number of poles) as $n \times 20\,\mathrm{dB}/10f$, we can follow those asymptotes from f_N back to the maximum signal level; the crossing point then defines the system cut off frequency f_{hn} for each of the three filter systems.

Since we do not have an explicit relation between the Bessel–Thomson filter cut off and its asymptote, we shall use Eq. 6.3.10 for Butterworth systems to find the frequency f_a at which the 5-, 7-, and 9-pole Butterworth filter would exhibit the $A = 2^{12}$ attenuation required. By using $f_h = f_N^2/f_a$ we can then find the Butterworth cut off frequencies. Then by using the modified Bessel–Thomson poles (those that have the same asymptote as the Butterworth system of comparable order) we can find the Bessel–Thomson cut off frequencies which satisfy the no-aliasing requirement.

```
A=2^12;     % ADC resolution limit sets the required attenuation
fs=2e+7;    % ADC sampling frequency, 20MHz
fN=fs/2;    % Nyquist frequency, 10MHz
M=1e+6;     % megahertz scale-factor

% the normalized 5-, 7- and 9-pole system asymptotes, all crossing
% the ADC resolution limit, 1/A, at fN, after Eq.6.3.10, will have
% the following cutoff frequencies :

fh5=fN/10^(log10(A^2-1)/(2*5));
fh7=fN/10^(log10(A^2-1)/(2*7));
fh9=fN/10^(log10(A^2-1)/(2*9));
disp(['fh5 = ', num2str(fa5/M), ' MHz'])
disp(['fh7 = ', num2str(fa7/M), ' MHz'])
disp(['fh9 = ', num2str(fa9/M), ' MHz'])
% the disp commands return the following values :

» fh5 = 1.8946 MHz
» fh7 = 3.0475 MHz
» fh9 = 3.9685 MHz
```

We now find the poles and the system bandwidth of the 5-, 7-, and 9-pole Bessel–Thomson systems, which have their responses normalized to the same asymptotes as the above Butterworth systems of equal order:

```
N=601;                  % number of frequency samples
f=fN*logspace(-2,0,N); % length-N frequency vector, from 2 decades
                        % below fN to fN, in log-scale
w=2*pi*f;               % angular frequency

[z5,p5]=bestap(5,'a'); % Bessel-Thomson asymptote-normalized systems
[z7,p7]=bestap(7,'a');
[z9,p9]=bestap(9,'a');

p5=p5*2*pi*fa5;         % Scaling-up the poles by the previously
p7=p7*2*pi*fa7;         % calculated cutoff frequencies, so that all
p9=p9*2*pi*fa9;         % three responses have 1/A attenuation at fN

M5=20*log10(abs(freqw(p5,w))); % Calculate magnitudes in dB ;
M7=20*log10(abs(freqw(p7,w)));
M9=20*log10(abs(freqw(p9,w)));
                        % plot magnitudes in dB vs. log frequency
db3=-3.0103;            % the -3dB reference level
                        % and the ADC resolution limit
semilogx( f/M, M5, '-r',...
          f/M, M7, '-g',...
          f/M, M9, '-b',...
          fN*[0.05, 0.35]/M, [db3, db3], '-k'...
          [f(1), f(N)]/M, [1/A, 1/A], '-k' )
xlabel( 'f [MHz]' ) ;
ylabel( 'Attenuation [dB]' )
```

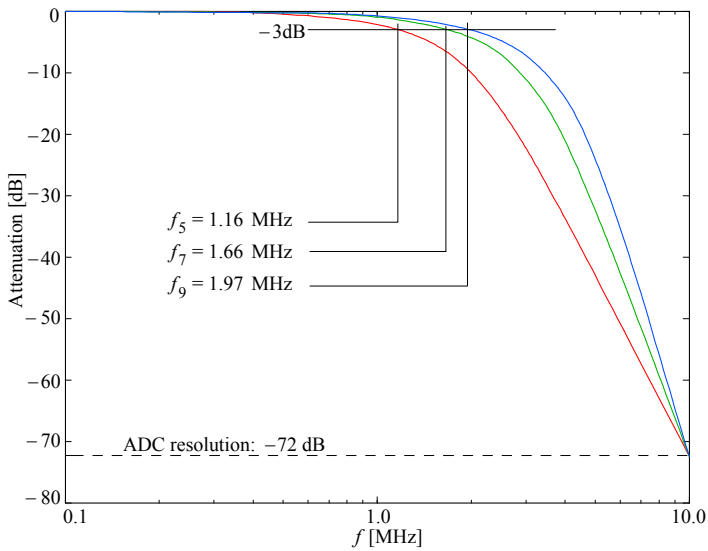


Fig. 7.2.7: Magnitude vs. frequency of Bessel–Thomson 5-, 7-, and 9-pole systems, normalized to the attenuation of 2^{-12} (−72 dB) at f_N (10 MHz).

Fig. 7.2.7 shows the frequency responses, calculated to have the same attenuation, equal to the relative ADC resolution level of −72 dB, at the Nyquist frequency. We now need their approximate −3 dB cut off frequencies:

```
m=abs(M5-3.0103);       % compare the magnitudes with the -3dB level
x5=find(m==min(m));     % and find the index of each frequency limit

m=abs(M7-3.0103);
x7=find(m==min(m));

m=abs(M9-3.0103);
x9=find(m==min(m));

[f5, f7, f9]=f([x5, x7, x9]);       % find the cutoff frequencies

% display cutoff frequencies of the Bessel-Thomson systems :
disp(['f5 = ', num2str(f5/M), ' MHz'])
disp(['f7 = ', num2str(f7/M), ' MHz'])
disp(['f9 = ', num2str(f9/M), ' MHz'])
» f5 = 1.166 MHz
» f7 = 1.660 MHz
» f9 = 1.965 MHz
```

Note that these values are much lower then the cutoff frequencies of the asymptotes, owing to the more gradual roll-off of Bessel–Thomson systems. Also, note that a greater improvement in performance is achieved by increasing the system order from 5 to 7 then from 7 to 9. We would like to have a confirmation of this fact from the step responses (later, we shall also see how these step responses would look when sampled at the actual sampling time intervals).

```
fs=2e+7;          % sampling frequency
t=(0:1:500)/fs;   % time vector (to calculate the rise times we need
                  % a much finer sampling then the actual 50 ns

S5=atdr(z5,p5,t,'s');         % Step responses
S7=atdr(z7,p7,t,'s');
S9=atdr(z9,p9,t,'s');

% plot the step responses
% and the 0.1 and 0.9 reference levels to compare the rise times :
x10=t(50,130), x90=t(150,300), y10=[0.1,0.1], y90=[0.9,0.9];

plot(t,S5,'-r', t,S7,'-g', t,S9,'-b', x10,y10,'-k', x90,y90,'-k' )
xlabel( 't [us]' )

% calculate the rise times :
x5a=find( abs(S5-y10) == min( abs(S5-y10) ) );
x5b=find( abs(S5-y90) == min( abs(S5-y90) ) );
x7a=find( abs(S7-y10) == min( abs(S7-y10) ) );
x7b=find( abs(S7-y90) == min( abs(S7-y90) ) );
x9a=find( abs(S9-y10) == min( abs(S9-y10) ) );
x9b=find( abs(S9-y90) == min( abs(S9-y90) ) );
Tr5=t(x5b)-t(x5a);
Tr7=t(x7b)-t(x7a);
Tr9=t(x9b)-t(x9a);
```

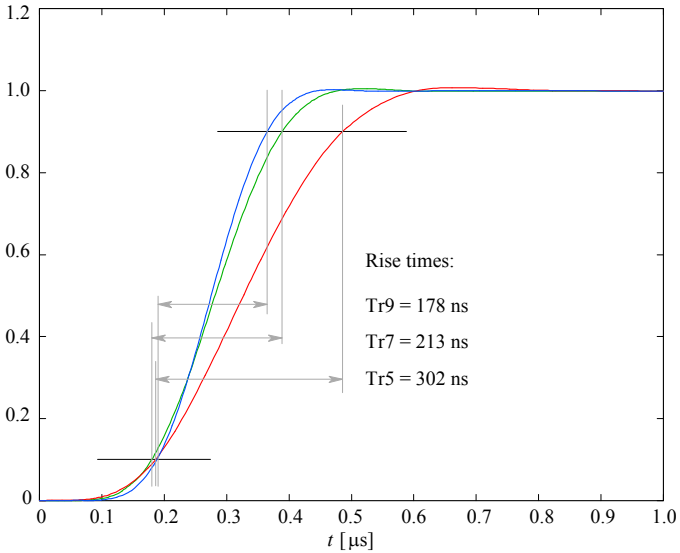


Fig. 7.2.8 : Step responses of the 5-, 7-, and 9-pole Bessel–Thomson systems, having equal attenuation at the Nyquist frequency. The rise times are calculated from the number of samples between the 10% and 90% of the final value of the normalized amplitude.

