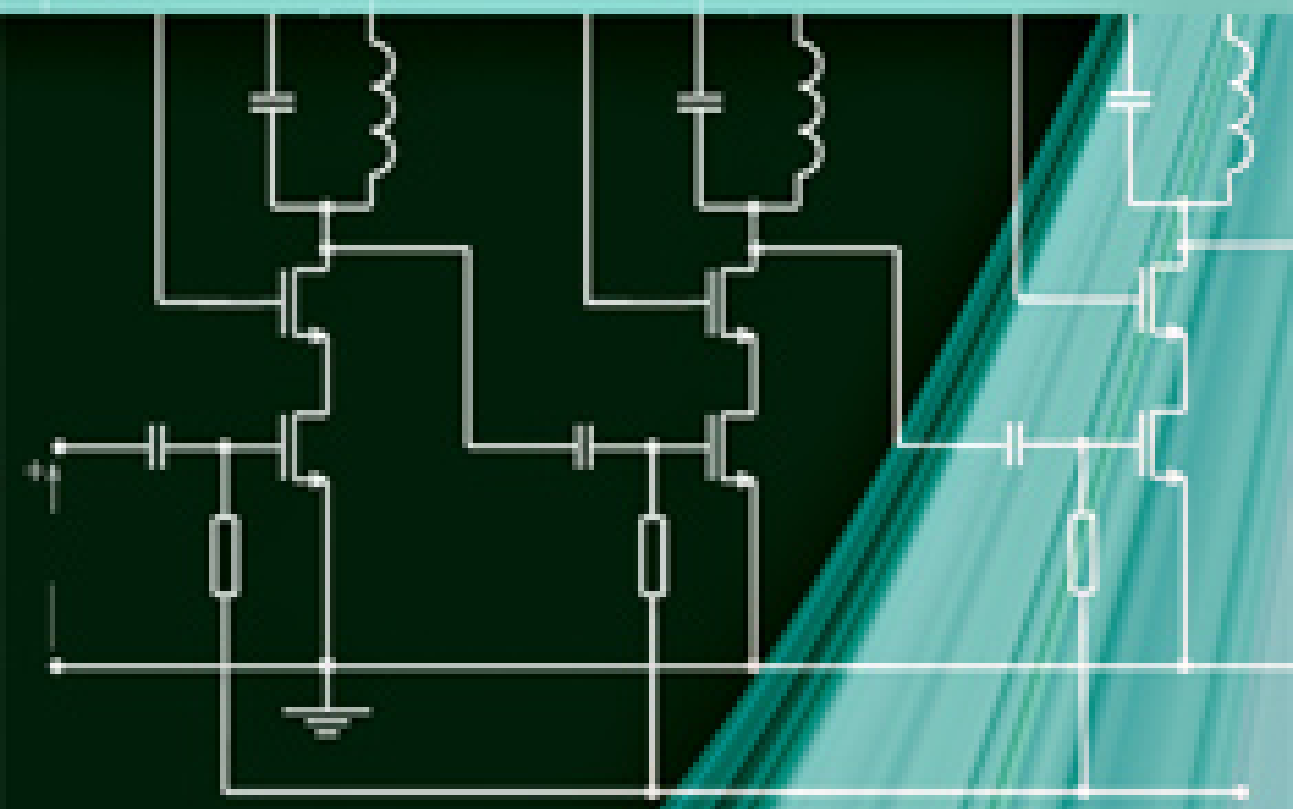
Fundamentals of High-Frequency CMOS Analog Integrated Circuits



Duran Leblebici
and Yusuf Leblebici

CAMBRIDGE

CAMBRIDGE www.cambridge.org/9780521513401

This page intentionally left blank

Fundamentals of High-Frequency CMOS Analog Integrated Circuits

With a design-centric approach, this textbook bridges the gap between fundamental analog electronic circuits textbooks and more advanced RF IC design texts. The structure and operation of the building blocks of high-frequency ICs are introduced in a systematic manner, with an emphasis on transistor-level operation, the influence of device characteristics and parasitic effects, and input–output behavior in the time and frequency domains.

Key features include:

- solved design examples to guide the reader through the decision process that accompanies each design task, with an emphasis on key trade-offs;
- coverage of the major issues that must be taken into account when combining analog and digital circuit building blocks;
- key criteria and parameters that are used to describe system-level performance;
- simple circuit models to enable a robust understanding of high-frequency design fundamentals;
- SPICE simulations that are used to check results and fine-tune the design.

This textbook is ideal for senior undergraduate and graduate courses in RF CMOS circuits, RF circuit design, and high-frequency analog circuit design. Analog integrated circuit designers and RF circuit designers in industry who need help making design choices will also find this a practical and valuable reference.

Duran Leblebici is Professor Emeritus of Electrical and Electronics Engineering at Istanbul Technical University (ITU). He has been teaching a range of undergraduate and graduate courses, from device electronics and fabrication technologies to integrated electronic circuits and RF IC design, for more than 40 years. He is the author of three textbooks in the field of electronics. He also established the first microelectronics laboratory and the first VLSI design house at ITU. He has received the Distinguished Service Award of the Turkish Scientific and Technological Research Council (TUBITAK) in 1992, in recognition of his services to microelectronics education.

Yusuf Leblebici is Director and Chair Professor of the Microelectronic Systems Laboratory at the Swiss Federal Institute of Technology in Lausanne (EPFL). He has previously worked as a faculty member at the University of Illinois at Urbana-Champaign, at Istanbul Technical University, and at Worcester Polytechnic Institute (WPI), where he established and directed the VLSI Design Laboratory, and also served as a project director at the New England Center for Analog and Mixed-Signal IC Design. He is a co-author of more than 150 scientific articles and three textbooks.

Fundamentals of High-Frequency CMOS Analog Integrated Circuits

Duran Leblebici
Istanbul Technical University (ITU)

Yusuf Leblebici
Swiss Federal Institute of Technology in Lausanne (EPFL)

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK
Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521513401

First published in print format 2009

ISBN-13 978-0-511-53976-3 eBook (EBL)

ISBN-13 978-0-521-51340-1 hardback

To
Yıldız Leblebici – wife, mother, and colleague.

Contents

Preface

In the first half of the twentieth century the radio was the main activity area of the electronics industry and, correspondingly, RF circuits occupied a considerable part in the electronic engineering curriculum and books published in this period. Properties of resonance circuits and electronic circuits using them, single-tuned amplifiers as the input stages of receivers, double-tuned circuits as IF amplifiers, RF sinusoidal oscillators and high-power class-C amplifiers have been investigated in depth. It must be kept in mind that the upper limit of the radio frequencies of those days was several tens of MHz, the inductors used in tuned circuits were air-core or ferrite-core coils with inductance values in the micro-henries to milli-henries range, having considerably high quality factors, ranging from 100 to 1000, and the tuning capacitors were practically lossless.

The knowledge developed for the vacuum-tube circuits easily adapted to the transistors with some modifications related to the differences of the input and output resistances of the devices. In the meantime, the upper limit of the frequency increased to about 100 MHz for FM radio and to hundreds of MHz for UHF-TV. The values of the inductors used in these circuits correspondingly decreased to hundreds to tens of nano-henries. But these inductors were still wound, high-Q discrete components.

In the second half of the twentieth century, the emergence of integrated circuits drastically increased the reach of electronic engineering. Digital electronics on one side and analog electronics using the potentials of the operational amplifiers on the other side forced the curricula and the textbooks to skip certain old and "already known" subjects (among them resonance circuits and tuned amplifiers), to open room to these new subjects. The development of (inductorless) active filters that replaced the conventional passive LC filters extensively used in telecommunication systems even decreased the importance of inductors.

Rapid development of CMOS technology and the steady decrease of the dimensions of the devices according to "Moore's Law" led to an increase in the complexity of ICs (from now on, the VLSI circuits) and the operating frequencies as well. In the digital realm this helped to improve the performances of digital computers and digital telecommunication systems. In the analog realm the operation frequencies of the circuits increased to the GHz region and correspondingly the inductance values decreased to below 20 nH, so that now it was possible to realize them as on-chip components. Hence the freedom from external bulky discrete inductors opened a new

horizon towards small and light-weight mobile systems; the mobile telephone, GPS systems, Bluetooth, etc.

But there is a problem related to this development; the quality factors of on-chip inductors are very low, usually around 10–100, and tuning capacitors are not lossless any more. Most of the earlier theory (and design practice) that was developed with very high-Q discrete components does not easily translate to such integrated high-frequency circuits and on-chip components with less than ideal characteristics. Instead of relying on the comprehensive theory and analytical design of high-frequency analog circuits that was in place many decades earlier, most of the new-generation designers were tempted to adopt rather ad-hoc design strategies not grounded in sound theory, and to explain away the inevitable inconsistencies as "secondary effects". To make matters even more complicated, systematic treatment of subjects such as high-frequency circuit behavior and resonance/tuned circuits has been missing from the electrical engineering curricula for several decades, and analog designers entering the field of RF/high-frequency design had to re-learn these subjects.

One of the objectives of this book is to fill this gap; to introduce the fundamental aspects of high-frequency circuit operation, to systematically discuss the behavior of key components (in particular, submicron MOSFETs and on-chip passive components), summarize the behavior of the series and parallel resonance circuits in detail and investigate the effects of the losses of on-chip inductors and capacitors that are usually not taken into account in formulas derived for high-Q resonance circuits.

Since all these circuits are being developed and used usually in the GHz range, sometimes close to the physical limits of the devices, it becomes necessary to recapitulate the behavior of MOS transistors together with their important parasitics and the frequency related secondary effects.[1] In Chapter 1 the basic current–voltage relations of MOS transistors are derived, taking into account the parabolic (not linear) shape of the inversion charge profile that leads to a different approach to understand the channel shortening effect and the gate-source capacitance of the transistor. The velocity saturation effect and the behavior of a MOS transistor under a velocity saturation regime are also investigated as an important issue for small geometry devices. Although it is not used extensively for HF applications, the sub-threshold regime is also investigated in brief.

In addition to the intrinsic behavior of MOS transistors the parasitics – that are inevitable and have severe effects on the overall behavior especially at high frequencies – have been discussed, mostly in connection with the BSIM3 parameters. The properties, limits and parasitics of the passive on-chip components, namely resistors, fixed value and variable capacitors (varactors) and inductors are also summarized in this chapter from a realistic and design-oriented point of view.

The subject of Chapter 2 is the DC properties of the basic analog MOS circuits that will be investigated in the following chapters. The interactive use of analytic

[1] This part of the book must not be considered as an alternative to the existing sound and comprehensive models such as EKV and others, but rather as an attempt to explain the behavior of MOS transistors, based on the basic laws of electrostatics and circuit theory that all electronics students already know.

expressions – that provide interpretable knowledge about the basic behavior of the circuit – and SPICE simulations – that give the designer a possibility to "experiment", to fine-tune and optimize the circuit, all secondary and parasitic effects included – are exemplified throughout the chapter. It is believed that to have the ability of using together the analytical expressions and the power of SPICE is a "must" for an analog designer.

In Chapter 3 the frequency-dependent behaviors of the basic circuits are given, not limited to the frequency characteristics of the gain but including the input and output impedances. The important properties of them, usually not dealt with in books, are investigated and their effects on the performance of wide-band circuits are underlined. The basics of the techniques used to enhance the gain; the additive approach (distributed amplifiers) and the cascading strategies to reach to the wide-band amplifiers (not only voltage amplifiers, but also current amplifiers, transadmittance amplifiers and transimpedance amplifiers) are systematically investigated.

In Chapter 4 first the resonance circuits are recapitulated with this approach and the behavioral differences of high-Q and low-Q resonance circuits are underlined. Afterwards, tuned amplifiers are systematically investigated taking into account the low-Q effects, not only for single-tuned amplifiers but also for double-tuned and staggered tuned amplifiers that are not covered in many new (and even older) books in detail. The LNA, that is one of the most important classes of tuned amplifiers, is also investigated in this chapter together with the noise behavior of MOS transistors, that is developed with a different approach.

LC sinusoidal oscillators are discussed in Chapter 5 with the negative resistance approach and the classical positive feedback approach as well, with emphasis on the effects of the low-Q components. The problems related to the frequency stability of LC oscillators are discussed and the phase noise in LC oscillators is investigated with a different approach.

The last chapter is devoted to a summary of the higher-level system view of HF analog circuits, especially in the context that virtually all such high-frequency circuits are eventually integrated with considerable digital circuitry for interface, post-processing and calibration purposes – and that such integration is increasingly done on the same silicon substrate. The traditional system-level view of high-frequency components and circuits is strongly influenced by conventional (all-analog) modulation and transmission systems modeling, which is based almost exclusively on the frequency domain. The behavior of all digital systems, on the other hand, is preferably described in the time domain. While the translation between these two domains is (in theory) quite straightforward, the designers must develop a sense of how some of their choices in the analog realm eventually influence the behavior of the digital part, and vice versa. Data converters (analog-to-digital and digital-to-analog converters) naturally play an important role in this translation between domains, and Chapter 6 attempts to summarize the key criteria and parameters that are used to describe system-level performance.

The target audience of this book includes advanced undergraduate and graduate-level students who choose analog/mixed-signal microelectronics as their area of

specialization, as well as practicing design engineers. The required background that is needed to follow the material is consistent with the typical physics, math and circuits background that is acquired by the third (junior) year of a regular electrical and computer engineering (ECE) curriculum.

Solved design examples are provided to guide the reader through the decision process that accompanies each design task, emphasizing key trade-offs and eventual approximations.

A number of individuals have contributed with their time and their efforts to the creation of this textbook. In particular, both authors would like to thank Mrs Yildiz Leblebici who read the entire manuscript, carefully checked the analytical derivations throughout all chapters, and provided valuable insight as an experienced electronics teacher. The authors also acknowledge the generous support of Mr Giovanni Chiappano from austriamicrosystems A.G.[2] for offering the use of transistor parameters in numerous examples.

The idea of this book was originally launched with the enthusiastic encouragement of Dr Philip Meyler of Cambridge University Press, who saw the need for a design-oriented text in this field and patiently followed through its early development. We are deeply grateful to Dr Julie Lancashire, our publisher, for her guidance, support and encouragement over the years, leading up to the final stages of production. The editorial staff of Cambridge University Press has been wonderfully supportive throughout this project. We would especially like to thank Ms Sarah Matthews and Mr Christopher Miller for their valuable assistance.

Last but not least, the authors would also like to thank all reviewers who read all or parts of the manuscript and provided very valuable comments.

[2] Device characteristics and technology parameters provided by austriamicrosystems A.G. will be labeled with the acronym "AMS" throught the text.

1 Components of analog CMOS ICs

1.1 MOS transistors

The basic structure of an n-channel metal–oxide–semiconductor (NMOS) transistor built on a p-type substrate is shown in Fig. 1.1. The MOS transistor consists of two disjoint p-n junctions (source and drain), bridged by a MOS capacitor composed of the thin gate oxide and the poly-silicon gate electrode. If a positive voltage with sufficiently large magnitude is applied to the gate electrode, the resulting vertical electric field between the substrate and the gate attracts the negatively charged electrons to the surface. Once the electron concentration on the surface exceeds the majority hole concentration of the p-type substrate, the surface (the channel) is said to be inverted, i.e., a conducting channel is formed between the source and the drain. The carriers, i.e., the electrons in an NMOS transistor, enter the channel region underneath the gate through the source contact, leave the channel region through the drain contact, and their movement in the channel region is subject to the control of the gate voltage.

To ensure that both p-n junctions are continuously reverse biased, the substrate potential is kept lower than the source and drain terminal potentials. Note that the device structure is symmetrical with respect to the drain and source regions; the different roles of these two regions are defined only in conjunction with the applied terminal voltages and the direction of the drain current. In an n-channel MOS (NMOS) transistor, the source is defined as the n+ region which has a lower potential than the other n+ region, the drain. This means that the current flow direction is from the drain to the source. By convention, all terminal voltages of the device are defined with respect to the source potential.

The value of the gate-to-source voltage (V_{GS}) necessary to cause surface inversion (to create the conducting channel) is called the threshold voltage V_T. This quantity depends on various device and process parameters such as the work function difference between the gate and the substrate, the substrate (surface) Fermi potential, the depletion region charge concentration, the interface charge concentration, the gate oxide thickness and oxide (dielectric) permittivity, as well as the concentration of the channel implantation that is used to adjust the threshold voltage level.

If the applied gate-to-source voltage exceeds the threshold voltage of the MOS transistor, a sufficiently high concentration of electrons is achieved in the channel region, leading to surface inversion. Thus, an n-type conducting channel is formed between the source and the drain, which is capable of carrying the drain (channel) current. If a small positive voltage is applied to the drain, a current proportional to this voltage will start to flow from the drain to the source through the conducting channel. The effective

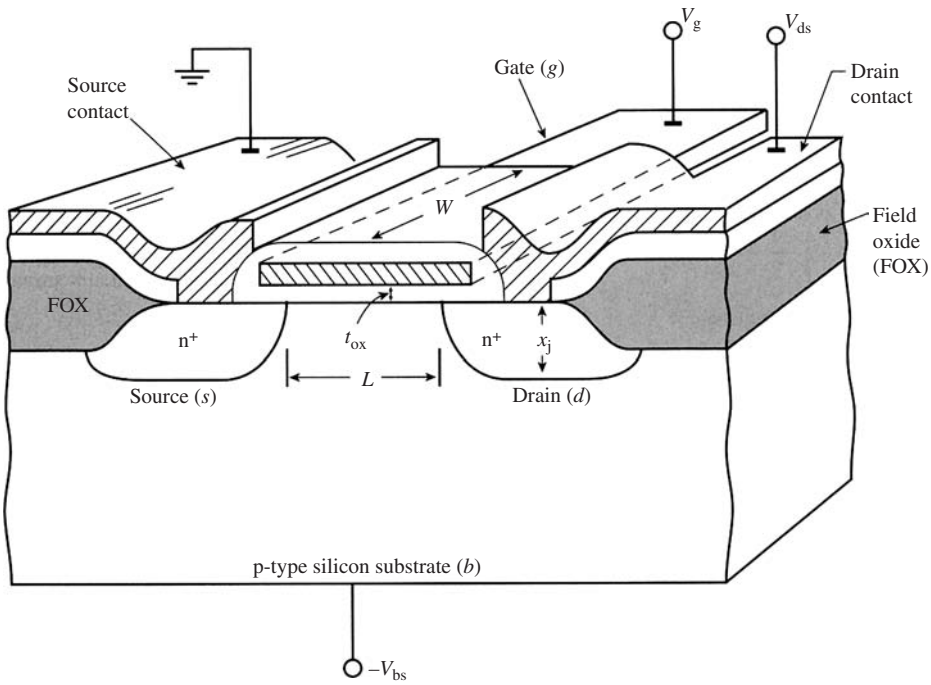


Figure 1.1 Simplified cross-section view of an n-channel MOS (NMOS) transistor (after Taur and Ning [13]).

resistivity of the continuous inversion layer between the source and the drain depends on the gate voltage. This operating mode is called the linear (or triode) mode, where the channel region acts as a voltage-controlled resistor. During this operating mode, the electron velocity in the channel is usually much lower than the drift velocity limit.

As the applied drain voltage is increased, the inversion layer charge and the channel depth at the drain end start to decrease. Eventually, when the drain voltage reaches a limit value called the saturation voltage ($V_{D(sat)}$), the inversion charge at the drain is reduced – theoretically – to zero, and the velocity of electrons – theoretically – reaches very high values, as discussed in the following sections. This event is named as the "pinch-off" of the channel. Beyond the pinch-off point, i.e., for drain voltage values larger than the saturation voltage, electrons travel in a very shallow pinched-off channel with a very high velocity, which is called the "saturation velocity". This operating regime is known as the saturation mode.

If the transistor is formed on an n-type substrate, using two p+ regions as source and drain, this structure is called a p-channel MOS (PMOS) transistor. In a PMOS transistor, the fundamental mechanisms of surface inversion and channel conduction are exactly the same as in NMOS transistors, although the majority of carriers consist of holes, not electrons. Thus, the gate-to-source voltage applied to the gate electrode to achieve surface inversion must be negative. Also, it should be taken into account that the hole mobility is considerably smaller than the electron mobility at room

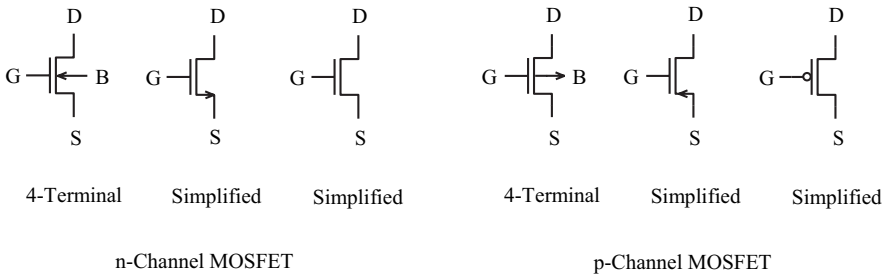


Figure 1.2 Commonly used circuit symbols for NMOS and PMOS transistors.

temperature, which leads to a smaller effective channel conductance for the PMOS transistor with the same channel dimensions. Nevertheless, the complementary nature of NMOS/PMOS biasing and operating conditions offers very useful circuit implementation possibilities, which underlines the importance and the wide-spread use of complementary MOS (CMOS) circuits in a very large range of applications.

Commonly used circuit symbols for n-channel and p-channel MOS transistors are shown in Fig. 1.2. While the four-terminal representation shows all external terminals of the device, the three-terminal symbol is usually preferred for simplicity. Unless noted otherwise, the substrate terminals are always assumed to be connected to the lowest potential for NMOS devices, and to the highest potential for PMOS devices.

1.1.1 Current–voltage relations of MOS transistors

The basic (so-called Level-1) current–voltage relations of a MOS transistor are given in most basic electronics textbooks. Since these relations contain a small number of parameters, they are convenient for hand calculations. The parameters of these expressions are:

- the mobility of electrons (or holes), μ;
- the gate capacitance per unit area, C_{ox};
- the threshold voltage of the transistor, V_T;
- the gate-length modulation coefficient, λ;
- and the aspect ratio of the transistor, (W/L).

In the following, these relations are derived with a different approach, to remind the reader of the fundamentals, and also to clarify the understanding of device behavior. In addition, the derivation presented here is based on a realistic profile of the channel-region inversion charge (as calculated from the fundamental electric field expressions), as opposed to the classical gradual channel approach which assumes linear charge profiles in the channel. This model represents the transistor under moderate to strong inversion conditions with reasonable accuracy for hand calculations, provided that the channel length is not too short and the transistor is not in the velocity saturation region.

For short-channel MOS transistors in which the carrier velocities reach saturation, i.e. approach a limit velocity, this model is no longer valid. Since a great majority of transistors realized in analog MOS integrated circuits today have channel lengths in

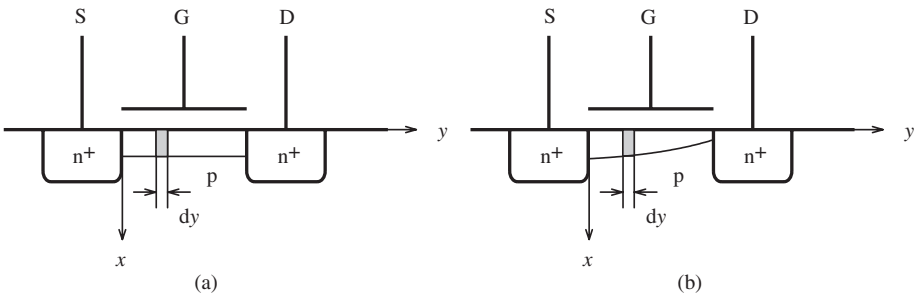


Figure 1.3 Inversion channel profiles of an NMOS transistor for (a) $V_{GS} > V_T$, $V_{DS} = 0$, (b) $V_{GS} > V_T$, $V_{DS} > 0$.

the sub-half-micron range, they may easily enter the velocity saturation region. Therefore it is necessary to derive rules to check whether a transistor is operating in the velocity saturation region or not, and to obtain expressions that are valid for velocity saturated transistors.

1.1.1.1 The basic current–voltage relations without velocity saturation

The cross-section of an NMOS transistor having an inversion layer along the channel owing to a gate–source voltage greater than the threshold voltage and zero source–drain voltage is shown in Fig. 1.3(a). Since there is no surface inversion for gate voltages smaller than the threshold voltage V_T, the value of the inversion charge density is

$$Q_i = -C_{ox}(V_{GS} - V_T)\,[\text{coulomb/cm}^2]$$

where $(V_{GS} - V_T)$ is the "effective gate voltage" for this case. The amount of the inversion charge of a transistor having a channel length L and a channel width W is

$$\bar{Q}_i = -C_{ox}WL(V_{GS} - V_T)$$

The minus signs in front of these expressions denote that this is a negative charge, since the carriers in the inversion layer of an NMOS transistor are electrons.

When we apply a positive drain voltage with respect to the source, a drain current (I_D) flows in the $-y$ direction and is constant along the channel (Fig. 1.3(b)). However, owing to the voltage drop on the channel resistance, the voltage along the channel is not constant. Although the effective gate voltage – inducing the inversion charge – at the source end of the channel is $(V_{GS} - V_T)$, it decreases along the channel and becomes equal to $(V_{GD} - V_T) = (V_{GS} - V_{DS} - V_T)$ at the drain end. If the channel voltage with respect to the source is denoted by $V_c(y)$, the effective gate voltage as a function of y can be written as

$$V_{eff}(y) = (V_{GS} - V_T) - V_c(y) \tag{1.1}$$

and the amount of the inversion charge in an infinitesimal channel segment dy is

$$d\bar{Q}_i(y) = -C_{ox}W[(V_{GS} - V_T) - V_c(y)]dy \tag{1.2}$$

Since the drain current is constant along the channel, for any y position the current can be expressed as

$$I_D = \frac{d\bar{Q}_i(y)}{dt} = \frac{d\bar{Q}_i(y)}{dy/v(y)} \tag{1.3}$$

where $v(y)$ is the velocity of electrons at position y, and can be expressed in terms of the electron mobility and the electric field strength at y:

$$v(y) = \mu_n E(y) = -\mu_n \frac{dV_c(y)}{dy} \tag{1.4a}$$

Using (1.2), (1.3) and (1.4a),

$$I_D = \mu_n C_{ox} W[(V_{GS} - V_T) - V_c(y)] \frac{dV_c(y)}{dy}$$

which gives the electric field strength as

$$E(y) = \frac{dV_c(y)}{dy} = \frac{I_D}{\mu_n C_{ox} W[(V_{GS} - V_T) - V_c(y)]} \tag{1.4b}$$

and

$$\frac{I_D}{\mu_n C_{ox} W} dy = [(V_{GS} - V_T) - V_c(y)] dV_c(y)$$

After integration from the source end ($y = 0$) to y we obtain

$$\frac{I_D}{\mu_n C_{ox} W} y = (V_{GS} - V_T) V_c(y) - \frac{1}{2} V_c^2(y) \tag{1.5}$$

From (1.5), the channel voltage $V_c(y)$ corresponding to a certain gate voltage and drain current can be deduced as

$$V_c(y) = (V_{GS} - V_T) \mp \sqrt{(V_{GS} - V_T)^2 - \frac{2I_D}{\mu_n C_{ox} W} y} \tag{1.6}$$

To fulfill the obvious physical condition $V_c(0) = 0$, the sign before the square-root term has to be taken as minus. The effective gate voltage is found from (1.1) and (1.6):

$$\begin{aligned} V_{eff}(y) &= (V_{GS} - V_T) - V_c(y) \\ &= (V_{GS} - V_T) \sqrt{1 - \frac{2I_D}{\mu_n C_{ox} W (V_{GS} - V_T)^2} y} \end{aligned} \tag{1.7}$$

It is useful to interpret (1.7) for certain cases.

(a) For $y = 0$ (at the source end of the channel) the effective channel voltage is $V_{eff}(0) = (V_{GS} - V_T)$, as expected.
(b) If $V_c(L) = V_{DS} = (V_{GS} - V_T)$, the effective channel voltage at $y = L$ is equal to zero and the channel is pinched-off at the drain end of the channel. For this case

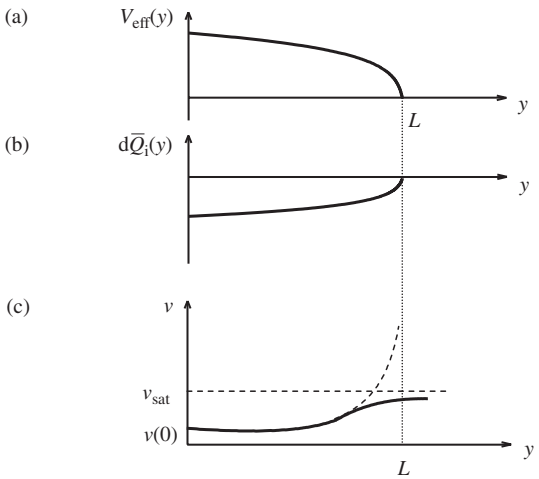


Figure 1.4 The variation of (a) the effective channel voltage, (b) the corresponding inversion charge and (c) the velocity of electrons along the channel for a transistor pinched-off at the drain end of the channel.

the value of the drain current can be found as

$$I_D = I_{D\,(sat)} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_T)^2 \tag{1.8a}$$

In this expression $\mu_n C_{ox}$ is a technology-dependent parameter and has the same value for all NMOS transistors on a chip.[1] For a given technology (1.8a) can be written as

$$I_D = I_{D\,(sat)} = \frac{1}{2}KP_n\frac{W}{L}(V_{GS} - V_T)^2 \tag{1.8b}$$

$I_{D(sat)}$ is called the "saturation current"[2] corresponding to a given gate voltage. The variations of the effective channel voltage along the channel and the corresponding inversion charge for a "saturated" transistor are plotted in Fig. 1.4(a) and (b), respectively, based on (1.2) and (1.7). Note that the inversion charge profile in Fig. 1.4(b) is found to be a second-order function of the distance y, and not a linear profile as usually assumed in the conventional gradual channel approach. While this does not influence the current–voltage relationship, the realistic charge profile will later be used for a more straightforward calculation of the channel capacitance.

[1] Similarly, $KP_p = \mu_p C_{ox}$ is a parameter common to all PMOS transistors on a chip.

[2] The phrase "saturation current" is also used to express the value of the drain current per 1 micron channel width for a certain technology, when the gate and the drain are both connected to the maximum permissible voltage for this technology. This "saturation current" takes into account all secondary effects discussed in the following sections, affecting the drain current. For example, for AMS 0.35 micron, 3.3 V technology the typical value of the saturation current is given as 540 μA/micron for NMOS transistors and 240 μA/micron for PMOS transistors, respectively.

It is useful to note and to interpret an important fact: the current remains constant along the channel but the electron density is decreasing. To maintain the current constant along the channel, i.e. to carry the same amount of charge in a certain time interval along the channel, the velocities of electrons have to increase from the source end to the drain end of the channel. Even more dramatically, the electron velocity theoretically has to reach infinity in the case of pinch-off of the channel, since the inversion charge decreases to zero. But it is known that the velocities of the electrons (and holes) cannot exceed a certain limit value and approach asymptotically this "saturation velocity", v_{sat}. In Fig. 1.4(c) the velocity of electrons along the channel is plotted. The dashed curve corresponds to the theoretical behavior, without any velocity limitation. The solid curve takes into account the velocity limitation. It is obvious that, owing to this limitation, the electron density does not decrease to zero but has a finite value to maintain the drain current with the limit velocity.

(c) If $V_c(L) = V_{DS} < (V_{GS} - V_T)$, the effective channel voltage is always positive along the channel. In other words, the channel does not pinch-off. From another point of view this can be interpreted such that the distance of the pinch-off point (L') is longer than the channel length. The drain current in this case can be solved from (1.6), for $y = L$ and $V_c(y) = V_{DS}$ as

$$I_D = \mu_n C_{ox} \frac{W}{L} \left[(V_{GS} - V_T).V_{DS} - \frac{1}{2} V_{DS}^2 \right] \tag{1.9a}$$

which reduces to (1.8a) for $V_{DS} = (V_{GS} - V_T)$, as expected. The variations of the effective channel voltage, the inversion (electron) charge density and the velocities of electrons along the channel are plotted, qualitatively, in Fig. 1.5.

For this "pre-saturation" region the variation of the drain current for small V_{DS} values can be written as

$$I_D \cong \mu_n C_{ox} \frac{W}{L} (V_{GS} - V_T) V_{DS} \quad \text{for } V_{DS} \ll (V_{GS} - V_T) \tag{1.9b}$$

This means that the drain current is proportional to the drain–source voltage. In other words the transistor acts as a resistor in this region. That is why this region is also called as the "resistive region".

(d) For $V_c(L) = V_{DS} > (V_{GS} - V_T)$, the transistor is in saturation. Assume that the transistor is pinched-off at the drain end of the channel and then the drain-source voltage increases by ΔV_{DS}. The effective channel voltage becomes equal to zero (the channel voltage becomes equal to $(V_{GS} - V_T)$) at a distance L' smaller than L. Since the current (I_D) and the electron velocity (v_{sat}) are constant in the interval $L' - L$, the field strength is approximately equal to the critical field strength, E_{cr}, which implies a linear variation of the potential along the pinched-off portion of the channel. Corresponding variations of the effective channel voltage and the charge density along the channel are shown in Fig. 1.6(a) and (b). Note that the critical field strength must have a value corresponding to the saturation

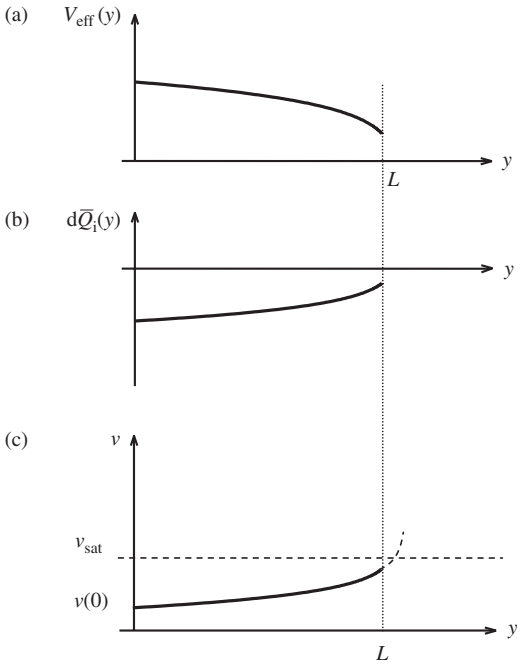


Figure 1.5 The variation of (a) the effective channel voltage, (b) the corresponding inversion charge and (c) the velocity of electrons along the channel for a transistor operating in the resistive (no pinch-off) region.

velocity, that is around $E_{cr} = 10^5$ V/cm for silicon.[3] Correspondingly, the velocity of electrons is equal to the saturation velocity along the pinched-off region of the channel, as shown in Fig. 1.6(c).

From these considerations we can conclude that

$$\Delta L = (L - L') = \frac{V_{DS} - (V_{GS} - V_T)}{E_{sat}} \tag{1.10}$$

The influence of the drain–source voltage on the drain current at a certain channel length can now be calculated:

$$\begin{aligned} I_{D\,(sat)} &= \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_T)^2 \\ \frac{dI_D}{dV_{DS}} &= \frac{dI_D}{dL}\times\frac{dL}{dV_{DS}} = \left(-I_D\frac{1}{L}\right)\times\left(-\frac{1}{E_{sat}}\right) = \Lambda \,.\, I_D \end{aligned} \tag{1.11}$$

[3] This field strength must not exceed the "breakdown field strength", which is approximately 3×10^5 V/cm for silicon. When the field strength approaches to this value, the silicon crystal structure tends to break down and the drain current increases further.

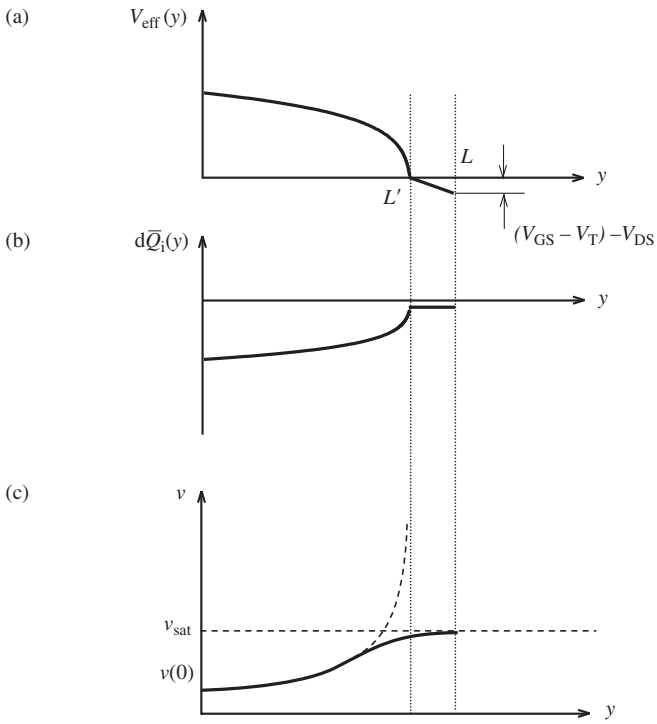


Figure 1.6 The variation of (a) the effective channel voltage, (b) the corresponding inversion charge and (c) the velocity of electrons along the channel for a transistor pinched-off before the drain end of the channel.

where

$$\Lambda = \frac{1}{L.E_{\text{sat}}} \tag{1.12}$$

Equation (1.11) gives the slope of the output characteristic curve corresponding to a certain V_{GS}, at the beginning of the saturation region, and is equal to the output conductance of the transistor (g_{ds}) for this point. Thus, the drain current corresponding to any drain–source voltage for the same gate–source voltage can be calculated as

$$I_D = I_{D\,(\text{sat})} + g_{ds}(V_{DS} - V_{DS\,(\text{sat})}) = I_{D\,(\text{sat})} + g_{ds}[V_{DS} - (V_{GS} - V_T)] \tag{1.13}$$

From (1.12) and (1.13) the drain current can be written as

$$I_D = I_{D\,(\text{sat})} \frac{1}{1 - \Lambda[V_{DS} - (V_{GS} - V_T)]} \cong I_{D\,(\text{sat})}\{1 + \Lambda[V_{DS} - (V_{GS} - V_T)]\} \tag{1.14a}$$

Using simple linear relations as shown in Fig. 1.7, it is possible to express the drain current in terms of the conventional "lambda parameter" as

$$I_D = I_{D\,(\text{sat})} \frac{1 + \lambda V_{DS}}{1 + \lambda(V_{GS} - V_T)} = \frac{1}{2}\mu_n C_{ox} \frac{W}{L}(V_{GS} - V_T) \frac{1 + \lambda V_{DS}}{1 + \lambda(V_{GS} - V_T)} \tag{1.14b}$$

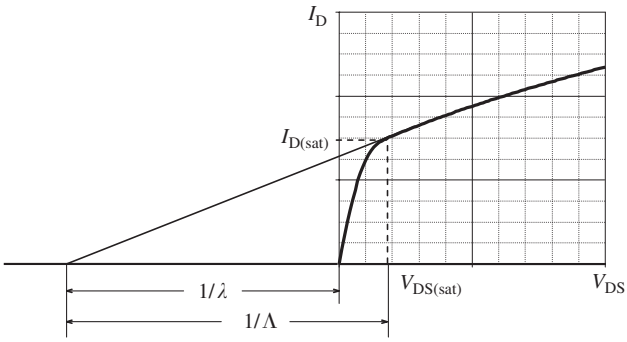


Figure 1.7 The output characteristic curve for a certain V_{GS} value and definitions of the Λ and λ parameters.

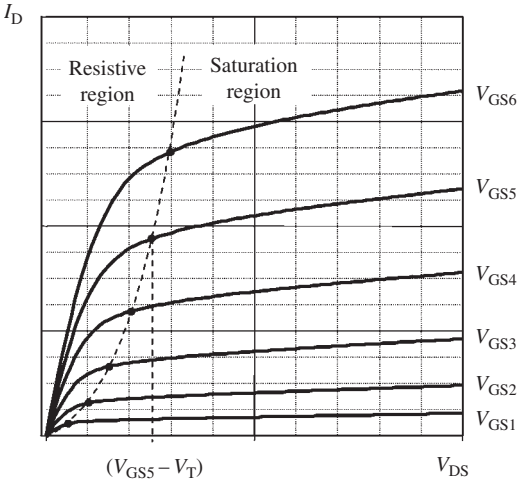


Figure 1.8 The output characteristic curves of a typical NMOS transistor. The border between the resistive region and the saturation region corresponding to $V_{DS} = (V_{GS} - V_T)$ is plotted as a dashed line.

For $V_{DS} \geq (V_{GS} - V_T)$, (1.14a) can be simplified as

$$I_D = I_{D\,(sat)}(1 + \lambda.V_{DS}) = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_T)(1 + \lambda.V_{DS}) \tag{1.15}$$

This expression is commonly used to model the "channel length modulation effect". However, it should be kept in mind that this approximation is not very accurate, especially for short-channel devices.

In Fig. 1.8 the output characteristic curves of an NMOS transistor covering these three characteristic features, namely the pre-saturation (or resistive) region, the onset of the pinch-off and the saturation region, are given. For PMOS transistors the

characteristic curves have a similar shape. But since in PMOS transistors the gate–source voltage must be negative to induce a p-type inversion layer, and consequently the polarities of the drain–source voltage and the drain current are negative, all voltages and currents on the characteristics must be marked as "negative".

The initial slope of the characteristic curve corresponding to a certain gate–source voltage can be calculated from (1.9a) as

$$\frac{dI_D}{dV_{DS}} \cong \mu_n C_{ox} \frac{W}{L}(V_{GS} - V_T)$$

and the inverse of this conductance is called the "on resistance, r_{on}" of the transistor:

$$r_{on} = \frac{1}{\mu_n C_{ox} \frac{W}{L}(V_{GS} - V_T)} \tag{1.16}$$

which expresses the series resistance exhibited by the transistor when it is used as a switch.

For analog applications, a MOS transistor is – almost – always used in the saturation region. Therefore, the parameters corresponding to this region have prime importance and will be investigated in detail, later on.

1.1.1.2 Current–voltage relations under velocity saturation

In Section 1.1.1.1 we have seen that the velocity of electrons reaches the "saturation velocity" at the drain end of the channel. In Fig. 1.9 the velocity of carriers in silicon is shown as a function of the lateral electric field strength. The saturation velocities of electrons and holes are approximately 10^7 cm/s and 8×10^6 cm/s, respectively, in bulk silicon. The initial slope of the velocity curve corresponds to the low field mobility of electrons and holes. It has been shown that the saturation velocities of electrons and holes in the inversion layer of a MOS structure are considerably lower than that in bulk silicon and are given as 6.5×10^6 cm/s and 5.85×10^6 cm/s, respectively [1].

We have also seen that velocity saturation can extend towards the source end of the channel under certain bias conditions. To understand this behavior, it is useful to consider the plot given in Fig. 1.9, which displays the generic behavior of carrier

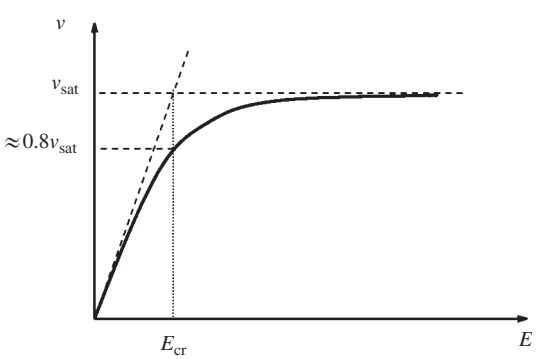


Figure 1.9 Velocity of carriers versus the lateral electric field strength in silicon.

velocity (electrons and holes) as a function of the lateral electric field strength in the channel. This plot indicates that:

- the velocity of electrons (and holes) increases proportionally with the lateral electric field strength (parallel to the direction of current) until it reaches the vicinity of the so-called "critical field strength" (E_{cr});
- the proportionality factor is called the "low-field mobility" of electrons and holes:

$$v_n = -\mu_n . E \qquad v_p = \mu_p . E \tag{1.17}$$

  The minus sign in the first expression indicates that the velocity of electrons is in the opposite direction of the electric field;
- as already mentioned, the saturation velocities of electrons and holes in the inversion layer of a MOS transistor are 6.5×10^6 cm/s and 5.85×10^6 cm/s, respectively;
- carrier mobility, critical field strength and the saturation velocity are related as

$$E_{crn} \cong \frac{v_{sat}}{\mu_n} \qquad E_{crp} \cong \frac{v_{sat}}{\mu_p} \tag{1.18}$$

- the carrier velocities corresponding to the critical field strengths are approximately equal to $0.8\times v_{sat}$;
- a final remark; the mobility of electrons (and holes) in the channel of a MOS transistor also depends on the transversal (perpendicular to the direction of the current) electric field strength [2], [3], [4]. This secondary effect must not be neglected for thin gate oxide (small geometry) devices (see Appendix A).

In a transistor operating in the pinched-off (saturated) regime, electrons travel under velocity saturation conditions at the drain end of the channel, as seen in Fig. 1.6(c). If the length of this region is only a small fraction of the channel length, the effects of the velocity saturation can be neglected. Especially for short channel devices, electrons can travel under velocity saturation conditions in a major part of the channel, even along the whole channel. In these cases the expressions derived under the assumption that the mobility of carriers is constant along the channel are no longer valid, and new expressions must be derived.

Now consider an NMOS transistor in which velocity saturation conditions are observed along the entire length of the channel, and consequently all electrons in the channel travel with the saturation velocity. Since the current is constant and the velocity of electrons is equal to the saturation velocity, the charge density also must be constant along the channel. This is equal to the charge density at the source end of the channel (1.2), and the total charge in the channel region is

$$\bar{Q}_i = WLC_{ox}(V_{GS} - V_T) \tag{1.19}$$

If this charge is being swept in t seconds, that is $t = L/v_{sat}$, the drain current under velocity saturation conditions becomes

$$I_{D(v-sat)} = WC_{ox}(V_{GS} - V_T)v_{sat} \tag{1.20}$$

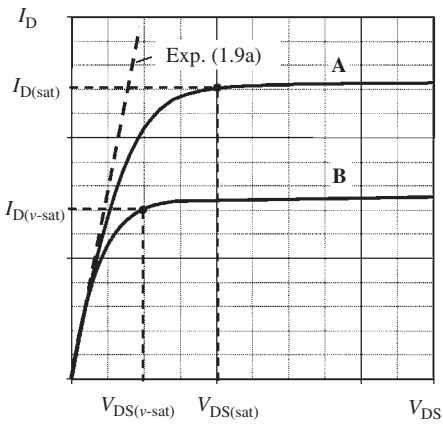


Figure 1.10 Comparison of normal saturation and velocity saturation.

From (1.18) and (1.20) we can write

$$I_{D(v-sat)} = WC_{ox}(V_{GS} - V_T)(\mu_n E_{crn}) \tag{1.21}$$

For smaller values of the field strength this expression reduces to

$$I_D = \mu_n C_{ox} \frac{W}{L}(V_{GS} - V_T)V_{DS} \tag{1.22}$$

which is identical to the current expression in the resistive region (1.9a) for small values of V_{DS}, plotted in Fig. 1.10 with a dashed line. For the rest of this curve there are two alternatives.

- If the velocity saturation does not exist, the transistor enters into the normal saturation region at $V_{DS(sat)} = (V_{GS} - V_T)$ and the current saturates, as shown with curve A, to a value given in (1.8).
- If the velocity of the electrons reaches saturation velocity, the current saturates to the value given in (1.20), as shown with curve B. For this case the drain–source voltage corresponding to the onset of the velocity saturation can be found from (1.21) and (1.22) as

$$V_{DS\,(v-sat)} = \frac{Lv_{sat}}{\mu_n} \tag{1.23a}$$

Until now we have assumed that the velocities of electrons in the channel of a velocity saturated NMOS transistor are equal to the saturation velocity that corresponds to the extreme velocity saturation. From an inspection of Fig. 1.9 we see that the saturation starts at a lower field strength and the velocity of electrons asymptotically approaches v_{sat}. Therefore it is more realistic to modify (1.23a) as

$$V_{DS\,(v-sat)} = \frac{Lkv_{sat}}{\mu_n} \tag{1.23b}$$

where k is a constant smaller than unity, (e.g., $k = 0.8$), corresponding to the critical field strength can be used as an appropriate value. Similarly, the drain current under velocity saturation conditions starts to saturate at approximately kv_{sat} and gradually increases to v_{sat}. Therefore, modifying (1.20) as

$$I_{D\,(v-sat)} = kWC_{ox}(V_{GS} - V_T)v_{sat} \tag{1.20a}$$

is more realistic.

Now we can interpret the results related to the velocity saturation.

- If $V_{DS\,(v-sat)}$ is smaller than $V_{DS\,(sat)}$, the transistor is in the velocity saturation regime. This leads us to an expression for checking if a transistor is in velocity saturation:

$$\frac{Lkv_{sat}}{\mu_n} < (V_{GS} - V_T) \rightarrow L < \frac{\mu(V_{GS} - V_T)}{kv_{sat}} \tag{1.24}$$

- According to (1.23a), for a transistor operating in the velocity saturation region, the saturation voltage is not only smaller than the normal saturation voltage, but is also independent of the gate–source voltage. This is a valuable property for circuits that have a tight supply voltage budget. For example, for an $L = 0.13\,\mu m$ transistor with $V_T = 0.25\,V$ and $\mu_n = 200\,cm^2/V.s$, the calculated value of $V_{DS\,(v-sat)}$ is 0.34 V.
- The velocity saturation is a small-geometry phenomenon. For example, assuming 0.2 V gate overdrive and $\mu_n = 400\,cm^2/V.s$ electron mobility, an NMOS transistor is subject to velocity saturation if the channel length is smaller than $0.15\,\mu m$.
- Since the hole mobility is always smaller than the electron mobility, velocity saturation occurs only for extremely short channel PMOS transistors. For example, assuming 0.2 V gate overdrive and $\mu_p = 150\,cm^2/V.s$ hole mobility value, a PMOS transistor can enter into the velocity saturation regime only if the channel length is smaller than 85 nm.
- For a certain gate length, the transistor may enter the velocity saturation regime if the gate overdrive voltage increases. For an $L = 0.13\,\mu m$ transistor with $V_T = 0.25\,V$ and $\mu_n = 200\,cm^2/V.s$, the velocity saturation occurs for $(V_{GS} - V_T) > 0.34\,V$, or for $V_{GS} > 0.59\,V$.
- According to (1.21), for a velocity saturated transistor the drain current is linearly related to the gate voltage. In Fig. 1.11(a) the gate–source voltage to drain-current transfer characteristics of a $W/L = 13\,\mu m/0.13\,\mu m$ and a $W/L = 130\,\mu m/1.3\,\mu m$ transistor are shown. The quadratic characteristic of the long channel transistor that is not subject to the velocity saturation is obvious. For the short channel transistor the characteristic is linear after the onset of velocity saturation.
- In Fig. 1.11(b) the output characteristics of a $0.13\,\mu m$ NMOS transistor are shown. Approximately equal intervals between the curves indicate the linear relation of the drain current to the gate–source voltage. The small and almost constant value of the drain–source saturation voltage is about 0.4 V (which approximately matches the calculated value). The nature of velocity saturation in small-geometry devices can be best appreciated when these curves are compared with the characteristics of a long channel (non-velocity saturated) transistor shown in Fig. 1.11(c).

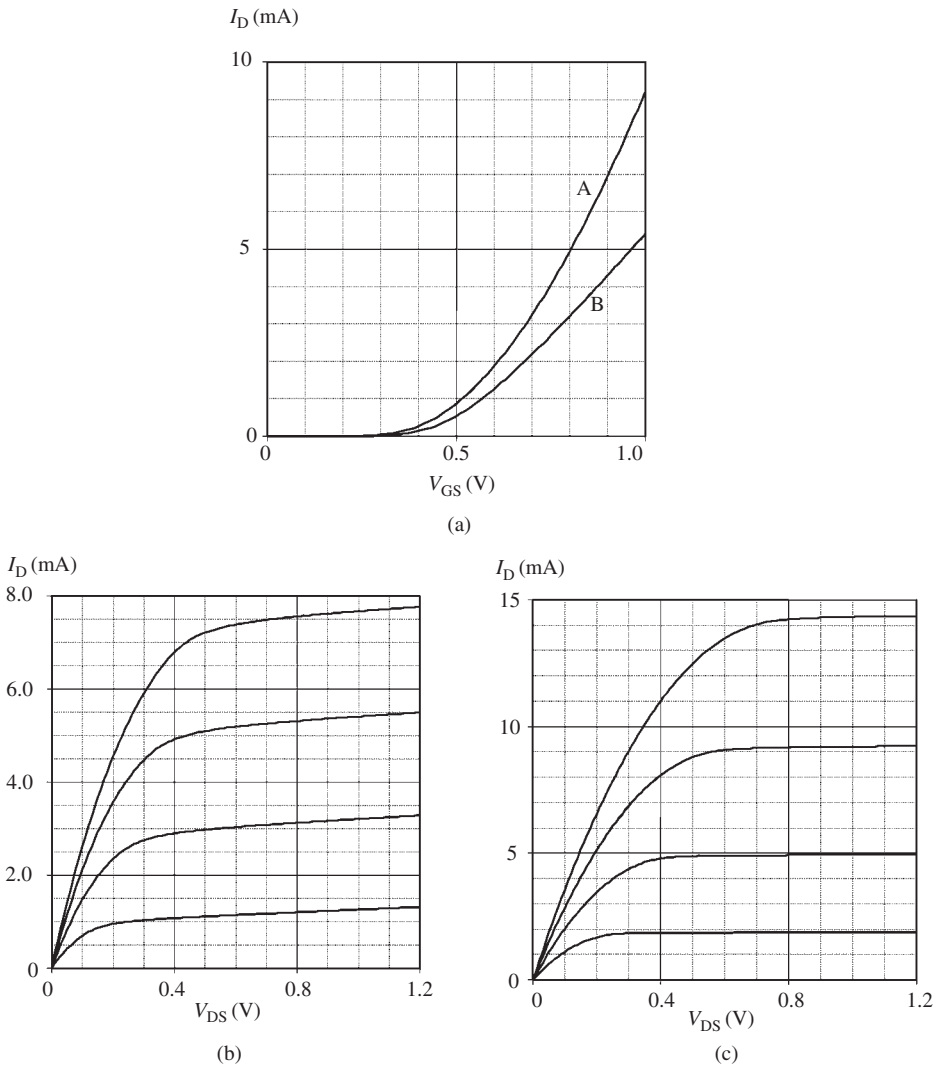


Figure 1.11 (a) $I_D = f(V_{GS})$ characteristic curves of a typical 0.13 micron technology transistor; (A) non-velocity saturated 130 µm/1.3 µm, (B) velocity/saturated 13 µm/0.13 µm transistor. (b) The output characteristics of a 13 µm/0.13 µm transistor. Note the effects of the velocity saturation. (c) The output characteristics of a 130 µm/1.3 µm transistor. ($V_{GS} = 0.6$ V to 1.2 V, with 0.2 V intervals.)

1.1.1.3 The sub-threshold regime

Until now it has been assumed that for gate–source voltages smaller than the threshold voltage V_T the drain current of the transistor is zero. But it is known that a very small drain current flows for gate–source voltages considerably smaller than the threshold voltage. This is called the "sub-threshold" current and it can be controlled by the gate–source voltage. In this section the sub-threshold regime will

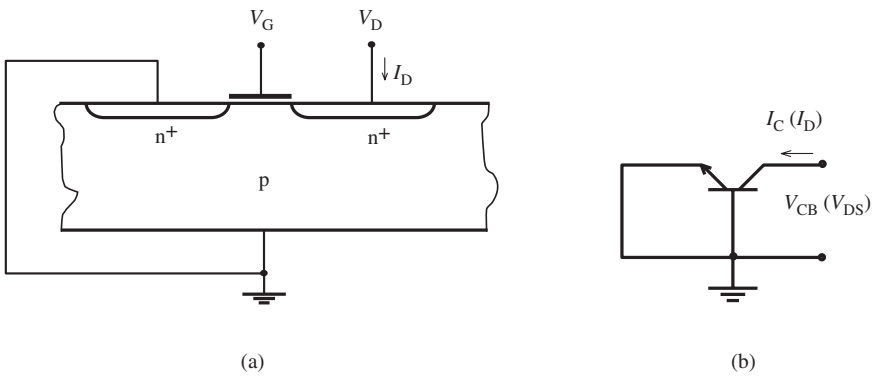


Figure 1.12 (a) Cross-section of an NMOS transistor, with its source connected to the bulk (or to the appropriate well). (b) The transistor is represented as a BJT, with its base region (bulk) connected to the emitter (the source region of the NMOS transistor).

be explained with a different approach based on the basic behaviors of the bipolar transistors.

Figure 1.12(a) shows the simplified cross-section of an NMOS transistor. When the channel is not inverted, this structure can also be interpreted as an npn bipolar junction transistor (BJT) such that the source, substrate and drain of the MOS transistor correspond to the emitter, base and collector of the bipolar transistor, respectively. The base of the BJT is connected to its emitter, as shown in Fig. 1.12(b).

From BJT theory it is known that the basic Ebers–Moll or Gummel–Poon expressions can be reduced to

$$I_C = -I_{CBS}(e^{-V_{CB}/V_T} - 1) \tag{1.25}$$

for $V_{BE} = 0$, where I_{CBS} is the reverse saturation current of the collector–base junction when the emitter is short-circuited to the base[4] [5]. I_{CBS} can be expressed as

$$I_{CBS} = qAn_i^2\left(\frac{D_n}{p_{p0}L_n} + \frac{D_p}{n_{n0}L_p}\right) \tag{1.26a}$$

where q is the unit charge, A is the cross-section of the collector junction, n_i is the intrinsic carrier density of silicon, D_n, D_p are the diffusion coefficients, L_n, L_p are the diffusion lengths of electrons and holes. p_{p0} and n_{n0} are the majority carrier concentration in the p-type and n-type regions that correspond to the base and collector regions of the bipolar transistor, also to the substrate and source/drain regions of the MOS transistor. Since the source and drain doping densities (corresponding to the majority carrier concentrations) are always much higher than that of the substrate, (1.26a) can be simplified as

$$I_{CBS} \cong qAn_i^2\frac{D_n}{p_{p0}L_n} = M\frac{1}{p_{p0}} \tag{1.26b}$$

[4] The V_T is equal to (kT/q) and must not be confused with the threshold voltage of the MOS transistor. At room temperature (kT/q) is approximately 26 mV. To prevent any confusion the threshold voltage of the MOS transistor will be shown as V_{Th} in this section.

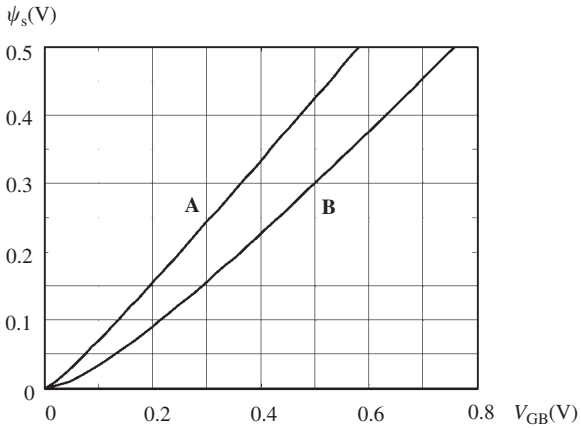


Figure 1.13 The variation of the surface potential as a function of the gate voltage for an NMOS structure ($C_{ox} = 5\times10^{-7}$ F/cm^2). The upper curve (A) was calculated for $N_A = 10^{16}$ cm^{-3}, and the lower curve (B) for $N_A = 10^{17}$ cm^{-3}.

which corresponds to the drain current of the NMOS transistor in sub-threshold.

It is known that the carrier concentrations in the channel region depend on the surface potential. The hole concentration on the surface of the channel region in terms of the surface potential (ψ_s) is given as

$$p'_{p0} = p_{p0}\, e^{(-\psi_s/V_T)} \tag{1.27}$$

For gate voltage values considerably smaller than the threshold voltage of the transistor, it can be shown that the surface potential is approximately proportional to the gate voltage, as shown in Fig. 1.13 [6]. Hence (1.27) can be written as

$$p'_{p0} = p_{p0}\, e^{(-nV_{GS}/V_T)} \tag{1.28}$$

where n has a value between 0.6 and 0.9. Equation (1.28) shows us that the gate voltage controls the hole concentration in the channel and the change of the gate voltage in the positive direction decreases the hole density in the "base region" of the npn bipolar transistor. In other words, p_{p0} in (1.26a) decreases to p'_{p0} as a function of the gate voltage. Now (1.26b) can be written as

$$I_{CBS} = \frac{M}{p_{p_0}}\, e^{nV_{GS}/V_T} \tag{1.29}$$

and from (1.25)

$$I_D = -\frac{W}{L} I_{D0}.\, e^{nV_{GS}/V_T}(e^{-V_{DS}/V_T} - 1) \tag{1.30}$$

where I_{D0} is a structural parameter (constant) of the MOS transistor and will be shown as I_0 in the following.[5]

[5] Note that (a) M is proportional to W and (b) the recombination rate of the electrons injected into the gate region – which reduces the number of electrons reaching the drain – is proportional to 4.

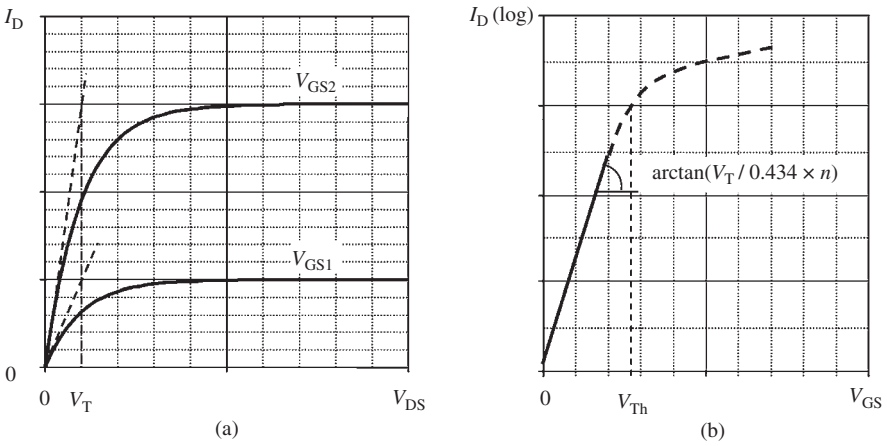


Figure 1.14 (a) The sub-threshold drain current as a function of the drain–source voltage for two different gate voltage values ($V_{GS2} > V_{GS1}$). (b) The drain current as a function of the gate–source voltage plotted on a logarithmic I_D scale.

This expression can be evaluated as follows.

- For a given value of V_{GS}, the drain current can be plotted as a function of V_{DS}, as shown in Fig. 1.14(a). The "corner" of this curve corresponds to $V_T = 26$ mV. The drain current starts to "saturate" after this point.
- The "family" of curves corresponding to different values of V_{GS} forms the "output characteristics" of the MOS transistor operating in the sub-threshold regime. The separations between these curves vary exponentially, similar to the output characteristics of a bipolar transistor driven with the base-emitter voltage (not base current).
- In Fig. 1.14(b) the variation of the drain current as a function of the gate–source voltage on a logarithmic axis is shown. For $V_{DS} \gg V_T$,

$$I_D \cong I_0 \, e^{nV_{GS}/V_T}$$

$$\log I_D = \log I_0 + \frac{nV_{GS}}{V_T} \log e = \log I_0 + \frac{nV_{GS}}{V_T} \times 0.434$$

and the inverse slope of the plot, that is usually expressed as the "sub-threshold slope" in [mV/decade]

$$S[\text{mV/dec.}] = \frac{V_T[\text{mV}]}{0.434 \times n} \tag{1.31}$$

The numerical values of the sub-threshold slope for $n = 0.7$ and 0.8 are $S = 85.2$ and 74.9 [mV/dec.], respectively.

- This slope is constant up to the vicinity of the threshold voltage. Then the transistor leaves the sub-threshold regime and enters the strong inversion regime. This part of the curve is plotted with a dashed line.

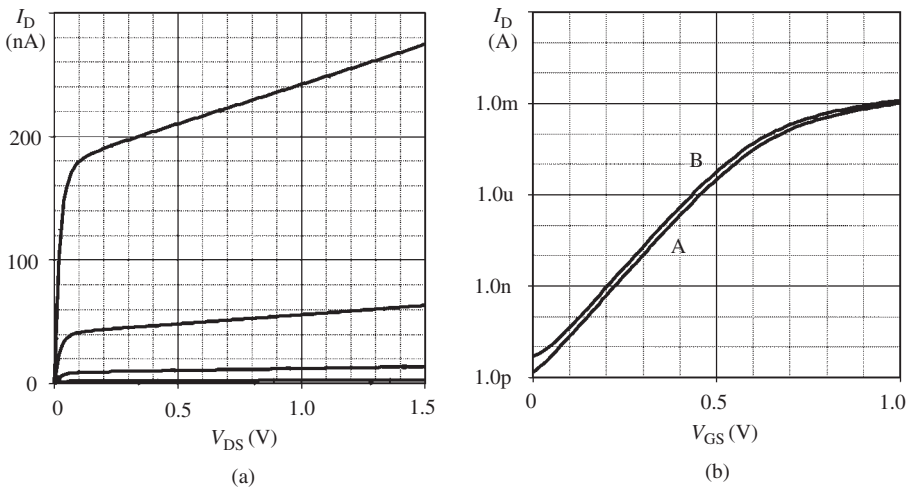


Figure 1.15 The sub-threshold characteristics of an AMS 0.35, NMOS transistor ($W = 35\,\mu m$, $L = 0.35\,\mu m$). (a) The output characteristics ($V_{GS} = 0.25\,V$ to $0.4\,V$, steps: $0.05\,V$). (b) $I_D = f(V_{GS})$ (A: $V_{DS} = 0.5\,V$, B: $V_{DS} = 3\,V$). The sub-threshold slope is 75 mV/dec.

In Figure 1.15(a) and (b) the output characteristic curves and the drain current-gate voltage characteristics of a $0.35\,\mu m$ NMOS transistor are given for operating conditions in the sub-threshold regime, as obtained from PSpice simulations. These curves are (in principle) in good agreement with the curves given in Figure 1.14. The width of the pre-saturation region on the output characteristic curves is of the order of V_T, as expected. This is valuable information for low supply-voltage budget cases.

The drain current of a given transistor under sub-threshold conditions is always considerably lower than that of the normal (in inversion) operation. Since the parasitic capacitances of the device are the same for both of these cases, the high-frequency performance of a MOS transistor operating in the sub-threshold regime is always inferior compared to the normal operation. Therefore, the sub-threshold operation must be considered as suitable only for low-frequency and low-power applications.

1.1.2 Determination of model parameters and related secondary effects

The current–voltage relations given in the previous sections are the most basic current–voltage relations of a MOS transistor. They are simple enough for hand calculations and suitable for understanding the basic behavior of devices and basic circuits containing MOS transistors. To use these expressions for hand calculations, the model parameters in them, namely μ, C_{ox}, V_T , Λ (or λ) and certainly the gate dimensions (W and L) must be known. In this section we will discuss how to determine the numerical values of these basic parameters, and related secondary effects that must be taken into account.

1.1.2.1 Mobility

As seen from (1.9b), the drain current of an NMOS transistor is directly influenced by the mobility of electrons (similarly, by the mobility of holes for a PMOS transistor) in the inversion channel. But the mobility is not a "constant". It depends on the manufacturing process and the bias conditions of the transistor. The low-field mobility values for a certain process are given by the manufacturer among the high-level parameter sets, for example BSIM3. The numerical values of the low-field mobility for electrons and holes are usually in the range of 200–500 cm^2/V.s and 70–150 cm^2/V.s, respectively.[6] Reliable average values of mobility for a specific process can be found in the high-level parameter sets (for example BSIM3) supplied by the manufacturers.

Another important issue is the traversal electrical field dependence of the mobility. It has been shown that carrier mobilities decrease with the transversal field strength that depends on the thickness of the gate oxide, the value of the threshold voltage and the gate voltage[7] [3], [4]. This effect is more pronounced for small-geometry (thin gate oxide) processes. For example, for $T_{ox} = 5$ nm, $V_T = 0.5$ V and $V_{GS} = 1$ V, the electron mobility decreases to approximately 75% of its low-field value. The obvious result of this important secondary effect is a considerable discrepancy of the drain current from the value calculated from (1.8a), assuming constant mobility. In Fig. 1.16 the $I_D = f(V_{GS})$ curve (the transfer characteristic) of a MOS transistor is given with and without the influence of the transversal field. From the comparison of these curves it can be seen that

- owing to the transversal field the drain current decreases considerably for high gate-voltage values;
- owing to this effect the transfer characteristic appears to be more "linear" than quadratic;
- the slope of the transfer characteristic at a certain operating point (i.e. the measure of the gate–source voltage dependence of the drain current, which is called the "transconductance parameter" of the transistor) decreases considerably.

This effect is obviously a small-geometry problem and must not be overlooked when the gate oxide thickness is smaller than 10 nm. For hand calculations, it is convenient to use the μ value corresponding to the actual gate–source bias voltage of the transistor in the circuit.

1.1.2.2 Gate capacitance

The gate capacitance corresponds to the maximum value (i.e. for accumulation or strong inversion) of the gate-bulk capacitance per unit area. Its value can be calculated as

$$C_{ox} = \frac{\varepsilon_0 \varepsilon_{ox}}{T_{ox}} \tag{1.32a}$$

[6] In the model parameter lists released by the manufacturers, the mobility is given in [cm^2/V.s] or [m^2/V.s]. Care is necessary to use the same dimension for length in calculations.

[7] See Appendix A.

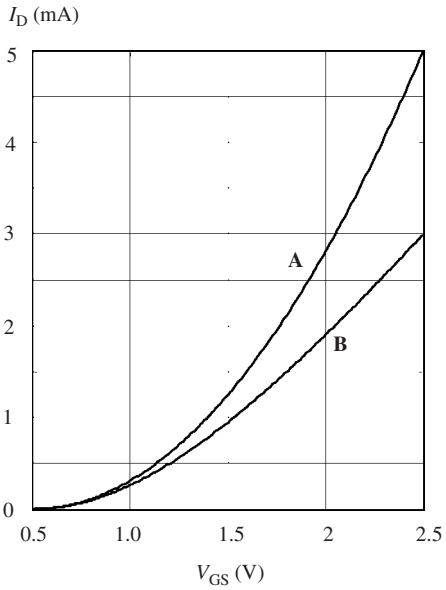


Figure 1.16 The transfer characteristic of an NMOS transistor ($\mu_{n0} = 500$ cm^2/V.s, $t_{ox} = 5$ nm, $V_T = 0.5$ V and $W/L = 0$) when the transversal field effect is neglected (A) and not neglected (B).

If we insert the values of ε_0 and ε_{ox} (8.85×10^{-14} F/cm and 3.9 for silicon dioxide as the gate insulator) and express the gate oxide thickness in nm for convenience, (1.32a) can be arranged as

$$C_{ox} = \frac{34.5}{T_{ox}[\text{nm}]} \times 10^{-7} \, [\text{F/cm}^2] \tag{1.32b}$$

The value of the gate capacitance is one of the most accurately determined parameters of a MOS transistor and depends only on the thickness of the gate oxide, T_{ox}. The value of T_{ox} is usually in the range of 2 to 8 nm for sub-micron transistors.[8] An important issue related to the gate oxide thickness is the electrostatic breakdown of the gate dielectric. The breakdown field strength for silicon dioxide is given as 10 MV/cm, which corresponds to 1 V per nm. This high sensitivity of the gate oxide with respect to breakdown (an irreversible device failure mechanism) usually requires special measures to prevent the gate-bulk voltage reaching the breakdown value.

1.1.2.3 Threshold voltage

The threshold voltage, from the point of view of the circuit designer, is the gate–source voltage for which an appreciable drain current starts to flow. It is well known from device physics that the value of the threshold voltage depends on several structural

[8] The production tolerance of the gate oxide thickness is usually within $\pm 10\%$. Consequently the "most robust" parameter of a MOSFET, C_{ox}, can have a value 10% higher or lower than the given value.

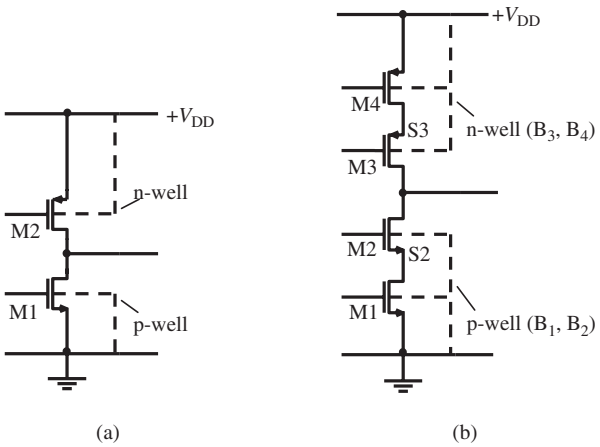


Figure 1.17 (a) Example where transistors (M1 and M2) are not subject to substrate bias effect. (b) Example where M2 and M3 are subject to substrate bias effect.

parameters such as the doping concentration of the bulk (or well), the type of the gate material, the oxide and oxide–silicon interface quality, the threshold adjustment implant dose, etc. However, many of these process parameters are considered as the "intellectual property" of the IC manufacturer, and hence not available to circuit designers. The "basic" threshold voltage is given in the parameter list of advanced simulation models as (VTH0). It is the threshold voltage of a large transistor when the source is connected to the bulk (or to the appropriate well). Several effects influencing the value of the threshold voltage have been theoretically investigated in detail [7], [8] and also included into the advanced models. These secondary effects can be ignored for the hand calculations, with one exception: the substrate bias effect.

The MOS transistors in a CMOS circuit are usually operated with the source connected to the n-well for NMOS transistors and to the p-well for PMOS transistors, as shown in Fig. 1.17(a). But in some cases a series connection of same-type transistors is needed, as shown in Fig. 1.17(b). In this case the source regions of M1 and M4 can be connected to the appropriate wells. But the source of M2 is connected to the drain of M1, and has to be separated from the well. This means that there is a positive voltage between the source of M2 and the p-well. Similarly, there is a negative voltage between the source of M3 and the n-well. This bias (V_{SB}) increases the magnitude of the threshold voltage. It has been shown that the increase of the threshold voltage depends on a number of structural parameters such as the gate oxide thickness, the doping concentration of the substrate and the doping properties of the channel region, and modeled in detail in advanced models.

For hand calculations the substrate bias effect can be ignored and left to the fine-tuning of the circuit as a whole with SPICE. If the accurate values of the threshold voltages are needed in the hand-calculation stage, it is possible to use SPICE simulations to obtain the actual threshold voltage of a transistor under a specified substrate

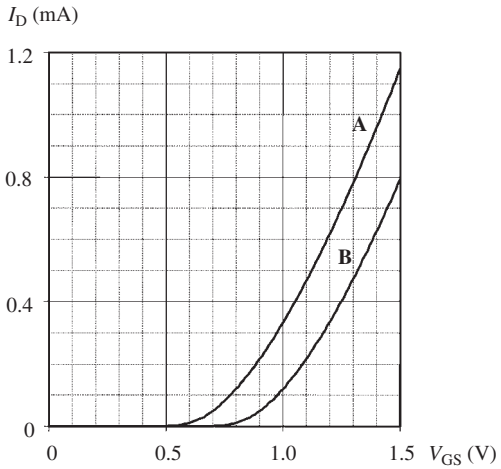


Figure 1.18 The $I_D = f(V_{GS})$ transfer curves of an $L = 0.35\,\mu m$, $W = 10\,\mu m$ NMOS transistor obtained from PSpice simulation with BSIM3-v3 parameters (A) for no substrate bias, (B) for 1 V substrate bias.

bias condition. Assume that the drain–source voltage of M1 in Fig. 1.17(b) is 1 V. It means that the source of M2 is 1 V positive with respect to ground. In Fig. 1.18 the $I_D = f(V_{GS})$ curves of a transistor for $V_{SB} = 0$ and $V_{SB} = 1$ V are shown, obtained from a simple SPICE simulation. From these curves the threshold voltages for no substrate bias and for a 1 V substrate bias can be found to be 0.5 V and 0.7 V, respectively. This example shows that the substrate bias can significantly change the value of the threshold voltage, and must not be easily ignored.

1.1.2.4 Channel length modulation factor

It is known that the drain current does not remain constant in the saturation region but exhibits a gradual increase with the drain–source voltage. This effect is modeled by the λ parameter for low-level models. In these models it is assumed that the saturation region tangents of the output characteristic curves intersect the horizontal axis at the same point and the voltage corresponding to this point is $-1/\lambda$. This effect is modeled in detail in advanced high-level models. In Fig. 1.19 the output characteristics of an NMOS transistor obtained from SPICE simulation with BSIM3-v3 parameters are shown. The attempt to determine the λ parameter (or similarly the Λ parameter defined in Section 1.1.1) from these curves shows that for a realistic transistor the tangents do not intersect the horizontal axis at the same point and therefore it is not possible to find a single λ (or Λ) parameter that is valid for all operating conditions.

On the other hand λ (or Λ) is a useful parameter for hand calculations to determine the DC operating points with better accuracy and to calculate the small-signal parameters corresponding to a certain operating point. A solution is to determine the value of λ corresponding to the actual operating point from the characteristic curves.

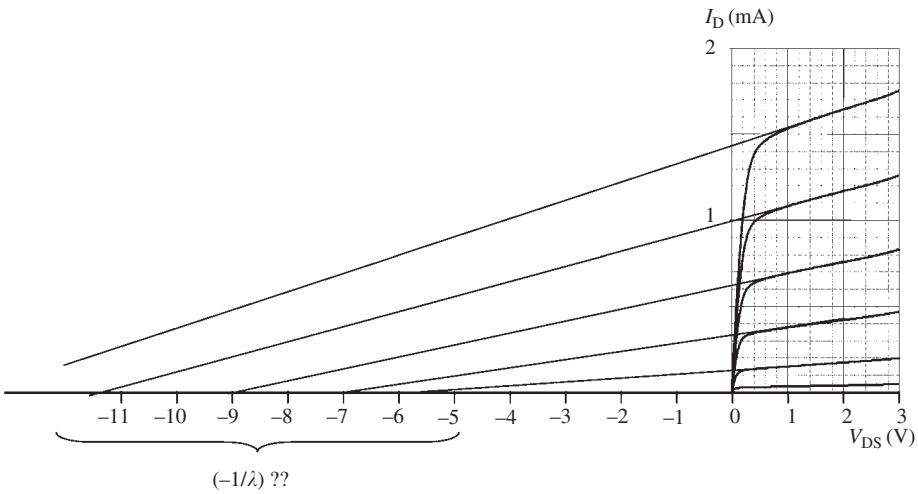


Figure 1.19 Determination of the λ parameter from the output characteristic curves obtained from SPICE simulation with BSIM3-v3 parameters. Note that it is not possible to find a unique λ parameter value that is valid for all operating points.

1.1.2.5 Gate length (L) and gate width (W)

The device parameters (such as mobility, threshold voltage, etc.) reviewed in the previous sections are usually not considered as design parameters; their values are mostly dictated by the process technology, and the circuit designer has little or no influence on the setting of their values. The only significant design freedom is found in choosing the dimensions of the gate, namely the gate length (L) and the gate width (W). These dimensions are usually called the "drawn" geometries, i.e. the dimensions on the mask layout. Note that the actual dimensions on the chip are always somewhat different from these values, as a result of the lithography and etching steps of fabrication.

It will be shown in the following sections that the high-frequency performance of a MOS transistor strongly depends on the gate length of the transistor; shorter channel lengths provide better high-frequency performance. Consequently it is wise to use the minimum possible channel length value[9] for a certain technology, if the high-frequency performance has prime importance. It is also known that the small-signal output resistance of a MOS transistor is smaller for shorter channel lengths. But a high output resistance is desired for certain types of circuits, e.g. for current sources. Therefore, a trade-off must be made and the designer has to decide on the appropriate channel length for each transistor, sometimes following several iterations guided by hand calculations and/or by simulations.

It is known that the saturation drain current is proportional to the aspect ratio (W/L) for a non-velocity saturated MOS transistor and proportional to the channel width (W)

[9] The usable minimum channel length is usually a part of the name of the technology. For example, TSMC018 indicates that for this technology offered by TSMC, the minimum "drawn" channel length is 0.18 micrometers.

for a velocity saturated transistor:

$$I_D = I_{D\,(sat)} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_T)^2 = \frac{1}{2}KP\frac{W}{L}(V_{GS} - V_T)^2 \tag{1.8}$$

$$I_{D\,(v-sat)} = kWC_{ox}(V_{GS} - V_T)v_{sat} \tag{1.20b}$$

The gate–source voltage dependence of the drain current, called the transconductance parameter, is also proportional to *W/L* for a non-velocity saturated transistor and proportional to the channel width *W* for a velocity saturated transistor:

$$g_{m(sat)} = \frac{dI_D}{dV_{GS}} = \mu_n C_{ox}\frac{W}{L}(V_{GS} - V_T) = \sqrt{2\mu C_{ox}\frac{W}{L}I_D} \tag{1.33}$$

$$g_{m(v-sat)} = kWC_{ox}v_{sat} \tag{1.34}$$

These four expressions indicate that a wider channel provides higher drain current and higher transconductance. But a wide channel means higher area consumption on the chip and higher parasitic capacitances that affect the high-frequency performance of the device. Therefore, there is another design trade-off related to the dimensions of the transistors in a circuit.

For a given process technology (i.e., for a given $KP = \mu C_{ox}$ value), (1.8) can be normalized as

$$K = \frac{I_D}{(KP/2)} = \frac{W}{L}(V_{GS} - V_T)^2 \tag{1.35}$$

In Fig. 1.20, the gate overdrive is plotted against the aspect ratio (*W/L*) for different values of the *K* parameter. From this figure, the necessary aspect ratio corresponding to an "acceptable" gate overdrive can be easily found, for a certain drain current (or vice versa). Note that the gate overdrive voltage is also the saturation voltage of the transistor, and must be "acceptably" small, especially if there are a number of transistors sharing the total DC supply voltage as in Fig. 1.17(b).

1.1.3 Parasitics of MOS transistors

The "parasitics" of a MOS transistor correspond to all non-intentional and non-avoidable passive or active devices that exist around the MOS transistor. Namely, they are the parasitic capacitances between different regions of the MOS transistor, the resistances associated with the several terminals of the device, the p-n junctions (for example the drain–well junction) that are the integral parts of the MOS transistor, and similar bipolar structures.[10] These parasitic p-n junctions and bipolar transistors are normally reverse-biased such that their currents are negligibly small.

[10] It is known that the so-called "latch-up" effect in CMOS inverters is a result of the parasitic bipolar transistors that are integral parts of the structure.

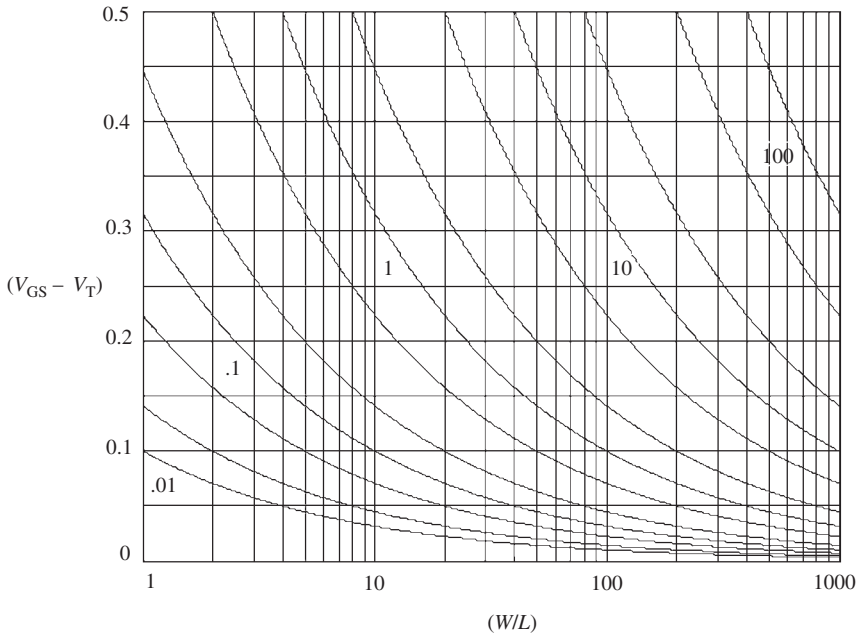


Figure 1.20 Variation of $(V_{GS} - V_T)$ as a function of the aspect ratio. The parameter $K = I_D/(KP/2)$ is plotted from $K = 0.01$ to $K = 100$ with intervals 1, 2, 5. Note that for PMOS transistors the absolute values of $(V_{GS} - V_T)$ and K must be used.

But for VLSI circuits containing millions of MOS transistors, the sum of these reverse-biased junction currents, also called the substrate currents, becomes a severe problem.

1.1.3.1 Parasitic capacitances

In Fig. 1.21 the parasitic capacitances and resistances of a MOS transistor are shown. These parasitic elements are all geometry dependent. Some of them, for example the junction capacitances, are also technology and bias dependent. In this section, the parasitic capacitances of MOS transistors will be examined under saturation conditions.

The total gate–source capacitance, C_{gs}

The gate–source capacitance of a MOS transistor biased in the saturation region is the sum of the two components; the gate capacitance (C_g) corresponding to the carrier charge in the inversion layer that is induced and controlled by the gate–source voltage and the gate–source overlap capacitance (C_{gso}).

The incremental inversion layer charge of a transistor operating in the saturation region as a function of the channel voltage, $V_c(y)$ was obtained as

$$d\bar{Q}_i(y) = -C_{ox}W[(V_{GS} - V_T) - V_c(y)]dy \tag{1.2}$$

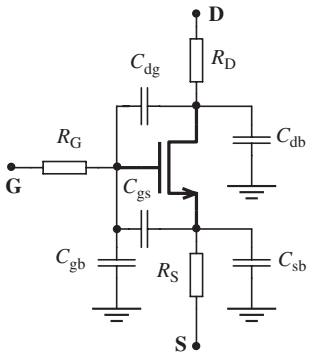


Figure 1.21 The parasitic capacitances and resistances of a MOS transistor.