We see that in this case the improvement from order 7 to order 9 is not so high as to justify the added circuit complexity and cost of one more amplifying stage.

So let us say we are temporarily satisfied with the 7-pole filter system. However, its 1.66 MHz bandwidth for a 12 bit ADC, sampling at 20 MHz, is simply not good enough. Even the 1.96 MHz bandwidth of the 9-pole system is rather low. The question is whether we can find a way around the limitations imposed by the anti-aliasing requirements?

Most ADC recording systems do not have to show the sampled signal in real time. To the human eye a screen refreshing rate of 10 to 20 times per second is fast enough for most purposes. Also many systems are intentionally made to record and accumulate large amounts of data to be reviewed later. So on most occasions there is plenty of time available to implement some sort of signal post–processing.

We are going to show how a digital filter can be combined with the analog anti-aliasing filter to expand the system bandwidth beyond the aliasing limit without increasing the sampling frequency.

Suppose we could implement some form of digital filtering which would suppress the alias spectrum below the ADC resolution and then we ask ourselves what would be the minimum required pass band attenuation of such a filter. The answer is simple: the filter attenuation must follow the inverse of the alias spectrum envelope. But if we were to allow the spectrum around the sampling frequency to alias, our digital filter would need to extend its attenuation characteristic back to DC. Certainly this is neither practical nor desirable. Therefore since $f_s = 2f_N$, our bandwidth improvement factor, let us call it B, must be lower than 2.

So let us increase the filter cut off by $B = 1.73$; the input spectrum would then contain frequencies only up to Bf_N, which would alias back down to $f_s - Bf_N$, in this case $2 - 1.73 = 0.27 f_N$. This frequency is high enough to allow the realization of a not too demanding digital filter.

Let us now study the shape of the alias spectrum which would result from taking our original 7-pole analog filter, denoted by F_{7o}, and pushing it up by the chosen factor B to F_{7b}, as shown in Fig. 7.2.9. The spectrum S_A between f_N and Bf_N is going to be aliased below the Nyquist frequency into S_B.

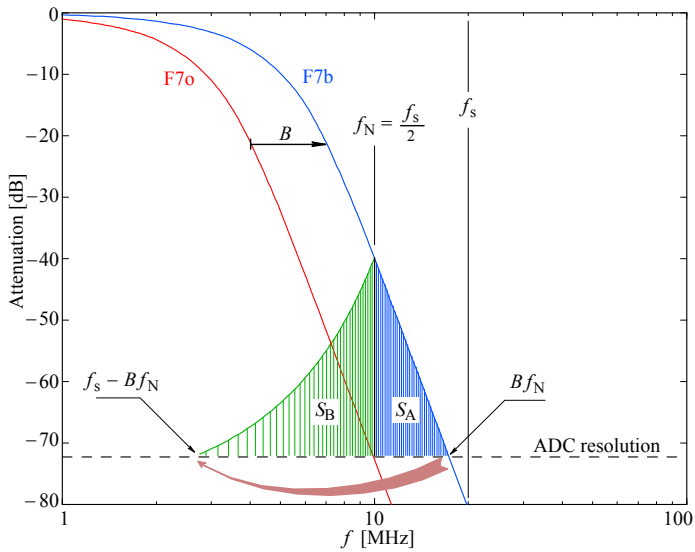


Fig. 7.2.9: Alias spectrum of a 7-pole filter with a higher cut off frequency. F_{7o} is our original 7-pole Bessel-Thomson analog filter, which crosses the 12-bit ADC resolution level of -72 dB at exactly the Nyquist frequency, $f_N = f_S/2 = 10$ MHz. This guaranties freedom from aliasing, but the bandwidth is rather low. If we move it upwards by a factor $B = 1.73$ to F_{7b}, the spectrum S_A will alias into S_B. Note the alias spectrum inversion: f_N remains in its place, whilst Bf_N is aliased to $f_s - Bf_N$. Note also that the alias spectral envelope has changed in comparison with the original: in the log-log scale plot a linearly falling spectral envelope becomes curved when aliased. This change of the spectral envelope is important, since it will allow us to use a relatively simple filter response shape to suppress the aliased spectrum below the ADC resolution.

Note that in the log frequency scale the aliased spectrum envelope is not linear, even if the original one is (as defined by the attenuation characteristic of the analog filter).

If we flip the spectrum S_B up, as in Fig. 7.2.10, the resulting spectral envelope, denoted by F_{rq}, represents the minimal attenuation requirement of a digital filter, which would restore freedom from aliasing.

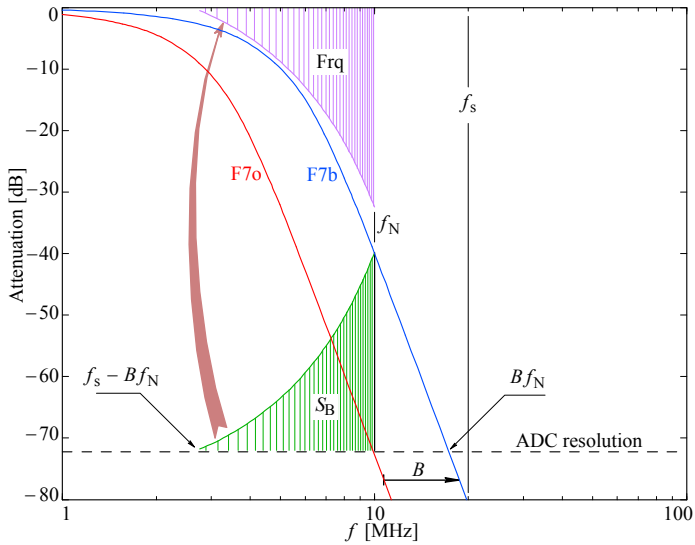


Fig. 7.2.10: If we invert the alias spectrum S_B the envelope of the resulting spectrum F_{rq} represents the minimum attenuation requirement that a digital filter should have in order to suppress the aliased spectrum below the ADC resolution.

The following procedure shows how to calculate and plot the various elements of Fig. 7.2.9 and Fig. 7.2.10, and F_{rq} in particular, starting from the previously calculated 7-pole Bessel–Thomson system magnitude F_{7o}:

```
     % the following calculation assumes a log frequency scale and
     % a linear in dB attenuation.
A=2^12;                 % number of levels resolved by a 12-bit ADC
a=20*log10(1/A);        % ADC resolution, -72dB
Nf=601;                 % number of frequency samples
f=logspace(6,8,Nf);     % frequency vector, 1 to 100 MHz
B=1.73;                 % chosen bandwidth increase (max. 1.8)
     % the original 7-pole filter magnitude crosses a at fN :
F7o=20*log10(abs(freqw(p7, 2*pi*f)));
     % F7o shifted up by B to F7b :
F7b=20*log10(abs(freqw(p7*B, 2*pi*f)));

fA=B*fN;                % F7b crosses ADC resolution (a) at fA
xn=min(find(f>=fN));    % index of fN in f
xa=min(find(f>=fA));    % index of fA in f
Sa=F7a(xn:xa);          % source of the alias spectrum
fa=f(xn:xa);            % frequency band of Sa
Sb=F7a(xa:-1:xn);       % the alias spectrum, from fs-fA to fN
fb=fs-f(xa:-1;xn);      % frequency band of Sb
Frq=a-Sa;               % min. required dig.filt. magnitude in dB
fr=fb;                  % in the same freq. range: fs-fa to fN
M=1e+6;                 % MHz scale factor
semilogx( f/M,F7o,'-r', f/M,'-b', fa/M,Sa,'-y', fb/M,Sb,'-c',...
          fr/M,Frq,'-m', [f(1),f(Nf)]/M,[a,a],'--k',...
          [fN,fN],[-72,-5],':k', [fs,fs],[-80,0],':b' )
xlabel('f [MHz]')
```

As can be seen in Fig. 7.2.10, the required minimum attenuation F_{rq} is broad and smooth, so we can assume that it can be approximated by a digital filter of a relatively low order; e.g., if the analog filter has 7 poles, the digital one could have only 6 poles. The combined system would then be effectively a 13-pole. Of course, the digital filter reduces the combined system's cut off frequency, but it would still be higher than in the original, non-shifted, analog only version. However, the main problem is that the cascade of two separately optimized filters has a non-optimal response and the shape of the transient suffers most. This can be solved by simply starting from a higher order system, say, a 13-pole Bessel–Thomson. Then we assign 7 of the 13 poles to the analog filter and 6 poles to the digital one.

The 6 poles of the digital filter must be transformed into appropriate sampling time delays and amplitude coefficients. This can be done either with 'z-transform' mapping, or simply by calculating its impulse response and use it for convolution with the sampled input signal, as we shall do here.

But note that since now the 7 poles of the analog filter will be taken from a 13-pole system, they will be different from the 7-pole system discussed so far (see a comparison of the poles in Fig. 7.2.11). Although the frequency response will be different, the shape of the alias band will be similar, since the final slope is the same in both cases. Nevertheless, we must repeat the calculations with the new poles.

The question is by which criterion do we choose the 7 poles from the 13. From F_{rq} in Fig. 7.2.9 we can see that the digital filter should not cut sharply, but rather gradually. Such a response could be achieved if we reserve the poles with the lower imaginary part for the digital filter and assign those with high imaginary part to the analog filter. But then the analog filter step response would overshoot and ring, compromising the dynamic range. Thus, the correct design strategy is to assign the real and every other complex conjugate pole pair to the analog filter, as shown below.

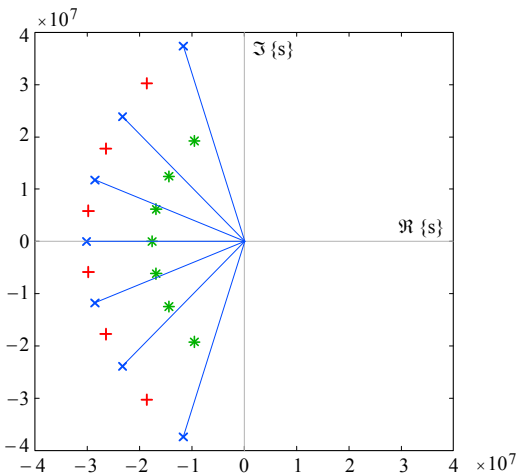


Fig. 7.2.11: The 13-pole mixed mode system, the analog part marked by '×', the digital by '+'; compared with the original 7-pole analog only filter, marked by '*'. Note the difference in pattern size (proportional to bandwidth!).

The values of the mixed mode filter poles for the analog and digital part are:

```
» p7a :  1e+7 *  -3.0181
                 -2.8572 - 1.1751i
                 -2.8572 + 1.1751i
                 -2.3275 - 2.3850i
                 -2.3275 + 2.3850i
                 -1.1637 - 3.7353i
                 -1.1637 + 3.7353i

» p6d :  1e+7 *  -2.9785 - 0.58579i
                 -2.9785 + 0.58579i
                 -2.6460 - 1.77250i
                 -2.6460 + 1.77250i
                 -1.8655 - 3.02640i
                 -1.8655 + 3.02640i
```

Here is a calculation of aliasing suppression using a 13-pole mixed mode filter by using the above poles:

```
fN=1e+7;                % Nyquist frequency
M=1e+6;                 % MHz-us conversion factor
fs=2*fn;                % sampling frequency
A=2^12;                 % ADC dynamic range
a=20*log10(1/A);        % ADC resolution limit in dB
Nf=601;                 % number of frequency samples
f=logspace(6,8,Nf);     % log-scaled frequency vector, 1 - 100 MHz

     % find the frequency normalization factor :
f7=fN/10^(log10(A^2-1)/(2*7));
     % the 7-pole analog-only filter, used as a reference :
[z7,p7]=bestap(7,'a');
p7=2*pi*f7*p7;          % denormalized poles of the reference
F7o=20*log10(abs(freqw(p7,2*pi*f)));  % magnitude of the reference

[z13,p13]=bestap(13,'a');    % the 13th-order Bessel-Thomson system
B=1.73;                 % chosen bandwidth increase
p13=B*2*pi*f7*p13;      % denormalize the 13 poles (f7 same as ref.)
p13=sort(p13);          % sort the poles in ascending abs. value
p7a=p13([1,4,5,8,9,12,13]);  % analog filter pole selection
p6d=p13([2,3,6,7,10,11]);    % digital-equivalent filter poles

F7a=20*log10(abs(freqw(p7a,2*pi*f))); % analog system magnitude
F6d=20*log10(abs(freqw(p6d,2*pi*f))); % digital-equivalent magnitude
F13=20*log10(abs(freqw(p13,2*pi*f))); % total a+d system magnitude

xa=max(find(F7a>=a));  % freq. index of F7a crossing at a
xn=max(find(f<=fN));   % Nyquist frequency index
fa=f(xn:xa);           % frequency band above Nyquist
Fa=F7a(xn:xa);         % response spectrum above Nyquist
fb=fs-f(xa:-1:xn);     % alias band
Fb=F7a(xa:-1:xn);      % alias spectrum
Frq=a-Fb;              % alias suppression requirement
semilogx( f/M,  F7o, '-k', ...
          f/M,  F7a, '-r', ...
          f/M,  F6d, '-b', ...
          f/M,  F13, '-g', ...
          fa/M, Fa,  '-y', ...
          fb/M, Fb,  '-c', ...
          fb/M, Frq, '-m', ...
          [f(1),f(Nf)]/M, [a,a],'--k' )
axis([1, 100, -80, 0]);
```

The result of the above plot operation (`semilogx`) can be seen in Fig. 7.2.12, where we have highlighted the spectrum under the analog filter F_{7a} beyond the Nyquist frequency, its alias, and the inverted alias, which represents the minimum required attenuation F_{rq} of the digital filter. Note how the 6-pole digital filter's response F_{6d} just touches the F_{rq}. Probably it is just a coincidence that the bandwidth increase factor B chosen for the analog filter is equal to $\sqrt{3}$ (we have arrived at this value by repeating the above calculation several times, adjusting B on each run). This process can be easily automated by iteratively comparing the samples of F_{rq} and F_{6d}, and increase or decrease the factor B accordingly.

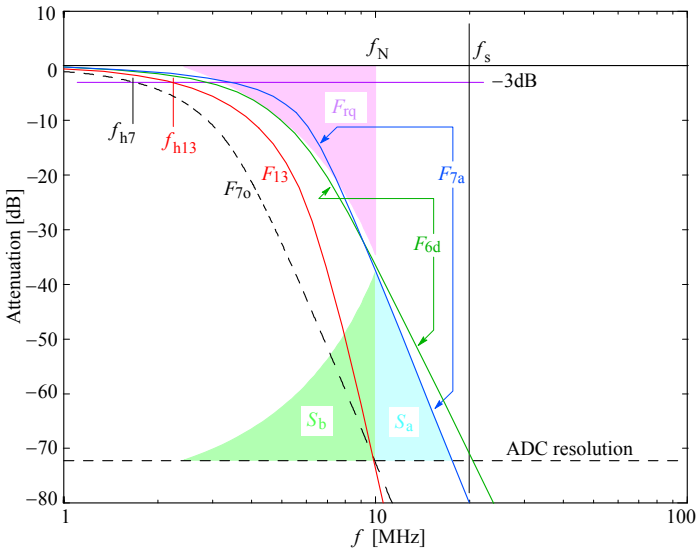


Fig. 7.2.12: The bandwidth improvement achieved with the 13-pole mixed mode filter. The F_{7o} is the original 7-pole analog only filter response, as in Fig. 7.2.9. The new analog filter response F_{7a}, which uses 7 of the 13 poles as shown in Fig. 7.2.11, was first calculated to have the same -72 dB attenuation at the Nyquist frequency f_N (as F_{7o}), and then all the 13 poles were increased by the same factor $B = 1.73$ as before. The resulting F_{7a} response generates the alias spectrum S_b from its source S_a. The envelope of the inverted alias spectrum F_{rq} sets the upper limit for the digital filter's response, F_{6d}, required for effective alias suppression. The response F_{13} is the total analog + digital 13-pole system, which crosses the ADC resolution limit at f_N, and suppresses the alias band below the ADC resolution level, which was the main goal. In this way the system's cut off frequency has increased from 1.66 MHz for F_{7o} to 2.2 MHz for the F_{13}. Thus, a bandwidth improvement of about 1.32 is achieved (not very much, but still significant — note that there are ways of improving this further!).