If we insert the channel voltage that was given with (1.6) into (1.2), we obtain

$$\mathrm{d}\bar{Q}_\mathrm{i}(y) = -C_\mathrm{ox}W(V_\mathrm{GS} - V_\mathrm{T})\sqrt{1 - \frac{2I_\mathrm{D}}{\mu_\mathrm{n}C_\mathrm{ox}W(V_\mathrm{GS} - V_\mathrm{T})^2}y}.\mathrm{d}y \tag{1.36}$$

At the onset of saturation, i.e. $y = L$, (1.36) can be simplified as

$$\mathrm{d}\bar{Q}_\mathrm{i}(y) = -C_\mathrm{ox}W(V_\mathrm{GS} - V_\mathrm{T})\sqrt{1 - \frac{y}{L}}.\mathrm{d}y \tag{1.37}$$

If we integrate (1.37) along the channel we obtain the total inversion charge in the channel as

$$\bar{Q}_\mathrm{i} = \frac{2}{3}C_\mathrm{ox}WL(V_\mathrm{GS} - V_\mathrm{T}) \tag{1.38}$$

and the corresponding capacitance as

$$C_\mathrm{g} = \frac{\mathrm{d}\bar{Q}_\mathrm{i}}{\mathrm{d}V_\mathrm{GS}} = \frac{2}{3}C_\mathrm{ox}WL \tag{1.39}$$

The physical parameters necessary to calculate the value of the gate–source overlap capacitance, C_gso, are usually not readily available. But the value of this capacitance per unit width of the channel (CGSO) is given among the model parameters, usually in F/m. Hence the total gate–source capacitance can be written as

$$C_\mathrm{gs} = \frac{2}{3}C_\mathrm{ox}WL + C_\mathrm{gso} = \frac{2}{3}C_\mathrm{ox}WL + (\mathrm{CGSO} \times W)$$

or for a certain channel length L,

$$C_\mathrm{gs} = W\left(\frac{2}{3}C_\mathrm{ox}L + \mathrm{CGSO}\right) = C_\mathrm{ox}WL\left(\frac{2}{3} + \frac{\mathrm{CGSO}}{C_\mathrm{ox}L}\right) = C_\mathrm{ox}WL \times k_\mathrm{ol} \tag{1.40}$$

Example 1.1 The capacitance-related parameters of a typical 0.18 micron technology are

gate oxide thickness: TOX = 4.2E-9 [m],
gate–source overlap capacitance: CGSO = 1.21E-10 [F/m].
Calculate the gate–source capacitance of a 20 μm/0.18 μm transistor.
Let us first find the value of the gate capacitance from (1.32a):

$$C_{ox} = \frac{34.5}{4.2} \times 10^{-7} = 8.2 \times 10^{-7}\ [\mathrm{F/cm^2}]$$

Since TOX and CGSO are given in terms of [m], C_{ox} must be converted to [F/m^2]:

$$C_{ox} = 8.2 \times 10^{-3} [\mathrm{F/m^2}]$$

From (1.40)

$$C_{gs} = 8.2 \times 10^{-3} \times (20 \times 10^{-6}) \times (0.18 \times 10^{-6}) \left(\frac{2}{3} + \frac{1.21 \times 10^{-10}}{(8.2 \times 10^{-3}) \times (0.18 \times 10^{-6})} \right)$$
$$= 29.5 \times 10^{-15} (0.749) = 22.1 \times 10^{-15}\ \mathrm{F} = 22.1\ \mathrm{fF}$$

An important piece of information obtained from this example is that the overlap capacitance is more than 10% of the total. It means that the overlap capacitance for short channel transistors is not a secondary component of C_{gs}, and must not be ignored.

Problem 1.1 *Derive an expression for the gate–source capacitance of a MOS transistor operating in the velocity saturation region.*
(Answer: $C_{gs} = C_{ox}WL + (CGSO \times W)$)

The drain–gate capacitance, C_{dg}

The drain–gate capacitance of a MOS transistor consists of only the drain–gate overlap capacitance. Although this is a small capacitance in magnitude, it has a very important influence on the high-frequency performance of the MOS transistors, as will be explained later on. Similar to the gate–source overlap capacitance, its value per unit width (usually per meter) is given in the model parameter lists as (CGDO). Therefore the value of the drain–gate capacitance of a MOS transistor can be found as

$$C_{dg} = (\mathrm{CGDO}) \times W\,[\mathrm{m}]\ [\mathrm{farad}] \tag{1.41}$$

The gate-substrate capacitance, C_{gb}

Its value is given as CGBO, i.e. the gate-substrate capacitance per unit length of the channel. Since the channel lengths of transistors are usually small, the gate-substrate capacitance always has a small numerical value compared to the other parasitic capacitances and can be easily neglected for hand calculations.

The drain–substrate and source–substrate capacitances, C_{db} and C_{sb}

C_{db} (and similarly C_{sb}) is the total junction capacitance of the drain–substrate (or drain–well) junction, that must be biased in the reverse direction. These junctions have two parts that are different in nature; the bottom junction and the side-wall junctions. It is known from the basic p-n junction theory that the capacitance of a reverse-biased junction decreases with the magnitude of the bias voltage. The variation of the capacitance depends on the doping properties of the junction:

$$C_j(V) = C_j(0)\left(1 - \frac{V}{\varphi_B}\right)^{-m}$$

where $C_j(0)$ is the value of the capacitance for zero bias, φ_B is the built-in junction potential and m is the grading coefficient of the junction, whose value varies between 1/3 and on 1/2, depending on the doping profiles.

In the model parameter lists, the value of the junction capacitance for zero bias is given in two parts: (CJ) is the bottom junction capacitance per unit area (usually in farad per square meter) and (CJSW) is the side-wall junction capacitance per unit length (usually in farad/meter). The grading coefficients for these junctions are also given separately as (MJ) and (MJSW). The built-in voltage is represented by (PB).

For hand calculations, this capacitance can be taken into account with some simplifying assumptions. Although somewhat pessimistic, using the zero-bias value of the junction capacitance is a simple but convenient approach.

Example 1.2 The simplified plan view of a MOS transistor fabricated with the AMS 0.35 micron geometry is given in the figure. The dimensions of the transistor are $L = 0.35\,\mu m$, $W = 20\,\mu m$. The widths of the source and drain regions are typically $X = 0.85\,\mu m$. The related model parameters are given next to the figure. Calculate the source and drain junction capacitances (a) for $V = 0$, (b) for $V = -1$ V.

(a) The bottom area of the junction:

$A = X \times W = (0.85\times10^{-6}) \times (20\times10^{-6}) = 17\times10^{-12}\ [m^2]$

The bottom junction zero bias capacitance:

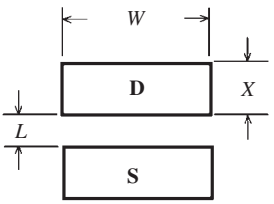
$C_j(0) = (CJ) \times A = (9.4\times10^{-4})(17\times10^{-12}) = 159.8\times10^{-16}\ [F] = 15.98\ [fF]$

The total length of the side-wall junction:

$L_{SW} = 2\times(X + W) = 2\times(0.85\times10^{-6} + 20\times10^{-6}) = 41.7\times10^{-6}\ [m]$

The total side-wall capacitance:

$C_{jsw}(0) = (CJSW) \times L_{SW} = (2.5\times10^{-10}) \times (41.7\times10^{-6}) = 10.425\ [fF]$



CJ	9.4 e-4
CJSW	2.5 e-10
MJ	3.4 e-1
MJSW	2.3 e-1
PB	6.9 e-1

The total zero bias junction capacitance:
$C_{jT}(0) = C_j(0) + C_{jsw}(0) = 26.4$ [fF]

(b) The bias-dependent factor for the bottom capacitance for 1 V reverse bias:

$$\left(1 - \frac{(-1)}{0.69}\right)^{-0.34} = 0.737$$

The bottom junction capacitance for 1 V reverse bias:
$C_j(-1) = 15.98 \times 0.737 = 11.78$ [fF].
The bias-dependent factor for the side-wall capacitance for 1 V reverse bias:

$$\left(1 - \frac{(-1)}{0.69}\right)^{-0.23} = 0.814$$

The side-wall junction capacitance for 1 V reverse bias:
$C_{jsw}(-1) = 10.425 \times 0.814 = 8.48$ [fF].
The total junction capacitance for 1 V reverse bias:
$C_{jT}(-1) = 11.78 + 8.48 = 20.26$ [fF].

That is 6.14 fF (23.3%) smaller than the zero bias value.

1.1.3.2 The high-frequency figure of merit

An important definition for a MOS transistor that is related to the parasitic capacitances of the device is the "high-frequency figure of merit", f_T. It is known that the low-frequency input current of a MOS transistor is practically zero, and consequently the low-frequency current gain is infinite. At higher frequencies, on the other hand, the capacitive current that flows into the gate terminal becomes non-negligible. Hence, the current gain decreases at high frequencies. f_T is defined as the frequency for which the magnitude of the current gain is equal to unity.

The capacitive (small-signal) input current of a MOS transistor can be expressed as

$$i_i = v_{gs}.sC_{gs}$$

where C_{gs} is the total gate–source capacitance of the transistor.

The maximum value of the output signal current can be written from the definition of the transconductance, as $i_o = g_m\, v_{gs}$. From these expressions the current gain can be written as

$$A_i = \frac{i_o}{i_i} = \frac{g_m}{sC_{gs}}$$

Consequently, f_T is found to be

$$f_T = \frac{1}{2\pi}\frac{g_m}{C_{gs}} \tag{1.42}$$

It is instructive to compare this expression for a non-velocity saturated and a velocity saturated transistor.

For a non-velocity saturated transistor the input capacitance that was given in (1.40) can be found as

$$C_{gs}\,WLC_{ox}k_{ol} \;=\; WLC_{ox}\left(\frac{2}{3}+\frac{\text{CGSO}}{LC_{ox}}\right)$$

The transconductance of a non-velocity saturated transistor was given in (1.33). From (1.40) and (1.33), f_T can be arranged as

$$f_T = \frac{1}{2\pi}\frac{\mu(V_{GS}-V_T)}{L^2\left(\frac{2}{3}+\frac{\text{CGSO}}{LC_{ox}}\right)} \tag{1.43a}$$

and in terms of the drain current,

$$f_T = \frac{1}{2\pi}A\sqrt{\frac{I_D}{W}} \tag{1.43b}$$

where A is a technology-dependent parameter and can be calculated from

$$A = \sqrt{\frac{2\mu}{k_{ol}^2 C_{ox}L^3}} \tag{1.43c}$$

This expression indicates that, for a non-velocity saturated transistor

- f_T increases with mobility. Therefore, NMOS transistors exhibit better high-frequency performance.
- f_T increases with the gate bias voltage (gate overdrive) and consequently with the DC current of the transistor.
- f_T strongly depends on the gate length. Therefore short channel devices are better for high-frequency applications.

For a velocity saturated transistor, using the transconductance expression given in (1.34) and the gate–source capacitance found in Problem 1.1, f_T can be expressed as

$$f_T = \frac{1}{2\pi}\frac{kv_{sat}}{L\left(1+\frac{\text{CGSO}}{LC_{ox}}\right)} \tag{1.44a}$$

and,

$$f_T = \frac{1}{2\pi}\frac{k}{k_{ol}}\frac{1}{L}v_{sat} \tag{1.44b}$$

From this expression we conclude that

- Since the saturation velocity of holes is only slightly smaller than that of electrons, the high-frequency performance of velocity saturated PMOS transistors is comparable to that of the NMOS transistors.
- The gate length dependence of f_T for a velocity saturated transistor is not as strong as a non-velocity saturated transistor.
- For a velocity saturated transistor f_T is independent of the DC operating conditions.

1.1.3.3 The parasitic resistances

The effects of the series source (or drain) resistances of small-geometry transistors can be examined in two parts: the intrinsic resistance and the extrinsic resistance. The

intrinsic resistance is the sum of the resistance of the accumulation layer in the source region induced by the gate voltage and the spreading resistance from the accumulation layer to the bulk of the source [9], [10], [11], and this component varies with the gate voltage. The extrinsic source (or drain) resistance is composed of the source region resistance and the resistances of the related silicon to metal contacts.

All these components are modeled in high-level simulation models and the associated model parameters are provided by the manufacturers. In these models the source and drain series resistances are usually evaluated and modeled together as a total series resistance [12], assuming that the structure is symmetrical and the source (or drain) series resistance is equal to one half of the total series resistance. It must be stressed that for analog design these two parasitics have to be considered as separate resistors. Although the magnitudes of these resistors are equal, their effects on the behavior of the device in a circuit are not the same. The obvious effect of the series drain resistance of a MOS transistor is the increase in the saturation voltage, owing to the voltage drop on this resistance. The different and important effects of the series source resistance are the decrease of the drain current corresponding to a certain gate–source voltage and the decrease of the transconductance, as investigated in Chapter 2.

The numerical value of the source (or drain) resistance of a MOS transistor can be obtained using the model parameters given in BSIM3-v3 parameter lists. In BSIM3-v3 the parameter related to the intrinsic part of the series resistance (RDSW) is given as the total series resistance per micron width of the transistor. For hand calculation purposes the gate voltage dependence of this resistor can be ignored and the approximate value of the series intrinsic source (or drain) resistance can be found as

$$R_{\mathrm{Si}} = \frac{1}{2}\frac{(\mathrm{RDSW})}{W[\mu\mathrm{m}]} \tag{1.45}$$

The total extrinsic resistance (except the contact resistances) is modeled with (RSH), the sheet resistances of the source and drain regions. If the contacts are placed at a distance of $X/2$ from the source (Fig. 1.22), the resistance from the edge of the source to the line of contacts can be calculated as

$$R_{\mathrm{Se}} = \frac{1}{2}(\mathrm{RSH}) \times (\mathrm{nrs}) \tag{1.46}$$

where (nrs) is defined by the geometry as nrs $= X/\mathrm{NW}$ for an N-finger transistor.

The other component of the extrinsic source resistance is the equivalent contact resistance that is inversely proportional to the number of parallel connected contacts. The value of a typical contact resistance is given by the manufacturer.

Example 1.3 Let us calculate the source series resistance of an AMS 0.35 PMOS transistor as shown in Fig. 1.22. The dimensions are $L = 0.35\,\mu\mathrm{m}$, $W = 5\,\mu\mathrm{m}$ and $X = 0.85\,\mu\mathrm{m}$.

The available parameter values from the data sheets of AMS are as follows:

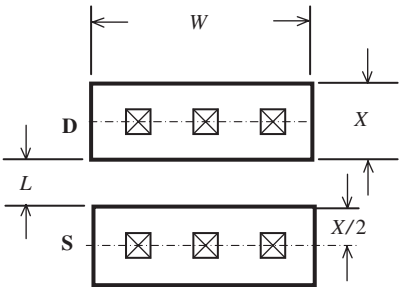


Figure 1.22 Plan view of the diffusion regions of a MOS transistor.

RDSW 1.033e+03
RSH 1.290e+02
Contact res. 60 ohm per 0.4 μm×0.4 μm contact from p-diff. to metal−1.
The intrinsic component of the source resistance from (1.45):

$$R_{\mathrm{Si}} = \frac{1}{2}\frac{1033}{5[\mu\mathrm{m}]} = 103.3\,\mathrm{ohm}$$

Source region body resistance from (1.46):

$$R_{\mathrm{Se}} = \frac{1}{2}129\times\left(\frac{0.85}{5}\right) = 10.96\,\mathrm{ohm}$$

The equivalent contact resistance:

$$R_{\mathrm{Sc}} = \frac{60}{3} = 20\,\mathrm{ohm}$$

Hence the total series source resistance is

$$R_{\mathrm{S}} = 103.3 + 10.96 + 20 = 134.26\,\mathrm{ohm}$$

Problem 1.2 *The parasitic resistance related parameters for AMS 0.35 NMOS transistors are given with their typical and maximum values as follows:*

RDSW 345
RSH 75 *typical,* 85 *maximum*
Contact res. 30 ohm *per* 0.4 μm × 0.4 μm *contact, typical,* 100 ohm *maximum.*
Dimension of the transistor is 20 μm/0.35μm. *Number of contacts on source and drain regions is* 10.
Calculate the "worst case" source and drain parasitic series resistances.

Since the DC gate current is negligible, the gate resistance of a MOS transistor has no effect on the DC (and low-frequency) performance. But at high frequencies, owing to the gate capacitance, the gate resistance affects the high-frequency performance of the device. It must be noted that the gate capacitance and the resistance of the gate conductor form a distributed R-C line along the width of the channel, and the gate signal applied from one end of the gate electrode propagates along the channel width, until it reaches the other end of the gate electrode. The propagation delay of this R-C line is given in [13]. For analog applications it is more interesting to investigate the signal attenuation along the channel width, as a function of the frequency. The [*y*] parameters of an R-C line that is a passive two-port are given in [14] as hyperbolic functions of $(l\sqrt{s.rc})$, where s is the complex frequency, l the length, and r and c are the resistance and capacitance of a short Δl sector. From these expressions, the voltage transfer coefficient of the line from its input to its unloaded output port can be found as

$$A = \frac{v_o}{v_i} = \frac{1}{\cosh(l\sqrt{s.rc})} \tag{1.47}$$

The magnitude of A is obviously equal to unity for low frequencies along the channel and decreases with frequency. This means that the control efficiency of the channel decreases from the near end of the gate (connected to the gate contact), toward the far end. Now we can calculate the frequency[11] at which the magnitude of A decreases to $1/\sqrt{2}$.

Now

$$\cosh(l\sqrt{s.rc}) = \sqrt{2}$$

corresponds to

$$(l\sqrt{s.rc}) = 0.882 \text{ or } (ls.l^2rc) = 0.78 \tag{1.48}$$

The total resistance and the total capacitance of the gate electrode are $R = r \times l$ and $C = c \times l$, respectively. R and C can be expressed, in terms of the sheet resistance of the gate electrode and C_{ox}, as

$$R = R_{sh}\frac{W}{L} \text{ and } C = C_{ox}WL \tag{1.49}$$

From the last two expressions, the complex frequency corresponding to $A = (1/\sqrt{2})$ can be calculated as

$$s_c = \frac{0.78}{R_{sh}C_{ox}W^2}$$

and the corresponding frequency as

$$f_c = \frac{0.78}{2\pi \times R_{sh}C_{ox}W^2} = \frac{0.124}{R_{sh}C_{ox}W^2} \tag{1.50}$$

[11] This frequency will be called the maximum usable frequency. Note that it is necessary to evaluate this frequency together with the high-frequency figure of merit, f_T, of the transistor.

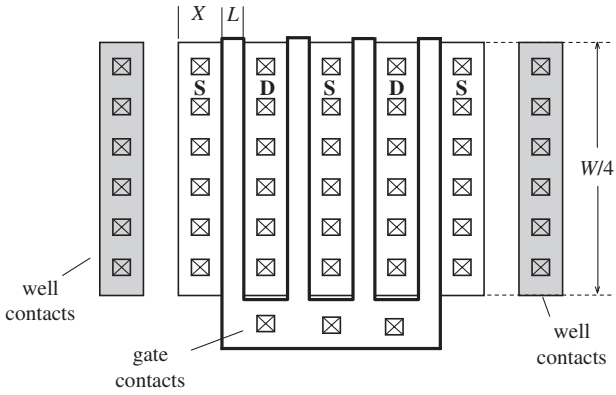


Figure 1.23 Plan view of a multi-finger MOS transistor.

This expression indicates that the gate width has a strong adverse effect on the usable maximum frequency. To overcome this problem the well-known solution is to use a finger-structure, as shown in Fig. 1.23.

Another advantage of this structure is the decrease of the source and drain junction areas and, consequently, the parasitic junction capacitances, as well as the source and drain series resistances. It can be easily seen that the total drain and source areas of an N finger transistor are equal to

$$A_D = \frac{N}{2} \times \frac{W}{N} \times X = \frac{WX}{2},\ A_S = \left(\frac{N}{2} + 1\right) \times \frac{W}{N} \times X = \frac{N+1}{N}\frac{WX}{2}$$

and that both of them approach one half of the junction areas of a 1-finger transistor for high N values.

Example 1.4 Calculate the maximum usable frequency for a 40 μm/0.35 μm AMS 0.35 micron NMOS transistor. The sheet resistance of the gate-poly is given as 8 ohm/ □ and the gate capacitance is 8.625 fF/cm^2.

From (1.50)

$$f_c = \frac{0.124}{8 \times (8.625 \times 10^{-7}) \times (40 \times 10^{-4})^2} = 1.123 \text{ GHz}$$

If the transistor has to operate at a frequency of up to 5 GHz, a multi-finger structure must be used. It can be easily calculated that the channel width corresponding to 5 GHz must not be bigger than 27.6 micron. To maintain the DC properties (i.e. the total channel width as 40 μm) the appropriate solution is to use a 4×10 μm structure with a certain safety margin, from the point of maximum usable frequency.

1.2 Passive on-chip components

To enable the integration of a complete analog circuit or system on a single chip, the passive components, namely resistors, capacitors and inductors compatible with the standard MOS transistor fabrication technologies must be available. To improve the quality of passive components, it may become necessary to add extra steps to the standard fabrication process, which certainly increases the complexity of the process and, consequently, the price. Figure 1.24 shows the cross-section of a typical CMOS chip with some of the passive components; a poly2 resistor, a poly1–poly2 capacitor and a metal–insulator–metal (MIM) capacitor. The on-chip inductors are usually formed using the uppermost thick metal layers.

1.2.1 On-chip resistors

On-chip resistors can be realized by using thin conductive films or doped silicon layers. One possibility for the thin films is to use the metal (usually aluminum) films that are normally used for interconnections. The sheet resistances of metal thin films are of the order of 50 to 100 mΩ/□. It is obvious that the metal films are only suitable for very low-value resistors, and are rarely used in ICs. The more important and frequently used option is the highly doped poly-silicon films that are normally used as the gate material of MOS transistors. For small geometries, they are usually silicided to decrease the resistance. The sheet resistance of the gate poly-silicon films

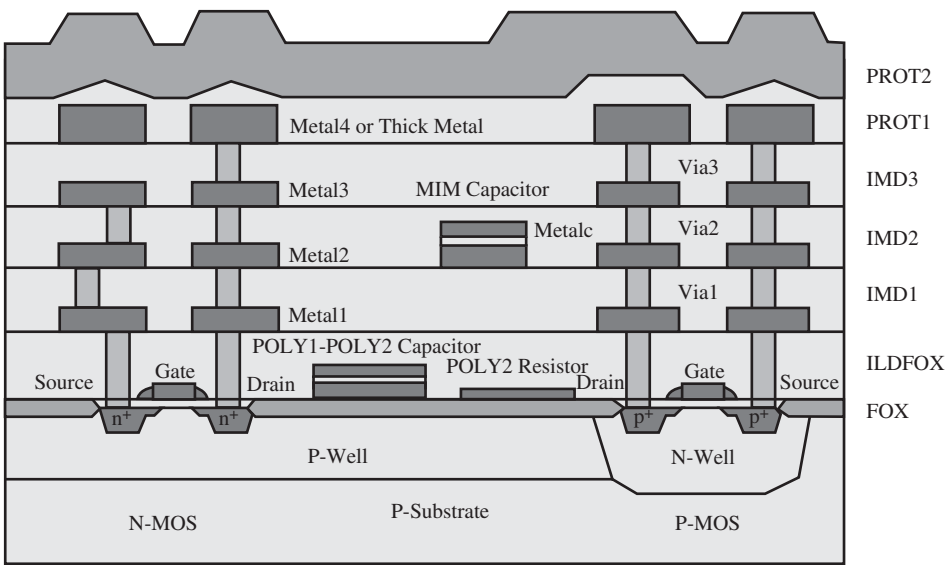


Figure 1.24 The partial cross-section of a 0.35 micron, four-metal layer CMOS chip. Some of the passive components are shown together with the NMOS and the PMOS transistors (courtesy of austriamicrosystems A.G.– AMS).

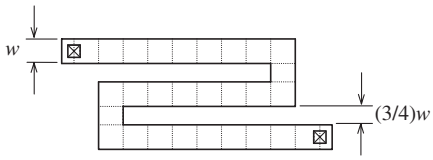


Figure 1.25 Plan view of a serpentine-shaped poly-silicon resistor. Note that $N_{sq} = 24$, $N_{cor} = 4$.

is usually in the range of 5 to 10 Ω/□. The plan view of a thin film resistor is shown in Fig. 1.25. The value of such a resistor can be calculated as

$$R \cong R_{sh}(N_{sq} + 0.6 \times N_{cor}) + 2R_c,$$

where R_{sh} is the sheet resistance of the conductive film in ohms per square, N_{sq} is the count of normal squares along the resistor, N_{cor} is the count of corner squares and R_c is the resistance of the end contacts.

Using the gate poly-silicon, resistor values of up to a few kilo-ohms can be realized. For higher value of resistors, the gate poly-silicon is not convenient owing to the adverse effects of the unavoidable distributed parasitic capacitance of the resistor and excessive area consumption. For analog applications, IC manufacturers usually provide an additional high resistivity poly-silicon film. The sheet resistance of this poly-silicon layer is usually in the range of 1 to 2 kΩ/□.

There are several important practical problems associated with the film resistors. One of them is the manufacturing tolerance of the sheet resistance of the film that is in the range of $\pm 20\%$. This corresponds to $\pm 20\%$ absolute tolerance for the resistors. But the relative tolerances of the resistors on a chip are considerably smaller. Another problem is the temperature coefficient of the resistors that is of the order of $+10^{-3}$/K for gate-poly and -0.5×10^{-3}/K for high resistance poly.

Another possibility for realizing resistors on a CMOS chip is to use several doped silicon layers, as was usually done in bipolar ICs. The n^+ and p^+ doped regions that are used to form the source and drain of the NMOS and PMOS transistors have sheet resistance values in the range of 50 to 150 Ω/□. The sheet resistance of a well is higher and in the range of 1 to 2 kΩ/□. The absolute tolerances of the doped silicon regions are usually of the order of $\pm 10\%$ and have temperature coefficients higher than $+10^{-3}$/K. Another drawback related to the doped silicon layers is the bias dependence of the sheet resistance values.[12]

The contact resistance certainly depends on the contact size and the technology. Contact sizes are usually of the order of the minimum geometry. The series resistance related to the contacts can be decreased by increasing the number (not size) of parallel contacts. This is also useful from the reliability point of view. As an

[12] Note that the diffusion region used as a resistor must not have a silicide layer on top, which would reduce the overall resistance to unusable levels. Many layout design tools provide the "silicide block" feature, to prevent the silicide layer occurring on top of the resistor areas.

example, for a 0.35 micron technology, the typical (and maximum) resistance values of $0.4\,\mu m \times 0.4\,\mu m$ contacts are given as follows:

metal-poly	2 (10) Ω/contact
metal-S/D (n-type)	30 (100) Ω/contact
metal-S/D (p-type)	60 (150) Ω/contact.

1.2.2 On-chip capacitors

There are two classes of on-chip capacitors: passive, fixed value capacitors and variable capacitors (varactors). The passive capacitors are – in principle – basic parallel plate capacitors. Varactors are solid state structures such as reverse-biased p-n junctions and MOS capacitors. A varactor obviously can be used as a fixed value capacitor if it is biased with a fixed bias voltage.

1.2.2.1 Passive on-chip capacitors

Two options to form parallel plate capacitors can be seen in Fig. 1.24. These options are the poly1–poly2 capacitor and the metal–insulator–metal (MIM) capacitor. The poly1–poly2 capacitors have to be realized at the bottom level of the structure. The MIM capacitors can be realized at any metal level in principle, but the practically possible level (or levels) for a certain technology are usually declared by the manufacturer.

The capacitance of a parallel plate capacitor is given as

$$C = \frac{A}{t}\varepsilon_0\varepsilon_{ox}$$

where ε_0 and ε_{ox} are the permittivity of the vacuum and dielectric constant of the insulating layer, respectively, A is the area of the parallel plates and t is the distance between the conducting plates that is filled with the insulating material, usually silicon dioxide. Inserting the numerical values of ε_0 and ε_{ox}, and expressing the area in square microns and t in nanometers, the capacitance can be found to be

$$C = \frac{A34.5}{t[\text{nm}]}[\text{fF}] \tag{1.51}$$

Owing to the fringe-field effect at the periphery of the capacitor, the accuracy of this expression decreases for smaller dimensions.

In poly and MIM capacitors the electric field between the electrodes is perpendicular to the surface. To increase the surface usage efficiency, structures using the lateral electric field are developed. The simplest example for a lateral field (or interdigitated) capacitor is shown in Fig. 1.26. The capacitance of this structure is proportional to the thickness of the metal layer (t_M) and the length of the serpentine-shaped insulating region (dielectric) between the two metal electrodes, and inversely proportional to the width of the insulator (t_{ins}). It is obvious that the most suitable metal layer is the thick metal layer. Several structures using both the lateral and perpendicular fields can be developed to increase the surface usage efficiency further [15], [16].

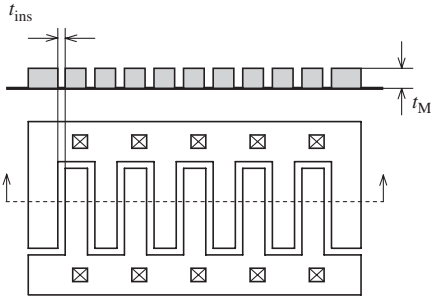


Figure 1.26 The plan view and cross-section of an interdigitated capacitor.

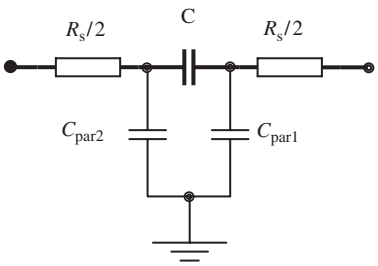


Figure 1.27 The dominant parasitics of an on-chip capacitor.

The parasitics of passive on-chip capacitors

The dominant parasitics of an on-chip capacitor are shown in Fig. 1.27, where C is the intended capacitance, C_{par1} and C_{par2} are the parasitic capacitances between the plates of the capacitor and ground, and R_s is the equivalent series resistance representing the total losses. For vertical field capacitors, for example the poly1–poly2 capacitor, the parasitic capacitance of the bottom plate is considerably bigger than the parasitic of the upper plate, and in some cases not negligibly small compared to C. In horizontal field capacitors, for example the capacitor shown in Fig. 1.26, the parasitic capacitances associated with the two electrodes are approximately equal, which is an advantage in certain cases.

The equivalent parasitic series resistance contains the resistances of the contacts and the plates. The latter is a distributed resistor in two dimensions; frequency dependent and complicated to calculate. The overall effect of this resistance is usually represented by the quality factor (Q) of the capacitor.[13] The quality factors of passive on-chip capacitors are usually in the range of 30 to 100 at 1 GHz. It must be noted that at

[13] The quality factor of a capacitor at a certain frequency ω can be defined as $Q = 1/R_sC\omega$, if the capacitor is in resonance at ω with an ideal (lossless) inductance (see Chapter 4).

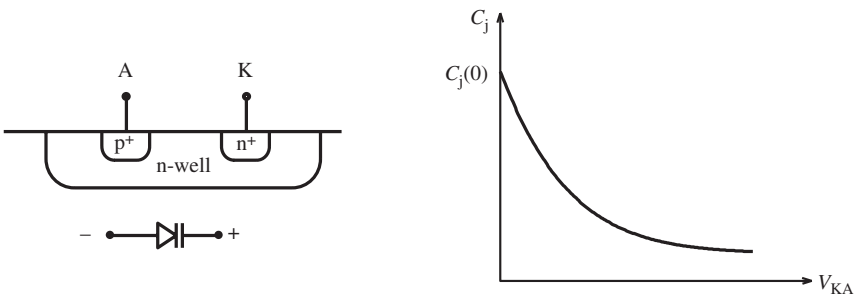


Figure 1.28 The cross-section and the typical control voltage versus capacitance curve of a p-n junction varactor.

higher frequencies, certainly depending on the geometry of the structure, the self-inductances of the plates and associated connections start to become effective and the capacitor can resonate with its parasitic inductance.

1.2.2.2 Varactors

The varactor is a voltage-tunable capacitor which is mostly used in high-frequency applications, to tune the frequency characteristics of circuits. One of the important classes of varactors is a reverse-biased p-n junction.[14] It is known from basic semiconductor device physics that the depth (and consequently the electric charge) of the depleted regions of a p-n junction varies with the applied voltage. This means that the amount of charge in the depletion regions can be controlled with the voltage, which, by definition, corresponds to an electrostatic capacitance. The value of this capacitance can be expressed as already mentioned in Section 1.1.3.

$$C_j(V) = C_j(0)\left(1 - \frac{V}{\varphi_B}\right)^{-m} \tag{1.52}$$

where $C_j(0)$ is the value of the capacitance for zero bias, φ_B is the built-in junction potential and m is the grading coefficient of the junction, whose value varies between 1/3 and 1/2, depending on the doping profiles. $C_j(0)$ depends on the junction area and the doping concentrations of the p-n junction. The junction barrier height, φ_B, is in the range 0.65 to 0.75 V and is given as (PB) in parameter lists, as already mentioned. There are several options to form a p-n varactor in a CMOS IC. The cross-section of one of them and the typical voltage–capacitance curve are shown in Fig. 1.28. The tuning range[15] depends on the range of the control voltage; its minimum value is zero and the maximum value is limited by the breakdown voltage of the junction and the available DC voltage value on the chip.

[14] The p-n junction type of variable capacitors are extensively used not only as an integral part of bipolar and MOS ICs, but also as discrete components as the tuning elements of radio and TV receivers and telecommunication systems.

[15] The tuning range is defined as $\gamma = (C_{max} - C_{min})/(C_{max} + C_{min})$.

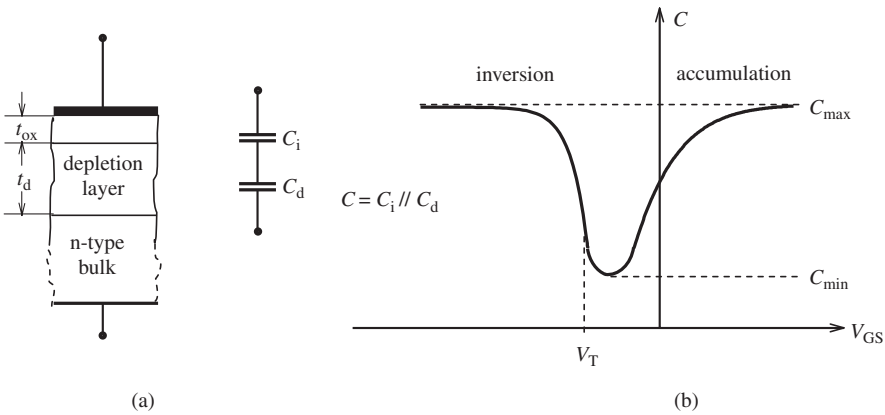


Figure 1.29 (a) The oxide and depletion capacitance components of a MOS capacitor. (b) Typical *C–V* curve of a PMOS transistor.

It is known from MOS theory that the gate-bulk capacitance of a MOS structure exhibits a typical variation. In Fig. 1.29 the variation of the capacitance of a PMOS transistor as a function of the DC voltage applied between the gate electrode and the bulk is shown. In one extreme case, where the gate voltage is sufficiently more negative than the threshold voltage V_T (which is negative for a PMOS transistor), there is a strong inversion layer that forms the lower "plate" of the MOS capacitor. The capacitance is equal to $C_i = WLC_{ox}$. In the other extreme, if the gate voltage is sufficiently more positive than the threshold voltage, the attracted majority carriers (electrons in this case) of the n-type bulk form a conductive n-type accumulation layer. The capacitance corresponding to this case is again equal to C_i. In between these two extreme cases, for a narrow interval of the gate voltage, the semiconductor surface just below the gate oxide is – practically – depleted from electrons and holes. The thickness of this depletion layer and consequently the fixed ion charge is subject to the control of the gate voltage, which corresponds to a capacitance $C_d = dQ_d/dV_G$. For this interval the total gate-bulk capacitance is the series equivalent of C_i and C_d : $C = C_iC_d/(C_i + C_d)$. To change the capacitance characteristic given in Fig. 1.29 to a more useful form that permits one to control the capacitance monotonically in one direction, some modifications are possible.

One of the solutions compatible with the standard CMOS technologies is shown in Fig. 1.30(a). In this structure the n-well (the bulk of the transistor) is connected to the highest positive potential in the circuit. Consequently, it is not possible to drive the device into the accumulation region. The source and drain regions are connected to each other. If the gate voltage is more negative than the threshold voltage, the device is in inversion mode and the capacitance is equal to C_i. This is the maximum value of the capacitance. For gate voltages sufficiently higher than the threshold voltage, inversion is not possible and the device is in depletion mode. The corresponding capacitance is at its minimum value and is equal to $C = C_iC_d/(C_i + C_d)$.

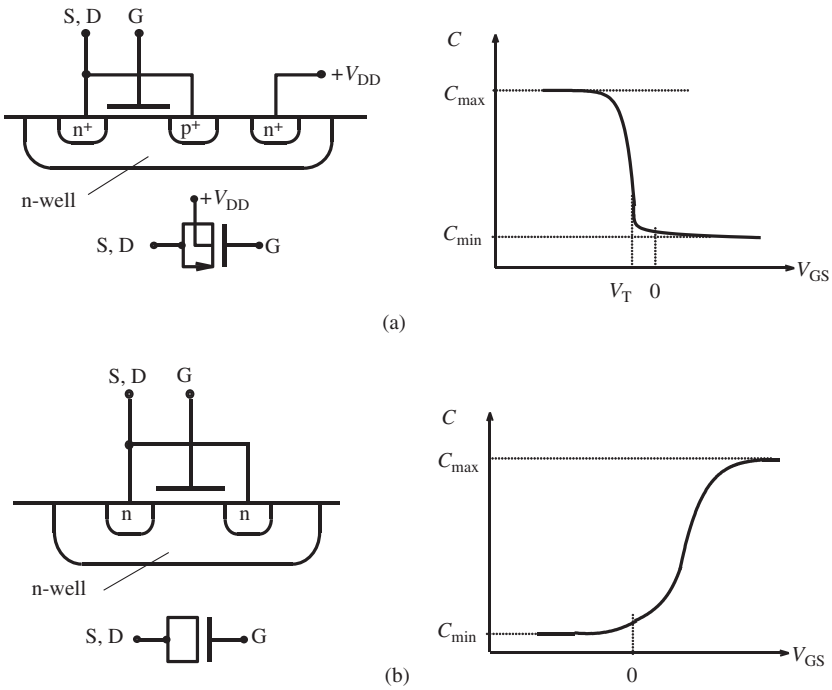


Figure 1.30 Cross-section and the *C–V* curve of (a) an inversion type MOS varactor and (b) an accumulation type MOS varactor.

The second solution is shown in Fig. 1.30(b). The bulk (well) again is n-type. For gate voltages higher than the threshold voltage an electron-rich accumulation layer is formed and connects the two high concentration n-type regions. Since this high conductivity layer is just below the oxide layer, the capacitance is equal to the oxide capacitance, $C_i = WLC_{ox}$. When the gate voltage decreases toward negative values, the accumulated electrons and then the majority electrons of the n-type bulk are repelled to form a depletion layer. For sufficiently low gate voltages the capacitance is equal to $C = C_iC_d/(C_i + C_d)$, and varies monotonically between these two extreme values.

For a certain technology (for a certain gate oxide thickness and doping concentration of the n-well) the minimum and maximum values of the capacitance are the same for the inversion-type and accumulation-type MOS varactor. Since the surface consumption is smaller, usually the accumulation type is preferred. The tuning ranges and the quality factors of the MOS varactors are usually in the range of 40 to 60% and 50 to 80 at 1 GHz, respectively.

An important problem associated with the varactors is the nonlinearity of the *C–V* curve. For small signals superimposed on to the bias voltage, the capacitance of the device has the value corresponding to this bias. It has been shown that for large signal amplitudes – in addition to a nonlinear distortion – the effective capacitance of the device decreases [17].

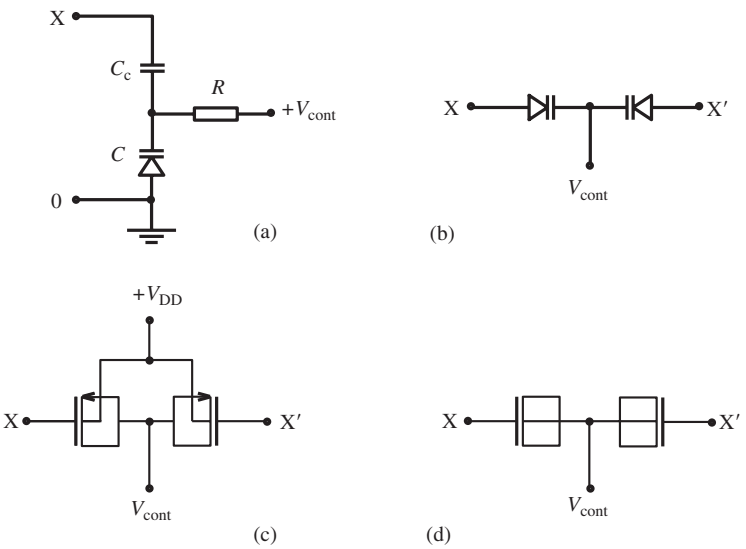


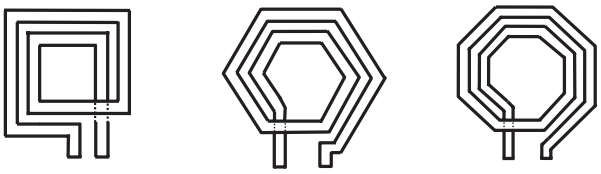
Figure 1.31 Biasing of varactors: (a) a one-end grounded varactor, (b) a p-n junction type symmetrical varactor pair, (c) an inversion-type MOS symmetrical varactor pair, (d) an accumulation-type MOS symmetrical varactor pair.

In Fig. 1.31 several examples are given related to the biasing of varactors. Figure 1.31(a) shows the biasing of a p-n junction varactor that forms a variable capacitor one of whose ends is grounded. The coupling capacitor C_c must be as high as possible to profit from the full tuning range of the varactor. The bias resistor R must be high enough not to affect the Q of the varactor.

In many applications varactors are used as floating tuning elements. In the examples shown in Fig. 1.31(b), (c), and (d) [18], the nodes X and X′ have the same DC voltage. For proper operation of the varactors in Fig. 1.31(b), the control voltage V_{cont} must remain positive with respect to V_X. The control voltage in Fig. 1.31(c) must be positive with respect to V_X. In Fig. 1.31(d), V_{cont} can be changed from negative to positive voltage with respect to V_X.

1.2.3 On-chip inductors

Throughout the history of electronic and telecommunication systems, inductors have always been among the most bulky and expensive passive components of the circuits. In the case of integrated circuits the situation was even worse; it was completely impossible to integrate larger-valued inductors as on-chip components. The goal of most of the active filters developed during the last five decades was to eliminate the inductors and to replace them with capacitors and active circuit combinations. Thanks to operational amplifiers and OTAs that are the outcome of the developments of microelectronics technology, active filters have effectively eliminated the inductors



Form	K_1	K_2
Square	2.34	2.75
Hexagon	2.33	3.82
Octagon	2.25	3.55

Figure 1.32 Typical forms of on-chip spiral inductors: (a) square, (b) hexagonal, and (c) octagonal. Inset: K_1 and K_2 parameters for these forms.

from many application areas. Eventually the maximum operation frequencies of systems increased up to the "gigahertz" region, where it was not possible to apply the conventional active filter techniques. But in parallel to the increase of the frequency, the necessary inductance values decreased down to the range of "nano-henries", which is small enough to be integrated on chip.

On-chip inductors are usually shaped by etching the uppermost thick metal layer, which is usually aluminum in standard technologies and has a thickness of around one micron. Copper is an alternative with a higher conductivity, and is provided by some foundries.

The on-chip inductors are mostly realized in the forms shown in Fig. 1.32. The value of such an inductor depends on the dimensions and number of turns of the spiral. Since the length of the conductor corresponding to a certain loop area is minimum for a circle, the quality factor of a polygon inductor is higher than that of a square inductor. There are several simple formulas in the literature to calculate the inductance of a planar spiral inductor. One of them, named the "modified Wheeler formula" that provides reasonable accuracy for simple first step calculations, is given in [19]:

$$L = K_1 \mu_0 \frac{n^2 d_{\text{avg}}}{1 + K_2 \rho} \tag{1.53}$$

where K_1 and K_2 are form-dependent coefficients that are given in Fig. 1.32, μ_0 is the magnetic permeability of the material, which is equal to $\mu_0 = 1.257 \times 10^{-8}$ [henry/cm] for silicon, and is equal to the permeability of vacuum, and n is the number of turns of the spiral. The average diameter, d_{avg}, and the fill factor, ρ, are defined as follows:

$$d_{\text{avg}} = \frac{d_{\text{out}} + d_{\text{in}}}{2} \qquad \rho = \frac{(d_{\text{out}} - d_{\text{in}})}{(d_{\text{out}} + d_{\text{in}})} \tag{1.54}$$

where d_{out} and d_{in} are the outer and inner diameter of the inductor, respectively.

The practically realizable range of the on-chip inductors is $L = 0.5$–10 nH, and the corresponding quality factors range between $Q = 3$–15.[16]

[16] The quality factor of an inductor at a certain frequency ω can be defined as $Q = L\omega/R_s$, if all losses are represented by a series resistor R_s and the inductor is in resonance at ω with an ideal (lossless) capacitor (see Chapter 4).

Example 1.5 Calculate the self-inductance of the spiral inductor shown in Fig. 1.33, and the quality factor at 10 GHz assuming that the only loss is related to the DC resistance of the aluminum strip.

The outer diameter of the coil is 100 μm. The inner diameter can be calculated as $d_{in} = 100 - (4\times10 + 2\times5) = 50$ μm. Then the average diameter and the fill factor are: $d_{avg} = 0.5(100 + 50) = 75$ μm, $\rho = (100 - 50)/(100 + 50) = 1/3$.

Using the values given in Fig. 1.32:

$$L = 2.34\times(1.275\times10^{-8})\frac{(2^2)\times(75\times10^{-4})}{1+(2.75\times0.333)} = 460\times10^{-12} = 0.46[\text{nH}].$$

The average length of the aluminum strip is $l = 620$ μm (620×10^{-4} cm) and the cross-section $A = (10\times10^{-4})\times(1\times10^{-4}) = 10^{-7}\,\text{cm}^2$. The specific resistance of aluminum is 2.73×10^{-6} ohm.cm. Now the resistance of the coil can be calculated as

$$R_s = (2.73\times10^{-6})\frac{620\times10^{-4}}{10^{-7}} = 1.69 \text{ ohm}.$$

With the assumption that the series resistance represents the only loss of the inductor, the quality factor at $f = 10$ GHz is

$$Q = \frac{L\omega}{R_s} = \frac{(0.46\times10^{-9})(2\pi\times10^{10})}{1.69} = 17.1.$$

There are several computer programs developed to calculate the inductance and quality factor of on-chip inductances with high precision [20], [21]. One of them, SPIRAL, calculates the value of the inductance shown in Fig. 1.33 and the quality factor of the inductor at 10 GHz as 0.42 nH and 5.2, respectively. The difference between the two values of the inductance can be considered acceptable. But there is a

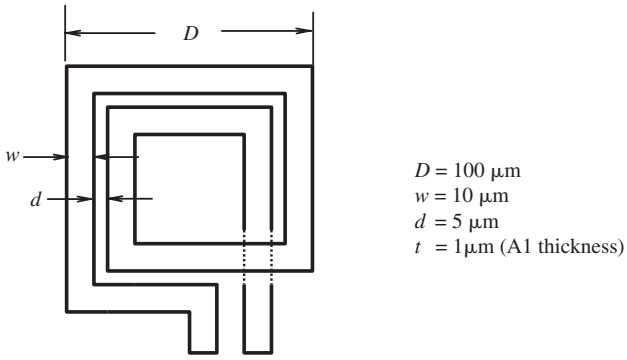


Figure 1.33 Plan view of a typical square inductor.

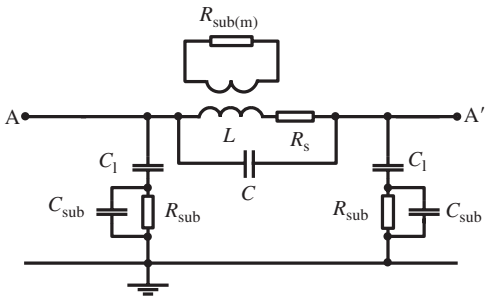
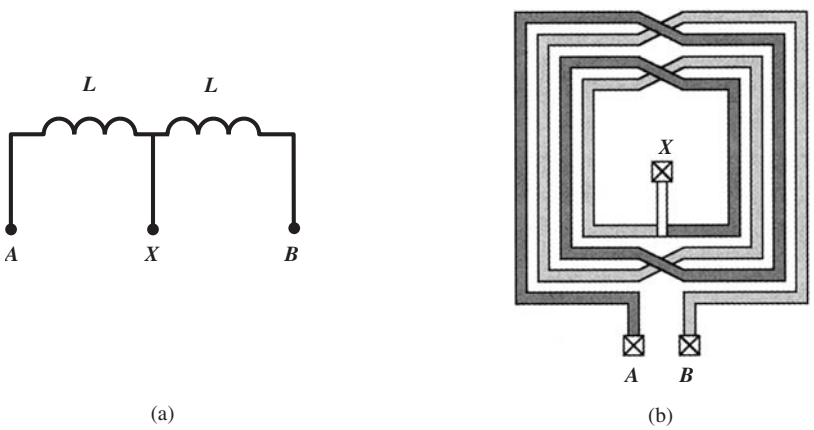


Figure 1.34 The equivalent circuit of an on-chip inductor connected between the nodes A and A′.



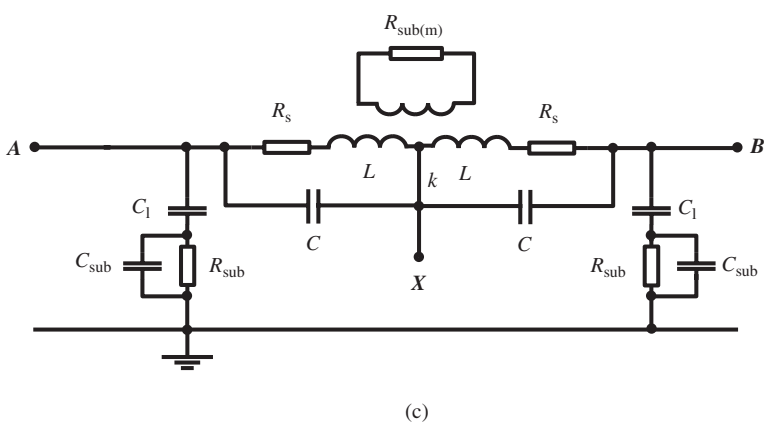


Figure 1.35 (a) A center-tapped (symmetrical) inductor. Note that $L_{AB} = 2L(1+k)$, where k is smaller than unity and depends on the geometry.[17] (b) Layout of the center-tapped inductor. (c) The equivalent circuit of the center-tapped, symmetrical inductor (k is the magnetic coupling coefficient of the two halves of the inductor).

[17] It is known that the equivalent inductance of two serially connected inductors (L_1 and L_2) can be expressed as $L_{eq} = L_1 + L_2 \mp 2k\sqrt{L_1 L_2}$, where k is the magnetic coupling coefficient and is always smaller than unity, and the sign is positive if the fields of the individual inductors support each other.

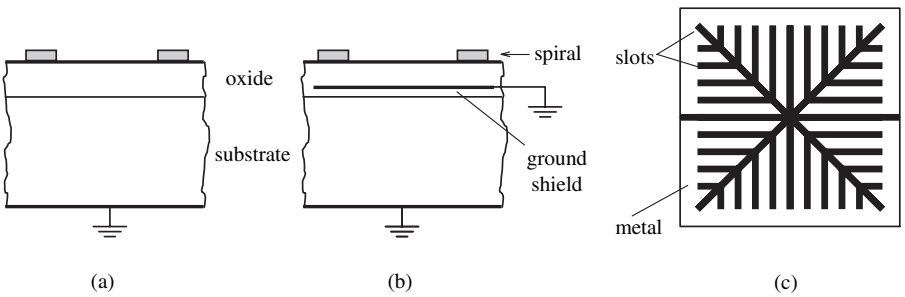


Figure 1.36 (a) Cross-section of a spiral on-chip inductor. (b) Spiral inductor with ground shield. (c) Plan view of a patterned ground shield.

very large difference between the calculated and simulated Q values. This means that there are other loss mechanisms in addition to the DC resistance of the aluminum strip. One of them is the skin effect that increases the resistance at high frequencies. Other mechanisms can be explained by the equivalent circuit shown in Fig. 1.34.

In Fig. 1.34 R_s represents the resistance of the metal strip that increases with frequency owing to the skin effect, as previously mentioned. C is the parasitic capacitance between the terminals of the inductor. The metal strip forming the inductor is separated from the silicon substrate by a thick silicon dioxide layer. Consequently, there is a capacitance between the metal strip and the substrate. This distributed capacitance is represented by two equal parasitic capacitances (C_l) lumped at the terminals of the inductor. R_{sub} is the series resistance of C_l and depends on the doping concentrations of the regions below the structure. Capacitors parallel to R_{sub} are included into the equivalent circuit to achieve a good fit to the measured data [22]. $R_{sub(m)}$ represents the loss owing to the magnetically induced current in the substrate. Another effect of this "image current" that is magnetically coupled to the inductor is to reduce the value of the effective inductance.

In several applications, for example differential LNAs, differential oscillators, etc., center-tapped symmetrical inductors are needed, as shown in Fig. 1.35(a). To ensure the symmetry of the parasitics of the two half-sections, a careful layout design has prime importance. A square shaped symmetrical inductance layout is shown in Fig. 1.35(b). For hexagonal and octagonal inductors, symmetrical layouts can be generated similarly.

The losses owing to R_{sub} can be reduced (a) by decreasing the current flowing through C_l, and (b) by decreasing the value of the resistance itself. Decreasing the capacitor current requires a thick oxide layer between the thick metal layer and the silicon surface that depends on the number of metal layers and the thickness of the inter-layer oxides (Fig. 1.36(a)). To decrease the resistance on the current path, a grounded, high conductivity metal or poly-silicon shield (ground shield) can be placed between the inductor and the substrate (Fig. 1.36(b)). This modification effectively reduces the losses associated with the capacitive substrate current. But owing to the image current induced on this layer (that is equivalent to the current of a short-circuited secondary, magnetically

coupled to the inductor) the effective inductance decreases strongly. To overcome this adverse effect a simple but efficient solution is to pattern the ground shield in such a way to prevent the image currents on the shield (Fig. 1.36(c)) [23]. But it must be kept in mind that this shield does not prevent the induced currents in the silicon substrate and cannot reduce the losses represented by $R_{\mathrm{sub(m)}}$.

The parasitic capacitances shown in Fig. 1.34 determine the "self resonance frequency" of the inductor (see Chapter 4). Note that for frequencies higher than this frequency, the "inductor" exhibits a "capacitive" impedance.

To increase the quality factor of an on-chip inductor one solution is to etch away the silicon dioxide layer under the spiral and thus to reduce the parasitic capacitive current. But this "suspended inductor" approach is not compatible with standard processes. The bonding wires also can be used to realize low-value inductors, with considerably high Q values.[18] However, owing to high tolerances and bad repeatability, they cannot be considered as standard components of ICs.

[18] The self-inductance of a 25 micron diameter bonding wire is approximately 1nH/mm. The Q value of a bonding wire is about 60 at 2 GHz.

2 Basic MOS amplifiers: DC and low-frequency behavior

Basic MOS amplifiers are the main building blocks of a vast array of analog signal processing systems as well as other analog electronic circuits. The overall performance of a complex circuit strongly depends on the performances of its basic building blocks. In this chapter the main properties of these basic circuits will be investigated.

2.1 Common source (grounded source) amplifier

The basic structure of a common source amplifier is shown in Fig. 2.1(a). The gate of the NMOS transistor, M, is biased with a DC voltage source V_{GS}, to conduct the appropriate DC (quiescent) current. A signal source (v_i) is connected in series with the bias voltage to control the drain current. The load resistor R_D helps to convert the drain current variations into output voltage variations (output signal). Since the output of a MOS amplifier is usually connected to the gate of another MOS amplifier, the DC and low-frequency load coming from the subsequent stage is negligible[1] and the only load is R_D. In Fig. 2.1(b), this "DC load line" with a slope equal to $(-1/R_D)$ is drawn on the output curves of M. For a given value of the gate bias voltage (for example V_{GS1}), the drain current is $I_{D1,}$ which corresponds to the intersection of the load line and the output curve corresponding to V_{GS1}. This intersection point is called as the "operating point" or the "quiescent point" and denoted by Q.

- From its high value of V_{DD} down to the lower limit of the saturation region ($V_{DS(sat)}$), the output voltage is proportional to the drain current, which has a quadratic relation to the input voltage. This means that there is a severe nonlinear distortion on the output voltage.[2]
- To obtain the maximum output voltage swing without excessive distortion or clipping, the quiescent drain–source voltage must be situated in the middle of the "dynamic range" of the output signal, as shown in Fig. 2.1.
- It is obvious that for a given input signal amplitude, higher R_D values result in higher output signal amplitudes, or in other words, higher voltage gains.

[1] The effects of the input capacitance and the input conductance that appears owing to the Miller effect at high frequencies will be investigated later on.

[2] In a velocity saturated transistor the drain current is almost linearly related to the gate voltage and consequently the nonlinear distortion is smaller.

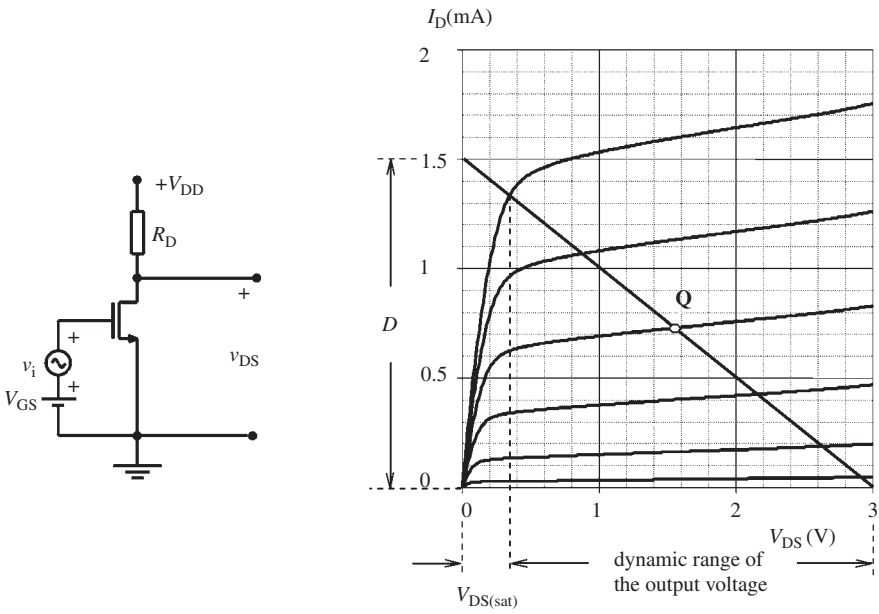


Figure 2.1 (a) The basic resistor loaded MOS amplifier. (b) The output characteristic curves, the DC load line and the operating point (Q). The drain current dynamic range is marked with D. The NMOS transistor (AMS 0.35) has a *W/L* ratio of (35 μm/0.35 μm). $R = 2$ k ohm and $V_{GS} = 0.9$ V.

- When the instantaneous value of the input voltage increases in the positive direction, the total gate–source voltage increases, resulting in an increase on the drain current and an increase on the voltage drop on R_D. This means that the output voltage decreases, or changes in the opposite direction to the input voltage, i.e. for a sinusoidal input signal the output signal is 180° out-of-phase with the input.

Resistors are not preferred components for CMOS integrated circuits, since they consume a large amount of area on the chip and are not directly compatible with standard CMOS processes. To eliminate the resistor, a MOS transistor can be used instead. There are several alternatives to this solution, but the most convenient approach is to use a PMOS transistor biased with an appropriate gate–source voltage (V_{GS2}), as the load of the NMOS input transistor (Fig. 2.2(a)). In this circuit the PMOS transistor acts as a nonlinear resistor controlled by its gate–source DC bias voltage (Fig. 2.2(b)). From another point of view, this load can be considered as a DC current source, provided that M2 is in the saturation region, i.e. $|V_{DS2}| > |(V_{GS2} - V_{T2})|$. The current of this source is equal to the quiescent drain current of M1, and its internal resistance is the output resistance of M2, for a given gate bias (Fig. 2.2(c)). Since M2 has no signal control mission and acts only as a passive load, this type of amplifiers will be called “passive MOS loaded amplifiers”.

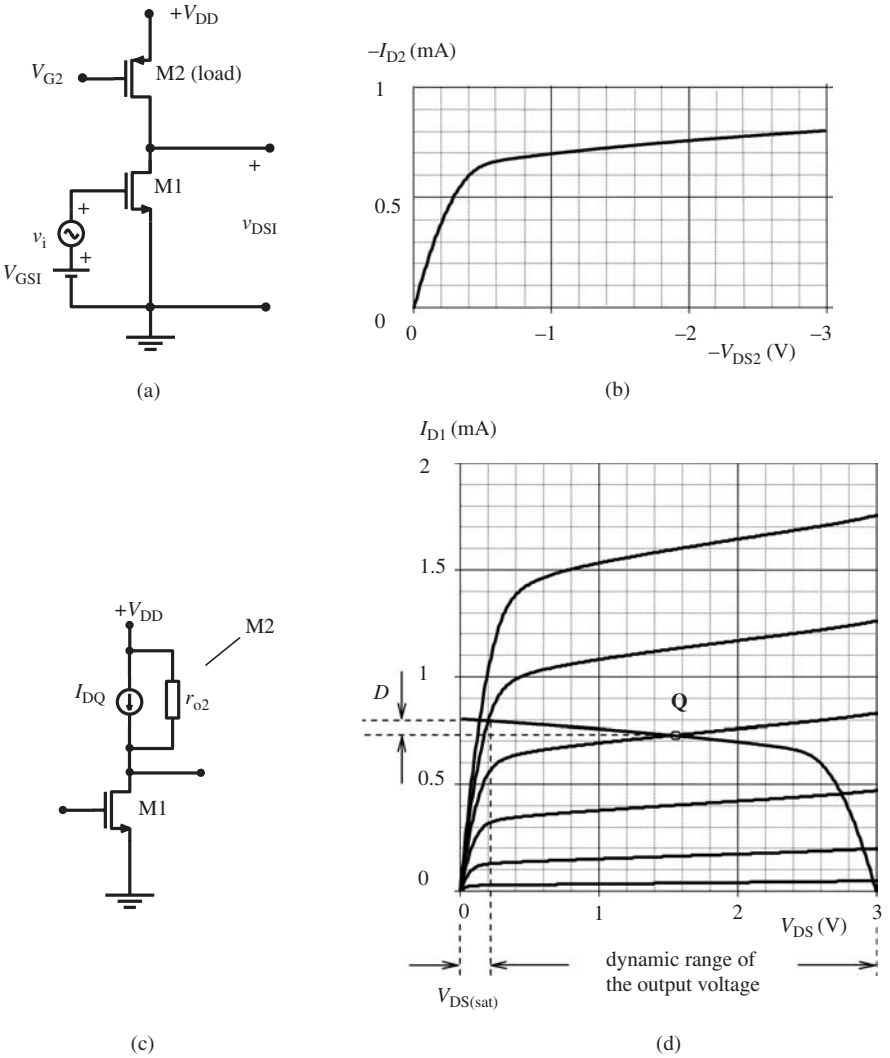


Figure 2.2 (a) Passive PMOS transistor-loaded common-source NMOS amplifier. (b) The nonlinear load line (the output characteristic curve of the load transistor for the specified V_{GS2} value). (c) The input transistor (M1) and the equivalent of the load transistor (M2). (d) The output characteristic curves of M1 together with its load. The drain current dynamic range is marked with D. M1 and M2 are AMS 0.35 transistors with aspect ratios of 35 µm/0.35 µm and 60 µm/0.6 µm, respectively. Gate biases are $V_{GS1} = 0.9$ V and $V_{GS2} = 1.3$ V ($V_{G2} = 1.7$ V).

Apart from the elimination of the resistor, this circuit has several advantages compared to the resistor-loaded circuit. In Fig. 2.2(d), the nonlinear load curve is drawn on to the output curves of M1 in such a way that $V_{DS1} + |V_{DS2}| \equiv V_{DD}$. Since the drain currents of M1 and M2 have equal magnitudes, the operating point corresponds

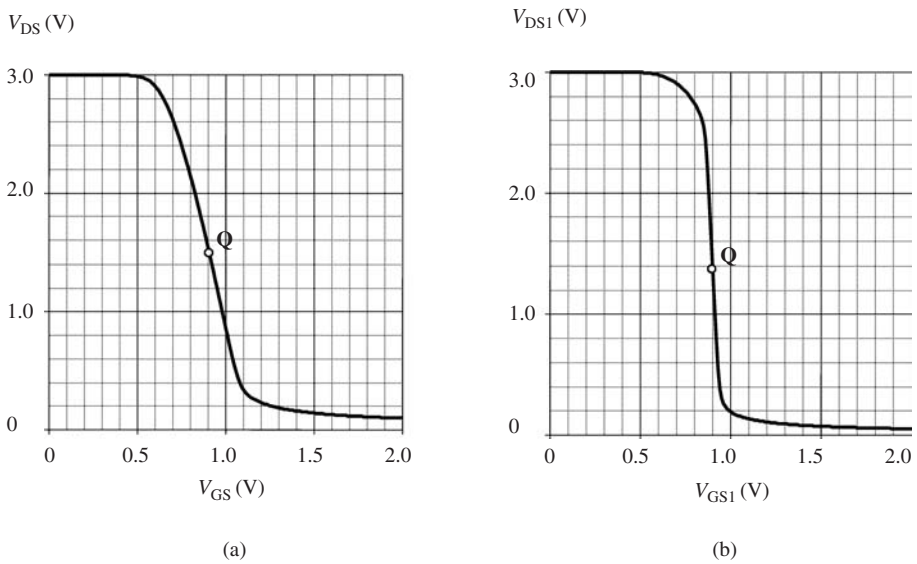


Figure 2.3 PSpice simulation results showing the input voltage to output voltage transfer curves of (a) the resistor-loaded amplifier shown in Fig. 2.1 and (b) the passive PMOS loaded amplifier shown in Fig. 2.2.

to the intersection of the characteristic curve of the load transistor (M2) and the curve of the input transistor corresponding to its bias voltage, V_{GS1}. To provide a symmetrical variation for the output voltage around the quiescent value, this point must be situated in the middle of the dynamic range, which is the region where both M1 and M2 are in saturation mode. This operating range has a width equal to $V_{DD} - (V_{DS1sat} + V_{DS2sat})$.

The amount of change (i.e., the dynamic range) in the drain current corresponding to the dynamic range of the output voltage is quite small. This means that the full swing of the output voltage can be obtained with a comparatively small swing of the input signal voltage. Consequently, (a) this circuit produces a higher voltage gain, and (b) since this circuit operates in a small portion of the nonlinear $I_D - V_{GS}$ curve, the linearity is better, in other words, the nonlinear distortion is smaller.

In Fig. 2.3 the voltage transfer curves of (a) a resistor-loaded and (b) a PMOS transistor-loaded common-source amplifier, operating under identical conditions, are given. The comparison of these curves shows that the small-signal voltage gain (that corresponds to the slope of the curve at the operating point) of the transistor-loaded amplifier is obviously higher than the gain of a resistor-loaded amplifier.

The passive transistor-loaded amplifiers have a serious problem: the appropriate biasing of the load transistor to maintain the operating point in the middle of dynamic range is very critical. As can be easily seen from Fig. 2.2(d), a small change of the load curve (that can arise from the value of the bias voltage of M2 or can be owing to the tolerances of M2 and/or M1) can shift the operating point, and the output voltage usable dynamic range can decrease considerably. To overcome this problem,

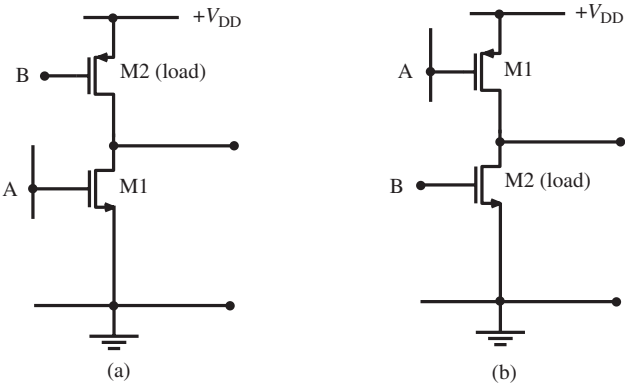


Figure 2.4 (a) NMOS input common-source amplifier. (b) PMOS input common-source amplifier. A indicates the output node of the previous stage, and B indicates the DC source to bias the load transistor.

the passive MOS loaded amplifiers are always used in circuits having an overall negative feedback to stabilize the operating points.[3]

2.1.1 Biasing

The DC gate bias voltage of a common-source amplifier is usually supplied by the output of the previous stage, together with the input signal. This means that the previous stage must be designed in such a way that the quiescent DC voltage on the output node has the appropriate value to bias the common-source stage at the predefined operating point. The type of the transistor of the common-source stage must be chosen depending on the level of the DC voltage on the output of the previous stage. Usually, an NMOS transistor is preferable for DC voltages close to ground (0 V), and a PMOS transistor is preferable for DC voltages close to the positive supply (Fig. 2.4).

For common-source input stages, the gate has to be independently biased with a DC voltage source. If capacitive coupling of the signal to the gate of the input transistor is permissible, the gate can be connected to the bias voltage source through a high-value resistor or an RF "choking coil" (in radio frequency applications), as shown in Fig. 2.5(a) and Fig. 2.5(b).

The bias voltage can be obtained from the main DC voltage source (for example V_{DD}) with a resistive or MOS voltage divider. If the input signal source has to be directly coupled to the gate of the transistor (the quiescent DC voltage on the input node is zero), the source must be connected to an appropriate negative voltage source, to maintain the appropriate gate–source DC bias (Fig. 2.5(c)).

[3] For example, the second single-ended gain stages of CMOS operational amplifiers are usually passive loaded MOS amplifiers.

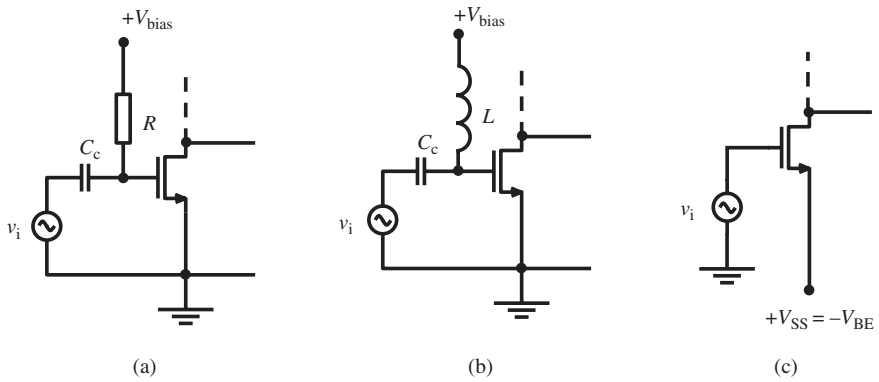


Figure 2.5 Biasing of the input transistor. (a) and (b) AC coupling, (c) DC coupling of the signal to the input of the amplifier.

For passive MOS transistor-loaded amplifiers, the fixed DC bias voltage of the load transistor can be obtained from V_{DD} using a suitable voltage divider.

2.1.2 The small-signal equivalent circuit

For an amplifier biased at an appropriate operating point, the characteristic curve representing the voltage–current relation of the device can be approximated by the tangent of this curve, as long as the magnitude of the signal applied to the input is small enough compared to the bias voltage at this point. This corresponds to the "linearization" of the device characteristics that permits us to use the simple and well-known network analysis procedures, instead of the complicated nonlinear solution methods that are practically impossible for hand calculations.

For a non-velocity saturated NMOS transistor biased in the normal saturation region, the drain current can be expressed in terms of the gate–source and drain–source voltages as

$$I_D = \frac{1}{2}\beta(V_{GS} - V_T)^2\left(\frac{1 + \lambda V_{DS}}{1 + \lambda(V_{GS} - V_T)}\right) \tag{2.1}$$

where $\beta = KP(W/L)$ and $KP = \mu C_{ox}$. It was mentioned in Chapter 1 that, especially for small geometries (for very thin gate oxides), the mobility strongly decreases with the transversal electric field in the gate region. Therefore, in (2.1) the mobility value corresponding to the quiescent gate–source voltage must be used. In addition, if there is a substrate bias, the threshold voltage corresponding to this condition must be used. With these precautions, the quiescent drain current can be calculated in terms of V_{DS} and V_{GS}.

If we expand (2.1) into a Taylor series around the operating point and assume that variations of voltages and current are sufficiently small compared to the quiescent values, we can neglect all higher order terms and express the drain current as

$$I_D + \Delta I_D \cong I_D + \left.\frac{\partial I_D}{\partial V_{GS}}\right|_{\Delta V_{DS}=0} \times \Delta V_{GS} + \left.\frac{\partial I_D}{\partial V_{DS}}\right|_{\Delta V_{GS}=0} \times \Delta V_{DS}$$

or

$$\Delta I_D \cong \left.\frac{\partial I_D}{\partial V_{GS}}\right|_{\Delta V_{DS}=0} \times \Delta V_{GS} + \left.\frac{\partial I_D}{\partial V_{DS}}\right|_{\Delta V_{GS}=0} \times \Delta V_{DS} \tag{2.2}$$

Equation (2.2) indicates that the change of the drain current is (approximately) a linear function of the variations of the gate and drain voltages. The coefficients of this linear expression are

$$g_m = \left.\frac{\partial I_D}{\partial V_{GS}}\right|_{\Delta V_{DS}=0} \quad \text{and} \quad g_o = \left.\frac{\partial I_D}{\partial V_{DS}}\right|_{\Delta V_{GS}=0} \tag{2.3}$$

and can be calculated as the partial first derivatives of (2.1) at the operating point. They are the previously defined "transconductance" and "output conductance" parameters of the transistor[4] for this operating point. Now (2.2) can be written as

$$\Delta I_D = g_m \Delta V_{GS} + g_o \Delta V_{DS} \tag{2.4}$$

Each of these two small-signal parameters of the MOS transistor for any (V_{GS}, V_{DS}, I_D) operating point can be calculated from (2.1) and (2.3). For the transconductance parameter, we use (see Chapter 1)

$$g_m = \beta(V_{GS} - V_T)(1 + \lambda V_{DS}) \frac{2 + \lambda(V_{GS} - V_T)}{2\left(1 + \lambda(V_{GS} - V_T)^2\right)} \tag{2.5a}$$

This expression can be simplified for small λ values to

$$g_m \cong \beta(V_{GS} - V_T) \tag{2.5b}$$

Although it is an approximate formula especially for short channel transistors (where the λ parameter is not negligibly small), (2.5b) is very convenient for hand calculations. Another simple and useful expression that gives the transconductance in terms of the drain DC current is

$$g_m \cong \sqrt{2\beta I_D} \tag{2.5c}$$

The corresponding expressions for PMOS transistors are

$$|I_D| = \frac{1}{2}\beta(V_{GS} - V_T)^2 \left(\frac{1 + \lambda|V_{DS}|}{1 + \lambda|V_{GS} - V_T|}\right)$$

$$g_m \cong \beta|(V_{GS} - V_T)|$$

$$g_m \cong \sqrt{2\beta|I_D|}$$

[4] Throughout this book, the output conductance and the output resistance of a transistor will be represented interchangeably by g_o or g_{ds} and r_o or r_{ds}, respectively.

In the case of velocity saturation (see Chapter 1), the drain current can be written as

$$|I_D| = kWC_{ox}(V_{GS} - V_T)v_{sat} = G_m(V_{GS} - V_T) \tag{2.6}$$

and the transconductance is found to be:

$$g_m = \frac{dI_D}{dV_{GS}} = kWC_{ox}v_{sat} \tag{2.7}$$

which is valid for both NMOS and PMOS transistors. Equation (2.7) indicates that for a short channel MOS transistor operating in the velocity saturation regime, the transconductance parameter does not depend on the operating point and is proportional to the gate width.

The output conductance parameter of an NMOS transistor calculated from (2.1) and (2.3) is

$$g_o = I_D \frac{\lambda}{(1 + \lambda V_{DS})} \tag{2.8a}$$

and can be simplified for small λ values to

$$g_o \cong I_D\lambda \tag{2.8b}$$

For PMOS transistors, owing to the negative polarity of the source–drain voltage and the drain current, the output conductance parameter must be expressed as

$$g_o = |I_D| \frac{\lambda}{(1 + \lambda|V_{DS}|)} \cong |I_D|\lambda \tag{2.8c}$$

From (2.4) we can also derive a different set of definitions for g_m and g_o that is useful for obtaining the values of these parameters by direct measurements, or from the characteristic curves of the transistor:

$$g_m = \left.\frac{\Delta I_D}{\Delta V_{GS}}\right|_{\Delta V_{DS}=0} \qquad \text{and} \qquad g_o = \left.\frac{\Delta I_D}{\Delta V_{DS}}\right|_{\Delta V_{GS}=0} \tag{2.9}$$

The transconductance is the most important parameter of the transistor and expresses the controllability of the gate–source voltage on the drain current. The definition of g_m given in (2.3) indicates that its value is equal to the slope of the tangent at the operating point of the $I_D = f(V_{GS})$ curve (drawn at the quiescent value of V_{DS}). Similarly, g_o is the slope of the tangent at the operating point of the $I_D = f(V_{DS})$ output characteristic curve, corresponding to the V_{GS} quiescent value.

In (2.4) the Δ variations of the voltages and the current can be replaced by the instantaneous values of the small-signal components on the quiescent DC values and the expression can be written as

$$i_d = g_m v_{gs} + g_{ds} v_{ds} \tag{2.10}$$

which can be represented by a circuit as shown in Fig. 2.6. This "small-signal equivalent circuit" is very useful for solving the relations of signal currents and

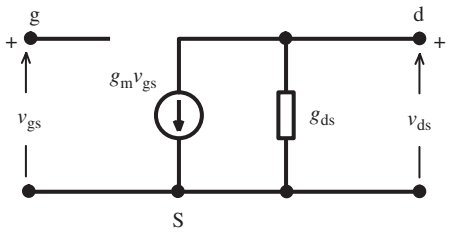


Figure 2.6 The small-signal equivalent circuit of a MOS transistor biased at a certain operating point in the saturation region.

voltages in a circuit containing MOS transistors. For high frequencies, it is necessary to include into this equivalent circuit other physical and parasitic components, as will be investigated in depth in the following chapters.

The basic rules of deriving the small-signal equivalent of a circuit can be summarized as follows.

- Calculate (determine) the DC quiescent voltages and currents of each transistor.
- Calculate the small-signal parameters (g_m and g_{ds}) corresponding to this operating point.
- Replace the transistors in the circuit with their small-signal equivalents.
- Include all passive components of the circuit.
- Short-circuit all ideal DC voltage sources (supply voltages, bias voltages, etc.).
- Open-circuit all ideal DC current sources.
- If the DC voltage and/or current sources are not ideal (i.e. if they have serial or parallel internal impedances), include them in the small-signal equivalent circuit.
- Connect the signal source(s), and solve the circuit using conventional linear circuit analysis methods.
- To obtain reasonable results from hand calculations, apply sufficiently small magnitudes to the driving signal sources such that all output signals remain in the dynamic ranges corresponding to these outputs.[5]

In Fig. 2.7(a) and (b) the small-signal equivalent circuits corresponding to the amplifiers shown in Fig. 2.1 and Fig. 2.2 are given. From Fig. 2.7(a), the output voltage (v_o) can be easily calculated in terms of the input signal voltage (v_i) and then the small-signal voltage gain of the circuit is

$$A_v = \frac{v_o}{v_i} = -g_m(r_{ds}//R_D) = -g_m \frac{r_{ds}R_D}{r_{ds} + R_D}, \tag{2.11a}$$

[5] Computer aided (for example SPICE) analyses have the same problem. Once the software forms the small-signal equivalent circuit, it assumes that the circuit is linear for all amplitudes of the input signal(s). It is the user's responsibility to apply reasonable input signals in order to keep the magnitudes of the output signals within the dynamic ranges. In case of doubt, it is advisable to run a "transient" simulation to be sure that there is no excessive nonlinear distortion or clipping on the output signals.

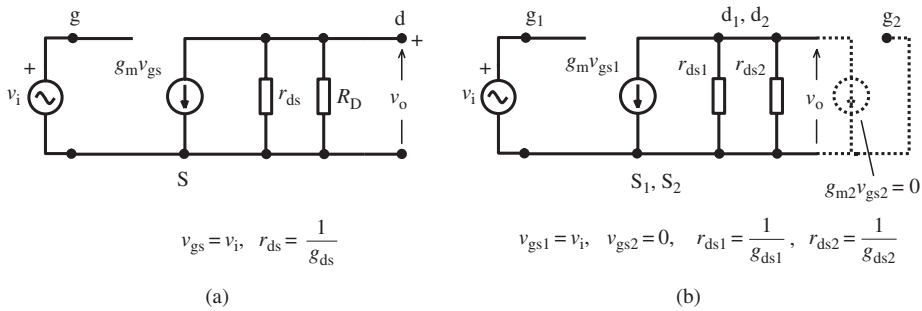


Figure 2.7 The small-signal equivalent circuits of (a) a resistor-loaded common-source amplifier, (b) a passive MOS transistor-loaded common-source amplifier.

which reduces to

$$A_v = -g_m R_D \tag{2.11b}$$

for $r_{ds} \gg R_D$.

The small-signal voltage gain of a passive-loaded common-source amplifier can be calculated from Fig. 2.7(b). Since there is no signal on the gate of the load transistor, it acts only as a resistor having a value equal to its output resistance at its operating point:

$$A_v = \frac{v_o}{v_i} = -g_{m1}(r_{o1} // r_{o2}) = -\frac{g_{m1}}{(g_{o1} + g_{o2})} \tag{2.12a}$$

If we insert the values of g_{m1}, g_{o1} and g_{o2} in terms of the DC drain current, we obtain

$$A_v \cong -\sqrt{2\beta_1 I_{D1}}\frac{1}{I_{D1}\lambda_1 + |I_{D2}|\lambda_2} = -\sqrt{\frac{2\beta_1}{I_{D1}}}\frac{1}{(\lambda_1 + \lambda_2)} \tag{2.12b}$$

This expression indicates that:

- the voltage gain of a common-source amplifier is higher for lower DC quiescent currents (in other words for lower power consumption);
- the voltage gain can be increased with load and/or input transistors having smaller λ parameters (larger channel lengths), certainly keeping the aspect ratios constant.

However, it must be kept in mind that the performance of the amplifier deteriorates at high frequencies owing to the increased output internal resistance and increased parasitic drain junction capacitances, as will be explained in Chapter 3.

Example 2.1 An amplifier as shown in Fig. 2.2(a) is designed for AMS 0.35 µ technology. The supply voltage is 3.2 V. Channel widths and lengths for both of the transistors are 35 µm and 0.35 µm, respectively.

(a) Calculate the gate bias voltages for M1 and M2, to fix the operating point in the middle of the operating range, i.e. 1.6 V, for 1 mA quiescent current.

Since the lambda parameters of the transistors (especially of M2) are considerably high and the supply voltage is low, it is appropriate to use (1.14a) to calculate the drain current, which is

$$|I_{D2}| \cong \frac{1}{2}\beta_2 (V_{GS2} - V_{Tp})^2 \frac{1 + \lambda_2 |V_{DS2}|}{1 + \lambda_2 |(V_{GS2} - V_{Tp})|}$$

for M2.

The parameters of M2 for AMS 0.35μ technology are given as $\mu_{p0} = 137$ [cm^2/V.s], $C_{ox} = 4.56 \times 10^{-7}$ [F/cm^2], $V_{Tp} = -0.7$ [V], and $\lambda_2 \approx 0.2$ [V^{-1}]. $(W/L)_2$ is given as 100 and V_{D2} must be in the middle of the output voltage dynamic range, i.e. 1.6 V. β_2 can be calculated as

$$\beta_2 = \mu_p C_{ox} (W/L)_2 = 137 \times (4.56 \times 10^{-7}) \times 100 = 6.25 \times 10^{-3} \ [\mathrm{A/V^2}]$$

Inserting $|(V_{GS2} - V_{Tp})| = V_{Go2}$ and $(1 + \lambda_2 |V_{DS2}|) = (1 + 0.2 \times 1.6) = 1.32 = a$, to simplify the expressions, V_{Go2} can be solved as

$$V_{Go2} = \frac{1}{2\beta_2 a}\left(2\lambda_2 |I_{D2}| \mp \sqrt{(2\lambda_1 |I_{D2}|)^2 + 8|I_{D2}|\beta_2 a}\right)$$

$$V_{Go2} = \frac{1}{2(6.25 \times 10^{-3})1.32} \times \left(2 \times 0.2 \times 10^{-3} \mp \sqrt{(2 \times 0.2 \times 10^{-3})^2 + 8 \times 10^{-3}(6.25 \times 10^{-3}) \times 1.32}\right) = 0.516 \text{ V}$$

Since M2 is a PMOS transistor, V_{GS2} and V_{Tp} are negative,

$$V_{GS2} = -V_{Go2} + V_{Tp} = -0.516 + (-0.7) = -1.216 \text{ V}.$$

Similarly for the NMOS M1 transistor with $\mu_{n0} = 475.8$ [cm^2/V.s], $V_{Tn} = 0.5$ [V], $\lambda_1 \approx 0.073$ [V^{-1}]. and $(W/L)_1 = 100$, the gate–source bias voltage of M1 can be calculated as $V_{GS1} = 0.79$ V.

(b) Calculate the voltage gain. From (2.12a)

$$A_v = -\sqrt{2 \times [475.8 \times (4.56 \times 10^{-7}) \times 100]} \frac{1}{10^{-3}} \frac{1}{(0.073 + 0.2)} = -24.13 \ (27.6 \text{ dB})$$

(c) Check the results with PSpice simulations.

It is common practice to use the calculated bias voltage values as inputs to SPICE. But owing to the approximations in the expressions, the hand-calculated bias voltage values do not provide the targeted drain currents in most cases and a fine-tuning becomes necessary. It is useful to develop a tuning strategy to prevent tedious iterations to converge on the appropriate operating conditions. The DC sweep feature of SPICE is a useful possibility. For this example we will apply the following procedure.

- Fix the drain voltage of the load transistor to the targeted value (connect a 1.6 V DC supply between the drain node of M2 and the ground).

- DC sweep the gate bias voltage in a reasonable interval. Note the value of the gate voltage (V_{G2}) corresponding to the targeted drain current (-1 mA in this case).
- Disconnect the 1.6 V supply.
- Apply V_{G2} to the gate of M2.
- DC sweep the gate bias voltage of M1 (V_{GS1}) and observe the drain node voltage of M1 as a function of V_{GS1}. This is the voltage transfer curve of the amplifier and the V_{GS1} value corresponding to $V_{DS1} = 1.6$ V is the appropriate bias voltage.

The obtained PSpice results are:

$$V_{GS1} = 0.9658 \text{ V},\ V_{G2} = 1.775 \text{ V}\ (V_{GS2} = -1.425 \text{ V}) \text{ and } A_v = -18.57\ (25.37 \text{ dB})$$

The discrepancies between the hand calculation and simulation results arise from several secondary effects that are not included in the expressions. For example, the mobility degradation owing to the transversal field and the series source resistance are the most important effects that result in the lowering of the β parameter of the transistor. Certainly it is possible to take this effect into account as explained in Chapter 1, but since a fine-tuning is unavoidable, it is wiser not to complicate the calculations further.

Another important fact to note is that the gate–source voltages of M1 and M2 are given with high precision, since the positions of the operating point and the quiescent current strongly depend on the bias voltages. Therefore, as already mentioned, this type of circuit must be used in an appropriate negative feedback loop, to keep the operating point in the middle of the voltage transfer curve.

(d) Repeat the simulation for $I_D = 1$ mA and $L_2 = 1$ μm, $W_2 = 100$ μm.

The PSpice simulation results obtained after a fine-tuning of the operating point (which is necessary to bring the operating point to the middle of the voltage transfer curve) are as follows:

$$V_{GS1} = 0.987 \text{ V}, V_{G2} = 1.775 \text{ V}\ (V_{GS2} = -1.425 \text{ V}) \text{ and } A_v = -33.2\ (30.42 \text{ dB})$$

The voltage gain is considerably (more than 5 dB) increased as expected, at the expense of higher area consumption and higher parasitic capacitance on the output node. Certainly it is possible to increase the dimensions of M1 as well, but the lambda parameters of 0.35 μm NMOS transistors are already small and to increase the gate length does not provide significant increase of the voltage gain. (The increase in the C_{gs} and C_{dg} parameters is another drawback affecting the high-frequency performance of the circuit, as will be explained in Chapter 3.)

A simple but important modification of a common-source amplifier is shown in Fig. 2.8(a), where a resistor is connected between the source terminal and the ground. From the small-signal equivalent circuit (Fig. 2.8(b)), the voltage gain of this circuit can be calculated as

$$A_v = -g_m \frac{R_D}{(1 + g_m R_S) + \frac{(R_D + R_S)}{r_o}} \tag{2.13a}$$

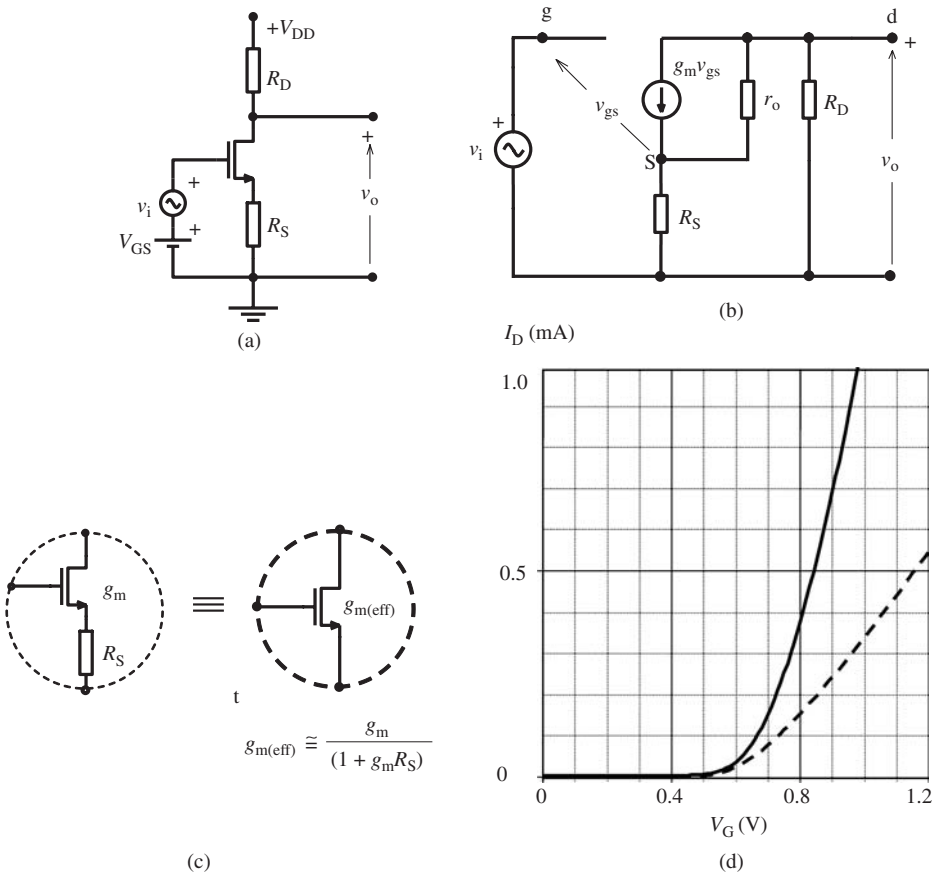


Figure 2.8 (a) Resistor-loaded common-source amplifier with a series source resistor. (b) The small-signal equivalent circuit. (c) The definition of the effective transconductance. (d) Transfer characteristic of a MOS transistor without R_S (solid line) and with $R_S = 500\ \Omega$ (dashed line)

which can be simplified for $r_o \gg (R_D + R_S)$ to

$$A_v \cong -\frac{g_m}{(1 + g_m R_S)} R_D \tag{2.13b}$$

If we compare (2.13b) with (2.11b) we conclude that a resistor connected in series to the source terminal reduces the transconductance of the transistor to a smaller value, which we will call the "effective transconductance":

$$g_{m(eff)} = \frac{g_m}{1 + g_m R_S} \tag{2.13c}$$

Another effect of R_S is the linearization of the input voltage–output current transfer characteristic and the consequent decrease of the nonlinear distortion. To visualize this effect, the transfer characteristics of a MOS transistor alone and the transfer characteristics of the transistor–resistor combination are shown in Fig. 2.8(d). Note the

decrease of the slope of the curve (which is the transconductance) and the improvement of the linearity. This technique is extensively used to improve the linearity (or decrease the nonlinearity related distortions) of MOS amplifiers at the expense of reducing the magnitude of the gain. It is obvious that the gate bias voltage of such a transistor must be

$$V_G = V_{GS} + I_D R_S$$

where V_{GS} is the gate–source voltage to flow I_D.

Example 2.2 Let us calculate the effects of the parasitic source resistance of a 35 μm/0.35 μm NMOS transistor. The *KP* parameter of this technology is given as 170 μA/V^2. Assume that the dominant part of the parasitic series resistance is the intrinsic component, and that the extrinsic source-region resistance and the contact resistances are negligible. The total series resistance for NMOS transistors is given as $RDSW = 345$ ohm per micron width.

(a) Calculate the parasitic series source resistance of the 35 μm/0.35 μm transistor.

$$R_S + R_D = \frac{RDSW}{W} = \frac{345}{35} = 9.86 \cong 10 \Rightarrow R_S \cong 5 \text{ ohm}$$

(b) For $I_D = 1$ mA, calculate the decrease of the transconductance owing to this series resistance and the necessary increase of the gate bias voltage to compensate the DC voltage drop on this resistance.
From (2.13b), the reduction of the transconductance is found to be

$$\frac{g_{m(eff)}}{g_m} = \frac{1}{1 + g_m R_S}$$

$$g_m R_S = \sqrt{2\beta I_D} = \sqrt{2KP\frac{W}{L} I_D} \times R_S$$

$$g_m R_S = \sqrt{2 \times (170 \times 10^{-6}) \times 100 \times 10^{-3}} \times 5 = 0.029$$

$$\frac{g_{m(eff)}}{g_m} = \frac{1}{1 + 0.029} = 0.97$$

The DC voltage drop on R_S is $5 \times 10^{-3} = 5$ mV.

(c) Repeat (b) for $I_D = 10$ mA.
A factor of 10 increase of I_D increases $g_m R_S$ by a factor of $\sqrt{10} = 3.16$, to $0.029 \times 3.16 = 0.09$.

Hence

$$\frac{g_{m(eff)}}{g_m} = \frac{1}{1 + 0.09} = 0.92$$

The DC voltage drop on R_S is $R_S I_D = 5 \times (10 \times 10^{-3}) = 50$ mV.

Problem 2.1 *Derive an expression for a velocity saturated transistor to calculate the reduction of the transconductance as a function of the source-region series parasitic resistance.*

2.2 Active transistor loaded MOS amplifier (CMOS inverter as analog amplifier)

The passive PMOS load of the NMOS input transistor in the amplifier shown in Fig. 2.2(a) can be "activated" by connecting its gate to the input node, instead of a constant bias voltage (Fig. 2.9(a)). This circuit is nothing but a CMOS logic inverter, which is investigated in depth in all books on digital CMOS circuits [24], [25]. The input voltage–output voltage transfer curve of such an inverter is given in Fig. 2.9(b), where the output logic (1) and logic (0) regions are marked with (A) and (C), respectively. In the transition region (B), the output voltage is – strongly and rather linearly – controlled by the input voltage. This means that if we bias the input node in such a way that the DC quiescent voltage of the output node is in the middle of this linear region (i.e. equal to $V_{DD}/2$), this circuit can be used as an analog amplifier. An even more advantageous solution is to design the circuit in such a way that the input DC bias voltage also is equal to zero, which permits direct coupling to single-ended signal sources and loads.

The voltage transfer curve of an inverter amplifier shown in Fig. 2.9(b) shows us that the output signal can change from zero to V_{DD} (rail to rail). But for analog amplification with low nonlinear distortion, it is convenient to exclude the upper and lower portions of the transfer curve and restrict the output signal dynamic range into the region where both M1 and M2 are in the saturation mode.

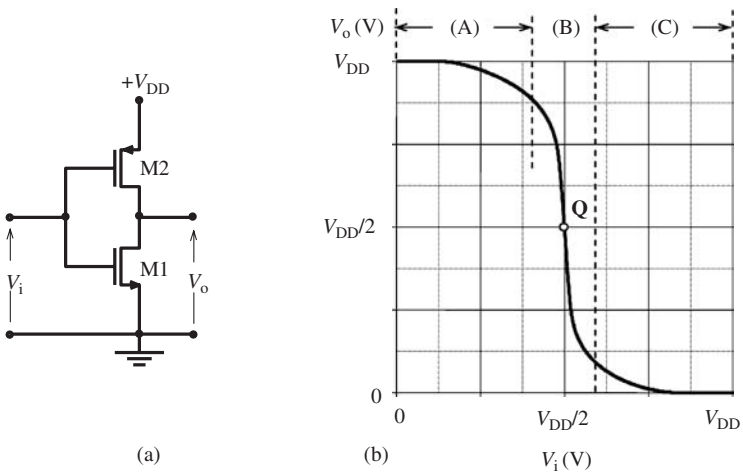


Figure 2.9 The CMOS inverter as a linear amplifier (a) and its voltage transfer curve (b).

In Fig. 2.9(a), assuming that both M1 and M2 are not in the velocity saturation regime, for an input voltage V_i and the corresponding output voltage V_o, the drain currents are

$$I_{D1} = \frac{1}{2}\mu_n C_{ox}\left(\frac{W_1}{L_1}\right)(V_i - V_{Tn})^2(1 + \lambda_n.V_o)$$
$$I_{D2} = -\frac{1}{2}\mu_p C_{ox}\left(\frac{W_2}{L_2}\right)(V_i - V_{DD} - V_{Tp})^2[1 + \lambda_p(V_{DD} - V_o)]$$

and they are equal in magnitude ($I_{D1} = -I_{D2}$). This yields an expression which is useful for calculating the necessary conditions to set the operating point in the middle of the transfer curve (i.e. V_i and V_o, both are equal to $V_{DD}/2$).

$$\frac{W_2}{W_1} = \frac{L_2}{L_1}\frac{\mu_n}{\mu_p}\left(\frac{V_i - V_{Tn}}{V_i - V_{DD} - V_{Tp}}\right)^2 \frac{1 + \lambda_n.V_o}{1 + \lambda_p(V_{DD} - V_o)} \tag{2.14}$$

Since the input voltage–output voltage transfer curve is very steep around the operating point, as shown in Fig. 2.9(b), the sensitivity of the output voltage with respect to the tolerances of the input voltage and the transistor parameters is very high. Therefore it is quite difficult to set the operating point at the desired position. The dimensions (ratios) calculated according to (2.14) usually do not give an operating point accurately (even sufficiently) in the middle of the linear operating range.[6]

A simple solution to improve the stability of the operating point is to connect a resistor between the input and output nodes, as shown in Fig. 2.10(a), thereby forcing the output voltage to approach the input DC voltage, which provides a current feedback to the input node. In Fig. 2.10(b) simulation results are shown for the transfer curves of a CMOS inverter amplifier – which is intentionally not optimized, with and without a feedback resistor. It can be seen that the operating point indeed moves toward the middle of the output range as expected, however, the price of this improvement is a reduction of the slope (the small-signal voltage gain of the amplifier) and the dynamic range.

Another possibility for extending the usability of this circuit is to use symmetrical twin power supplies, as shown in Fig. 2.11(a). This solution provides two important advantages: (a) since the input DC bias voltage is zero, the AC (or pulse) signal source can be directly connected to the input node, (b) the load of the amplifier – which can be resistive or reactive – can be directly connected between the output node and the ground (Fig. 2.11(b)). It must be kept in mind that transistors must be able to deliver the necessary current to swing the voltage of the output node up or down to its maximum or minimum value. In the case of a capacitive load, since the transistors are operating in the saturation region and hence acting as current sources, the current supplied by M1 or M2 discharges or charges the load capacitor with a constant slope.

[6] Especially for small geometries, for which the NMOS transistors usually operate in the velocity saturation region (but not the PMOS transistors, as explained in Chapter 1), (2.1) is not valid any more. Therefore a fine-tuning with SPICE simulation becomes compulsory.

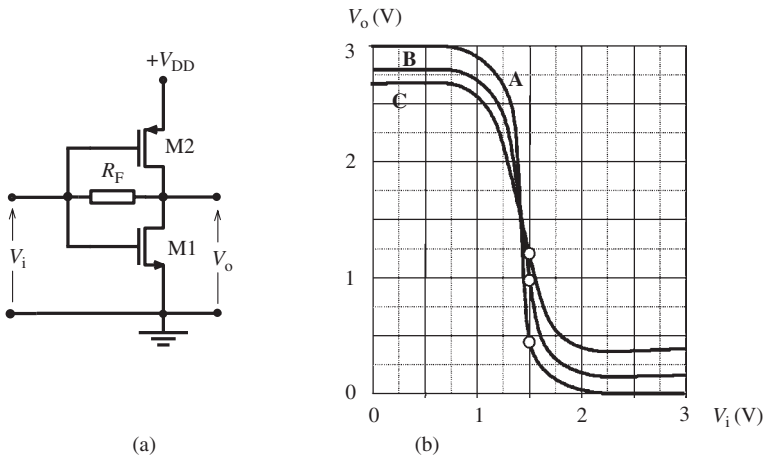


Figure 2.10 (a) A CMOS inverter amplifier with R_F feedback resistor. (b) PSpice simulation results with no feedback resistor (A), with $R_F = 10$ k (B) and with $R_F = 5$ k (C). (Transistors: $L = 0.6\ \mu$, $W_1 = 6\ \mu$, $W_2 = 15\ \mu$.)

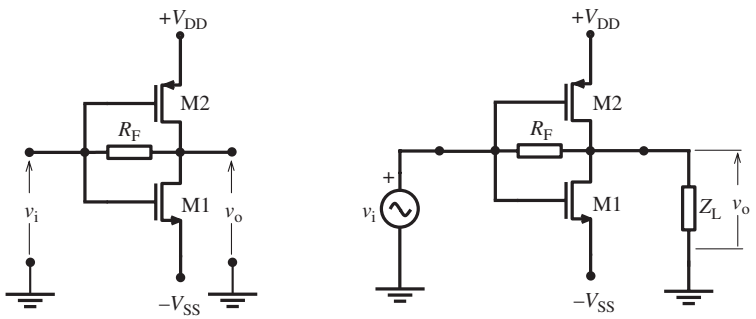


Figure 2.11 CMOS inverter amplifier biased by two DC power supplies to force the input and output quiescent voltages to zero.

Consequently, the output voltage can reach its minimum or maximum value within a certain time. This "slew-rate" effect will be investigated in Chapter 3.

The small-signal equivalent of this circuit and re-arranged forms of it to facilitate the solution are given in Fig. 2.12(a), (b) and (c), respectively. From 2.12(c) the voltage gain can be easily obtained as

$$A_v = -\frac{\bar{g}_m - G_F}{\bar{G}_L + G_F} \tag{2.15}$$

which reduces to

$$A_v = -\frac{\bar{g}_m}{\bar{G}_L} \tag{2.16}$$

for $G_F = 0$. Comparing (2.15) and (2.16), one can conclude that there are two separate signal paths from input to output: one amplifier path and a passive divider path. To prevent the domination of the passive path the condition of $\bar{g}_m \succ G_F$ must be fulfilled.

Owing to the R_F feedback resistor, the input conductance of the amplifier is not zero any more. It can be calculated from the small-signal equivalent circuit given in Fig. 2.12(b) as

$$g_i = \frac{\bar{G}_L + \bar{g}_m}{\bar{G}_L + G_F} G_F \tag{2.17}$$

For $G_F \gg \bar{G}_L$ the input conductance is approximately equal to $\bar{G}_L + \bar{g}_m \cong \bar{g}_m$. It means that this is a low-input resistance circuit; therefore this circuit can also be considered as a "transresistance amplifier", suitable for use with high internal impedance sources.

The value of the transresistance can be calculated as

$$R_m = \frac{v_o}{i_i} = \frac{v_o}{v_i} \cdot \frac{v_i}{i_i} = A_v . r_i = -\frac{\bar{g}_m - G_F}{\bar{g}_m + \bar{G}_L} \frac{1}{G_F} \tag{2.18}$$

and reduces to

$$R_m \simeq -R_F$$

for $G_F, \bar{G}_L \ll \bar{g}_m$

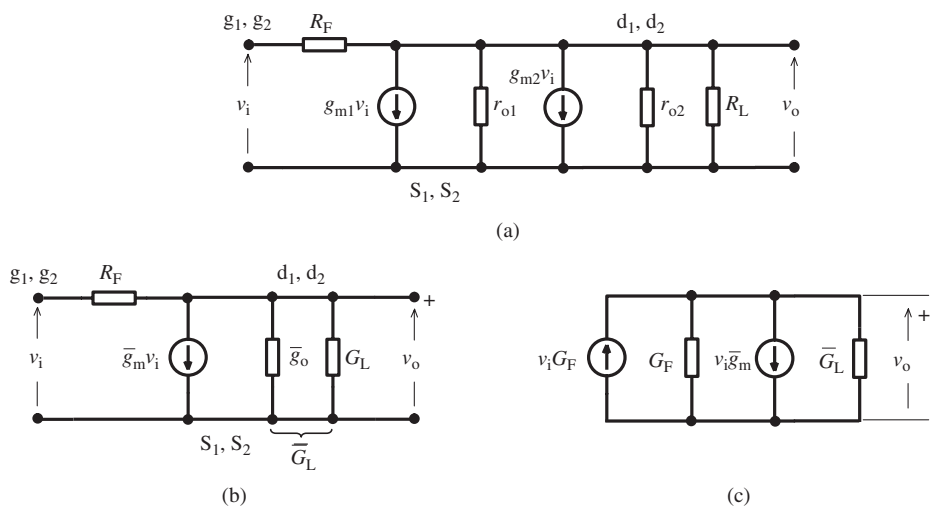


Figure 2.12 The small-signal equivalent circuit of the inverter amplifier given in Fig. 2.10.

Example 2.3 A CMOS inverter amplifier will be designed. The DC supply is a ± 1.5 V twin voltage source. The quiescent DC current must be 0.5 mA.[7] The design will be made for AMS 0.35 µ technology. The channel lengths of M1 and M2 will be chosen as 0.35 µm to maximize the RF and pulse performance.

(a) Calculate the channel widths of M1 and M2.
From the drain current expression of M1,

$$\left(\frac{W_1}{L}\right) = \frac{2I_{D1}}{(KP)_n(V_{GS1} - V_{Tn})^2(1 + \lambda_n V_{DS1})}, \left(\frac{W_1}{L}\right)$$
$$= \frac{2 \times (0.5 \times 10^{-3})}{(170 \times 10^{-6})(1.5 - 0.5)^2(1 + 0.073 \times 1.5)} = 5.3 \rightarrow W_1 \cong 1.86\,\mu\text{m}$$

The channel width of M2 can be calculated similarly, or from (2.14) as $W_2 \cong 6.4\,\mu$m.

(b) Calculate the voltage gain.
To calculate the voltage gain from (2.16), the transconductances and the output conductances of M1 and M2 must be found.

The β parameters of M1 and M2 are

$$\beta_1 = (KP)_n(W_1/L) = (170 \times 10^{-6}) \times 5.5 = 0.9 \times 10^{-3}\ [\text{A/V}^2]$$
$$\beta_2 = (KP)_p(W_2/L) = (58 \times 10^{-6}) \times 18.9 = 1.096 \times 10^{3}\ [\text{A/V}^2]$$

Transconductances:

$$g_{m1} = \sqrt{2 \times (0.9 \times 10^{-3}) \times (0.5 \times 10^{-3})} = 0.95\,\text{mS}$$
$$g_{m2} = \sqrt{2 \times (1.1 \times 10^{-3}) \times (0.5 \times 10^{-3})} = 1.05\,\text{mS}$$
$$\bar{g}_m = 0.95 + 1.05 = 2\,\text{mS}$$

Output conductances:

$$g_{o1} \cong \lambda_n I_{D1} = 0.073 \times (0.5 \times 10^{-3}) = 36.5 \times 10^{-6}\,\text{S}$$
$$g_{o2} \cong \lambda_p |I_{D2}| = 0.2 \times (0.5 \times 10^{-3}) = 100 \times 10^{-6}\,\text{S}$$

Since there is no external resistive load,

$$\bar{G}_L = (36.5 + 100) \times 10^{-6} = 0.1365\,\text{mS}$$

Now the voltage gain:

$$A_v = -\frac{2 \times 10^{-3}}{0.1365 \times 10^{-3}} = -14.7\,(23.4\,\text{dB})$$

[7] The value of the DC quiescent current is related to the slew-rate of the amplifier and must be high for a high slew-rate.

(c) Compare these results with PSpice simulation results.

With the calculated channel widths, PSpice gives a zero offset voltage transfer curve, but the quiescent current of the circuit is 0.2 mA, which is considerably lower than the targeted value. To increase the DC current to 0.5 mA, a procedure as explained in Example 2.1 must be applied. At the end of this fine-tuning the channel widths are found to be 2.5 μm and 15.5 μm, respectively, and the voltage gain is 18.4 (25.3 dB).

2.3 Common-gate (grounded-gate) amplifier

We have seen that the basic principle of a common-source amplifier is to control the drain current with the gate–source voltage. This principle is also valid for the "common-gate" or the "grounded-gate" amplifier given in Fig. 2.13(a). V_{SG} biases the transistor in inversion mode and provided that $V_{GS} > V_T$ a channel exists and a current (I_D) flows through this channel. The DC voltage of the output node is $V_{DG} = V_{DD} - I_D R_D$. We know that the condition to keep the transistor in the saturation region is $V_{DS} \geq (V_{GS} - V_T)$. This condition can be expressed in terms of the output node quiescent voltage as

$$V_{DG} \geq -V_T. \tag{2.19}$$

This means that for a common-gate amplifier the DC voltage can decrease down to zero, and can even become negative. Consequently, the output voltage dynamic range of this circuit is larger than that of the common-source amplifier. The appropriate output quiescent voltage is the mid-point of the output dynamic range.

The circuit has an input current exactly equal to the output current, in other words the current gain is equal to unity. If the input current is kept constant, the output of a common-gate circuit acts as an almost ideal current source. The utilization of this property for an important class of amplifiers, namely the "cascode" circuits, will be explained later.

The small-signal equivalent circuit of a common-gate amplifier (Fig. 2.13(b)) can be drawn with a re-arrangement of the equivalent circuit given in Fig. 2.7, and Fig. 2.13(c) can be obtained with a Norton–Thévenin transformation. From Fig. 2.13(c) the input current can be

$$i_i = v_i \frac{1 + g_m . r_o}{r_o + R_D} \tag{2.20}$$

Since $v_o = i_i R_D$, the voltage gain can be calculated as

$$A_v = \frac{v_o}{v_i} = \frac{1 + g_m r_o}{(r_o + R_D)} R_D \tag{2.21a}$$

which can be reduced to

$$A_v \cong +g_m R_D \tag{2.21b}$$

for $g_m \gg (1/r_o)$ and $r_o \gg R_D$

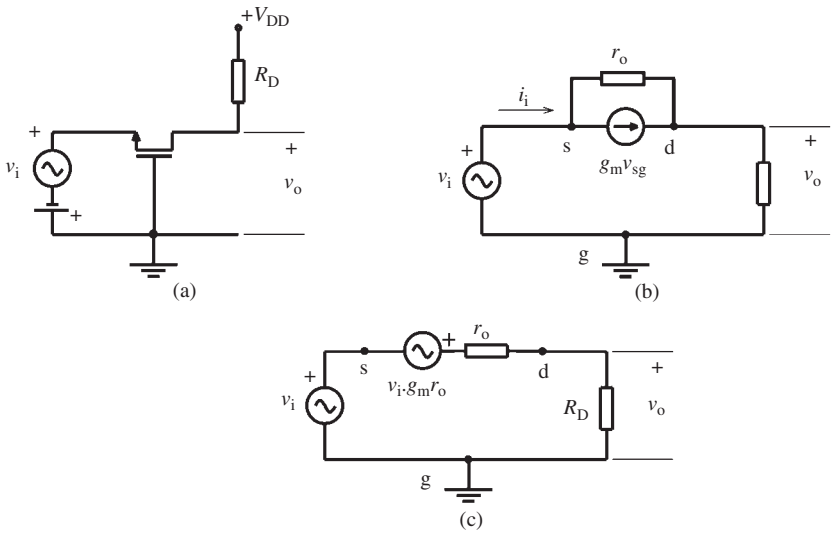


Figure 2.13 (a) The resistor-loaded common-gate amplifier, (b) the small-signal equivalent circuit and (c) the modified equivalent circuit following a Thévenin transformation.

The comparison of (2.11b) and (2.21b) indicates that:

- the magnitude of the small-signal voltage gain of a common-gate amplifier is (at least, approximately) equal to the gain of a common-source amplifier using the same transistor dimensions, operating point and load;
- unlike the common-source amplifier, the sign of the gain of a common-gate amplifier is positive. This means that the output signal is in phase with the input signal.

Another important difference of the common-gate amplifier is its input resistance. It can be intuitively understood that, owing to the high input current equal to the output current, the common-gate amplifier has a low input resistance. From (2.20) the small-signal input resistance can be calculated as

$$r_{\mathrm{i}} = \frac{r_{\mathrm{o}} + R_{\mathrm{D}}}{1 + g_{\mathrm{m}}.r_{\mathrm{o}}} \tag{2.22a}$$

and reduces to

$$r_{\mathrm{i}} \cong \frac{1}{g_{\mathrm{m}}} \tag{2.22b}$$

for $g_{\mathrm{m}} > (1/r_{\mathrm{o}})$ and $r_{\mathrm{o}} > R_{\mathrm{D}}$. This low-input resistance circuit can also be considered as a transresistance amplifier for high internal resistance signal sources, with

$$R_{\mathrm{m}} = R_{\mathrm{D}}. \tag{2.23}$$

2.4 Common-drain amplifier (source follower)

The third basic one-transistor amplifier configuration is the common-drain amplifier, generally called the "source follower". The circuit diagram of a source follower is shown in Fig. 2.14(a). The drain is directly connected to V_{DD} and the load resistance is connected between the source terminal and the ground. The DC current flowing through R_S is equal to the drain current in magnitude. The input DC bias voltage has to be equal to the sum of V_{GS} and the voltage drop on R_S owing to the corresponding drain current.

The initial small-signal equivalent circuit and a re-arranged form of it are given in Fig. 2.14(b) and Fig. 2.14(c), respectively. From Fig. 2.14(c) the small-signal voltage gain can be found as

$$A_v = \frac{g_m.\bar{R}}{1 + g_m.\bar{R}} \tag{2.24a}$$

where $\bar{R}$ is the parallel equivalent of the load resistor (R_S) and the output internal resistance of the transistor, corresponding to the operating point. Equation (2.24a) can be reduced to

$$A_v \cong +1 \tag{2.24b}$$

for $g_m.\bar{R} \gg 1$, which is valid for many practical cases. If we interpret (2.24b) together with the fact that there is a constant DC voltage difference equal to V_{GS} between the input terminal and the output terminal, we understand why this circuit is called a "voltage follower".

To better understand why the source follower has extensive use in electronic circuits in spite of unity voltage gain, the output internal resistance of the circuit must be calculated. In Fig. 2.14(d) the equivalent circuit is arranged to calculate the output internal resistance according to the definition. From this circuit the small-signal output internal resistance can be easily found as

$$r_o = \frac{v}{i_o} = \frac{\bar{R}}{1 + g_m.\bar{R}} \tag{2.25a}$$

which can be reduced to

$$r_o \cong \frac{1}{g_m} \tag{2.25b}$$

for $g_m.\bar{R} \gg 1$. We know that the transconductances of a MOS transistor operating in the non-velocity saturated regime and velocity saturated regime are

$$g_m \cong \sqrt{2\beta I_D} = \sqrt{2\mu C_{ox}\frac{W}{L}I_D}$$

$$g_m = k.C_{ox}W.v_{sat}$$

respectively. This means that it is possible to obtain small-output internal resistance values using appropriate W and I_D values for a non-velocity saturated transistor. For a

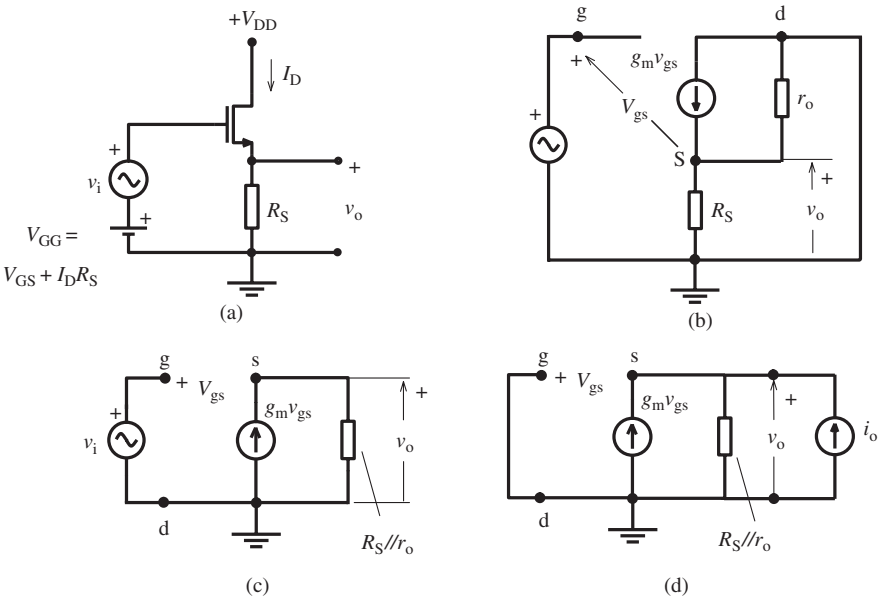


Figure 2.14 (a) The basic source follower. (b) The small-signal equivalent circuit and (c) its rearranged form. (d) The small-signal equivalent circuit to calculate the output internal resistance.

velocity saturated transistor the solution is even simpler as the output internal resistance of the circuit is inversely proportional to the gate width. As already explained in principle, a high input impedance–low internal output impedance circuit is useful as a buffer, for efficient signal transfer from a high internal resistance (non-ideal) voltage source to a low impedance load.

A different configuration of the source follower uses a current source instead of the resistor, as shown in Fig. 2.15(a). If there is a resistive load to be driven by this amplifier, it is usually more convenient to use twin DC power supplies to bring the quiescent voltage of the output node to zero and hence to prevent any DC current flowing through the load resistor (Fig. 2.15(b)). The DC current source connected to the source of the transistor is usually a MOS transistor, as shown in Fig. 2.15(c). This transistor (M2) must be biased to conduct a current equal to the drain DC current of M1, when the voltage of the output node is equal to zero. Certainly, both of these transistors must remain in the saturation region for the whole output voltage dynamic range.

For a negative input signal the drain (and source) current of M1 decreases by $i_d(-)$. To maintain the current of the current source ($I_S = I_{D2}$) constant, a current equal to $i_d(-)$ has to flow through R_L. Owing to the direction of this current, the output voltage decreases. Depending on the value of the signal amplitude, I_{D1} can decrease down to zero. For this extreme case, $i_d(-)$ becomes equal to I_S. To maintain M2 in the saturation region

$$V_{SS} - I_{D2}.R_L = V_{DS2} \geq (V_{GS2} - V_T) \tag{2.26}$$

must be satisfied for a given value of R_L. The output voltage corresponding to this case (the negative peak value of the output voltage) is $\check{v}_o = -I_S.R_L$.

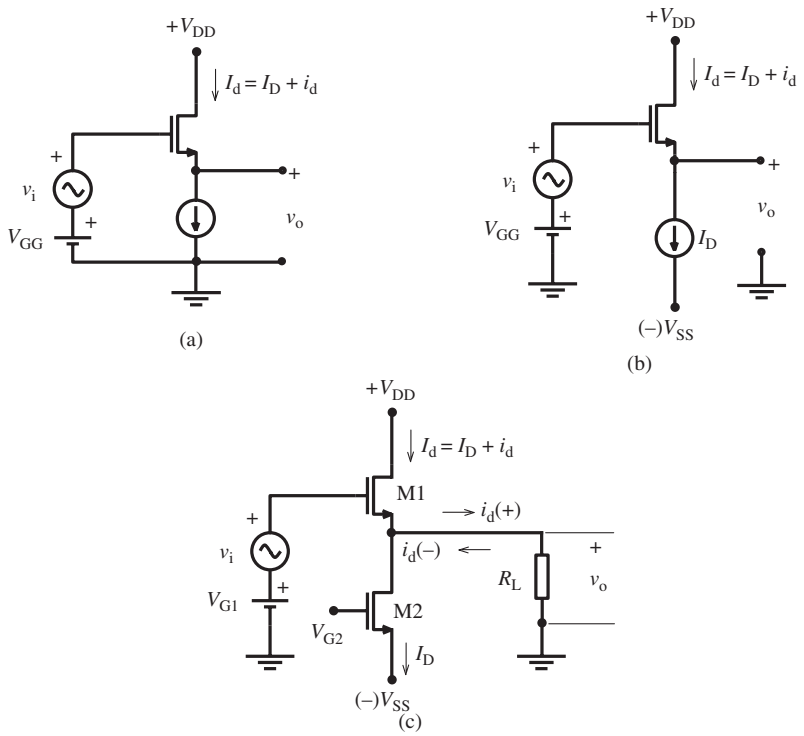


Figure 2.15 (a) The source follower with a current source load. (b) Using twin power supplies for zero output DC voltage. (c) Using a MOS transistor as current source.

For a positive signal voltage the drain current of M1 increases by $i_d(+)$. This difference current flows through R_L, and the output voltage increases. For $\hat{v}_o = |\check{v}_o|$, the maximum value of $i_d(+)$ must be equal to $2I_S$. For this case the drain current of M1 becomes equal to $2I_S$. To maintain M1 in the saturation region,

$$(V_{DD} - v_o) \geq (\hat{V}_{GS1} - V_T) \tag{2.27}$$

must be satisfied, where $\hat{V}_{GS1}$ is the gate–source voltage of M1 corresponding to $I_{D1} = 2I_S$.

Example 2.4 Design a source follower as shown in Fig. 2.16(a) for AMS 0.35 micron technology. The design goals are zero output quiescent voltage and minimum ± 1 V output swing. The channel lengths will be chosen as 0.35 micron for good high-frequency performance. Calculate (a) channel widths and gate bias voltages, (b) the voltage gain, (c) the output internal resistance of the amplifier. (d) With the calculated channel width values perform a PSpice simulation. Fine-tune the bias voltages to obtain the calculated currents necessary for the design goal.

(a) Since the output quiescent voltage is zero and the output swing is ± 1 V, the minimum value of the output voltage is $\check{v}_o = -1$ V. To keep the power consumption

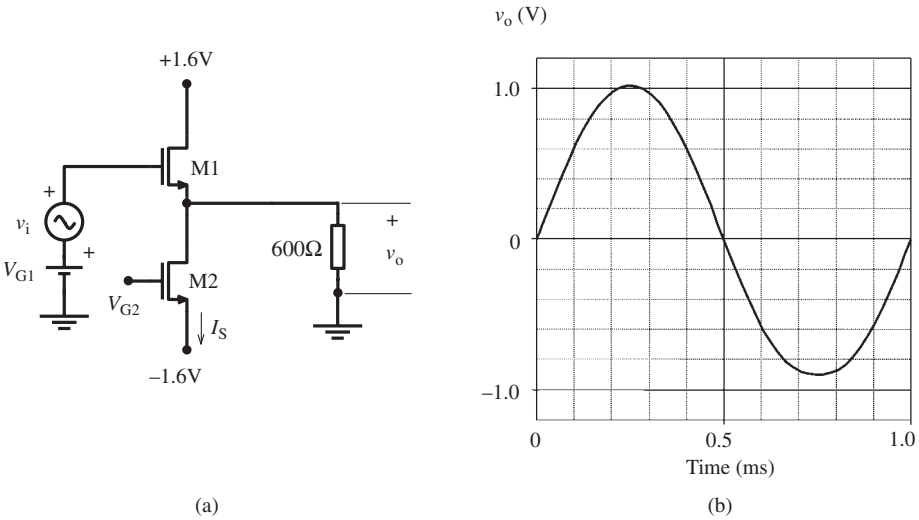


Figure 2.16 (a) The schematic diagram of the source follower to be designed. (b) The PSpice TRAN output for a 1.5 V amplitude, 1 kHz sinusoidal input signal.

minimum, at this extreme point of operation M1 must be driven to cut-off and the full current of the current source (M2) must flow from R_L. Therefore, $\check{v}_o = -I_S R_L$ and $I_S = I_{D2} = 1/600 = 1.67\,\text{mA}$. Under these conditions M2 must fulfill the condition given with (2.26), which is necessary to maintain M2 in the saturation region:

$$(V_{GS2} - V_T) \leq 1.6 - [(1.67 \times 10^{-3}) \times 600] = 0.6\,\text{V}$$

With a 0.1 V safety margin we will prefer to use $(V_{GS2} - V_T) = 0.5\,\text{V}$, since $V_T = 0.5\,\text{V}$, $V_{GS2} = 1\,\text{V}$ and $V_{G2} = -0.6\,\text{V}$. This means that M2 must conduct 1.67 mA for $V_{GS2} = 1\,\text{V}$. From the drain current expression of M2 the aspect ratio can be calculated as

$$\left(\frac{W}{L}\right)_2 = \frac{2 \times I_S}{(KP)_n (V_{GS2} - V_T)^2} = \frac{2 \times (1.67 \times 10^{-3})}{(170 \times 10^{-6}) \times (0.5)^2} \cong 78.5$$

which corresponds to $W_2 = 22.5\,\mu\text{m}$.

At the maximum of the output voltage ($\hat{v}_o = +1\,\text{V}$), a current of 1.67 mA must flow in the opposite direction through the load, R_L. For this extreme case the current of M2 must be equal to the sum of this current and the current of M2, i.e. $\hat{I}_{D1} = 3.34\,\text{mA}$, and M1 must remain in the saturation region according to (2.27):

$$(\hat{V}_{GS1} - V_T) \leq 1.6 - 1 = 0.6\,\text{V}$$

where $\hat{V}_{GS1}$ is the voltage corresponding to $\hat{I}_{D1}$. With a 0.1 V safety margin we will use 0.5 V. Hence

$$\left(\frac{W}{L}\right)_1 = \frac{2 \times \hat{I}_{D1}}{(KP)_n (\hat{V}_{GS1} - V_T)^2} = \frac{2 \times (3.34 \times 10^{-3})}{(170 \times 10^{-6}) \times (0.5)^2} \cong 157$$

which corresponds to $W_1 = 55\,\mu\text{m}$.

Now the bias voltage of M1 under the quiescent condition must be calculated, corresponding to a drain current that is equal to I_S, in order to keep the load current equal to zero and to fulfill the zero output quiescent voltage condition,

$$(V_{GS1} - V_T)^2 = \frac{2 \times (1.67 \times 10^{-3})}{(170 \times 10^{-6}) \times 157} = 0.125$$

which gives the bias voltage of M1 as $V_{G1} = V_{GS1} = 0.85\,\text{V}$.

(b) The voltage gain can be calculated from (2.24a). $\bar{R}$ in this expression is the parallel equivalent of the output resistances of M1 and M2, and the load resistance, that is approximately equal to R_L for this case. Therefore

$$A_v \cong \frac{g_{m1} R_L}{1 + g_{m1} R_L}$$

The transconductance of M1 can be calculated as

$$g_{m1} = \sqrt{2(KP)_n (W/L)_1 I_{D1}} = \sqrt{2 \times (170 \times 10^{-6}) \times 157 \times (1.67 \times 10^{-3})} = 9.4\,\text{mS}$$

Then the voltage gain can be found as:

$$A_v \cong \frac{(9.4 \times 10^{-3}) \times 600}{1 + [(9.4 \times 10^{-3}) \times 600]} = 0.85$$

(c) The output internal resistance from (2.25a):

$$r_o \cong \frac{600}{1 + (9.4 \times 10^{-3}) \times 600} \cong 90\,\text{ohm}$$

(d) The PSpice simulation performed with the calculated values does not initially match the targeted current values. Therefore, a fine-tuning of the bias voltages is necessary, with the procedure used in Example 2.1. The bias voltages for a quiescent current of 1.67 mA are obtained as $V_{G1} = 1.34\,\text{V}$ and $V_{G2} = -0.3\,\text{V}$. Figure 2.16(b) shows the transient simulation result for a 1 kHz sinusoidal input voltage with 1.5 V amplitude. This figure indicates that the positive and negative half periods of the output voltage are not equal, indicating a nonlinearity, which is normal for maximum voltage swing.

Problem 2.2 *A source follower as shown in Fig. 2.16(a) will be used to drive a transmission line having a characteristic impedance of* 600 ohm. *To prevent signal reflections, a transmission line must be terminated with resistors equal to its characteristic impedance, at both ends. The input impedance of a transmission line properly terminated at the output end is equal to the characteristic impedance, which is purely resistive. (a) Design the circuit. (b) Calculate the value of the maximum output swing. (c) Calculate the voltage gain. Improve the design with PSpice.*

2.5 The "long tailed pair"

Probably the most important and most frequently used basic amplifier circuit is the "long tailed pair". Figure 2.17 shows several types of the long tailed pair, along with a short evolution history. All of these circuits permit zero DC quiescent voltages for both inputs, provided that twin DC supplies (V_{DD} and V_{SS}) are used. In the circuits shown in Fig. 2.17, the input transistors are NMOS transistors. It is also possible (and in some cases advantageous) to use PMOS transistors as input transistors. In this case it is obviously necessary to replace the PMOS load transistors with NMOS ones.

The basic structure of the long tailed pair is shown in Fig. 2.17(a). The circuit is symmetrical in nature; M1 and M11 are identical transistors and the load resistors are equal. Under quiescent conditions, M1 and M11 equally share the tail current, I_T, which is supplied by a passive MOS current source, M3.

Instead of the drain resistors, passive PMOS transistors (M2, M12) biased in their saturation regions can be used (Fig. 2.17(b)). It is obvious that, similar to the passive transistor-loaded common-source amplifier, the biasing of the load transistors is critical. To overcome this serious problem, and to fix the DC quiescent voltages of the output nodes, the – so-called – "common-mode feedback, CMFB" is an effective and extensively used technique, and will be investigated later.

In the configuration shown in Fig. 2.17(c), the input transistors are loaded with diode-connected (low-impedance) PMOS transistors (M2, M12) and their currents are mirrored to the load resistors via M4 and M14, to reduce the Miller effect. It is obvious that the mirroring coefficient (current gain) can be higher than unity, which serves to increase the voltage gain. Note that in this circuit both the input and the output quiescent voltages can be set to zero.

In the circuit shown in Fig. 2.17(d), the load resistors are replaced with passive NMOS current sources (M5 and M15). The currents of these sources must be equal to $B.(I_T/2)$, where B is the current gain of the PMOS current mirrors loading the input transistors. It is advantageous to bias M5 and M15 from the same DC voltage source that biases M3. In this case the gate widths of M5 and M15 – in principle – must be equal to $(B/2).W_3$. Since the λ parameters of NMOS and PMOS transistors are – mostly – not equal, to bring the output quiescent voltages to the desired value (to zero for the twin DC power-supply case), usually a fine-tuning on the channel widths of M5 and M15 becomes necessary. It is also important to mention that the sensitivity of the output DC quiescent voltage of this configuration is considerably lower than that of the circuit given in Fig. 2.17(b). Therefore the need for CMFB is not as severe.

If a symmetrical or differential output is not needed, there is an appropriate solution to obtain a single-ended output (Fig. 2.17(e)). This "active loaded long tailed pair" configuration is extensively used as the input stage of operational amplifiers.

All these configurations (except the circuit given in Fig. 2.17(e)) are fully symmetrical in nature, and can be used in different operating modes.

- When a signal is applied to one of its input terminals and the other input is grounded, the circuit can be used as a single-ended input amplifier providing two

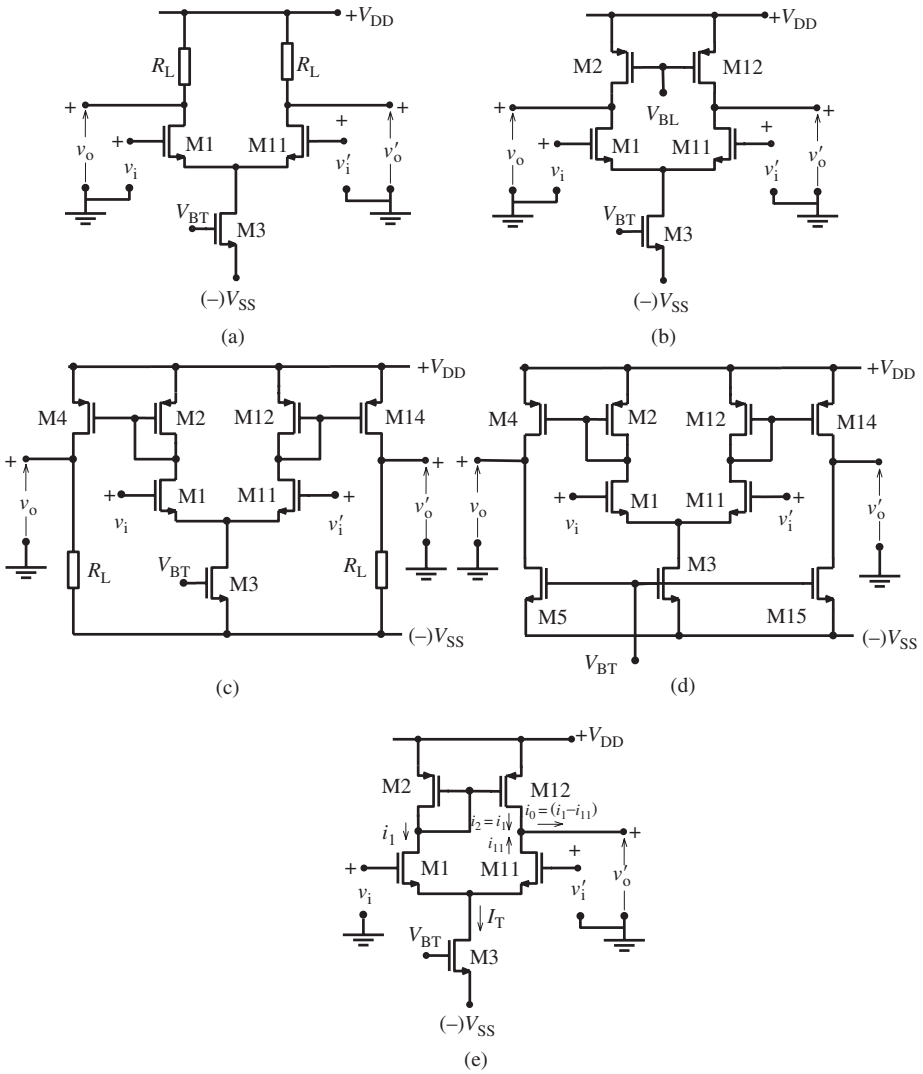


Figure 2.17 Evolution of the basic differential amplifier. (a) Resistance-loaded differential input–differential output long tailed pair. (b) The load resistors replaced with passive PMOS loads. (c) Mirroring the load resistors in (a) to reduce the Miller effect. (d) Replacement of the load resistors in (c) with passive NMOS loads. (e) The differential input–single ended output amplifier.

symmetrical output signals, v_o and v'_o simultaneously, with opposite phase and equal amplitudes, each with respect to ground, as shown in Fig. 2.18(a).

- It is possible to use the difference of these two output signals as a floating differential output voltage: $v_{od} = (v_o - v'_o)$ (Fig. 2.18(b)).
- The long tailed pair can be used as a "difference amplifier" providing a differential (or twin single-ended) output voltage proportional to the difference of two single-ended, independent input signals, $(v_i - v'_i)$ (Fig. 2.18(c)).

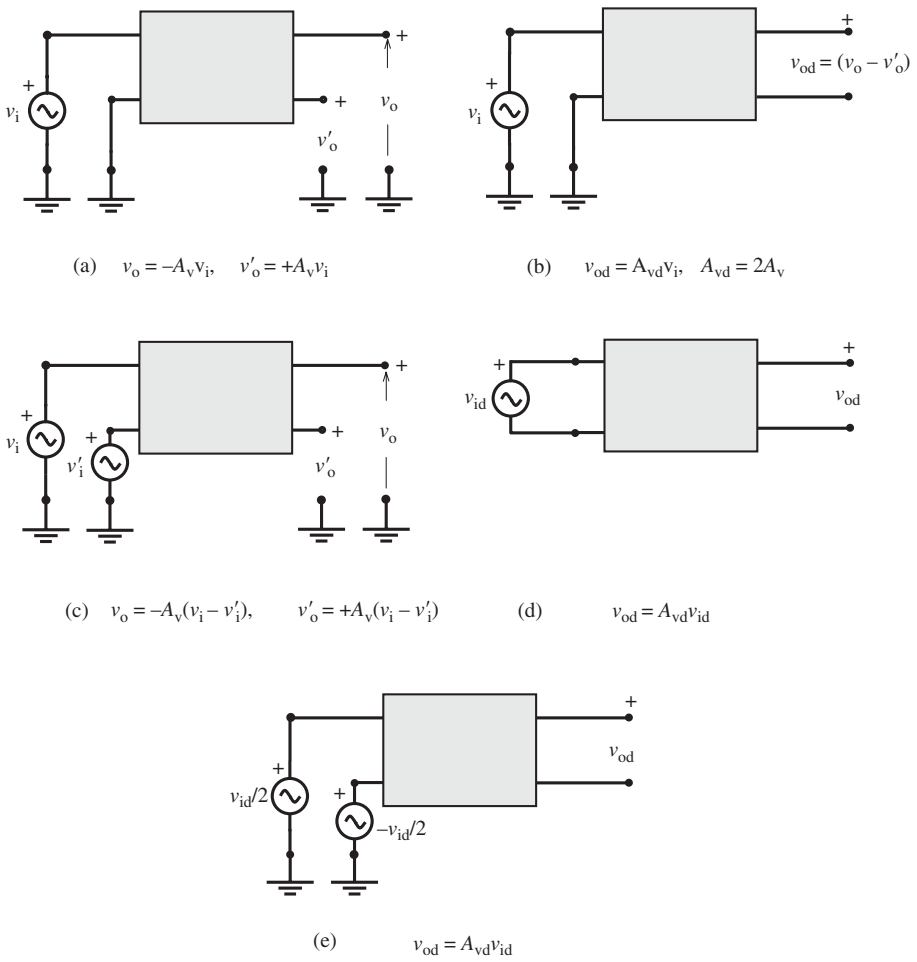


Figure 2.18 (a) Single-ended input, twin single-ended outputs. (b) Single-ended input, differential output. (c) Difference amplification for two independent input signals. (d) Floating differential input. (e) Differential inputs with respect to ground.

- A floating input signal (Fig. 2.18(d)) or an input signal differential with respect to the ground (Fig. 2.18(e)) can be applied between two inputs. For the floating input signal case, to provide DC bias, the gates of M1 and M11 must be connected to ground with two high-value resistors. Since the output signal is also differential, the circuit is called a "fully differential" or "differential input–differential output" amplifier.
- In the case when the output signals are defined as currents, the circuit – by definition – is called a "transconductance amplifier", or "operational transconductance amplifier (OTA)".

To find the gain expressions of the configurations given in Fig. 2.17, it is possible to draw the small-signal equivalent circuits and then calculate the gain.

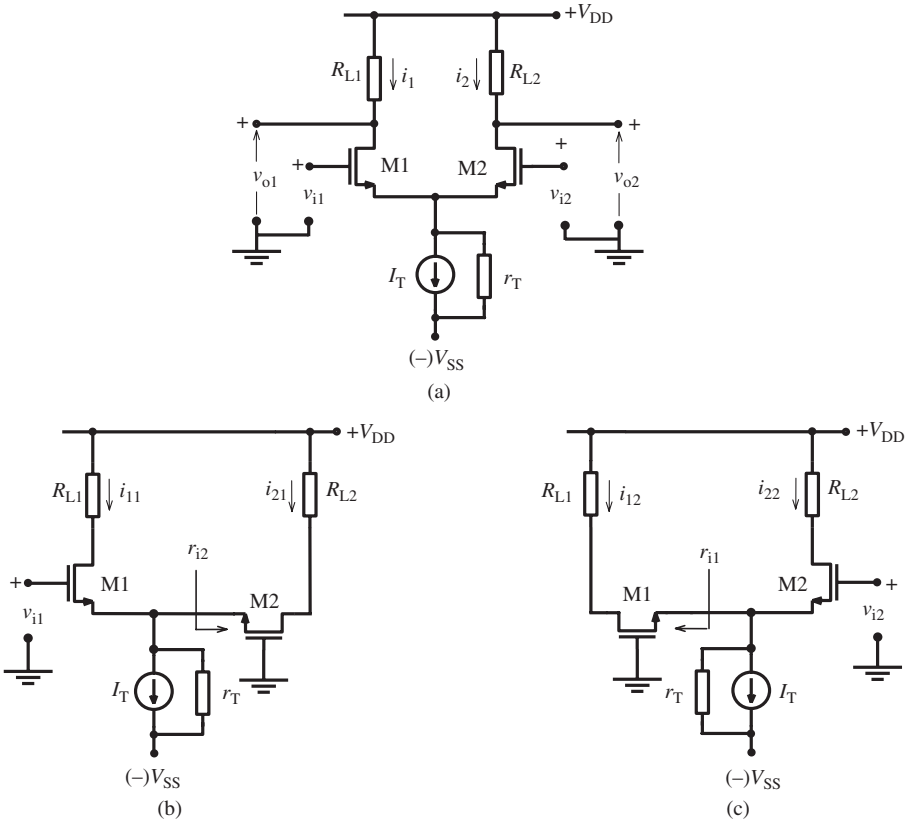


Figure 2.19 Calculation of the signal currents in a long tailed pair via superposition. (a) The basic circuit. (b) Circuit driven from the gate of M1 while the other input is grounded. (c) Circuit driven from the gate of M2 while the other input is grounded.

Here we will prefer a different approach, using the knowledge acquired in previous sections.

In Fig. 2.19(a), the schematic diagram of the basic form of the long tailed pair is given. To enhance the generality of the results, the load resistors of the transistors are chosen with different values as R_{L1} and R_{L2}. The tail current source is represented by an ideal DC current source I_T, and its internal resistance (r_T). Under quiescent conditions, the drain currents of M1 and M2 are equal to $I_T/2$, and, consequently, their transconductances are equal to $g_m = \sqrt{\beta_n I_T}$.

For the case of a signal applied to the gate of M1 where the other input is grounded, the circuit can be re-drawn as shown in Fig 2.19(b), to view the circuit from a different point. We can consider M1 as a transistor with a resistor (r_{i2}) in series with its source terminal. Here, r_{i2} is the small-signal input resistance of M2 operating in the grounded gate configuration,[8] which according to (2.22b) is equal to $1/g_m$. The effective

[8] The internal resistance of the tail current source is always much higher than the input resistance of the grounded gate circuit, therefore it can be neglected.

transconductance of a MOS transistor with a resistor in series with its source terminal (see (2.13c)) was found to be

$$g_{m(eff)} = \frac{g_m}{1 + g_m R_S}$$

In this case, since the source resistance is equal to $1/g_m$, the effective transconductance of M1 is $g_{m(eff)} = g_m/2$. Hence the signal current of M1 that is the current flowing from R_{L1} becomes

$$i_{11} = v_{i1}\frac{g_m}{2} \quad \text{for } v_{i1} \neq 0 \text{ and } v_{i2} = 0 \tag{2.28a}$$

Since the drain current of a grounded-gate circuit is equal to its source current, the current flowing from R_{L2} is

$$i_{21} = -i_{11} = -v_{i1}\frac{g_m}{2} \tag{2.28b}$$

Similarly, for the case of $v_{i2} \neq 0$ and $v_{i1} = 0$ (Fig. 2.19(c)), the currents flowing from R_{L1} and R_{L2} are

$$i_{22} = v_{i2}\frac{g_m}{2} \tag{2.29a}$$

and

$$i_{12} = -i_{22} = -v_{i2}\frac{g_m}{2} \tag{2.29b}$$

Since under small-signal conditions the circuit can be considered as linear, the superposition principle is applicable. Hence, the drain signal current components of M1 and M2 become

$$i_1 = i_{11} + i_{12} = \frac{g_m}{2}(v_{i1} - v_{i2}) \tag{2.30a}$$

and

$$i_2 = i_{21} + i_{22} = -\frac{g_m}{2}(v_{i1} - v_{i2}) \tag{2.30b}$$

These results can be applied to the circuits given in Fig. 2.17, to calculate the voltage gain or the transconductance.

For the resistance-loaded circuit shown in Fig. 2.17(a), the output voltages are

$$v_o = -\bar{R}_L i_1 = -\bar{R}_L \frac{g_m}{2}(v_i - v_i') \tag{2.31a}$$

where $\bar{R}_L$ is the parallel equivalent of R_L and the output resistance of M1.[9] Similarly,

$$v_o' = -\bar{R}_L i_1' = \bar{R}_L \frac{g_m}{2}(v_i - v_i') \tag{2.31b}$$

[9] The load of such an amplifier is usually the gate input of the following stage, which is usually pure capacitive. If there exists a resistive component of the load, certainly it must be taken into account as an additional parallel component to $\bar{R}_L$.

and the differential output voltage,

$$(v_o - v'_o) = -\bar{R}_L g_m (v_i - v'_i) \tag{2.31c}$$

Therefore, the single-ended voltage gains,

$$A_v = \frac{v_o}{(v_i - v'_i)} = -\frac{1}{2} g_m \bar{R}_L \tag{2.32a}$$

$$A'_v = \frac{v'_o}{(v_i - v'_i)} = \frac{1}{2} g_m \bar{R}_L \tag{2.32b}$$

and the differential voltage gain

$$A_{vd} = \frac{(v_o - v'_o)}{(v_i - v'_i)} = -g_m \bar{R}_L \tag{2.32c}$$

These expressions can be easily applied to the configurations given in Fig. 2.18 and interpreted accordingly.

For the passive transistor-loaded long tailed pair shown in Fig. 2.17(b), the gain expressions can be readily obtained by replacing R_L by the output resistance of M2.

In the circuit shown in Fig. 2.17(c), the currents of M1 and M11 are mirrored to the load resistors via M2–M4 and M12–M14 current mirrors. If the mirroring coefficients of these mirrors are B, where B can be unity or larger, the gain expressions become

$$A_v = \frac{v_o}{(v_i - v'_i)} = -\frac{1}{2} g_m B \bar{R}_L \tag{2.33a}$$

$$A'_v = \frac{v_o}{(v_i - v'_i)} = \frac{1}{2} g_m B \bar{R}_L \tag{2.33b}$$

$$A_{vd} = \frac{(v_o - v'_o)}{(v_i - v'_i)} = -g_m B \bar{R}_L \tag{2.33c}$$

For the voltage gain expressions of the circuit shown in Fig. 2.17(d), $\bar{R}_L$ must be replaced by the output resistances of the passive load transistors M5 and M15. This circuit is more suitable to use as an OTA. Since under the quiescent conditions the currents of M4 and M5 (similarly M14 and M15) are equal, the output signal currents i_o and i'_o are zero. For v_i and v'_i input drives, the output currents are

$$i_o = -B_{i1} = \frac{1}{2} g_m B (v_i - v'_i) \tag{2.34a}$$

$$i'_o = -B'_{i1} = -\frac{1}{2} g_m B (v_i - v'_i) \tag{2.34b}$$

Therefore the differential transconductance is calculated as

$$G_{md} = \frac{(i_o - i'_o)}{(v_i - v'_i)} = g_m B \tag{2.35}$$

The output internal resistances of this OTA are obviously the parallel equivalent of the output resistances of M4 and M5, and M14 and M15, respectively.

For the single-ended circuit given in Fig. 2.17(e), the signal current of M1 is mirrored to the output of M11. Since the current of M11 is equal in magnitude, but opposite in phase, the total signal current flowing through the load is equal to $2i_1$. Hence, the output voltage and the voltage gain become

$$v'_o = 2i_1 R'_L = g_m R'_L (v_i - v'_i) \tag{2.36a}$$

$$A_v = \frac{v'_o}{(v_i - v'_i)} = g_m R'_L \tag{2.36b}$$

Example 2.5 A fully symmetrical OTA as shown in Fig. 2.17(d) will be designed for AMS 0.35 micron technology. The target value of the transconductance is 3 mS. For a good high-frequency performance, the channel lengths of the active transistors, the quiescent current of the circuit and the B factor are chosen as 0.35 μm, 4 mA and 1, respectively. The design will be made with hand calculations, and then fine-tuned with PSpice simulations.

Since $B = 1$, the drain DC current of the input transistors is equal to 1 mA, and the transconductance of the input transistors must be 3 mS. Then, the β coefficient and the aspect ratio of M1 and M11 can be calculated from $g_m = \sqrt{2\beta I_D}$ as

$$\beta_1 = \frac{g_{m1}^2}{2I_{D1}} = \frac{(3 \times 10^{-3})^2}{2 \times (10^{-3})} = 4.5 \times 10^{-3} \ [\mathrm{A/V^2}]$$

and

$$\left(\frac{W}{L}\right)_1 = \frac{\beta_1}{(KP)_n} = \frac{4.5 \times 10^{-3}}{170 \times 10^{-6}} = 26.5 \ \rightarrow \ W_1 = 9.3\,\mu\mathrm{m}$$

From the drain current expression,

$$(V_{GS1} - V_{Tn})^2 \cong \frac{2I_{D1}}{(KP)_n (W/L)_1} = \frac{2 \times (10^{-3})}{(170 \times 10^{-6}) \times 26.5} = 0.44 \quad (V_{GS1} - V_{TN}) = 0.67\,\mathrm{V}.$$

The gate–source bias voltage of M1 (and M11) is found to be $V_{GS1} = 1.17$ V. Since the quiescent voltage of both inputs is equal to zero, the DC voltage of the source of M1 and M11 is $V_{S1} = -1.17$ V.

We know that all transistors in this circuit must operate in the saturation region, even under worst-case conditions. The worst-case condition for M1 and M2 occurs when the full tail current (2 mA in our case) flows over this branch, under a high positive drive on the input of M1. For $\hat{I}_{D1} = I_T = 2$ mA, the gate drive can be found as $(\hat{V}_{GS1} - V_{Tn}) = 0.94$ V. Therefore, to maintain the saturation condition of M1, the condition of $V_{DS1} \geq (\hat{V}_{GS1} - V_{Tn})$ must be satisfied. With a safety margin, we choose $V_{DS1} = 1$ V under maximum drive of M1. The voltage of the drain nodes of M1

and M2 can be found to be $V_{D1} = V_{DS1} - V_{S1} = -0.17\,\text{V}$. Then the gate–source voltage is –1.77 V and the drain current of M2 with this gate drive must be 2 mA. From these considerations, the aspect ratio and the gate width of M2 (and M12) can be calculated as

$$\left(\frac{W}{L}\right)_2 \cong \frac{2|I_{D2}|}{(KP)_P(V_{GS2} - V_{TP})^2} = \frac{2 \times (2 \times 10^{-3})}{(58 \times 10^{-6}) \times [-1.77 - (-0.7)]} = 60.5 \;\rightarrow\; W_2 = 21.2\,\mu\text{m}.$$

To obtain a high internal resistance, the channel length of the tail current source (M3) will be chosen as 1 μm. The aspect ratio and the channel width of M3 can be calculated such that the transistor is in the saturation region for $I_{D3} = I_T = 2\,\text{mA}$. The drain–source voltage of M3 is $V_{DS3} = V_{S1} - V_{SS} = -1.17 - (-1.6) = 0.43\,\text{V}$. Therefore $V_{DS3} \geq (G_{GS3} - V_{TN})$ must be satisfied. With a safety margin, we can choose $V_{GS3} - V_{TN} = 0.4\,\text{V}$, therefore $V_{GS3} = 0.9\,\text{V}$ and $V_{G3} = -0.7\,\text{V}$. Now it is possible to find the aspect ratio of M3.

$$\left(\frac{W}{L}\right)_3 = \frac{2 \times (2 \times 10^{-3})}{(170 \times 10^{-6}) \times (0.4)^2} = 147 \;\rightarrow\; W_3 = 51.5\,\mu\text{m}.$$

Since $B = 1$, the dimensions of M4 (and M14) are the same as M2.[10] Note that M5 and M15 are not on the signal path. Therefore it is possible to use a longer channel length value, for example 1 μm as in the tail transistor. The channel widths can be calculated for $I_{D5} = 1$ mA, $V_{DS5} = 1.6\,\text{V}$ and $V_{G5} = V_{G3} = -0.7\,\text{V}$ as $W_5 = 73.5\,\mu\text{m}$.

As in the previous examples, simulations performed with the calculated values do not initially match the targeted drain currents, but they provide a reasonable starting point for iterations. To maintain the saturation conditions of transistors, it is wise to keep the calculated bias voltages and fine-tune the widths of the transistors for targeted currents. The PSpice netlist obtained after the fine-tuning procedure is given below. The circuit diagram and the input voltage to output current transfer curves are shown in Fig. 2.20.

```
*DESIGN:OTA-AMS 035*
.lib "cmos7tm.mod"
VDD 100 0 1.6
VSS 200 0 -1.6
M1 6 1 3 200 modn L=.35U W=24U ad=20.5e-12 as=20.5e-12 Pd=48u
Ps=48u
```

[10] If the drain–source voltages of M2 and M4 are not equal as in our example owing to the channel shortening effect, the drain currents are not exactly equal. This only acts on the value of B, and consequently the final value of the transconductance. Certainly it is possible to fine-tune the width of M4 and M14 during simulations.

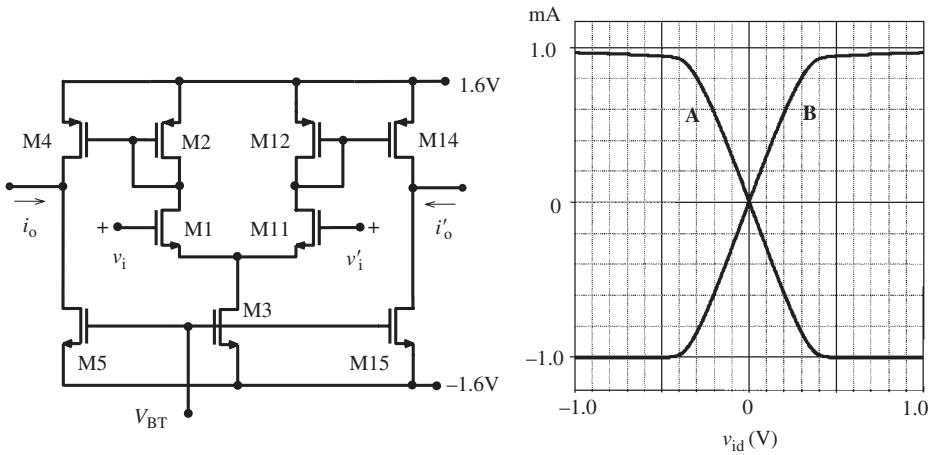


Figure 2.20 (a) The circuit diagram and dimensions of transistors. (b) The input voltage (v_{id}) to the short-circuit output current transfer curve of the OTA; Curve A: I(R7), Curve B: I(17). The value of the transconductance is 3.016 mS.

```
  M11 5 2 3 200 modn L=.35U W=24U ad=20.5e-12 as=20.5e-12
Pd=48u Ps=48u
  M2 6 6 100 100 modp L=.35U W=50U ad=42.5e-12 as=42.5e-12
PD=100U PS=100U
  M12 5 5 100 100 modp L=.35U W=50U ad=42.5e-12 as=42.5e-12
PD=100U PS=100U
  M3 3 4 200 200 modn L=1U W=230U ad=195.5e-12 as=195.5e-12
PD=460U PS=460U
  M4 7 5 100 100 modp L=.35U W=48U ad=41e-12 as=41e-12 PD=96U
PS=96U
  M5 7 4 200 200 modn L=1U W=111U ad=92.4e-12 as=92.4e-12
PD=222U PS=222U
  M14 17 6 100 100 modp L=.35U W=48U ad=41e-12 as=41e-12
PD=96U PS=96U
  M15 17 4 200 200 modn L=1U W=111U ad=92.4e-12 as=92.4e-12
PD=222U PS=222U
  R7 7 0 1
  R17 17 0 1
  vtb5 4 0 -0.7
  VIN1 1 0 dc 0 ac 1m
  e2 2 0 1 0 -1
  .DC VIN1 -1 1 10M
  .AC DEC 20 10MEG 10G
  .PROBE
  .END
```

Problem 2.3 *It is intended to increase the transconductance of the circuit designed in Example 2.5 to* 10 mS. *One of the solutions is to increase the B factors of the load current mirrors. Modify the design in this way.*

Problem 2.4 *To improve the linearity of the circuit designed in Problem 2.3, an appropriate resistor can be connected in series to the source terminals of the input transistors, at the price of the reduction of the transconductances. Modify the design in this way to reduce the total transconductance of the circuit to* 3 mS. *Compare the transfer curves with that of Example 2.5.*

Problem 2.5 *Repeat the design given in Example 2.5 with PMOS input transistors. Discuss the advantages and disadvantages of this new design.*

2.5.1 The large signal behavior of the long tailed pair

From Fig. 2.20(b) we see that the linear dependence between the input and the output voltages is limited to a certain region around the quiescent point. For larger differential input voltage swings, first the linearity of the output signal variation deteriorates, then the control ability of the input signal completely disappears. Obviously, the calculated small-signal behavior of the amplifier is valid only in the region where the output signal is a linear function of the input signal. For example, it can be seen from Fig. 2.20(b) that for the circuit designed above, the linear operation range is approximately $\pm 0.3\,\mathrm{V}$ around the quiescent point. For larger input signal amplitudes, nonlinear distortion and even clipping on the output signal is unavoidable.

In addition to this observation, Fig. 2.20(b) shows that for a sufficiently large differential input signal (approximately $\pm 0.4\,\mathrm{V}$ for our example) the output current switches – almost completely – from one output to another. This is important information indicating that a long tailed pair can be used as a switching circuit.

Since all differential circuits based on the long tail structure exhibit the same basic properties, we will investigate the large signal behavior of the basic long tailed pair shown in Fig. 2.21(a). M1 and M2 are biased in their saturation region and equally share the tail current under quiescent condition (i.e. $v_{\mathrm{id}} = 0$). The drain currents can be written in terms of the gate–source voltages as

$$I_1 \cong \frac{1}{2}\beta(V_{\mathrm{gs1}} - V_{\mathrm{T}})^2 \qquad I_2 \cong \frac{1}{2}\beta(V_{\mathrm{gs2}} - V_{\mathrm{T}})^2$$

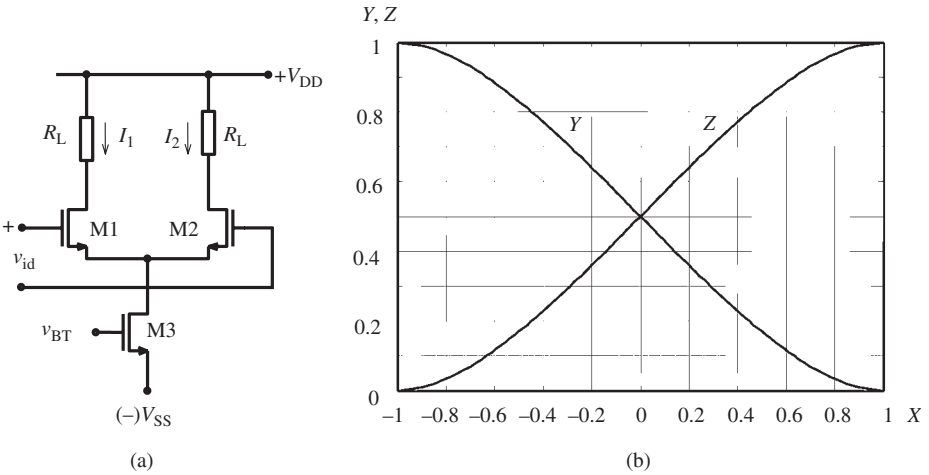


Figure 2.21 (a) The basic long tailed pair. (b) The normalized drain currents of M1 and M2 as a function of the differential input voltage. ($X = v_{id}/\sqrt{2I_T/\beta}$, $Y = I_1/I_T$, $Z = 1 - Y$)

then

$$(V_{gs1} - V_T) \cong \sqrt{\frac{2I_1}{\beta}}, \qquad (V_{gs2} - V_T) \cong \sqrt{\frac{2I_2}{\beta}}$$

The differential input voltage:

$$v_{id} = V_{g1} - V_{g2} = V_{gs1} - V_{gs2} = \sqrt{\frac{2I_1}{\beta}} - \sqrt{\frac{2I_2}{\beta}} \tag{2.37a}$$

with $I_2 = I_T - I_1$:

$$\frac{v_{id}}{\sqrt{2I_T/\beta}} = \sqrt{\frac{I_1}{I_T}} - \sqrt{1 - \frac{I_1}{I_T}} \tag{2.37b}$$

Using this normalized expression, the variations of I_1 and I_2 as a function of the differential input voltage v_{id} can be plotted as shown in Fig. 2.21(b). As expected, for $v_{id} = 0$, $I_1 = I_2 = (I_T/2)$ monotonically increases with v_{id} and I_2 decreases, such that the sum of these currents remains equal to I_T. For

$$v_{id} = \sqrt{2I_T/\beta} \tag{2.38}$$

the tail current is completely switched to M1 and the drain current of M2 becomes zero. Similarly, for $v_{id} = -\sqrt{2I_T/\beta}$ the drain current of M2 becomes equal to I_T and I_1 reduces to zero. This means that the circuit can be used as a two-position switch, to toggle a current (the tail current) between two branches. According to (2.38), to switch a certain current, the input voltage switching interval is narrower for higher β values.

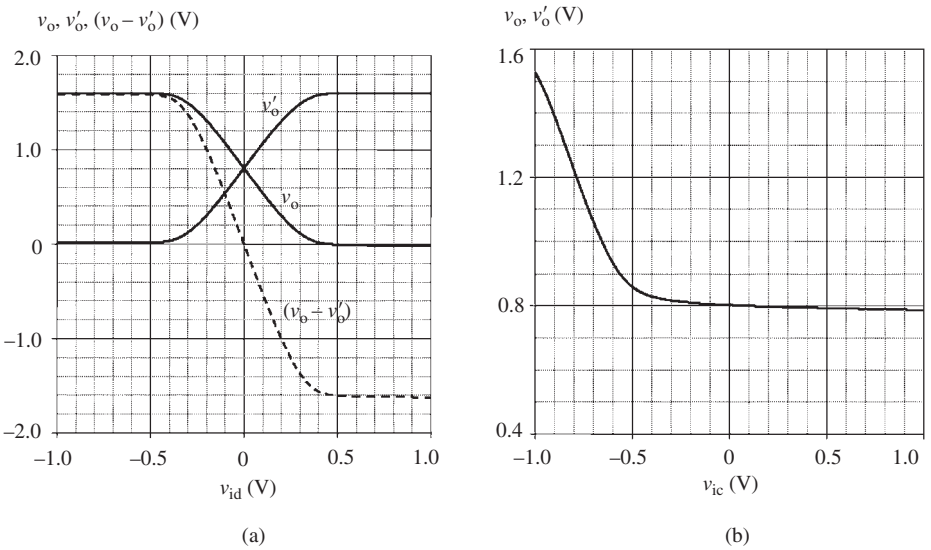


Figure 2.22 (a) Variations of the single-ended and differential output voltages of a resistance-loaded long tailed pair for DC sweep of the differential input voltage (input transistors: 35 μm/0.35 μm, the tail transistor: 90 μm/1 μm, the tail current: 0.8 mA, load resistors: 2 k ohm). (b) Variations of the single-ended output voltages for common-mode input voltage.

The simulation results of the single-ended and differential output voltages of a resistance-loaded long tailed pair are given in Fig. 2.22(a), as a function of the differential input voltage. The linear control regions of the transfer characteristics and the switching property of the circuit are clearly shown. The small-signal voltage gain can be derived as the slope of the transfer curve at the quiescent point.

One of the important aspects of a long tailed pair is its behavior when the voltages applied to the inputs are identical, i.e. when the inputs are driven by a "common-mode signal". According to (2.31a) and (2.31b), under common-mode input signal conditions (i.e. $v_i = v'_i$), the signal components of the output voltages must be zero. In other words, the voltages of output nodes must remain constant. But in reality, as shown by the simulation results in Fig. 2.22(b), v_o and v' vary with the common-mode input signal, and their variations are identical. As seen from this figure, there are two different regions of these curves. In the vicinity of the quiescent operating point (under small-signal conditions) the variations are small. But for large negative input voltages the output voltages decrease sharply, which is owing to the transition of the tail transistor to the pre-saturation (resistive) region.

The rate of change of the single-ended output voltage with respect to the common-mode input signal under small-signal conditions is called the "common-mode gain" of the amplifier and can be calculated from the small-signal equivalent circuit shown in Fig. 2.23(b), as the ratio of one of the output voltages to the common input voltage:

$$A_{vc} = \frac{v_{o2}}{v_{ic}} = -g_m R_L \frac{1}{(1 + 2r_{oT}g_m) + \frac{2r_{oT}+R_L}{r_{o2}}} \approx -\frac{R_L}{2r_{oT}} \tag{2.39}$$

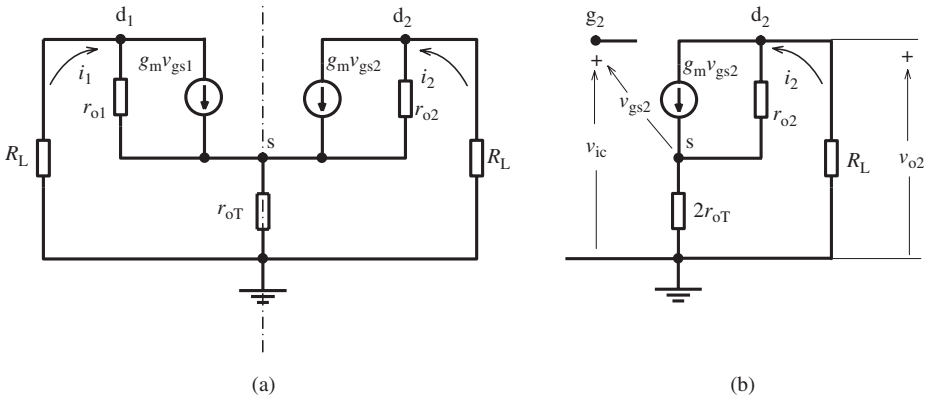


Figure 2.23 (a) The small-signal equivalent circuit of a resistance-loaded long tailed pair (r_{oT} represents the internal resistance of the tail current source). (b) The half-circuit obtained by cutting the circuit from its symmetry axis for common-mode drive.

It must be noted that for common-mode input signals, the difference between the output signals is zero and consequently, the "common-mode input to differential output voltage gain" is zero, provided that the symmetry of the circuit is perfect.

For a mixed input signal that has differential as well as common-mode components, the differential and common-mode gains can be independently calculated using the superposition principle since the circuit is assumed to be linear under small-signal conditions. The ratio of the "differential input to differential output voltage gain" to the "common-mode voltage gain" is defined as the "common-mode rejection ratio (CMRR)":

$$CMRR = \frac{A_{vd}}{A_{vc}} \approx 2g_m r_{oT} \tag{2.40}$$

Equation (2.40) indicates that for a high CMRR, the small-signal internal resistance of the tail current source must be as high as possible. Especially for small geometry technologies (small supply-voltage values) it is not convenient to use complicated high internal resistance current sources (for example cascode circuits) that contain more than one transistor in series. Since the tail current source is not in the signal path, it is possible to use long channel devices that provide higher internal resistance values. But since the channel widths also must be proportionally high, the parasitic capacitances are higher, which leads to lower internal impedances at high frequencies that deteriorate the CMRR at high frequencies, accordingly.

Problem 2.6 *Calculate the transistor channel widths of a long tailed pair that will be used to switch a* 1 mA *current, with a total input switching range of* 200 mV. *The parameters of this* 0.18 µm *technology are given as* $V_T = 0.4$ V, $\mu_n = 300$ cm^2/V.s *and* $C_{ox} = 8 \times 10^{-7}$ F/cm^2.

Problem 2.7 *Derive (2.39).*

2.5.2 Common-mode feedback

We have seen that the DC components of the output signals of a differential amplifier may be different from the targeted value. This difference can arise from the biasing of the passive load transistors, which is critical, or can be the result of the non-zero common-mode gain under common-mode input signals. To eliminate, or at least to reduce, the adverse effects of these unwanted DC components that occur equally on both outputs, a technique called "common-mode feedback" (CMFB) is a useful tool.[11] The common-mode feedback must be arranged in such a way that it has no effect on the differential gain of the amplifier, but forms an effective negative feedback to reduce the common-mode gain and to control the DC components that are common for both outputs.

There are several CMFB circuits shown in the literature [24], [25]. The generally used approach for the common-mode feedback is summarized in Fig. 2.24. The output signals having differential signal components and DC common-mode components[12] are shown in Fig. 2.24(a). Assuming that the symmetry of the circuit is perfect, the difference of these signals, which is the differential output signal of the amplifier, has no common-mode component, as shown in Fig. 2.24(c). The sum of the output signals is shown in Fig. 2.24(b). Since the differential signal components are in opposite phase, they cancel each other. The result is the sum of the common-mode components that is compared with a reference, and the resulting error signal is fed back to an appropriate node of the circuit in correct magnitude and phase, to eliminate – or reduce – the error (Fig. 2.24(d)).

If the output signals of the amplifier are voltages (as in a differential voltage amplifier) the feedback circuit must be designed to be effective on the common-mode voltages for an appropriate control of the output common-mode component. But for OTAs, where the output signals of the circuit are currents by definition, the CMFB circuit must be designed to minimize the common-mode components of the output currents.

It must be noted that, although the CMFB is effective at DC and at the low end of the frequency range, owing to the additional load it impairs the high-frequency performance of the amplifier.

Example 2.6 As a practical example, a passive PMOS transistor-loaded long tailed pair having a non-ideal tail current source is shown in Fig. 2.25(a). The circuit was designed for 0.45 V DC level at the output nodes. The bias voltage of the load

[11] Owing to the tolerances of the devices, the symmetry of the circuit can be not perfect. For this case the CMFB is only partially effective.

[12] The common-mode signal does not have to be DC, but it must be within the bandwidth of the feedback circuit.

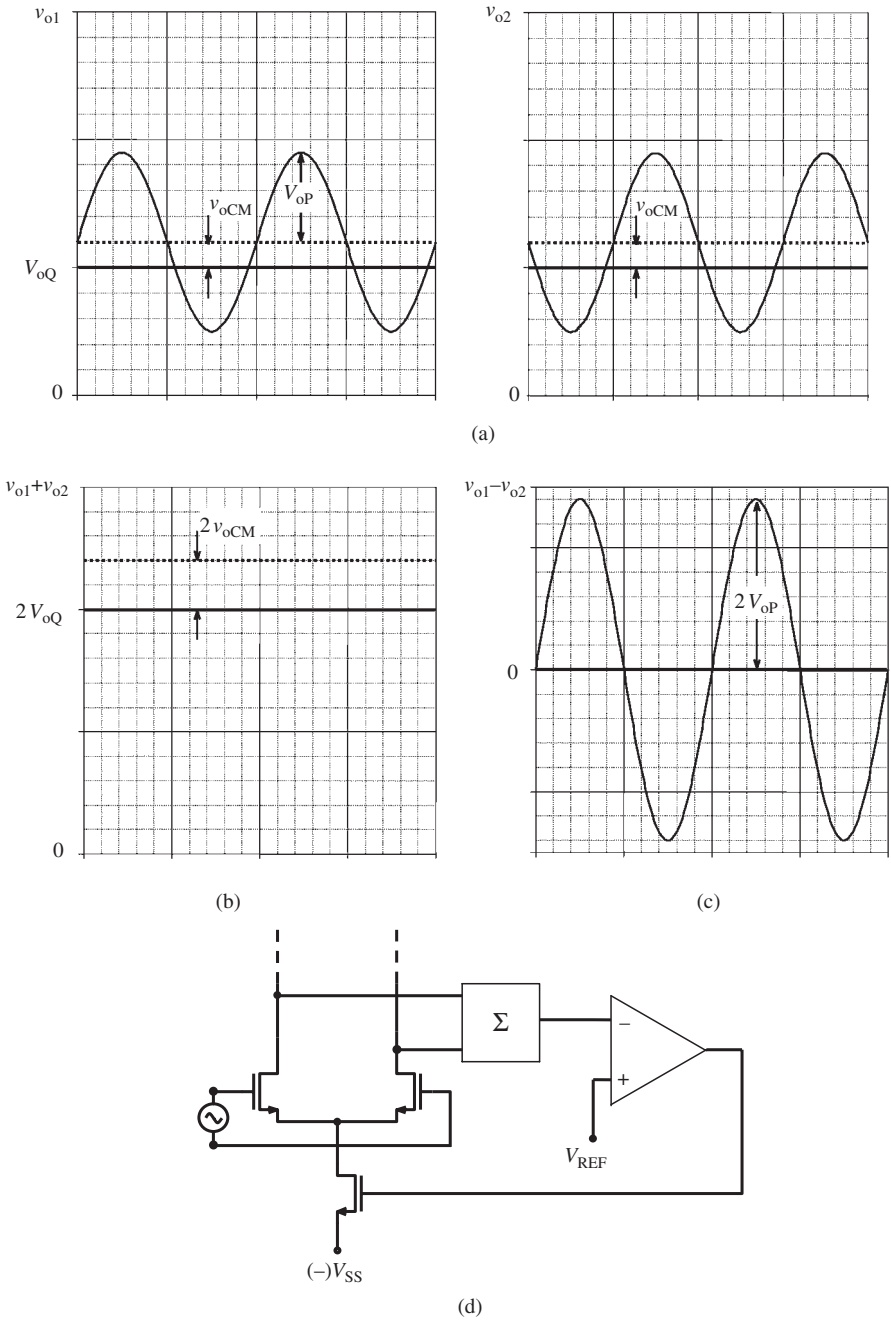


Figure 2.24 (a) The voltages of the output nodes of a differential voltage amplifier. V_{oQ} and v_{oCM} represent the quiescent DC voltage and the CM component of the output voltage, respectively. (b) The sum and (c) the difference of the output voltages. (d) Schematic of the CMFB circuit.

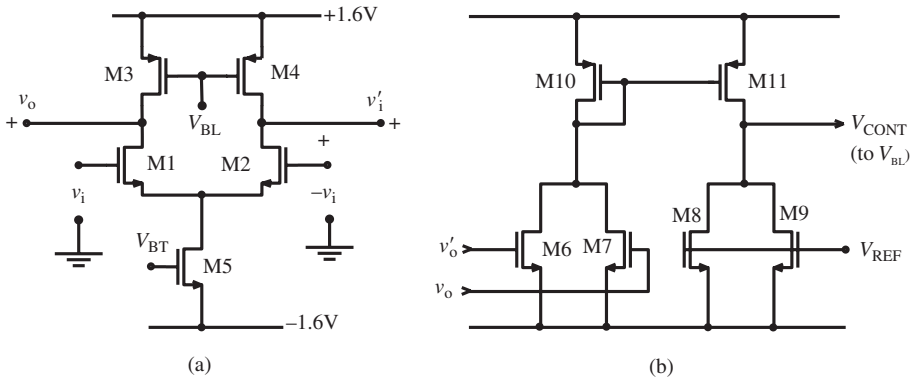


Figure 2.25 (a) A passive transistor-loaded long tailed pair. Transistors are 0.35 μ AMS transistors. All channel lengths are 0.35 μm. Channel widths: for the input transistor 35 μm, for the load transistors 60 μm and the tail transistor 10 μm. (b) The simple voltage CMFB circuit. M6 to M9 are the same with $L = 0.35\,\mu m$ and $W = 5\,\mu m$. To increase the gain of the current differencing amplifier long channel-length devices are used as M10 and M11 (300 μm/2 μm).

transistors was tuned as $V_{BL} = 0.674\,V$ to obtain this quiescent voltage. Fig. 2.26(a) shows the very high sensitivity of the output quiescent voltages with respect to this bias voltage, and it can be understood that this circuit is very sensitive to the variations of the bias voltage, as well as to the parameter tolerances of the transistors, and consequently, is not a very robust circuit as it is.

A simple CMFB circuit to control the output common-mode voltages is given in Fig. 2.25(b). M6 and M7 in this circuit measure the voltages of the output nodes and produce quadratically related drain currents. These currents are summed on M10 and produce a gate–source voltage that is an inverse quadratic function of the current. This voltage is not a perfectly linear function of the output voltage but there is an inherent nonlinearity compensation to some extent. M8 and M9 are the same as M6 and M7, and biased with a reference voltage (V_{REF}). When the DC voltages of the output nodes are equal to V_{REF}, the drain voltage of M11 is equal to V_{G10}, which has the appropriate value to bias the load transistors. M10 and M11 work together as a current differencing amplifier[13] and produce the control voltage (V_{CONT}). If a CM voltage occurs on the output nodes, an error voltage component approximately proportional to the difference of the drain currents of M10 and M11 is produced and controls the load transistors to compensate the CM signal. From another point of view, this circuit compares the output common-mode voltage with the reference voltage and forces the circuit to reduce the error.