As expected, the digital filter has reduced the system bandwidth below that of analog filter; however, the analog + digital system's response F_{13} has a cut off frequency f_{h13} well above the f_{h7} of the original analog only 7-pole response, the F_{7o}:

$$f_{7o} = 1.66\,\text{MHz} \qquad f_{13} = 2.2\,\text{MHz}$$

Therefore the total bandwidth improvement factor is $f_{13}/f_{7o} = 1.32$.

The digital filtering process which we are going to use is actually a convolution of discrete signal samples with the filtering coefficients which represent the filter impulse response (sampled) in the time domain. Therefore, if we are going to implement the digital filter in software, we should pay attention to the number of samples of the digital filter impulse response. A high number of samples means a longer time to calculate the convolution. Luckily, the pole selection used will have a nearly Gaussian, non-oscillating, impulse response with only a small undershoot, so a suitable digital filter can be made with only 11 samples (see Fig. 7.2.15).

Let us now calculate and compare the two step responses to examine the effect of bandwidth improvement in time domain:

```
t=2e-9*(0:1:500);              % time vector, 2 ns resolution
M=1e+6;
g7o=atdr(z70,p70,t,'s');       % reference 7-pole system step-response
g13=atdr(z13,p13,t,'s');       % 13-pole step-response

% plot the step responses with the 0.1 and 0.9 reference levels :
plot( t*M, g7o, '-k',...
      t*M, g13, '-r',...
      t([5, 25])*M, [0.1, 0.1], '-k',...
      t([15, 45])*M, [0.9, 0.9], '-k' )
xlabel('t [us]')
```

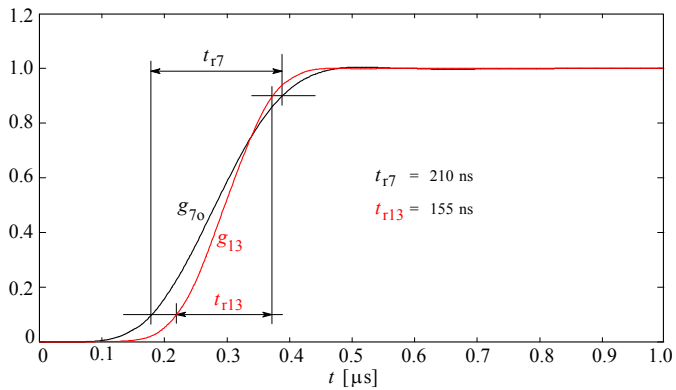


Fig. 7.2.13: Step response comparison of the original 7-pole analog only filter g_{7o} and the improved 13-pole mixed mode (7-pole analog + 6-pole digital) Bessel–Thomson filter, g_{13}. Note the shorter rise time and longer time delay of g_{13}. The resulting rise time is also better than that of the 9-pole analog only filter (see Fig. 7.2.8).

The greatest improvement, however, can be noted in the group delay; as shown in Fig. 7.2.14, the mixed mode system more than doubles the linear phase bandwidth, thus putting a lower constraint on spectrum analysis.

```
                               % continuing from previous calculations:
gd7o=gdly(z7,p7,2*pi*f);       % group-delay for the original 7-pole and
gd13=gdly(z13,p13,2*pi*f); % the mixed-mode 13-pole system
semilogx( f/M, gd7o*M, '-k', f/M, gd13*M, '-r')
```

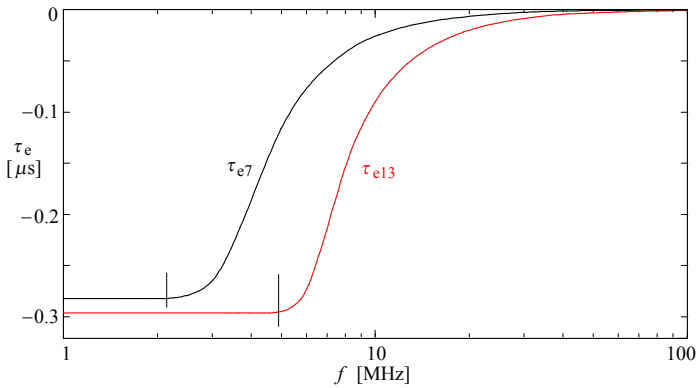


Fig. 7.2.14: Envelope delay comparison of the original 7-pole analog only filter τ_{e7} and the mixed mode 13-pole Bessel–Thomson filter τ_{e13}. The 7-pole analog only filter is linear to a little over 2 MHz, while the 13-pole mixed mode filter is linear well into the stop band, up to almost 5 MHz, more than doubling the useful linear phase bandwidth.

The reader is encouraged to repeat all the above calculations also for the 5- and 9-pole case to examine the dependence of the bandwidth improvement factor on the analog filter's slope.

As we have seen, the bandwidth improvement factor is very sensitive to the steepness of the attenuation slope, so the 9-pole analog filter assisted by an 8-pole equivalent digital filter may be found to be attractive now, extending the bandwidth further.

7.2.4 Gain Optimization

We have said that we need a total gain of 100. Since the analog 7-pole filter can be realized with a three-stage amplifier (one 3-pole stage and two 2-pole stages, see Fig. 7.2.22), the gain of each stage can be optimized by taking the third root of the total gain, $100^{1/3} = 4.6416$ or 13.3 dB (compare this to a two-stage 5-pole filter, where each stage would need $100^{1/2} = 10$ or 20 dB). The lower gain should allow us to use opamps with lower f_T and still have a good phase margin to prevent pole drifting from the optimum (because of the decreasing feedback factor at high frequencies). This is important, since the required 12 bit precision is difficult to achieve without local feedback, and the cost of fast precision amplifiers can be high.

As calculated before, for the sampling frequency of 20 MHz the bandwidths are 1.660 MHz for the 7-pole analog only filter and 2.188 MHz for the 13-pole mixed mode system. In addition to the 13.3 dB of gain, each amplifier stage will need at least 20 dB of feedback at these frequencies, to prevent the response peaking which would result from the finite amplifier bandwidth (if it were too low). By accounting for the amplifier gain roll–off of 20 dB/decade we conclude that we need amplifiers with a unity gain bandwidth of at least 70 MHz for a 9-pole filter and 120 MHz for a 7-pole filter. Note also that amplifiers with a unity gain bandwidth of 160 MHz would be required for the two stages of the 5-pole filter with 20 dB of gain per stage and a system cutoff frequency of only 1.160 MHz!

7.2.5 Digital Filtering Using Convolution

Before we set off to design the analog filter part, let us check the digital filtering process. If we take the output of the analog filter and convolve it with the impulse response of the 6-pole equivalent digital filter, we obtain the response of the composite 13-pole filter. We would also like to see how the step response would look when sampled at the actual ADC sampling frequency of 20 MHz:

```
dt=2e-9;                        % 2 ns time resolution for plotting
Nt=501;                         % length of time vector
t=dt*(0:1:Nt-1);                % time vector
M=1e+6;                         % microsecond normalization factor
fs=2e+7;                        % actual ADC sampling frequency (20MHz)
r=1/(fs*dt);                    % num. of dt's in the sampling period

g7a=atdr(z7,p7a,t,'s');         % step-response of the analog part
h6d=atdr(z7,p6d,t,'n');         % normalized impulse resp. digital part
g13=conv(g7a,h6d);              % digital filtering = convolution
g13=g13(1:max(size(t)));        % take only length(t) samples

plot( t*M, g7a, '-r',...
      t*M(1:r:Nt), g7a(1:r:Nt), 'or',...
      t*M, Nt*h6d, '-g',...
      t*M(1:r:Nt), Nt*h6d(1:r:Nt), 'og',...
      t*M, g13, '-b',...
      t*M(1:r:Nt), g13(1:r:Nt), 'ob' )
xlabel('t [us]')
```

The plot can be seen in Fig. 7.2.15. The dot markers indicate the first 15 time samples of the analog filter step response, the digital filter impulse response and the composite mixed mode filter step response.