To apply CMFB to the amplifier shown in Fig. 2.25(a), the outputs of the amplifier must be connected to the gates of M6 and M7, and the control voltage output of the CMFB circuit must be replaced by the bias voltage source of the load transistors.

[13] This circuit is inspired by the Norton Amplifier structure used in bipolar ICs in earlier years [25].

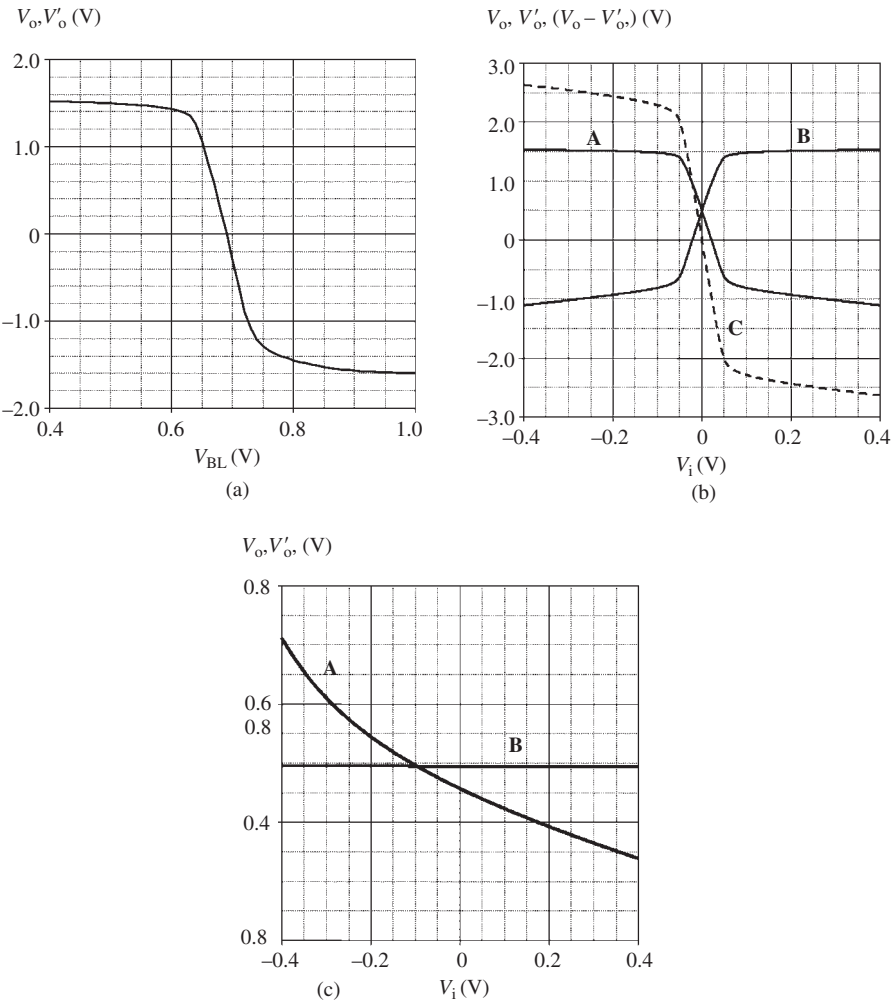


Figure 2.26 (a) The variation of the single-ended output voltages of the long tailed pair shown in Fig. 2.25(a), as a function of the bias voltage of the load transistors. (b) The single-ended (A and B) and differential (C) output voltages of the CMFB applied amplifier for a differential drive. (c) The voltages of the output nodes as a function of the common-mode input signal, (A) without CMFB, (B) with CMFB.

The single-ended and differential output voltages of this combination are shown in Fig. 2.26(b) under a differential input drive ($v_{i2} = - v_{i1}$). In Fig. 2.26(c), the output voltages under a common-mode signal ($v_{i2} = v_{i1}$), without (A) and with (B) common-mode feedback are shown. It can be clearly seen that the CMFB effectively reduces the common-mode gain and clamps the output DC levels to the reference voltage. The small and relatively harmless difference between the single-ended output voltages and the reference voltage is the result of the finite-loop gain of the feedback.

Example 2.7 One of the important classes of differential amplifiers is the fully differential OTA, which is extensively used in the gyrator-based active filters. Since the output signals of OTAs are currents by definition, the common-mode feedback signal must be sampled from the sum of the output currents and compared to the targeted quiescent value (i.e. zero in most cases). Then, the produced error voltage must be applied to an appropriate node, in the correct phase. The output currents must be sampled without affecting the load currents. To solve this problem, in the circuit shown in Fig. 2.27(a), the output transistors (M6, M7 and M16, M17) are duplicated (M26, M27 and M36, M37) and the replicated second set of output currents are used as the input signals of the feedback circuit shown in Fig. 2.27(b). In this circuit the replicated output currents are summed on a resistor (R51) and produce a feedback signal proportional to the sum of the output currents and obviously independent from the actual load of the amplifier (which can be nonlinear and/or reactive). The rest of the circuit is a simple differential amplifier, comparing the feedback signal with the target quiescent voltage and producing a suitable bias voltage for M7, M17, M27 and M37.

As an example for this current-mode CMFB (CCMFB), the OTA shown in Fig. 2.27(a) was designed. The tail current was chosen as 100 μA. The target transconductance value was 1 mS. In order to improve the linearity of the amplifier, series source resistances were used, as shown in Fig. 2.28(a). The core OTA was intentionally not well optimized to obtain zero output offset, such that the influence of CMFB can be observed more clearly. The results are as follows:

- The output offset currents without CCMFB : 11.27 μA
  with CCMFB : 1.115 μA

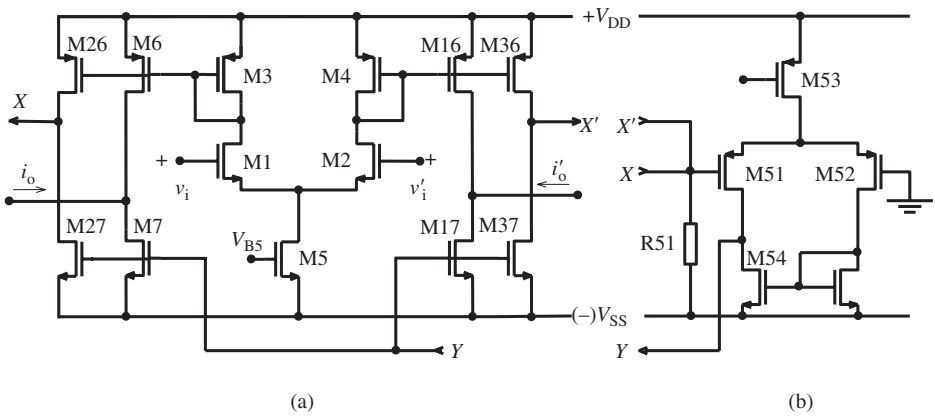


Figure 2.27 (a) The schematic of a fully symmetric OTA that has separate replicated current outputs for feedback. (b) The simple current CMFB circuit used in this example.

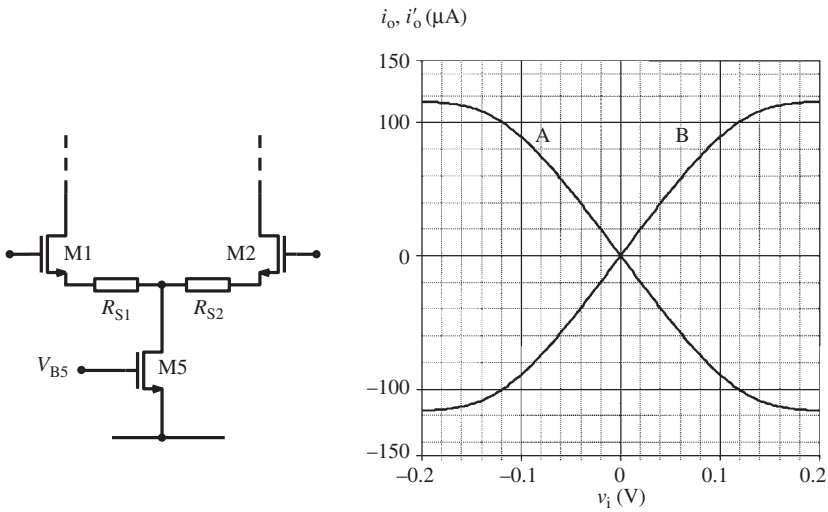


Figure 2.28 (a) The source series resistances of the input transistors that decrease the effective transconductance but improve linearity as mentioned in Section 2.1. (b) The DC transfer curves of the OTA with the CMFB circuit given in Fig. 2.27. The input transistors have series source resistances. The dimensions of the transistor are as follows: M1, M2: 50 μm/0.35 μm; M3, M4, M6, M16, M26: 50 μm/0.35 μm; M5: 23 μm/1 μm; M7, M17, M27, M37: 13 μm/1 μm; M51, M52: 10 μm/1 μm; M54, M55: 2 μm/0.35 μm; M53: 40 μm/1 μm; $R_{S1} = R_{S2} = 385$ ohm, $R_{51} = 1$ k ohm. Bias voltages of M7, M17, M27, M37 are –0.7 V.

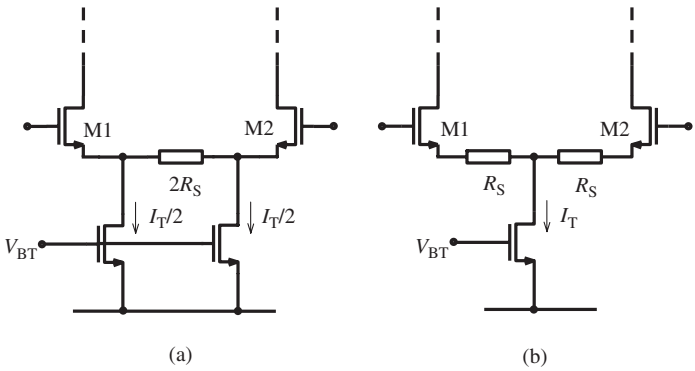
$$\text{- The common-gain} \quad \begin{matrix} \text{without CCMFB} : 3.40\,\mu\text{S}\ (\text{CMRR} = 49.4\,\text{dB}) \\ \text{with CCMFB} : 0.13\,\mu\text{S}\ (\text{CMRR} = 77.7\,\text{dB}). \end{matrix}$$

The improvements on the values of the output offset current, the CMRR, and the shape of the simulated transfer curve given in Fig. 2.28(b) show the effectiveness of the common-mode feedback.

Problem 2.8 *Discuss how we can improve the linearity and the CMRR of the OTA even further in the example, without (and with) increasing the current consumption.*

Problem 2.9

(a) *An alternative solution to improve the linearity of a long tailed pair is given in the figure, part (a). Show that this is equivalent to the classical approach given in the figure part (b).*



(b) *Compare and discuss the advantages and disadvantages of these two solutions.*

3 High-frequency behavior of basic amplifiers

It was already discussed in earlier chapters that the amplitude and the phase responses of amplifiers change with frequency, either intentionally or non-intentionally. In some applications, a specific frequency response is desired, for example a band-pass characteristic for an LNA (here, LNA stands for "low-noise amplifier" – the low-noise input stage of a receiver). To shape the frequency response according to our needs, we use reactive components; such as inductors and capacitors. For other applications a flat frequency response is required; however, gain inevitably drops at higher frequencies. The reason for this "non-intentional" change of the frequency response is the "parasitic" components of the circuit; i.e. all non-avoidable reactive (usually capacitive) components related to the devices and the interconnections. To investigate the essential frequency-dependent behaviors of amplifiers, it is necessary to improve the small-signal equivalent circuit of a MOS transistor developed in Chapter 2.

The small-signal equivalent circuit of a MOS transistor biased at a certain operating point in the saturation region was given in Chapter 2, Fig. 2.6. To extend the usability of this equivalent circuit to high frequencies (in other words, to radio frequencies), it is necessary to add the parasitic capacitances and the parasitic resistances that were discussed in Chapter 1, as shown in Fig. 3.1(a). In this equivalent circuit R_D, R_S and R_G are the series resistances of the corresponding regions of the device, where d', s' and g' represent the so-called "internal nodes" that are inaccessible from the device terminals; d, s and g represent the external terminals. $C_{g's'}$ is the gate–source capacitance, $C_{d'g'}$ is the drain–gate overlap capacitance. C_{sb} and C_{db} are the junction capacitances of the source and drain regions and R_{sb} and R_{db} are the series resistances of them.

Although the equivalent circuit given in Fig. 3.1(a) models the small-signal behavior of a MOS transistor in the RF region with good accuracy, it is not convenient for hand calculations. To be able to obtain manageable and interpretable expressions, it is necessary to simplify this equivalent circuit. The reduced equivalent circuit shown in Fig. 3.1(b) is simple enough for hand calculations and provides reasonable accuracy to enable us to understand the basic behavior of the circuits, prior to the detailed SPICE simulations.

The small-signal equivalent circuits of an arbitrary electronic circuit, valid at high frequencies, can be constructed according to the basic rules given in Chapter 2, and then solved using conventional network analysis techniques. One of these techniques is the well-known method called the "Miller Theorem" that has been extensively used

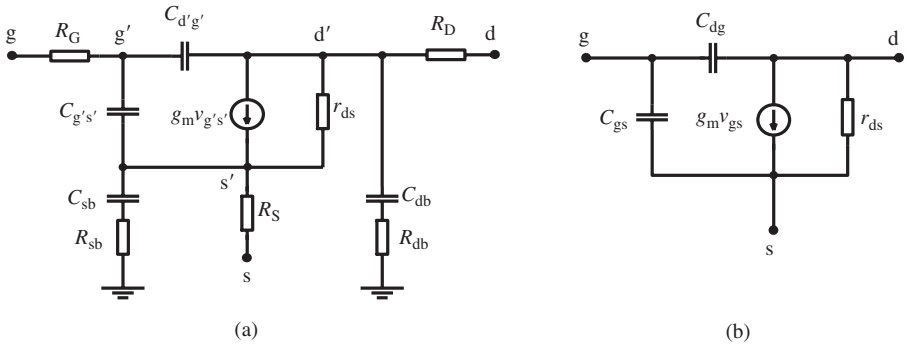


Figure 3.1 (a) The RF small-signal equivalent circuit of a MOS transistor that contains all parasitics. (b) The simplified version for hand calculations.

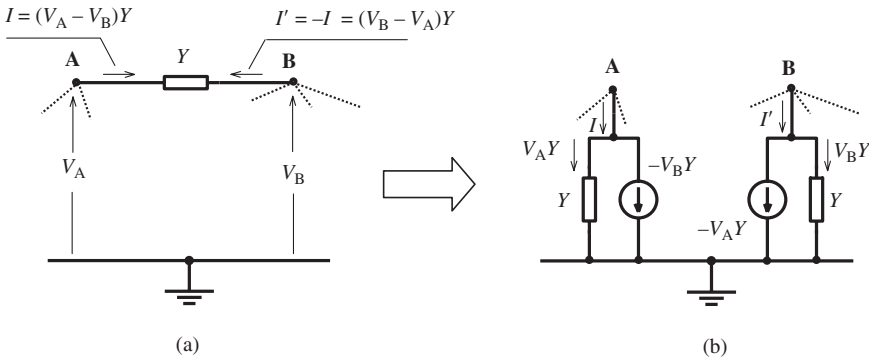


Figure 3.2 Development of the (modified) Miller conversion.

since the early days of the RF electronics [26]. The method was developed to understand and calculate the unexpected effects of the anode-to-grid capacitance of a triode tube, on the input impedance of an amplifier. Before going any further, a modified form of this method, useful for straightforward calculation of the gain function as well as the input and output impedances, will be presented below.

Assume that there is an admittance Y between any two nodes (say A and B) of a circuit. It is possible to *replace* this admittance with two branches, connected between these nodes and the ground, without changing the total current balance of these nodes. The development of this conversion is given step-by-step in Fig. 3.2. The original circuit that contains an admittance Y between the nodes A and B is shown in Fig. 3.2(a). The out-going currents from these nodes owing to Y are shown as I and I', where $I = -I'$. In the converted circuit, the parallel branches replacing Y have currents equal to I and I', respectively. One of the components of these currents is proportional to the node voltages ($V_A Y$ and $V_B Y$) and can be represented by two admittances equal to Y,

connected between these nodes and the ground. The other component of the currents on the parallel branches is proportional to the voltages of the other node, and must be represented by voltage-controlled current sources $(-V_B Y)$ and $(-V_A Y)$, as shown in Fig. 3.2(b).

At first sight it seems that this transformation increases the complexity of the circuit. But it will be seen that it simplifies the calculation of gain and input and output impedances of amplifiers containing a path between the input and output nodes (that almost always exists).

3.1 High-frequency behavior of a common-source amplifier

The high-frequency small-signal equivalent circuit of a common-source amplifier (either resistive-loaded as shown in Fig. 2.1(a), or passive transistor-loaded as shown in Fig. 2.2(a)) is given in Fig. 3.3(a). The load of the amplifier is shown as Y_L which can be any resistive or reactive (containing inductors and/or capacitors as well) admittance. Applying the Miller transformation as explained above, the equivalent circuit can be re-arranged as shown in Fig. 3.3(b) and Fig. 3.3(c). From Fig. 3.3(c), the output voltage v_{ds} can be directly calculated in terms of the input voltage v_{gs} as

$$v_{ds} = -\frac{g_m - sC_{dg}}{Y_o} v_{gs} \tag{3.1}$$

where Y_o is the sum of the load admittance and the output internal admittance of the amplifier which is the parallel equivalent of g_{ds} and C_{dg}. As a more realistic approach, it is useful to add the parasitic capacitance between the drain and the ground C_{op} that is composed of the drain-bulk capacitance, C_{db} and the interconnection parasitics. Now the voltage gain of the circuit can be written as

$$A_v = -\frac{g_m - sC_{gs}}{Y_o} \tag{3.2}$$

Using (3.1), the current source on the left hand side of Fig. 3.3(b) can be expressed in terms of v_{gs}, as shown in Fig. 3.3(c):

$$i = v_{gs}(g_m - sC_{dg})\frac{sC_{dg}}{Y_o} = -v_{gs}A_v sC_{dg}$$

In Fig. 3.4 the equivalent circuit is re-arranged by replacing the current source with the "Miller admittance" y_{mi} connected between the gate and the source nodes. From Fig. 3.4 the input admittance of the amplifier can be found to be

$$y_{mi} = sC_{dg}\frac{g_m - sC_{dg}}{Y_o} = sC_{dg}(-A_v) \tag{3.3a}$$

$$y_i = s(C_{gs} + C_{dg}) + y_{mi} = s(C_{gs} + C_{dg}) + sC_{dg}\frac{g_m - sC_{dg}}{Y_o} \tag{3.3b}$$

From (3.2) and (3.3a) it is obvious that the voltage gain and the input admittance depend on the load admittance, which in many cases is the dominant part of Y_o.

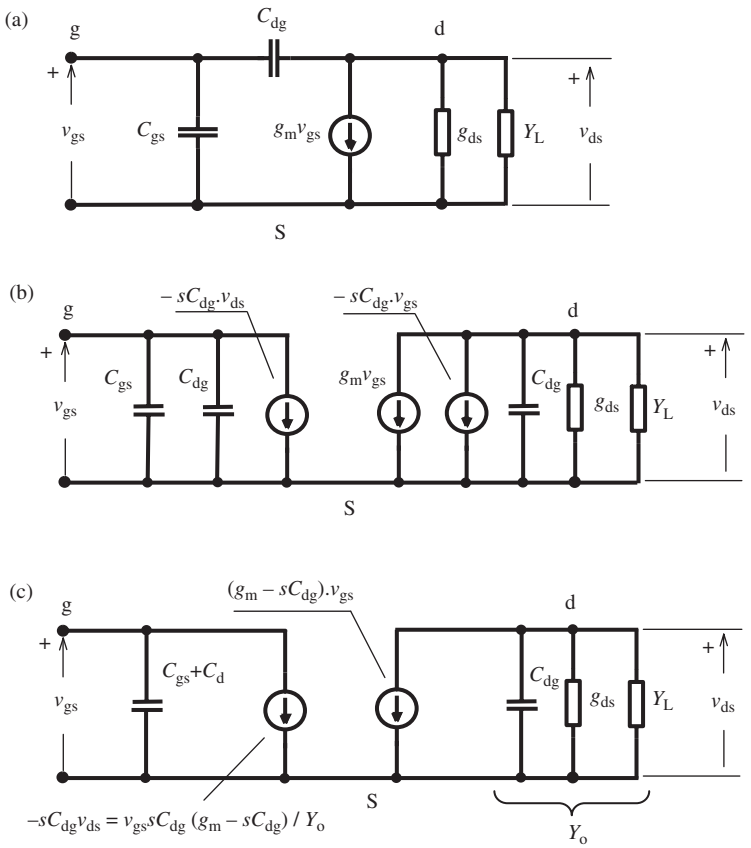


Figure 3.3 Development of the small-signal equivalent circuit of the common-source amplifier: (a) original equivalent circuit, (b) equivalent circuit after Miller transformation, (c) equivalent circuit after simplification.

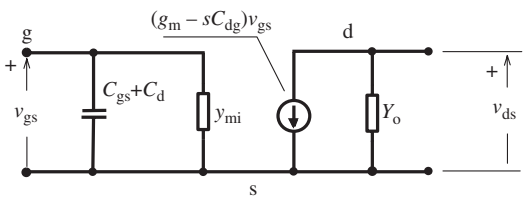


Figure 3.4 Modification of the equivalent circuit for calculation of input admittance.

Therefore, it is necessary to investigate the frequency-dependent behavior of the gain and the input admittance for typical cases of the load. The most important case is the R-C load and will be investigated in the following section. The other important case, the L-C load (tuned load) will be dealt with in detail in Chapter 4.

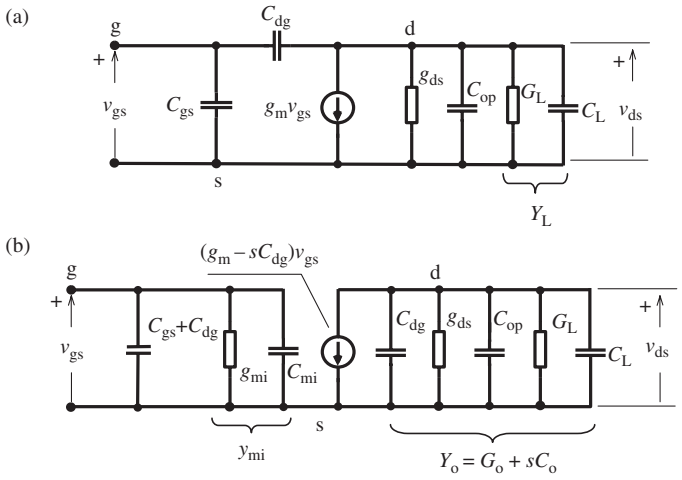


Figure 3.5 The small-signal equivalent circuit of an R-C loaded amplifier (a) before Miller transformation, (b) after Miller transformation.

3.1.1 The R-C load case

The basic small-signal equivalent circuit and the circuit after application of the Miller transformation of an R-C loaded common-source amplifier are given in Fig. 3.5(a) and (b), respectively. Note that even for the case of pure resistive load at the output node, the actual conditions correspond to the R-C loaded amplifier case owing to C_{dg} and the output parasitics. The total admittance at the output node ($Y_o = G_o + sC_o$) is the parallel equivalent of the external load, $Y_L = G_L + sC_L$, and $g_{ds} + s(C_{dg} + C_{op})$, where C_{op} represents the total parasitic capacitance of the output node.

Now the voltage gain of the amplifier can be calculated from (3.2):

$$A_v = -\frac{g_m - sC_{dg}}{G_o + sC_o} = \frac{C_{dg}}{C_o}\frac{(s - s_0)}{(s - s_p)} \tag{3.4a}$$

where

$$s_0 = +\frac{g_m}{C_{dg}} \quad \text{and} \quad s_p = -\frac{G_o}{C_o} = -\frac{(g_{ds} + G_L)}{(C_{op} + C_{dg} + C_L)} \tag{3.4b}$$

are the zero and the pole of the voltage gain function. The low-frequency ($s \to 0$) voltage gain of the amplifier can be easily obtained from (3.4a) as

$$A_v(0) = -\frac{g_m}{G_o} = -\frac{g_m}{(G_L + g_{ds})} \tag{3.5}$$

The magnitude and phase characteristics of the amplifier can be obtained from the pole–zero diagram of the gain function (Fig. 3.6).

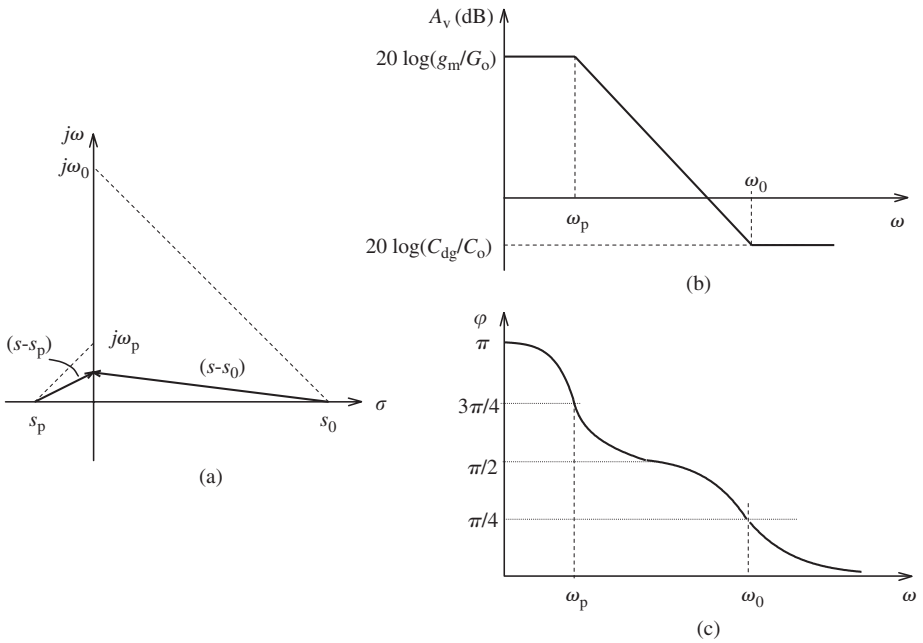


Figure 3.6 (a) The pole–zero diagram of the R-C loaded amplifier. (b) The magnitude characteristic of the voltage gain with its asymptotes. (c) The approximate phase characteristic.

The following conclusions can be drawn from the characteristics.

- The magnitude of the gain decreases with frequency and drops to $1/\sqrt{2}$ (or 3 dB below) of its low-frequency value at ω_p corresponding to the pole.
- For $\omega > \omega_p$, the magnitude characteristic of an R-C loaded amplifier decreases with a 20 dB/decade slope, up to the frequency corresponding to the "zero" of the gain function, which is usually much higher than the pole frequency (or the –3 dB frequency).
- The –3 dB frequency of the amplifier is inversely proportional to the total output node capacitance. This frequency (or the bandwidth of the amplifier) can be set to a lower value by using an appropriate parallel capacitor connected to the output.
- The phase of the gain is 180° at low frequencies, decreases to (approximately) 135° at the pole frequency and tends asymptotically to 90°. But owing to the existence of the right half-plane zero, it continues to decrease and reaches (approximately) 45° at the zero frequency and becomes asymptotic to zero at very high frequencies. This additional 90° phase shift, compared to the hypothetical case of $C_{dg} = 0$, can be the source of certain problems in feedback amplifiers.

The gain–bandwidth product of the amplifier can be calculated from (3.4b) and (3.5) as

$$GBW = |A_v| \times \frac{1}{2\pi}|s_p| \cong \frac{g_m}{2\pi C_o} \tag{3.6}$$

The input admittance can be calculated from (3.3a), fixing the value of the Miller admittance for this case. The Miller admittance can be calculated from (3.3a) as

$$y_{\text{mi}} = sC_{\text{dg}}\frac{g_{\text{m}} - sC_{\text{dg}}}{G_{\text{o}} + sC_{\text{o}}} \tag{3.7}$$

and in the frequency domain,

$$y_{\text{mi}}(\omega) = \text{j}\omega C_{\text{dg}}\frac{g_{\text{m}} - \text{j}\omega C_{\text{dg}}}{G_{\text{o}} + \text{j}\omega C_{\text{o}}} = g_{\text{mi}}(\omega) + \text{j}b(\omega) \tag{3.8}$$

The real and imaginary parts of this admittance are

$$g_{\text{mi}}(\omega) = \frac{C_{\text{dg}}}{C_{\text{o}}}\left(g_{\text{m}} + G_{\text{o}}\frac{C_{\text{dg}}}{C_{\text{o}}}\right)\frac{1}{(\omega_{\text{p}}/\omega) + 1} \tag{3.9a}$$

$$b_{\text{mi}}(\omega) = \omega\frac{C_{\text{dg}}^2}{C_{\text{o}}}\frac{\omega_0\omega_{\text{p}} - \omega^2}{\omega_{\text{p}}^2 + \omega^2} = \omega C_{\text{mi}} \tag{3.9b}$$

where ω_{p} and ω_0 are the frequencies corresponding to the pole and the zero of the voltage gain function.

Equation (3.9a) shows that there is an *unexpected* real part of the input admittance, which manifests itself as a frequency-dependent input conductance. The variation of this conductance as a function of ω is plotted in Fig. 3.7(a). The value of this

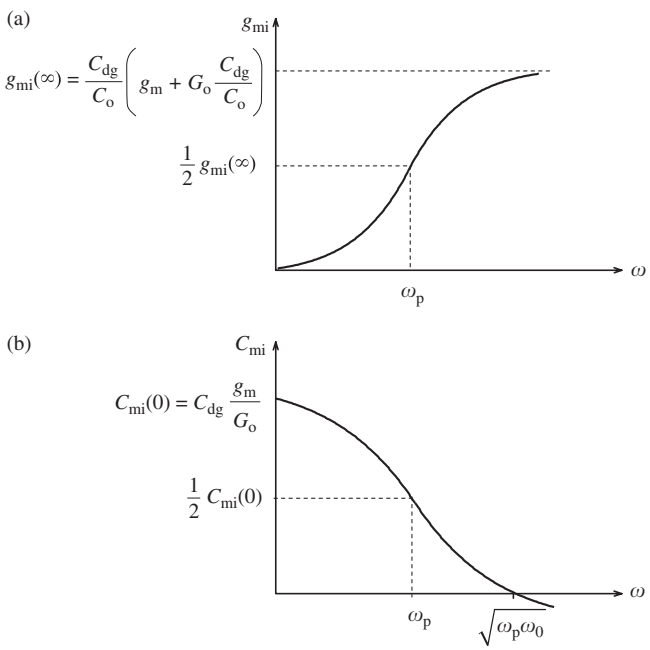


Figure 3.7 Variation of (a) the input conductance and (b) the input capacitance of the common-source amplifier.

conductance is zero for $\omega = 0$, and increases with frequency up to

$$g_{mi}(\omega \to \infty) = \frac{C_{dg}}{C_o}\left(g_m + G_o\frac{C_{dg}}{C_o}\right) \tag{3.10}$$

The input conductance at the pole frequency of the amplifier is

$$g_{mi}(\omega_p) = \frac{g_{mi}(\omega \to \infty)}{2} = \frac{1}{2}\frac{C_{dg}}{C_o}\left(g_m + G_o\frac{C_{dg}}{C_o}\right) \tag{3.11}$$

Since the gate is isolated from all other parts of the device by the gate oxide which is – almost – an ideal insulator, the input conductance of a MOS transistor is generally assumed to be zero. This is true for the DC case, and the conductance is negligibly small at low frequencies. But owing to the existence of the drain–gate parasitic capacitance as seen from (3.9), this assumption is not true for the AC case, in general. Especially at higher frequencies approaching the pole frequency of the amplifier, the input conductance can reach surprisingly high values.

Example 3.1 For the amplifier given in Fig. 3.8, the DC drain current (bias) of the transistor is $I_D = 150\,\mu A$, therefore $V_{DS} = 1.5$ V. For this operating point the small-signal parameters are $g_m = 2$ mS, $C_{dg} = 20$ fF, $C_{gs} = 90$ fF, and $g_{ds} = 30\,\mu S$. The parasitic capacitance of the output node is assumed to be 10 fF.

The low-frequency gain and the pole frequency can be calculated from (3.5) and (3.4b) as

$$A_v(0) = -\frac{2\times 10^{-3}}{(3\times 10^{-6}\times 10^{-4})} = -15.38 \Rightarrow 23.74\ \text{dB}$$

$$f_p = \frac{1}{2\pi}\frac{(30\times 10^{-6} + 10^{-4})}{(10\times 10^{-15} + 20\times 10^{-15} + 50\times\ 10^{-15})} = 258.7\ \text{MHz}$$

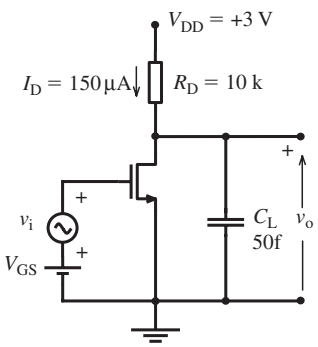


Figure 3.8 The R-C loaded MOS amplifier investigated in the example.

The input conductance for this frequency calculated from (3.11) is

$$g_{mi}(f_p) = \frac{1}{2}\frac{20\times10^{-15}}{80\times10^{-15}}\left(2\times10^{-3} + 10^{-4}\frac{20\times10^{-15}}{80\times10^{-15}}\right) = 0.253\times10^{-3} \text{ siemens}$$

which corresponds to an input resistance of 3.95 kΩ! Even for one-tenth of the pole frequency, the input conductance is still high and can be calculated from (3.9a) as 4.95 μS, which corresponds to an input resistance of 200 kΩ.

This example proves that the input conductance of a common-source MOS amplifier has a frequency-dependent resistive component and this resistance can drop to considerably low values – which influences the output load of the previous stage or the signal source. It is obvious that especially for high internal impedance signal sources, this frequency-dependent input conductance of the amplifier has to be taken into account.

Note that the input capacitance of the amplifier is the sum of C_{gs}, C_{dg} and C_{mi}. In Fig. 3.7(b) the variation of the Miller capacitance is plotted as a function of ω. It shows that the third component of the input capacitance has a value proportional to g_m/G_o that is the magnitude of the Miller capacitance at low frequencies, and is usually high.

For the example given above, the value of the Miller capacitance is

$$C_{mi}(0) = C_{dg}\frac{g_m}{G_o} = 20\times10^{-15}\,\frac{2\times10^{-3}}{130\times10^{-6}} = 307.7 \text{ fF}$$

and the total input capacitance is

$$C_i = C_{gs} + C_{dg} + C_{mi} = 90 + 20 + 307.7 = 417.7 \text{ fF}$$

of which the dominant part is the Miller capacitance. C_i capacitively loads the previous stage (or the driving signal source). This high capacitive load affects the high-frequency performance of the previous stage and is the main reason for the deterioration of the high-frequency gain of multi-stage amplifiers.

In Fig. 3.9, the PSpice simulation results are shown for the amplifier given in Fig. 3.8. The transistor is an AMS 0.6 μ NMOS device with $L = 0.6$ μm, $W = 60$ μm. The drain current is 150 μA. Small-signal parameters of this transistor are approximately equal to the values used in the example. The frequency axes for the magnitude and phase characteristics (Fig. 3.9(a) and 3.9(b)) are intentionally drawn up to 30 GHz, to see the effects of the zero of the gain function. The variations of the input admittance and the input capacitance, which is the sum of C_{gs}, C_{dg} and the C_{mi} Miller capacitance are given in Fig. 3.9(c) and (d). The simulation results reasonably agree with the values calculated using the analytical expressions.

3.2 The source follower amplifier at radio frequencies

As explained in Chapter 2, the source follower is typically used as a buffer to couple a high internal impedance source to a low impedance load. The simplified circuit diagram of a source follower is given in Fig. 3.10(a). The load can have any form; a capacitive (R-C) load, an inductive (R-L) or a tuned (L-C) load. In this section

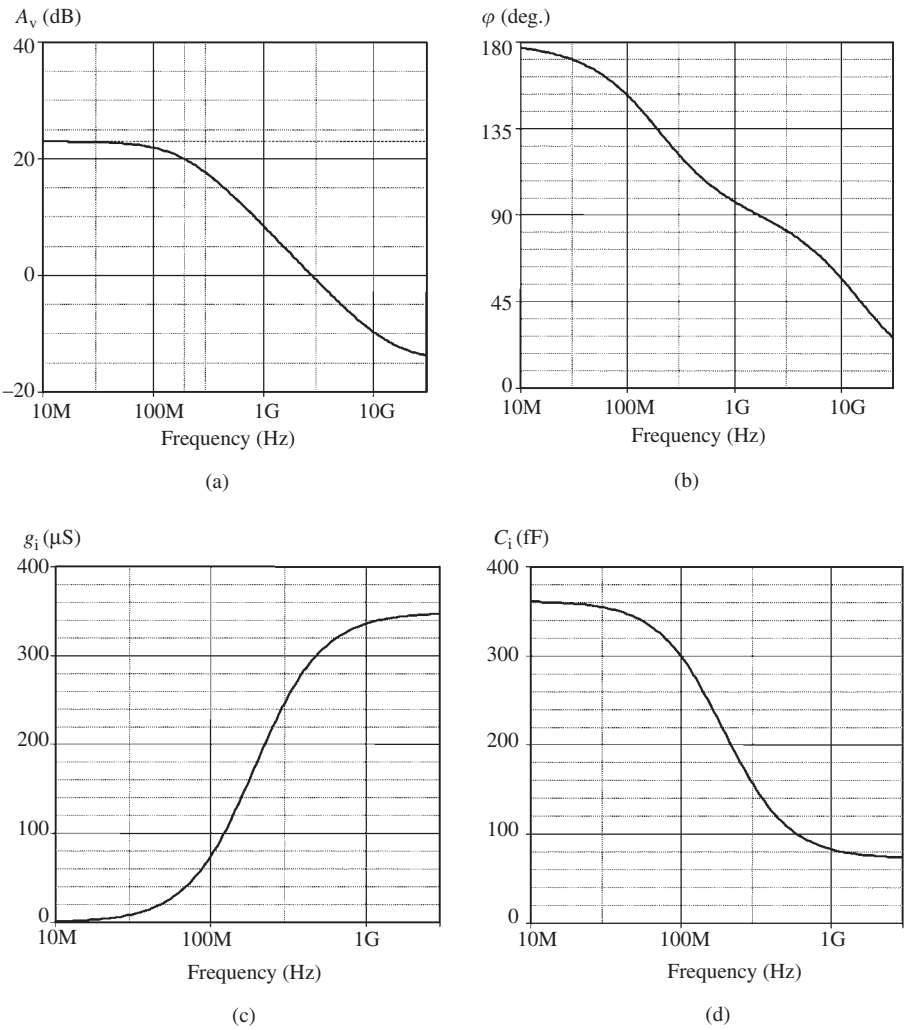


Figure 3.9 PSpice simulation results for the circuit shown in Fig. 3.8: (a) the magnitude characteristic, (b) the phase characteristic of the voltage gain, (c) the variation of the input conductance and (d) the variation of the input capacitance as a function of frequency. The important figures of performance are: $A_v(0) \cong 23\,\text{dB}$, $f_p \cong 192\,\text{MHz}$, $g_i(f_p) \cong 230\,\mu\text{S}$, $C_i(0) \cong 360\,\text{fF}$.

we will deal with the R-C load only and leave the tuned load case to Chapter 4 (Frequency-selective RF circuits). The small-signal equivalent circuit of an R-C loaded source follower is given in Fig. 3.10(b), and re-arranged in Fig. 3.10(c) to ease the solution, where

$$G = G_s + g_{ds} \text{ and } C = C_s + C_{op}$$

From Fig. 3.10(c), the output and input voltages can be written as

$$v_o = \frac{i_i + g_m v_{gs}}{Y} \quad v_i = v_{gs} + v_o \tag{3.12}$$

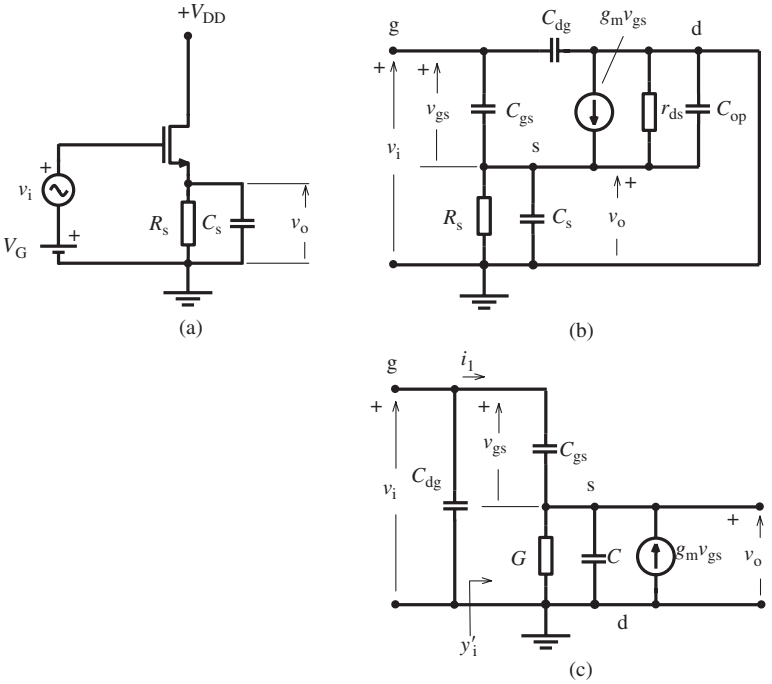


Figure 3.10 The source follower amplifier circuit: (a) simplified circuit diagram, (b) small-signal equivalent circuit and (c) re-arranged equivalent circuit for calculation of gain and input impedance.

where

$$v_{gs} = i_i / sC_{gs} \text{ and } Y = G + sC \tag{3.13}$$

From (3.12) and (3.13) the voltage gain can be solved as

$$A_v = \frac{g_m + sC_{gs}}{(g_m + G) + s(C_{gs} + C)} \tag{3.14a}$$

which can also be expressed in terms of the pole and zero frequencies as

$$A_v = \frac{C_{gs}}{(C_{gs} + C)} \frac{(s - s_0)}{(s - s_p)} \tag{3.14b}$$

where

$$s_0 = -\frac{g_m}{C_{gs}} \quad \text{and} \quad s_p = -\frac{(g_m + G)}{(C_{gs} + C)} \tag{3.15}$$

The low-frequency gain can be seen from (3.14a) to be

$$A_v(0) = \frac{g_m}{(g_m + G)}$$

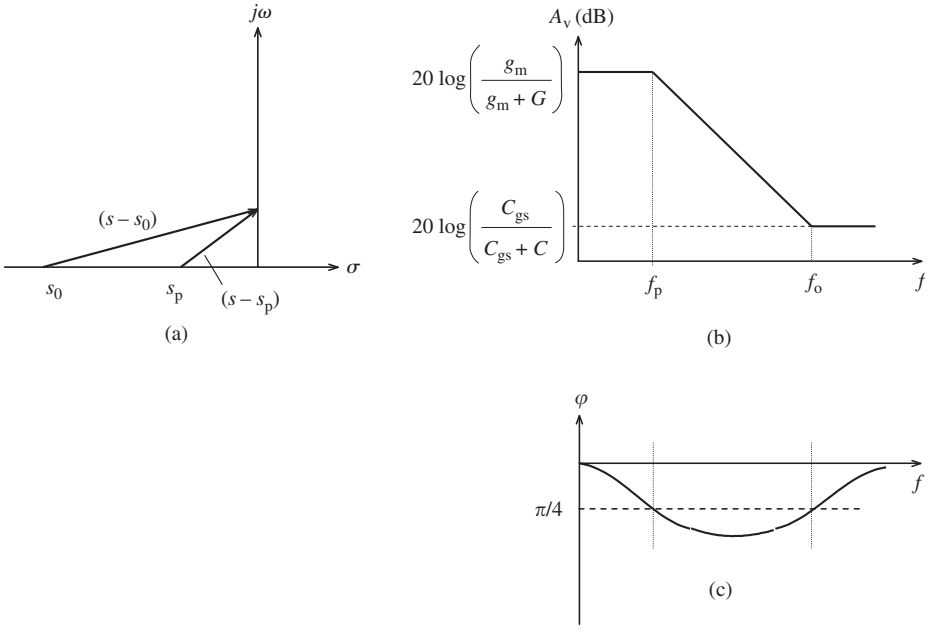


Figure 3.11 (a) The voltage gain pole–zero diagram of the R-C loaded source follower. (b) The magnitude characteristic of the voltage gain with its asymptotes. (c) The approximate phase characteristic.

which – naturally – fits to (2.24a). The magnitude and phase characteristics of the amplifier can be obtained from the pole–zero diagram of the gain function (Fig. 3.11).

From Fig. 3.10(c) it can be seen that the input admittance of the source follower is the sum of the admittance shown with y_i' and sC_{dg}. y_i' is i_i/v_i, and i_i can be calculated as

$$i_i = (v_i - v_o)sC_{gs} = v_i(1 - A_v)sC_{gs}$$

Using the value of A_v from (3.14a), the input admittance (C_{dg} excluded) in the ω domain can be calculated as

$$y_i(\omega) = \left(1 - \frac{(g_m + j\omega C_{gs})}{(g_m + G) + j\omega(C_{gs} + C)}\right) \cdot j\omega C_{gs} = g_i + j\omega C_i \tag{3.16}$$

The input conductance, which is the real part of this expression, can be arranged as

$$g_i(\omega) = \frac{C_{gs}}{(C_{gs} + C)^2}(GC_{gs} - g_m C)\frac{1}{1 + (\omega_p/\omega)^2} \tag{3.17a}$$

or

$$g_i(\omega) = g_i(\infty)\frac{1}{1 + (\omega_p/\omega)^2} \tag{3.17b}$$

A careful analysis of these expressions leads to the following important conclusions.

- For an R-C loaded source follower, if $GC_{gs} \succ g_m C$ the input conductance is positive and similar to that of an R-C loaded common-source amplifier; it increases with frequency.
- In the case of $GC_{gs} = g_m C$, the input conductance is zero for all frequencies. This is an important feature and useful for the design of very low input conductance (high input impedance) amplifiers.
- In the case of $GC_{gs} < g_m C$ the input conductance of a capacitive-loaded source follower is negative and this negative conductance increases with frequency. This property can be used when a negative conductance is needed.[1] If there is a parasitic inductance on the gate connection of the transistor, this negative conductance, together with this inductance and the input capacitance of the transistor, can result in ringing behavior on the signal, or it can even lead to oscillation. For on-chip source follower input stages, the inductance of the bonding wire can cause such ringing or oscillation. In such cases, this negative conductance must be compensated with an appropriate resistance connected in parallel or series to the gate.

The total input capacitance of a source follower can be calculated as the sum of C_{dg} and the capacitance calculated from (3.16) as

$$C_{iT} = C_{dg} + C\frac{\omega_0'\omega_p + \omega^2}{\omega_p^2 + \omega^2} \tag{3.18}$$

where ω_p and ω_0' are the pole and zero frequencies of the input admittance function. The pole frequency is equal to the pole frequency of the gain function as given in (3.15), and the zero frequency is $\omega_0' = (G/C)$. The values of the input capacitance for low frequencies and for very high frequencies are

$$C_{iT}(0) = C_{dg} + (C_{gs} + C)\frac{G}{(g_m + G)} \quad C_{iT}(\infty) = C_{dg} + C$$

The frequency-dependence of the input conductance and the input capacitance of a source follower for $GC_{gs} \ll g_m C$ are plotted in Fig. 3.12(a) and (b), respectively.

In Fig. 3.13(a), the PSpice simulation results are given for different capacitive loads of a source follower. It is seen that for small capacitive loads the input conductance is positive, for an appropriate load it is equal to zero for all frequencies, and negative for higher capacitive loads. Figure 3.13(b) shows the output waveform of this amplifier, driven by a 50-ohm input resistance square-wave signal source and having a 5 nH series gate inductance (that corresponds to approximately 5 mm long bonding wire). The observed ringing corresponds to the resonance frequency of the inductance and the input capacitance of the amplifier.

Source followers usually drive low-impedance loads. In some applications, for example if the source follower drives a transmission line, the output internal

[1] See Chapter 4, Example 4.2 for a practical application of this property to increase the quality factor of an inductor.

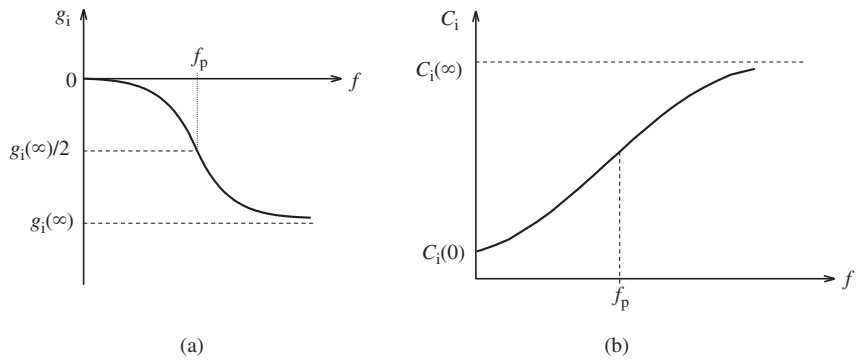


Figure 3.12 Frequency-dependence of (a) the input conductance and (b) the input capacitance of a source follower for the condition $GC_{gs} \ll g_m C$ which is usually valid.

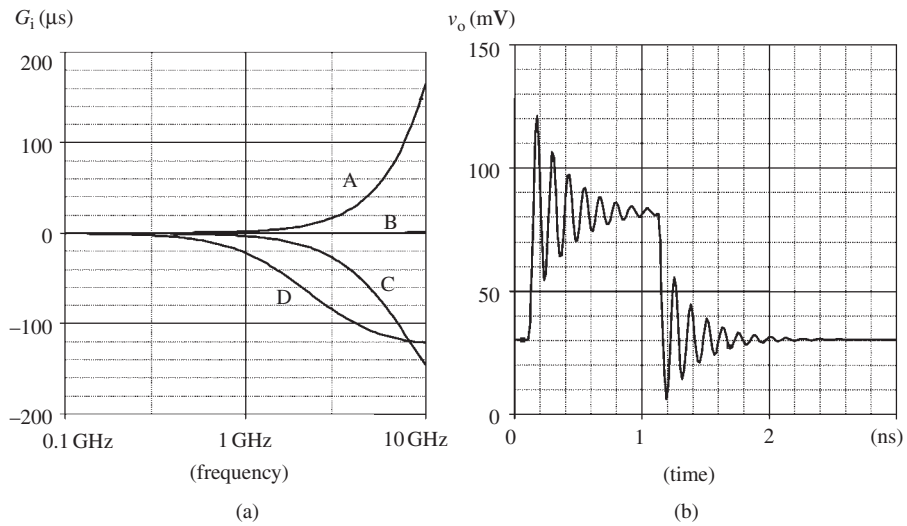


Figure 3.13 (a) The frequency-dependence of the input conductance of a source follower for different load capacitance values for (A) 20 fF, (B) 80 fF, (C) 200 fF and (D) 1pF. (b) The waveform of the output voltage for a square-wave input signal with an $L = 5$ nH inductor in series to the gate and 200 fF source load capacitance, with the internal resistance of the pulse source being 50 Ω. (The transistor is an AMS 0.35 μ NMOS device with $W = 35\,\mu$m $L = 0.35\,\mu$m. $R_L = 100$ ohm and $I_D = 300\,\mu$A.)

impedance must keep a pre-defined value in a wide frequency range in order to maintain a perfect matching for optimum power transfer and to prevent reflections. Therefore, the behavior of the output impedance (or admittance) at high frequencies has to be investigated.

The simplified schematic and the high-frequency small-signal equivalent circuit of a source follower to calculate the output impedance (or admittance) are given in

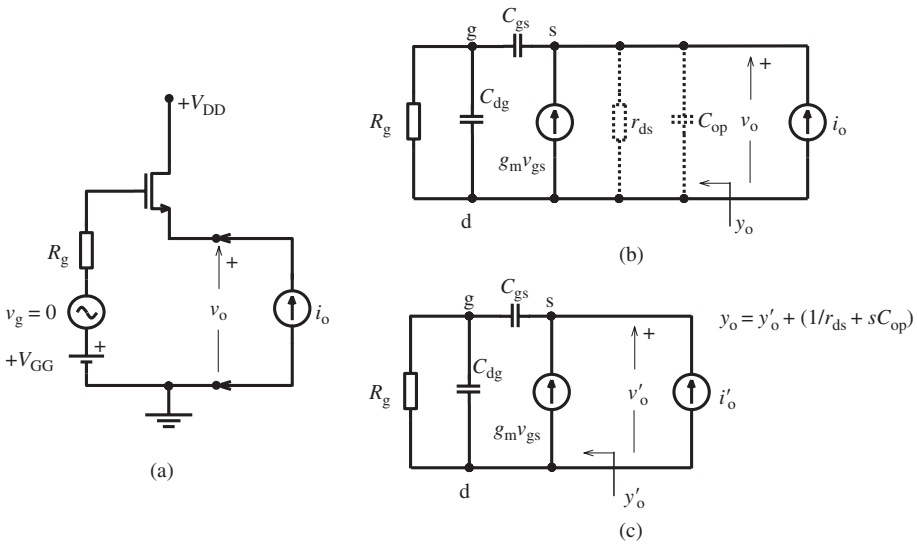


Figure 3.14 (a) The conceptual arrangement to calculate the output impedance of the circuit. (b) The equivalent circuit of the source follower amplifier used for the calculation of output impedance (admittance). (c) Simplified equivalent circuit.

Fig. 3.14(a) and (b). It is assumed that the internal impedance of the signal source driving the input is resistive and is shown with R_g. A current source (i_o) is connected to the output port to find the output internal admittance,[2] and the voltage of this port will be calculated. It can be easily seen that the parallel components r_o and C_{op} can be excluded to ease the calculation (Fig. 3.14(c)), and then re-included.[3] The output admittance of the reduced circuit can now be calculated as

$$y'_o = \frac{i'_o}{v'_o} = \frac{(G_g + sC_{dg})(g_m + sC_{gs})}{G_g + s(C_{dg} + C_{gs})} \tag{3.19}$$

This admittance function has two zeros and one pole described as

$$s_{01} = -\frac{G_g}{C_{dg}} \quad s_{02} = -\frac{g_m}{C_{gs}} \quad s_p = -\frac{G_g}{(C_{dg} + C_{gs})} \tag{3.20}$$

and the low-frequency value of the output admittance is equal to g_m, as already shown in Chapter 2.

The frequency characteristic of y'_o depends on the relative positions of the zeros and the pole. It is obvious that $s_{01} > s_p$ and the position of s_{02} depends on the value of g_m. In Fig. 3.15 the pole–zero diagram and the frequency characteristics of $|y'_o|$ corresponding to $|s_{01}| > |s_{02}| > |s_p|$ are shown. It can be seen that, under appropriate conditions, there

[2] For the sake of simplicity of the calculations, the output admittance is preferred.

[3] It will be seen later that the admittance corresponding to these components is very small compared to the admittance of the reduced circuit, consequently, it is negligible.

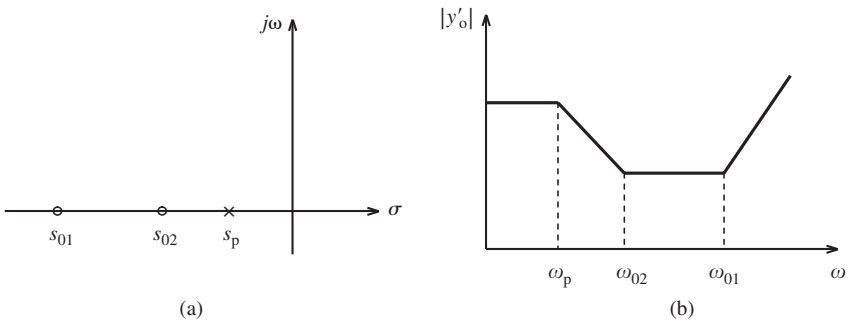


Figure 3.15 (a) The pole–zero diagram of y'_o of a source follower. (b) The corresponding frequency characteristic.

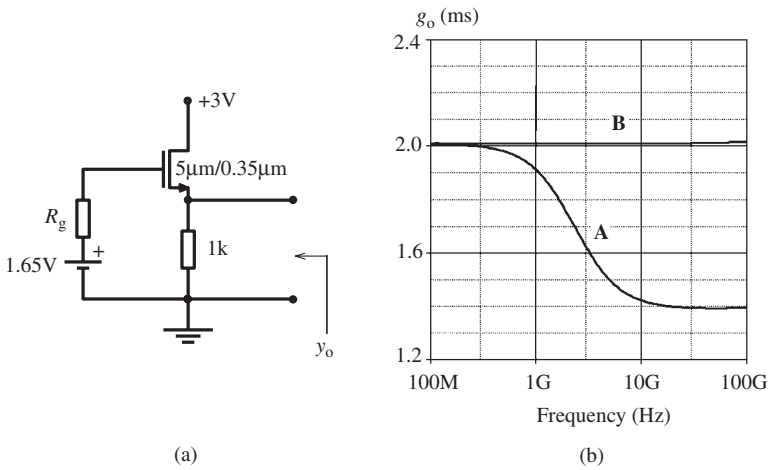


Figure 3.16 (a) The simulated source follower. (b) The output conductance of the follower (A) without pole–zero cancellation and (B) with pole–zero cancellation. The transistor is an AMS 0.35 micron NMOS transistor with $W = 5$ µm, $L = 0.35$ µm and $I_D = 0.7$ mA. The value of the driving signal source internal resistance is 10 kΩ for (A) and 1 kΩ for (B).

is a possibility of pole–zero cancellation. For example, if $|s_{02}| = |s_p|$ they cancel each other and the output conductance remains constant up to the frequency corresponding to s_{01}. In Fig. 3.16, the PSpice simulation results for a source follower are shown, without and with pole–zero cancellation.

3.3 The common-gate amplifier at high frequencies

The schematic diagram and the small-signal equivalent circuit of a common-gate amplifier are given in Fig. 3.17(a) and (b). Similar to the previously investigated

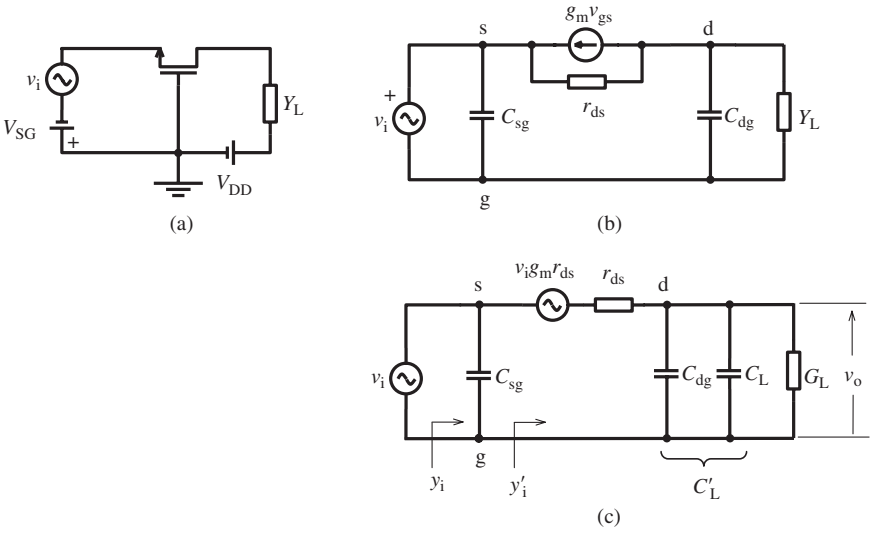


Figure 3.17 (a) The circuit schematic of the common-gate amplifier. (b) The small-signal equivalent circuit of the common-gate amplifier. (c) Equivalent circuit with R-C load.

common-source and source follower circuit, the investigation of the case of tuned load will be left to Chapter 4. For an R-C (or G-C) load, the equivalent circuit is given in Fig. 3.17(c), with a Norton–Thévenin transformation to ease the solution, where

$$Y_L = G_L + sC_L \quad Y'_L = Y_L + sC_{dg} = G_L + s(C_L + C_{dg}) = G_L + sC'_L$$

From Fig. 3.17(c), the voltage gain of the amplifier can be solved as

$$A_v = \frac{v_o}{v_i} = \frac{g_m + g_{gs}}{C'_L} \frac{1}{(s - s_p)} \tag{3.21}$$

where the pole of the gain function that corresponds to the 3 dB frequency (or bandwidth) of the amplifier is

$$s_p = -\frac{G_L + g_{ds}}{C'_L} \tag{3.22}$$

Using (3.21) and (3.22), the low-frequency gain and the gain–bandwidth product can be calculated as

$$A_v(0) = \frac{g_m + g_{ds}}{G_L + g_{ds}} \cong \frac{g_m}{G_L + g_{ds}} \tag{3.23}$$

and

$$GBW \cong \frac{1}{2\pi} \frac{g_m}{C'_L} \tag{3.24}$$

which is equal to that of the common-source amplifier.

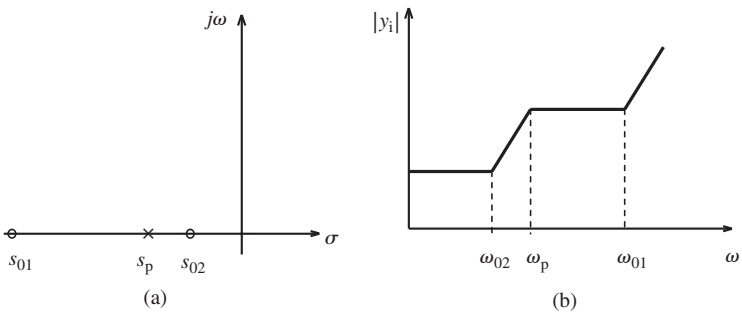


Figure 3.18 (a) The pole–zero diagram, and (b) the frequency characteristic of the input admittance of a common-gate amplifier.

Although the magnitudes of the voltage gain and the gain–bandwidth product of common-source and common-gate amplifiers are equal, there are significant differences between their input and output admittances.

The input admittance of a common-gate amplifier can be calculated from Fig. 3.17(c), as the sum of sC_{gs} and y_i':

$$y_i = \frac{s^2 C_{gs} C_L' + s[C_{gs}(g_{ds} + G_L) + C_L'(g_m + g_{ds})] + G_L(g_m + g_{ds})}{(g_{ds} + G_L) + sC_L'} \tag{3.25a}$$

which can be simplified for $C_{gs}(g_{ds} + G_L) < C_L'(g_m + g_{ds})$ and $g_{ds} < g_m$, that are usually valid:

$$y_i = \frac{s^2 C_{gs} C_L' + s g_m C_L' + G_L g_m}{(g_{ds} + G_L) + sC_L'} = C_{gs} \frac{(s - s_{01})(s - s_{02})}{(s - s_p)} \tag{3.25b}$$

where

$$s_p = -\frac{G_L + g_{ds}}{C_L'} \quad s_{01} \cong -\frac{g_m}{C_{gs}} \quad s_{02} \cong -\frac{G_L}{C_L'} \tag{3.26}$$

The pole–zero diagram and the variation of the magnitude of the input admittance with frequency are shown in Fig. 3.18.

From Fig. 3.18(b) we can see that:

- the magnitude of the input admittance is constant (i.e. the input admittance itself is a real quantity) up to the vicinity of $\omega_{02} = G_L/C_L'$ (which is approximately equal to the 3 dB frequency of the voltage gain);
- if $G_L \gg g_{ds}$, the magnitude of the input admittance for low frequencies (also up to the vicinity of the 3 dB frequency of the amplifier) is equal to the transconductance of the transistor. Since g_m is usually of the order of several mS, this means that the common-gate circuit is a low-input impedance amplifier.

The output admittance (y_o) of a common-gate amplifier will be calculated from Fig. 3.19(a), as the sum of sC_{dg} and y_o':

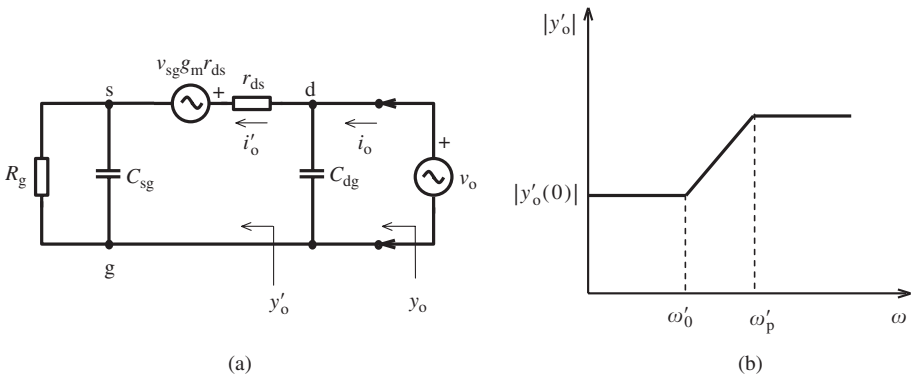


Figure 3.19 (a) The small-signal equivalent circuit of the common-gate amplifier used for calculation of output admittance. (b) Frequency characteristic of the output admittance.

$$y'_o = \frac{i'_o}{v_o} = \frac{1}{r_{ds}} \frac{s - s'_0}{s - s'_p} \tag{3.27a}$$

where

$$s'_0 = -\frac{1}{R_g C_{gs}} \quad \text{and} \quad s'_p = -\frac{1 + \frac{R_g}{r_{ds}}(1 + g_m r_{ds})}{R_g C_{gs}} \tag{3.27b}$$

and obviously the magnitude of the zero is smaller than that of the pole. The magnitude of the output admittance for zero frequency (as well as for low frequencies) is

$$y'_o(0) = \frac{1}{r_{ds}(1 + g_m R_g) + R_g} \tag{3.28}$$

The frequency characteristic of y'_o with its asymptotes is given in Fig. 3.19(b). From (3.28) and (3.27a) we can conclude that

- the output resistance, $r_o = 1/\mathrm{Re}(y'_o) = 1/\mathrm{Re}(y_o)$ of a common-gate amplifier can have very high values for high driving-source internal resistance and high transconductance values; in other words the output acts as an – almost – ideal current source;
- the frequency corresponding to $r_o C_{dg}$ may dominate over the zero of y'_o and determines the corner frequency of y_o;
- since a common-gate stage has a low input impedance and a high output impedance, and the output current is equal to the input current, it can be considered as a unity gain – almost – ideal current amplifier, and can be used as an intermediate stage to transfer efficiently the current of a moderate (not too high, not too low) internal impedance current source to the load.

If we evaluate the already investigated properties of a common-gate stage; namely its low input impedance, its very high output internal impedance and its unity current gain, we can conclude that it is suitable to be used as a transimpedance amplifier. In

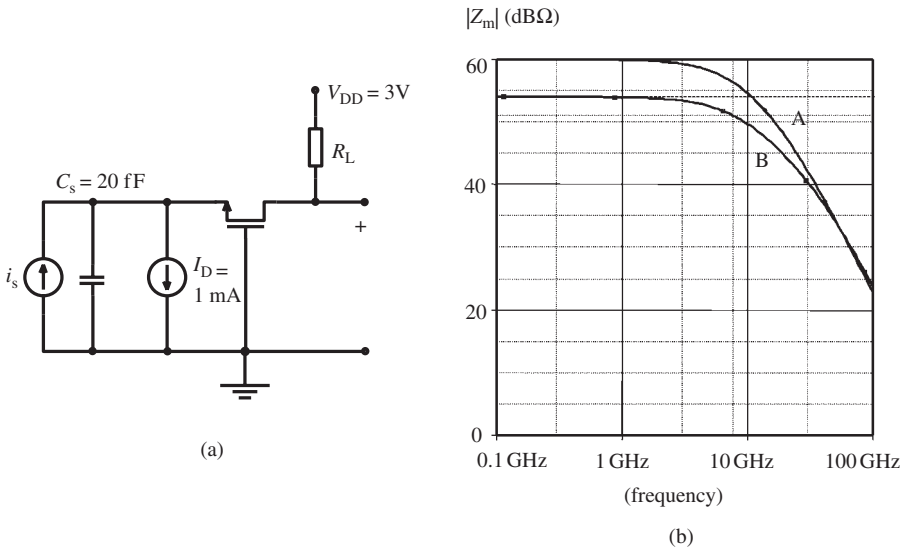


Figure 3.20 (a) Example of a common-gate transimpedance amplifier, (b) PSpice simulation results showing the frequency characteristics for 1 k ohm (A) and 500 ohm (B) load resistance. The transistor is an AMS 0.35 μ NMOS device with $L = 0.35\,\mu m$, $W = 10\,\mu m$.

Fig. 3.20, the schematic diagram and the PSpice simulation results of a common-gate transimpedance amplifier are given. The current signal source and the parallel capacitance represent the input signal source, for example a photodiode. The low-frequency value of the transimpedance is equal to the load resistance. The 40 dB/decade slope of the frequency response indicates that there are two poles in the high-frequency region, apparently one from the input resistance, which is approximately equal to $1/g_m$, and the total capacitance parallel to the input, and the other one from the output side, and equal to $1/R_L C_{db}$. The magnitude of the transimpedance at low frequencies, and the 3 dB frequency for a load of $R_L = 1$ k ohm and $R_L = 500$ ohm, are 1k ohm/5.8 GHz and 500 ohm/6.72 GHz, respectively.[4]

3.4 The "cascode" amplifier

We have seen that two of the basic one-transistor amplifiers, namely the common-source amplifier and the common-gate amplifier, provide high voltage gain. But both of them have severe drawbacks. Owing to the output-to-input (drain-to-gate) capacitance and the Miller effect, the input admittance of a common-source amplifier deteriorates (increases) at high frequencies. The input admittance of a common-gate

[4] These figures indicated the "intrinsic" bandwidth of the amplifier (i.e. with no external load capacitance). In the case of an external load capacitance, the pole related to the output becomes $1/R_L$ ($C_{db}+C_L$) and dominates the frequency response.

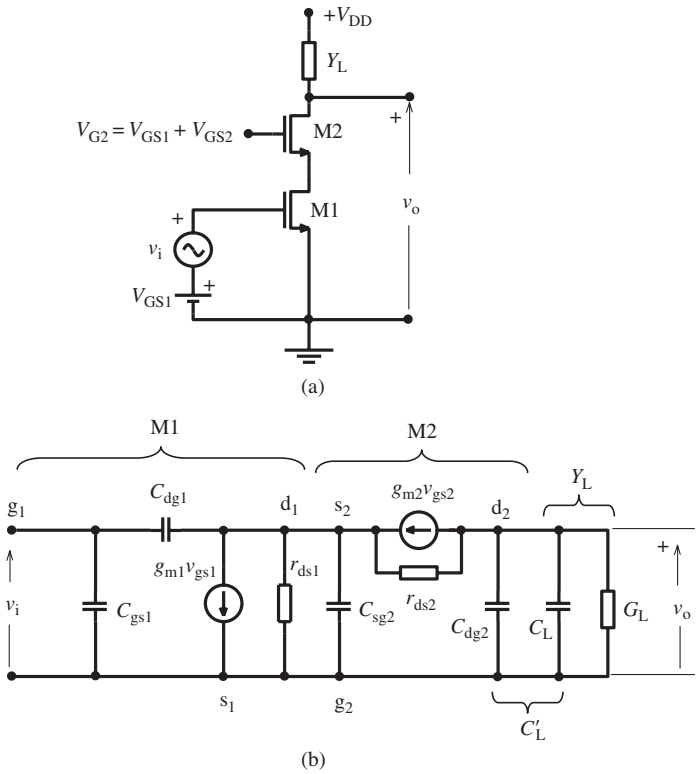


Figure 3.21 (a) The circuit schematic of a cascode amplifier and (b) its small-signal equivalent circuit.

amplifier is inherently small, therefore is not suitable to be used as an efficient voltage amplifier, especially if the signal source has high internal impedance.

The solution is the so-called "cascode" amplifier[5] that provides low output-to-input feedback, high input impedance and high voltage gain. In a cascode amplifier a common-source amplifier and a common-gate amplifier are combined in such a way that the common-source stage is loaded with the input of the common-gate amplifier (Fig. 3.21(a)).

The load is connected to the output of the common-gate stage. The basic operation principles are as follows.

- Since the input conductance of the common-gate stage is approximately equal to $1/g_{m2}$, the voltage gain of the input (common-source) stage is $A_{v1} \cong -(g_{m1}/g_{m2})$, which is small in magnitude and equal to unity for $g_{m1} = g_{m2}$ (and close to unity for many cases). Consequently the Miller admittance is small and equal to sC_{dg} up to the pole frequency of the voltage gain (see Expression (3.3)). This means that the

[5] The cascode configuration was used in early days of vacuum tube amplifiers to overcome the high plate-to-grid capacitance of triodes. The term was used for the first time in 1939 [27].

high input capacitance and high input conductance problems arising from the Miller effect are effectively eliminated.

- Consequently the input admittance is low (the input impedance is high) in a broader frequency range.
- Since the common-gate stage is driven by a high-impedance signal source; namely the output impedance of the common-source stage, the output internal impedance of the circuit is very high.
- Since the current gain of the common-gate circuit is equal to unity, the output current of the common-source stage is directly transferred to the load through the common-gate stage. Therefore, the overall voltage gain of the circuit is equal to the gain of the first stage, as if the load were directly connected to the output M1.

To investigate the properties of a cascode amplifier for a capacitive load in detail, the small-signal equivalent circuit shown in Fig. 3.21(b) will be used.

According to (3.25a), the input admittance of M2, which is the load of M1, is

$$y_{i2} = C_{gs2}\frac{(s - s_{01})(s - s_{02})}{(s - s_p)} \tag{3.29}$$

where

$$s_p = -\frac{G_L + g_{ds2}}{C'_L} \quad s_{01} \cong -\frac{g_{m2}}{C_{gs2}} \quad s_{02} \cong -\frac{G_L}{C'_L} \tag{3.30}$$

From (3.3), the voltage gain of M1 loaded with y_{i2} can be written as

$$A_{v1} = -\frac{g_{m1} - sC_{dg1}}{y_{i2}} = -C_{dg1}\frac{(s - s_{011})}{y_{i2}} \tag{3.31}$$

where, $s_{011} = -\,g_{m1}/C_{dg1}$. Combining (3.31) and (3.29) and re-naming the poles and zeros of A_{v1} to prevent any confusion, we obtain

$$A_{v1} = -\frac{C_{dg1}}{C_{gs2}}\frac{(s - s'_{01})(s - s'_{02})}{(s - s'_{p1})(s - s'_{p2})} \tag{3.32}$$

where

$$\begin{aligned} s'_{01} &= s_{011} = -\frac{g_{m1}}{C_{dg1}} \\ s'_{02} &= s_p = -\frac{G_L + g_{ds2}}{C'_L} \\ s'_{p1} &= s_{01} \cong -\frac{g_{m2}}{C_{gs2}} \\ s'_{p2} &= s_{02} \cong -\frac{G_L}{C'_L} \end{aligned} \tag{3.33a}$$

According to (3.21), the voltage gain of the common-gate stage, M2, is

$$A_{v2} = \frac{v_o}{v_{g2}} = \frac{g_{m2} + g_{ds2}}{C'_L}\frac{1}{(s - s'_{p3})} \cong \frac{g_{m2}}{C'_L}\frac{1}{(s - s'_{p3})} \tag{3.34}$$

where

$$s'_{p3} = -\frac{G_L + g_{ds2}}{C'_L} \tag{3.34b}$$

Now the overall voltage gain of the cascode circuit can be written as

$$A_v = A_{v1} \cdot A_{v2} = -\frac{C_{dg1}}{C_{gs2}} \frac{g_{m2}}{C'_L} \frac{(s - s'_{01})(s - s'_{02})}{(s - s'_{p1})(s - s'_{p2})(s - s'_{p3})} \tag{3.35a}$$

Noting that $s_{02}{}' = s_{p3}{}'$, the gain expression (3.35a) can be simplified as

$$A_v = -\frac{C_{dg1}}{C_{gs2}} \frac{g_{m2}}{C'_L} \frac{(s - s'_{02})}{(s - s'_{p1})(s - s'_{p2})} \tag{3.35b}$$

This expression can be interpreted as follows.

- The low-frequency value of the gain is

$$A_v(0) = -\frac{C_{dg1}}{C_{gs2}} \frac{g_{m2}}{C'_L} \frac{(-s'_{01})}{(-s'_{p1})(-s'_{p2})} = -\frac{g_{m1}}{G_L} \tag{3.36}$$

  If we compare this expression with (3.5), we can see that the low-frequency gain of a cascode circuit is greater than the gain of M1 would be if it were directly loaded with Y_L:

$$\frac{A_v(0)_{cascode}}{A_v(0)_{c.source}} = 1 + \frac{g_{ds1}}{G_L}$$

  This is the result of the very high output internal resistance of the common-gate stage, since it is driven by a high-impedance source (the output of the common-source stage).
- To evaluate the high-frequency performance of the circuit it is useful to compare the zero and the poles of (3.35a):

$$s'_{01} = s_{011} = -\frac{g_{m1}}{C_{dg1}}$$

$$s'_{p1} = s_{01} \cong -\frac{g_{m2}}{C_{gs2}}$$

$$s'_{p2} = s_{02} \cong -\frac{G_L}{C'_L}$$

  Since the zero is negative and obviously very high in magnitude compared to the magnitudes of s'_{p1} and s'_{p2}, its effect on the magnitude and phase of the gain is usually negligible. There are two poles affecting the magnitude and phase of the gain at high frequencies. Provided that $|s'_{p1}| \gg |s'_{p2}|$, the 3 dB frequency of the cascode circuit approaches to that of the simple common-source amplifier having the same output load:

$$f_{3dB} \cong \frac{G_L}{2\pi C'_L} \tag{3.37}$$

and the voltage-gain–bandwidth product becomes

$$GBW \cong \frac{g_{m1}}{2\pi C'_L}, \tag{3.38}$$

which is equal to that of a common-source amplifier.

- If the magnitudes of s'_{p1} and s'_{p2} are comparable, the high-frequency roll-off occurs earlier and at the 3 dB frequency the phase shift exceeds $\pi/4$. It means that owing to the additional pole corresponding to the input R-C of M2, the high-frequency performance of the cascode circuit deteriorates and the voltage-gain–bandwidth product decreases.

3.5 The CMOS inverter as a transimpedance amplifier

In Section 2.2, it was mentioned that a CMOS inverter containing a feedback resistor connected between the output and input nodes is suitable to be used as a transimpedance amplifier. The schematic diagram, the small-signal equivalent circuit and the modified equivalent circuit after the application of the Miller transformation are given in Fig. 3.22, where $\bar{g}_m$ and $\bar{g}_{ds}$ represent the sums of the corresponding parameters of M1 and M2 and C_i and C_o the total parallel capacitance to the input and output nodes, respectively. As explained earlier, the resistance connected between the output and input nodes helps to improve the stability of the operating point. But there is an unavoidable capacitance, namely the sum of the C_{dg} of M1 and M2, which is parallel to this resistor. During the analysis it will be assumed that there is a capacitance (C_F)

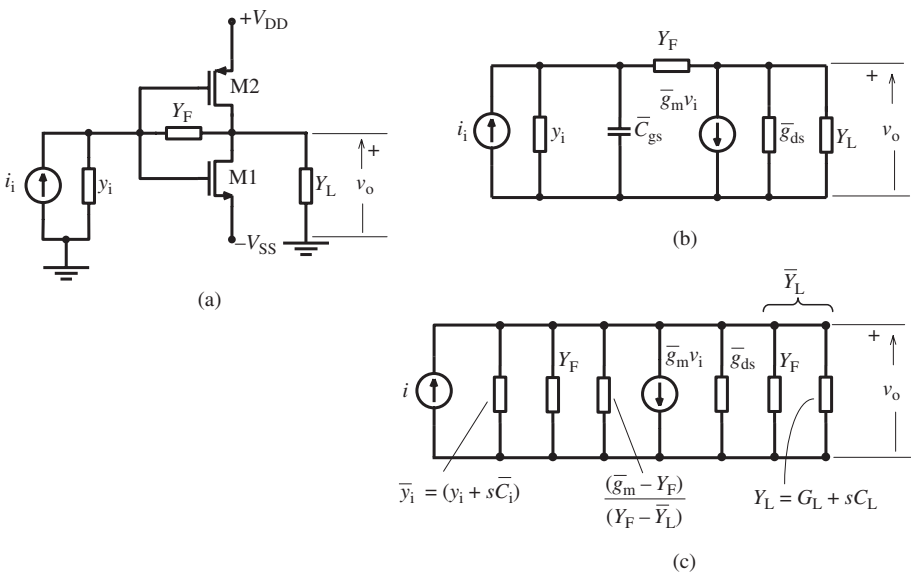


Figure 3.22 (a) The CMOS inverter as a transimpedance amplifier: the input signal is i_i (current) and the output signal is v_o (voltage). (b) The small-signal equivalent circuit. (c) The equivalent circuit after Miller transformation.

parallel to the feedback resistor, equal to the sum of the total drain–gate capacitance, the parasitic capacitance and the parallel external capacitance (if there is any).

From Fig. 3.22(c) the transimpedance can be calculated as

$$Z_{\mathrm{m}} = -\frac{(\bar{g}_{\mathrm{m}} - Y_{\mathrm{F}})}{Y_{\mathrm{F}}(\bar{g}_{\mathrm{m}} + \bar{y}_{\mathrm{i}}) + \bar{Y}_{\mathrm{L}}(\bar{y}_{\mathrm{i}} + Y_{\mathrm{F}})} \tag{3.39}$$

and arranged in the s domain as

$$Z_{\mathrm{m}} = \frac{sC_{\mathrm{F}} - (\bar{g}_{\mathrm{m}} - G_{\mathrm{F}})}{s^2(C_{\mathrm{i}}C_{\mathrm{o}} + C_{\mathrm{i}}C_{\mathrm{F}} + C_{\mathrm{o}}C_{\mathrm{F}}) + s[C_{\mathrm{F}}(g_{\mathrm{i}} + \bar{g}_{\mathrm{ds}} + \bar{g}_{\mathrm{m}}) + C_{\mathrm{i}}(\bar{g}_{\mathrm{ds}} + G_{\mathrm{F}}) + C_{\mathrm{o}}(g_{\mathrm{i}} + G_{\mathrm{F}})] + [G_{\mathrm{F}}(g_{\mathrm{i}} + \bar{g}_{\mathrm{ds}} + \bar{g}_{\mathrm{m}}) + g_{\mathrm{i}}\bar{g}_{\mathrm{ds}}]}$$

This expression can be simplified as

$$Z_{\mathrm{m}} \cong \frac{sC_{\mathrm{F}} - (\bar{g}_{\mathrm{m}} - G_{\mathrm{F}})}{s^2C_{\mathrm{o}}(C_{\mathrm{i}} + C_{\mathrm{F}}) + s[\bar{g}_{\mathrm{ds}}(C_{\mathrm{F}} + C_{\mathrm{i}}) + G_{\mathrm{F}}(C_{\mathrm{i}} + C_{\mathrm{o}}) + (\bar{g}_{\mathrm{m}}C_{\mathrm{F}})] + G_{\mathrm{F}}(\bar{g}_{\mathrm{ds}} + \bar{g}_{\mathrm{m}})} \tag{3.40}$$

The low-frequency value of the transimpedance (the transresistance) is

$$Z_{\mathrm{m}}(0) = -\frac{(\bar{g}_{\mathrm{m}} - G_{\mathrm{F}})}{G_{\mathrm{F}}(\bar{g}_{\mathrm{ds}} + \bar{g}_{\mathrm{m}})}$$

and reduces to $(1/G_{\mathrm{F}})$ for $\bar{g}_{\mathrm{ds}}$, $G_{\mathrm{F}} \ll \bar{g}_{\mathrm{m}}$, as already shown in Chapter 2.

The frequency characteristic of the amplifier depends on the relative positions of the zero and the two poles of this gain function. To investigate the possibilities, it is convenient to write the expression in a closed form as

$$Z_{\mathrm{m}} \cong \frac{s.D + E}{s^2.A + s.B + C}$$

The gain function has a positive-real zero:

$$s_{\mathrm{z}} = -\frac{E}{D} = \frac{(\bar{g}_{\mathrm{m}} - G_{\mathrm{F}})}{C_{\mathrm{F}}} \tag{3.41}$$

The poles can be solved as the roots of the quadratic nominator:

$$s_{\mathrm{p1,p2}} = \frac{B}{2A}\left(-1 \mp \sqrt{1 - \frac{4AC}{B^2}}\right) \tag{3.42}$$

These roots are:

$$\begin{aligned} &\text{(a) two separate negative-real poles for} && \frac{4AC}{B^2} < 1 \\ &\text{(b) two equal negative-real poles for} && \frac{4AC}{B^2} = 1 \\ &\text{(c) one complex-conjugate pair for} && \frac{4AC}{B^2} > 1 \end{aligned} \tag{3.43}$$

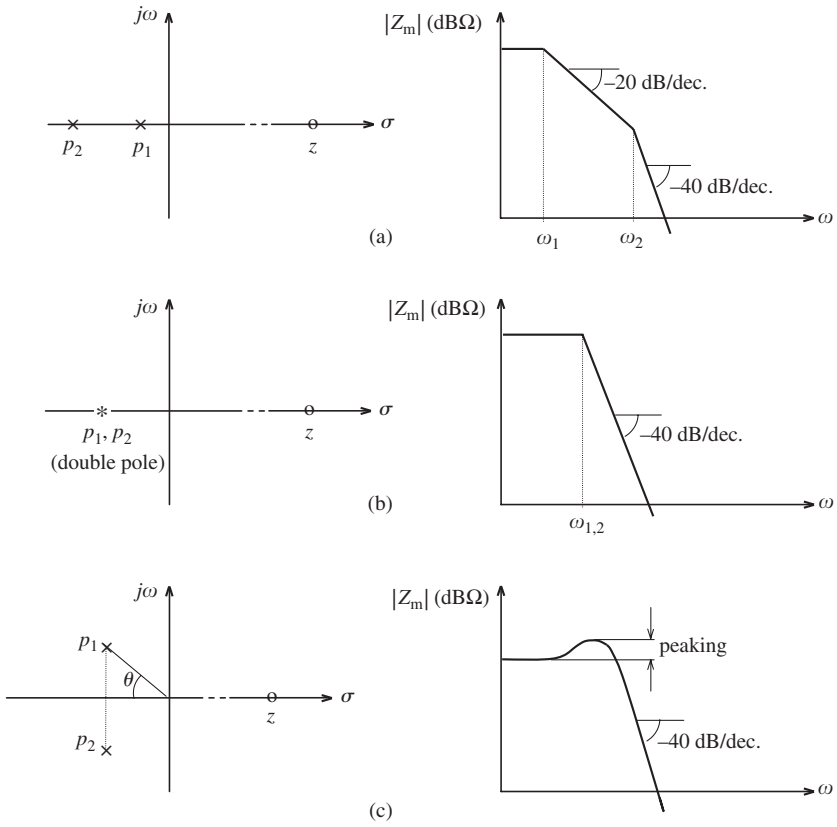


Figure 3.23 Possible pole–zero diagrams and the corresponding frequency characteristics (the magnitude of the zero is assumed much higher than the magnitude of the poles).

In Fig. 3.23, the pole–zero positioning for these cases and the corresponding shapes of the frequency characteristics are shown. The important points related to these three cases will be briefly discussed below.

(a) Two separate negative-real poles

In this case, provided that the two poles are sufficiently apart from each other and the right half-plane zero is far away, the 3 dB frequency is determined by the dominant pole, the slope of the magnitude curve is –20 dB/decade and the phase shift of the output signal does not exceed $\pi/2$, up to the vicinity of the second pole. It is known from feedback theory that, if this amplifier is in a negative feedback loop, one of the poles must dominate to guarantee the stability, i.e. the first pole frequency must be sufficiently smaller in magnitude than the second one. It can be shown that if the gain of an amplifier (A_o) is reduced to A_f by negative feedback, in order to have a flat final frequency response,

$$\frac{\omega_{p2}}{\omega_{p1}} \geq \frac{A_o}{A_f} - 1$$

must be satisfied, where ω_{p1} and ω_{p2} are the frequencies corresponding to the poles of the no-feedback (or open-loop) gain of the amplifier (see Chapter 4). It means that in the case of a strong negative feedback, to reduce the gain substantially, the magnitude of one of the poles must be much smaller than the other. For such a case $4AC/B^2 \ll$ and (3.42) can be written as

$$s_{p1,p2} \cong \frac{B}{2A}\left[-1 \mp \left(1 - \frac{1}{2}\frac{4AC}{B^2}\right)\right]$$

$$s_{p1} = -\frac{C}{B} \text{ (lower frequency "dominant" pole)}$$

$$s_{p2} = -\frac{B}{A} \text{ (higher frequency "far" pole)}$$

The values of these poles in terms of the circuit parameters can be calculated as

$$s_{p1} = -\frac{G_F(\bar{g}_{ds} + \bar{g}_m)}{C_F(\bar{g}_{ds} + \bar{g}_m) + C_o G_F} = -\frac{1}{\frac{C_F}{G_F} + \frac{C_o}{(\bar{g}_{ds}+\bar{g}_m)}} \tag{3.44}$$

$$s_{p2} = -\frac{C_F(\bar{g}_{ds} + \bar{g}_m) + G_F C_o}{C_o C_F} = -\left(\frac{(\bar{g}_{ds} + \bar{g}_m)}{C_o} + \frac{G_F}{C_F}\right) \tag{3.45}$$

(b) Two equal negative-real poles (double-poles)

In this case, the drop of the gain is 6 dB at the frequency corresponding to the double-pole[6] and the slope of the decrease is 40 dB/decade. This configuration is apparently not suitable to be used in a feedback loop, but can be used as a flat frequency response wide-band amplifier stage suitable to give a rail-to-rail output voltage.

Inserting the circuit parameters into (3.43(b)), the condition to obtain such a frequency characteristic and the magnitude of the frequency corresponding to the pole can be solved as

$$C_F \cong C_o \frac{G_F}{(\bar{g}_{ds} + \bar{g}_m)} \tag{3.46}$$

$$\omega_p = \frac{B}{2A} = \frac{\bar{g}_{ds}(C_F + C_i) + G_F C_o + \bar{g}_m C_F}{2C_o(C_i + C_F)} \tag{3.47}$$

(c) Complex-conjugate poles

The pole–zero diagram of an amplifier having two complex-conjugate poles and a positive-real zero is shown in Fig. 3.23(c). From network theory it is known that the shape of the frequency response depends on the θ angle. For $\theta > \pi/4$ (or $\xi = \cos\theta < 0.707$) the frequency characteristic exhibits a peak, and the 3 dB frequency slightly increases (Fig. 3.24(a)). The response of the amplifier to a step input signal correspondingly exhibits an over-shoot that also increases with θ (Fig. 3.24(b)),

[6] The 3 dB frequency of a multiple-pole amplifier can be calculated as $\omega_{3dB} = \omega_p\sqrt{2^{1/n} - 1}$, where n is the number of the coincident poles. In the case of $n = 2$, $\omega_{3dB} = 0.614\ \omega_p$.

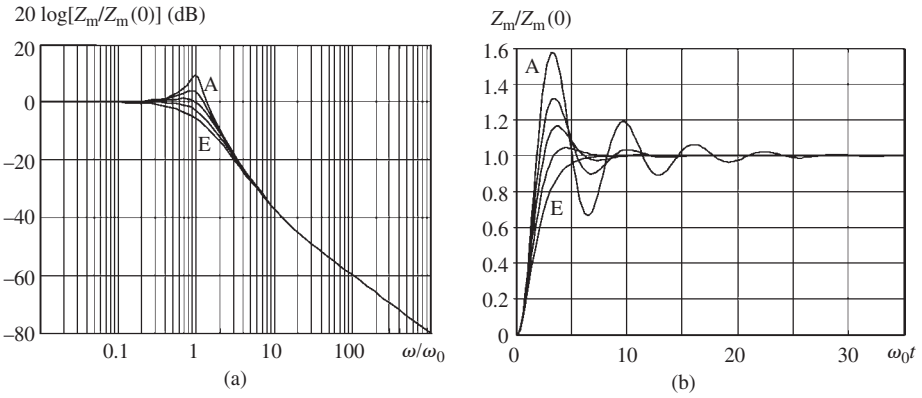


Figure 3.24 (a) The normalized frequency responses and (b) the normalized step responses for different values of θ. Curves A to E correspond to $\theta = 80$, 70, 60, 45 and 0 degrees, respectively. ($\omega_0^2 = C/A$. Curves are drawn for $\omega_z/\omega_0 = 10$. Note the decrease of the slope owing to the zero.)

and is considered harmful for many applications. The effects of the positioning of the complex-conjugate poles of an amplifier are investigated in several electronics textbooks in detail, and can be easily adapted to this case, whenever necessary.

Example 3.2 A transimpedance amplifier will be designed with a low-frequency "gain" of 2000 ohm (66 dBΩ), flat up to 1 GHz, delivering a rail-to rail output voltage swing to a $C_L = 1$ pF external capacitive load. The internal impedance of the driving current signal source is 10 k ohm parallel to 50 fF (the output impedance of the previous stage). The design will be made using the AMS 0.35 micron CMOS technology parameters. As one of the candidate configurations, we will design a CMOS inverter type circuit. Since any peaking on the frequency response (any over-shoot on the pulse response) is not acceptable, a double-pole solution is targeted. The design will be made using the expressions derived in this section, and then fine-tuned with PSpice.

The circuit diagram is as shown in Fig. 3.25. The input capacitance (C_i) in the expressions is the sum of the signal source internal capacitance (which is given as 50 fF) and the gate–source capacitances of the transistors (which are not known yet). Similarly, the output capacitance (C_o) is the sum of the load capacitance (which is given as 1 pF) and the junction capacitances of the transistors (which are not known). The only known component is G_F, which is approximately equal to $1/Z_m(0) = 0.5$ mS. Another aspect that must be kept in mind is that the width of the PMOS transistor must be approximately (μ_{no}/μ_{po}) times bigger than that of the NMOS transistor (see Chapter 2). This ratio is $475.8/137 \cong 3.5$ for AMS 0.35 micron CMOS technology (see Appendix A).

For maximum high-frequency performance, the gate lengths of both transistors will be chosen as 0.35 μm.

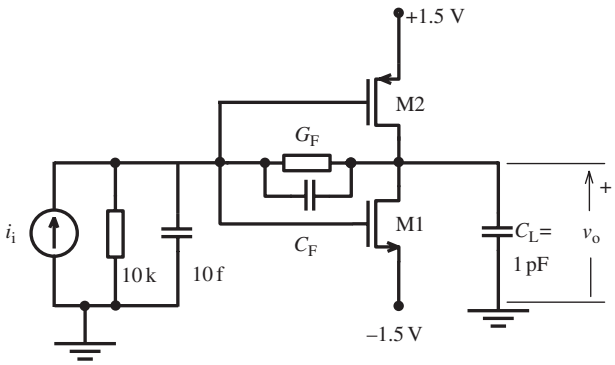


Figure 3.25 The circuit diagram of the transimpedance amplifier to be designed.

To estimate the dimensions of the transistors, expressions (3.46) and (3.47) can be used. If we insert (3.46) into (3.47) and make some simplifications assuming that $C_F < C_i$ and $\bar{g}_{ds} \ll \bar{g}_m Z$ (both being realistic), we obtain

$$\omega_p \cong \frac{\bar{g}_{ds}}{2C_o} + \frac{G_F}{C_i} \tag{3.48}$$

The bandwidth (3 dB frequency) was given as 1 GHz. Since the amplifier has a double-pole, the frequency corresponding to the pole is $f_p = 1\text{ GHz}/0.614 = 1.628$ GHz, or $\omega_p = 2\pi\ 1.628 \times 10^9 \cong 10^{10}$ rad/s. But since all parameters – except G_F – are geometry dependent, (3.48) does not lead to the solution.

Another attack point is to consider the slew-rate. The rise-time of a 1 GHz bandwidth amplifier is approximately $t_r = 0.35/f_{3dB} = 0.35/10^9 = 0.35$ ns (Chapter 2). For a rail-to-rail output step, the PMOS transistor that acts as a current source must be capable of charging the total load capacitance up to $V_{DD} = 1.5$ V in 0.35 ns:

$$I_P = C_o \frac{dV}{dT} = C_o \frac{1.5}{0.35 \times 10^{-9}} \tag{3.49}$$

C_o is the sum of the external load capacitance (1 pF) and the total parasitic capacitance of the output node. This parasitic capacitance can be calculated (see Appendix 1) in terms of the widths of the transistors as

$$C_{jDN}(0) = 0.8W_N + 0.5(X + W_N) \cong 1.3W_N \text{ [fF, } W_N \text{ in } \mu\text{m]}$$
$$C_{jDP}(0) = 0.8W_P + 0.5(X + W_P) \cong 1.3W_P \text{ [fF, } W_N \text{ in } \mu\text{m]}$$

and the total parasitic capacitance[7]

$$C_{jDT}(0) \cong 1.3(W_N + W_P) = 1.3(1 + 3.5)W_N = 5.85W_N \text{ [fF, } W_N \text{ in } \mu\text{m]}$$

[7] This is the maximum (pessimistic) value of the parasitic capacitance, since it corresponds to the zero bias voltage on the drain junctions. In reality the junction is reverse-biased and the parasitic capacitance is correspondingly smaller.

Now we can make an estimation for the total output capacitance and take $C_o = 1.2$ pF, which corresponds to $W_N = 34$ micron (which must be checked later on).

Now from (3.49), I_{DP} can be calculated as

$$|I_{DP}| = 1.2 \times 10^{-12} \frac{1.5}{0.35 \times 10^{-9}} = 5.14 \times 10^{-3} = 5.14\,\text{mA}$$

This is the drain current of the PMOS which flows under a gate voltage equal to $-V_{DD}$ (–1.5 V). Now the gate width of the PMOS transistor can be calculated:

$$|I_{DP}| \cong \frac{1}{2}\mu_p C_{ox} \frac{W_P}{L}(-V_{DD} - V_{TP})^2(1 + \lambda_P V_{DD})$$

$$5.14 \times 10^3 = \frac{1}{2}137(4.56 \times 10^{-7})\frac{W_P}{0.35}(-1.5 + 0.7)^2(1 + 0.2 \times 1.5)$$

$$W_P \cong 70\,\mu\text{m}$$

The corresponding NMOS transistor channel width to conduct the same current under quiescent conditions can be calculated from (2.14) as $W_N \cong 15\,\mu$m.

We know that the hand-calculated current values corresponding to a certain bias condition are usually higher than the currents in reality. Therefore it is necessary to check the drain currents with DC simulations and fine-tune to the target values and adjust the quiescent voltage of the output node, as performed in Example 4.3. The PSpice simulation shows that the drain currents for the calculated channel widths are 1.93 mA, which is considerably smaller than the 5.14 mA target value. A tuning procedure as described before gives $W_1 = 45\,\mu$m and $W_2 = 154\,\mu$m. Hence, the DC currents become equal to the target value and the DC voltage of the output node becomes zero (i.e. no offset).

Now, as a final step, the value of C_F can be calculated from (3.46). But before this step, C_o, $\bar{g}_m$ and $\bar{g}_{ds}$ must be derived.

The output parasitic capacitance, which is the sum of the junction capacitances of the drain regions, can be found for the calculated dimensions as:

$$C_{jDT}(0) \cong 5.85 W_N = 5.85 \times 45 \cong 263\ \text{fF}\ C_o = C_L + C_{jDT} = 1.263\,\text{fF}$$

$\bar{g}_m$ is the sum of the transconductances and can be calculated as

$$\bar{g}_m = g_{mN} + g_{mP} = \sqrt{2\frac{W_1}{L}\mu_n C_{ox} I_D} + \sqrt{2\frac{W_2}{L}\mu_p C_{ox} I_D} = 21.1\,\text{mS}$$

$\bar{g}_{ds}$ is the sum of the output conductances of the transistors:

$$\bar{g}_{ds} \cong I_D(\lambda_N + \lambda_P) = 5.14 \times 10^{-3}(0.073 + 0.2) = 1.4\,\text{mS}$$

Now the value of C_F can be calculated from (3.46):

$$C_F = 1.263 \frac{0.5 \times 10^{-3}}{22.6 \times 10^{-3}} = 0.028\,\text{pF} = 18\,\text{fF}$$

It must be noted that the sum of the drain-to-source capacitances is the intrinsic component of C_F. If we calculate $(C_{dgN}+C_{dgP})$ we find that

$$(C_{dgN} + C_{dgP}) = (W_N + W_P)C_{DGO} = (45 + 154)0.12 = 23.9\,\text{fF}$$

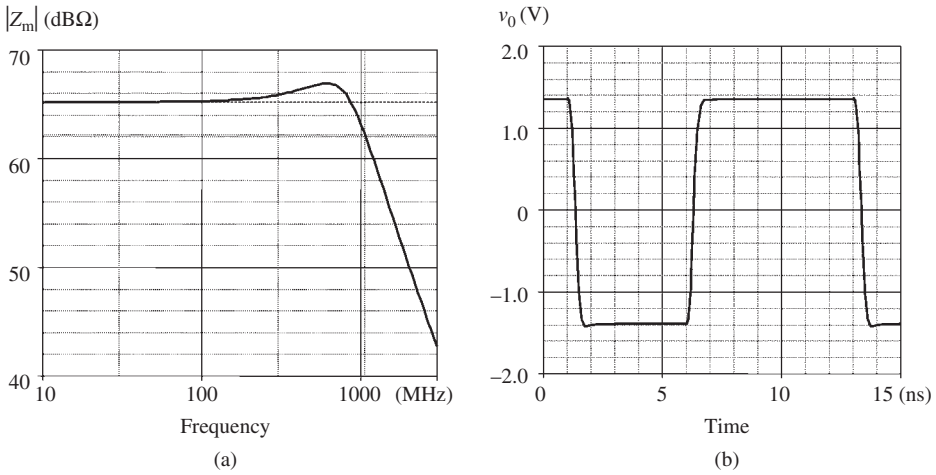


Figure 3.26 (a) Frequency response, (b) pulse response of the designed transimpedance amplifier.

This means that owing to the intrinsic C_F, the circuit is slightly over-compensated. The PSpice simulation results with the calculated design parameters are given in Fig. 3.26. Obviously there are certain discrepancies from the target values. Let us interpret the results and obtain hints for the further fine-tuning of the design.

From the frequency characteristic given in Fig. 3.26(a), we see the following.

(a) The low-frequency gain is 65.2 dBΩ. If this 0.8 dBΩ discrepancy is not tolerable, R_F can be increased to reach the target value. But (3.48) tells us that the increase of R_F decreases the pole frequency, and consequently the 3 dB frequency.

(b) The 3 dB frequency is 1.05 GHz, but there is a 1.7 dB peak, which may not be acceptable for some applications. To decrease the peaking, C_F must be increased (see (3.40) and (3.43)). Indeed, if we use a feedback capacitor equal to $C_F = 50$ fF, the peak disappears, the response becomes flat, but the 3 dB frequency decreases to 795 MHz. To increase the 3 dB frequency, according to (3.48) R_F must be decreased. Observations in (a) and (b) indicate that there is a trade-off problem.

(c) If the magnitude of the gain has the prime importance, the gain can be increased to the target value with $R_F = 2.2\,k\Omega$. The frequency response is flat and the 3 dB frequency is 764 MHz for $C_F = 45$ fF, has 1 dB peak and 935 MHz bandwidth for $C_F = 12$ fF.

(d) If a flat frequency response with 1 GHz bandwidth has prime importance, this can be obtained at the price of gain decrease. For $R_F = 1.2\,k\Omega$ and $C_F = 60$ fF, the gain is 60.6 dBΩ and the 3 dB frequency is 1.004 GHz, with 0.2 dB peak, which is acceptable for many applications.

(e) If both the bandwidth and the gain conditions have to be fulfilled without any compromise from the original specifications, the DC current must be increased or the circuit must be designed using an alternative technology with smaller feature sizes to decrease the parasitic capacitances.

(f) The pulse response given in Fig. 3.26(b) shows that the circuit has no slew-rate limitation for the original design (the rise-time is 0.35 ns) and reasonably nice wave shape. The PSpice netlist for the original design is given below.

```
*AMS-035u INVERTER TRANSIMPEDANCE AMPLIFIER*
  vdd 10 0 1.5
  vss 20 0 -1.5
  cin 2 0 50f
  rin 2 0 10k
  *iin 0 2 pulse −1m 1m 1n 10p 10p 5n 12n
  iin 0 2 ac 1u
  M1 3 2 20 20 modn w=45U l=.35U, ad=38.2e-12 as=38.2e-12, pd=90e-6,
ps=90e-6
  M2 3 2 10 10 modp w=154U l=.35U, ad=131e-12 as=131e-12, pd=308e-6,
ps=308e-6
  RF2 3 2 2k
  *vx 3 0 0
  *cf 3 2 15f
  CL 3 0 1P
  .lib "cmos7tm.mod"
  .ac dec 50 .01g 3G
  .DC iin −1m 1m 10u
  .TRAN .01N 15N
  .probe
  .end
```

3.6 MOS transistor with source degeneration at high frequencies

In Section 2.1, we saw that a resistor in series to the source terminal of a MOS transistor decreases the transconductance to so-called "effective transconductance". It is useful to generalize this effect for an impedance placed in series with the source terminal and to deal not only with the transconductance, but with all of the *y* parameters. In Fig. 3.27(a), a MOS transistor is shown, together with the impedance (the source degeneration impedance) connected in series to the source, which can be external, or the parasitic (internal) impedance of the transistor. Now we will calculate the *y* parameters of the "equivalent transistor" M_e shown in Fig. 3.27(b), in terms of the small-signal parameters of M and its source impedance, Z_s. For the sake of the simplicity of the derivations, C_{dg} will be excluded from the *y*-parameter calculations, and it can be re-inserted whenever necessary.[8]

The y_{11} and y_{21} parameters, both defined under shorted output conditions, can be solved from Fig. 3.28(a):

[8] Hint: the total equivalent *y* parameters of two parallel connected two-ports correspond to the sums of the *y* parameters of the individual two-ports.

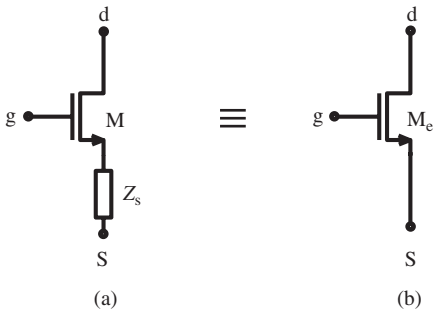


Figure 3.27 (a) A MOS transistor with source degeneration impedance. (b) The equivalent transistor.

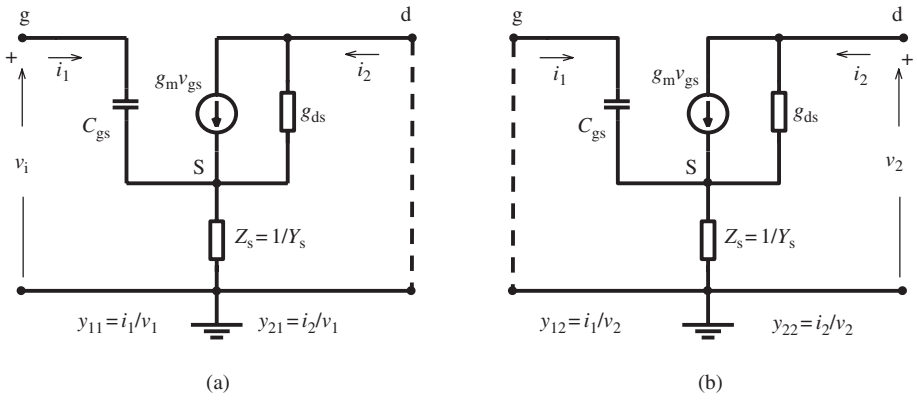


Figure 3.28 Small-signal equivalent circuits used to derive expressions (a) for y_{11} and y_{21}, (b) for y_{12} and y_{22}.

$$y_{11} \cong \frac{sC_{gs}}{1 + (g_m + sC_{gs})Z_s} \tag{3.50}$$

$$y_{21} \cong g_m \frac{1}{1 + g_m Z_s} \tag{3.51}$$

Similarly, the y_{12} and y_{22} parameters, both defined under shorted input conditions, can be derived from Fig. 3.28(b):

$$y_{12} \cong -\frac{sC_{gs}g_{ds}}{sC_{gs} + Y_s + g_m} \tag{3.52}$$

$$y_{22} \cong \frac{g_{ds}}{1 + \frac{g_m}{(sC_{gs} + Y_s)}} \tag{3.53}$$

These expressions are valid for any type of source degeneration impedance. For an inductance (L_s) connected in series to the source, e.g. as shown in Fig. 3.29 (assuming that the effect of y_{12} is negligible), the input admittance can be written as

$$y_{in} \cong y_{11} \cong \frac{sC_{gs}}{1 + (g_m + sC_{gs})sL_s}$$

$$z_{in} = \frac{1}{y_{in}} = \frac{1 + (g_m + sC_{gs})sL_s}{sC_{gs}} = \frac{1}{sC_{gs}} + \frac{g_m L_s}{C_{gs}} + sL_s \tag{3.54}$$

which represents a series combination of C_{gs}, L_s and a frequency-independent resistive component. For a specified frequency, the capacitive component can be eliminated by resonance with an appropriate inductance connected in series to the gate. This is a useful possibility for the design of resistive, low-input impedance circuits like LNAs, which will be investigated in Chapter 4 in further detail.

Problem 3.1 *The circuit shown in Fig. 3.29(a) will be used as the input stage of an amplifier operating at 2 GHz. The parameters of the transistor are given as:* $\mu = 300$ $cm^2/V.s$, $T_{ox} = 6$ *nm, CDGO* $= 1.2 \times 10^{-10}$ F/m, $L = 0.25$ *and* $W = 50$ μm. *The transistor is operating in saturation and biased for* $I_D = 2$ mA. *The quality factors of inductors on the chip are* $Q = 10$ *at* 2 GHz.

(a) Calculate the transconductance of the transistor.
(b) Calculate the value of the input capacitance of the transistor.
(c) To make the real part of the input impedance equal to 100 ohms, *calculate the inductance of the inductor.*
(d) Calculate the value of the imaginary part of the input impedance and the value of the corresponding reactive element.
(e) Calculate the value of the reactance to be connected in series to the gate to make the input impedance pure resistive.

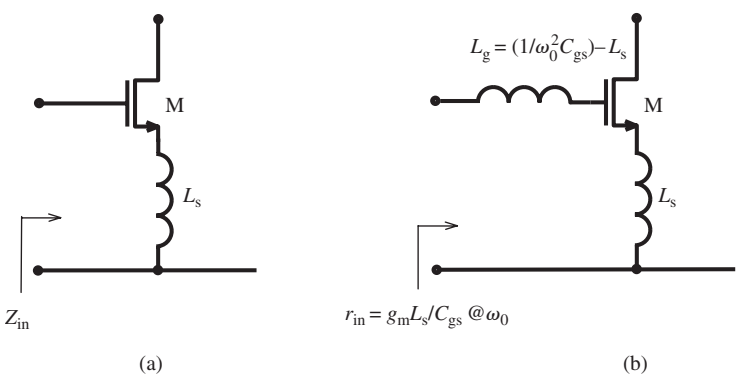


Figure 3.29 (a) Inductive source degeneration. (b) Arrangement to obtain a resistive and small input impedance for a certain frequency.

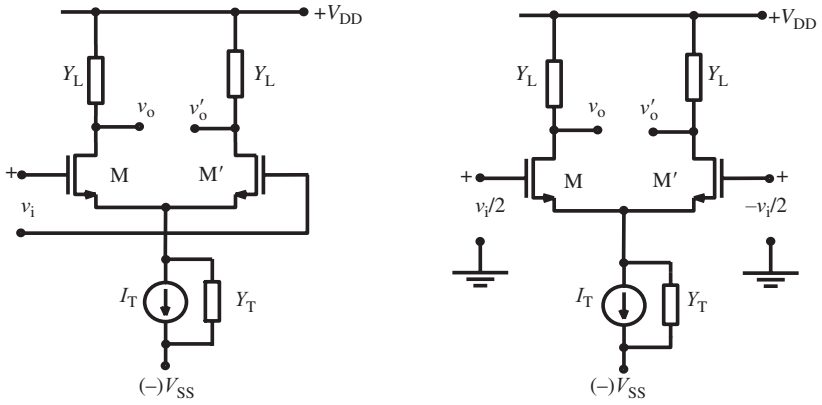


Figure 3.30 (a) Circuit diagram of a differentially driven long tailed pair. (b) Circuit with the equivalent individual single-ended input signals.

3.7 High-frequency behavior of differential amplifiers

In Chapter 2 we have seen that several varieties of differential amplifiers were developed over time. In this section we will investigate the frequency-dependent behavior of the three most important and basic differential amplifier configurations; the R-C loaded long tailed pair, the fully differential active loaded pair and the active current mirror loaded differential input–single ended output long tailed pair.

3.7.1 The R-C loaded long tailed pair

The circuit diagram of a long tailed pair loaded with equal impedances is given in Fig. 3.30(a). The transistors M and M′ are matched devices, therefore all corresponding parameters are assumed to be equal. The admittance Y_T represents the internal admittance of the tail DC current source. The circuit is differentially driven by a voltage source, v_i. Owing to the complete symmetry of the circuit, the two inputs are assumed to be driven by $v_i/2$ and $-v_i/2$, respectively (see Chapter 2), as shown in Fig. 3.30(b). The individual output voltages of M and M′ are shown by v_o and v_o'. Therefore the differential output voltage is $v_{od} = (v_o - v_o')$.

The small-signal equivalent of the circuit is given in Fig. 3.31(a). Since the signal voltage on the common-source node is zero owing to the symmetry of the circuit, it is possible to deal with each of the symmetrical halves separately, as shown in Fig. 3.31(b). In Fig. 3.31(c) this equivalent circuit is re-drawn to facilitate the network equations.[9]

From Fig. 3.31(c) the output voltage and the voltage gain from one of the inputs to the output of the corresponding transistor can be easily written as

[9] Note that C_{gs} is excluded from Fig. 3.31(b) since it is parallel to a voltage source.

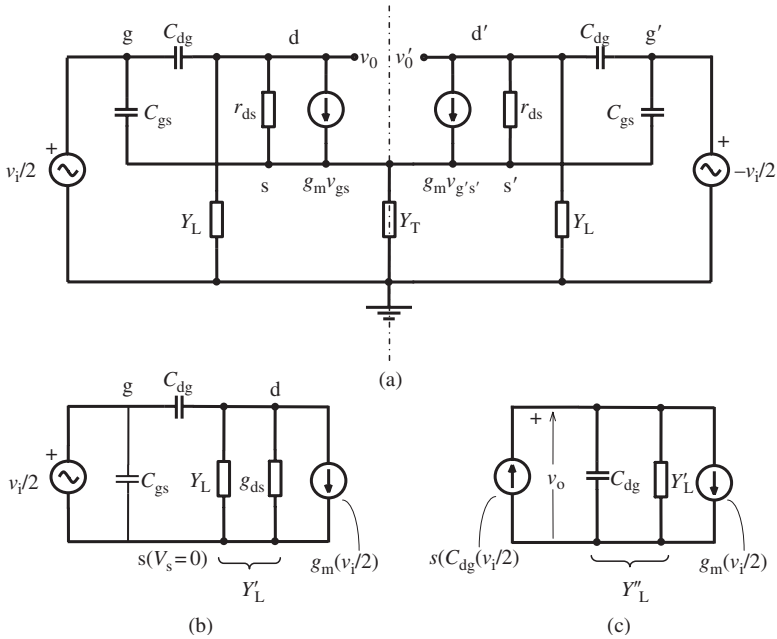


Figure 3.31 (a) The small-signal equivalent circuit. (b) Simplification based on the symmetry of the circuit. (c) The equivalent circuit after a Norton–Thévenin conversion.

$$v_o = \frac{v_i}{2}\left(sC_{dg} - g_m\right)\frac{1}{Y''}$$
$$A_v = \frac{v_o}{(v_i/2)} = \left(sC_{dg} - g_m\right)\frac{1}{Y''} \tag{3.55}$$

For an R-C load as $Y_L = G_L + sC_L$, $Y_L^{''}$ becomes

$$Y_L' = (G_L + sC_L) + g_{ds}$$
$$Y_L'' = Y_L' + sC_{dg} = (G_L + g_{ds}) + s(C_L + C_{dg})$$

and the voltage gain

$$A_v = \frac{\left(sC_{dg} - g_m\right)}{(G_L + g_{ds}) + s(C_L + C_{dg})} \tag{3.56}$$

which can be written in terms of its pole and zero as

$$A_v = \frac{C_{dg}}{(C_L + C_{dg})}\frac{(s - s_0)}{(s - s_p)} \tag{3.57}$$

where

$$s_p = -\frac{(G_L + g_{ds})}{(C_L + C_{dg})} \quad \text{and} \quad s_0 = +\frac{g_m}{C_{dg}} \tag{3.58}$$

and the low-frequency gain becomes

$$A_v(0) = \frac{C_{dg}}{(C_L + C_{dg})} \times \frac{-s_0}{-s_p} = -\frac{g_m}{(G_L + g_{ds})}$$

as expected (see Chapter 2). It is obvious that if series (and equal) resistors are connected to the sources to improve the linearity as mentioned in Chapter 2, g_{meff} must be used instead of g_m in the gain expressions.

Expression (3.57) is the voltage gain from the input of M which is driven by $(v_1/2)$ to its output. Similarly, the voltage gain of M′ is

$$A'_v = \frac{v'_o}{-(v_i/2)} = \frac{C_{dg}}{(C_L + C_{dg})}\frac{(s - s_0)}{(s - s_p)} \tag{3.59}$$

Now the voltage gain from the differential input (v_i) to the differential output can be written as

$$A_{vdd} = \frac{(v_o - v'_o)}{(v_i/2) - (-v_i/2)} = \frac{v_{od}}{v_i} = \frac{C_{dg}}{(C_L + C_{dg})}\frac{(s - s_0)}{(s - s_p)} \tag{3.60}$$

Expression (3.60) indicates that:

- an R-C loaded long tailed pair has a pole that is determined by the capacitance and the total equivalent conductance parallel to each output node. The frequency corresponding to this pole is the 3 dB frequency of the voltage gain:

$$f_{3dB} = \frac{1}{2\pi}\frac{(G_L + g_{ds})}{(C_L + C_{dg})} \tag{3.61}$$

- it must be noted the drain junction capacitances (C_{dj}) are parallel to the external load capacitances (C_L). For small external load capacitances, C_{dj} must be taken into account and the effective load capacitance must be calculated as $C_L + C_{dg} + C_{dj}$;
- the gain function has a positive zero at very high frequencies. The magnitude of this zero can be calculated as

$$s_0 = \frac{g_m}{C_{dg}} = \mu\frac{1}{L}\frac{C_{ox}}{C_{DGO}}(V_{GS} - V_T) \tag{3.62}$$

which is independent of the gate width. For example, with AMS 0.35 micron technology and 0.1 V gate overdrive, (3.62) gives the frequency corresponding to this pole as 82.2 GHz, which is too high to be taken into account.

As a by-product of these calculations, the signal current flowing through the total load admittance of one of the transistors, which is nothing else but the current of the voltage-controlled current source in the small-signal equivalent circuit, can be written as

$$i_s = -v_o[(G_L + g_{ds}) + s(C_L + C_{dg})] \tag{3.63a}$$

From (3.56) and (3.63a), i_s can be solved and simplified as

$$i_s = -(v_i/2)C_{dg}(s - s_0) \tag{3.63b}$$

This current can be assumed constant (frequency-independent) and equal to

$$i_s \cong i_s(0) = \frac{v_i}{2} g_m \tag{3.63c}$$

since for all practical cases $\omega < \omega_0$. This knowledge is valuable and will be used for the investigation of current-mirror loaded circuits. It is also useful to note that for arbitrary loads – for example RLC (tuned) loads – the output voltage can be calculated simply as $v_o = -i_s Z_L$.

3.7.2 The fully differential, current-mirror loaded amplifier

As already discussed in Chapter 2, one of the most important versions of differential amplifiers is the fully differential, current-mirror loaded amplifier shown in Fig. 3.32. The circuit can be considered either as a voltage amplifier when loaded with equal impedances, or as a transconductance (transadmittance) amplifier when loaded with impedances sufficiently smaller than the output internal impedance (see Chapter 2).

The DC operation of the circuit was investigated in Chapter 2. Now we will calculate the output currents of the circuit under a differential input voltage. According to (3.63b) the drain current is

$$i_1 = \frac{v_1}{2} g_{m1}$$

which is the input signal current of the current mirror (M2, M3). We know that the output signal current of this mirror can be expressed as (see Appendix D)

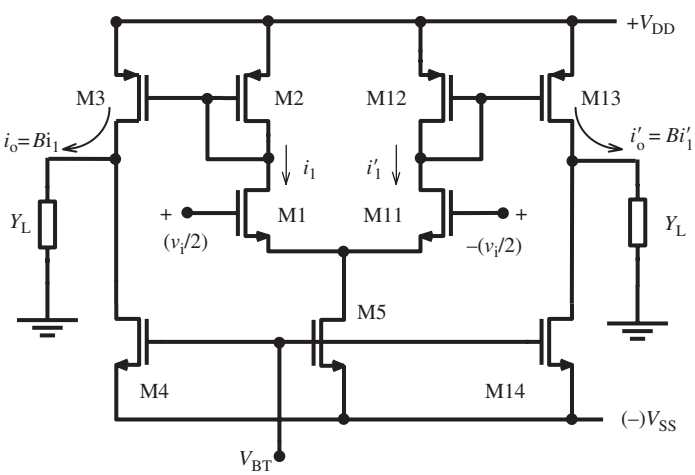


Figure 3.32 Circuit diagram of a fully differential, current-mirror loaded amplifier.

$$i_3 = i_1 B = i_1 B_0 \frac{s_p}{s_0} \frac{(s - s_0)}{(s - s_p)} \tag{3.64}$$

where B_0 is the DC (or low-frequency) current transfer ratio, which is equal to W_3/W_2 and s_0 and s_p are the zero and pole of B:

$$s_0 = +\frac{g_{m3}}{C_{dg3}} \quad \text{and} \quad s_p = -\frac{g_{m2}}{(C_{gs2} + C_{gs3}) + C_{xp}} \tag{3.65}$$

where C_{xp} represents the extra parasitics, the sum of the drain junction capacitances of M1 and M2 and the gate-to-source overlap capacitances of M2 and M3. M4 is a DC current source, biased by V_B. Since M4 is biased such that its DC current is equal to the DC current of M3, and M4 has a high small-signal output impedance, the output current (i_o) flowing through the low-impedance load is equal to i_3. From (3.63b) and (3.64) the output current can be expressed in terms of the input signal voltage, and then the transadmittance can be written as

$$y_m = \frac{i_o}{v_1} = \frac{1}{2} g_{m1} B_0 \frac{s_p}{s_0} \frac{(s - s_0)}{(s - s_p)}$$

and since $|s_0| \gg |s_p|$,

$$y_m \cong \frac{1}{2} g_{m1} B_0 s_p \frac{1}{(s - s_p)} \cong g_m s_p \frac{1}{(s - s_p)} \tag{3.66}$$

where g_m is the low-frequency value of y_m.

This is a typical single-pole gain function where the magnitude characteristic has a 3 dB frequency corresponding to the pole, and a –20 dB/decade slope. The phase characteristic reaches $\pi/4$ at the 3 dB frequency, and theoretically becomes asymptotic to $\pi/2$. But at the high-frequency end of the characteristic, owing to the effects of the neglected zero or the neglected parasitics, the phase shift can exceed $\pi/2$ and can cause problems if the amplifier is in a feedback loop.

If we calculate the 3 dB frequency of the amplifier from (3.65) as

$$f_{3dB} = \frac{1}{2\pi} \frac{g_{m2}}{(1 + B_0) C_{gs2} + C_{xp}} = \frac{1}{2\pi} \frac{\mu_p C_{ox} (W_2/L) |V_{GS2} - V_{TP}|}{\frac{2}{3}(1 + B_0) W_2 L C_{ox} + C_{xp}} \tag{3.67a}$$

and assume $C_{xp} < (1 + B_0) C_{gs2}$, we obtain:

$$f_{3dB} \approx \frac{1}{2\pi} \frac{\mu_p}{(1 + B_0) L^2} (|V_{GS2} - V_{TP}|) \tag{3.67b}$$

Although this assumption has limited validity for very small geometries, it provides a valuable design hint: using NMOS current mirrors for M2 and M3 and, consequently, using matched PMOS transistors as the input pair results in an improvement of the 3 dB frequency, by approximately (μ_n/μ_p) times, for comparable gate overdrive voltage values. In addition, using a PMOS input stage is usually more advantageous from the noise point of view.

Equation (3.67a) also indicates that the bandwidth decreases with B_0. On the other hand, as explained in Chapter 2 and seen from (3.66), the low-frequency value of the transadmittance is proportional to B_0. For the design of a high-gain, high-bandwidth transadmittance amplifier, there is a trade-off related to B_0, in terms of the total power consumption, total area consumption, the low-frequency value of the transadmittance and the bandwidth.

The output voltages on equal output loads can be calculated as

$$v_o = \frac{i_o}{Y_L} = v_i \frac{y_m}{Y_L} \quad v'_o = \frac{i'_o}{Y_L} = \frac{-i_o}{Y_L} = -v_1 \frac{y_m}{Y_L}$$

where the differential output voltage corresponding to a differential input is found to be:

$$v_{odd} = (v_o - v'_o) = 2v_i \frac{y_m}{Y_L} = v_i \frac{1}{Y_L} g_{m1} B_o s_p \frac{1}{(s - s_p)}$$

Thus, the differential-input-to-differential-output voltage gain can be written as

$$A_{vdd} = \frac{1}{Y_L} g_{m1} B_o s_p \frac{1}{(s - s_p)} \tag{3.68}$$

For example, with an R-C load of $Y_L = G_L + sC_L$ the differential voltage gain becomes

$$A_{vdd} = \frac{1}{C_L} g_{m1} B_o s_p \frac{1}{(s - s_p)(s - s_{pL})} \tag{3.69a}$$

where

$$s_{pL} = -\frac{G_L}{C_L} \tag{3.69b}$$

Consequently, the R-C loaded fully differential amplifier has two negative-real poles that influence the frequency characteristics. In most cases s_{pL} dominates and the gain becomes

$$A_{vdd} \cong \frac{1}{C_L} g_{m1} B_o \frac{1}{(s - s_{pL})} \tag{3.70}$$

which is safe for feedback applications. However, as mentioned in Chapter 2, it is usually necessary to apply a common-mode feedback (CMFB) in such circuits to maintain the stability of the operating points. Although this feedback does not directly involve the differential signal path, additional phase shifts on the differential signal may occur owing to the unavoidable parasitics, leading to unexpected stability problems (i.e. oscillations).

Example 3.3 Let us calculate the 3 dB frequency of the differential OTA (Fig. 2.20), designed in Chapter 2. The design goals were: transconductance $G_m = 3$ mS, $B_o = 1$ and the DC current of the input transistors $I_{D1} = 1$ mA (consequently the total current consumption, 4 mA). The calculated dimensions (in micrometers) for 0.35 micron AMS CMOS technology were

```
M1, M11: 24/0.35,
  M2, M12: 50/0.35,
  M4, M14: 48/0.35,
  M5, M15: 111/1,
  M3: 230/1.
```

From (3.65), the 3 dB frequency (frequency corresponding to the pole of the gain function) can be written as

$$f_{3\text{dB}} = \frac{1}{2\pi}\frac{g_{\text{m2}}}{(C_{\text{gs2}} + C_{\text{gs3}}) + C_{\text{xp}}}$$

$$g_{\text{m2}} = \sqrt{2(KP)_{\text{p}}(W/L)_2|I_{\text{D2}}|}, (KP)_{\text{p}}$$

$$= \mu_{\text{p}}C_{\text{ox}} = 137 \times 4.56 \times 10^{-7} = 0.624 \times 10^{-4}\ [\text{A/V}^2]$$

$$g_{\text{m2}} = \sqrt{2 \times (0.624 \times 10^{-4}) \times (50/0.35) \times 1} = 4.22 \times 10^{-3}\text{S} = 4.22\ \text{mS},$$

$$(C_{\text{gs2}} + C_{\text{gs4}}) = \frac{2}{3}(W_2 + W_4)LC_{\text{ox}} + (W_2 + W_4)\text{CDGO}$$

Inserting dimensions in micrometers, specific capacitances in fF/μm^2 and fF/μm:

$$(C_{\text{gs2}} + C_{\text{gs4}}) = \frac{2}{3}(50 + 48) \times 0.35 \times 4.56 + (50 + 48) \times 0.12 \cong 110\ \text{fF}$$

$$C_{\text{xp}} = (W_1 + W_2)XC_{\text{j}} + (W_1 + W_2 + 2X)C_{\text{jsw}} + (W_1 + W_2 + W_4)\text{CDGO}$$

$$C_{\text{xp}} = (24 + 50) \times 0.85 \times 0.94 + 2 \times (24 + 50 + 1.7) \times 0.25$$

$$+(24 + 50 + 48) \times 0.12 \cong 111.6\ \text{fF}$$

$$f_{3\text{dB}} = \frac{1}{2\pi}\frac{4.3 \times 10^{-3}}{(110 + 111.6) \times 10^{-15}} = 3.09\ \text{GHz}$$

The frequency characteristic obtained with PSpice simulation is shown in Fig. 3.33. The hand calculation and simulation results for the low-frequency value of the transconductance are in perfect agreement (3 mS and 3.02 mS, respectively). The

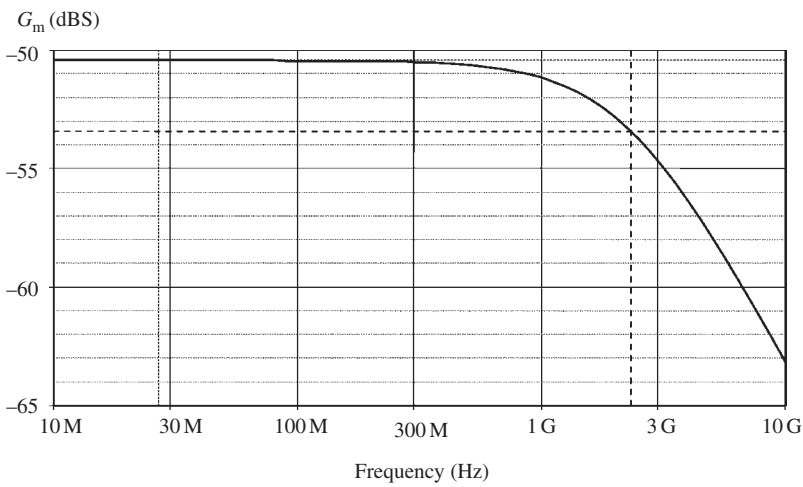
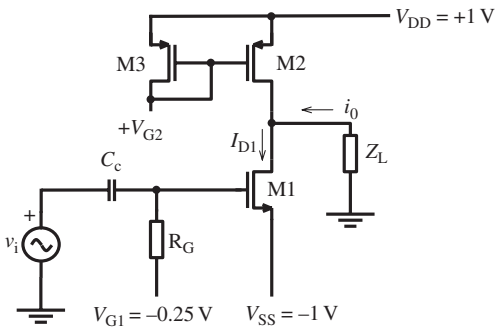


Figure 3.33 Simulated frequency response of the amplifier. The transconductance, expressed in dBS=20 log (i_{o1}/v_1), is −50.448 dBS, which corresponds to 3.02 mS. The 3 dB frequency is 2.36 GHz and the slope is less than –20 dB/dec, up to 10 GHz.

disagreement of the 3 dB frequency (3.09 GHz vs. 2.36 GHz) is mainly owing to the approximations in the analytical expressions and the neglected parasitics.

Problem 3.2 *A simple, single-ended transadmittance amplifier is shown in the figure. The parameters of the transistors are given as follows:*



$L = 0.25\,\mu$m, $T_{ox} = 5.4$ nm,
NMOS: $V_T = 0.45$ V, CDGO = CGSO$= 2.5\times10^{-10}$ F/m, $C_j = 1.8\times10^{-3}$ F/m^2, $C_{jsw} = 4.2\times10^{-10}$ F/m, $X = 0.5\,\mu$m.
PMOS: $V_T = -0.45$ V, CDGO = CGSO $= 4.5\times10^{-10}$ F/m, $C_j = 1.8\times10^{-3}$ F/m^2, $C_{jsw} = 3.5\times10^{-10}$ F/m, $X = 0.5\,\mu$m.
The PMOS current mirror is biased for 1 mA *drain current.*

(a) Find the width of M1 *for* $i_o = 0$ *at quiescent point* ($v_i = 0$).
(b) Z_L *is very small compared to the output internal impedance of the circuit. Calculate the low-frequency value of the transadmittance.*
(c) Calculate the 3 dB *frequency of the transadmittance.*

3.7.3 Frequency response of a single-ended output long tailed pair

The circuit schematic of an NMOS differential pair loaded with a PMOS unity gain current mirror is given in Fig. 3.34. This type of differential amplifier is mostly used as the input stage of operational amplifiers, where the subsequent stage is usually a common-source gain stage. Therefore, the load of the single-ended input stage is the capacitive input admittance of the gain stage. Although the single-ended long tailed pair behaves as a transconductor, we will prefer to consider it as a voltage amplifier owing to this commonly used configuration.

We know that under a small differential AC input signal, the resulting AC drain current components of M1 and M11 (i_1 and i_{11}) are equal in magnitude and opposite in phase. The drain current of M2, which is i_1, is mirrored as $i_{12}=Bi_1$, where B is the

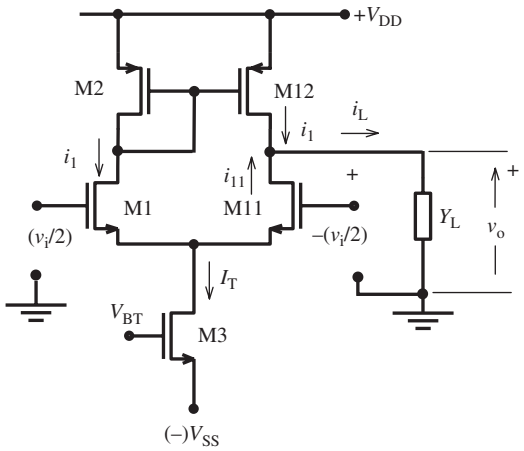


Figure 3.34 NMOS differential amplifier with single-ended output, loaded with PMOS current mirror.

frequency-dependent mirroring factor (current gain) of the M2–M12 current mirror. Hence the total load current can be written as

$$i_L = i_1 + i_{12} = i_1(B+1) \tag{3.71}$$

The drain signal current of M1 is $i_1 = g_{m1}v_{gs1}$. Since $v_{gs1} = (v_i/2)$ under small-signal conditions, the load current, the output voltage and the small-signal voltage gain can be written as

$$i_L = \frac{1}{2}v_i g_{m1}(B+1) \quad v_o = \frac{i_L}{Y_L} \quad A_v = \frac{v_o}{v_i} = \frac{1}{2}\frac{g_{m1}}{Y_L}(B+1) \tag{3.72}$$

The current gain of a current mirror can be expressed as (see Appendix D)

$$B = B_0\frac{s_p}{s_0}\frac{(s-s_0)}{(s-s_p)} \tag{3.73a}$$

where B_0 is the low-frequency current transfer ratio that is equal to unity if the transistors of the mirror are identical as in Fig. 3.34. For this case B becomes

$$B = \frac{s_p}{s_0}\frac{(s-s_0)}{(s-s_p)} \tag{3.73b}$$

The pole and the zero of this function are

$$s_0 = +\frac{g_{m12}}{C_{dg12}} = \frac{g_{m2}}{C_{dg2}} \tag{3.74a}$$

$$s_p = -\frac{g_{m1}}{(C_{gs1} + C_{gs12} + C_{pT})} = \frac{g_{m1}}{(2C_{gs1} + C_{pT})} \tag{3.74b}$$

where C_{pT} represents the total parasitics of the drain node of M1, i.e. the sum of the drain junction capacitances of M1 and M2 and the gate–source overlap capacitance of M12.

Now $(B+1)$ in (3.72) can be written as

$$(B+1) = 1 + \frac{s_p}{s_0}\frac{(s-s_0)}{(s-s_p)} = \frac{s_0(s-s_p)+s_p(s-s_0)}{s_0(s-s_p)}$$

and simplified with $s_0 \gg |s_p|$ as

$$(B+1) \cong \frac{(s-2s_p)}{(s-s_p)} \tag{3.75}$$

Now the voltage gain from (3.72) becomes

$$A_v \cong \frac{1}{2}\frac{g_{m1}}{Y_L}\frac{(s-2s_p)}{(s-s_p)} \tag{3.76}$$

The total load admittance (Y_L) is usually a parallel combination of a capacitance and a conductance:

$$y_L = g_L + sC_L$$

where the total parallel capacitance (C_L) is the sum of the external load capacitance and the junction capacitances of the drain regions of M11 and M12, and the total conductance (g_L) is the sum of the output internal admittances of M2 and M4 and (if there is any) the external load admittance. Hence the gain expression becomes

$$A_v \cong \frac{1}{2}\frac{g_{m1}}{C_L}\frac{(s-2s_p)}{(s-s_p)(s-s_L)} \tag{3.77}$$

where

$$s_L = -\frac{g_L}{C_L} \tag{3.78}$$

This gain expression can be interpreted as follows.

- The gain function has one negative-real zero and two negative-real poles.
- The phase shift exceeds $\pi/2$ around the second pole but approaches back to $\pi/2$ at higher frequencies.
- The gain at low frequencies ($s \to 0$) is

$$A_{vo} \cong \frac{g_{m1}}{g_L} \tag{3.79}$$

- With no external load, the pole related to the output reaches its maximum value that is determined by the intrinsic conductance and capacitance of the output node as $s_{L(max)} = -g_o/C_{djT}$, where g_o is the sum of the output conductances, and C_{djT} the sum of the drain region junction capacitances of M11 and M12. It can be seen that the magnitude of this pole is always smaller than that of the pole related to the current mirror; in other words, s_L dominates.
- With an external load, which is usually capacitive, this pole dominates more strongly and determines the bandwidth of the amplifier.

- The gain–bandwidth product of the amplifier without external load can be used as a figure of merit for optimization purposes:

$$GBW_{(\max)} \cong A_{vo} f_{L(\max)} = A_o \left(\frac{1}{2\pi} \left| s_{L(\max)} \right| \right) \tag{3.80}$$

Example 3.4 A differential amplifier is designed for a typical 0.18 μm CMOS technology with $L_1 = L_{11} = 0.18\,\mu m$, $L_2 = L_{12} = 0.18\,\mu m$, $W_1 = W_{11} = 18\,\mu m$ and $W_2 = W_{12} = 18\,\mu m$. The lengths of the drain areas are 0.5 μm. The tail current source of $I_T = 400\,\mu A$ is assumed ideal. The key device model parameters are listed as:

$T_{ox} = 4.2$ nm ($C_{ox} = 8.2 \times 10^{-7}$ F/cm^2), $V_{ToP} = -0.43$ V, $\mu_{Po} = 71.2$ cm^2/V.s, $V_{ToN} = 0.315$ V, $\mu_{No} = 326$ cm^2/V.s, $C_{GDO} = C_{GSO} = 1.58 \times 10^{-12}$ F/m, $C_{joP} = 1.14 \times 10^{-8}$ F/cm^2, $C_{swP} = 1.74 \times 10^{-10}$ F/cm, $C_{joN} = 1.19 \times 10^{-8}$ F/cm^2, $C_{sw} = 1.6 \times 10^{-10}$ F/cm.

The sum of the junction capacitances of M11 and M12 for zero bias can be calculated as $C_{jdT} = 33$ fF. Under a bias of approximately 1 V, the capacitance value is $C_{jdT} = 23$ fF. The output conductances are given as $g_{o11} = 91.5\,\mu S$ and $g_{o12} = 41.25\,\mu S$.

The dominant pole frequency with no external load can be calculated as:

$$f_{L(\max)} = \frac{1}{2\pi} \frac{g_o}{C_o} = \frac{1}{2\pi} \frac{(g_{o11} + g_{o12})}{C_{jdT}}$$

$$f_{L(\max)} = \frac{1}{2\pi} \frac{(91.5 + 41.25) \times 10^{-6}}{23 \times 10^{-15}} = 918.6 \text{ MHz}$$

It is obvious that the external load strongly affects the bandwidth. For an external load of only 50 fF and 1 pF, the calculated values of the pole frequencies are 289.4 MHz, and 20.6 MHz, respectively.

g_{m1}, which is necessary to find the low-frequency voltage gain, is calculated as

$$g_{m1} = \sqrt{2 I_D \mu_n C_{ox} (W_1/L_1)} = \sqrt{2 \times (2 \times 10^{-4}) \times 245 \times 8.3 \times 10^{-7} \times 100} = 2.85 \text{ mS}$$

where $\mu_n = 245$ cm^2/V.s corresponds to a 0.5 V gate overdrive as a reasonable value (see Appendix A).

Now, A_{vo} can be calculated as:

$$A_{vo} = \frac{g_{m1}}{g_L} = \frac{2.85 \times 10^{-3}}{132.75 \times 10^{-6}} = 21.5 \text{ (26.6 dB)}$$

These results lead to the following observations about the circuit. The bandwidth is primarily determined by the dominant pole related to the output node, which depends on the sum of the output conductances of M11 and M12, the sum of the junction capacitances of the drain regions of M11 and M12 (both being geometry-dependent) and the load capacitance. The low-frequency gain also depends on the device geometries. Therefore, it is reasonable to investigate the effects of the input and load transistor geometries on the gain–bandwidth product.

As we know, the output conductance of a MOS transistor can be approximately calculated as $g_{ds} = \lambda I_D$. The value of the channel length modulation factor (λ) is inversely proportional to the channel length [28]:

$$\lambda \cong \frac{1}{B_2.L.\sqrt{N}} = h.\frac{1}{L}$$

where h is a structural parameter. Using this expression, the output conductance can be calculated in terms of the tail current, channel lengths of the input transistors, the ratio of the channel lengths of M11 and M12 ($l = W_{11}/W_{12}$), and the h parameter:

$$g_L = h\frac{I_T}{2}\left(\frac{1}{L_{11}} + \frac{1}{L_{12}}\right) = h\frac{I_T}{2}\frac{1}{L_{11}}(1+l) \tag{3.81}$$

The total output node capacitance can be written as

$$C_L = C_{jd11} + C_{jd12} + C_{L(ext)} = C_{jdT} + C_{L(ext)} \tag{3.82}$$

Hence the dominant pole frequency of the amplifier becomes[10]

$$f_L = \frac{1}{2\pi}\left(h\frac{I_T}{2}\frac{1}{L_1}(1+l)\right)\frac{1}{C_{jdT} + C_{L(ext)}}$$

On the other hand, the low-frequency gain can be written as

$$A_{vo} = \frac{g_{m1}}{g_L} = \frac{\sqrt{I_T\mu_n C_{ox}(W_1/L_1)}}{h\frac{I_T}{2}\frac{1}{L_1}(1+l)} = \frac{2}{h(1+l)}\sqrt{\frac{1}{I_T}\mu_n C_{ox} W_1 L_1}$$

and the gain–bandwidth product

$$GBW = \frac{1}{2\pi}\sqrt{I_T\mu_n C_{ox}\frac{W_1}{L_1}}\frac{1}{C_{jdT} + C_{L(ext)}} = \frac{g_{m1}}{C_{jdT}}\frac{1}{1 + \frac{C_{L(ext)}}{C_{jdT}}} \tag{3.83}$$

Expression (3.83) provides useful hints to maximize the gain–bandwidth (GBW) product.

(a) GBW strongly depends on the ratio of the external load capacitance to the sum of the drain junction capacitances of M11 and M12.
(b) Especially for $C_{L(ext)} \gg C_{jdT}$, increasing the tail current helps to increase the GBW at the expense of more power consumption.
(c) If the external load capacitance is comparable to the total output node junction capacitances, decreasing the widths of the transistors of the current mirror load increases the GBW. (But in this case V_{GS12} increases and affects the DC supply voltage budget. Therefore there is a lower limit for W_2, related to the DC operating conditions of the circuit.)
(d) GBW increases with mobility (using NMOS input transistors is advantageous).

[10] Since M1 and M11, M2 and M12 are identical transistors, parameter symbols of M1 and M2 are used for both members of the pairs, hereafter.

(e) GBW changes inversely with L_1 (using minimum channel length transistors is advantageous).

3.7.4 On the input and output admittances of the long tailed pair

The two input admittances of a long tailed pair have frequency-independent capacitive components corresponding to the physical gate–source capacitances. In addition to these components, there are frequency-dependent real and imaginary components of the input admittances resulting from the Miller effect, similar to that of a common-source amplifier or a source follower. These components do not have a significant effect on the behavior of the amplifier, provided that the internal impedances of the driving signal sources are very small compared to the input impedances of the amplifier. However, in many practical cases this condition is not valid, and the interaction of the source impedance and the input impedance of the amplifier affects the amplitude and phase characteristics of the amplifier. The input capacitance is well known and can be easily taken into account, whenever necessary. But the input *conductance* is usually ignored. Therefore we will concentrate on the real components of the input admittances of differential amplifiers, in the following.

We have already seen that, although the input of a common-source or a common drain amplifier is the gate terminal that is isolated from the rest of the circuit, the input admittance acquires a (positive or negative) real part at high frequencies. In this section, we will use the knowledge developed in Sections 3.2 and 3.3 to investigate the input conductances of different types of differential amplifiers, under different driving conditions.

We know that for a differentially driven long tailed pair, the source node is considered to be at the ground potential. Therefore, the behavior of the input admittance is the same as that of a grounded source amplifier. In Section 3.1, we found that the input admittance of the common-source amplifier has a real part that increases with frequency, as a result of the Miller effect. For higher values of the voltage gain (i.e. for higher load resistance values), the magnitude of the input conductance increases. The simulation results showing the input conductance of a resistively loaded and differentially driven long tailed pair (Fig. 3.35(a)) are in agreement with this reasoning.

If a long tailed pair is driven from one of its inputs and the other input is grounded, the source of the input transistor is not at the ground potential any more. Instead, the source has a load equal to the input admittance of the second transistor that is acting in this case as a passive common-gate circuit. We know from Section 2.3 that the input conductance of a common-gate amplifier is equal to g_m. This conductance and its parallel capacitance, which is the sum of the source region junction capacitances of M1 and M11 and the input capacitance of M11, form an R-C load to the source of M1. From Section 3.2, we know that such a source load induces a *negative input conductance* component, provided that the drain is at the ground potential, or the value of the drain load impedance is low. The curve A in Fig. 3.35(b) corresponds to such a situation. For higher values of the drain resistance and corresponding higher gate-to-

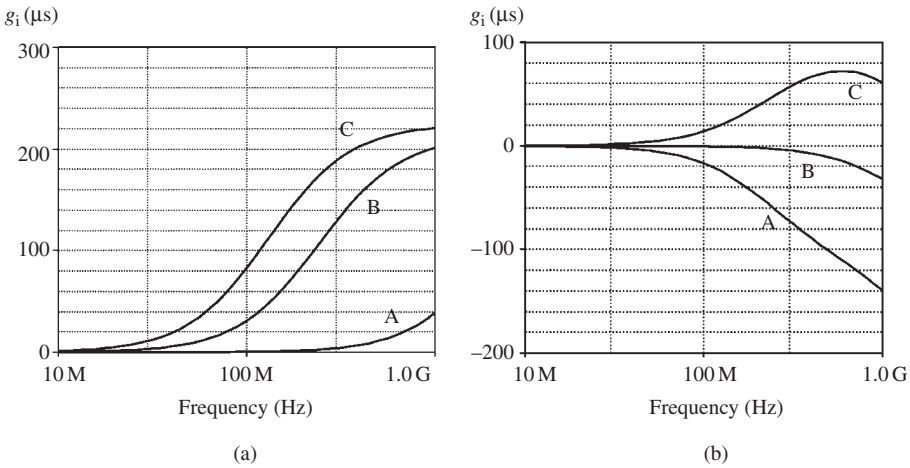


Figure 3.35 Variation of the input conductance of a resistively loaded long tailed pair with frequency ($L = 0.35$ μm, $W = 100$ μm, $I_T = 200$ μA). (a) Differential drive, (b) single-ended drive. A: $R_L = 500$ ohm, B: $R_L = 5$ k ohm, C: $R_L = 10$ k ohm. C_L is 50 fF for all cases.

drain voltage gain, the positive Miller component of the input conductance starts to become effective, eventually compensating (curve B) or dominating (curve C) the negative component.

These results imply that the two input admittances of a single-ended output long tailed pair as shown in Fig. 3.34, which is driven from one of its inputs while the other input is grounded, are not equal. In the case when the circuit is driven from the gate of M1, which is loaded with a diode-connected (low-impedance) drain load, the Miller component is negligible and the input conductance at high frequencies is negative. If the circuit is driven from the gate of M11, which has a high impedance drain load, the Miller component dominates and the input conductance at high frequencies becomes positive. The simulation results given in Fig. 3.36 illustrate this reasoning.

The output admittance of a long tailed pair depends on the circuit configuration. For a resistively loaded or passive-transistor loaded amplifier as shown in Fig. 2.17(a) and (b), the output admittance is the sum of the output admittance of the input transistor and the load admittance, together with the parasitic drain region junction capacitances. The direct way to increase the output resistances is to increase the channel lengths. To keep the main parameters (DC operating conditions, the transconductances of the input transistors) of the circuit, the aspect ratios must be maintained, leading to the increase of the junction capacitances. Therefore, there is a trade-off between a high-output impedance and the 3 dB frequency. Another possibility is to use a cascode circuit configuration as the passive load, whenever the DC voltage budget permits.

In the OTA configuration shown in Fig. 2.17(d), the current mirrors loading the input transistors have dominating roles on the 3 dB frequency of the circuit. Since the

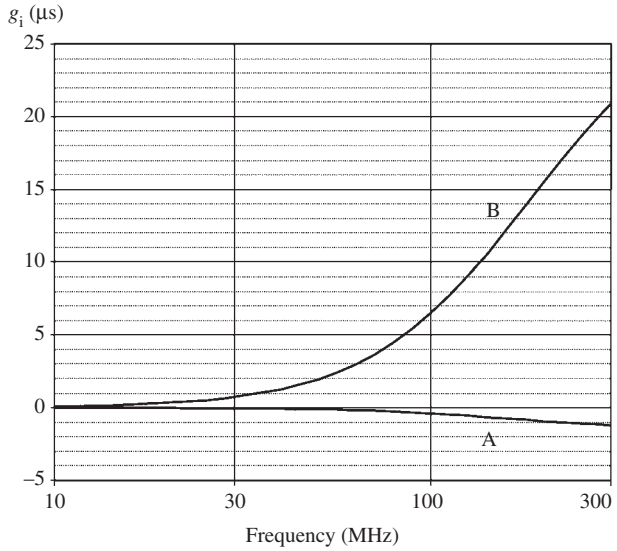


Figure 3.36 Single-ended input conductance of a single-ended output long tailed pair ($L = 0.35\,\mu m$, $W_1 = W_{11} = 10\,\mu m$, $W_2 = W_{12} = 30\,\mu m$, $I_T = 200\,\mu A$, $C_L = 50\,fF$). A: input conductance of M1, B: input conductance of M11.

3 dB frequency of a current mirror strongly depends on the channel length, increasing the channel lengths of the load transistors is not feasible for high-frequency applications. As M5 and M15 (as well as M3) are not on the signal path, it is possible to increase the channel lengths (and maintain the aspect ratios) of these transistors. But this leads to an increase of the drain region junction capacitances that increases the internal parallel output capacitance of the circuit and affects the high-frequency performance to some extent.

We have seen that the input admittance of a differential amplifier has a frequency-dependent capacitive component and a (positive or negative) real part that can have significant values, even at moderate-to-high frequencies. The effects of these components on tuned amplifiers and gyrator-based tuned circuits will be investigated in Chapter 4.

3.8 Gain enhancement techniques for high-frequency amplifiers

The gains of the basic amplifiers investigated in the previous sections may not be sufficiently high for certain applications. The gain can be increased by cascading several amplifying stages, which yields an overall gain equal to the product of the gains of the cascaded stages. In this case it is obvious that both the gain and the bandwidth of a certain stage will be affected by the input impedance of the succeeding stage and the overall cut-off frequency is smaller than the cut-off frequencies of the

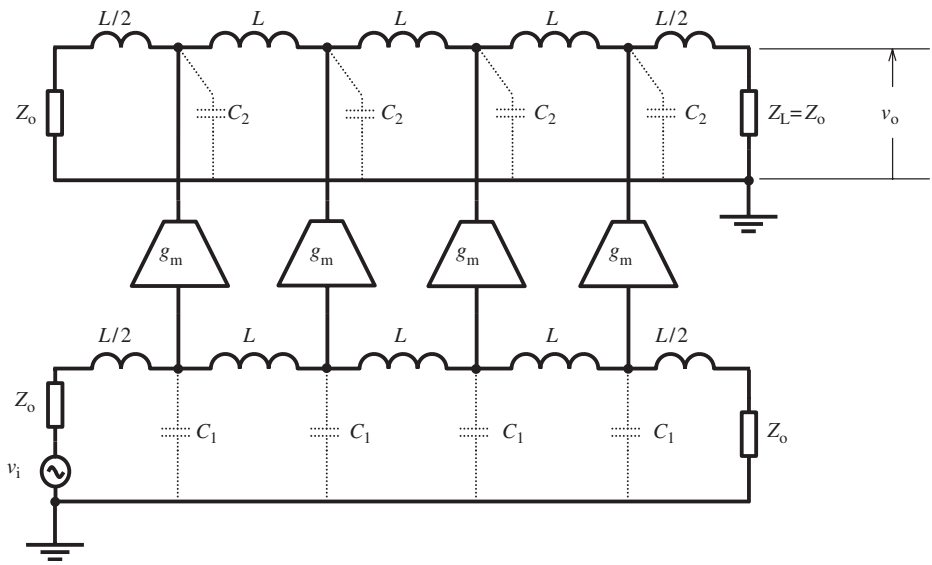


Figure 3.37 Conceptual schematic of a distributed amplifier.

individual stages. This "multiplicative" approach is extensively used for the design of high-gain amplifiers and to obtain maximum cut-off frequency for a certain overall gain. In this section, strategies on the cascading of different types of amplifiers will be discussed.

Another possibility is to increase the gain by "adding" the gains of the individual stages, which is usually called "distributed amplification". In this case the overall cut-off frequency becomes equal to the cut-off frequency of the identical gain stages.

3.8.1 "Additive" approach: distributed amplifiers

The distributed amplifier concept was first developed in the early years of electronics engineering, initially for vacuum tubes [29], and subsequently applied to bipolar transistors, MOS transistors and MESFETs to boost the cut-off frequency of amplifiers[11] – a feat which is not easily possible with conventional cascading techniques. The basic principle of a distributed amplifier is shown in Fig. 3.37. The circuit contains two artificial L-C transmission lines having identical signal delays τ per sector. These lines are terminated at the input and output ends with resistors equal to the characteristic impedance of the lines, Z_{o1} and Z_{o2}, respectively, to eliminate any signal reflections. Although Z_{o1} and Z_{o2} do not have to be equal, it is common practice to use $Z_{o1} = Z_{o2} = Z_o = 50$ ohm.

[11] For example, [30], [31], [32].

The basic parameters of an artificial L-C transmission line can be calculated in terms of the self inductance and parallel capacitance of a sector of the line as follows:

$$\text{characteristic impedance:} \quad Z_o = \sqrt{L/C},$$

$$\text{cut-off frequency:} \quad \omega_0 = \frac{2}{\sqrt{LC}}, \quad f_0 = \frac{1}{\pi\sqrt{LC}},$$

$$\text{signal delay per sector:} \quad \tau = \frac{2}{\omega_0} = \sqrt{LC}.$$

The input capacitances of the identical amplifiers constitute the parallel capacitances of the input transmission line (C_1). Similarly, the output capacitances of the amplifiers (C_o) – partially – constitute the parallel capacitances of the output transmission line, such that the symmetry condition $(C_o + C_2) = C_1$ is satisfied. To minimize the resistive loading of the line, the input and output resistances of the amplifiers must be as high as possible with respect to the characteristic impedance of the line, which imposes the use of transadmittance amplifiers as the most appropriate type.

To understand the principle of operation of a distributed amplifier, assume that the instantaneous value of the input voltage at $t = 0$ is v_i. The input voltage propagates along the line and v_i appears at the input of the first amplifier, A_1 at $t_1 = \tau_h$,[12] at the input of the second amplifier, A_2, at $t_2 = (\tau_h + \tau)$, at the input of the third amplifier, A_3, at $t_3 = (\tau_h + 2\tau)$, and so on. The output current $i_o(v_i) = g_m v_i$ of A_1 corresponding to v_i that appears at $t_1 = \tau_h$ on node 1 of the output line propagates forward and backward along the line. The forward component of this current that is equal to $i_f(v_i) = i_o(v_i)/2$ reaches the node 2 at $t = t_2$, simultaneously with the output current of A_2 corresponding to v_i. Hence the current corresponding to v_i that is propagating in the forward direction along the output line becomes the sum of the individual components, namely $2 \times i_f(v_i)$. At $t = t_3$ the forward current corresponding to v_i becomes $3 \times i_f(v_i)$, etc. Finally, the current corresponding to v_i reaching to the load, Z_L, becomes $4 \times i_f(v_i)$.

From these considerations, the output voltage and the voltage gain of a distributed amplifier containing n cells can be found to be

$$v_o = n \times \frac{1}{2} g_m v_i Z_o \qquad A_v = \frac{v_o}{v_i} = \frac{1}{2} g_m Z_o n \tag{3.84}$$

To obtain a high voltage gain one possibility is to increase n. Another possibility is to increase the gain to a reasonable value and then to cascade the distributed amplifier blocks to reach the targeted total voltage gain,

$$A_{vT} = (A_v)^m = \left(\frac{1}{2} g_m Z_o n\right)^m \tag{3.85}$$

where m is the number of the cascaded distributed blocks.

The bandwidth of a distributed amplifier is determined by the cut-off frequency of the individual transadmittance amplifiers and that of the transmission lines. Any type

[12] τ_h is used for the signal delay of the input (or output) half-sector of the line.

of transadmittance amplifier can be used as the amplifier block, provided that the bandwidth of the amplifier fulfills the requirement. One of the most frequently used amplifier topologies in distributed configurations is a simple MOS transistor biased in the saturation region. Since the voltage gain of an individual amplifier loaded with $Z_o/2$ is low, the adverse effects of Miller feedback are usually negligible. The cascode configuration is certainly another possibility for implementing the amplifier block.

3.8.2 Cascading strategies for basic gain stages

As already mentioned, cascading of multiple gain stages (i.e. connecting the output port of one amplifier stage to the input port of the next stage) is the most frequently used technique for building high-gain amplifiers. The types and numbers of the cascaded stages depend on the type and overall gain of the targeted multi-stage amplifier. Here the term "gain" is used in a broader meaning; it can be power gain, voltage gain, current gain, transfer admittance or transfer impedance.

In multi-stage power amplifiers, the input impedance of the next stage must be equal to the complex-conjugate of the output impedance of the previous stage, in order to allow the most efficient power transfer from the output of the previous amplifier to the input of the following stage.[13] This means that, if the output impedance of the first stage is $Z_{o1} = r_{o1} + jx_{o1}$, to satisfy the maximum power transfer condition, the input impedance of the second stage must be equal to

$$Z_{i2} = \bar{Z}_{o1} = r_{o1} - \mathrm{j}x_{o1}$$

Note that the real components of these two impedances are identical and the imaginary components are in series resonance. This condition dictates that the maximum power transfer condition can be fully satisfied at a certain frequency only (the resonance frequency of the two complementary reactances), and approximately satisfied in the bandwidth of the resonance circuit. The maximum power transfer condition can also be expressed in terms of the output and input admittances, i.e. $Y_{i2} = \bar{Y}_{o1} = g_{o1} - \mathrm{j}b_{o1}$, which corresponds to parallel resonance of the parallel reactive components, in addition to the condition $g_{i2} = g_{o1}$.

It can be seen from this brief explanation that the cascading strategy based on the maximum power transfer condition is useful only for narrow-band applications, but not for wide-band applications, for example from DC to several GHz.

For wide-band applications it is necessary to *efficiently* transfer the signal (a voltage or a current) from the output port of the previous stage to the input port of the following stage.

If the signal to be transferred from the output of the previous stage to the input of the following stage is a voltage, the necessary condition for efficient signal transfer is $z_{i2} \gg Z_{o1}$. Similarly, if the signal to be transferred from the output of the previous stage to the input of the following stage is a current, the necessary condition for efficient

[13] This is one of the basic theorems of network theory.

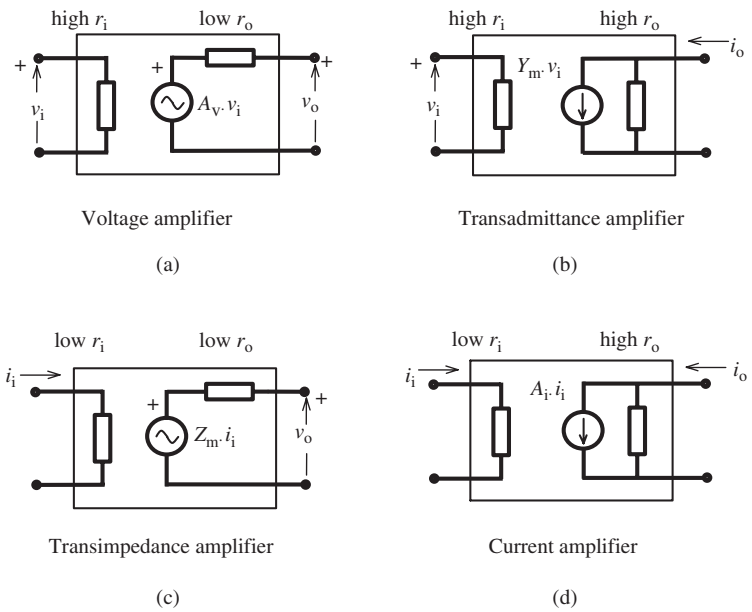


Figure 3.38 The four basic amplifier configurations.

signal transfer is $z_{i2} < z_{o1}$. The parallel reactive components of these impedances usually consist of unavoidable parasitic capacitances and become effective at the high end of the frequency band, in other words, they determine the limit of the usable frequency range of the amplifier.

From these considerations, we can state that for efficient and wide-band signal transfer at a cascading node:

- the parallel resistance – i.e., the parallel equivalent of the output resistance of the first stage and the input resistance of the following stage – must be as small as possible, which increases the pole frequency related to this cascading node;
- the input resistance of the next stage must be as high as possible compared to the output resistance of the previous stage, which ensures efficient voltage transfer at this cascading node;
- alternatively, the input resistance of the next stage must be as small as possible compared to the output resistance of the previous stage, which ensures efficient current transfer at this cascading node.

In Fig. 3.38 the four possible amplifier configurations are shown, indicating their input and output resistances. The equivalent circuit given in Fig. 3.38(a) is called a typical “voltage amplifier”, with a high input resistance and a low internal output resistance.

The configurations shown in Fig. 3.38(b), Fig. 3.38(c) and Fig. 3.38(d) are called, depending on the magnitude of the input and output resistances (i.e., depending on the

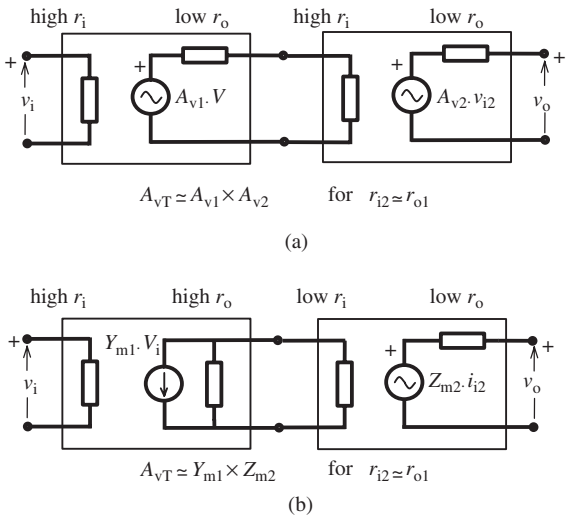


Figure 3.39 Two appropriate configurations for a two-stage voltage amplifier. (a) Cascaded two-voltage amplifiers, (b) a trans-admittance amplifier cascaded with a transimpedance amplifier.

appropriate type of the input and output signal) "transadmittance amplifier", "trans-impedance amplifier" and "current amplifier", respectively.

To construct a two-stage cascaded wide-band voltage amplifier, we can choose one of the two possible solutions as shown in Fig. 3.39, by applying the conditions stated above to the cascading node for efficient signal transfer, as well as for high bandwidth: either cascading two voltage amplifiers, or cascading a transadmittance amplifier in the first stage with a transimpedance amplifier in the second stage.

Similarly, the possible configurations for two-stage transadmittance amplifiers, two-stage transimpedance amplifiers and two-stage current amplifiers are given in Figs. 3.40, 3.41 and 3.42, respectively.

Appropriate configurations for a certain type of amplifier containing more than two cascaded stages can be found by using the same systematic approach.

Problem 3.3 *Draw the block diagrams of possible configurations of a three-stage high-gain, high-bandwidth transimpedance amplifier and compare their properties.*

3.8.3 An example: the "Cherry–Hooper" amplifier

To construct a high-bandwidth voltage amplifier, we usually prefer the well-known basic form of a MOS voltage amplifier, as shown in Fig. 3.43. The low-frequency voltage gain of such an amplifier is

$$A_v(0) = -g_m R_L$$

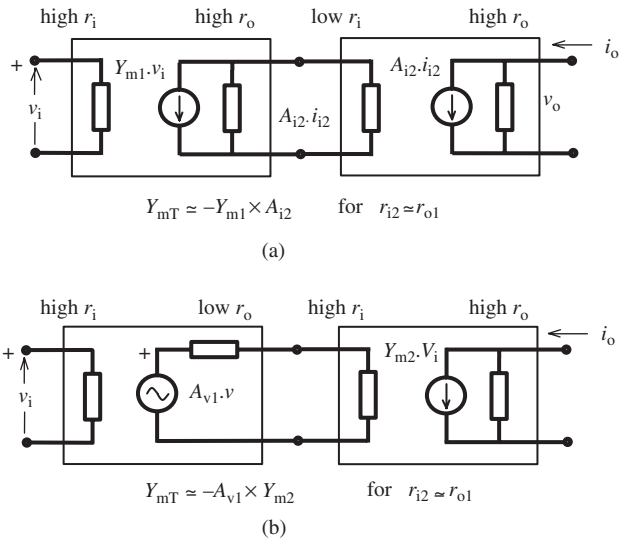


Figure 3.40 Two appropriate configurations for a two-stage transadmittance amplifier: (a) a transadmittance amplifier cascaded with a current amplifier, (b) a voltage amplifier cascaded with a transadmittance amplifier.

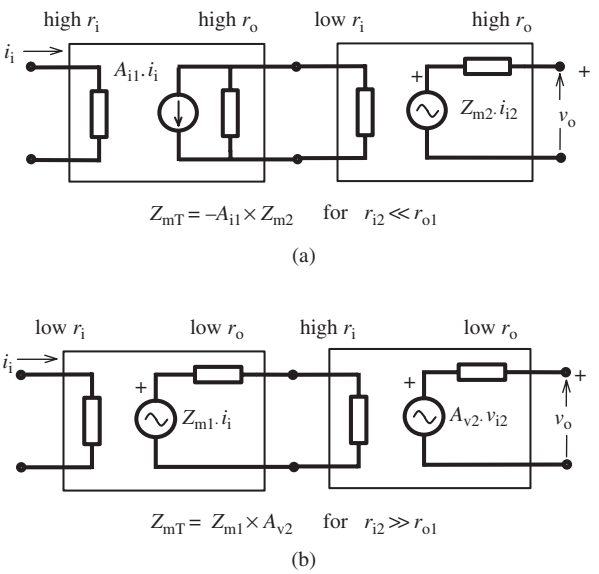


Figure 3.41 Two appropriate configurations for a two-stage transimpedance amplifier: (a) a current amplifier cascaded with a transimpedance amplifier, (b) a transimpedance amplifier cascaded with a voltage amplifier.

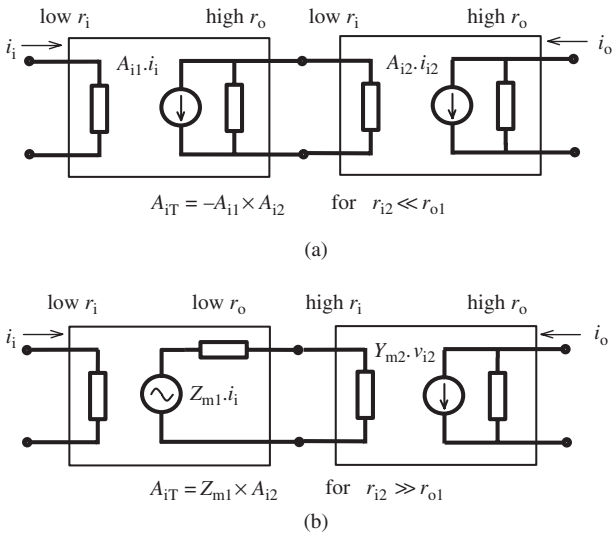


Figure 3.42 Two appropriate configurations for a two-stage current amplifier: (a) cascaded two-current amplifiers, (b) a transimpedance amplifier cascaded with a transadmittance amplifier.

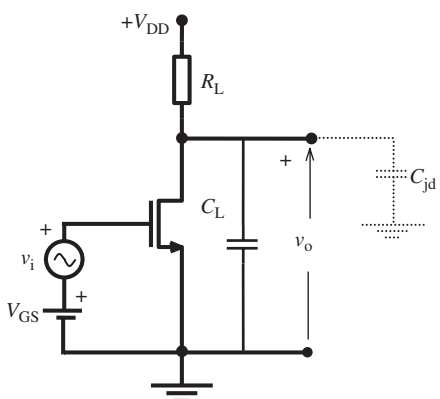


Figure 3.43 The basic (simplest) form of a single-stage voltage amplifier.

and the cut-off frequency is

$$f_{3dB} = \frac{1}{2\pi} \frac{1}{R_L C_o}, \quad C_o = C_L + C_{jd},$$

where C_o is the sum of the output load capacitance C_L and the drain parasitic capacitance, C_{jd} (the junction capacitance of the drain junction). Note that both of these capacitances are controllable by the designer. In order to increase the bandwidth, R_L

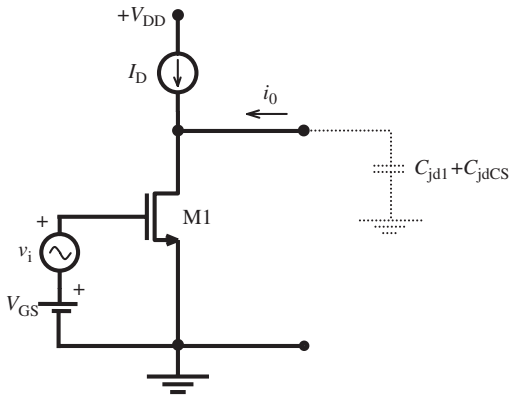


Figure 3.44 The basic (simplest) form of a transadmittance amplifier.

has to be decreased, which also decreases the voltage gain. The straightforward approach to increase the gain to the targeted value is to cascade two similar voltage amplifiers as shown in Fig. 3.39(a). In this case, the pole frequency of the cascading node becomes

$$\omega_{pC} = \frac{1}{R_L C_o} \tag{3.86}$$

where C_o is the sum of the drain parasitic capacitance C_{jd} and the gate–source capacitance of the second stage. At the output port of the amplifier, the pole frequency is determined by R_L and the load capacitance as

$$\omega_{pL} = \frac{1}{R_L C_L} \tag{3.87}$$

Note that there is another possibility: to cascade a transadmittance amplifier and a transimpedance amplifier, as shown in Fig. 3.39(b). The simplest topology of a transadmittance amplifier is shown in Fig. 3.44 where the current supplied by the DC current source is equal to the drain current of M1. The low-frequency transadmittance of the circuit is g_{m1}, corresponding to the transconductance of M1. The output parasitic capacitance is the sum of the drain parasitic capacitance C_{jd1} and the output capacitance of the DC current source, C_{jdCS}, which is usually the drain-bulk capacitance of a PMOS transistor and is of the same order as C_{jd1}.

As the second stage, we need to build a low-input resistance transimpedance amplifier, having a low input resistance to obtain a high pole frequency at the cascading junction and a low output resistance to obtain a high pole frequency at the output node. Unfortunately, there is no basic MOS amplifier configuration fulfilling the low input impedance and low output impedance at the same time. A possibility is to use an amplifier with applied parallel voltage feedback, as shown in Fig. 3.45(a).

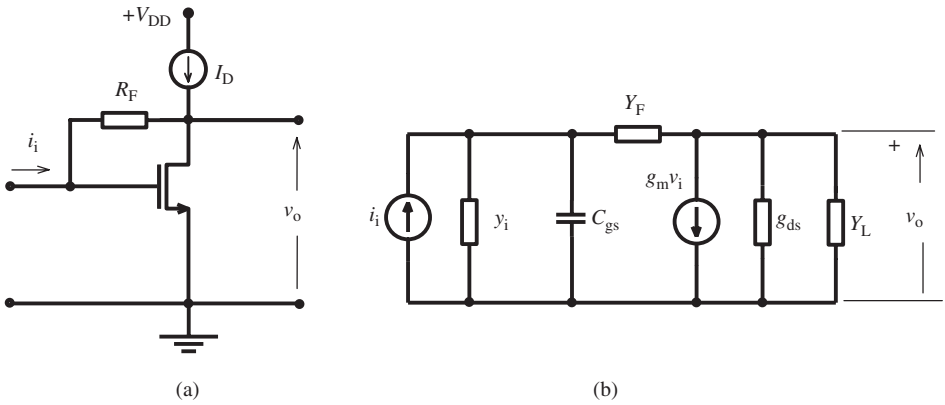


Figure 3.45 (a) The parallel voltage feedback applied to a MOS amplifier as a transimpedance amplifier. (b) The small-signal equivalent circuit.

The small-signal equivalent circuit of this amplifier is shown in Fig. 3.45(b), where y_i represents the output admittance of the signal source (in our case the output admittance of the transadmittance amplifier), Y_F is the parallel equivalent of the feedback resistance, R_F, the drain–gate capacitance of the transistor and the parallel capacitance connected to R_F (if there is any).