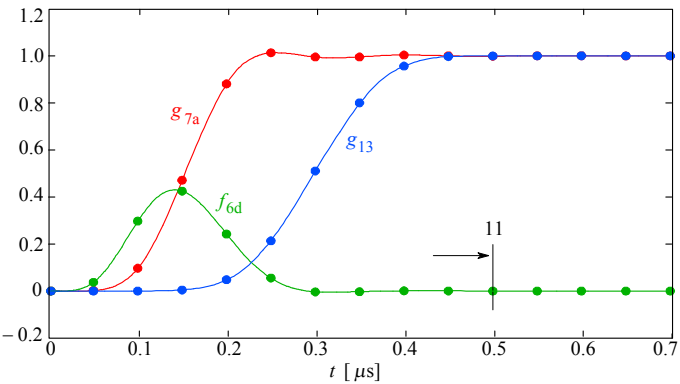


Fig. 7.2.15 : Convolution as digital filtering for a unit step input. The 7-pole analog filter step response g_{a7} is sent to the 6-pole equivalent digital filter, having the impulse response of f_{d6}. Their convolution results in the step response g_{13} of the effectively 13-pole mixed mode composite filter. The marker dots represent the actual time quantization as would result from the ADC sampling at 20 MHz. The impulse response of the digital filter is almost zero after only 11 samples, so the digital filter needs only the first 11 samples as its coefficients for convolution. Note that even if the value of the first coefficient is zero, it must nevertheless be used in order to have the correct response.

Of course, to simulate the actual digitalization, we should have multiplied the normalized signal by 2^{12} and then rounded it to integers to obtain a vertical range of 4096 discrete levels. However, the best computer monitors have a vertical resolution of only $1/4$ of that. If viewed on screen the signal would be seen as if it were digitized with a 10 bit resolution at best (but with lower quantization noise). On paper, with a printer resolution of 600 dots per inch, the vertical size of this figure would have to be 6.5 inches (16.5 cm) in order to appreciate the full ADC resolution.

Nevertheless, measurement precision is always of prime importance, particularly if additional operations are required to extract the information from the recorded signal. In such a case the digital filtering can help, since it will additionally suppress the quantization noise, as well as the amplifier's noise, by a factor equal to the square root of the number of filter coefficients. Disregarding the analog input noise for the moment, if the quantization noise of the 12 bit ADC were ± 2 LSBs at the maximum sampling frequency, its precision would be effectively 10 bits. So, if the impulse response of our digital filter was 11 samples long the quantization noise would be averaged by an additional $\sqrt{11}$, or about 3.3 times, resulting in an almost 2 bit improvement in precision. Effectively, the internal signal precision in memory would be restored to nearly 12 bits.

A final note on the step response: we have started the design of the mixed mode filter because we have assumed that aliasing would be a problem. However, aliasing is a problem only for periodic signals, not for the step! Nevertheless, we are interested in the step response for two reasons:

- one is that even if we were not to care for it, our customers and the people using our instrument would want to know the rise time, the settling time, etc., and, also very important, our competitors would certainly not spare their means of checking it if they can show they are better, or just very close but cheaper!
- the other reason is that the step response will give us a direct evidence of the system phase linearity, a parameter of fundamental importance in spectrum analysis.

7.2.6 Analog Filters With Zeros

By using a 13-pole mixed mode filtering we have so far achieved a bandwidth improvement of about 1.37 compared to a 7-pole analog only filter. But even greater improvements are possible if we use analog filters with zeros. Combining a Bessel low pass response with zeros in the stop band is difficult but not impossible to achieve. The zeros must be in pairs and purely imaginary, and they must be placed at the Nyquist frequency and its first harmonic, the sampling frequency (beyond f_s the filter attenuation is high, so any aliasing from there will already be below the ADC resolution). In this way the zeros effectively prevent aliasing down to DC, and they also modify the filter slope to become very steep near them, allowing us to increase the bandwidth further still: a factor of up to 1.87 can be achieved.

One such example, calculated from the same initial assumptions as before, is shown in the following figures. Fig. 7.2.x5 shows the poles and zeros, Fig. 7.2.x6 shows the frequency responses, and Fig. 7.2.x7 shows the alias spectrum in relation to the filter responses and the resulting alias suppression by the digital filter.

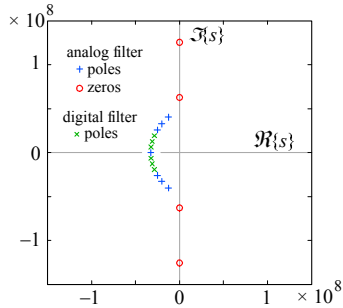


Fig. 7.2.16: An example of a similar mixed mode filter, but with the analog filter using two pairs of imaginary zeros, one pair at the sampling frequency and the other pair at the Nyquist frequency.

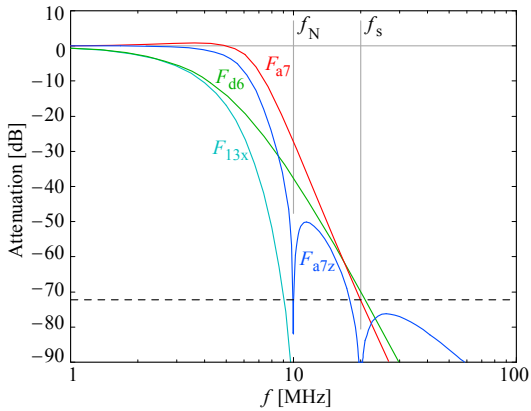


Fig. 7.2.17: By adding the zeros the analog filter frequency response in modified from F_{a7} to F_{a7z}. F_{d6} is the digital filter response and F_{13x} is the composite mixed mode response.

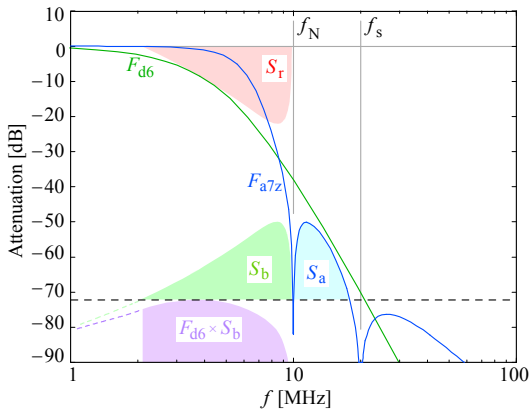


Fig. 7.2.18: The spectrum S_a is aliased into S_b. The difference between the ADC resolution and S_b (both in dB) gives S_r, the envelope of which determines the minimum attenuation required by F_{d6} to suppress S_b below the ADC resolution.

Fig. 7.2.19 shows the time domain responses. Note the influence of zeros on the analog filter response.

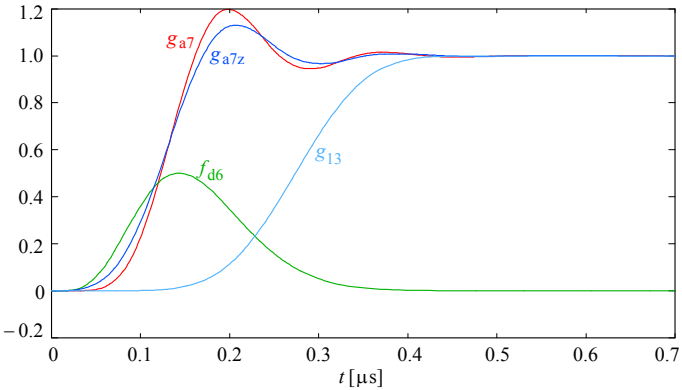


Fig. 7.2.19: The step response of the 7-pole analog filter g_{a7} is modified by the 4 zeros into g_{a7z}. Because the alias spectrum is narrower the bandwidth can be increased. The mixed mode system step response g_{13} has a rise time of less than 150 ns (in contrast with the 180 ns for the case with no zeros).