It can be seen that this equivalent circuit is the same as that of the CMOS inverter investigated in Section 3.5. Therefore, the expressions and results can be directly applied using the appropriate parameter values. The low-frequency value of the transimpedance was found

$$Z_m(0) = -\frac{(g_m - G_F)}{G_F(g_{ds} + g_m)} \tag{3.88a}$$

which can be reduced to

$$Z_m(0) \cong -\frac{(g_m - G_F)}{g_m G_F} = -\frac{1}{G_F} + \frac{1}{g_m} \cong -\frac{1}{G_F} = -R_F \tag{3.88b}$$

for $g_{ds} < g_m$ and $G_F < g_m$. The low-frequency value of the input impedance can be calculated from the equivalent circuit as

$$Z_i(0) = \frac{R_F + r_{ds}}{1 + g_m r_{ds}} \cong \frac{1}{g_m} \tag{3.89}$$

for $r_{ds} \gg R_F$ and $g_m r_{ds} \gg 1$. Similarly, the low-frequency value of the output impedance,

$$Z_o(0) = \frac{r_{ds}}{1 + g_m r_{ds}} \cong \frac{1}{g_m} \tag{3.90}$$

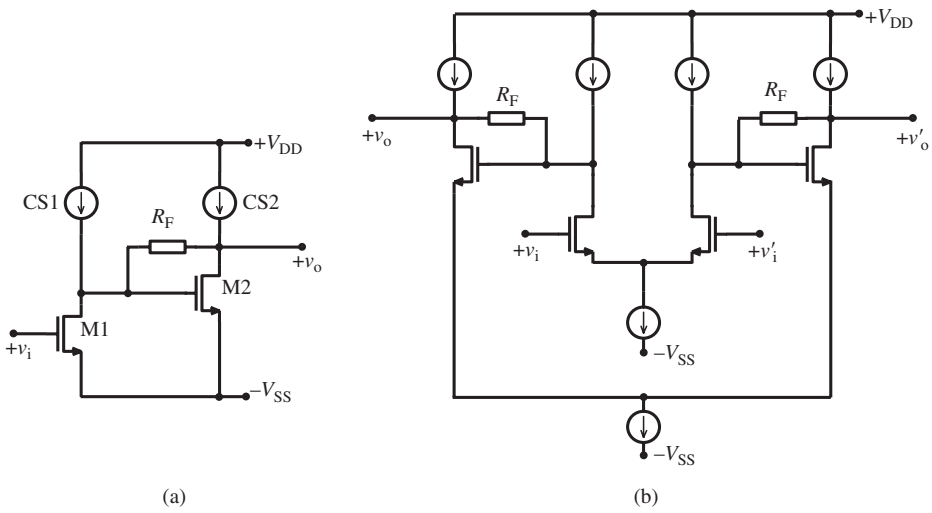


Figure 3.46 (a) A voltage amplifier composed of a transadmittance stage and a transimpedance stage. (b) The differential version of this circuit that is known as the Cherry–Hooper amplifier.

According to (3.89) and (3.90), the input and output resistances of the amplifier are approximately equal to $(1/g_m)$, provided that $r_{ds} \gg R_F \gg (1/g_m)$ and $g_m r_{ds} \gg 1$. These input and output resistances can be made small enough to increase the pole frequencies of the cascading node, as well as the output node, compared to that of the two-stage cascaded voltage amplifier, given in (3.86) and (3.87).

The schematic of this two-stage voltage amplifier is given in Fig. 3.46(a). The low-frequency voltage gain is

$$A_{vT}(0) = Y_{m1}(0) \times Z_{m2}(0) \cong -g_{m1} R_F$$

The pole frequency corresponding to the cascading node is

$$\omega_{pC} = \frac{g_{m2}}{C_{jd1} + C_{jdCS1} + C_{gs2}}$$

where C_{jdCS1} is the drain-bulk junction capacitance of the PMOS current source of M1. Similarly, the pole frequency corresponding to the output node is

$$\omega_{pL} = \frac{g_{m2}}{C_{jd2} + C_{jdCS2} + C_L}$$

Another factor affecting the overall frequency characteristic is the parasitic capacitance parallel to the feedback resistor. As discussed in Section 3.5, this capacitance has an important effect of the transimpedance of the second stage and it may be possible to fine-tune the overall frequency characteristic with an additional capacitance in parallel to C_{dg2}.

The differential version of this circuit is given in Fig. 3.46(b), which corresponds to the well-known and extensively used wide-band voltage amplifier topology; the "Cherry and Hooper" amplifier [33]. This example demonstrates that in order to construct a certain type of wide-band amplifier, all alternative solutions given in Figs. 3.39 to 3.42 must be considered in a systematic manner.

4 Frequency-selective RF circuits

During the early days of radio design, the tuned amplifier was one of the most important subjects of electronic engineering. It was used as the input RF amplifier of radio receivers tunable in a certain frequency range and as an intermediate frequency (IF) amplifier tuned to a fixed frequency. From the 1930s to 1950s RF amplifiers using electron tubes were one of the most important and interesting research areas of electronics engineering and were investigated in depth. TV and all other wireless systems were other application areas for tuned amplifiers.

After the emergence of transistor circuits in the 1950s, the knowledge already acquired was sufficient for transistorized tuned amplifiers, because the frequency range was still limited to several hundreds of MHz. Active filters, one of the benefits of analog ICs, largely replaced another class of frequency-selective circuits, i.e. L-C filters. As a result of these developments the importance of inductors and circuits containing inductors decreased and eventually these subjects disappeared from many electronic engineering curricula and textbooks.

The rapid expansion of wireless personal communication and data communication during the recent decade, and the developments in IC technology that extended the operation frequencies into the GHz range, resulted in the "re-birth" of frequency-selective circuits containing inductors. In parallel to the increase of operating frequencies reaching up to multi-GHz range, the necessary inductance values decreased to "nanohenry" level that are now possible to realize as an integral part of ICs, as mentioned in Chapter 1. But the quality factor of these "on-chip inductors" is typically in the range of 5 to 15, that is considerably smaller compared to the quality factor of classical discrete "wound" inductors.[1] It must be kept in mind that the theory of tuned circuits developed in older textbooks is based on high-Q inductors. Therefore, a re-consideration of tuned circuits for low-Q circuits is necessary.

In this chapter, first a summary of the resonance circuits will be given, including basic definitions and behaviors with special emphasis on low-quality factor circuits. Then, single-tuned amplifiers, stagger tuning and amplifiers containing coupled resonance circuits will be presented. Active and passive filters are considered outside the scope of this book, but owing to the increasing importance of a class of active filters at high frequencies, gyrator-based g_m-C circuits will be investigated in detail.

[1] The quality factors of discrete wound inductors are in the range of 100 to 1000, depending on the structure and material of the coil and the frequency.

4.1 Resonance circuits

Resonance is one of the most important phenomena that occurs in many physical systems that are comprised of components capable of storing energy as potential as well as kinetic energy. These systems start to oscillate when excited, i.e., if a small amount of energy is injected into the system, for example as potential energy. This energy transitions between "fully potential" and "fully kinetic" phases, in the system.[2] The frequency of oscillations depends on the parameters of the system. The pendulum is one of the fundamental oscillatory systems that is easiest to understand. An oscillatory system loses its energy in time, if there is (it would be better to say: there always is) a reason to consume energy in the system (for example the air friction for a pendulum), the amplitude of the oscillation eventually decreases down to zero.

In electrical circuits, the capacitor is the component that is capable of storing potential energy in the form of charge. The energy stored in a capacitor is $E_p = (1/2)CV^2$ if the voltage is V. An inductor, on the other hand, stores kinetic energy in the form of flux. If a current I is flowing through an inductor, the energy stored in the inductor is $E_k = (1/2)LI^2$. All resistors in a circuit consume energy. The consumed (dissipated) energy in a time interval t is $E_d = I^2Rt$, in terms of the current, and $E_d = (V^2/R)t$, in terms of voltage.

The components of an electrical resonance circuit are an inductor and a capacitor. The series and/or parallel parasitic resistances, for example the resistance of the inductor material and the parallel dielectric losses of the capacitor, are the energy-consuming components of the system. There are two possibilities for forming a resonance circuit: to connect the inductor and the capacitor in series or in parallel. It will be shown that the resonance effect of a parallel resonance circuit is pronounced when it is driven by a current source. In contrast, a series resonance circuit exhibits the resonance effect when it is driven by a voltage source.

4.1.1 The parallel resonance circuit

A parallel resonance circuit is shown in Fig. 4.1. The inductor is modeled by its self inductance (L) and the series resistance (r_L) that represents all losses related to the inductor (for an on-chip inductor, this includes the parasitic resistance of the strip, the losses of the magnetically induced currents and the substrate resistance, as shown in Fig. 1.32). C is the value of the capacitance. For on-chip capacitors, since the insulator is silicon dioxide, the parallel dielectric losses are negligibly small, but there is a considerable series resistance as shown in Fig. 1.25 that is the main cause of the low Q value of an on-chip capacitor.[3] R_p represents the internal resistance of the signal (current) source, and other parallel losses if there are any. The impedance seen by the

[2] For an excellent reading on resonance see [34].

[3] Note that for discrete capacitors in classical resonance circuits, the series resistance of the capacitor is neglected during the derivation of the expressions. Therefore, these expressions do not sufficiently represent the behavior of the on-chip resonance circuits.

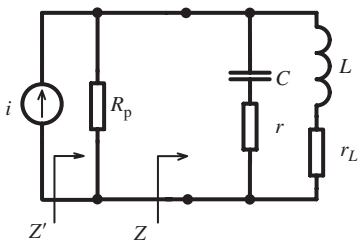


Figure 4.1 A parallel resonance circuit with the series resistances of the inductor and the capacitor.

current source (Z') is the parallel equivalent of R_p and Z (Fig. 4.1). In the following, we will first calculate Z and then include the contribution of R_p.

The impedance Z can be calculated as

$$Z = \frac{r_L + s(L + Cr_C r_L) + s^2 LCr_C}{1 + sC(r_C + r_L) + s^2 LC}$$

and in the ω domain,

$$Z(\omega) = \frac{(r_L - \omega^2 LCr_C) + j\omega(L + Cr_C r_L)}{(1 - \omega^2 LC) + j\omega C(r_C + r_L)} \tag{4.1}$$

The real and imaginary parts of (4.1) are

$$\mathrm{Re}\{Z\} = \frac{(r_L - \omega^2 LCr_C)(1 - \omega^2 LC) + \omega^2 C(r_L + r_C)(L + Cr_L r_C)}{(1 - \omega^2 LC)^2 + \omega^2 C^2 (r_L + r_C)^2} \tag{4.2}$$

$$\mathrm{Im}\{Z\} = \frac{\omega(L + Cr_L r_C)(1 - \omega^2 LC) - \omega C(r_L - \omega^2 LCr_C)(r_L + r_C)}{(1 - \omega^2 LC) + \omega^2 C^2 (r_L + r_C)^2} \tag{4.3}$$

From these expressions, some information related to the following special cases can be extracted.

- For $\omega \to 0$ the imaginary part of the impedance becomes zero and the real part is equal to r_L, as can be intuitively seen from Fig. 4.1.
- For $\omega \to \infty$ the imaginary part of the impedance again becomes zero and the real part is equal to r_C.
- The imaginary part of the impedance becomes zero (the impedance is resistive) for

$$\omega^2_{(\mathrm{Re})} = \frac{(L - Cr_L^2)}{LC(L - Cr_C^2)} \tag{4.4}$$

In Fig. 4.2(a), the root locus of Z is shown for $r_C \ll r_L$, which corresponds to the conventional resonance circuits where the series resistance of the capacitor is negligible. As seen from the figure, the impedance is resistive and equal to r_L at $\omega = 0$. The

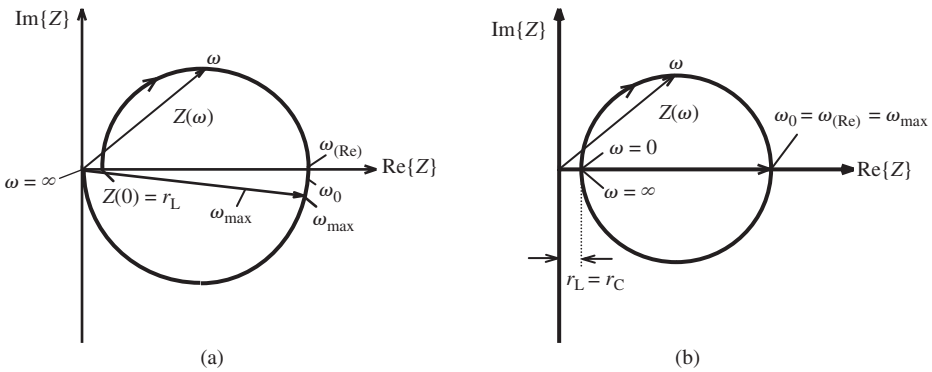


Figure 4.2 The root locus of the impedance of a parallel resonance circuit; (a) for $r_C \ll r_L$, (b) for $r_C = r_L$.

impedance is inductive up to a frequency for which the impedance again becomes fully resistive. The frequency corresponding to this case was found in (4.4) and can be simplified to

$$\omega_{(\mathrm{Re})}^2 = \frac{1}{LC} - \frac{r_L^2}{L^2} \tag{4.5}$$

for $r_C \ll r_L$. Note that this frequency is not the same as the ω_0 natural frequency[4] that is defined as

$$\omega_0^2 = \frac{1}{LC} \tag{4.6}$$

In addition, it can be shown that the frequency for which the magnitude of the impedance is maximum is equal to

$$\omega_{\max}^2 = \frac{1}{LC}\sqrt{1 + 2r_L^2 C} - \frac{r_L^2}{L^2} \tag{4.7}$$

for $r_C \ll r_L$. The differences between these three frequencies that are characteristic for a resonance circuit are small but may have important effects on the behavior of the circuits containing resonance circuits. If they are assumed to be equal, this assumption may hide certain delicate properties of the circuit, as will be exemplified below.

Considering the example for an L-C oscillator, the circuit oscillates at a frequency for which the loop gain satisfies the gain condition according to the Barkhausen criterion, and the total phase shift on the loop is equal to zero. It is obvious that, ideally, the frequency for which the magnitude of the impedance (and consequently the gain) is maximum and the frequency for which the phase shift is zero (the impedance is resistive) will coincide; with the notation we used above, $\omega_{\max} = \omega_{(\mathrm{Re})}$. The conventional way to satisfy (at least to approach to) this condition is to use a low-loss (high-Q) resonance circuit as the resonator of the oscillator.

[4] The term "natural frequency" is used for the resonance frequency of the lossless *L-C* combination.

For on-chip resonance circuits, there is a possibility to equate ω_{max} and $\omega_{(Re)}$. It can be seen from (4.4) that for $r_C = r_L$, $\omega_{(Re)}$ becomes equal to ω_0 and the beginning of the root locus ($\omega = 0$) coincides with its end ($\omega = \infty$) at $\mathrm{Re}\{Z\} = r_C = r_L$. The root locus corresponding to this special case is shown in Fig. 4.2(b). It can be seen that owing to the symmetry of the plot not only ω_0 and $\omega_{(Re)}$, but also ω_{max} coincide. This may be a valuable hint for the phase noise minimization of L-C oscillators.

Example 4.1 To check these results let us simulate a resonance circuit for two different cases.

(a) The inductance is $L = 2nH$ and the series resistance of the inductance is $r_L = 10$ ohm. C has the appropriate value to tune the circuit to $f_0 = 5$ GHz. The series resistance of the capacitor is negligibly small.
(b) The inductance and the resonance frequencies are the same as (a). But the series resistances of L and C are the same and equal to 5 ohm.

Draw the variations of the magnitude and phase of the impedance as a function of frequency and compare.

The value of the capacitance can be calculated as 506.7 fF from (4.6). The PSpice results of the magnitude and phase are shown in Fig. 4.3(a). As can be seen from these plots, the frequencies corresponding to the maximum of the magnitude and the frequency where the impedance is resistive (i.e., where the phase angle is zero) are not

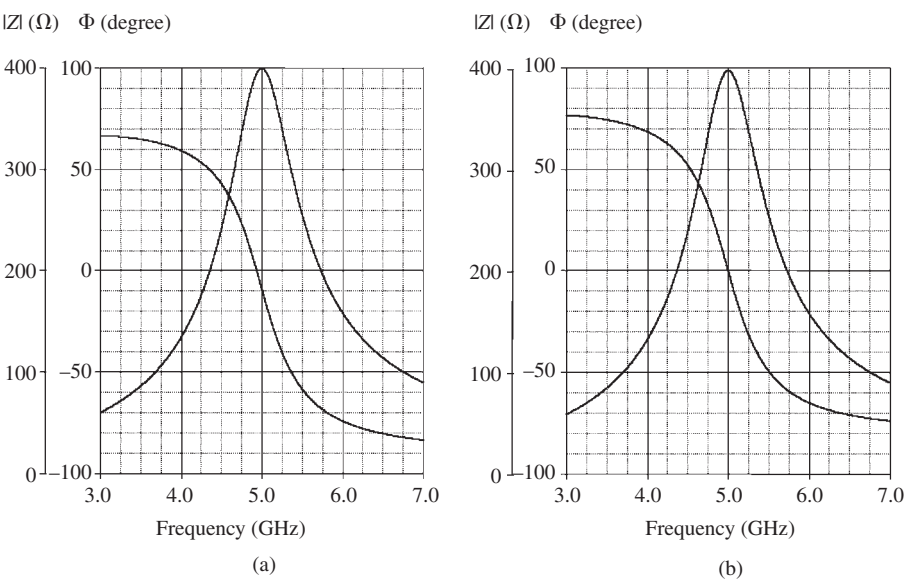


Figure 4.3 The phase and magnitude curves of a parallel resonance circuit: (a) with a lossy inductor and lossless capacitor, (b) with a lossy inductor and lossless capacitor, if the series resistances of L and C are equal.

the same. From another important viewpoint, the phase angle corresponding to the maximum value of the impedance is not zero, but $\Phi = -8.05$ degrees.

The simulation results for case (b) are shown in Fig. 4.3(b). As can be seen from these plots, the frequencies corresponding to the maximum of the impedance and to the zero phase shift are the same, or in other words, the phase shift corresponding to the maximum value of the impedance is zero, as expected.

4.1.1.1 The quality factor of a resonance circuit

Before proceeding further, let us review the basic definition of the quality factor (Q) of any oscillatory system, and apply this definition to calculate the quality factor of the circuit shown in Fig. 4.4.

The quality factor of an oscillatory system is defined as

$$Q = 2\pi \frac{\text{the total energy of the system}}{\text{the energy lost in one period}} \tag{4.8}$$

Since the total energy in an oscillatory system swings back and forth between fully potential energy and fully kinetic energy, to find the total energy of the system it is convenient to calculate the value of the potential or kinetic energy at one of these extreme conditions. For the circuit shown in Fig. 4.4 the circuit is excited with a sinusoidal voltage, $v = V \sin \omega_0 t$, whose frequency is equal to the natural frequency ω_0 of the resonance circuit. The maximum value of the potential energy stored in the capacitor corresponds to the case when the voltage on this capacitor is maximum, i.e. the peak value of v. Then the maximum of the potential energy, which is equal to the total energy of the system, is

$$E = \frac{1}{2} C V^2 \tag{4.9}$$

The energy lost in one period (T) on the parallel resistor R_p is

$$E_p = T \times \frac{1}{2} \frac{V^2}{R_p} = \frac{1}{2f_0} \frac{V^2}{R_p} \tag{4.10}$$

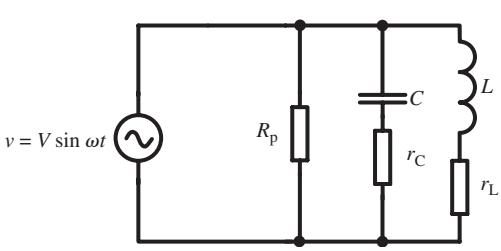


Figure 4.4 The parallel resonance circuit in its most general form.

To calculate the energy lost on r_L, the peak value of the sinusoidal current flowing through the inductor branch must be calculated:

$$I_L = \frac{V}{\sqrt{r_L^2 + \omega_0^2 L^2}}$$

Then the energy lost in one period on r_L is

$$E_L = \frac{1}{2f_0} V^2 \frac{r_L}{r_L^2 + \omega_0^2 L^2} \tag{4.11}$$

Similarly, the energy consumed on r_C in one period can be found as

$$E_C = \frac{1}{2f_0} V^2 \frac{r_C}{r_C^2 + \frac{1}{\omega_0^2 C^2}} \tag{4.12}$$

The total energy lost in one period is the sum of (4.10), (4.11) and (4.12). Hence from (4.8) and (4.9) the effective quality factor containing all of the losses of the circuit can be obtained as

$$\frac{1}{Q_{\text{eff}}} = \frac{1}{Q_p} + \frac{1}{Q_L} + \frac{1}{Q_C} \tag{4.13}$$

Here:

$$Q_L = \frac{r_L^2 + \omega_0^2 L^2}{\omega_0 L r_L} \cong \frac{\omega_0 L}{r_L} \tag{4.14}$$

which corresponds to the losses on r_L, in other words it is the quality factor of the inductor at ω_0,

$$Q_C = \frac{\omega_0^2 C^2 r_C^2 + 1}{\omega_0 C r_C} \cong \frac{1}{\omega_0 C r_C} \tag{4.15}$$

which corresponds to the losses on r_C, i.e. it is the quality factor of the capacitor at ω_0, and

$$Q_p = \omega_0 C R_p = \frac{R_p}{\omega_0 L} \tag{4.16}$$

which corresponds to the losses on R_p.

Note that the approximate values of Q_L and Q_C are only valid for small values of the corresponding resistances and that neglecting them can cause considerable error for Q values smaller than 5.

Problem 4.1 *Calculate the error of the approximate form of* (4.14) *for* $Q_L = 50$, $Q_L = 10$, $Q_L = 5$ *and* $Q_L = 3$.

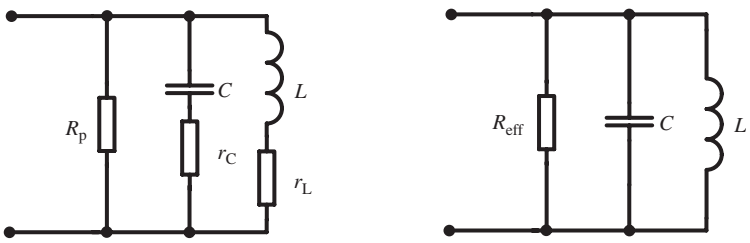


Figure 4.5 (a) Parallel resonance circuit with its lossy components. (b) The effective parallel resistance representing all losses of the circuit at f_0.

It is common practice to represent the total losses of a resonance circuit with a lumped parallel resistor (effective resistance) R_{eff}, as shown in Fig. 4.5.

If we express the losses of the inductor branch with a parallel resistor R_{Lp}, from (4.14) and (4.16) we can obtain

$$R_{\text{Lp}} = r_{\text{L}} + \frac{\omega_0^2 L^2}{r_{\text{L}}} \cong r_{\text{L}}(1 + Q_{\text{L}}^2) \cong \frac{L}{r_{\text{L}} C} \tag{4.17}$$

Similarly, the parallel resistor representing the losses of the capacitor branch is

$$R_{\text{Cp}} = r_{\text{C}} + \frac{1}{\omega_0^2 C^2 r_{\text{C}}} \cong r_{\text{C}}(1 + Q_{\text{C}}^2) \cong \frac{L}{r_{\text{C}} C} \tag{4.18}$$

Then the effective resistance can be expressed as the parallel equivalent of R_{Lp}, R_{Cp} and R_{p}:

$$R_{\text{eff}} = (R_{\text{Lp}} // R_{\text{Cp}} // R_{\text{p}}) \tag{4.19a}$$

or in terms of parallel conductances,

$$G_{\text{eff}} = G_{\text{Lp}} + G_{\text{Cp}} + G_{\text{p}} \tag{4.19b}$$

Then the effective quality factor containing all of the losses becomes

$$Q_{\text{eff}} = \frac{1}{\omega_0 L G_{\text{eff}}} = \frac{\omega_0 C}{G_{\text{eff}}} \text{ or } Q_{\text{eff}} = \frac{R_{\text{eff}}}{\omega_0 L} = \omega_0 C R_{\text{eff}} \tag{4.20}$$

This equivalence considerably decreases the complexity of the expressions and is extensively used in the literature, but it also hides some properties of the resonance circuit. For example, although the natural frequency (f_0), the frequency where the impedance is real ($f_{(\text{Re})}$) and the frequency corresponding to the maximum magnitude of the impedance ($f_{(\text{max})}$) are not the same for the original circuit shown in Fig. 4.5(a), they are all the same for Fig. 4.5(b) and the skew of the frequency characteristics exemplified in Example 4.1 is now hidden.

4.1.1.2 The quality factor from a different point of view

As previously mentioned, an oscillatory system starts to oscillate once it is excited, but the amplitude of oscillation gradually decreases owing to the losses. Since the rate of this amplitude degradation and the quality factor of the system are both defined by the losses of the system, there must be an inter-relation between them.

Consider a parallel resonance circuit simplified as shown in Fig. 4.5(b). The impedance of this circuit can be expressed in the s-domain as

$$Z = \frac{1}{C}\frac{s}{s^2 + s\frac{1}{R_{\text{eff}}C} + \frac{1}{LC}} = \frac{1}{C}\frac{s}{(s - s_{\text{p1}})(s - s_{\text{p2}})} \tag{4.21}$$

The poles of the impedance function are

$$s_{\text{p1,p2}} = -\frac{1}{2R_{\text{eff}}C} \mp \sqrt{\left(\frac{1}{2R_{\text{eff}}C}\right)^2 - \frac{1}{LC}}$$

or

$$s_{\text{p1,p2}} = \sigma \mp \text{j}\sqrt{\omega_0^2 - \sigma^2} = \sigma \mp \text{j}\bar{\omega}_0 \tag{4.22a}$$

where

$$\sigma = -\frac{1}{R_{\text{eff}}C},\ \omega_0 = \frac{1}{\sqrt{LC}} \text{ and } \bar{\omega}_0 = \sqrt{\omega_0^2 - \sigma^2} \tag{4.22b}$$

If the circuit is excited with a current pulse, the voltage between the terminals of the circuit can be written in the s-domain as

$$V = \left(\frac{I}{s}\right)\frac{1}{C}\frac{s}{(s - s_{\text{p1}})(s - s_{\text{p2}})}$$

and the voltage in the *time*-domain as

$$v(t) = I\frac{1}{C}\frac{1}{(s_{\text{p1}} - s_{\text{p2}})}(\text{e}^{s_{\text{p1}}t} - \text{e}^{s_{\text{p2}}t})$$

which can be arranged as

$$\begin{aligned} v(t) &= I\frac{1}{C}\frac{1}{2\text{j}\bar{\omega}_0}\text{e}^{\sigma t}(\text{e}^{\text{j}\bar{\omega}_0 t} - \text{e}^{-\text{j}\bar{\omega}_0 t}) \\ &= I\frac{1}{C}\frac{1}{\bar{\omega}_0}\text{e}^{\sigma t}\sin\bar{\omega}_0 t, \end{aligned}$$

and with $\sigma = -(\omega_0/2Q_{\text{eff}})$ and $\omega_0 \cong \bar{\omega}$,

$$v(t) = V_{\text{init}}\text{e}^{-(\bar{\omega}_0/2Q_{\text{eff}})t}\sin\bar{\omega}_0 t, \tag{4.23}$$

where V_{init} indicates the initial voltage amplitude. Clearly, the rate of attenuation (damping) of the voltage depends inversely on the quality factor Q.

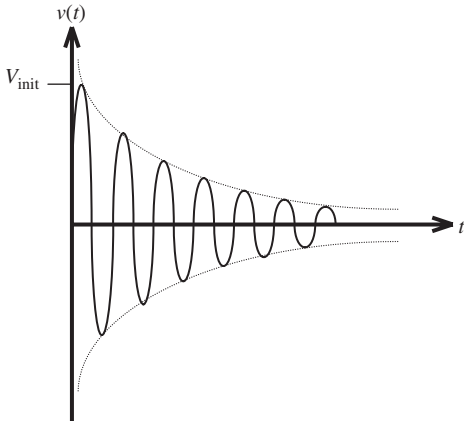


Figure 4.6 The damped oscillation of an excited parallel resonance circuit.

The variation of the voltage of an excited parallel resonance circuit is shown in Fig. 4.6. Two important properties of this damped oscillation can be extracted from (4.23):

- at the end of $n = Q_{\text{eff}}$ oscillations, the amplitude decreases from V_{init} to $V_{\text{init}}e^{-\pi}$;
- the number of oscillations[5] corresponding to $\hat{v} = (1/e)V_{\text{init}}$ is equal to (Q_{eff}/π).

4.1.1.3 The "*Q* enhancement"

There is a relatively straightforward possibility for increasing the low quality factor of an on-chip resonance circuit: connecting a "negative conductance" in parallel to the resonance circuit. We know that the Q factor of the resonance circuit at its resonance frequency (ω_0) can be expressed as $Q_{\text{eff}} = 1/L\ \omega_0 G_{\text{eff}}$. If we connect a negative conductance ($-G_{\text{n}}$) in parallel to G_{eff}, the enhanced effective parallel conductance becomes $G'_{\text{eff}} = (G_{\text{eff}} - G_{\text{n}})$. This decrease of the parallel conductance consequently increases the Q factor to $Q'_{\text{eff}} = 1/L\ \omega_0\ (G_{\text{eff}} - G_{\text{n}})$.[6]

Although this negative conductance, which is usually introduced by an appropriate electronic circuit, adds an additional noise and nonlinearity, it can be effectively used to increase the quality factor of a resonance circuit, [26], [27]. One of the simplest negative resistance circuits is a capacitive loaded source follower. In Section 3.2 we have seen that the input resistance of a source follower can be negative under appropriate conditions, and can be used to enhance the quality factor of a resonance circuit.

[5] In the 1960s Rohde and Schwarz introduced a Q-meter (QDM) whose operation principle was based on these interesting relations.

[6] Note that for $G_{\text{n}} = G_{\text{eff}}$ the quality factor becomes infinite. According to (4.23), once the circuit is excited, the magnitude of the oscillation remains constant at its initial value. It means that the circuit operates as a sinusoidal oscillator. It can be shown that in all types of L-C oscillator there exists a negative conductance parallel to the L-C circuit introduced by a positive-feedback circuit or by a device, like a tunnel diode, that inherently exhibits a negative conductance (see Chapter 5).

Example 4.2 Consider the problem of increasing the $Q = 10$ of an on-chip parallel resonance circuit to 20. The resonance frequency, the inductance, and the capacitance are given as $f_0 = 2$ GHz, $L = 10$ nH and $C = 0.633$ pF, respectively. The effective parallel conductance of the resonance circuit can be calculated as

$$G_{\text{eff}} = \frac{1}{L\omega_0 Q} = \frac{1}{10\times10^{-9}\times2\pi\times(2\times10^{9})} = 7.96\times10^{-4}$$

which corresponds to $R_{\text{eff}} = 1256.6$ ohm. It is obvious that to increase the Q factor to 20, the effective conductance must be decreased to $G'_{\text{eff}} = 3.98\times10^{-4}$ S and the necessary negative conductance to achieve this is $G_{\text{n}} = -3.98\times10^{-4}$ S.

From (3.17a) we see that the negative input conductance of a source follower is maximum for $g_{\text{m}}C \gg GC_{\text{gs}}$ and for $\omega \gg \omega_{\text{p}}$, and becomes equal to

$$g_{\text{i}}(\infty) = -\frac{CC_{\text{gs}}}{(C + C_{\text{gs}})^2} g_{\text{m}}$$

and has its maximum value for $C = C_{\text{gs}}$. Under this condition the pole frequency given in (3.15) becomes

$$\omega_{\text{p}} = \frac{g_{\text{m}} + G}{2C_{\text{gs}}} \cong \frac{g_{\text{m}}}{2C_{\text{gs}}} \cong \frac{3}{2}\frac{\mu(V_{\text{GS}} - V_{\text{T}})}{L^2}$$

for $g_{\text{m}} \gg G$, which can be satisfied by using a current source instead of R, as shown in Fig 2.15. From the last two expressions we obtain, for $C = C_{\text{gs}}$

$$g_{\text{i}}(\infty) \cong -\frac{g_{\text{m}}}{4}$$

which offers the first design hint.

From Fig. 3.12, we see that ω_0 must be chosen where the slope of the curve is small, or in other words, close to the asymptote, in order to obtain a high negative conductance with a small sensitivity. For $\omega > 3\omega_{\text{p}}$ the slope of the curve (or the sensitivity of the negative conductance against the frequency) is sufficiently small. This means that for $\omega_0 > 3\omega_{\text{p}}$, the value of the negative conductance becomes approximately equal to $g_{\text{i}}(\infty)$.

The aspect ratio of the transistor can be calculated in terms of G_{n} and the gate bias voltage. From (3.17d) and (1.33)

$$g_{\text{m}} = 4|G_{\text{n}}| \cong \mu_{\text{n}}C_{\text{ox}}(W/L)(V_{\text{GS}} - V_{\text{T}})$$
$$(W/L) = \frac{4|G_{\text{n}}|}{\mu_{\text{n}}C_{\text{ox}}(V_{\text{GS}} - V_{\text{T}})}$$

which indicates a trade-off between gate overdrive and the aspect ratio. $(V_{\text{GS}} - V_{\text{T}}) = 1$ V is a reasonable value for the AMS 0.35 micron technology, for which the maximum supply voltage is given as 3.3 V, and the aspect ratio can be calculated as

$$(W/L) = \frac{4|G_{\text{n}}|}{\mu_{\text{n}}C_{\text{ox}}(V_{\text{GS}} - V_{\text{T}})} = \frac{4(3.98\times10^{-4})}{374(4.54\times10^{-7})1} = 9.37$$

For this aspect ratio and $(V_{GS} - V_T) = 1\,V$, the drain current is

$$I_D \cong \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_T) = \frac{1}{2}374\times 9.3(4.54\times 10^{-7}) = 0.79\,mA$$

Now L can be calculated from (3.15a). With $\omega_0 = 4\omega_p$ as a value that satisfies $\omega_0 > 3\omega_p$,

$$L = \sqrt{\frac{4.\frac{3}{2}\mu_n(V_{GS} - V_T)}{\omega_0}} = \sqrt{\frac{6\times 374\times 1}{2\pi\times(2\times 10^9)}} = 4.2\times 10^{-4}cm \cong 4.2\,\mu m$$

and

$$W = 9.37\times 4/2 \cong 39\,\mu m$$

The circuit diagram together with the current source is given in Fig. 4.7. The V_{DD} supply is 3.3 V. The gate of M1 is directly biased from V_{DD}. To satisfy $(V_{GS1} - V_T) = 1\,V$ (or $V_{GS1}) = 1.5\,V$, the DC voltage of the source of M1 must be 1.8 V, which is the drain–source voltage of the current source transistor, M2. M2 must be biased such that its current is also 0.79 mA and is in the saturation region. For $V_{GS2} = 1\,V$, since $I_{D1} = I_{D2}$, the aspect ratio of M2 can be calculated as

$$\left(\frac{W}{L}\right)_2 = \left(\frac{W}{L}\right)_1 \frac{(V_{GS1} - V_T)(1 + \lambda_n.V_{DS1})}{(V_{GS2} - V_T)(1 + \lambda_n.V_{DS2}} = 9.37\frac{(1.5 - 0.5)^2(1 + 0.073\times 1.5)}{(1 - 0.5)^2(1 + 0.073\times 1.8)} \cong 37$$

and with $L_2 = 0.35\,\mu m$, $W_2 = 13\,\mu m$.

The final value to be calculated is C which is equal to C_{gs1}:

$$C_{gs1} \cong \frac{2}{3}W_1L_1C_{ox} = (39\times 10^{-4})(3.5\times 10^{-4})(4.54\times 10^{-7}) = 495\,fF$$

The designed Q-enhancement circuit is simulated with PSpice. To adjust the current of M1 to 0.79 mA, it is necessary to adjust the gate voltage of the current source M2 to 1.22 V. The frequency characteristics of the original resonance circuit and the Q-enhanced circuit are given in Fig. 4.7(b) as curve A and curve B, respectively. The quality factor calculated from curve B is 17.8, which corresponds to 78% increase of the quality factor, but not 100% as targeted. One of the reasons is the loss associated with the body resistances of the devices that were not included in the analytical expressions. The second reason is the value of g_m, which is usually smaller than the calculated value, as discussed in Chapter 1. To compensate it the channel width can be increased in order to reach the target value.

Another valuable possibility for this simple Q-enhancement circuit shown in Fig. 4.7(a) is to control the transconductance of M1 (and hence the negative conductance) with the gate bias voltage of the current source, M2. In Fig. 4.7(c) the frequency characteristics of the circuit for different values of V_{G2} are shown. Note that

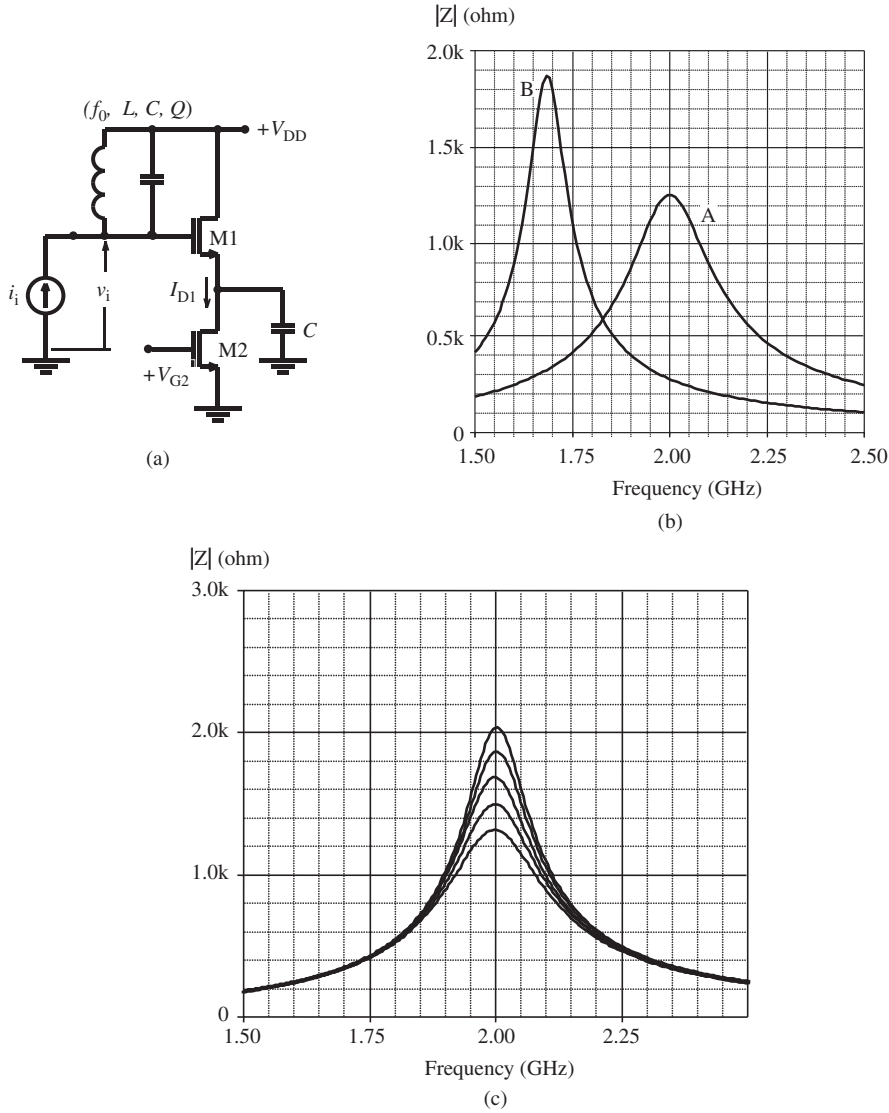


Figure 4.7 (a) The complete circuit diagram of the *Q*-enhancement circuit. (b) The frequency characteristic of the original resonance circuit (A) and the *Q*-enhanced circuit (B). (c) *Q* adjustment of the resonance circuit (curves correspond to $V_{G2} = 0.6$ V to 1.4 V with 0.2 V intervals).

the value of *C* has been decreased to compensate the effect of the input capacitance of M1 and to bring the resonance frequency to the target value.

Problem 4.2. *Make transient simulations of the original circuit and the Q-enhanced circuit in Example 4.1 for sinusoidal input currents with an amplitude of 1 mA and 5 mA. Compare and interpret the results.*

4.1.1.4 Bandwidth of a parallel resonance circuit

We already know that a parallel resonance circuit has frequency-selective behavior. Its impedance is maximum at its resonance frequency and decreases below and above this frequency. Owing to the series resistances of the inductance and/or capacitance branches there may exist a skew, as seen in Fig. 4.3(a). We also know that this asymmetry is associated with a shift of the zero crossing frequency of the phase, from the top of the magnitude curve. Keeping in mind these imperfections, we prefer the conventional approach to obtain universal expressions for the bandwidth, assuming that the quality factor of the circuit is sufficiently high, or the series resistances of the inductance and capacitance branches sufficiently balance each other, as seen in Section 4.1.1.1. Since the equivalent circuit in Fig. 4.5(b) can represent both of these cases, we will base our derivations on this circuit.

The impedance of the circuit was obtained as (4.21), which can be written in the frequency domain as

$$Z = R_{\text{eff}} \frac{1}{1 + \text{j}\frac{R_{\text{eff}}}{L\omega}(\omega^2 LC - 1)}$$

With $LC = 1/\omega_0^2$ and $Q = R_{\text{eff}}/L\omega_0$ this expression can be arranged as

$$Z = R_{\text{eff}} \frac{1}{1 + \text{j}Q\frac{\omega_0}{\omega}\left(\frac{\omega^2}{\omega_0^2} - 1\right)} \cong R_{\text{eff}} \frac{1}{1 + \text{j}\beta Q} \tag{4.24a}$$

where

$$\beta = 2\Delta\omega/\omega_0 \tag{4.24b}$$

and $\Delta\omega$ is the difference from the resonance frequency. From (4.24) the magnitude and phase of the impedance can be written as

$$|Z| = Z(\omega_0) \frac{1}{\sqrt{1 + (\beta Q)^2}} \quad \Phi = -\arctan(\beta Q) \tag{4.24c}$$

It can be seen from these expressions that:

- at resonance, the magnitude of the impedance is maximum and equal to R_{eff}, and from (4.20)

$$Z(\omega_0) = R_{\text{eff}} = Q_{\text{eff}} L\omega_0$$

- if the effective quality factor is known, a higher inductance value results in a higher resonance impedance. This is an important fact that is useful to obtain higher voltage gain from a tuned amplifier, as will be shown later on;
- for $\beta Q = \pm 1$ the magnitude of the impedance decreases to $Z(\omega_0)/\sqrt{2}$, which corresponds to the -3 dB frequencies of the magnitude curve:

$$f_{(-3\text{dB})} = f_0 \pm \frac{f_0}{2Q} \text{ and } B = \frac{f_0}{Q} \tag{4.25}$$

- the phase angle of the impedance is zero at the resonance frequency (the impedance is resistive), as expected;
- below the resonance frequency the phase angle is positive (impedance is inductive) and above f_0 the impedance is capacitive. The phase angles corresponding to the band ends (-3 dB frequencies) are $\mp\pi/4$. The normalized magnitude and phase characteristics of Z are given in Fig. 4.8.

4.1.1.5 Currents of L and C branches of a parallel resonance circuit

Assume that the input current of the parallel resonance circuit shown in Fig. 4.9 under resonance is $i(\omega_0)$. The voltage between the terminals of the circuit is $v(\omega_0) = i(\omega_0)R_{\text{eff}}$. The current of the capacitive branch is

$$\begin{aligned} i_{\text{C}}(\omega_0) &= v(\omega_0)(\text{j}C\omega_0) = i(\omega_0)R_{\text{eff}}(\text{j}C\omega_0) \\ &= \text{j}[i(\omega_0) \times Q_{\text{eff}}] \end{aligned} \tag{4.26}$$

Note that:

- the current flowing through the capacitance branch leads the input current by 90°;
- the magnitude of the current flowing through the capacitance branch is Q_{eff} times larger than the input current.

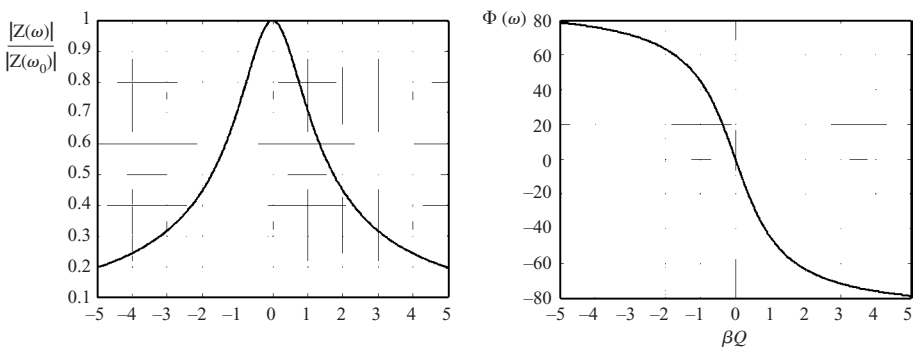


Figure 4.8 The normalized magnitude and phase characteristics of a parallel resonance circuit.

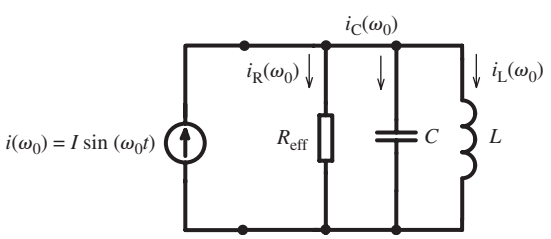


Figure 4.9 The branch currents of a parallel resonance circuit under resonance.

Similarly the current of the inductive branch can be written as

$$i_{\mathrm{L}}(\omega_0) = \frac{v(\omega_0)}{\mathrm{j}L\omega_0} = \frac{i(\omega_0)R_{\mathrm{eff}}}{\mathrm{j}L\omega_0} = -\mathrm{j}[i(\omega_0)\times Q_{\mathrm{eff}}] \tag{4.27}$$

Note that:

- the current flowing through the inductance branch lags 90° behind the input current;
- the magnitude of the current flowing through the inductance branch is Q_{eff} times larger than the input current. This phenomenon can lead to a long-term effect called "electromigration", which means the weakening of a conductor owing to the momentum transfer between the electrons and the metal atoms where the current density exceeds a certain value (approximately 2 mA/μm^2 for aluminum[7]). Note that a similar problem also exists for the interconnect lines of the capacitor in the resonance circuit.

Problem 4.3 *Calculate the branch currents of the circuit shown in Fig. 4.5(a), assuming realistic values for the on-chip passive components and their resistive parasitics.*

4.1.2 The series resonance circuit

A series resonance circuit is formed by connecting a capacitor and an inductor, as shown in Fig. 4.10(a). The series resistances representing the losses of the inductor and the capacitor are shown as r_{L} and r_{C}, and r_{S} is the internal resistance of the voltage source. The circuit can be simplified as shown in Fig. 4.10(b), where r_{eff} is the sum of r_{L}, r_{C} and r_{S}. The impedance of the circuit is

$$Z = (r_{\mathrm{eff}} + sL + \frac{1}{sC}) \tag{4.28}$$

and the admittance

$$\begin{aligned} Y &= \frac{1}{r_{\mathrm{eff}} + sL + \frac{1}{sC}} = \frac{1}{L}\frac{s}{s^2 + s\frac{r_{\mathrm{eff}}}{L} + \frac{1}{LC}} \\ &= \frac{1}{L}\frac{s}{(s - s_{\mathrm{p1}})(s - s_{\mathrm{p2}})} \end{aligned} \tag{4.29}$$

The poles of this admittance function are the same as the poles of the impedance of a parallel resonance circuit:

$$s_{\mathrm{p1,p2}} = \sigma \mp \mathrm{j}\sqrt{\omega_0^2 - \sigma^2} \tag{4.30}$$

where

$$\omega_0^2 = \frac{1}{LC}, \quad \sigma = -\frac{r_{\mathrm{eff}}}{2L} = -\frac{\omega_0}{2Q_{\mathrm{eff}}}, \quad Q_{\mathrm{eff}} = \frac{\omega_0}{2\sigma} = \frac{L\omega_0}{r_{\mathrm{eff}}} \tag{4.31}$$

[7] This value is given for DC currents. It is known that the electromigration is less effective for AC, and at high frequencies. But current crowding owing to the skin effect and the proximity effects must be considered.

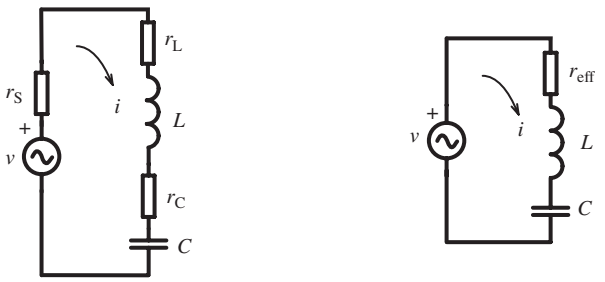


Figure 4.10 (a) The series resonance circuit, (b) its simplified form.

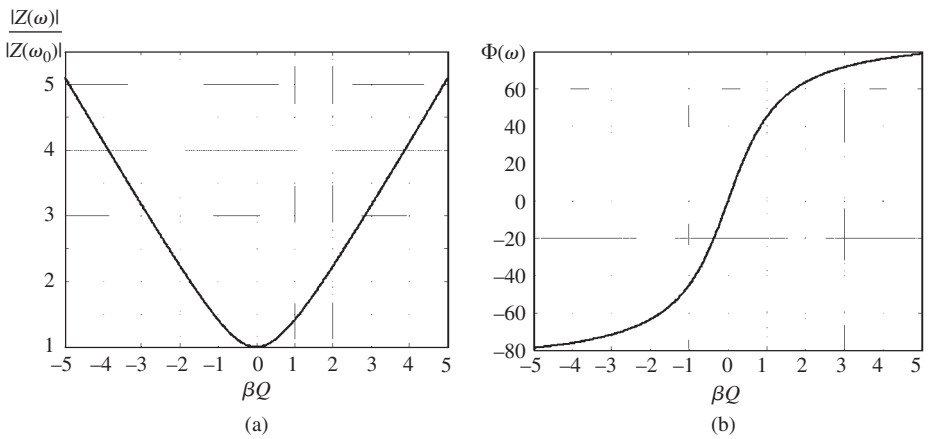


Figure 4.11 Normalized magnitude and phase characteristics of a series resonance circuit.

Now, the impedance of the circuit can be written as

$$Z(\omega) = Z(\omega_0)(1 + j\beta Q_{\text{eff}}) \tag{4.32}$$

where $\beta = 2\Delta\omega/\omega_0$ and $\Delta\omega$ represents the difference from the resonance frequency, similar to the parallel resonance circuit. The magnitude and phase of the impedance are

$$|Z(\omega)| = Z(\omega_0)\sqrt{1 + (\beta Q_{\text{eff}})^2} \quad \Phi(\omega) = \arctan(\beta Q_{\text{eff}}) \tag{4.33}$$

The normalized variations of the magnitude and the phase of the impedance are given in Fig. 4.11. From expressions (4.33) and Fig. 4.11 it can be seen that:

- the magnitude of the impedance is minimum and equal to r_{eff} at ω_0;
- for $\beta Q = \pm 1$ the magnitude of the impedance increases to $\sqrt{2}.Z(\omega_0)$. This can be expressed as the fact that the $+3$ dB frequencies of the magnitude are

$$f_{(+3\text{dB})} = f_0 \pm \frac{f_0}{2Q} \text{ and the bandwidth } B = \frac{f_0}{Q} \tag{4.34}$$

- the phase angle of the impedance is zero at the resonance frequency (the impedance is resistive), as expected;
- below the resonance frequency the phase angle is negative (impedance is capacitive) and above f_0 the impedance is inductive. The phase angles corresponding to the band ends (+3 dB frequencies) are $\mp\pi/4$.

4.1.2.1 Component voltages in a series resonance circuit

Consider a series resonance circuit as shown in Fig. 4.12. The current at resonance is $i(\omega_0) = v/r_{\text{eff}}$, where $r_{\text{eff}} = r_S + r_L + r_C$. Then the voltage between the terminals of the inductor can be written as

$$\begin{aligned} v_L(\omega_0) &= i(\omega_0)(r_L + \mathrm{j}L\omega_0) = v(\omega_0)\frac{r_L + \mathrm{j}L\omega_0}{r_{\text{eff}}} \\ &= \mathrm{j}.v(\omega_0)\left(\frac{r_L}{r_{\text{eff}}} + Q_{\text{eff}}\right) \cong \mathrm{j}.v(\omega_0)\times Q_{\text{eff}} \end{aligned} \tag{4.35}$$

and similarly,

$$v_C(\omega_0) \cong -\mathrm{j}.v(\omega_0)\times Q_{\text{eff}} \tag{4.36}$$

Note that

- the voltage between the terminals of the inductor is approximately Q_{eff} times larger than the input voltage and leads the input voltage 90°;
- the voltage on the capacitor is approximately Q_{eff} times larger than the input voltage and lags 90° behind the input voltage. This is an important phenomenon which has to be taken into account to prevent the breakdown of the capacitor dielectric, especially if the capacitor is a MOS capacitor or MOS varactor.

4.2 Tuned amplifiers

In all wireless applications, we need "tuned" or "narrow-band" amplifiers that provide gain at a certain frequency and in a narrow band around this frequency. The tuning

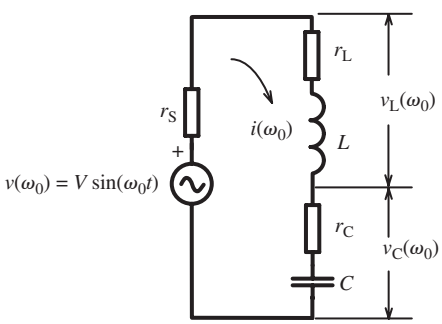


Figure 4.12 Voltages in a series resonance circuit at resonance.

frequency and the bandwidth depend on the area of application. Since the early days of the radio, resonance circuits have been the main components of tuned amplifiers, owing to their frequency-selective nature.

We have seen that the voltage gain of any amplifier is proportional to the total load impedance, which is the parallel equivalent of the external load and the output impedance of the amplifier. We also know that the impedance of a parallel resonance circuit is maximum at a certain frequency that is approximately equal to the natural frequency (or the resonance frequency) of the circuit. Then the easiest way to form a tuned amplifier is to use a parallel resonance circuit as the load of the amplifier such that the resonance frequency of the load together with the output impedance of the amplifier is equal to the desired tuning frequency. In some applications, the tuning frequency must be fixed as in the intermediate frequency (IF) amplifier of a receiver. But in some other applications, such as the input amplifier of a radio, the tuning frequency must be adjustable in a certain frequency band. To tune the frequency of an amplifier, the most commonly used way is to incorporate a suitable varactor into the resonance circuit.[8]

The bandwidth of the amplifier is certainly determined by the frequency characteristic of the overall (effective) load. For some applications the bandwidth and the shape of the frequency characteristic in this band do not fulfill the needs of the application. There are several techniques to improve the shape of the frequency characteristic and to increase the relative bandwidth in certain applications. The use of coupled resonance circuits instead of the simple parallel resonance circuits as load and the "staggered tuning" of the stages of an amplifier are among the well known and extensively used solutions.

In principle, any high-output impedance amplifier configuration can be used as the gain block of a tuned amplifier. The simplest and one of the most frequently used configuration, especially at the lower end of the RF spectrum, is the common-source tuned amplifier.

4.2.1 The common-source tuned amplifier

The circuit diagram of a common-source tuned amplifier is given in Fig. 4.13(a). The parallel L-C circuit is connected between the drain of the transistor and V_{DD}. The DC load of the transistor is the DC resistance of the inductor, and the AC load is the impedance of this load at the frequency of operation of the circuit. The input signal source is represented by a non-ideal voltage source having an internal resistance R_S, and is connected to the gate of the transistor via a coupling capacitor C_c, which exhibits a low reactance at the frequency of operation, and can be considered short-circuited for the signal. The gate DC bias voltage of the transistor is applied via a high value resistance R_G that can be considered open-circuited for the signal. The input impedance of the following stage at the frequency of operation is represented by a parallel combination r_{i2},C_{i2}.

[8] Formerly, mechanically controlled "variable capacitors" were being used in non-integrated tuned amplifiers.

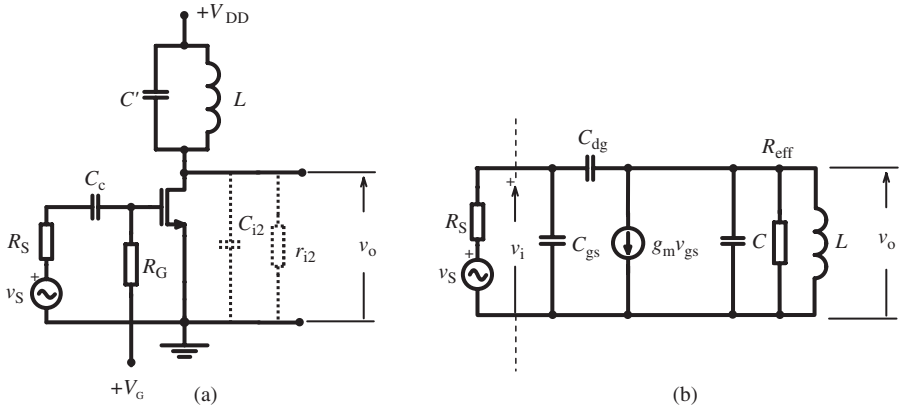


Figure 4.13 (a) Schematic diagram of a tuned common-source amplifier. (b) The small-signal equivalent circuit. C represents the total capacitance (the sum of the external capacitor C', the output capacitance of the transistor, the input capacitance of the following stage and the parasitics). R_{eff} is the parallel equivalent of the output resistance of the transistor, the input resistance of the following stage and the parallel resistance corresponding to the losses of the resonance circuit.

The small-signal equivalent circuit of the amplifier is given in Fig. 4.13(b), where the signal voltage at the input of the transistor is shown by v_{i}. This voltage certainly depends on the input resistance (in general, impedance) of the signal source and the input impedance of the amplifier, at the frequency of operation. Therefore, to characterize the amplifier, the voltage gain alone is not sufficient; the input impedance (or admittance) also must be investigated. The voltage gain ($v_{\text{o}}/v_{\text{i}}$) of a common-source amplifier for any type of load was found to be (see Eqn. (3.2)):

$$A_{\text{v}} = -\frac{g_{\text{m}} - sC_{\text{dg}}}{Y_{\text{o}}}$$

where Y_{o} represents the total output load admittance, that is the parallel equivalent of C, L and R_{eff} for our circuit;[9]

$$A_{\text{v}} = -\frac{g_{\text{m}} - sC_{\text{dg}}}{sC + \frac{1}{sL} + G_{\text{eff}}} \tag{4.37}$$

To derive the variation of the magnitude and the phase of the gain with frequency, it is possible to write (4.37) in the frequency domain, or to use the pole–zero diagram of the gain function. In the following, we prefer to examine the pole–zero diagram.

[9] It is obvious that for $\omega_0 C_{\text{dg}} \ll g_{\text{m}}$, the voltage gain can be written as $A_{\text{v}} = -g_{\text{m}}/Y_{\text{o}} = -g_{\text{m}}Z_{\text{o}}$. Consequently the frequency characteristic and the –3 dB frequencies of the amplifier are the same as those of the total load impedance, Z_{o}. Here a less straightforward approach will be used to enable us to discuss the effects of the "zero" of the gain function and to prepare the reader for the concepts that will be used to investigate the staggered tuning in Section 4.3.

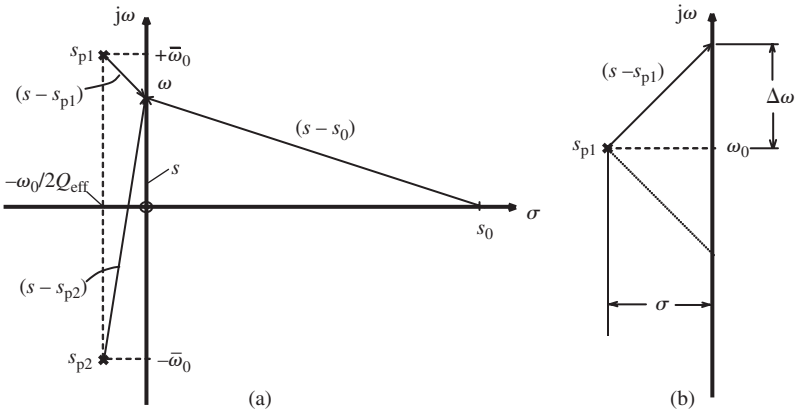


Figure 4.14 (a) The pole–zero diagram of the gain function of a tuned amplifier, (b) its simplified form for the vicinity of the resonance frequency (note that $(s - s_{p1})$ is drawn for the upper 3 dB frequency of the gain).

Equation (4.37) can be arranged in terms of its poles and zero as

$$A_v = \frac{C_{dg}}{C}\frac{s(s - s_0)}{(s - s_{p1})(s - s_{p2})} \tag{4.38}$$

where

$$s_0 = +\frac{g_m}{C_{dg}},\quad s_{p1,\,p2} = -\frac{G_{eff}}{2C} \mp j\sqrt{\frac{1}{LC} - \left(\frac{G_{eff}}{2C}\right)^2} \tag{4.39a}$$

and with

$$\sigma = -\frac{G_{eff}}{2C} = -\frac{\omega_0}{2Q_{eff}} \quad \text{and } \omega_0^2 = \frac{1}{LC}\ s_0 = +\frac{g_m}{C_{dg}},\quad s_{p1,p2} = \sigma \mp j\sqrt{\omega_0^2 - \sigma^2} \tag{4.39b}$$

The pole–zero diagram of the gain function is given in Fig. 4.14(a). For any ω value the magnitude and phase of the gain can be obtained as

$$|A| = \frac{C_{dg}}{C}\frac{|s||s - s_0|}{|s - s_{p1}||s - s_{p2}|} \tag{4.40a}$$

$$\Phi = \Phi_s + \Phi_{(s-s_0)} - \Phi_{(s-s_{p1})} - \Phi_{(s-s_{p2})}. \tag{4.41a}$$

Provided that $|s_0| \gg \omega_0$ and $\omega_0 \gg \sigma$, which are valid for most practical cases, for the vicinity of ω_0 the magnitude and phase can be approximated as

$$|A| \cong \frac{C_{dg}}{C}\frac{|\omega||s_0|}{|s - s_{p1}|2\omega_0} = \frac{g_m}{2C|s - s_{p1}|} \tag{4.40b}$$

$$\Phi \cong \frac{\pi}{2} + \pi - \Phi_{(s-s_{p1})} - \frac{\pi}{2} = \pi - \Phi_{(s-s_{p1})} \tag{4.41b}$$

The simplified form of the pole–zero diagram of the amplifier corresponding to (4.40b) and (4.41b), valid for and around the resonance frequency, is shown in Fig. 4.14(b). From this figure the 3 dB frequencies and the bandwidth of the amplifier can be found as

$$f_{(-3\text{dB})} = f_0 \pm \Delta f = f_0 \pm \frac{f_0}{2Q_{\text{eff}}} \qquad B = 2\Delta f = \frac{f_0}{Q_{\text{eff}}} \tag{4.42}$$

which are the same as that of a parallel resonance circuit.

The magnitude of the gain corresponding to the resonance frequency can be calculated from (4.40b) with $|s - s_{\text{p1}}| = \sigma = \omega_0/2Q_{\text{eff}}$:

$$|A(\omega_0)| = g_{\text{m}} \frac{Q_{\text{eff}}}{\omega_0 C} = g_{\text{m}} R_{\text{eff}} \tag{4.43}$$

and similarly the phase angle for ω_0,

$$\Phi(\omega_0) = \pi \tag{4.44}$$

For the calculations above we assumed that the zero of the voltage gain related to the drain–gate capacitance is positioned far away on the right half-plane and is therefore negligible. This assumption corresponds to $\omega_0 C_{\text{dg}} \ll g_{\text{m}}$, which is usually valid. Under this assumption the voltage gain (2.28) can be written as

$$A(\omega) = -\frac{g_{\text{m}}}{Y_{\text{o}}} = -g_{\text{m}} Z_{\text{o}} \tag{4.45}$$

where Z_{o} is the effective impedance of the parallel resonance circuit.

Example 4.3 Check the validity of the assumption of $\omega_0 C_{\text{dg}} \ll g_{\text{m}}$ for a typical 0.13 micron NMOS transistor operating in the velocity saturation regime. The operating frequency of the amplifier is 3 GHz.

Under velocity saturation (which is the case for a 0.13 micron transistor, as shown in Chapter 1), the transconductance is $g_{\text{m(v-sat)}} = kWC_{\text{ox}}v_{\text{sat}}$ and the drain–gate capacitance $C_{\text{dg}} = W \times \text{CDGW}$. Therefore,

$$\frac{\omega_0 C_{\text{dg}}}{g_{\text{m}}} = \frac{\omega_0 \times \text{CDGW}}{C_{\text{ox}} v_{\text{sat}}}$$

which is independent of the gate width. The related parameter values for this 0.13 micron technology are $\text{TOX} = 2.3 \times 10^{-9}$ [m] (which corresponds to $C_{\text{ox}} = 15 \times 10^{-3}$ f/m^2), and $\text{CDGW} = 5.18 \times 10^{-10}$ [f/m]. The saturation velocity of electrons in the channel was given in Chapter 1 as 6.5×10^4 [m/s]. Therefore

$$\frac{\omega_0 C_{\text{dg}}}{g_{\text{m}}} = \frac{2\pi \times (3 \times 10^9) \times (5.18 \times 10^{-10})}{(15 \times 10^{-3}) \times (6.5 \times 10^4)} \cong 10^{-2}$$

which corresponds to an error of 1% on the magnitude of the gain and an excess phase shift of only 0.57°.

Problem 4.4 *An amplifier tuned to 1 GHz is designed using an AMS 0.35 NMOS transistor with W = 100 μm and L = 0.35 μm. The drain current is 5 mA. Calculate the gain and phase errors of this amplifier if the influence of C_{dg} is neglected.*

The numerical results of this example show that the effects of the drain–gate capacitance are negligibly small. The question can then arise as to why this capacitance is notorious for its adverse effects on tuned amplifiers. The answer to this question is related to the effects of C_{gs} on the input admittance of the amplifier and the risk of oscillation under certain conditions, as examined in the following.

The input admittance of a common-source amplifier was found as

$$y_i = s(C_{gs} + C_{dg}) + y_{mi} = s(C_{gs} + C_{dg}) + sC_{dg}\frac{g_m - sC_{dg}}{Y_o} \tag{3.3b}$$

The first term is apparently capacitive, but the second term (the Miller component) needs to be investigated. Let us write the Miller admittance in the frequency domain, and then calculate the real and imaginary parts:

$$y_{mi}(\omega) = j\omega C_{dg}\frac{g_m - j\omega C_{dg}}{G + j\omega C + \frac{1}{j\omega L}} = \omega^2 LC_{dg}\frac{-g_m + j\omega C_{dg}}{(1 - \omega^2 LC) + j\omega LG}$$

$$\mathrm{Re}\{y_{mi}\} = \frac{(\omega/\omega_0)^2 \times \frac{C_{dg}}{C}}{\left(1 - (\omega/\omega_0)^2\right)^2 + \omega^2 L^2 G^2}\left(-g_m\left(1 - (\omega/\omega_0)^2\right) + (\omega/\omega_0)^2\frac{C_{dg}}{C}G\right) \tag{4.46}$$

$$\mathrm{Im}\{y_{mi}\} = \frac{(\omega/\omega_0)^2 \times \frac{C_{dg}}{C}}{\left[1 - (\omega/\omega_0)^2\right]^2 + \omega^2 L^2 G^2}\omega\left\{C_{dg}\left[1 - (\omega/\omega_0)^2\right] + LGg_m\right\} \tag{4.47}$$

From (4.47) it is possible to see that

- at the resonance frequency (for $\omega = \omega_0$) the input conductance is

$$\mathrm{Re}\{y_{mi}(\omega_0)\} = \frac{\omega_0^2 C_{dg}^2}{G}$$

  which strongly depends on C_{dg};
- above the resonance frequency (where the load impedance is capacitive), the input conductance is positive and varies with frequency;
- below the resonance frequency (where the load impedance is inductive), the input conductance has a dominant negative component and varies with frequency.

This frequency-dependent input conductance and especially its negativity below the ω_0 resonance frequency of the output load is important from different points of view.

- In the case of a non-ideal input signal source, the signal voltage on the gate of the transistor changes with frequency. Therefore, the overall frequency characteristic is

determined not only by the output load, but also by the internal impedance of the signal source.

- If a tuned circuit exists in parallel to the input, owing to the positive parallel conductance above ω_0 and the negative parallel conductance below ω_0, the quality factor of this circuit decreases above ω_0 and increases below ω_0. The result is the skew of the frequency characteristic of the input resonance circuit, which affects the overall frequency characteristic of the circuit.
- The negative conductance component of the input admittance can over-compensate the losses of the input resonance circuit and can lead the circuit to oscillate.

According to (4.47) the imaginary part of the input admittance of a tuned amplifier also depends on the frequency. However, this is not as severe as the varying, negative input conductance. It only acts on the tuning of the input resonance circuit, if there is any.

To exemplify these observations the PSpice simulation results of a simple tuned amplifier are given in Fig. 4.15, using an AMS 0.35 NMOS transistor with $L = 0.35$ μm and $W = 200$ μm. Supplies are $V_{DD} = 3$ V and $V_G = 0.8$ V. $L = 10$ nH and C is trimmed to 2.34 pF to tune the circuit to $f_0 = 1$ GHz. The parallel resistance representing the total losses of the resonance circuit is 1 k ohm, which corresponds to $Q = 15.9$.

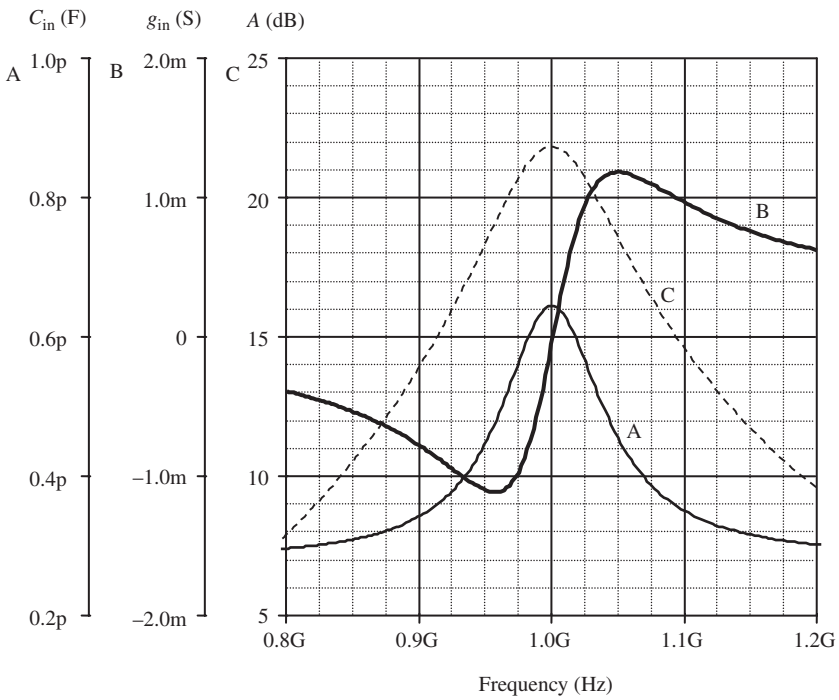


Figure 4.15 Variations of the input capacitance (A), the input conductance (B) and the voltage gain of the amplifier (C). Note the fluctuations of the input capacitance and the input conductance, and especially negativity of the input conductance below the resonance frequency.

Curves A and B show the variations of the input capacitance and the input conductance, respectively. The negative input conductance below f_0 and positive input conductance above f_0 are seen from curve B. The maximum values of the input conductance are 1.2 mS and correspond approximately to the 3 dB frequencies of the gain that is plotted as curve C.

These dramatic variations of the input admittance of a tuned MOS amplifier are obviously owing to the drain–gate capacitance of the device, which is unavoidable and whose adverse effects increase with frequency. Consequently, a circuit as shown in Fig. 4.13 can be used only at the lower end of the RF spectrum. For high-frequency RF amplifiers, the extensively used solution is the "cascode" circuit that was investigated in general in Section 3.4.

4.2.2 The tuned cascode amplifier

A cascode circuit loaded with a parallel resonance circuit is shown in Fig. 4.16. M1 and M2 are biased in the saturation region with V_{G1} and V_{G2}. The load of M1 is the input impedance of M2, operating as a common-gate circuit.

We know that the input impedance of a common-gate circuit is approximately equal to the parallel equivalent of $1/g_m$ and C_{sg}. Consequently the voltage gain of M1 is low and equal to $(-g_{m1}/g_{m2})$ up to the frequencies close to g_{m2}/C_{sg2}. Therefore the Miller component of the input admittance of M1 is considerably smaller compared to that of a high-gain common-source amplifier. In addition, since the output resistance of a cascode circuit is higher than the output resistance of a common-source circuit, the effective Q of the load becomes higher.

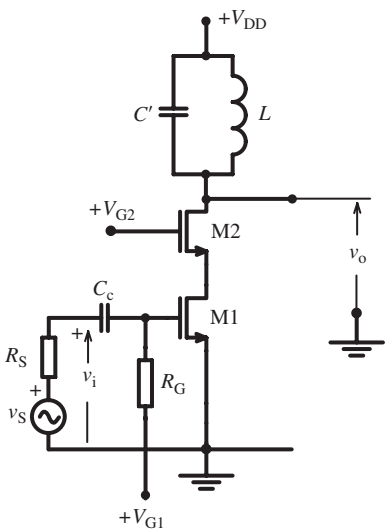


Figure 4.16 Schematic diagram of a tuned cascode amplifier.

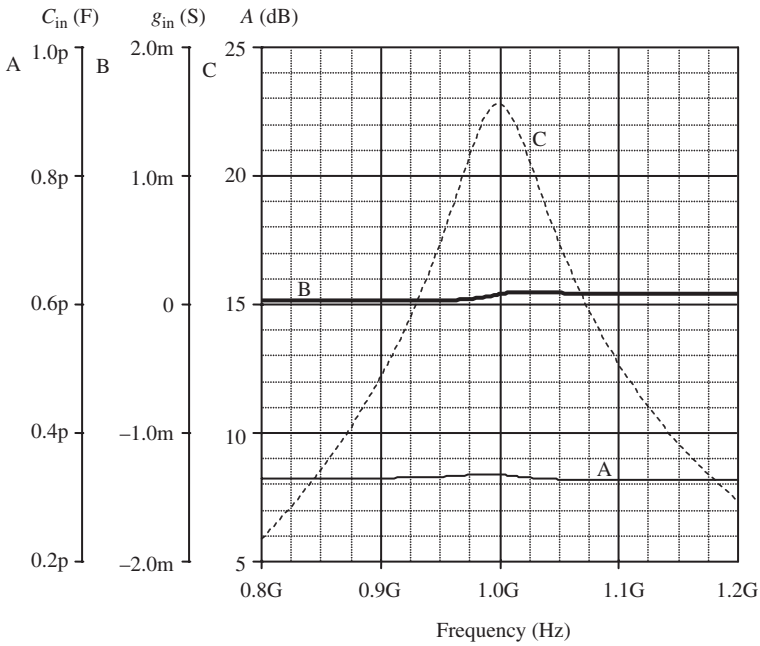


Figure 4.17 Variations of the input capacitance (A), the input conductance (B) and the voltage gain (C) of the tuned cascode amplifier. Note the almost constant input capacitance and the input conductance.

To visualize the benefits of the cascode configuration, the simulation results of a cascode amplifier are given in Fig. 4.17. The parameters of the circuit are the same as the parameters of the common-source amplifier, whose simulation results were given in Fig. 4.15:

M1 and M2: AMS 0.35 NMOS transistor. $L = 0.35$ μm, $W = 200$ μm.
Supplies: $V_{DD} = 3$ V and $V_{G1} = 0.8$ V, $V_{G2} = 1.5$ V,
$L = 10$ nH, $C' = 2.34$ pF ($f_0 = 1$ GHz).
Parallel resistance representing the total losses of L and C is 1 k ohm.

To ease the comparison, the vertical axes in Fig. 4.17 are intentionally chosen as the same as those of Fig. 4.15. The obvious advantages of the cascode circuit can be summarized as follows.

- The input capacitance is almost constant in the entire frequency band and equal to the input capacitance of M1.
- The input conductance is positive and almost constant in the entire frequency band. This means that the input of the circuit is a well-defined load for the driving signal source (or the previous stage) and has no adverse effect if there is another tuned circuit parallel to the input.
- The bandwidth of the gain is smaller (the effective Q is higher) compared to that of the reference common-source circuit. This is the result of the high output resistance of the common-gate output transistor M2, as expected.

4.3 Cascaded tuned stages and the staggered tuning

The voltage gain of a tuned amplifier in the s-domain was given as

$$A_{\mathrm{v}} = \frac{C_{\mathrm{dg}}}{C}\frac{s(s-s_0)}{(s-s_{\mathrm{p1}})(s-s_{\mathrm{p2}})} \tag{4.38}$$

From Fig. 4.14(a) it can be seen that in the vicinity of the resonance frequency and for $|s_0| \gg |s_{\mathrm{p1}}|$, $(s-s_{\mathrm{p2}}) \cong 2s$ and $(s-s_0) \cong -s_0 = \frac{g_{\mathrm{m}}}{C_{\mathrm{dg}}}$. Therefore (4.38) can be simplified as

$$A_{\mathrm{v}} \cong -\frac{C_{\mathrm{dg}}}{2C}\frac{s_0}{(s-s_{\mathrm{p1}})} \tag{4.48}$$

and with $A_{\mathrm{v}}(\omega_0) = -g_{\mathrm{m}}R_{\mathrm{eff}}$ and $Q_{\mathrm{eff}} = \omega_0 C R_{\mathrm{eff}}$,

$$A_{\mathrm{v}} \cong A_{\mathrm{v}}(\omega_0)\frac{\omega_0}{2Q_{\mathrm{eff}}}\frac{1}{(s-s_{\mathrm{p1}})} = A_{\mathrm{v}}(\omega_{\mathrm{p1}})\frac{\omega_{\mathrm{p1}}}{2Q_{\mathrm{eff}}}\frac{1}{(s-s_{\mathrm{p1}})} \tag{4.49}$$

If n identical stages are connected in cascade, the total voltage gain becomes

$$A_{\mathrm{vT}} \cong \left[A_{\mathrm{v}}(\omega_{\mathrm{p1}})\right]^n \left(\frac{\omega_0}{2Q_{\mathrm{eff}}}\right)^n \frac{1}{(s-s_{\mathrm{p1}})^n}$$

and the bandwidth of the amplifier shrinks to

$$B_{\mathrm{T}} = B\sqrt{2^{1/n}-1} \tag{4.50}$$

This is the appropriate solution if a high-gain and narrow-band amplifier is needed. But in some cases a relatively broad bandwidth and a flat frequency characteristic in this band are needed. It can be understood intuitively that tuning the stages of this multi-stage amplifier to slightly different frequencies around the center frequency of the band can lead to a feasible solution. In this case the gain of this multi-stage amplifier can be written as

$$A_{\mathrm{vT}} \cong A_{\mathrm{v1}}(\omega_{\mathrm{p1}}) \cdots A_{\mathrm{v}n}(\omega_{\mathrm{p}n})\frac{\omega_{\mathrm{p1}}}{2Q_{\mathrm{eff1}}} \cdots \frac{\omega_{\mathrm{p}n}}{2Q_{\mathrm{eff}n}}\frac{1}{(s-s_{\mathrm{p1}}) \cdots (s-s_{\mathrm{p}n})}$$

which has n poles.

The appropriate positions of the poles of the transfer function (the voltage gain in our case) to obtain a desired frequency characteristic are investigated in depth in classical filter theory [28]. It is known that the number of poles and their relative positions determine the bandwidth and the shape of the frequency characteristics. Among several possibilities for the distribution of the poles in the s-domain, the Butterworth distribution and the Chebyshev distribution have prime importance and extensive use in practice.

The Butterworth distribution provides a "maximally flat" frequency characteristic in the band. It has been shown that to obtain a Butterworth-type frequency characteristic, poles must be on a semi-circle whose center is at ω_0 on the vertical axis of the s-plane and they must be symmetrically positioned with respect to the horizontal diameter of

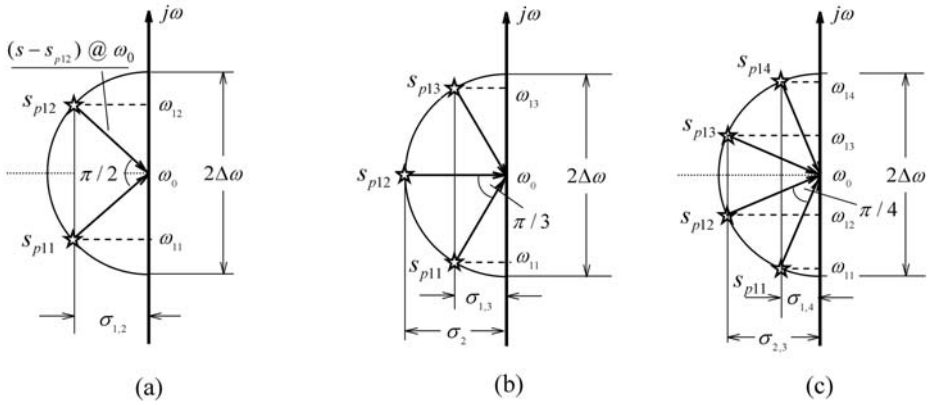


Figure 4.18 The appropriate positions of the poles of an (a) 2-pole, (b) 3-pole and (c) 4-pole circuit that has a maximally flat frequency characteristic.

the circle, with equal distances. The diameter of the circle on the $j\omega$ axis corresponds to the bandwidth of the circuit in angular frequency $2\Delta\omega$. The appropriate positions of the poles for $n = 2$, 3 and 4 are shown in Fig. 4.18,[10] where

$$\sigma_{1,2} = -\frac{\omega_{11}}{2Q_1} = -\frac{\omega_{12}}{2Q_2} \text{ for a 2-pole circuit}$$

$$\sigma_2 = -\frac{\omega_0}{2Q_2} \text{ and } \sigma_{1,3} = -\frac{\omega_{11}}{2Q_1} = -\frac{\omega_{13}}{2Q_3} \text{ for a 3-pole circuit}$$

$$\sigma_{1,4} = -\frac{\omega_{11}}{2Q_1} = -\frac{\omega_{14}}{2Q_4} \text{ and } \sigma_{2,3} = -\frac{\omega_{12}}{2Q_2} = -\frac{\omega_{13}}{2Q_3} \text{ for 4-pole circuit}$$

In the case of small relative bandwidths, where $2\Delta\omega \ll \omega_0$, the sigmas can be written as

$$\sigma_{1,2} \cong -\frac{\omega_0}{2Q_1} = -\frac{\omega_0}{2Q_2} \Rightarrow Q_1 \cong Q_2 \text{ for a 2-pole circuit}$$

$$\sigma_2 = -\frac{\omega_0}{2Q_2}, \sigma_{1,3} \cong -\frac{\omega_0}{2Q_1} = -\frac{\omega_0}{2Q_3} \Rightarrow Q_1 \cong Q_3 \text{ for a 3-pole circuit}$$

$$\sigma_{1,4} \cong -\frac{\omega_0}{2Q_1} = -\frac{\omega_0}{2Q_4}\ \sigma_{2,3} \cong -\frac{\omega_0}{2Q_2} = -\frac{\omega_0}{2Q_3}$$

$$\Rightarrow Q_1 \cong Q_4, Q_2 \cong Q_3 \text{ for a 4-pole circuit}$$

Example 4.4 A three-stage staggered tuned amplifier having a Butterworth-type frequency characteristic will be designed. The center frequency of the frequency characteristic is 1 GHz. The maximum possible effective Q value for the resonance circuits – without any Q-enhancement feature – is given as 20.

[10] It must not be overlooked that these diagrams are the simplified versions of the pole–zero diagrams, for the vicinity of the center frequency as shown in Fig. 4.14. The full pole–zero diagrams contain the complex conjugates of the poles shown in Fig. 4.18.

(a) What is the realizable bandwidth?
(b) Calculate the tuning frequencies of the stages.
(c) Calculate the appropriate effective Q values of the resonance circuits.

Solution:

The pole–zero diagram of the voltage gain function of the amplifier is shown in Fig. 4.18(b). From the geometry of the figure it can be easily seen that the magnitude of the real part of the center pole, p_{12}, must be equal to the half of the bandwidth:

$$|\sigma_2| = \left|\frac{\omega_0}{2Q_2}\right| = \Delta\omega \Rightarrow Q_2 = \frac{\omega_0}{2\Delta\omega} = \frac{f_0}{2\Delta f}$$

Similarly, the real parts of p_{11} and p_{13} must be equal in magnitude to $\Delta\omega/2$:

$$|\sigma_1| = \left|\frac{\omega_0}{2Q_1}\right| = \frac{\Delta\omega}{2} \Rightarrow Q_1 = \frac{\omega_{01}}{\Delta\omega} = \frac{f_{01}}{\Delta f}$$
$$|\sigma_3| = \left|\frac{\omega_0}{2Q_3}\right| = \frac{\Delta\omega}{2} \Rightarrow Q_3 = \frac{\omega_{03}}{\Delta\omega} = \frac{f_{03}}{\Delta f}$$

ω_{01} and ω_{03} can be calculated from the geometry:

$$\omega_{01} = \omega_0 - \Delta\omega\cos(\pi/6)$$
$$\omega_{03} = \omega_0 + \Delta\omega\cos(\pi/6)$$

Since ω_{03} is the highest among the three tuning frequencies, the quality factor corresponding to this resonance circuit is the highest and must be equal to the possible maximum Q value, which is 20:

$$Q_3 = \frac{\omega_{03}}{\Delta\omega} = \frac{\omega_0 + \Delta\omega\cos(\pi/6)}{\Delta\omega} = \frac{\omega_0}{\Delta\omega} + \cos(\pi/6) = \frac{f_0}{\Delta f} + 0.866 = 20$$

which yields $\Delta f = 52.26$ MHz ($\Delta\omega = 328.36$ rad/s). Now the tuning frequencies and the quality factors can be calculated as

$$f_{01} = f_0 - \Delta f\cos(\pi/6) = 1000 - 52.26\times 0.866 = 954.74\ [\text{MHz}]$$
$$Q_1 = \frac{f_{01}}{2\Delta f} = \frac{954.74}{52.26} = 18.27$$

$$f_{02} = f_0 = 1000\ [\text{MHz}]\quad Q_2 = \frac{f_0}{2\Delta f} = \frac{1000}{2\times 52.26} = 9.57$$

$$f_{03} = f_0 + \Delta f\cos(\pi/6) = 1000 + 52.26\times 0.866 = 1045.26\ [\text{MHz}]$$
$$Q_3 = \frac{f_{03}}{\Delta f} = \frac{1045.26}{52.26} = 20$$

Note that since the quality factors are not high and the relative bandwidth ($2\Delta f/f_0$) is not small, Q_1 and Q_3 are not equal.

The normalized frequency characteristic of the circuit calculated and plotted with MATLAB is given in Fig. 4.19.

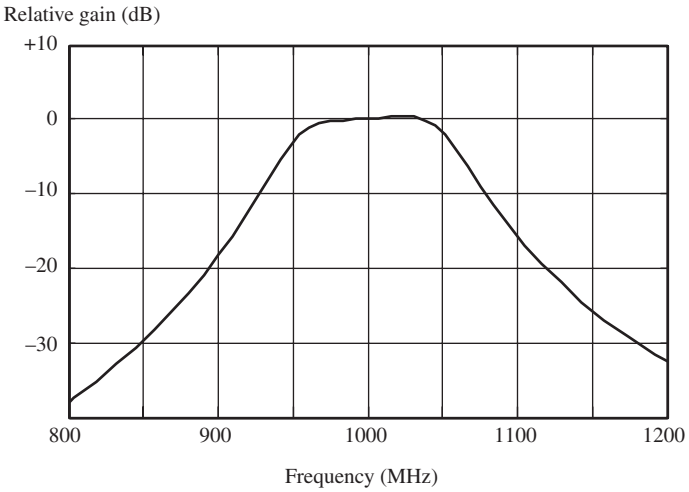


Figure 4.19 The calculated frequency characteristic of the circuit, calculated with MATLAB.

Problem 4.5 *The center frequency and the bandwidth of a 4-stage, staggered tuned amplifier are 2 GHz and 80 MHz, respectively. Calculate the tuning frequencies and the quality factors to obtain a Butterworth-type frequency characteristic.*

The second important type of pole distribution provides a Chebyshev-type frequency characteristic. The side-walls of a Chebyshev-type (or equi-ripple) characteristic are steeper than those of a same-order Butterworth-type characteristic, but exhibit a typical ripple on the top of the curve, as shown in Fig. 4.20(a). The number of ripples depends on the order of the circuit.

The poles of a Chebyshev-type circuit are positioned on an ellipse, whose longer axis is on the $j\omega$ axis and the length of the longer axis corresponds to the bandwidth of the circuit. It has been shown that the appropriate positions of the poles for a certain amount of ripple can be obtained from the positions of a Butterworth-type circuit that has the same bandwidth. As shown in Fig. 4.20(b), the tuning frequencies of the resonance circuits are the same, but the real parts of the poles of the Chebyshev-type circuit are smaller. For an nth order Chebyshev-type circuit with r(dB) ripple, the magnitude of σ of the Chebyshev pole can be calculated in terms of the corresponding Butterworth pole from

$$(\sigma_i)_C = \tanh a \times (\sigma_i)_B \tag{4.51a}$$

where

$$a = \frac{1}{n}\sinh^{-1}\frac{1}{\sqrt{\varepsilon}} \quad \varepsilon = \log^{-1}\frac{r(\text{dB})}{10} - 1 \tag{4.51b}$$

For convenience, the values of $\tanh a$ for $n = 2$, 3 and 4 and for several ripple values are given in Table 4.1.

Table 4.1

r (dB)	$n=2$	$n=3$	$n=4$
0.05	0.898	0.750	0.623
0.1	0.859	0.696	0.567
0.2	0.806	0.631	0.505
0.3	0.767	0.588	0.467
0.4	0.736	0.556	0.439
0.5	0.709	0.524	0.416

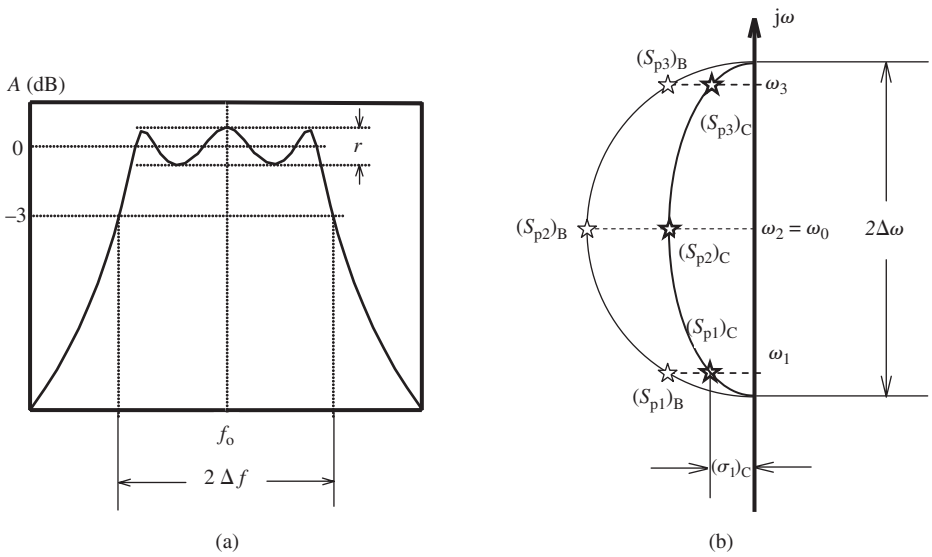


Figure 4.20 (a) The frequency characteristic of a third-order Chebyshev-type circuit. (b) The positioning of the poles of a Butterworth-type circuit and a Chebyshev-type circuit that have the same bandwidth.

Example 4.5 A three-stage staggered tuned amplifier will be designed with a center frequency of 2 GHz, having a voltage gain of 40 dB and a Chebyshev-type frequency characteristic with 0.5 dB ripple and 380 MHz bandwidth. The design will be made for a technology similar to the 0.35 micron AMS technology, but with an additional thick metal layer, allowing 10 nH inductors with a quality factor of 10 at 2 GHz. The Q values can be further increased with a Q-enhancement circuit similar to the circuit given in Example 4.1.

From Fig. 4.20(b) we see that the bandwidth of the circuit is $B = 2\Delta\omega = 2\left|(\sigma_2)_B\right|$, where $(\sigma_2)_B$ is the negative real center pole of a Butterworth-type circuit having the same center frequency and bandwidth. The center (real) Chebyshev pole can be calculated from (Fig. 4.18b) and Table 4.1:

$$B = 2\Delta\omega = 2\left|(\sigma_2)_B\right|$$

On the other hand,

$$\left|(\sigma_2)_C\right| = 0.524\left|(\sigma_2)_B\right| = 0.524\times\Delta\omega$$

Now Q_2 can be calculated as

$$Q_2 = \frac{\omega_0}{2\Delta\omega}\frac{1}{0.524} = \frac{f_0}{2\Delta f}\frac{1}{0.524} = 10.04 \cong 10$$

This means that the center pole can be realized without any Q-enhancement.

From Fig. 4.18(b) and Fig. 4.20(b) the tuning frequency and the quality factor corresponding to $(s_{p1})_C$ can be calculated as

$$f_1 = f_0 - \Delta f\times\cos(\pi/6) = 2000 - 190\times 0.866 = 1835.46\,[\text{MHz}]$$

$$Q_1 = \frac{\omega_1}{\left|2(\sigma_1)_C\right|},\ (\sigma_1)_C = 0.524(\sigma_1)_B,\ (\sigma_1)_B = \Delta\omega\times\sin(\pi/6)$$

$$Q_1 = 18.4$$

which correspond to a resonance impedance (effective parallel resistance) of

$$R_{1(\text{eff})} = L_1\omega_1 Q_1 = (10\times 10^{-9})\times(2\pi\times 1835.46\times 10^6)\times 18.35 = 2121\,\Omega$$

Similarly, the tuning frequency, the quality factor and the effective parallel resistance corresponding to $(s_{p3})_C$ can be found as

$$f_3 = 2164.5\,\text{MHz},\ Q_3 = 21.7\ R_{3(\text{eff})} = 2949\ \text{ohm}$$

Since the resonance circuit tuned to 2000 MHz was intended to be used without any Q-enhancement, its quality factor and the effective parallel resistance are $Q = 10$ and $R_{2(\text{eff})} = 1256$ ohm, respectively.

The basic circuit diagram is given in Fig. 4.21(a). As the individual stages of this amplifier, the cascode configuration is the natural choice, as explained in 4.2.2. The tuning frequencies of these stages will be f_1, $f_2 = f_0$ and f_3. To obtain the targeted voltage gain of 100 at the center frequency ($f_0 = f_2$), as well as frequencies corresponding to $(s_{p1})_C$ and $(s_{p3})_C$, the gains of the individual stages must be properly determined.

The total voltage gain of the amplifier at $f_0 = 2000$ MHz is equal to the multiplication of the voltage gains of the individual stages at this frequency. The DC currents and consequently the transconductances of the stages are chosen equal for convenience. The voltage gains of the individual stages at f_0 are

$$|A_{v1}(f_0)| = g_m|Z_1(f_0)|,\ |A_{v2}(f_0)| = g_m|Z_2(f_0)| \text{ and } |A_{v3}(f_0)| = g_m|Z_3(f_0)|$$

respectively. The load impedance of the second stage, which is tuned to f_0, is equal to the parallel equivalent resistance of the resonance circuit:

$$|Z_2(f_0)| = R_{2(\text{eff})} = L\omega_0 Q_2 = (10\times 10^{-9})\times(2\pi\times 2\times 10^9)\times 10 = 1257\ \text{ohm}$$

and the voltage gain

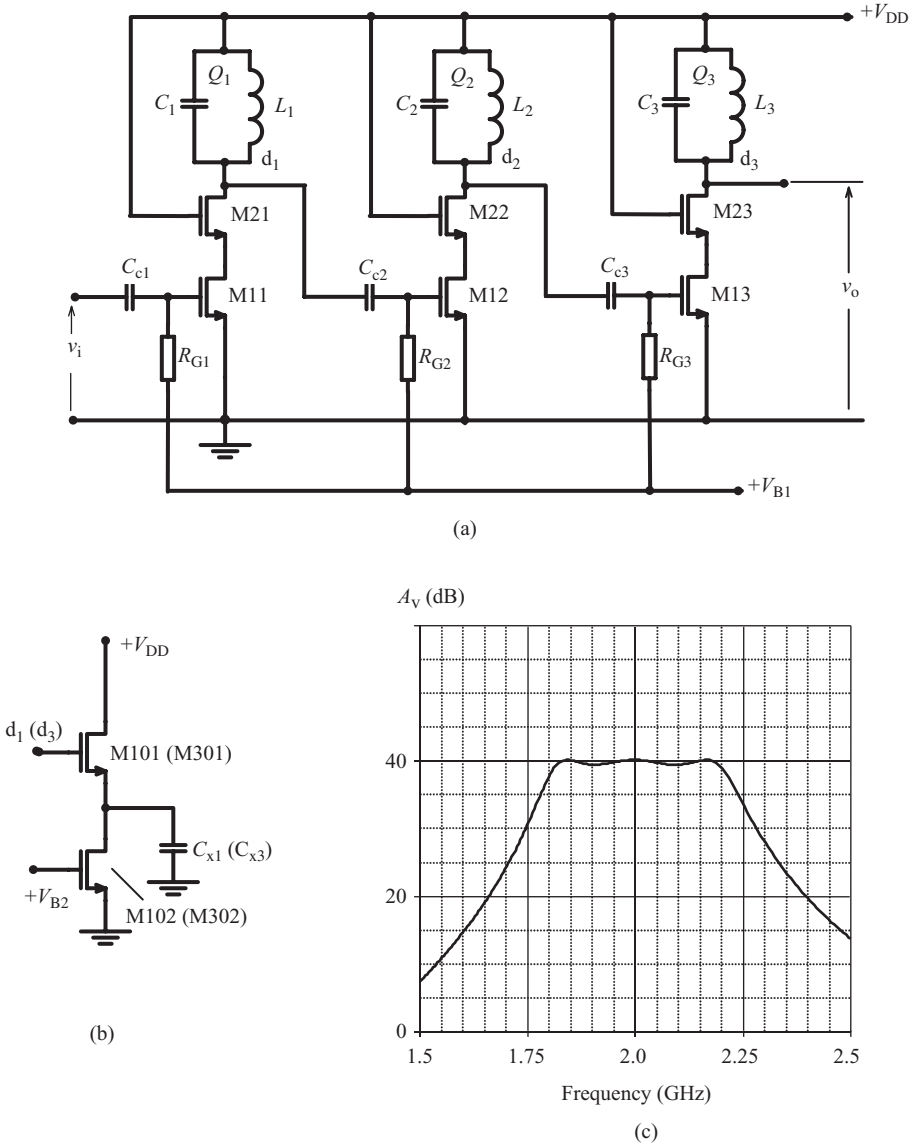


Figure 4.21 (a) Schematic diagram of the amplifier. (b) The Q-enhancement circuits. (c) The frequency characteristic of the amplifier after a fine-tuning with PSpice. ($V_{DD} = 3.3$ V, $V_{B1} = V_{B2} = 1.15$ V, $I_{tot} = 9.23$ mA.)

$$g_m = \sqrt{2\mu_n C_{ox}(W/L)I_D}$$

To find the gain of the first stage at f_0, first $|Z_1(f_0)|$ must be calculated. From (4.24a)

$$|Z_1(f_0)| = |Z_1(f_1)| \frac{1}{\sqrt{1 + (\beta_1 Q_1)^2}}$$

where

$$\beta_1 = 2\Delta f_1/f_1 \text{ and } \Delta f_1 = f_0 - f_1 = 2000 - 1835.46 = 164.54 \text{ MHz}$$

and

$$|Z_1(f_1)| = R_{1(\text{eff})} = 2121 \text{ ohm.}$$

Hence

$$|Z_1(f_0)| = 1121 \times \frac{1}{\sqrt{1 + \left(\frac{2 \times 164.54}{1835.76} \times 18.3\right)^2}} = 615.3 \text{ ohm}$$

and the gain of the first stage

$$|A_{v1}(f_0)| = g_m \times 615.3$$

Similarly the voltage gain of the third stage at f_0

$$|A_{v3}(f_0)| = g_m \times 853$$

Now the total voltage gain can be written as

$$|A_{vT}(f_0)| = (g_m)^3 \times 615.3 \times 1257 \times 853 = (g_m)^3 \times (0.659 \times 10^9)$$

which must be equal to 100. Then the transconductance of the stages can be calculated as

$$g_m = 5.34 \text{ mS}$$

According to (1.33) the transconductance of a non-velocity saturated transistor is

$$g_m = \sqrt{2\mu_n C_{ox}(W/L)I_D}$$

which implies the use of the minimum possible channel length, 0.35 μm in our case. Equation (1.33) also implies that there is a trade-off between the channel width and the drain current. Keeping in mind that a high channel width increases the parasitic capacitances, we will choose $W = 40$ μm. With the parameter values of the AMS 0.35 micron technology, (1.15) gives the drain current as 0.84 mA. But it must be kept in mind that owing to certain secondary effects such as the series parasitic resistance of the source and the mobility degradation owing to the transversal field in the channel region, to obtain the targeted transconductance value it is mostly necessary to increase the drain current.

The PSpice simulation result of the circuit after a fine-tuning is shown in Fig. 4.21(c). To adjust the total voltage gain to 40 dB, the drain currents of the stages were increased. In addition, to use the same bias voltages for the gain stages and both of the Q-enhancement circuits, the channel widths of the gain stages were changed to appropriate values and the gate bias resistor of the second stage is used to adjust the Q of the first stage. The parameters corresponding to the frequency characteristic shown in Fig. 4.21(c) are as follows.

M11, M21, M12, M22, M13, M23: $W = 44$ μm, $L = 0.35$ μm

M101, M301: $W = 46$ μm, $L = 2.4$ μm

M102, M302: $W = 20$ μm, $L = 0.35$ μm
$L_1 = L_2 = L_3 = 10$ nH, $Q = 10$
$C_1 = 445$ fF $C_2 = 520$ fF, $C_3 = 235$ fF
$R_{G1} = R_{G3} = 40$ k, $R_{G2} = 14$ k
$C_{C1} = C_{C2} = C_{C3} = 1$ pF
$C_{x1} = 250$ fF $C_{x3} = 410$ fF.

4.4 Amplifiers loaded with coupled resonance circuits

Amplifiers loaded with a pair of "coupled" resonance circuits, individually tuned to the center frequency of the band, were extensively used to obtain a reasonably flat frequency characteristic in a limited band, since the early days of radio. Although the resonance circuits are tuned to the same frequency, the transfer function of the circuit exhibits two pairs of conjugate poles having different imaginary parts, as in Fig. 4.18(a).

In earlier realizations, the main application area of these "double-tuned amplifiers" was to intermediate frequency (IF) amplifiers of all types of super-heterodyne receivers. The standard IF frequency was around 450 kHz with a bandwidth of 9 kHz for AM receivers and 10.7 MHz with a bandwidth of 150 kHz for FM receivers. The advantages of this approach were a reasonably flat response within this relatively narrow band, with only one amplifying stage, and the ease of the tuning procedure. Magnetic coupling is usually preferred to couple the resonance circuits, but it is possible to show that any approach that provides interaction among the resonance circuits gives similar results and it is possible to use the same basic equations, after an appropriate parameter conversion.

With the recent availability of computer-based tools, on-chip inductors can now be designed with reasonable precision together with their parasitics, and consequently, the capacitively or inductively coupled resonance circuits can be considered as components of integrated circuits.

4.4.1 Magnetic coupling

The basic form of a magnetically coupled circuit is shown in Fig. 4.22(a). The resonance frequencies of the two resonance circuits are the same and will be shown as ω_0,

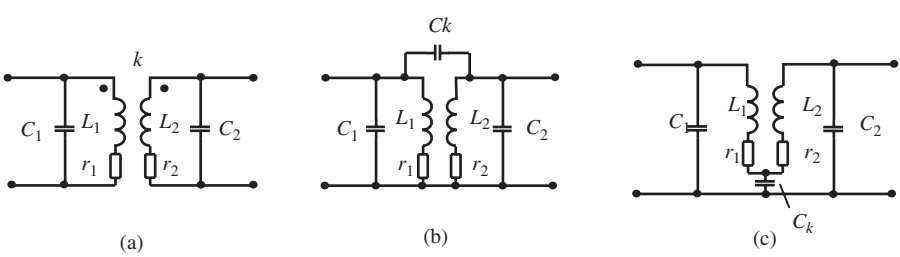


Figure 4.22 The basic form of (a) a magnetic-coupled, (b) a capacitive voltage-coupled and (c) a capacitive current-coupled double-tuned circuit.

but L_1 and L_2 are not necessarily equal. The coupling coefficient is shown as k and equal to $M/\sqrt{L_1L_2}$, by definition. The current source driving the first resonance circuit can be considered as the voltage-dependent current source of any type of high-output resistance amplifying stage. The losses of the input side and the output side were lumped as r_1 and r_2.

The transfer impedance of the circuit given in Fig. 4.22(a) can be written as

$$\frac{V_2}{I_1} = \frac{sM}{(1 + sr_1C_1 + s^2L_1C_1)(1 + sr_2C_2 + s^2L_2C_2) - s^4M^2C_1C_2} \tag{4.52}$$

with

$$\omega_0^2 = \frac{1}{L_1C_1} = \frac{1}{L_2C_2} \quad Q_1 = \frac{1}{\omega_0C_1r_1} \quad Q_2 = \frac{1}{\omega_0C_2r_2}$$

Equation (4.52) can be arranged as

$$\frac{V_2}{I_1} = \frac{sM\omega_0^4Q_1Q_2}{(s^2Q_1 + s\omega_0 + \omega_0^2Q_1)(s^2Q_2 + s\omega_0 + \omega_0^2Q_2) - s^4k^2Q_1Q_2} \tag{4.53}$$

To ease the calculation of the poles the fourth-order term of the denominator can be arranged as

$$s^4Q_1Q_2(1 - k^2) = s^4Q_1Q_2(1 - k)(1 + k)$$

which permits us to re-write the denominator as

$$\left[(s^2Q_1(1 - k) + s\omega_0 + \omega_0^2Q_1)\right]\left[(s^2Q_2(1 + k) + s\omega_0 + \omega_0^2Q_2)\right]$$

which yields two pairs of conjugate poles as

$$s_1, s_1' = -\frac{\omega_0}{2Q_1(1 - k)} \mp \mathrm{j}\frac{\omega_0}{Q_1(1 - k)}\sqrt{4Q_1^2(1 - k) - 1}$$
$$s_2, s_2' = -\frac{\omega_0}{2Q_2(1 + k)} \mp \mathrm{j}\frac{\omega_0}{Q_2(1 + k)}\sqrt{4Q_2^2(1 + k) - 1}$$

Since $4Q_1^2(1 - k) \gg 1$ and $4Q_2^2(1 + k) \gg 1$ for small values of k, the poles can be written as

$$s_1, s_1' \cong -\frac{\omega_0}{2Q_1(1 - k)} \mp \mathrm{j}\frac{\omega_0}{\sqrt{1 - k}}$$
$$s_2, s_2' \cong -\frac{\omega_0}{2Q_2(1 + k)} \mp \mathrm{j}\frac{\omega_0}{\sqrt{1 + k}}$$

Since usually $k \ll 1$, these expressions can be even further simplified as

$$s_1, s_1' \cong -\frac{\omega_0}{2Q_1(1 - k)}\omega_0\left(1 + \frac{k}{2}\right) \tag{4.54}$$

$$s_2, s_2' \cong -\frac{\omega_0}{2Q_2(1 + k)}\omega_0\left(1 - \frac{k}{2}\right) \tag{4.55}$$

with

$$(1-k) \cong 1,\ (1+k) \cong 1 \quad 1/\sqrt{1-k} \cong 1+(k/2) \text{ and } 1/\sqrt{1+k} \cong 1-(k/2)$$

From (4.54), (4.55) and Fig. 4.18(a) it can be concluded that;

- to obtain pole pairs appropriate for Butterworth or Chebyshev responses the real parts of the poles must be equal:

$$\sigma = -\frac{\omega_0}{2Q_1(1-k)} = -\frac{\omega_0}{2Q_2(1+k)} \tag{4.56a}$$

and for $k \ll 1$,

$$Q_1 \cong Q_2 = Q \tag{4.56b}$$

- to obtain a second-order Butterworth-type frequency characteristic the positive imaginary part poles must be positioned as shown in Fig. 4.23(a). The corresponding k values and the bandwidth can be calculated from the geometry of the figure:

$$|\sigma| = \frac{\omega_0}{2Q} = \omega_0 \frac{k}{2} \Rightarrow \quad k = \frac{1}{Q} \tag{4.57}$$

$$2\Delta\omega = 2 \times \left(\sqrt{2}|\sigma|\right) = \sqrt{2}\frac{\omega_0}{Q} \Rightarrow B = 2\Delta f = \sqrt{2}\frac{f_0}{Q} \tag{4.58}$$

From the comparison of (4.58) and (4.25) we see that the bandwidth of a double-tuned circuit with maximally flat frequency response is $\sqrt{2}$ times larger than that of a single-tuned circuit, having the same Q value.

The positioning of poles for a Chebyshev-type frequency characteristic is shown in Fig. 4.23(b). The shift of the poles in the horizontal direction with respect to the poles of a Butterworth-type response having the same bandwidth, corresponding to a certain

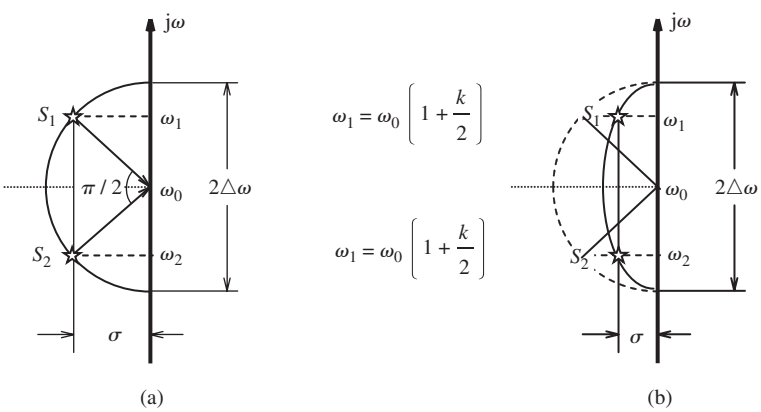


Figure 4.23 Positions of the poles (a) for a maximally flat (Butterworth-type), (b) for an equi-ripple (Chebyshev-type) frequency characteristic.

ripple value, was given in Section 4.3. For example, to obtain a Chebyshev-type frequency characteristic with 0.5 dB ripple, from the geometry of the figure,

$$(\sigma)_C = \frac{\omega_0}{2Q} = 0.709 \times \omega_0 \frac{k}{2} \Rightarrow k = \frac{1}{0.709}\frac{1}{Q} = \frac{\sqrt{2}}{Q} \tag{4.59}$$

$$2\Delta\omega = 2 \times \sqrt{2} \times \omega_0 \frac{k}{2} = \sqrt{2} \times \omega_0 \times \frac{1}{0.709}\frac{1}{Q} \Rightarrow B = 2\Delta f = 2 \times \frac{f_0}{Q} \tag{4.60}$$

From the comparison of (4.60) and (4.25) we see that the bandwidth of a double-tuned circuit with 0.5 dB ripple Chebyshev-type frequency response is twice as large as that of a single-tuned circuit, having the same Q value.

Example 4.6 A double-tuned amplifier has a load as shown in Fig. 4.24(a). The center frequency of the response is 2 GHz. Effective quality factors of both sides were adjusted to 20 with an appropriate Q-enhancement circuit, which corresponds to a parallel effective resistance of $R_{p1} = R_{p2} = 2512$ ohm.

(a) Calculate the magnetic coupling coefficient for a maximally flat response and corresponding bandwidth.

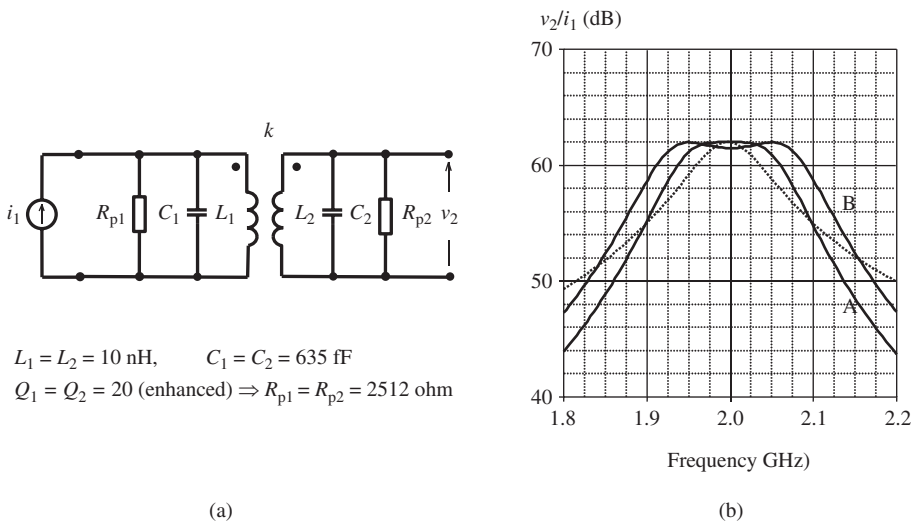


Figure 4.24 (a) The equivalent circuit of the output side of the amplifier. (b) Frequency characteristics of the double-tuned circuit for (A) $k = 1/Q$ (maximally flat response) and (B) $k = 1.41/Q$ (0.5 dB ripple Chebyshev response). The response of a single-tuned circuit (C) is also given for comparison.[11]

[11] It can be shown that the magnitude of the transfer impedance corresponding to $kQ = 1$ at ω_0 is half of the resonance impedance of the single-tuned circuit. To ease the comparison, the response of the single-tuned circuit is shifted by -6 dB.

The value of the coupling coefficient can be calculated from (4.57):

$$k = \frac{1}{Q} = \frac{1}{20} = 0.05$$

According to (4.58) the bandwidth is

$$B = \sqrt{2}\frac{f_0}{Q} = \sqrt{2}\frac{2\times 10^9}{20} = 141 \text{ MHz}$$

(b) Calculate the magnetic coupling coefficient for a 0.5 dB ripple Chebyshev-type response and corresponding bandwidth.
From (4.59)

$$k = \frac{\sqrt{2}}{Q} = \frac{\sqrt{2}}{20} = 0.0707$$

From (4.60)

$$B = 2\times\frac{f_0}{Q} = 2\times\frac{2\times 10^9}{20} = 200 \text{ MHz}$$

(c) Plot the frequency responses with PSpice simulation.
Results are given in Fig. 4.24(b) that fit the calculated values.

Problem 4.6 *A cascode amplifier stage with a double-tuned circuit load will be designed. The target values are:*

- *center frequency: $f_0 = 1$ GHz,*
- *voltage gain at 1 GHz: 20 dB,*
- *bandwidth: $2\Delta f = 120$ MHz,*
- *type of the response: 0.4 dB ripple, Chebyshev-type,*
- *technology: AMS 0.35,*
- *available on-chip inductors: 5 nH ($Q = 6$ at 1 GHz), 10 nH ($Q = 5.7$ at 1 GHz), 15 nH ($Q = 4.7$ at 1 GHz),*
- *quality factors of the on-chip capacitors: $Q = 50$,*
- *power supply: +3.2 V.*

(a) Determine the positions of poles of the gain function and then calculate the effective quality factors corresponding to these poles.
(b) Calculate the value of the negative resistances necessary to enhance the actual Q values to the calculated Q values (assume the output resistance of the cascode circuit is negligibly high).
(c) Calculate the magnetic coupling coefficient, k.
(d) Calculate the resonance capacitances, the parasitics included.

(e) Calculate the transconductance of the cascode stage.
(f) Simulate the circuit and plot the frequency characteristic (the idealized equivalent circuit of the cascode amplifier can be used).
(g) Check the effects of the ± 10 % and ± 20 % spread of the k value.

4.4.2 Capacitive coupling

To establish an interaction among the two resonance circuits there are other possibilities, as shown in Fig. 4.22(b) and (c). The circuit in Fig. 4.22(b) is called the capacitive voltage-coupled double-tuned circuit. The expressions related to this case have similar forms to the expressions derived for magnetic coupling. The coupling capacitance value corresponding to a certain value of the magnetic coupling coefficient, k, is given as [29]

$$C_k = -\frac{k}{\omega_0^2\sqrt{L_1L_2}} \tag{4.61}$$

The negative sign indicates that the capacitive voltage coupling is equivalent to a negative magnetic coupling.

For the capacitive current coupling shown in Fig. 4.22(c), C_k can be calculated as

$$C_k = -\frac{1}{k\omega_0^2\sqrt{L_1L_2}} \tag{4.62}$$

It must be noted that

- expressions given for the capacitive coupling were obtained for high-Q circuits, therefore a fine-tuning with SPICE simulation is needed for low-Q circuits;
- the center frequency shifts downward for the capacitive voltage coupling and upward for the capacitive current coupling, owing to the effect of the coupling capacitor.

Example 4.7 Calculate the value of the coupling capacitance for the circuit shown in Fig. 4.24(a) to obtain a 0.5 dB ripple Chebyshev-type response. Assume that there is no magnetic coupling.

From (4.61)

$$C_k = \frac{0.0707}{(2\pi\times 2\times 10^9)^2\sqrt{(10^{-8})(10^{-8})}} = 45 \text{ fF}$$

4.5 The gyrator: a valuable tool to realize high-value on-chip inductances

A tuned amplifier can be interpreted as a "band-pass filter" that amplifies (passes) the frequencies in a certain band and does not amplify (stops) all other frequencies.

Several types of filters, especially band-pass filters and low-pass filters, are being used extensively in telecommunication systems. Filters were one of the main subjects of network theory during the 1940s and 1950s and have been investigated in depth. Several methods have been developed to design a filter that fulfills the requirements related to the pass-band, stop-band and impedances. These filters were "passive" filters that were composed of passive components: capacitors, inductors and resistors.

Analog and digital "active" filters have been the result of the developments of IC technology. For one of the types of analog active filter the methodology of the well-investigated passive filter theory can be directly applied; the inductive elements in these filters are not "real" inductors, but are "emulated" with an electronic circuit called a "gyrator" and a capacitor. With this approach, relatively high-value inductances (not possible to realize as an on-chip component) can be integrated, and filters can be realized at relatively high frequencies that could not be reached with op-amp based active filters.

As will be shown below, a gyrator can be easily realized with a number of transconductance amplifiers (OTAs). That is why the gyrator-based filters are also called "G_m-C filters", as specific members of the broader "G_m-C filters" family.

The "gyrator" was described (or invented) by B. D. H. Tellegen with a purely theoretical approach in 1948 [30]. After a systematic classification of linear two-ports, Tellegen noticed that there must exist a yet-unknown two-port satisfying the expression

$$\begin{pmatrix} i_1 \\ i_2 \end{pmatrix} = \begin{pmatrix} 0 & G \\ -G & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \tag{4.63}$$

He named this hypothetical two-port the "gyrator" and offered a symbol shown in Fig. 4.25(a).

It can be shown that a gyrator can be realized with two ideal voltage-controlled current sources, as shown in Fig. 4.25(b), or with two ideal OTAs, as shown in Fig. 4.25(c), and in addition, the G parameters of the current sources do not have to be equal [31]. Consequently, the describing expression of the circuit shown in Fig. 4.25(c) becomes

$$\begin{pmatrix} i_1 \\ i_2 \end{pmatrix} = \begin{pmatrix} 0 & g_{m2} \\ -g_{m1} & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \tag{4.64}$$

The input impedance of a gyrator loaded with a load admittance Y_2 can be easily calculated:

$$i_2 = -Y_2 v_2 = -g_{m1} v_1, \;\; i_1 = g_{m2} v_2 \quad \Rightarrow \quad Z_1 = \frac{v_1}{i_1} = \frac{Y_2}{g_{m1} g_{m2}} \tag{4.65}$$

The practical importance of the gyrator can be best appreciated by connecting a capacitor to the output port as shown in Fig. 4.26, and then calculating the input impedance of the circuit. In this case, since $Y_2 = sC$, (4.65) can be arranged as

$$Z_1 = s \frac{C}{g_{m1} g_{m2}} \;\; \Rightarrow \;\; L = \frac{C}{g_{m1} g_{m2}} \tag{4.66}$$

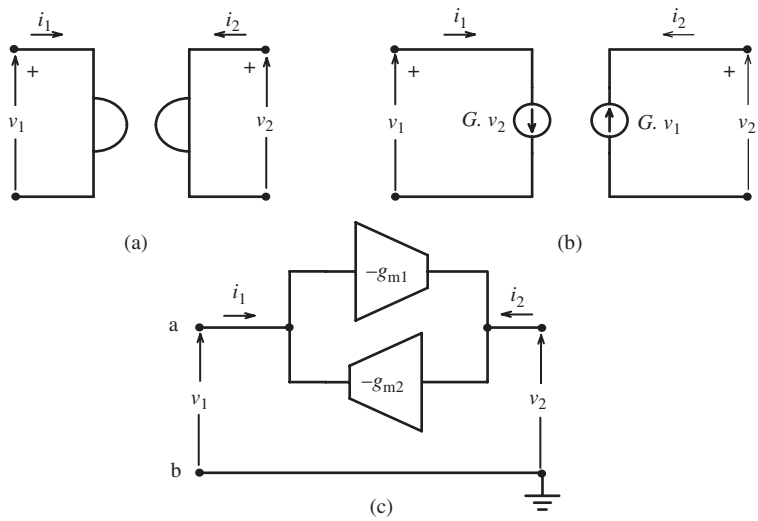


Figure 4.25 (a) The original symbol of the gyrator. (b) Realization of a gyrator with two voltage-controlled current sources and (c) with two OTAs.

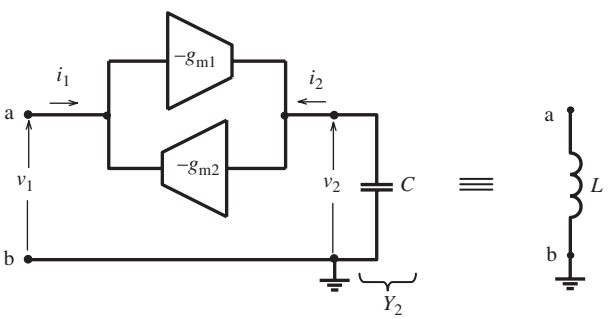


Figure 4.26 Gyrator-based inductance realization.

Expression (4.66) shows us that the input impedance of a capacitance-loaded ideal gyrator is equal to the impedance of an ideal inductor. Therefore, the input port of the circuit can be used to replace an inductor. This means that the high inductance values that are not possible to realize with on-chip inductors can be realized with a gyrator and an appropriate load capacitor on the chip.

Example 4.8 A 10 μH inductance is needed. The transconductance values of the OTAs that will be used to implement the gyrator are 0.5 mS each. Calculate the value of the capacitor to be connected to the output port of the gyrator.

From (4.66)

$$C = Lg_{m1}g_{m2} = (10\times10^{-6})(0.5\times10^{-3})(0.5\times10^{-3}) = 2.5\ \text{pF}$$

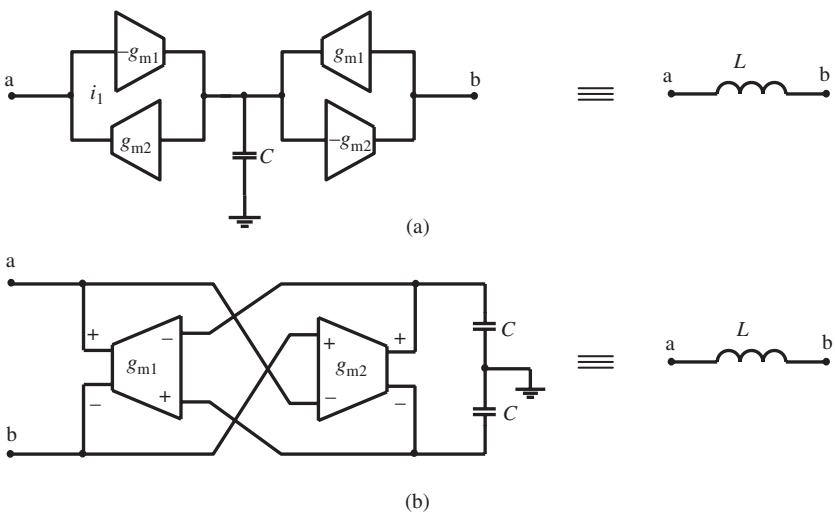


Figure 4.27 Realization of floating inductances: (a) with single-ended input, single-ended output OTAs, (b) with symmetrical OTAs.

It must be noted that one terminal of this "emulated" inductor is connected to ground. In many applications, however, floating inductors are needed. A floating inductor can be realized between two circuit nodes a and b, as shown in Fig. 4.27(a) and (b) [32].

4.5.1 Parasitics of a non-ideal gyrator

Up until this point we assumed that the OTAs used to compose a gyrator were ideal. In reality they have a finite input capacitance (C_i), a finite output capacitance (C_o) and a non-zero output conductance. In addition, the transadmittance function does not have a flat frequency characteristic but rolls-off at high frequencies, which can be expressed as a one-pole gain function as given in (3.76). Consequently, all these non-idealities must be taken into account to characterize the realized inductance.

In Fig. 4.28 a "real" gyrator is represented with an "ideal" gyrator and the input and output parasitics of the circuit. For simplicity, we will assume that OTAs are realized with the circuit given in Fig. 3.32. The output capacitance C_o is the sum of the input capacitance of OTA2 (the gate capacitance of the input transistor) and the sum of the drain capacitances of the transistors connected to the output node of OTA1. The output conductance g_o is the sum of the output conductances of the output transistors of OTA1. Therefore the total load admittance of the gyrator, Y_2 is the sum of these components and the external load capacitance, C_L.

The input parasitics, C_i and g_i, are similar to their output counterparts. Since they are connected in parallel to the input impedance Z_1 to be calculated, they will first be excluded to simplify the calculations, and will be added later.

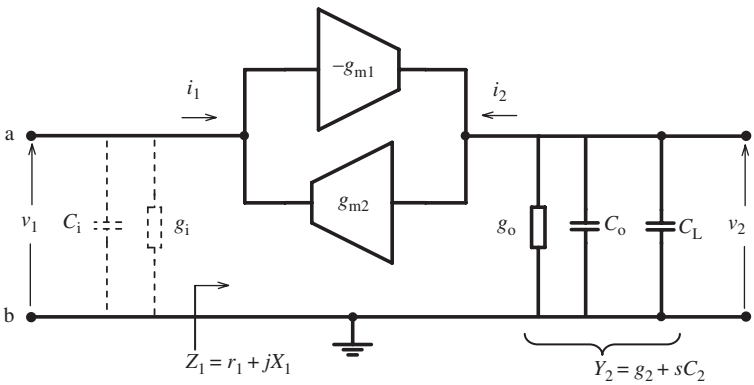


Figure 4.28 Gyrator with its input and output parasitics. Note that $g_2 = g_o$ and $C_2 = (C_o + C_L)$.

The gain of a typical OTA can be expressed using a one-pole gain function as given in (3.66):

$$y_m = g_m s_p \frac{1}{(s - s_p)}$$

Now with $Y_2 = g_2 + sC_2$ (4.65) can be written as

$$\begin{aligned} Z_1 &= \frac{Y_2}{y_{m1} y_{m2}} \\ &= \frac{1}{g_{m1} g_{m2} s_{p1} s_{p2}} (s - s_{p1})(s - s_{p2})(g_2 + sC_2) \qquad (4.67a) \\ Z_1(\omega) &= \frac{1}{g_{m1} g_{m2} \omega_{p1} \omega_{p2}} (j\omega + \omega_{p1})(j\omega + \omega_{p2})(g_2 + j\omega C_2) \end{aligned}$$

It is convenient to simplify this expression assuming that the poles of the two OTAs are identical:

$$Z_1(\omega) = \frac{1}{g_{m1} g_{m2} \omega_p^2} (j\omega + \omega_p)^2 (g_2 + j\omega C_2) \qquad (4.67b)$$

The imaginary part of the input impedance can be calculated as

$$\mathrm{Im}\{Z_1\} = \omega \left(\frac{C_2}{g_{m1} g_{m2}} + 2 \frac{g_2}{\omega_p g_{m1} g_{m2}} \right) \qquad (4.68a)$$

and for $\omega \gg \omega_p$,

$$\mathrm{Im}\{Z_1\} \cong \omega \left(\frac{C_2}{g_{m1} g_{m2}} \right) = \omega L \qquad (4.68b)$$

The real part of the input impedance, which corresponds to the series resistance of L, is

$$\mathrm{Re}\{Z_1\} = \frac{g_2}{g_{m1} g_{m2}} - \frac{\omega^2}{\omega_p^2} \frac{1}{g_{m1} g_{m2}} (g_2 + 2\omega_p C_2) = r_L$$

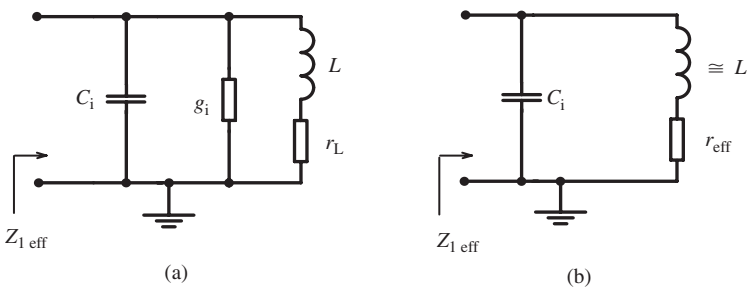


Figure 4.29 (a) The equivalent circuit of the input impedance of a gyrator, loaded with a capacitor, the input parasitics included. (b) The simplified equivalent circuit.

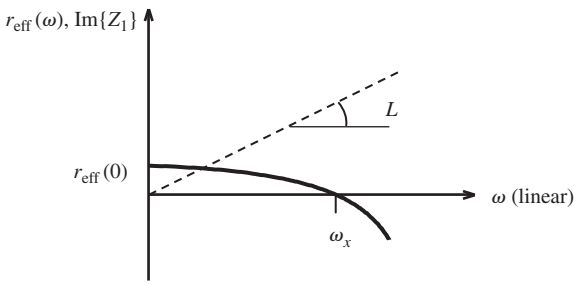


Figure 4.30 Variations of r_{eff} (solid line) and the imaginary part of Z_1 (dashed line) with frequency, for $g_i < (2/\omega_p L)$. Note that the frequency axis was chosen as linear to show the linear dependence of the imaginary part. The slope of the dashed line is equal to L.

since usually $2\omega_p C_2 \gg g_2$, r_L can be simplified as

$$r_L \cong \frac{g_2}{g_{m1}g_{m2}} - \frac{\omega^2}{\omega_p}\frac{2C_2}{g_{m1}g_{m2}} \tag{4.69}$$

which represents the losses of the inductor that appears at the input port of the gyrator. The total (effective) input impedance of the gyrator, including the input parasitics, is shown in Fig. 4.29(a). It is apparent that g_i adds on to the total losses of the inductor and decreases the quality factor and C_i leads to resonance at an angular frequency equal to $\omega_0 = \sqrt{LC_i}$.

The effective resistance of the inductor together with the effect of g_i can be calculated as

$$\begin{aligned} r_{eff} &\cong r_L + \omega^2 L^2 g_i \\ &= \frac{g_2}{g_{m1}g_{m2}} - \omega^2 L^2\left(\frac{2}{\omega_p L} - g_i\right) \end{aligned} \tag{4.70}$$

In Fig. 4.30 the imaginary part of Z_1 from (4.68b) that is proportional to L and the effective series resistance of L from (4.70) for $g_i < (2/\omega_p.L)$ are plotted as a function of ω. From these plots it can be seen that

- the slope of $\mathrm{Im}\{Z_1\}$ is constant and equal to L;
- for $g_i < (2/\omega_p L)$, $r_{eff}(\omega)$ decreases with frequency and crosses zero at ω_x, which can be calculated as

$$\omega_x = \sqrt{\frac{g_2}{g_{m1}g_{m2}} \frac{1}{L^2\left(\frac{2}{\omega_p L} - g_i\right)}} \tag{4.71}$$

  and then becomes negative. If L resonates with a capacity at a frequency higher than ω_x, the circuit oscillates

- for $g_i > (2/\omega_p L)$, $r_{eff}(\omega)$ is always positive; therefore there is no risk of instability
- the quality factor of the inductor for any ω is

$$Q_{eff}(\omega) = \frac{L\omega}{r_{eff}(\omega)} = \frac{L\omega}{\frac{g_2}{g_{m1}g_{m2}} - \omega^2 L^2\left(\frac{2}{\omega_p L} - g_i\right)} \tag{4.72}$$

From (4.72) it can be concluded that

- the effective quality factor of the emulated inductor increases with the transconductances of OTAs and decreases with the input and output conductances;
- for $g_i < (2/\omega_p L)$ the circuit has an inherent Q-enhancement feature. For frequencies approaching ω_x the effective quality factor sharply increases and reaches infinity at $\omega = \omega_x$, where the effective resistance of the inductor becomes zero;
- for this case, owing to the increased sensitivity of Q_{eff}, operating with this excessive Q-enhancement (at frequencies close to ω_x) is not convenient.

Example 4.9 A gyrator as shown in Fig. 4.28 is formed using two OTAs similar to the circuit given in Example 3.3. OTAs are used in single-ended input, single-ended output configuration. The transconductances are $g_{m1} = g_{m2} = 1$ mS, $g_{o1} = g_{o2} = 84$ μS and $f_{p1} = f_{p2} = 2.62$ GHz. The external output load capacitance is $C_L = 1$ pF.

In Fig. 4.31 the PSpice simulation results for the real and imaginary parts of the input impedance are given. The imaginary part linearly increases with frequency according to (4.68) as expected.[12] The inductance value calculated from the initial slope of the imaginary part curve is 1.21 μH. According to (4.68b), if there were no parasitic capacitances (i.e. $C_2 = C_L$), the value of the inductance should be 1 μH. The difference arises from the output parasitic capacitance of the gyrator, which can be easily calculated as 0.21 pF.

The real part of the input impedance for $f = 0$ that corresponds to $r_{eff}(0)$ is 83.4 ohm, which is in agreement with the value calculated from (4.69).

[12] Note that for frequencies approaching the resonance frequency of the inductance with the input parasitic capacitance of the gyrator, the imaginary part of the input impedance does not increase linearly any more.

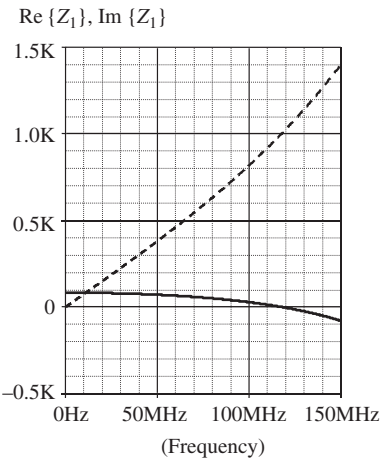


Figure 4.31 The variations of the real part (solid line) and the imaginary part (dashed line) of the input impedance of the gyrator.

Problem 4.7

(a) Design a 10 μH inductor with the OTAs used in Example 4.9.
(b) What is the frequency limit of usability for this inductor?
(c) Calculate the effective quality factor at 20 MHz and 50 MHz.
(d) Compare the results with the results of PSpice simulation.

4.5.2 Dynamic range of a gyrator-based inductor

The voltage swing between the terminals of a gyrator-emulated inductor is limited by the maximum output voltage of OTA1, which depends on (and is usually approximately equal to) the supply voltage, or by the maximum input voltage of OTA2, whichever is smaller. A similar limitation holds for the voltage of the output terminal. To ensure proper operation, the input port swing must not be limited owing to the output port dynamic range, and vice versa. This means that the maximum output swings of the input and output ports must be equal, i.e. $\hat{V}_1 = \hat{V}_2 = \hat{V}$ (see Fig. 4.32).

From Fig. 4.32 the output port voltage can be written in terms of the input port voltage as

$$v_2 = -\frac{i_2}{Y_2} = -i_2 \frac{1}{sC_2 + g_2} = v_1 g_{m1} \frac{1}{sC_2 + g_2}$$

and the magnitude in the ω domain is

$$V_2 = V_1 \frac{g_{m1}}{\sqrt{\omega^2 C_2^2 + g_2^2}} \cong V_1 \frac{g_{m1}}{\omega C_2} \tag{4.73}$$

Similarly the input port voltage can be written as

$$v_1 = i_1 Z_1 = i_1 (sL + r_L) = v_2 g_{m2} (sL + r_L)$$

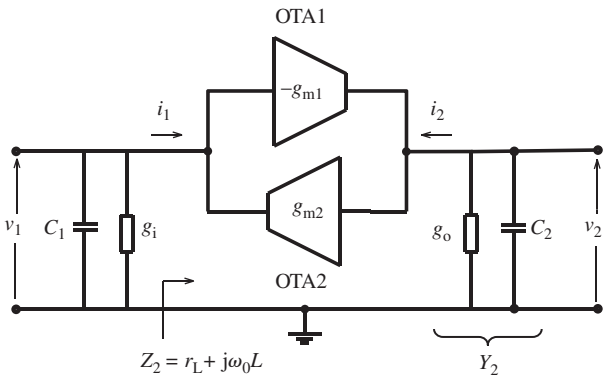


Figure 4.32 Input and output port voltages of a capacitive loaded gyrator.

and the magnitude in the ω domain:

$$V_1 = V_2 g_{m2}\sqrt{\omega^2 L^2 + r_L^2} \cong V_2 g_{m2}\omega L \tag{4.74}$$

The inductance is typically used in resonance with a parallel capacitor, C_1. Equations (4.73) and (4.74) can be written in terms of the resonance frequency, which is equal to $1/\sqrt{LC_1}$,

$$V_2 \cong V_1 \frac{g_{m1}}{\omega_0 C_2} \tag{4.75}$$

and

$$V_1 \cong V_2 \omega_0 L = V_2 \frac{g_{m2}}{\omega_0 C_1} \tag{4.76}$$

From the condition of equality of the magnitudes of the input and output voltages as stated above and from (4.75) and (4.76),

$$\frac{g_{m1}}{g_{m2}} = \frac{C_2}{C_1} \tag{4.77}$$

To verify this expression, a PSpice simulation is performed. The gyrator given in Example 4.9 is driven with a 10 μA AC current source from the input port. The input and output port voltages are plotted for different C_2/C_1 ratios (see Fig. 4.33). It is seen clearly that the input and output port voltages at resonance are equal for $C_2/C_1 = 1$ as expected according to (4.77), since $g_{m1} = g_{m2}$ in the example.

4.6 The low-noise amplifier (LNA)

LNA is the acronym used for a class of amplifiers, the "low-noise amplifiers", commonly employed as the input stages of wireless receivers. Since the incoming signal from an antenna is usually weak, the unavoidable noise generated in the amplifier must

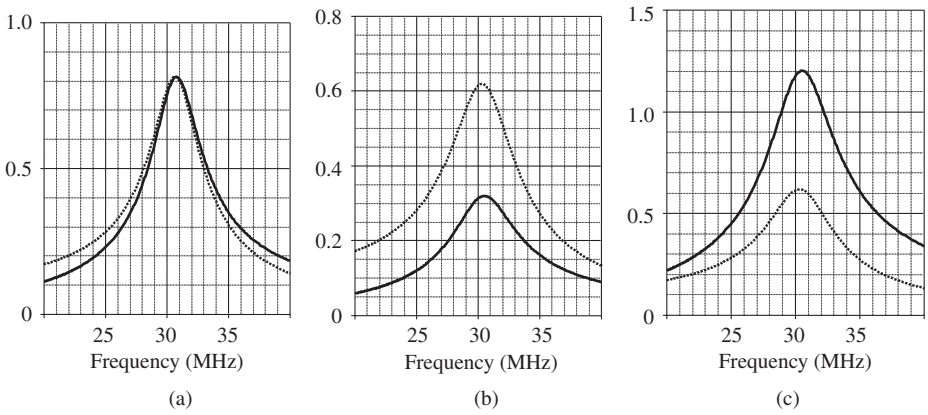


Figure 4.33 The input (solid line) and output port (dashed line) voltages (in volts) of the gyrator for different C_2/C_1 ratios: (a) $C_1 = C_2 = 5$ pF, (b) $C_1 = 10$ pF, $C_2 = 2.5$ pF, (c) $C_1 = 2.5$ pF, $C_2 = 10$ pF.

be as low as possible, to obtain an acceptable "signal-to-noise ratio" at the output of the amplifier. Therefore, one of the key design goals for the LNA is a low noise contribution to the input signal, together with a good impedance matching to the signal source, a sufficiently large output signal dynamic range and – certainly – a low power consumption.

Since LNAs are being used as the input stage of receivers, they must be tuned (or be tunable) to the carrier frequency of the transmitter that we intend to receive. Therefore, LNAs are inherently considered to be "tuned amplifiers". The bandwidth of the amplifier must be large enough to cover the side-bands of the modulated carrier. But owing to the low Q values of the on-chip inductors, the bandwidth usually becomes larger than necessary and the signals that remain outside of the modulation bandwidth must be eliminated by the succeeding stages of the amplifier.

4.6.1 Input impedance matching

The input signal source of the LNA is usually an antenna with the appropriate shape and dimension to provide the desired radiation pattern (omni-directional or directional with a specified beam-pattern) at the frequency of interest. The internal impedance of the antenna at this frequency (which is normally the resonance frequency of the antenna) has a dominant resistive component corresponding to the radiation losses[13] and a usually small reactive component corresponding to the connection parasitics.

[13] From antenna theory it is known that an antenna is reciprocal in one sense; it exhibits similar behavior to a transmitting or receiving antenna. Since the concept of "radiation loss" is more understandable, this term has been used even for receiving antennas. It must be noted that the radiation loss is not constant; it is maximum when the radiation is maximum, which corresponds to the maximum of the current or voltage, in other words, to resonance. On both sides of the resonance peak, the radiation and, correspondingly, the antenna loss decrease, leading to further narrowing of the resonance curve.

If the antenna is located at a distance from the input terminal of the amplifier, a low-loss transmission line has to be used to connect the antenna to the amplifier. In this case, to maximize the signal power transfer from the antenna to the amplifier, we must satisfy:

(a) impedance matching between the antenna and the line,
(b) impedance matching between the line and the input of the amplifier.

Coaxial lines with 50 ohm characteristic impedance[14] are the most frequently used transmission lines[15] for single-ended antennas such as monopole antennas. For symmetrical antennas such as dipole antennas, folded-dipole antennas, etc., symmetrical transmission lines can be used to connect to the inputs of a differential LNA. Another solution is to use a "balun"[16] to connect the output of a symmetrical antenna to the input of a coaxial line. If the antenna impedance is not resistive and not equal to the characteristic impedance of the line at the operation frequency of the amplifier, an impedance-matching circuit has to be used to provide a resistive impedance equal to Z_o (or R_o) to maximize the power transfer from the antenna to the transmission line. If the resistive component of the antenna impedance (R_A) is equal to R_o, reactive components can be easily compensated by a series matching inductor or matching capacitor, as shown in Fig. 4.34(a) and (b).

For $R_A < R_o$ and $R_A > R_o$, there are several simple solutions. For one of the solutions for $R_A < R_o$, the circuit diagram, the corresponding phasor diagram and the design formulas are given in Fig. 4.35. Note that this solution is especially advantageous for antennas having an inductive component that can be considered as a part of L_M.

Problem 4.8 *Derive matching circuits similar to that given in Fig. 4.35, which are suitable for an antenna*

(a) having a series capacitive component and $R_A < R_o$,
(b) having a parallel capacitive component and $R_A > R_o$,
(c) having a parallel inductive component and $R_A > R_o$.

The use of impedance-converting transformers is another possibility to match the real part of the antenna impedance to the line.

[14] The characteristic impedance (Z_o) of a transmission line is the impedance for which no signal reflection occurs when the line is terminated with this impedance. The characteristic impedance of a low-loss transmission line is equal to $\sqrt{L/C}$, where L and C correspond to the series inductance and parallel capacitance of the line per unit length, and is resistive. To underline its resistive character, R_o will be used for the characteristic impedance, whenever necessary.

[15] Other standard characteristic impedance values for coaxial lines are 60 ohm and 75 ohm.

[16] "balun" is an abbreviation for "balanced-to-unbalanced" converting circuit.

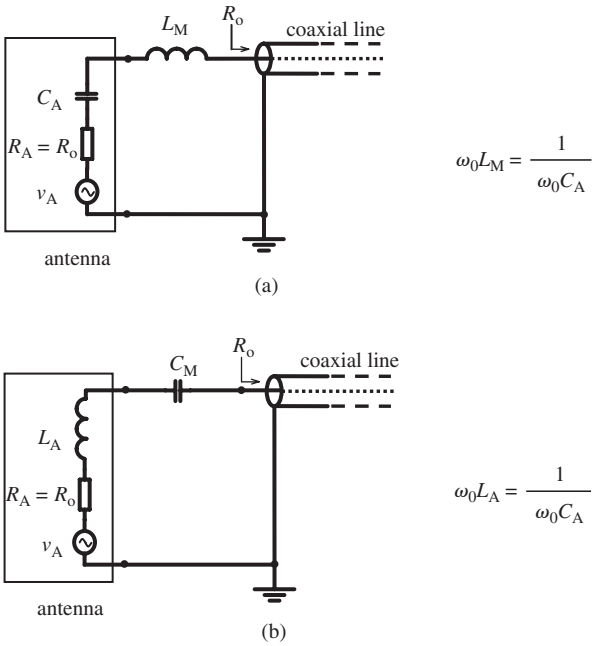


Figure 4.34 Impedance matching of a single-ended antenna when the real part of the antenna impedance is equal to the characteristic impedance of the coaxial line (a) for a capacitive antenna, (b) for an inductive antenna.

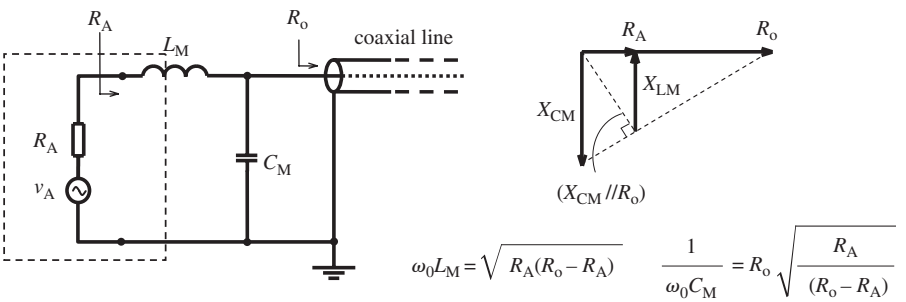


Figure 4.35 Impedance-matching circuit for $R_A < R_o$, suitable for antennas having a small series inductive component.

To provide an impedance matching between the transmission line and the input of the amplifier, the amplifier can be designed such that the real part of the input impedance is equal to R_o, which is usually 50 ohms.

The reactive component of the input impedance (which is usually capacitive) can be easily compensated with a series inductance, as shown in Fig. 4.36. This is the classical approach for the design of LNAs.

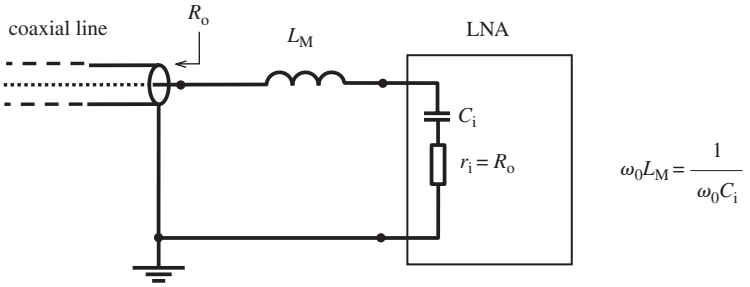


Figure 4.36 Impedance matching between a 50 ohm coaxial line and an LNA.

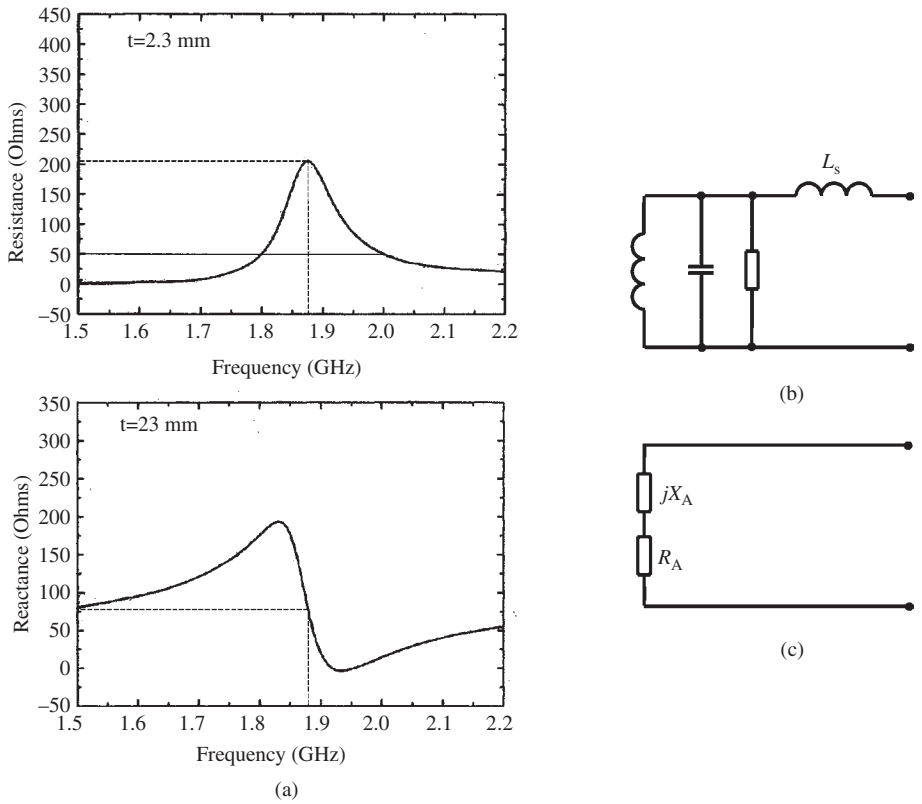


Figure 4.37 (a) The variation of the real and imaginary parts of the impedance of the inverted-F antenna given in [33]. (b) The model suitable to represent the behavior of the antenna. L_s represents the positive shift of the reactance curve. (c) The model representing the resistive and reactive parts of the antenna impedance.

In the widest application area of LNAs today, namely the wireless systems such as mobile telephones, GPS receivers, etc., the antenna is built on the same board with the LNA chip and close enough to eliminate the need for any transmission line in between. This means that the adherence to the 50 ohm standard for the input impedance of

LNAs has no meaning any more (except the ease of measurement and characterization). Therefore, LNAs must be designed to satisfy the matching with the impedance of the antenna that is intended to be used. An even more realistic approach is the co-design (or, at least interactive design) of the antenna and the amplifier. The following discussion serves to illustrate the importance of such considerations.

The most frequently used on-board antennas in mobile systems are folded-dipole and folded-loop antennas to drive the differential LNAs, and inverted-F antennas to drive the single-ended LNAs. There are many publications about these types of antenna, but they are mostly focused on the radiation pattern and the return-loss (or the SWR) when the antenna is driven with a 50 ohm line. Variation of the impedance with frequency, however, is investigated only in few publications.

One of the extensively used types of on-board single-ended antenna is the "planar inverted-F antenna". These types of antenna are suitable to be realized on the multilayer board of the receiver and provide a fairly omni-directional radiation pattern. The resonance frequency and the impedance of inverted-F antennas strongly depend on the structure and dimensions. For one of the published inverted-F antennas [33], the real and imaginary parts of the impedance are given as shown in Fig. 4.37(a).

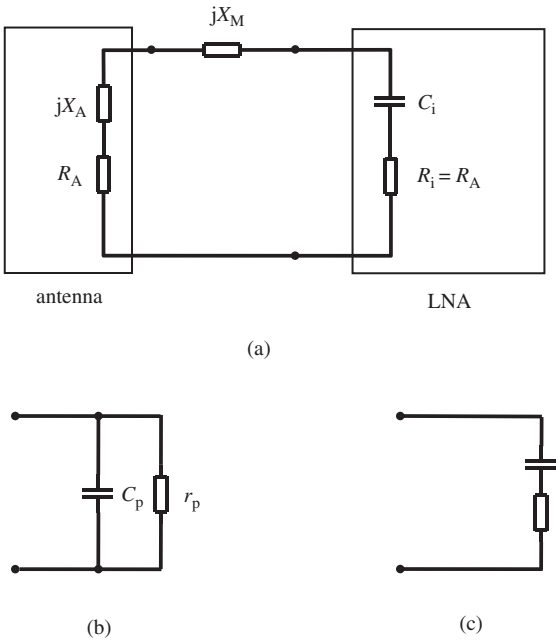
It can be seen that there is an apparent resemblance between these curves and the curves of the real and imaginary parts of the impedance of a parallel resonance circuit. The positive shift of the reactive part indicates that the model includes a series inductance, as shown in Fig. 4.37(b). It must be noted that the parallel resistance representing the radiation losses is not constant; it is minimum at the resonance frequency and increases with de-tuning on both sides of the peak. The equivalent circuit of the antenna in terms of its resistive and reactive components is given in Fig. 4.37(c).

From the curves given in Fig. 4.37(a), it can be seen that

- the resonance frequency of the antenna is 1.87 GHz, where the antenna is most effective;
- the resistive and reactive components of the antenna impedance at resonance frequency are $R_A = 210$ ohm and $X_A = +75$ ohm, respectively;
- therefore, for maximum signal transfer from the antenna to the input of the LNA, the input impedance of the LNA at 1.87 GHz must be $r_i = 210 - j75$ ohm, in other words 210 ohm in series with a 1.147 pF capacitance;
- if f_0, the center frequency of the LNA, is not equal to 1.87 GHz, the antenna must be modified to shift the resonance frequency to f_0;
- from the resistance curve it can be seen that at approximately 2 GHz the antenna impedance is purely resistive. If the antenna is loaded with a 50 ohm line, the return loss certainly becomes minimum at this frequency, as given in [33]. But since the antenna is not operating at the resonance frequency, the overall signal efficiency becomes considerably lower than the efficiency of the antenna operating at resonance.

4.6.2 Basic circuits suitable for LNAs

The real part of the impedance of an antenna at the resonance frequency is low; usually in the range of some tens of ohms to hundreds of ohms. To enable maximum signal



$$jX_A + jX_M + \frac{1}{j\omega_0 C_i} = 0$$

Figure 4.38 (a) The input loop of an LNA. (b) The usually valid input impedance equivalent of an LNA. (c) The series R-C equivalent of the input impedance that can be directly used in (a).

power transfer from the antenna to the input of the amplifier, the real part of the input impedance of the amplifier must be equal to that of the antenna. This means that the LNA must be a low input impedance amplifier. The imaginary parts can be eliminated by resonance at the frequency of operation, f_0, with the addition of an inductive or capacitive reactance into the loop, as shown in Fig. 4.38(a).

It must be noted that the input impedance of an amplifier is usually a parallel combination of a capacitance and a resistance, as shown in Fig. 4.38(b). To investigate the matching conditions, it is convenient to use the series equivalent of the input impedance, as shown in Fig. 4.38(c). The conversion formulas can be easily derived as

$$r_i = r_p \frac{1}{1 + (\omega \cdot C_p r_p)^2} \quad C_i = C_p \left(1 + \frac{1}{(\omega \cdot C_p r_p)^2}\right) \tag{4.78}$$

As seen from (4.78), r_i and C_i are frequency-dependent and must be calculated for $\omega = \omega_0$.

The output of the LNA drives the output resonance circuit that is also tuned to f_0. To prevent the loading of the output resonance circuit, the output resistance of the amplifier must be as high as possible. Consequently, the appropriate configuration for an LNA is a low input impedance–high output impedance circuit. The voltage gain of the LNA is determined by the transconductance and the load impedance. Therefore, it is appropriate to characterize the amplifier with its input and output impedances and the transconductance, g_m.

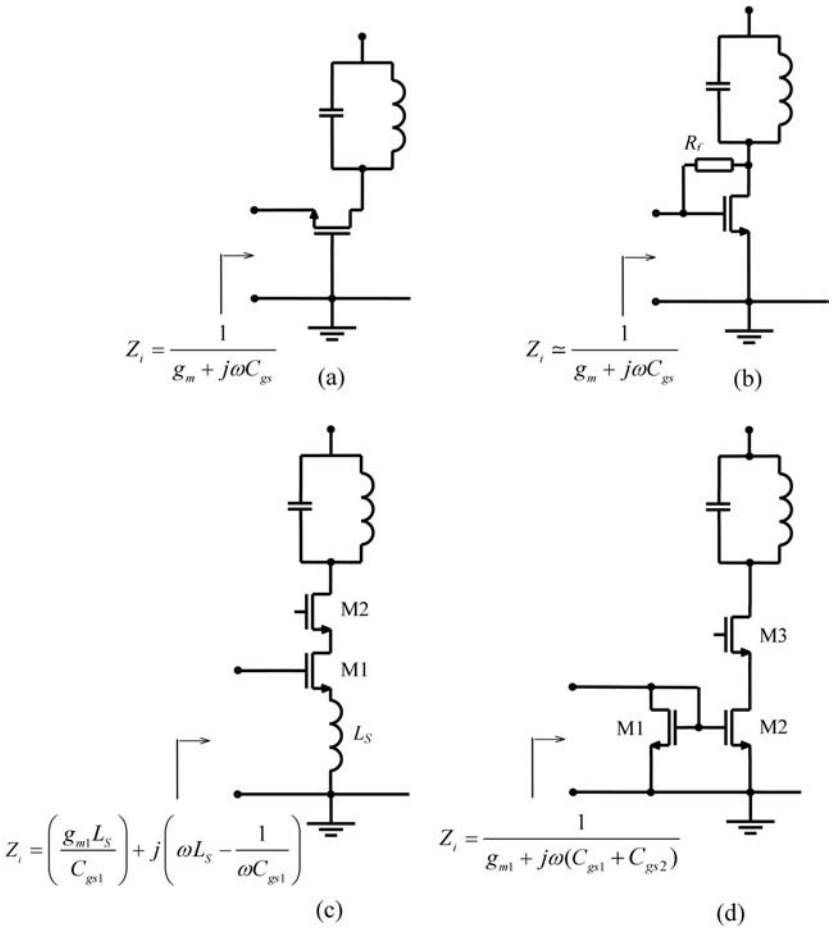


Figure 4.39 The basic circuits suitable to use as an LNA: (a) the common-gate amplifier, (b) reduction of the input impedance of a common-source amplifier with parallel voltage feedback, (c) the source degenerated and cascoded LNA, (d) the current mirror input cascoded circuit.

There are several circuit topologies that fulfill the low input impedance and a high output impedance conditions. Additional factors include high transconductance, low internal feedback, low power consumption and low noise.

In Fig. 4.39, the most basic circuits that fulfill the impedance conditions are shown. The basic properties of these circuits are summarized below. Differential (symmetrical) versions of these amplifiers are convenient for differential antennas, i.e. all versions of dipole antennas. The noise performances of these amplifiers will be compared later on, after the investigation of the noise performances of amplifiers.

The input impedance of the common-gate amplifier shown in Fig. 4.39(a) is the parallel combination of $1/g_{\mathrm{m}}$ and the input capacitance of the transistor. According to (4.78), the real part of the input impedance is approximately equal to $1/g_{\mathrm{m}}$ for $f_0 \ll f_{\mathrm{T}}$, where f_0 and f_{T} are the frequency of operation of the amplifier and the high-frequency

figure of merit of the transistor, respectively. The real part of the input impedance can be made equal to the real part of the antenna impedance, with an appropriate design of the aspect ratio and the quiescent current of the transistor. The output resistance is high as derived in Section 3.3, and its loading, on the output resonance circuit is negligible. Since there are no direct parasitics between the output and input nodes, the internal feedback also is negligible. Consequently, the common-gate amplifier must be considered as one of the usable candidates for LNAs.

It can be shown that the input impedance of a parallel voltage feedback amplifier shown in Fig. 4.39(b) becomes equal to the parallel combination of $1/g_m$ and the input capacitance of the transistor, provided that $g_{ds} \ll g_m$ and $(1/R_F) \ll g_m$. This means that it is a low-input impedance circuit. But it must be kept in mind that the parallel voltage feedback also decreases the output impedance of the amplifier. Therefore, owing to the excessive loading of the output resonance circuit, this is not a good candidate as a tuned LNA, but can be used as a wide-band LNA with an appropriate resistive load.

The circuit shown in Fig. 4.39(c) and its differential versions are the most widely used circuits as LNAs. The use of the cascode configuration to reduce the output to input feedback is useful, even necessary, as explained in Section 3.4. Since the cascode circuit exhibits a high output resistance, the loading of the output resonance circuit is also negligible. As already derived in Section 3.6, the input impedance of the circuit has a real part equal to $g_{m1}L_S/C_{gs1}$ that can be designed to obtain a resistance value equal to the real part of the antenna impedance with appropriate g_m and L_S values.

The reactive components of the input impedance can be eliminated by resonance together with the reactive part of the antenna impedance and an additional series reactance, as shown in Fig. 4.38(a). An additional advantage of the circuit is the voltage increase at the G-S port of the transistor owing to the series resonance, as explained in Section 4.1.2.

The input impedance of the circuit proposed in Fig. 4.39(d) is determined by M1, of which the input impedance is the parallel combination of the $1/g_m$ and the input capacitances of M1 and M2. M2 is the second transistor of the current mirror that is cascoded with M3 to reduce the internal feedback. The gain of the circuit can be enhanced when the current transfer ratio of the current mirror is higher than unity. The obvious disadvantage of the circuit is higher power consumption owing to the second DC current path and the additional noise owing to the third transistor.

4.6.3 Noise in amplifiers

At this point, it is worthwhile to examine the notion of noise, and modeling of noise in fundamental circuits. Amplifiers are being used to enhance the signal that is generated by a weak signal source; for example the signal generated by a microphone (which is proportional to the sound pressure), connected to the input of an audio amplifier, or the signal at the output of an antenna applied to the input of a receiver, which is proportional to the electromagnetic field strength at the tuning frequency of the receiver. In addition to the original signal to be amplified, there is usually a physical phenomenon that disturbs the amplified signal in a certain way: for example the

high-frequency "hiss" that we hear from the loudspeaker connected to the output of the audio amplifier when the input signal level is low, or the random speckles we see on the screen of a TV receiver when we receive from a distant transmitter. This phenomenon is called "noise" and is generated in all kinds of resistors and devices like diodes and transistors, owing to the random motions of electrons. In the scope of this book we will concentrate on the noise of LNAs. But it must be kept in mind that the physical mechanisms and the basic definitions are common for all kinds of amplifiers.

LNAs are being used to receive and amplify weak signals transmitted by distant transmitters, which may be fixed or mobile, or installed on board a satellite. The total signal received at the input of the amplifier does not only consist of the signal sent by the transmitter,[17] but in addition, it includes the unavoidable noise signal originating from the internal resistance of the antenna. To obtain a sufficiently high level of signal power with a reasonable signal-to-noise ratio (*S/N*) at the output of the LNA, the noise inherently generated in the amplifier must be kept as low as possible.

The noise performance of an amplifier is usually expressed by the "noise factor", *F*, that is defined as

$$F = \frac{(S/N)_{\text{input}}}{(S/N)_{\text{output}}} \tag{4.79}$$

The LNA amplifies equally the incoming signal and the input noise generated in the signal source. For a noiseless amplifier the output signal power and the output noise power are equal to the gain multiplied by the input signal and the input noise power, respectively. Therefore the noise factor of a noiseless amplifier is equal to unity. In a noisy amplifier, on the other hand, the output noise is the sum of the gain times the input noise and the output noise component representing the noise generated in the amplifier. From this consideration and (4.79) it can be concluded that

$$F = \frac{\text{total output noise power}}{\text{output noise component owing to the input noise power}} \tag{4.80}$$

and

$$F = 1 + \frac{\text{output noise power component generated by amplifier}}{\text{gain} \times \text{input noise power}} \tag{4.81}$$

The "noise figure", *NF*, which is also being used to express the noise performance of an amplifier, is the noise factor expressed in "decibels":

$$NF(\text{dB}) = 10 \times \log F \tag{4.82}$$

The noise generated in an amplifier mainly originates from the random movements of charge carriers in resistors and devices, owing to their thermal energy. Apart from

[17] In addition to the signal of interest sent by the transmitter, certain natural (atmospheric, cosmic, etc.) and man-made (originating from switching of power lines, corona discharges, etc.) noises can reach the input of the amplifier. Since these "external" noises are all sporadic and in some cases avoidable to some extent, they will be kept out of this discussion.

this "thermal noise", there are several other types of noise: the partition noise, the multiplication noise, the flicker (or 1/*f*) noise, etc. But since these other noise components are usually effective at lower frequencies and/or dominated by the thermal noise, they can be ignored for LNAs.

The thermal noise was first observed and measured by J. B. Johnson in 1928 and interpreted by H. Nyquist who derived an expression giving the value of the noise power in a conductor owing to the random movements of electrons:

$$P_{\mathrm{n}} = 4kTB \tag{4.83}$$

where k is the Boltzmann constant (1.36×10^{-23} joules/K), T is the temperature of the conductor in K, and B is the bandwidth of interest. B can be placed anywhere on the frequency axis, indicating the "white" (frequency-independent) character of the noise. Expression (4.83) shows that the noise power is the same for all conductors for a certain temperature and for a bandwidth of interest, *regardless of the material and shape of the conductor*.

4.6.3.1 Thermal noise of a resistor

If the electrical resistance of a conductor is R, the noise power can be expressed in terms of the mean square noise voltage between the terminals of the resistor, or the mean square noise current flowing through the resistor as

$$P_{\mathrm{n}} = \bar{i}_{\mathrm{n}}^2 \times R \quad \text{or} \quad P_{\mathrm{n}} = \bar{v}_{\mathrm{n}}^2 \times \frac{1}{R}$$

Consequently, the root mean square (effective) values of the noise current and the noise voltage can be written as

$$\bar{i}_{\mathrm{n}} = \sqrt{4kTB\frac{1}{R}} \text{ and } \bar{v}_{\mathrm{n}} = \sqrt{4kTB\times R} \tag{4.84}$$

and the noise equivalent circuit of a resistor can be drawn as seen in Fig. 4.40.

To ease the comparison of noise behavior of different devices (or circuits) and to conform with the existing noise measurement systems that usually measure the noise in a narrow band, it is common practice to express the noise for a 1 Hz bandwidth.

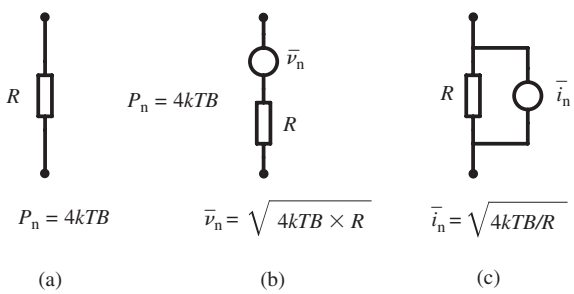


Figure 4.40 (a) A resistor and the associated noise power. The noise equivalent circuit with the noise voltage source (b), and with the noise current source (c).

Hence, the mean square noise voltage and the mean square noise current for a 1 Hz bandwidth can be written as

$$\bar{v}_n^2 \Big|_{B=1\text{Hz}} = 4kT \times R = S_v \quad \text{and} \quad \bar{i}_n^2 \Big|_{B=1\text{Hz}} = 4kT/R = S_i, \tag{4.85}$$

called the "spectral density" of the mean square noise voltage and noise current, respectively.

Problem 4.9

(a) Calculate the thermal noise voltage and thermal noise current of a 50 ohm resistance (i) for 30°C, (ii) for 100°C (the bandwidth of interest is 10 MHz).
(b) Calculate the thermal noise voltage and thermal noise current of a 1000 ohm resistance for the same temperatures and the same bandwidth.
(c) Compare and discuss the results.

4.6.3.2 Thermal noise of a MOS transistor

The thermal noise of a MOS transistor was investigated and modeled by A. van der Ziel in 1986 [34]. In the physical structure of a MOS transistor there are several "resistances" as shown in Fig. 4.41, that generate noise according to (4.84).

The total resistance between the external source node (S in Fig. 4.41) to the external drain node (D) of the transistor is the sum of the source series resistance (R_S), the channel resistance (R_{ch}) and the drain series resistance (R_D). The source series resistance (and similarly the drain resistance) is the sum of the intrinsic and extrinsic components and the equivalent contact resistance, as explained in Chapter 1. The source and drain series resistances are obviously technology- and geometry-dependent and can be more than one hundred ohms for small transistors and several ohms for large transistors. The gate series resistance, another noise source, is also technology- and geometry-dependent and can be minimized with appropriate finger structures [35].

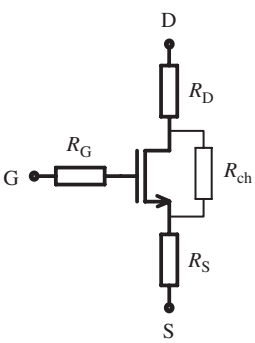


Figure 4.41 The MOS transistor with its noise generating resistances.

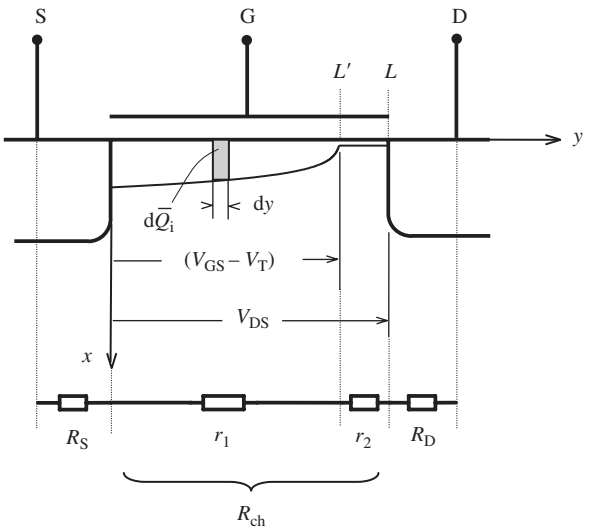


Figure 4.42 The cross-section of a MOS transistor in saturation and components of the series resistances of the drain current path.

It is known that the strongly dominant part of the MOS transistor noise is the thermal noise associated with the channel resistance [36]. In several publications this noise is investigated as the sum of the noises of the pre-pinch-off region of the channel and that of the pinched-off region [37], [38], [39]. On the other hand, it has been shown that the effect of V_{DS}, consequently the contribution of the pinched-off region on the drain current noise, is negligible [40], [41].

In this section, the noise associated with the channel resistance of a MOS transistor operating in the saturation region will be derived with a similar but more straightforward approach based on the expression derived in Chapter 1 for the inversion charge.

The cross-section of a non-velocity saturated NMOS transistor in the saturation region is shown in Fig. 4.42. According to (1.7), the effective gate voltage (and the inversion charge that is proportional to the effective gate voltage) varies with the square root of y, until the pinch-off point, L', that corresponds to a channel voltage of $(V_{DS} - V_T)$. Along the saturation region (from L' to L) electrons travel with the saturation velocity, v_{sat}, and the inversion charge density is constant (see Fig. 1.6). It is obvious that these two sections of the channel resistance shown as r_1 and r_2 are different in nature and, as mentioned above, from the point of view of noise, the strongly dominant part of the channel resistance is r_1. The value of r_1 corresponding to R_{ch} can be calculated based on the modified gradual channel approach developed in Chapter 1. The resistance of a channel element dy before the pinch-off point can be written as

$$dr(y) = \frac{dV_c}{I_D} \tag{4.86}$$

and the drain current, which is constant along the channel, is

$$I_D = \frac{d\bar{Q}_i(y)}{dt} \tag{4.87}$$

where $d\bar{Q}_i(y)$ is the inversion charge in the dy channel element that was calculated in Chapter 1 (1.37) as

$$d\bar{Q}_i(y) = -C_{ox}W(V_{GS} - V_T)\sqrt{1 - \frac{y}{L'}}\ dy$$

dt in (4.87) can be written as

$$dt = \frac{dy}{v} = \frac{dy}{\mu E(y)} = \frac{dy}{-\mu(dV_c/dy)} \tag{4.88}$$

From (4.86), (4.87) and (1.37) the resistance of a channel element dy can be calculated:

$$dr(y) = \frac{dy}{\mu C_{ox}W(V_{GS} - V_T)\sqrt{1 - \frac{y}{L'}}} \tag{4.89}$$

The integral of d$r(y)$ from $y = 0$ to $y = L'$ gives the value of the first (pre-pinch-off) section of the channel resistance:[18]

$$R_{ch} = \frac{2}{\mu C_{ox}\frac{W}{L'}(V_{GS} - V_T)} \tag{4.90}$$

where L' must be considered as

$$L' = L\frac{1 + \lambda(V_{GS} - V_T)}{1 + \lambda V_{DS}} \tag{4.91}$$

according to (1.14b). It must be also noted that μ in (4.90) is a function of V_{GS}, as mentioned in Section 1.1.2.1. and Appendix A.

Equation (4.90) can be arranged as

$$R_{ch} = \frac{(V_{GS} - V_T)}{I_D} \tag{4.92}$$

which is – not surprisingly – the DC resistance of the inversion region.

If we insert $(V_{GS} - V_T)$ in terms of I_D, (4.92) can be re-written as

$$R_{ch} = \sqrt{\frac{2}{I_D\mu\ C_{ox}\frac{W}{L}}} \tag{4.93}$$

or in terms of g_m,

$$R_{ch} = \frac{2}{g_m} \tag{4.94}$$

[18] Note that this is equal to $2/g_{do}$, where g_{do} is the output conductance corresponding to $V_{DS} \to 0$, in the original noise expressions of van der Ziel [34].

These expressions can be interpreted as follows.

- R_{ch} decreases with the square root of the drain current, I_D.
- R_{ch} decreases with the square root of the aspect ratio.
- R_{ch} decreases with the square root of the mobility. The bias dependence of the mobility must not be ignored for small-geometry devices.
- Since the mobility of holes is considerably smaller than the mobility of electrons, the channel inversion resistance of a PMOS transistor is higher than that of an NMOS transistor having the same geometry and the same drain current.
- R_{ch} decreases with the square root of C_{ox}. This means that the channel inversion resistance is inferior for a smaller-geometry transistor having the same aspect ratio.

Now the noise current and the noise spectral density corresponding to the channel resistance of a MOS transistor can be written as

$$\bar{i}_{nd}^2 = \frac{4kTB}{R_{ch}} = 2kTBg_m \ [A^2], \tag{4.95a}$$

$$S_{ich} = \bar{i}_{nd}^2 \ \Big|_{B=1Hz} = \frac{4kT}{R_{ch}} = 2kTg_m \ \ [A^2/Hz]. \tag{4.95b}$$

This expression and the interpretation related to the channel resistance show us that the channel current noise

- increases with the drain current,
- increases with the aspect ratio,
- is higher at small-geometry devices and
- the channel current noise of a PMOS transistor is lower than that of an NMOS transistor having the same geometry and the same drain current.

As an example, the variation of R_{ch} and S_{ich} for an AMS 0.35 ($W = 35$ μm, $L = 0.35$ μm) transistor as a function of the drain current is plotted in Fig. 4.43. Figure 4.43(a)

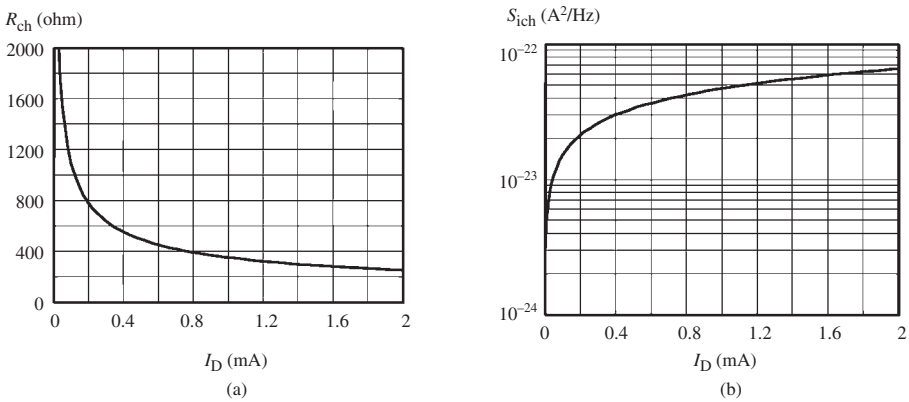


Figure 4.43 (a) The inversion channel resistance and (b) the drain noise current spectral density of a 35 μm/0.35 μm AMS transistor. (The V_{GS} dependence of μ is taken into account.)