7.2.7 Analog Filter Configuration

So far we have inspected the system performance in detail, without knowing the circuit's actual configuration. Now it is time to select the analog filter configuration and calculate the component values needed to comply with the required performance. To achieve this we have to consider both the system's accuracy (which has to be equal or better than the 12 bit ADC resolution) and the gain–bandwidth product of each stage, which we have already accounted for in part.

In general, any filter topology could be used if only the transfer function is considered. However for high frequency applications the filter topology which suits us best is the 'Multiple–Feedback' type (MFB), built around operational amplifiers. MFB uses a phase inverting, virtual ground amplifier with the filtering components in its feedback loop, as shown in Fig. 7.2.20 and 7.2.21. Without feedback, the 12 bit precision, offered by the ADC, would be impossible to obtain and preserve in the analog circuit. As a bonus, this topology also offers lowest distortion.

In fact, the non-inverting configurations, such as the '*Sallen–Key*' (SK) filter type or the '*Frequency Dependent Negative Resistance*' (FDNR) type, must support a common mode signal as high as the input signal, whilst the filtering is done on the small differential signal. The mismatch between the inverting and non-inverting opamp input impedance is inevitable in filters, so the finite and frequency dependent common mode rejection plays an important role regarding signal fidelity. As a 'rule of thumb' for high speed opamps: **invert if you can; if you can not, invert twice!**

Another factor, which becomes important in high frequency active filters, is the impedance of the filtering components; the MFB configuration allows us to use low resistance and moderate capacitance values, thus minimizing the influence of strays. In

the figures below are the schematic diagrams of a 3-pole and a 2-pole stage. We shall use the 3+2+2 cascade to implement the 7-pole filter required.

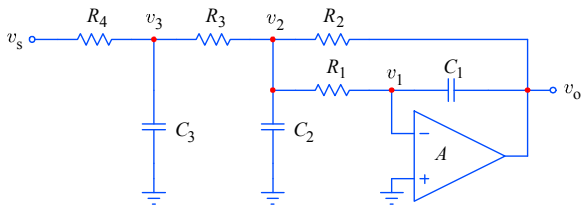


Fig. 7.2.20. The 'Multiple–Feedback' 3-pole (MFB-3) low pass filter configuration

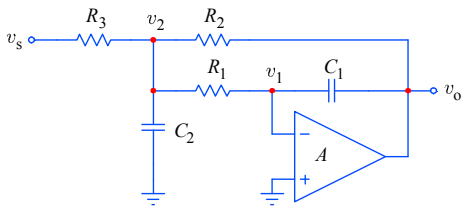


Fig. 7.2.21: The 'Multiple–Feedback' 2-pole (MFB-2) low pass filter configurations

7.2.8 Transfer Function Analysis of the MFB-3 Filter

Our task is to find the transfer functions of these two circuits and then relate the time constants to the poles so that the required component values can be found.

The 3rd-order stage will be the first in the cascade, so let us calculate its components first. For the complete analysis please refer to Appendix 7.1; here we write only the resulting transfer function, which in the general case is:

$$F(s) = A_0 \frac{-s_1 s_2 s_3}{(s-s_1)(s-s_2)(s-s_3)}$$

$$= A_0 \frac{-s_1 s_2 s_3}{s^3 - s^2(s_1+s_2+s_3) + s\,(s_1 s_2 + s_1 s_3 + s_2 s_3) - s_1 s_2 s_3} \qquad (7.2.2)$$

For the MFB-3 circuit the coefficients at each power of s are:

$$-(s_1+s_2+s_3) = \frac{1}{C_3 R_4}\left(1+\frac{R_4}{R_3}\right) + \frac{1}{C_2 R_3}\left(1+\frac{R_3}{R_2}+\frac{R_3}{R_1}\right) \qquad (7.2.3)$$

$$s_1 s_2 + s_1 s_3 + s_2 s_3 = \frac{1+(R_3+R_4)\left(\frac{1}{R_2}+\frac{1}{R_1}\right)}{C_2 C_3 R_3 R_4} + \frac{\frac{R_3}{R_2}}{C_1 C_2 R_1 R_3} \qquad (7.2.4)$$

$$-s_1 s_2 s_3 = \frac{R_3}{R_2}\left(1+\frac{R_4}{R_3}\right)\frac{1}{C_1 C_2 C_3 R_1 R_3 R_4} \qquad (7.2.5)$$

and the DC gain A_0 is:

$$A_0 = -\frac{R_2}{R_3 + R_4} \tag{7.2.6}$$

By examining the coefficients and the gain we note that we can optimize the component values by making the resistors R_1, R_3, and R_4 equal:

$$R_1 = R_3 = R_4 = R \tag{7.2.7}$$

The coefficients and the gain equations now take the following form:

$$-(s_1 + s_2 + s_3) = \frac{2}{C_3 R} + \frac{1}{C_2 R}\left(2 + \frac{R}{R_2}\right) \tag{7.2.8}$$

$$s_1 s_2 + s_1 s_3 + s_2 s_3 = \frac{1}{C_2 C_3 R^2}\left(3 + \frac{2R}{R_2}\right) + \frac{1}{C_1 C_2 R^2} \cdot \frac{R}{R_2} \tag{7.2.9}$$

$$-s_1 s_2 s_3 = \frac{2R}{R_2} \cdot \frac{1}{C_1 C_2 C_3 R^3} \tag{7.2.10}$$

$$A_0 = -\frac{R_2}{2R} \tag{7.2.11}$$

To simplify our expressions we shall substitute the term R/R_2 by:

$$M = \frac{R}{R_2} = -\frac{1}{2A_0} \tag{7.2.12}$$

so the coefficients can be written as:

$$K_2 = -(s_1 + s_2 + s_3) = \frac{2}{C_3 R} + \frac{2 + M}{C_2 R} \tag{7.2.13}$$

$$K_1 = s_1 s_2 + s_1 s_3 + s_2 s_3 = \frac{3 + 2M}{C_2 C_3 R^2} + \frac{M}{C_1 C_2 R^2} \tag{7.2.14}$$

$$K_0 = -s_1 s_2 s_3 = \frac{2M}{C_1 C_2 C_3 R^3} \tag{7.2.15}$$

Next we can normalize the resistance ratios and the RC time constants in order to eliminate the unnecessary variables. After the equations are solved and the component ratios are found we shall denormalize the component values to the actual cut off frequency, as required by the poles. We can thus set the normalization factor:

$$N = \frac{1}{R} \tag{7.2.16}$$

so instead of R we have the normalized resistance R_N:

$$R_N = NR = 1 \tag{7.2.17}$$

Accordingly, the capacitance values are now also normalized, and to distinguish the new values from the original ones we shall label the capacitors as:

$$C_\mathrm{a} = \frac{C_1}{N} \qquad C_\mathrm{b} = \frac{C_2}{N} \qquad C_\mathrm{c} = \frac{C_3}{N} \tag{7.2.18}$$

With these changes we obtain:

$$K_2 = \frac{2}{C_\mathrm{c}} + \frac{2+M}{C_\mathrm{b}} \tag{7.2.19}$$

$$K_1 = \frac{3+2M}{C_\mathrm{b}C_\mathrm{c}} + \frac{M}{C_\mathrm{a}C_\mathrm{b}} \tag{7.2.20}$$

$$K_0 = \frac{2M}{C_\mathrm{a}C_\mathrm{b}C_\mathrm{c}} \tag{7.2.21}$$

We can now express, say, C_a from Eq. 7.2.21:

$$C_\mathrm{a} = \frac{2M}{K_0 C_\mathrm{b}C_\mathrm{c}} \tag{7.2.22}$$

which we insert into Eq. 7.2.20:

$$K_1 = \frac{3+2M}{C_\mathrm{b}C_\mathrm{c}} + \frac{K_0 C_\mathrm{c}}{2} \tag{7.2.23}$$

From this we express C_b:

$$C_\mathrm{b} = \frac{2(3+2M)}{2K_1C_\mathrm{c} - K_0C_\mathrm{c}^2} \tag{7.2.24}$$

Now we can eliminate C_b from Eq. 7.2.19:

$$K_2 = \frac{2}{C_\mathrm{c}} + \frac{2+M}{\dfrac{2(3+2M)}{2K_1C_\mathrm{c} - K_0C_\mathrm{c}^2}} \tag{7.2.25}$$

and we remain with only C_c, for which we have a third-order equation:

$$C_\mathrm{c}^3 - 2\frac{K_1}{K_0}\,C_\mathrm{c}^2 + \frac{2(3+2M)K_2}{(2+M)K_0}\,C_\mathrm{c} - \frac{4(3+2M)}{(2+M)K_0} = 0 \tag{7.2.26}$$

By substituting the coefficients of this equation with p, q, and r:

$$p = -2\frac{K_1}{K_0} \tag{7.2.27}$$

$$q = \frac{2(3+2M)K_2}{(2+M)K_0} \tag{7.2.28}$$

$$r = -\frac{4(3+2M)}{(2+M)K_0} \tag{7.2.29}$$

we can rewrite Eq. 7.2.23 as:

$$C_c^3 + p\,C_c^2 + q\,C_c + r = 0 \tag{7.2.30}$$

The real general solution for this third-order equation (see Appendix 2.1) is:

$$C_c = -\frac{2}{3}\sqrt{p^2 - 3q}\,\sin\left\{\frac{1}{3}\arctan\left[\frac{-\sqrt{-3}\,(2\,p^3 - 9\,p\,q + 27\,r)}{9\sqrt{4\,r\,p^3 - p^2q^2 - 18\,p\,q\,r + 4\,q^3 + 27\,r^2}}\right]\right\} - \frac{p}{3} \tag{7.2.31}$$

By inserting the poles s_1, s_2, and s_3 into the expressions for K_0, K_1, K_3, and the expression for gain in M, and then using it all in the expressions for p, q, and r, we finally obtain the value of C_c. The explicit expression would be too long to write here, and, anyway, it is only a matter of simple substitution, which in a numerical algorithm would not be necessary. With the value of C_c known we can go back to Eq. 7.2.24 to find the value of C_b:

$$C_b = \frac{2\left(3 - \dfrac{1}{A_0}\right)}{\left[2(s_1s_2 + s_1s_3 + s_2s_3)C_c + s_1s_2s_3C_c^2\right]} \tag{7.2.32}$$

and from Eq. 7.2.20 we express C_a:

$$C_a = -\frac{1}{A_0} \cdot \frac{1}{-s_1s_2s_3\,C_bC_c} \tag{7.2.33}$$

Now that the normalized capacitor values are known, the denormalization process makes use of the circuit's cut off frequency, ω_{3h} (which, to remind you, is different from the cut off frequency ω_{7h} of the 7-pole filter, and also different from the total system cut off, ω_{13h}); ω_{3h} relates to K_0 and the poles as:

$$K_0 = \omega_{3h}^3 = -s_1s_2s_3 \tag{7.2.34}$$

From ω_{3h} we can denormalize the values of R and the capacitors to acquire some suitable values, such that the opamp of our choice can easily drive its own feedback impedance and the input impedance of the following stage. For high system bandwidth, it is advisable to choose R as low as possible, usually in the range between 150 and 750 Ω. Let us say that we can set $R = 220\,\Omega$. We use the inverse:

$$N = \frac{1}{R} = \frac{1}{220} \tag{7.2.35}$$

to denormalize the capacitors accordingly:

$$C_1 = NC_a \tag{7.2.36}$$

$$C_2 = NC_b \tag{7.2.37}$$

$$C_3 = NC_c \tag{7.2.38}$$

The cut off frequency is:

$$f_{3h} = \frac{\omega_{3h}}{2\pi} = \frac{1}{2\pi R}\sqrt[3]{\frac{1}{-A_0}\cdot\frac{1}{C_1C_2C_3}} \tag{7.2.39}$$

From the system gain we obtain the value of R_2:

$$R_2 = -2\,R\,A_0 \tag{7.2.40}$$

By inserting the first three poles from `p7a` for s_1, s_2, and s_3, we obtain the following component values:

```
% The poles of the 7th-order analog filter:

p7a:     1e+7 *  -3.0181
                 -2.8572 - 1.1751i
                 -2.8572 + 1.1751i
                 -2.3275 - 2.3850i
                 -2.3275 + 2.3850i
                 -1.1637 - 3.7353i
                 -1.1637 + 3.7353i     [rad/s]

% The first three poles of p7a are assigned to the MFB-3 circuit:

s1 = -3.0181 * 1e+7                    [rad/s]
s2 = ( -2.8572 - 1.1751i ) * 1e+7      [rad/s]
s3 = ( -2.8572 + 1.1751i ) * 1e+7      [rad/s]

% The single stage gain is:

Ao = 100^(1/3) = 4.642

% ------ MFB-3 components: ------------------------------

               R=R4=R3=R1 = 220 Ohm     R2 = 2042 Ohm

  C3 = 102.8 pF         C2 = 274.5 pF    C1 = 24.9 pF

% The cut off frequency is:

 f3h = 4.879 MHz
```

7.2.9 Transfer Function Analysis of the MFB-2 Filter

The derivation for the two second-order stages, which are both of the form shown in Fig. 7.2.21, is also reported in detail in Appendix 7.2. Again, here we give only the resulting transfer function:

$$\frac{v_o}{v_s} = -\frac{R_2}{R_3}\cdot\frac{\dfrac{R_3}{R_2}\cdot\dfrac{1}{R_1R_3C_1C_2}}{s^2 + s\left(\dfrac{R_3}{R_1}+\dfrac{R_3}{R_2}+1\right)\dfrac{1}{R_3C_2} + \dfrac{R_3}{R_2}\cdot\dfrac{1}{R_1R_3C_1C_2}} \tag{7.2.41}$$

By comparing this with the general form:

$$H(s) = A_0 \frac{s_1 s_2}{(s - s_1)(s - s_2)} = A_0 \frac{s_1 s_2}{s^2 - s\,(s_1 + s_2) + s_1 s_2} \tag{7.2.42}$$

we find the system gain:

$$A_0 = -\frac{R_2}{R_3} \tag{7.2.43}$$

and the component values are found from the denominator polynomial coefficients, in which, in order to optimize the component values, we again make $R_1 = R_3 = R$:

$$s_1 s_2 = \frac{1}{-A_0\, R^2\, C_1 C_2} \tag{7.2.44}$$

$$-(s_1 + s_2) = \frac{1}{R\, C_2} \left(2 + \frac{1}{-A_0}\right) \tag{7.2.45}$$

Solving for C_2 and C_1, respectively, we have:

$$C_2 = \frac{1}{-R\,(s_1 + s_2)} \left(2 + \frac{1}{A_0}\right) \tag{7.2.46}$$

$$C_1 = \frac{-(s_1 + s_2)}{s_1 s_2} \frac{1}{R(2A_0 + 1)} \tag{7.2.47}$$

Again we introduce the normalization factor $N = 1/R$, so that $R_N = NR = 1$ and we accordingly normalize the capacitor values:

$$C_\mathrm{a} = \frac{C_1}{N} \qquad\qquad C_\mathrm{b} = \frac{C_2}{N} \tag{7.2.48}$$

Then:

$$C_\mathrm{b} = \frac{1}{-(s_1 + s_2)} \left(2 + \frac{1}{A_0}\right) \tag{7.2.49}$$

$$C_\mathrm{a} = \frac{-(s_1 + s_2)}{s_1 s_2} \frac{1}{2A_0 + 1} \tag{7.2.50}$$

Let us say that here, too, we use the same values for R and N, as before ($R = 220\,\Omega$, $N = 1/200$; note however that in general we can use a different value if for whatever reason we find it more suitable — one such reason can be the preferred values of capacitors). Thus:

$$C_1 = NC_\mathrm{a} \qquad\qquad C_2 = NC_\mathrm{b} \tag{7.2.51}$$

The cut off frequency of the MFB-2 circuit is simply:

$$f_\mathrm{2h} = \frac{\omega_\mathrm{2h}}{2\pi} = \frac{\sqrt{s_1 s_2}}{2\pi} = \frac{1}{R\sqrt{-A_0\, C_1 C_2}} \tag{7.2.52}$$

From `p7a` we use the 4th and the 5th pole for the first MFB-2 stage and the 6th and 7th pole for the second MFB-2 stage. With those we obtain the following component values:

```
% The first MFB-2 circuit:
 s1 = ( -2.3275 - 2.3850i ) * 1e+7     [rad/s]
 s2 = ( -2.3275 + 2.3850i ) * 1e+7     [rad/s]
f2h = 5.304 MHz

------ component values: -----------------------------
 Ao = 4.642       R=R3=R1 = 220 ohm      R2 = 1021 ohm
                      C2 = 216.3 pF     C1 = 18.5 pF

% The second MFB-2 circuit:
 s1 = ( -1.1637 - 3.7353i ) * 1e+7     [rad/s]
 s2 = ( -1.1637 + 3.7353i ) * 1e+7     [rad/s]
f2h = 6.227 MHz

------ component values: -----------------------------
 Ao = 4.642       R=R3=R1 = 220 ohm      R2 = 1021 ohm
                      C2 = 432.7 pF     C1 = 6.7 pF
```

We can now finally draw the complete circuit schematic diagram with component values:

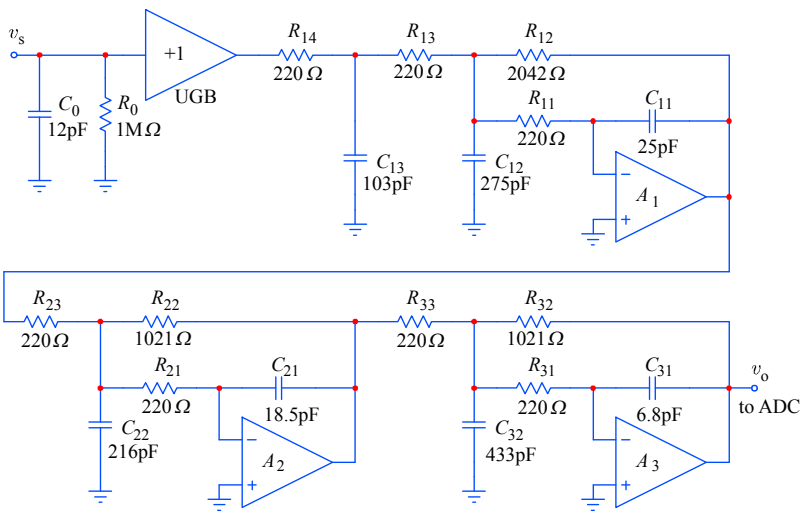


Fig. 7.2.22: The realization of the 7-pole analog filter for the ADC discussed in the aliasing suppression example. The input signal is separated from the filter stages by a high impedance (1 MΩ, 12 pF) unity gain buffer (UGB). The first amplifier stage A_1 with a gain of 4.64 is combined with the third-order filter, the components are calculated from the first three poles taken from the 7-pole analog part of the 13-pole mixed-mode system. The following two second-order filter stages A_2 and A_3 have the same gain as the first stage, whilst their components are calculated from the next two complex conjugate pairs of poles from the same bunch of 7. All three amplifying stages are inverting, so the final inversion must be done either at the signal display, the digital filter routine, or simply by taking the 2's complement of the ADC's digital word.

7.2.10 Standardization of Component Values

More often than not, multi-stage filters will have awkward component values, far from either of the closest preferred standard E12 or E24 values. In addition the greater the number of stages, the larger will be the ratio of maximal to minimal values, forcing the use of components with very tight tolerance.

Fortunately we are not obliged to use exactly the calculated values; indeed, we are free to adjust them, paying attention that each stage keeps its time constants as calculated. I.e., the capacitors will be probably difficult to obtain in E24 values, but resistors are easily available in E48 and even E96 values, so we might select the closest E12 values for the capacitors and then select the resistors from, say, E48.

Considering for example the first 2-pole stage (A_2) we could use 18 pF for C_{21} (instead of 18.5); then C_{22} would be 210 pF (say, 180 and 30 pF in parallel), the resistors R_{21} and R_{23} can be set to 226 Ω and R_{22} can be set to 1050 Ω.

The other two stages can be changed in a similar way. It is advisable not to depart from the initial values too much, in order to keep the impedances close to the driving capability of the amplifiers (remember that each amplifier has to drive both the input impedance of the following stage and the impedance of its own feedback network) and also to remain well above the influence of stays (low value capacitances and the amplifier inverting inputs are the most sensitive in this respect).

7.2.11 Concluding Remarks

The initial design requirement for the procedure described is probably an overkill, since we shall very rarely have the noise level or some other interfering signal as high as the full ADC amplitude range at the Nyquist frequency limit or above. If we are confident that the maximum disturbance level at the Nyquist frequency will always be some 30 dB lower than the full range amplitude, we can relax the filter specifications accordingly, either by having a less steep and less complicated filter, or by raising the filter's cut off frequency so that the attenuation at the Nyquist frequency intersects the level 30 dB above the ADC resolution.

For the example above, the maximum input signal was 40 mV, so, in the case of an interfering signal of, say, 1.3 mV (−30 dB), our filter would need an attenuation of about −42 dB at the Nyquist frequency. For the 7th-order analog part of the filter, having the attenuation slope of $-140\,\mathrm{dB}/10f$, this would result in a nearly doubled bandwidth (the digital part remains the same), but note that the alias spectrum can now extend down to DC, since its source spectrum includes the sampling frequency. Also, as we have seen at the beginning of this section, the $(\sin\omega T_\mathrm{s})/\omega T_\mathrm{s}$ spectral envelope allows us a further 12–13 dB at about $0.7f_\mathrm{s}$. In such a case an additional filter stage, with zeros at f_N and f_s and the third harmonic of f_N (all on the imaginary axis), such as an inverted (type II) Chebyshev filter, could be used to improve the alias rejection at low frequencies without spoiling the upper roll off slope of the frequency response by much, thus also preserving the smooth step response.

Résumé and Conclusion

We have shown only a small part of the vast possibilities of use offered by the application of numerical routines, either for the purpose of system design or for implementing them within the system's digital signal processing.

Some additional information and a few examples can be found on the disk included with the book, in form of the '*.M' files to be run within Matlab. Many of those files were used to produce the various figures in the book and can be used as a starting point for further routine development.

One of the problems of writing a book about a fast developing subject is that by the time the writers have finished the editing, several years might have passed and the book is no longer at the forefront of the technology's development.

We have tried to prevent the book from becoming outdated too soon by including all the necessary theory and covering the fundamental design principles which are independent of technology, and thus of lasting value. We have also tried to cover some of the most important steps in the development from a historical point of view, to show how those theoretical concepts and design principles have been applied in the past and how they have evolved to today's forms.

Although the importance of staying alert and following the new developments and ideas is as high as ever, the knowledge of the basic theory and its past applications helps us to identify more quickly the correct paths to follow and the aims to pursue.

References:

[7.1] *J.N. Little* and *C.B. Moller*, The MathWorks, Inc.:
PC-MATLAB For Students (containing disks with Matlab program)
Prentice-Hall, 1989

MATLAB-V For Students (containing a disk with Matlab program)
Prentice-Hall, 1998

Web link : <http://www.mathworks.com/>

See also the books on electronics + Matlab at:
<http://www.mathworks.com/matlabcentral/link_exchange/MATLAB/Books/Electronics/>

[7.2] *C.E. Shannon*, Collected Papers,
IEEE Press, Cat. No.: PC 0331

See also:
<http://en.wikipedia.org/wiki/Nyquist-Shannon_interpolation_formula>

[7.3] *A.V. Oppenheim* and *R. W. Schafer*, Digital Signal Processing,
Prentice-Hall, 1975

[7.4] *K. Azadet*, Linear-phase, continuous-time video filters based on mixed A/D structure,
ECCTD 1993 - Circuit Theory and Design, pp. 73–78, Elsevier Scientific Publishing

[7.5] *E. Margan*, Anti-aliasing with mixed-mode filters,
Electronics World and Wireless World, June, 1995, pp. 513–518
<http://www.softcopy.co.uk/electronicsworld/>

Index