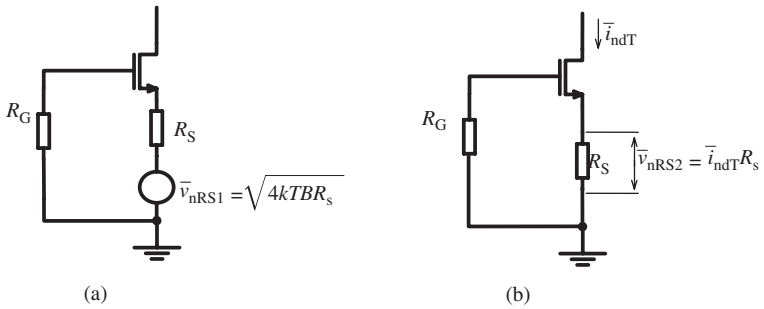


Figure 4.44 (a) The noise voltage of the source series resistance. (b) The noise voltage owing to the total drain noise current flowing through R_S.

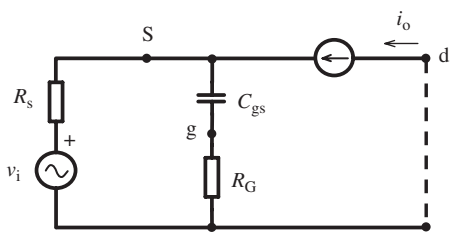


Figure 4.45 The small-signal equivalent circuit used to calculate $g_{mSD(eff)}$.

shows that the inversion channel resistance acquires considerably low values for moderate to high drain currents. Therefore, if the series source and drain resistances are not sufficiently small compared to R_{ch}, they must not be ignored.

Since the noise contributions of the parasitic internal resistances (R_S, R_D and R_G) depend not only on the transistor but also on the circuit, they must be investigated for different configurations separately. For this example, R_D can be considered as the effective load resistance. To gain insight about the contribution mechanisms of R_S and R_G, we will calculate their effects for the most basic common-source amplifier.

The source series resistance has two effects on the noise behavior of the transistor.

(a) The noise voltage, $\overline{v}_{nRS1}$, generated by the source series resistance of the transistor, is shown in Fig. 4.44(a). This voltage adds a component to the drain noise current, equal to $\overline{i}_{ndS1} = \left|g_{mSD(eff)}\right|.\overline{v}_{nRS1}$, where $g_{mSD(eff)}$ is the effective transconductance from the source noise source-to-drain, and can be calculated from Fig. 4.45 as

$$g_{mSD(eff)} = -\frac{g_m}{R_S(g_m + j\omega C_{gs}) + (1 + j\omega C_{gs}R_G)} \tag{4.96a}$$

$$\left|g_{mSD(eff)}\right| = \frac{g_m}{\sqrt{(1 + g_m R_S)^2 + \omega^2 C_{gs}^2 (R_S + R_G)^2}} \tag{4.96b}$$

(b) The noise voltage drop on R_S owing to the noise current flowing through R_S, which is $\overline{v}_{nRS2} = \overline{i}_{ndT}R_S$. This noise voltage provokes a drain noise current equal

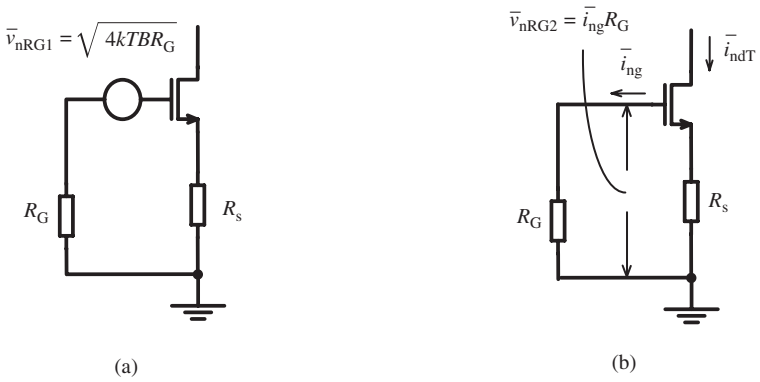


Figure 4.46 (a) The noise voltage source owing to the gate series resistance. (b) The noise voltage source owing to the gate noise current.

to $\bar{i}_{nDS2} = \left|g'_{m(S'D)(eff)}\right| . \bar{v}_{nRS2}$, where $g_{mSD}(\text{eff})'$ is the effective transconductance of the circuit for $R_S = 0$:

$$g'_{mSD(eff)} = -\frac{g_m}{(1 + j\omega C_{gs}R_G)} \tag{4.97a}$$

$$\left|g'_{mSD(eff)}\right| = \frac{g_m}{\sqrt{1 + (\omega C_{gs}R_G)^2}} \tag{4.97b}$$

Another important parasitic resistance of a MOS transistor is the gate resistance, whose value is strongly layout- and process-dependent. Owing to its distributed capacitance along the width of the transistor, or in the case of a multi-finger structure along the fingers, the effective value of the gate resistance decreases with frequency. However, since the externally connected series resistances to the gate usually dominate, the decrease of the inherent gate resistance of the transistor can be neglected.

The contribution of the gate series resistance to the noise is two-fold.

(a) It generates a thermal noise voltage equal to $\bar{v}_{nRG1} = \sqrt{4kTBR_G}$ that is in the input loop of the transistor and provokes a drain noise current component equal to $\bar{i}_{ndG1} = \left|g_{m(eff)}\right|\bar{v}_{nR_G}$.[19] This is a thermal (white) noise and is not co-related to $\bar{i}_{ndT}$. If there is an external resistance connected in series to the gate, it must be considered together with the inherent gate resistance of the transistor (see Fig. 4.46(a)).

(b) At high frequencies a noise current owing to the noise voltage on the inversion channel flows over the gate capacitance. This noise current ($\bar{i}_{ng}$) is not "white" since it increases with frequency, and is co-related to the inversion channel noise. The noise voltage drop on R_G owing to this current, $\bar{v}_{nRG2} = \bar{i}_{ng}R_G$, is another

[19] $g_{m(eff)}$ is the effective transconductance defined in (2.13c). If there is an impedance (Z_S) connected in series to R_S, (2.13c) must be modified as $g_{m(eff)} = 1/[1 + g_m(R_S + Z_S)]$.

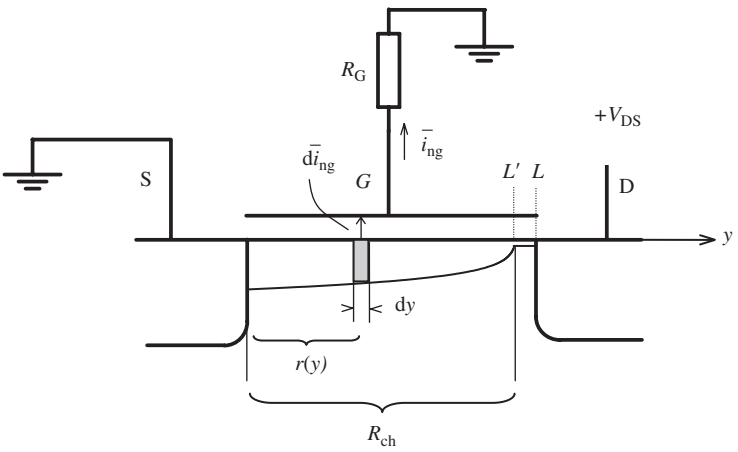


Figure 4.47 The gate noise current as the sum of the incremental components.

noise voltage source in the input loop and adds another component on to the drain noise current $\bar{i}_{ndG2} = \left|g_{m(eff)}\right|\bar{v}_{nRG2}$ (see Fig. 4.46(b)).

The approach used to calculate the gate noise current ($\bar{i}_{ng}$) is shown in Fig. 4.47. The noise voltage of a channel element dy at position y is

$$\bar{v}_n(y) = \bar{i}_{nd}.r(y)$$

where $\bar{i}_{nd}$ is the channel noise current and $r(y)$ is the resistance of the channel segment from the source end of the channel ($y = 0$) to y, that is equal to the integral of dr given in (4.89), from zero to y:

$$r(y) = \frac{2}{\mu C_{ox}\frac{W}{L'}(V_{GS} - V_T)}\left(1 - \sqrt{1 - \frac{y}{L'}}\right) \tag{4.98}$$

$\bar{v}_n(y)$ induces an incremental noise current over the incremental capacitance $dC_g = C_{ox}W dy$, that is

$$d\bar{i}_{ng} = j\omega(dC_g) \times \bar{v}_n(y) = j\omega(C_{ox}W dy) \times r(y)\bar{i}_{nd}$$

and hence

$$\bar{i}_{ng} = j\omega C_{ox}W\frac{2}{\mu C_{ox}\frac{W}{L'}(V_{GS} - V_T)}\int_0^{L'}\left(1 - \sqrt{1 - \frac{y}{L'}}\right)dy \times \bar{i}_{nd} \tag{4.99}$$

For $L \cong L'$, and from (1.33) and (1.40) this expression can be arranged as[20]

$$\left|\bar{i}_{ng}\right| \cong \omega\frac{C_{gs}}{g_m} = \omega\frac{1}{A}\sqrt{\frac{W}{I_D}} \times \bar{i}_{nd} \tag{4.100}$$

[20] It must be noted that this derivation is valid provided that $(1/\omega C_{gs}) \gg r_d$, in other words the noise current deviated to the gate is sufficiently smaller than the drain noise current.

where A, which was given in (1.43c), as

$$A = \sqrt{\frac{2\mu}{C_{ox}L^3k_{ol}^2}} \tag{4.101}$$

From (4.100) and (4.101), it can be seen that the gate noise current

- is proportional to frequency,
- is proportional to the drain inversion channel noise,
- strongly increases with L,
- is higher for thin-gate oxide transistors,
- is higher for wide transistors,
- is smaller for higher drain DC currents (for smaller channel resistances).

We have seen that in the composition of the total drain mean square noise current there are two classes of components: (a) those that are uncorrelated to each other or to the total noise current and (b) those that are correlated to the total drain noise current. The sum of the uncorrelated components can be calculated with superposition, since each of them exists even when the others are zero.

$$\overline{i}_{ndU}^2 = \overline{i}_{nd}^2 + \overline{i}_{ndS1}^2 + \overline{i}_{ndG1}^2 \tag{4.102}$$

where

$$\overline{i}_{nd}^2 = \frac{4kTB}{R_{ch}} = 2kTBg_m \tag{4.103}$$

$$\overline{i}_{ndS1}^2 = \left|g_{mSD(eff)}\right|^2.\overline{v}_{nR_S}^2 = \left|g_{mSD(eff)}\right|^2 \cdot 4kTB_S \tag{4.104}$$

$$\overline{i}_{ndG1}^2 = \left(\left|g_{m(eff)}\right|.\overline{v}_{nR_G}\right)^2 = \left|g_{m(eff)}\right|^2.4kTBR_G \tag{4.105}$$

The correlated components can be written as

$$\overline{i}_{ndS2}^2 = \left|g'_{mSD(eff)}\right|^2 R_S^2 \cdot \overline{i}_{ndT}^2 = a_1 \cdot \overline{i}_{ndT}^2 \tag{4.106a}$$

where

$$a_1 = \left|g'_{mSD(eff)}\right|^2 R_S^2 \tag{4.106b}$$

$$\begin{aligned}\overline{i}_{ndG2}^2 &= \left(\left|g_{m(eff)}\right|\overline{v}_{nRG2}\right)^2 = \left|g_{m(eff)}\right|^2 R_G^2 \cdot \overline{i}_{ng}^2 \\ &= \left|g_{m(eff)}\right|^2 \left(\omega\frac{1}{A}\sqrt{\frac{W}{I_D}}\right)^2 R_G^2 \cdot \overline{i}_{ndT}^2 = a_2 \cdot \overline{i}_{ndT}^2\end{aligned} \tag{4.107a}$$

where

$$a_2 = \left|g_{m(eff)}\right|^2 \left(\omega\frac{1}{A}\sqrt{\frac{W}{I_D}}\right)^2 R_G^2 \tag{4.107b}$$

To find the total drain mean square noise current, all these uncorrelated and correlated components must be superposed:

$$\begin{aligned}\overline{i}_{\mathrm{ndT}}^2 &= \overline{i}_{\mathrm{ndTU}}^2 + \overline{i}_{\mathrm{ndS2}}^2 + \overline{i}_{\mathrm{ndG2}}^2 \\ &= \overline{i}_{\mathrm{ndTU}}^2 + \alpha_1 \cdot \overline{i}_{\mathrm{ndT}}^2 + \alpha_2 \cdot \overline{i}_{\mathrm{ndT}}^2\end{aligned} \tag{4.108}$$

This expression can be arranged as

$$\overline{i}_{\mathrm{ndT}}^2 = \frac{\overline{i}_{\mathrm{ndU}}^2}{[1-(\alpha_1+\alpha_2)]} = \frac{\overline{i}_{\mathrm{nd}}^2 + \overline{i}_{\mathrm{ndS1}}^2 + \overline{i}_{\mathrm{ndG1}}^2}{[1-(\alpha_1+\alpha_2)]} \tag{4.109}$$

or in terms of its parameters,

$$\overline{i}_{\mathrm{ndT}}^2 = 4kTB\frac{(g_{\mathrm{m}}/2) + \left|g_{\mathrm{mSD(eff)}}\right|^2 R_{\mathrm{S}} + \left|g_{\mathrm{m(eff)}}\right|^2 R_{\mathrm{G}}}{[1-(\alpha_1+\alpha_2)]} \tag{4.110}$$

It must be noted that the gate noise current generates a noise voltage also on the signal source resistance, and adds to the thermal noise of the signal source. For the noise figure calculations, this component must be excluded from the signal source noise and included in the amplifier noise.

Example 4.10 In the following, the contribution of different components of the drain noise current will be calculated for a 35 µm (7×5 µm) / 0.35 µm AMS NMOS transistor at $f = 3$ GHz for 10 MHz bandwidth. The DC operating point of the transistor is $I_{\mathrm{D}} = 2$ mA, $V_{\mathrm{DS}} = 3$ V.

The parasitic source and drain series resistances of a 5 µm/0.35 µm transistor were calculated as 134 ohm each, in Example 1.3. The 35 µm width transistor is composed of seven 5 µm transistors as a multi-finger transistor, as shown in Fig. 1.23. Since source and drain regions are shared by the neighboring transistor and since they are connected in parallel, the source resistance of the 35 µm transistor is approximately 134/14 ≈ 10 ohm.

Since the poly gate sheet resistance for this technology is given as $R_{\mathrm{sh}} = 7$ ohm/□, the resistance of one of the gate stripes is 7×(5/0.35) = 100 ohm. If the gate stripes are parallel, as shown in Fig. 1.23, the equivalent resistance is 100/7 = 14.3 ohm. Poly to metal contact resistance is given as 2 ohm/0.4×0.4 micron contact. Assuming 10 contacts along the collecting stripe, the equivalent contact resistance is 0.2 ohm. Together with the resistance of the collecting stripe, the total resistance series to the gate can be taken as 15 ohm.

The gate-to-source voltage for a 2 mA drain current and the mobility corresponding to this voltage can be found as 1 V and 325 cm^2/V.s, respectively. The transconductance of the transistor for a 2 mA drain current is

$$g_{\mathrm{m}} = \sqrt{2\times 325\times(4.54\times10^{-7})\times(2\times10^{-3})} = 7.53\times10^{-3} \simeq 7.5\,\mathrm{mS}$$

and the effective transconductance with $R_{\mathrm{S}} = 10$ ohm,

$$g_{\mathrm{m(eff)}} = \frac{7.5\times10^{-3}}{1+7.5\times10^{-3}\times10} = 6.98\times10^{-3} \cong 7\,\mathrm{mS}$$

The magnitudes of $g_{mSD(eff)}$ and $g'_{mSD(eff)}$ can be calculated from (4.96b) and (4.97b) as

$$g_{mSD(eff)} = 6.97\,\text{mS} \quad g'_{mSD(eff)} = 6.48\,\text{mS}$$

The A, α_1 and α_2 parameters from (4.101), (4.106b) and (4.107b),

$$A = 2.47\times10^{11} \quad \alpha_1 = 4.2\times10^{-3} \quad \alpha_2 = 8.7\times10^{-5}$$

Now the total drain mean square noise current can be calculated from (4.110) with

$$4kTB = 4\times\left(1.38\times10^{-23}\right)\times300\times\left(10\times10^{6}\right) = 1.656\times10^{-13}\ \text{W}$$

$$\bar{i}^2_{ndT} = 1.656\times10^{-13}\,\frac{3.75\times10^{-3} + \left[(6.97\times10^{-3})^2\times10\right] + \left[(7\times10^{-3})^2\times15\right]}{1-(4.2\times10^{-3}+8.7\times10^{-5})} = 8.27\times10^{-16}\left[\text{A}^2\right]$$

which corresponds to $S_{nd} = 8.27\times10^{-23}$ [A^2/Hz].

This example shows that the major components of the total mean square drain noise current are related to the inversion channel resistance and the series resistances of the gate and source electrodes. The indirect and correlated influences of the drain noise current are comparatively small. Therefore, utmost care is necessary on the layout of the device, and the external series resistance must be kept as low as possible during the design of the circuit.

Following the same procedure, the variation of the drain noise spectral density as a function of the drain DC current of the same transistor has been calculated, taking into account the variations of the mobility, and plotted in Fig. 4.48, with and without contributions of the resistive device parasitics.

Another aspect that must not be overlooked is the "temperature". The "T" in the noise expressions is the temperature of the resistance, in the case of the MOS transistor the temperature of the channel region can be considerably higher than the ambient temperature and the average surface temperature of the die.[21] This fact imposes higher noise for higher power densities, consequently higher device temperature. For example, at 400 K the mean square noise currents increase approximately 33%. In Fig. 4.49 the total drain noise current spectral density of the transistor in Example 4.10 is plotted for 300 K and 400 K.

Problem 4.10. *Derive an expression to calculate approximately the gate noise current assuming that the noise voltage along the channel is constant and equal to the noise voltage at the mid-point of the channel. Compare the result with (4.100) and discuss.*

[21] From the publications related to the thermal simulation and mapping of ICs [42], [43], it can be seen that the temperatures of the directly heated micro-regions (hot spots) can be considerably higher than the average temperature of the die.

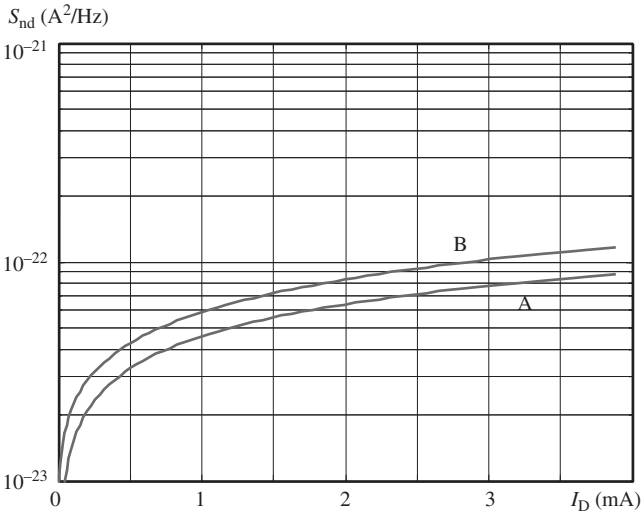


Figure 4.48 The drain current noise spectral density of an AMS 35 μm/0.35 μm NMOS transistor (A) with the channel inversion resistance only, (B) with the contributions of R_S and R_G (for $V_{DS} = 2$ V, $f = 3$ GHz and $T = 300$ K).

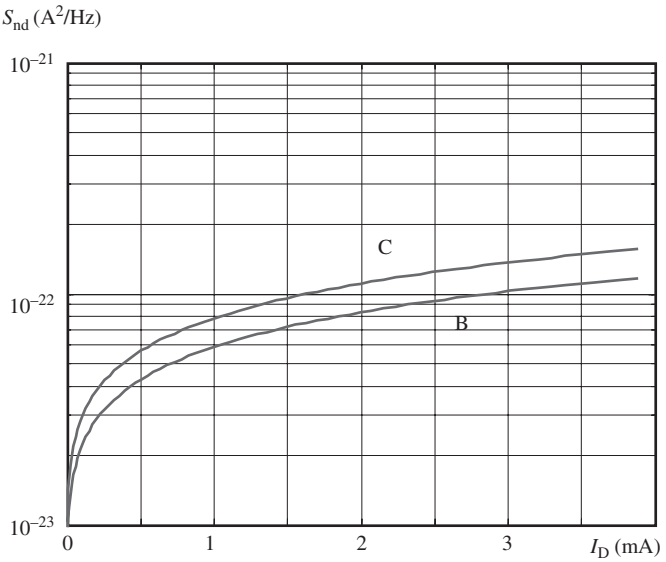


Figure 4.49 The drain current noise spectral density of the transistor in Example 4.10 as a function of the drain DC current: (B) for $T = 300$ K, (C) for $T = 400$ K.

Problem 4.11. *The die area of an integrated circuit is 10 mm^2 and its thickness is 0.5 mm. The die is mounted into an Amkor MLF, 44 lead miniature package, whose thermal resistance from the ambient to the bottom of the die is 24°C/W. The power consumption of the circuit is 1 W and the ambient temperature is 30 °C. Calculate the average surface temperature of the die. (The specific thermal conductance of silicon is 1.5 W/cm.°C.)*

Answer: 54.33 °C! (It is obvious that the temperatures of the channel regions of the individual MOS transistors on the die are considerably higher than this value, depending on their power densities.)

4.6.4 Noise in LNAs

As already mentioned, the noise performance of an amplifier depends on the device and its operating conditions, and the structure of the circuit. Therefore the noise figure of the typical LNA circuits shown in Fig. 4.39 must be investigated separately.

In Fig. 4.50(a) the schematic diagram of the most widely used type of tuned LNAs, the source degenerated cascode amplifier, is shown with its input impedance matching circuit. R_A and X_A represent the resistive and reactive components of the antenna and v_A the open circuit antenna output signal voltage. R_S is the series effective resistance of the source degeneration inductance L_S, that helps to obtain a low input resistance compatible with the internal resistance of a typical antenna or the characteristic impedance of the transmission line connecting the antenna to the amplifier. Z_G is necessary to fulfill the impedance matching, as will be explained later on.[22]

From (3.50), the input impedance seen from the gate terminal of an inductive source degenerated transistor that is shown as z_i' can be written as

$$\begin{aligned} z_i' \cong \frac{1}{y_{in}} &= \frac{1+(g_m+sC_{gs})(sL_S+R_S)}{sC_{gs}} \\ &= \left(\frac{g_m L_S}{C_{gs}}+R_S\right)+sL_S+\frac{1+g_m R_S}{sC_{gs}}, \end{aligned} \tag{4.111}$$

which has a real part equal to

$$r_i' = \left(\frac{g_m L_S}{C_{gs}}+R_S\right) \tag{4.112}$$

a series capacitance equal to

$$C_i' = \frac{C_{gs}}{1+g_m R_S} \tag{4.113}$$

and a series inductance apparently equal to L_S. The total input impedance of the amplifier in the ω domain, together with the matching impedance Z_G, can be written as

$$z_i(\omega) = z_i' + Z_G = (r_i' + R_G) + j\left(X_G + \omega L_S - \frac{1}{\omega C_i'}\right) \tag{4.114}$$

We have seen that for maximum signal power transfer from antenna to the amplifier, the input impedance of the amplifier must be equal to the conjugate of the antenna impedance, in other words, the real part of the input impedance of the amplifier must

[22] If the source and gate series resistances of the transistor are not negligibly small, they must be considered as parts of R_S and R_G.

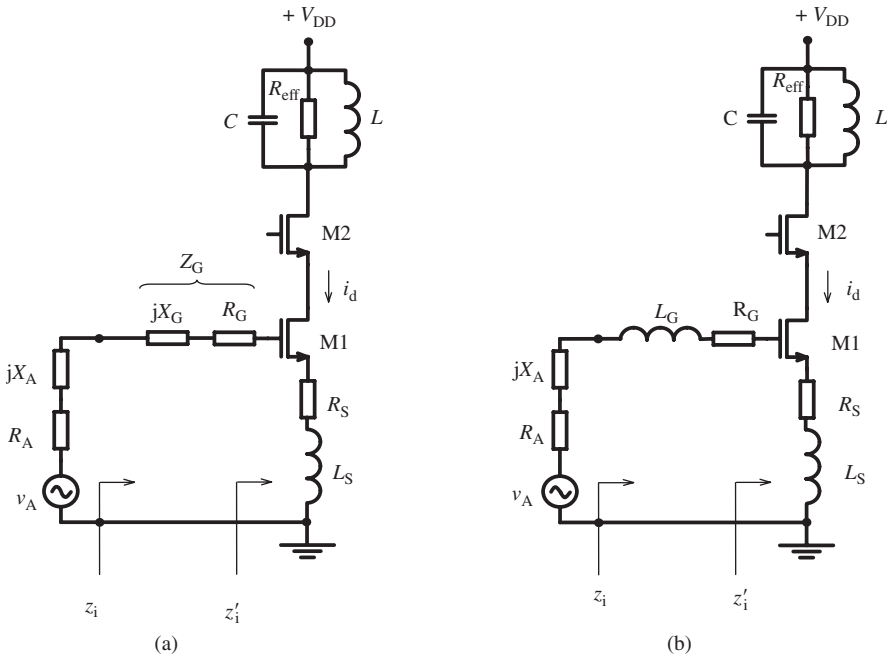


Figure 4.50 The source degenerated LNA: (a) the general form, (b) the most frequently encountered case, an LNA with a gate series inductor.

be equal to R_A and the reactive part must be in resonance with the reactive part of the antenna at ω_0, the operating frequency of the amplifier.

The equality condition of the real parts implies

$$\frac{g_m}{C_{gs}}L_S = R_A - (R_S + R_G) \tag{4.115}$$

which can be arranged as

$$A\sqrt{\frac{I_D}{W}}L_S = R_A - (R_S + R_G) \tag{4.116}$$

where A, a technology-dependent parameter, was given in (1.43c).

$$A = \sqrt{\frac{2\mu}{C_{ox}L^3k_{ol}^2}}$$

From the resonance condition,

$$\begin{aligned} &\left(X_A(\omega_0) + X_G(\omega_0) + \omega_0 L_S - \frac{1}{\omega_0 C'_i}\right) = 0 \\ &\Rightarrow X_G = \left(\frac{1}{\omega_0 C'_i} - \omega_0 L_S - X_A(\omega_0)\right) \end{aligned} \tag{4.117}$$

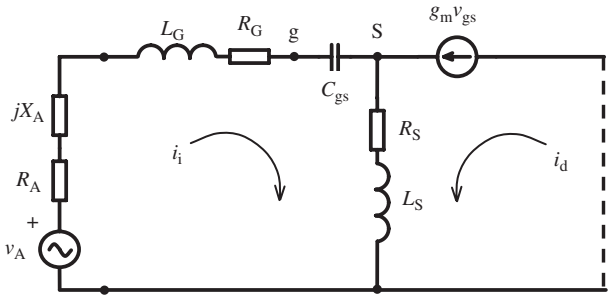


Figure 4.51 The simplified small-signal equivalent circuit of the input transistor.

This expression tells us that the magnitude and sign of X_G depend on the antenna impedance. For resistive, capacitive or slightly inductive antennas, which are the mostly encountered cases, X_G becomes positive, in other words the gate matching impedance must be an inductor:

$$L_G = \left(\frac{1}{\omega_0^2 C_i'} - L_S - \frac{X_A(\omega_0)}{\omega_0}\right) \tag{4.118}$$

R_G, the effective series resistance of L_G, already appears in (4.114). The schematic of the source degenerated LNA for this usual case is shown in Fig. 4.50(b) and will be used for further developments.

It is useful to calculate the transadmittance of M1 from a signal source in the gate loop to the drain current, which helps us to find the voltage gain of the circuit, as well as the noise contributions of R_G and R_S on the drain current. In Fig. 4.51 the simplified small-signal equivalent circuit of M1 is shown. The simplifications are based on the facts that (a) the load of M1 is the low input impedance of M2 that operates as a common-gate circuit, (b) the output internal resistance of M1 is very high compared to the input impedance of M2.

Noting that the reactances in the input loop cancel out according to (4.117), the transadmittance of the circuit for $\omega = \omega_0$ can be calculated as

$$Y_{mG}(\omega_0) = \frac{i_d}{v_A} = \frac{g_m}{g_m R_S + j\omega_0\left[C_{gs}(R_A + R_G + R_S) + g_m L_S\right]} \tag{4.119}$$

Under matching conditions given in (4.114), the transadmittance becomes

$$Y_{mG}(\omega_0) = \frac{g_m}{g_m R_S + j\omega_0 C_{gs}(2R_A)}$$

and its magnitude

$$|Y_{mG}(\omega_0)| = \frac{g_m}{\sqrt{(g_m R_S)^2 + (2\omega_0 C_{gs} R_A)^2}} \tag{4.120a}$$

If the second term of the denominator dominates, this expression can be simplified to

$$|Y_{mG}(\omega_0)| \cong \frac{g_m}{2\omega_0 C_{gs} R_A} \tag{4.120b}$$

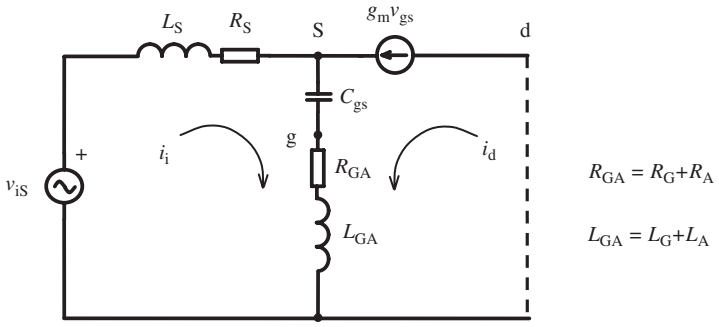


Figure 4.52 Equivalent circuit to calculate the transadmittance for a signal source series to the source of the transistor.

and from (4.115) and (4.116),

$$|Y_{mG}(\omega_0)|^2 \cong \frac{1}{(2\omega_0 R_A)^2} A^2 \left(\frac{I_D}{W_1}\right) \tag{4.120c}$$

Since the common-gate M2 has a current gain equal to unity up to the vicinity of its f_T, the voltage gain of the amplifier can be written as

$$|A_v(\omega_0)| = |Y_{mG}(\omega_0)| \times |Z_L(\omega_0)| = |Y_{mG}(\omega_0)| \times Q_{eff} L\omega_0 \tag{4.121}$$

where L is the inductance of the parallel resonance circuit of the output load and Q_{eff} its effective quality factor.

The contributions of thermal noises of R_G and R_A on the mean square drain noise current of M1 can be calculated as

$$\overline{i}^2_{ndG1} = (\overline{v}_{nR_G} \times |Y_{mG}(\omega_0)|)^2 = 4kTBR_G \times |Y_{mG}(\omega_0)|^2 \tag{4.122}$$

$$\overline{i}^2_{ndA} = (\overline{v}_{nR_A} \times |Y_{mG}(\omega_0)|)^2 = 4kTBR_A \times |Y_{mG}(\omega_0)|^2 \tag{4.123}$$

To calculate the noise contribution of R_S, it is useful to use Y_{ms}, i.e. the transadmittance from source to drain. The small-signal equivalent circuit corresponding to this case is given in Fig. 4.52. The transadmittance from v_{iS} to i_d for $\omega = \omega_0$ can be calculated as

$$Y_{mS}(\omega_0) = \frac{i_d}{v_{iS}} = -\frac{g_m}{g_m R_S + j\omega_0 C_{gs}(2R_A)}$$

which is same in magnitude as the gate transadmittance but opposite in sign, and its magnitude is apparently equal to (4.120a):

$$|Y_{mS}(\omega_0)| = \frac{g_m}{\sqrt{(g_m R_S)^2 + (2\omega_0 C_{gs} R_A)^2}} \tag{4.124}$$

Consequently the contribution of the thermal noise of R_S to the mean square drain noise current of M1 can be written as

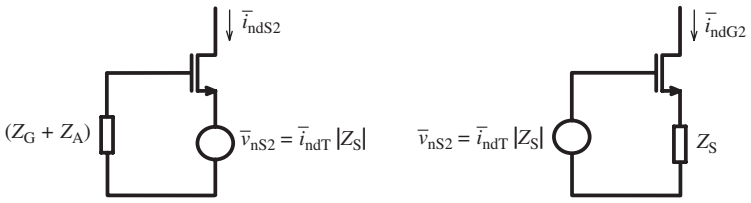


Figure 4.53 Schematics used to calculate (a) Y'_{mG} and (b) Y'_{mS}.

$$\begin{aligned}\bar{i}^2_{ndS1} &= \left(\bar{v}_{nR_S} \times |Y_{mS}(\omega_0)|\right)^2 = 4kTBR_S \times |Y_{mS}(\omega_0)|^2 \\ &= 4kTBR_S \times |Y_{mG}(\omega_0)|^2\end{aligned} \tag{4.125}$$

The major part of the mean square thermal (uncorrelated) noise of the drain current of M1 is the sum of three components: the thermal noise of the inversion channel that was given in (4.95a), the noise owing to the gate resistance R_G and the noise owing to the source resistance R_S.

$$\begin{aligned}\bar{i}^2_{ndU} &= \bar{i}^2_{nd} + \bar{i}^2_{ndG1} + \bar{i}^2_{ndS1} \\ &= 4kT_{IC}B \times (g_m/2) + 4kT_{IC}BR_G \times |Y_{mG}(\omega_0)|^2 + 4kT_{IC}BR_S \times |Y_{mG}(\omega_0)|^2\end{aligned} \tag{4.126}$$

where T_{IC} denotes the temperature of the chip.[23]

There are two additional components of the total drain noise that are the indirect results of the total channel noise, $\bar{i}_{ndT1}$:

(a) the noise component corresponding to the voltage drop on Z_S, owing to the total channel noise current of M1, namely $\bar{i}^2_{ndS2}$;
b) the noise component related to the gate noise current voltage drop in the gate-loop, namely $\bar{i}^2_{ndG2}$.

$\bar{i}^2_{ndS2}$ can be calculated from the schematic given in Fig. 4.53(a) and its small-signal equivalent:

$$\bar{i}^2_{ndS2} = \left(\bar{i}_{ndT1}|Z_S(\omega_0)| \times |Y'_{mS}|\right)^2 = a_1 \cdot \bar{i}^2_{ndT1} \tag{4.127}$$

where $|Y'_{mS}|$ is the transadmittance from $\bar{v}_{nS2}$ to $\bar{i}_{ndS2}$ and can be calculated as

$$|Y'_{mS}| = \frac{g_m}{\sqrt{\left[1 - \omega_0^2 C_{gs}(L_G + L_A)\right]^2 + \left[\omega_0 C_{gs}(R_G + R_A)\right]^2}} \tag{4.128}$$

and

$$a_1 = |Z_S(\omega_0)|^2 \times |Y'_{mS}|^2 \tag{4.129}$$

[23] Note that the channel temperature can be considerably higher than the average temperature of the chip, as mentioned in Problem 4.11.

The gate noise current given in (4.100) flows over the matching impedance connected in series to the gate (that is usually an inductor) and the antenna impedance. Taking into account that C_{gs} is – nearly – in resonance with the gate inductance and the reactive part of the antenna (since usually $L_S \ll L_G$), we can conclude that the noise voltage drop owing to the gate noise current is

$$\bar{v}_{nG2} = \bar{i}_{ng}(R_G + R_A)$$

which can be written as

$$\bar{v}_{nG2} = \omega_0 \frac{1}{A}\sqrt{\frac{W}{I_D}}(R_G + R_A) \times \bar{i}_{ndT}$$

This voltage provokes a drain current noise component:

$$\bar{i}^2_{ndG2} = \left(\bar{v}^2_{nG2} \times \left|Y'_{mG}\right|\right)^2 \tag{4.130}$$

where $\left|Y'_{mG}\right|$ is the transadmittance from $\bar{v}_{nG2}$ to $\bar{i}_{ndG2}$:

$$\left|Y'_{mG}\right| = \frac{g_m}{\sqrt{(1 + g_m R_S)^2 + (\omega_0 g_m L_S)^2}} \tag{4.131}$$

and from (4.107b)

$$a_2 = \left|Y'_{mG}\right|^2 \left(\omega_0 \frac{1}{A}\sqrt{\frac{W}{I_D}}\right)^2 (R_G + R_A)^2 \tag{4.132}$$

Finally, the noise of R_A can be calculated as

$$\bar{i}^2_{ndA} = \bar{v}^2_{nA}|Y_{mG}|^2 = 4kT_A B R_A |Y_{mG}|^2 \tag{4.133}$$

where T_A denotes the temperature of the antenna.

The total mean square noise of the drain current of M1, $\bar{i}^2_{ndT1}$, is the sum (superposition) of all these components:

$$\bar{i}^2_{ndT1} = \bar{i}^2_{nd} + \bar{i}^2_{ndG1} + \bar{i}^2_{ndS1} + \bar{i}^2_{ndS2} + \bar{i}^2_{ndG2}$$

which can be arranged as

$$\bar{i}^2_{ndT1} = \frac{\bar{i}^2_{nd} + \bar{i}^2_{ndG1} + \bar{i}^2_{ndS1}}{[1 - (a_1 + a_2)]} \tag{4.134}$$

and can be arranged in terms of its components:

$$\bar{i}^2_{ndT1} = \frac{4kT_{IC}B}{[1 - (a_1 + a_2)]}\left[\frac{g_{m1}}{2} + |Y_{mG}|^2(R_G + R_S)\right] \tag{4.135}$$

The noise contribution of M1 at the output is

$$P_{nM1(out)} = \bar{i}^2_{ndT1} R_L$$

where R_L is the effective parallel resistance of the output resonance circuit, which can be written in terms of the inductance of the load and its effective quality factor as

$$R_L = L\omega_0 Q_{eff}$$

The total noise power at the output owing to the amplifier, the contribution of the noise current of M2 and the inherent noise of R_L included, becomes

$$\begin{aligned} P_{n(out)} &= (\bar{i}^2_{ndT1} + \bar{i}^2_{nd2})R_L + 4kT_{IC}B \\ &= 4kT_{IC}B\left\{\frac{1}{[1-(a_1+a_2)]}\left[\frac{g_{m1}}{2} + |Y_{mG}|^2(R_G+R_S)\right]R_L + \frac{g_{m2}}{2}R_L + 1\right\} \end{aligned} \tag{4.136}$$

The noise power at the output only owing to the noise of R_A is

$$P_{nA(out)} = \bar{i}^2_{ndA}R_L = 4kT_A B R_A |Y_{mG}|^2 R_L \tag{4.137}$$

Now the noise factor of the amplifier according to (4.81) can be arranged as

$$F = 1 + \frac{T_{IC}}{T_A}\frac{1}{|Y_{mG}|^2 R_A}\left\{\frac{1}{[1-(a_1+a_2)]}\left[\frac{g_{m1}}{2} + |Y_{mG}|^2(R_G+R_S)\right] + \frac{g_{m2}}{2} + \frac{1}{R_L}\right\} \tag{4.138}$$

Now the expression (4.137) together with (4.120c), (4.129) and (4.132) can be interpreted as follows.

- The noise factor increases if the temperature of the chip exceeds the temperature of the antenna, which is valid for almost all practical cases.
- The noise factor increases with frequency owing to the frequency dependences of $|Y_{mG}|^2$, a_1 and a_2.
- The noise factor decreases with g_{m1} for small values of g_{m1} (as well as the drain current) and tends to remain constant for larger values of g_{m1}. For higher values of g_{m1}, owing to the decrease of the mobility, the noise factor starts to increase.
- Smaller transconductance values for M2 are better for smaller noise factors. But a smaller gate width and consequently higher voltage drop on M2 results in the decrease of the output signal dynamic range.
- The noise factor is smaller for higher R_L values.
- R_G and R_S, the effective series resistances of L_G and L_S (the parasitic gate and source resistances included) must be small. Their values depend on the corresponding inductances and their Q values, as well as the technology and layout.
- The noise factor increases for higher values of W_1.
- The effect of the value of R_A is also interesting. According to (4.138), the noise factor is smaller for higher R_A values. But it must be noted that for high R_A values, a_2 increases and impairs the noise factor.

These interpretations show that the design of an LNA is not a "one-shot" process; it needs a series of iterations and trade-offs to fulfill the input impedance, gain, power consumption, dynamic range and minimum noise considerations.

Example 4.11 A source degenerated cascode LNA as shown in Fig. 4.50(b) will be designed for $f_0 = 1.86$ GHz ($\omega_0 = 1.17 \times 10^{10}$ rad/s). The restrictions are as follows.

- The technology is similar to the AMS 0.35 micron technology but has a thick metal layer that permits us to realize reasonably high-Q on-chip inductors. The main parameters related to the NMOS transistors are $C_{ox} = 4.54 \times 10^{-7}$ F/cm^2, $k_{ol} = 0.74$,[24] $\mu_n = 325$ (as an average value that corresponds to $V_{GS} = 1$ V[25]). The numerical value of the A parameter is calculated as 2.47×10^{11}.
- The DC current shall not exceed 4 mA.
- The on-chip inductors will be chosen from a library that contains a number of well-characterized inductors: $L = 0.5$ nH, 1 nH, 2 nH, 5 nH and 10 nH and $Q = 10$ at 1.86 GHz.
- The antenna is a "folded loop antenna" designed for 1.86 GHz, whose impedance is $Z_A = 44.7 - j\ 9.73$ ohm, for unbalanced use.[26]
- The temperature of the antenna and the average temperature of the chip are assumed to be 300 K and 330 K, respectively.

The design can be initiated at one of several possible starting points. Let us start with (4.116), with a DC current smaller than the allowed maximum value to reserve a margin for probable iterations, and the possible minimum L_S value; $I_D = 3$ mA, $L_S = 0.5$ nH. It is necessary to estimate a value for $R_A - (R_S + R_G)$ in (4.116), and check its validity later on. Let us assume $R_A - (R_S + R_G) = 30$ ohm. With these starting values (4.116) yields $W = 508$ μm that corresponds to $C_{gs} = 597$ fF according to (1.40).

Now L_G can be calculated from (4.118):

$$L_G = \frac{1}{(1.17 \times 10^{10})^2 (597 \times 10^{-15})} - 0.5 \times 10^{-9} - \left(\frac{-9.73}{(1.17 \times 10^{10})} \right) = 12.53\,\text{nH}$$

Since we have to use inductors from the library, it is necessary to perform an iteration with 10 nH, the closest inductor value to 12.53 nH. With $L_G = 10$ nH (4.118) yields $C_{gs} = 755$ fF, which corresponds to $W = 642$ μm. To satisfy (4.116) with this value and $L_S = 0.5$ nH, I_D must be 3.79 mA which is smaller than the allowed maximum value.

Now we must calculate the source and drain series resistance and check if the condition on $(R_S + R_G)$ is reasonably close to the initial assumption.

R_S is the sum of the effective series resistance of L_S and the inherent source series resistance of M1. The series resistance R_{SL} of L_S can be calculated from (4.14) as

$$R_{SL} = \frac{L_S \omega_0}{Q} = \frac{(0.5 \times 10^{-9})(1.17 \times 10^{10})}{10} \cong 0.585\,\text{ohm}$$

[24] See Expression (1.40).
[25] See Appendix A.
[26] See Reference [44].

The inherent source series resistance of M1, R_{SM1} is calculated as in Example 4.10 and found to be 0.545 ohm. Therefore,

$$R_S = R_{SL} + R_{SM1} = 1.13 \text{ ohm}$$

Similarly, the total series gate resistance, R_G, which is the sum of the effective series resistance of L_G and the inherent gate series resistance of M1 can be found as 12.52 ohm. With these values

$$R_A - (R_S + R_G) = 44.7 - (1.13 + 12.52) = 31.05 \text{ ohm}$$

which is reasonably close to the initial assumption.

The parameters in (4.138) now can be calculated.

From (1.26)

$$g_{m1} = \sqrt{2\mu_n C_{ox}(W/L)I_D} = \sqrt{2\times 325\times(4.54\times 10^{-7})(642/0.35)(3.8\times 10^{-3})} = 44.7 \text{ mS}$$

From (4.120)

$$|Y_{mG}(\omega_0)|^2 = \frac{(44.7\times 10^{-3})^2}{[(44.7\times 10^{-3})\times 1.13]^2 + [2\times(1.17\times 10^{10})(755\times 10^{-15})\times 44.7]^2} = 3.2\times 10^{-3}\left[S^2\right]$$

From (4.128)

$$|Y'_{mS}(\omega_0)|^2 = \frac{(44.7\times 10^{-3})^2}{\left[1-(1.17\times 10^{10})^2(755\times 10^{-15})(10-0.83)\times 10^{-9}-.83)\right]^2 + [(1.17\times 10^{10})(755\times 10^{-15})\times(12.82+44.7)]^2} = 7.7\times 10^{-3}\left[S^2\right]$$

The source impedance

$$|Z_S(\omega_0)| = \sqrt{R_S^2 + (\omega_0 L_S)^2} = \sqrt{(1.13)^2 + [(1.17\times 10^{10})(0.5\times 10^{-9})]^2} = 5.96 \text{ ohm}$$

From (4.129)

$$a_1 = |Z_S(\omega_0)|^2 \times |Y'_{mS}(\omega_0)|^2 = (5.96)^2 \times (7.7\times 10^{-3}) = 0.27$$

From (4.131)

$$|Y'_{mG}(\omega_0)|^2 = \frac{(44.7\times 10^{-3})^2}{[1+(44.7\times 10^{-3})\times 1.13]^2 + [(1.17\times 10^{10})(44.7\times 10^{-3})(0.5\times 10^{-9})]^2} = 1.7\times 10^{-3}\left[S^2\right]$$

From (4.132)

$$a_2 = (1.7\times 10^{-3})\left[(1.17\times 10^{10})\frac{1}{2.47\times 10^{11}}\sqrt{\frac{642\times 10^{-4}}{3.8\times 10^{-3}}}\right]^2 (12.82+44.7)^2 = 0.2$$

The width of M2 is chosen as the half of the width of M1. Therefore, $g_{m2} = g_{m1}/\sqrt{2}$. The load impedance of the circuit at f_0 is $R_L \cong L\omega_0 Q = 1170$ ohm, under the assumption that the output resistance of M2 and the parallel resistance of C are high and do not affect the quality factor of the output resonance circuit.

Now all these values can be inserted into (4.138) and the noise factor can be found:

$$F = 1 + \frac{330}{300}\frac{1}{(3.2\times10^{-3})44.7} \left\{\frac{1}{(1-0.47)}\left[(22.35\times10^{-3}) + (3.2\times10^{-3})\times13.65\right] + (15.8\times10^{-3}) + \frac{1}{1170}\right\}$$
$$= 1 + 1.1\times\frac{1}{143}\left\{\frac{1}{0.53}[22.3 + 43.68 + 15.8 + 0.85]\right\} = 2.08$$

which corresponds to $NF = 3.18$ dB.

Inspecting the components of this last expression, the contributions of the parameters on the noise factor can be easily observed and the improvements to reduce the noise factor can be developed.

It is also important to note the effect of the temperature difference. For example, if the temperatures of the chip and that of the antenna were equal, the noise figure should be 2.96 dB.

The total capacitance resonating the load inductor can be calculated as

$$C_T = \frac{1}{\omega_0^2 L} = \frac{1}{(1.17\times10^{10})^2\times(10^{-8})} = 0.73 \text{ pF}$$

which is the sum of the input capacitance of the following stage, the drain junction capacitance and the capacitor connected parallel to the inductor, which usually is a MOS varactor that can be used to fine-tune the circuit.

Finally, the voltage and the power gain of the amplifier can be calculated as follows.

The voltage gain from v_A to v_0 is

$$A_v = |Y_{mG}|\times R_L = (56.56\times10^{-3})\times1170 = 66.18$$

which corresponds to 36.4 dB.

As a result of the impedance matching, the signal power delivered from the antenna is

$$P_i = \frac{(v_A/2)^2}{R_A}$$

and the signal power delivered to the load resistance

$$P_o = \frac{v_o^2}{R_L}$$

Consequently,

$$A_p = \frac{P_o}{P_i} = \left(\frac{v_o}{v_A/2}\right)^2 \times \frac{R_A}{R_L} = 4 \times A_v^2 \times \frac{R_A}{R_L}$$
$$= 4 \times (66.18)^2 \times \frac{44.7}{1170} = 669.3$$

which corresponds to 28.25 dB.

4.6.5 The differential LNA

For some applications, designing the LNA as a differential amplifier is more convenient.

(a) If the antenna is balanced (differential), i.e. the signal voltage of either of the output nodes is not at ground potential, but they have equal magnitude and opposite phase with respect to the ground.[27]
(b) If the circuit following the LNA has a differential input.

Another advantage of a differential LNA is that its even harmonics are small (theoretically zero), owing to the symmetrical structure of the circuit.

The schematic of a typical differential LNA is given in Fig. 4.54(a). The DC tail current source, I_T, provides better symmetry of the drain DC currents of the input transistors and lower common-mode gain, in other words less sensitivity against the external disturbing signals.

The circuit is re-drawn in Fig. 4.54(b), which helps us to consider the circuit as two single-ended LNAs and use the expressions derived in Section 4.6.4. In Fig. 4.54(b) the external resonance capacitor is composed of two series connected capacitors (or varactors).

The inductance of the center-tapped inductor that is composed of two equal parts depends on the coupling of them. If the coupling of the two halves of the inductor is zero or negligibly small, the total inductance is equal to $L_T = 2L$. This case corresponds to two separate identical inductors that are not physically close to each other. If these inductors are closely coupled, for example designed as a center-tapped inductor as shown in Fig. 1.35(b), the total inductance becomes $L_T = 2L(1+k)$, where k is the coupling coefficient. It means that if the coupling is strong and k is close to its theoretical maximum value of 1, the total inductance increases considerably (approaches to $4L$).[28] This is an advantage from the point of view of the area and the effective series resistance of the inductor.

However, there is another effect that must not be overlooked. The noise component of the drain current of M12 is applied to the resonance circuit of the right half of the

[27] Another solution is to connect a balanced antenna to the input of a single-ended LNA via a balun, a simple structure that converts a balanced signal to an unbalanced signal, and vice versa, at the expense of extra parasitics and noise.
[28] Typical values of k for center-tapped inductors are in the range of 0.6 to 0.7.

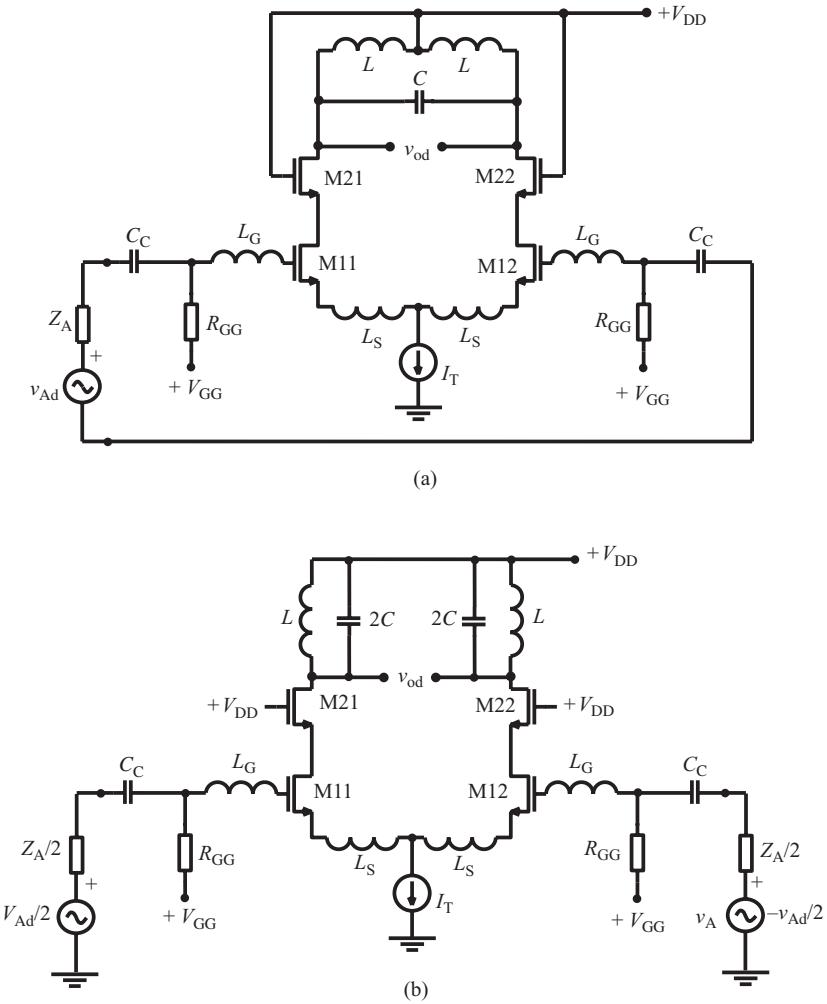


Figure 4.54 (a) Schematic of a typical differential LNA. (b) The re-arranged schematic that helps us to use the expressions derived for the single-ended LNA (coupling coefficient assumed $k \ll 1$).

circuit. If we show the root mean square value of this noise current at or in the vicinity of the resonance frequency as $\overline{i}_{nd2}$, the noise current flowing through L is equal to $Q \times \overline{i}_{nd2}$, which induces a noise current equal to $k \times Q \times \overline{i}_{nd2}$ on the inductor of the left half of the circuit. This noise current is not correlated to the self noise current of the left part of the circuit; therefore it increases the total noise of this part, and vice versa.

5 L-C oscillators

Oscillators, especially sinusoidal oscillators, are among the main components of all RF systems. They are being used as so-called "pilot" oscillators to generate the carrier frequencies of transmitters and as local oscillators in receivers to generate the signals that are necessary for frequency conversion purposes. The most important features of sinusoidal oscillators are the frequency of oscillation and its spectral purity, i.e., the frequency stability and the phase stability. The RF sinusoidal oscillators are almost exclusively based on the resonance effect.

In certain applications it is necessary to adjust the frequency of oscillation in a certain range, or to fine-tune the frequency to a pre-determined value. Since this feature is usually realized by a varactor in the resonance circuit of the oscillator, these circuits are called VCOs (voltage controlled oscillators). In some other applications, the frequency of oscillation must be fixed at a certain frequency with the maximum possible precision and stability. For these cases, quartz crystal oscillators are preferred. If the target frequency is higher than the range supported by the crystal oscillator, a higher frequency VCO can be "locked" to the frequency of the crystal oscillator.

As already explained in Chapter 4, an L-C circuit with external excitation will swing at a frequency determined by the values of its components. The magnitude of the swing decreases in time, owing to the losses of the system. If the losses of the system are compensated in some way, the magnitude of the swing remains constant, in other words, the system "oscillates". To compensate the losses of a resonance circuit, and to understand this compensation mechanism, we will investigate two well-known methods: the negative resistance approach and the feedback approach.

5.1 The negative resistance approach to L-C oscillators

Consider the series resonance circuit in Fig. 5.1. As was shown in Section 4.1.2, the natural frequency of the circuit is

$$\omega_0 = \sqrt{\frac{1}{LC}}$$

and the effective series resistance is r_{eff}, which represents the total losses of L and C. At ω_0 the reactances of L and C cancel out each other and the input impedance becomes equal to its real part, i.e. r_{eff}. If we connect a "negative resistance" in series,

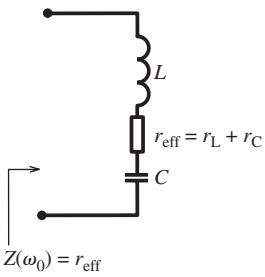


Figure 5.1 The input impedance of a series resonance circuit at resonance.

equal in magnitude to r_{eff}, then the total resistance of the circuit becomes zero and the circuit is ready to oscillate under *any* excitation – for example, noise, that is always present.

A "negative resistance" does not exist as a physical reality but some devices or circuits exhibit a negative resistance behavior in a limited range of their current–voltage characteristics. This means that the slope of the current–voltage characteristic is negative, in a certain operating range.

There are a number of two-terminal negative resistance devices: for example, the "tunnel diode" which is also known as the "Esaki diode". The tunnel diode is a semiconductor p-n junction, where both regions are doped up to the degenerate level, i.e. the Fermi levels of the p and n type regions are shifted into the valence band and the conduction band, respectively. Owing to the "quantum-mechanical tunneling" of electrons in a range of the bias voltage, the current–voltage characteristic exhibits a negative slope, as shown in Fig. 5.2(a).[1]

If a tunnel diode is biased at a point in the middle of the negative resistance region, the corresponding small-signal resistance becomes negative and equal in magnitude to the inverse of the slope of the curve at this point. The value of this resistance is usually small, in the range of tens of ohms, and therefore it is suitable to compensate the effective series resistance of a series resonance circuit. Since the parasitics can be made very small, tunnel diodes can be used at higher frequencies, up to several tens of GHz.[2]

The schematic of a tunnel diode oscillator is shown in Fig. 5.2(b). As already mentioned, oscillation starts if the negative resistance corresponding to the slope of the characteristic curve at the DC operating point Q is equal to the effective resistance of the resonance circuit. But there is a risk of the oscillation ceasing owing to the variations of the parameters that may decrease the value of the negative resistance. To guarantee a sustained oscillation, the magnitude of the negative resistance must be higher than the effective resistance of the resonance circuit. In this case, the amplitude of oscillation tends to increase steadily. But since the range of the negative resistance

[1] For a detailed explanation of the tunnel diode, refer to [45].

[2] Tunnel diodes are fabricated and packaged as discrete devices. But there are serious efforts to integrate tunnel diodes into silicon ICs (see the publications of A. Seabaugh).

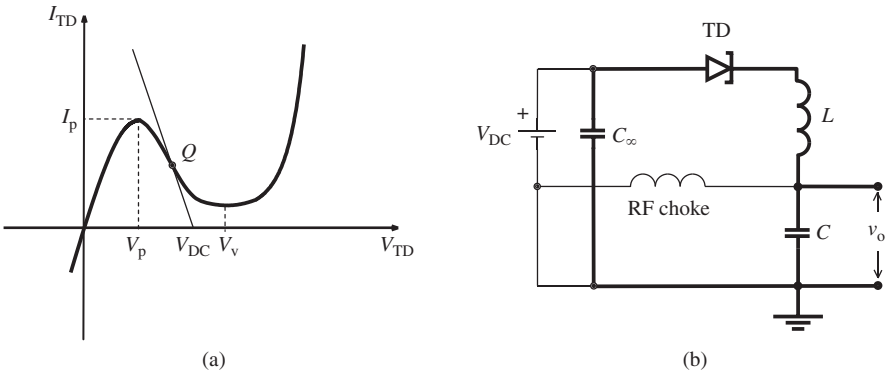


Figure 5.2 (a) Typical current–voltage characteristic of a tunnel diode. The peak and valley voltages, V_p and V_v, depend on the semiconductor and are in the range of hundreds of millivolts. (b) Schematic of a tunnel diode oscillator. The bias circuit is drawn with fine lines. The series resistances of L and the RF choking coil determine the DC load line shown in (a).

is limited, the amplitude can increase up to the point where the magnitude of the negative resistance drops to r_{eff} and stabilizes itself at the magnitude that corresponds to this operating point.

There are a number of transistor circuits that exhibit a negative resistance port. One of them is the input port of a source follower under appropriate operating and loading conditions, which was investigated in Section 3.2. and used for Q enhancement in Example 4.2. Another and most frequently used one is the cross-coupled differential negative resistance circuit. Since the magnitude of the negative resistance of this circuit is relatively high, it is convenient to use it to compensate the relatively high effective parallel resistance of a parallel resonance circuit at resonance frequency.

The general topology of a parallel resonance circuit is shown in Fig. 5.3(a). The frequency at which the input impedance is real is denoted by $\bar{\omega}_0$. The input impedance of the circuit at $\bar{\omega}_0$ can be represented in terms of its parallel components, as shown in Fig. 5.3(b), where

$$L' = L + \frac{R_L^2}{\bar{\omega}_0^2 L} \tag{5.1}$$

$$C' = \frac{C}{1 + \bar{\omega}_0^2 C^2 R_C^2} \tag{5.2}$$

$$\bar{\omega}_0 = \sqrt{\frac{1}{L'C'}} = \frac{1}{LC} \times \sqrt{\frac{L - CR_L^2}{L - CR_C^2}} \tag{5.3}$$

$$R_{\text{eff}} \cong (Q_L^2 + 1)\frac{R_L^2}{R_L + R_C} \tag{5.4}$$

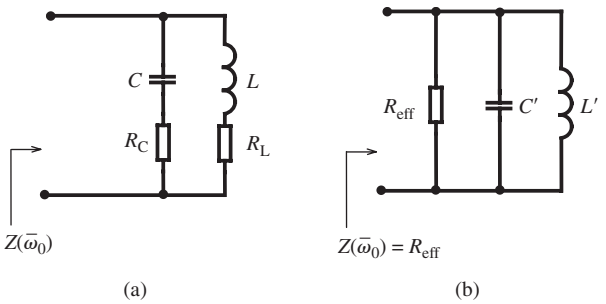


Figure 5.3 (a) A typical on-chip parallel resonance circuit with lossy inductance and capacitance. (b) The same circuit in terms of its parallel components at $\bar{\omega}_0$.

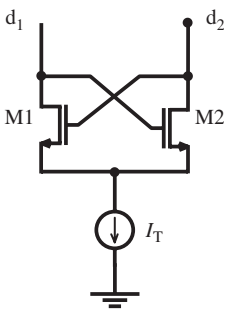


Figure 5.4 The cross-connected differential negative resistance circuit.

If a negative resistance equal in magnitude to R_{eff} is connected in parallel to this circuit, it cancels out R_{eff} and the remaining parallel resonance circuit oscillates at $\bar{\omega}_0$, which is the oscillation frequency of the circuit and will be shown by ω_{osc}.

The schematic diagram of a cross-connected negative resistance circuit is given in Fig. 5.4. It can be easily seen that the circuit has two robust stable operating points: M1 is off, M2 is on and the whole I_T tail current flows through T2, and vice versa. In addition, there is another operating point that is practically unstable, which corresponds to $I_{D1} = I_{D2} = (I_T/2)$. Any small (even infinitesimal) change in this condition impairs the stability of the operating points and the circuit switches to one of its robust operating points.

For this critical operating point, the small-signal admittance seen from the d_1–d_2 output port can be calculated from the equivalent circuit shown in Fig. 5.5(a) as

$$y_o = -\frac{1}{2}(g_m - g_{ds}) + \frac{1}{2}sC_{gs} \tag{5.5}$$

which corresponds to a negative resistance of

$$r_o = -\frac{2}{(g_m - g_{ds})} \cong -\frac{2}{g_m} \tag{5.6}$$

parallel to a capacitance equal to $C_o = C_{gs}/2$.

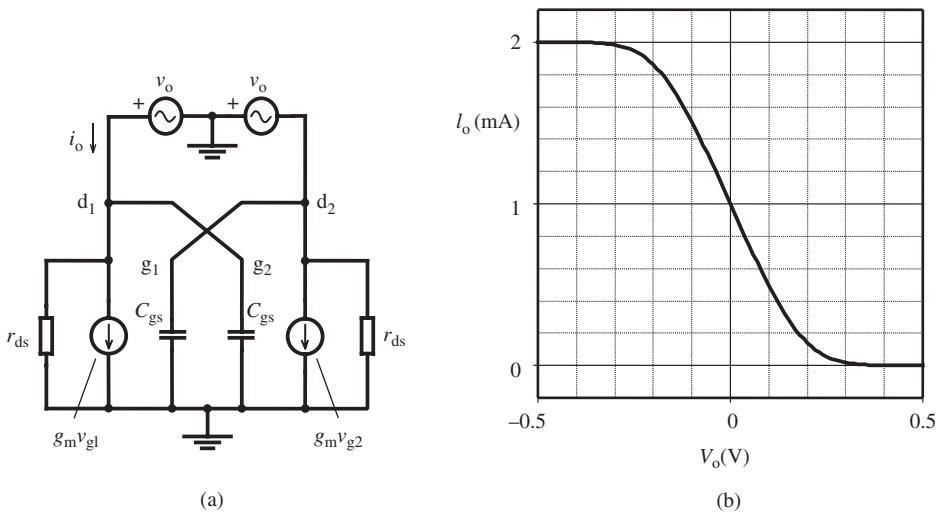


Figure 5.5 (a) The small-signal equivalent to calculate the output admittance of a cross-connected differential negative resistance circuit. (b) PSpice simulation result of the output current–voltage characteristic (AMS 0.35, 200 μ/0.35 μ, $I_T = 2$ mA).

The g_m in (5.6) depends on technology, dimensions of the transistors, the DC drain currents and also if the transistor is operating in the velocity saturated region or not. For velocity saturation, the transconductance is considerably smaller and cannot be controlled by the tail current. Note that expression (1.33) must be used for non-velocity saturated operation and (1.34) for velocity saturated operation to calculate the transconductance.

This negative resistance calculated from the linearized small-signal equivalent circuit is valid only in the middle of the operating range that corresponds to $I_{D1} = I_{D2}$, and increases to infinity at both ends of the range. To observe this nonlinear behavior of the output resistance, a simulation result is given in Fig. 5.5(b).

It is seen from this curve that the negative resistance is minimum in the mid-point that corresponds to the value calculated from the small-signal equivalent circuit, and increases symmetrically on both sides. For a parallel resonance circuit whose effective parallel resistance is higher than the negative resistance corresponding to the mid-point of the curve, the circuit oscillates. As explained for the tunnel diode oscillator, the nonlinearity of the curve limits the amplitude of the oscillation at a certain level that depends on g_m and can be controlled by I_T. The average negative resistance corresponding to a secure oscillation can be considerably higher than the mid-point resistance.

As an example, the negative resistance at the mid-point is 200 ohms and the resistance corresponding to the onset of limitation is about 300 ohms in Fig. 5.5(b). Therefore, to obtain a sustained oscillation with a safety margin, the resonance impedance of the parallel resonance circuit connected parallel to the d_1-d_2 port must not be smaller than 300 ohms.

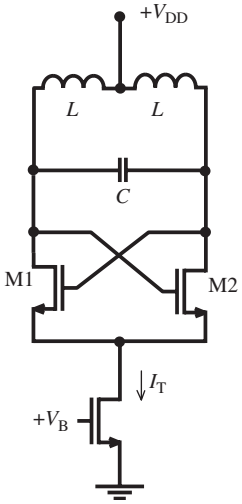


Figure 5.6 The differential negative resistance oscillator.

Another aspect (and advantage) of this circuit is that owing to the symmetry of the curve with respect to the mid-point, the even harmonics in the spectrum of the drain currents – and, consequently, on the differential output voltage – are zero.

The circuit schematic of an oscillator that uses a cross-connected negative resistance circuit is given in Fig. 5.6, which is also known as a differential oscillator. As explained for the differential LNAs, the two halves of the inductor can be coupled or not. The total equivalent value of the inductance is equal to

$$L_T = 2L(1 + k),$$

where k is the coupling coefficient and its value is approximately 0.7 for a center-tapped inductor as shown in Fig. 1.35, and can be considered as zero for two separate inductors on the chip.

Example 5.1 A differential oscillator will be designed for $f_0 = 1.6$ GHz. The inductor to be used will be selected from a library in which there are 1 nH, 2 nH, 5 nH and 10 nH inductors with $Q_{eff} = 7$. The transistors have characteristics similar to AMS 0.35 micron technology. It is assumed that a 1 V bias voltage is available on the chip. The allowed DC supply current is 2 mA.

The schematic of the circuit together with the tail current source is given in Fig. 5.6. Since the output voltage swings around V_{DD}, the supply voltage must be chosen lower than the allowed maximum value by the expected amplitude of the oscillation with a safety margin, and must be checked later on. We will use $V_{DD} = 2.5$ V, which corresponds to 1 V maximum amplitude, including the safety margin.

To obtain the oscillation with a relatively high negative resistance that corresponds to a small transconductance (i.e., small supply current and small transistor width), the

resonance impedance must be as high as possible. Since the resonance impedance of the resonance circuit is $R_p = L\omega_0 \times Q_{eff}$, the first (seemingly obvious) choice would have been to use the maximum inductance value, that is $10 + 10 = 20$ nH in this case. But this corresponds to a resonance capacitance that is equal to

$$C = \frac{1}{\omega_0^2 L} = \frac{1}{(2\pi \times 1.6 \times 10^9)^2 \times (20 \times 10^{-9})} = 495 \text{ fF}$$

Note that this capacitance represents the sum of $(C_{gs}/2)$, $(C_{db}/2)$, the resonance capacitor (or varactor), and the input capacitance of the following circuit which, on its own, can be in the range of several hundreds of femtofarads. Consequently, using the maximum inductance value would impose a very limited value for the total capacitance that may not be reasonable for practical applications. In our case, it is preferable to work with a higher resonance capacitance that dictates the use of a $5 + 5 = 10$ nH inductor, resonating at 1.6 GHz with 905 fF.

The effective parallel resistance of this resonance circuit at 1.6 GHz is

$$R_{eff} = L\omega_0 Q_{eff} = 10^{-8} \times (1.05 \times 10^{10}) \times 7 = 735 \text{ ohm}$$

Since the circuit operates with a considerable safety margin, we can neglect the output resistance of the transistor and the losses of the resonance capacitance. The value of the negative resistance must be smaller than this theoretical value to guarantee sustained oscillation, so that the amplitude reaches the saturation end of the negative resistance curve. Thus, we take $r_o = -350$ ohms, which corresponds to $g_m = 5.7$ mS according to (5.6).

The aspect ratio to obtain this transconductance value with 1 mA drain DC current can be calculated from (1.33):

$$\frac{W}{L} = \frac{g_m^2}{2\mu C_{ox} I_D} = \frac{(5.7 \times 10^{-3})^2}{2 \times 325 \times (4.54 \times 10^{-7}) \times 10^{-3}} = 110$$

Therefore the gate width must be 38 μm.

The gate width of the tail current source can be sized to conduct 2 mA under 1 V gate bias voltage, which is found as 62 μm for 0.35 μm channel length.

The PSpice transient simulation file is given below:

```
*DIFFERENTIAL NEGATIVE RESISTANCE OSC.*
VDD 100 0 2.5
.LIB "CMOS7TM.MOD"
M1 1 2 3 0 MODN w=38u l=.35u ad=13.3e-8 as=13.3e-8 pd=40u
ps=40u nrd=.01 nrs=.01
M2 2 1 3 0 MODN w=38u l=.35u ad=13.3e-8 as=13.3e-8 pd=40u
ps=40u nrd=.01 nrs=.01
MT 3 4 0 0 MODN w=62u l=.35u ad=21.7e-8 as=21.7e-8 pd=70u
ps=70u nrd=.01 nrs=.01
VB 4 0 1
L1 1 11 5N
RL1 11 100 7.5
```

```
L2 2 22 5N
RL2 22 100 7.5
C 1 2 .9p
.IC v(1)=2.6v
.TRAN .03N 51N 50N .01N
.probe
.end
```

The simulation results with calculated values are shown in Fig. 5.7. It must be noted that the amplitudes of the voltage swing on the drain nodes are one half of the differential output voltage v_o. This means that the maximum drain voltages are well below the permitted maximum voltage, 3.5 V.

For the transient simulation of oscillators, a number of factors have to be taken into account that also relate to the specific limitations of circuit simulator algorithms. Many circuit simulators are notorious for difficulties that may be encountered in simulating oscillatory circuit behavior, in the time domain. In particular:

- To start the oscillation it is necessary to impose a suitable "initial condition" to the circuit. For example, the initial voltage of d_1 was given as 2.6 V for this circuit.
- To make sure that the oscillation is sustained, it is necessary to observe the output signal after a reasonable time from the beginning of the transient simulation.
- The distortion (flattening of the peaks) on the currents of M1 and M2 indicates that the circuit is operating in the entire range of the negative resistance region. It is wise to adjust the circuit parameters such that the peaks are slightly flattened. This means that the circuit has a safety margin against parameter changes and at the same time, the distortion component reflected to the output voltage waveform is not excessive.

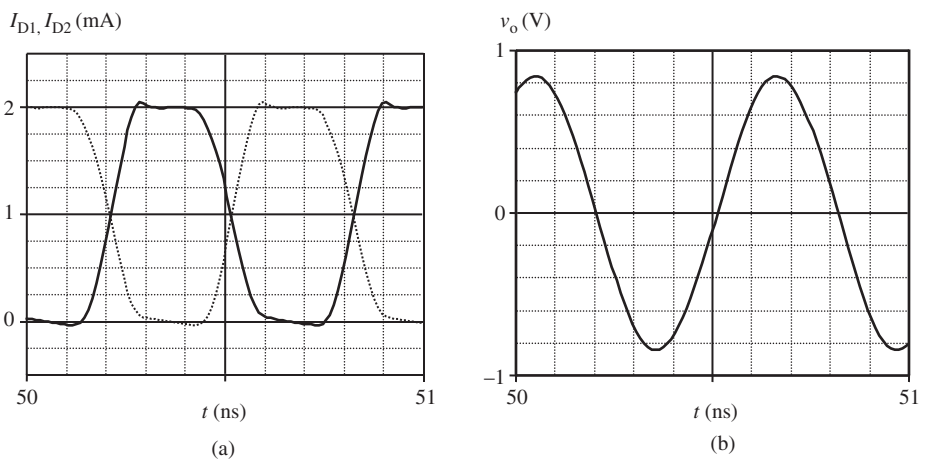


Figure 5.7 (a) The drain current waveforms of the drain currents. (b) The differential output voltage waveform.

5.2 The feedback approach to L-C oscillators

The classical approach to study oscillator circuits is based on the oscillation criterion stated by H. Georg Barkhausen in 1921 for feedback amplifiers. The general topology of a single loop linear feedback amplifier is shown in Fig. 5.8. Here, the gain block is usually an amplifier with a gain function A that can be a voltage gain, a current gain, a transadmittance or a transimpedance. The feedback block β is usually (but not necessarily) a passive two-port. The signals a_1, a_1', a_f and a_2 can be either voltage or current, but a_1, a_1' and a_f obviously must have the same dimension. The basic relations among the signals are

$$A = \frac{a_2}{a_1'} \quad \beta = \frac{a_f}{a_2} \quad a_1' = a_1 + a_f \text{ and } A_f = \frac{a_2}{a_1} \tag{5.7}$$

where A is the gain of the amplifier, β is the transfer function of the feedback block and A_f is the gain of the circuit as a whole, which is called the "gain with feedback" or the "gain of the feedback amplifier".

From the basic relations given in (5.7), the gain of the feedback amplifier can be found to be

$$A_f = \frac{A}{(1 - \beta A)} \tag{5.8}$$

where βA is called the "loop gain".

From this expression it can be seen that:

- If the magnitude of $(1 - \beta A)$ is larger than unity, the overall gain of the feedback amplifier is smaller than A, in other words, the gain is reduced by feedback. This case is also called "negative feedback".
- If the magnitude of $(1 - \beta A)$ is smaller than unity, the overall gain of the feedback amplifier is larger than A, in other words, the gain is increased by feedback. This case is called "positive feedback".
- As a special case of positive feedback, if

$$(1 - \beta A) = 0 \text{ or, equally, } \beta A = 1 \tag{5.9}$$

  then the gain of the feedback amplifier goes to infinity.

Since the gain of the amplifier block and the transfer function of the feedback block are frequency dependent, in other words, since A and β are complex quantities in the frequency domain, (5.9) must be written as

$$\begin{gathered} \beta A = 1 + \text{j}.0 \text{ or} \\ \text{Re}\{\beta A\} = 1 \text{ and } \text{Im}\{\beta A\} = 0 \text{ or} \\ |\beta A| = 1 \text{ and } \varphi(\beta A) = 0 \end{gathered} \tag{5.10}$$

According to the Barkhausen criterion, the meaning of infinite gain is that the circuit oscillates at the frequency where (5.10) is satisfied. The amplitude of the oscillation

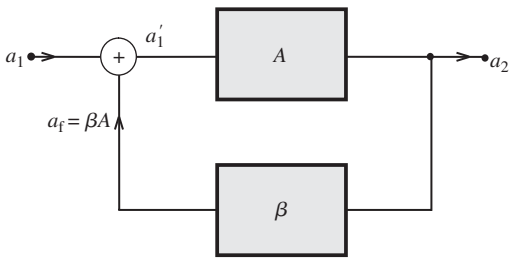


Figure 5.8 The block diagram of a feedback amplifier.

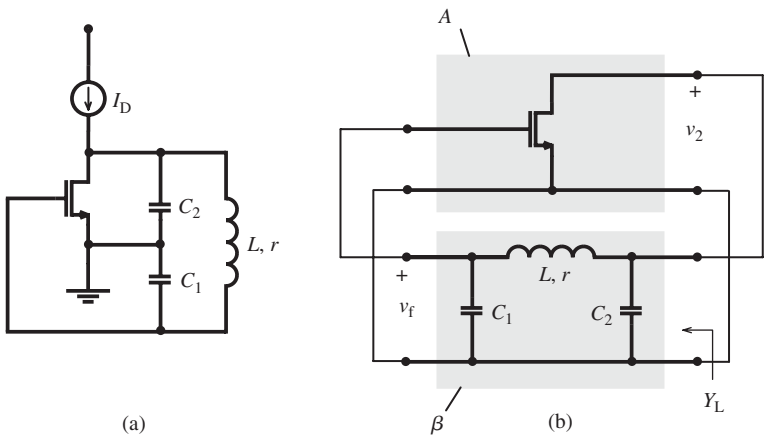


Figure 5.9 (a) The simplified schematic of a MOS Colpitts oscillator. (b) The circuit re-drawn with the amplifier block and the feedback block.

certainly is limited in a certain range by the amplifier, where the assumption of linearity is – at least approximately – valid.

As an example to this approach, the MOS transistor version of one of the classical oscillator circuits will be examined in the following: the Colpitts oscillator, which was first demonstrated in 1919 – certainly with vacuum tubes – by Edwin H. Colpitts. The simplified schematic of the circuit and its block diagram representation are shown in Fig. 5.9(a) and (b), respectively. The input capacitance of the transistor is included in C_1. The load of the amplifier, the admittance shown from the port that C_2 is connected to is shown by Y_L, and can be calculated as

$$\begin{aligned} Y_L &= sC_2 + \frac{1}{sL + r + \frac{1}{sC_1}} \\ &= \frac{s^3 LC_1C_2 + s^2 C_1C_2 r + s(C_1 + C_2)}{s^2 LC_1 + sC_1 r + 1} \end{aligned} \tag{5.11}$$

Then the voltage gain of the transistor is

$$A_v = -g_m \frac{1}{(g_{ds} + Y_L)}$$

and in open form

$$A_v = -g_m \frac{s^2 LC_1 + sC_1 r + 1}{s^3 LC_1 C_2 + s^2 (C_1 C_2 r + LC_1 g_{ds}) + s[(C_1 + C_2) + rC_1 g_{ds}] + g_{ds}} \tag{5.12}$$

The voltage transfer ratio of the feedback block can be calculated as

$$\beta = \frac{1}{s^2 LC_1 + sC_1 r + 1} \tag{5.13}$$

Hence the loop gain βA_v becomes

$$\beta A_v = -g_m \frac{1}{s^3 LC_1 C_2 + s^2 (C_1 C_2 r + LC_1 g_{ds}) + s[(C_1 + C_2) + rC_1 g_{ds}] + g_{ds}}$$

If we convert this expression into the ω domain and apply the Barkhausen criterion, stating that the loop gain must be real and equal to unity, from $\mathrm{Im}\{\beta A_v\} = 0$ the oscillation frequency that satisfies this condition can be calculated as

$$\omega_{osc} = \omega_0 \sqrt{1 + r \cdot g_{ds} \frac{C_2}{C_1 + C_2}} \tag{5.14}$$

Note that the oscillation frequency is not equal to the natural frequency of the resonance frequency but slightly different, which will be discussed later.

Similarly, the oscillation condition that gives the minimum value of the transconductance can be found from $\mathrm{Re}\{\beta A_v\} = +1$ or $|\beta A_v| = 1$ as

$$g_m \cong \frac{1}{r_{eff}} \frac{C_1 + C_2}{C_2} \tag{5.15a}$$

It must be noted that to guarantee sustained oscillations, g_m must be higher than this value to provide a safety margin. It is convenient to modify (5.15a) as

$$g_m = k_s \frac{1}{r_{eff}} \frac{C_1 + C_2}{C_2} \tag{5.15b}$$

where k_s is a safety factor, usually 1% to 10% larger than unity, i.e. between 1.01 and 1.1.

Another interesting example is to calculate the oscillation condition of a cross-connected oscillator that was given in Fig. 5.6, with the feedback approach. The circuit is re-drawn in Fig. 5.10(a) where the resonance circuit is divided into two identical parts as in Fig. 4.54. In Fig. 5.10(b) the circuit is drawn as a feedback amplifier. The total voltage gain of the two cascaded stages is positive and equal to the multiplication of the gains of the identical stages:

$$A_{vT}(\omega_0) = (-g_m \cdot R'_{eff})^2$$

where R'_{eff} is the effective parallel resistance of the loads at resonance and equal to

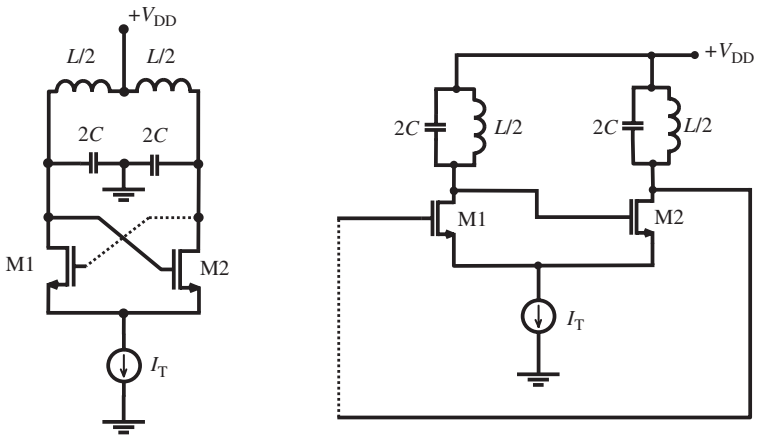


Figure 5.10 (a) The cross-connected oscillator. (b) The same circuit drawn as a feedback oscillator.

$$R'_{\text{eff}} = \left(\frac{L}{2}\right)\omega_0 \times Q_{\text{eff}}$$

Therefore the total gain becomes

$$A_{\text{vT}}(\omega_0) = g_{\text{m}}^2 \frac{1}{4}(L\omega_0 \times Q_{\text{eff}})^2 = g_{\text{m}}^2 \frac{1}{4} R_{\text{eff}}^2$$

where R_{eff} is the effective parallel resistance of the circuit composed of L and C.

Since the output of the amplifier is directly connected to its input, β is equal to unity. Then the loop gain is

$$\beta A_{\text{vT}}(\omega_0) = g_{\text{m}}^2 \frac{1}{4} R_{\text{eff}}^2$$

If we apply the Barkhausen criterion and write the loop gain equal to unity, we obtain

$$g_{\text{m}} = \frac{2}{R_{\text{eff}}}$$

which is equivalent to the expression (5.6) obtained with the "negative resistance approach".[3]

Problem 5.1 *The schematic of a common-drain Colpitts oscillator is shown in Fig. 5.11.*

(*a*) *Derive expressions to calculate the oscillation frequency and the minimum value of the transconductance for sustained oscillations.*

[3] This is an interesting example to see how different approaches converge for a certain physical reality.

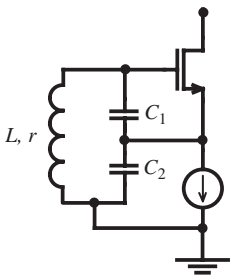


Figure 5.11 The common-drain version of the Colpitts oscillator.

(*b*). *Compare and discuss the circuit with the negative resistance circuit used for Q-enhancement given in Fig. 4.7.*

5.3 Frequency stability of L-C oscillators

The oscillation frequency of an L-C oscillator is mainly determined by the values of the inductor and the capacitor of the resonance circuit, but also influenced by some secondary factors, namely the losses of the resonance circuit and certain parameters of the transistor. For example, the oscillation frequency of a MOS Colpitts oscillator was given as (5.14). If we combine this expression with (5.15b), it is possible to arrange it as follows:

$$\omega_{\text{osc}} = \frac{1}{\sqrt{LC}}\sqrt{1 + k_{\text{s}}\frac{g_{\text{ds}}}{g_{\text{m}}}\frac{1}{Q^2}\frac{C_1}{C_2}} \tag{5.16}$$

where C is the series equivalent of C_1 and C_2, and Q is the quality factor of the resonance circuit.

This frequency is a function of a number of parameters and can vary depending on the tolerances or the variations of these parameters owing to some effects, for example temperature and supply voltage fluctuations. The influences of L and C tolerances are well defined and can be compensated with the trimming of frequency, for example with the aid of a varactor used as part of the resonance capacitance. The variations of other parameters are usually unpredictable and consequently more severe.

The effects of the small changes of parameters on the oscillation frequency can be expressed with an exact differential. For a Colpitts oscillator, for example, the change of the oscillation frequency owing to the small variations of the parameters is

$$\mathrm{d}\omega_{\text{osc}} = \frac{\partial\omega_{\text{osc}}}{\partial L}\mathrm{d}L + \frac{\partial\omega_{\text{osc}}}{\partial C_1}\mathrm{d}C_1 + \frac{\partial\omega_{\text{osc}}}{\partial C_2}\mathrm{d}C_2 + \frac{\partial\omega_{\text{osc}}}{\partial g_{\text{m}}}\mathrm{d}g_{\text{m}} + \frac{\partial\omega_{\text{osc}}}{\partial g_{\text{ds}}}\mathrm{d}g_{\text{ds}} + \frac{\partial\omega_{\text{osc}}}{\partial Q}\mathrm{d}Q$$

The change of the value of an on-chip inductance is mainly related to the dimensional changes owing to temperature variations. The resistance of the inductor that affects the

quality factor is also subject to change, not only owing to the dimensional changes but also to the temperature coefficient of the material (aluminum or copper).

The stability of an on-chip capacitor depends on its construction. The values of the MIM and poly1–poly2 capacitors are subject to variations owing to thermal expansion. The capacitances of the varactor capacitors are sensitive to the bias voltage variations.

The small-signal output conductance and the transconductance of the transistor depend on the voltage and the current of the operating point.

As a conclusion, it can be stated that for the stability of the oscillation frequency of a Colpitts oscillator, the variation of the operating point of the transistor and the bias voltages of varactors must be well stabilized, and the variations of the temperature must be small. It can be shown that these conclusions are equally valid for other types of feedback oscillators.

Equation (5.16) also shows that if the effective quality factor of the resonance circuit is high, the oscillation frequency approaches the natural frequency of the resonance circuits that depends only on L and C, and the overall sensitivity of the oscillation frequency decreases.[4]

If we solve C_1/C_2 from (5.15) and insert to (5.16), we obtain

$$\omega_{\text{osc}} = \frac{1}{\sqrt{LC}}\sqrt{1+\frac{g_{\text{ds}}}{g_{\text{m}}}\frac{1}{Q^2}\left(g_{\text{m}}R_{\text{eff}}-k_{\text{s}}\right)} \tag{5.17}$$

According to this expression, if the condition $g_{\text{m}}R_{\text{eff}} = k_{\text{s}}$ is fulfilled, the oscillation frequency becomes equal to the natural frequency of the resonance circuit, which corresponds to the maximum available frequency stability.

As another example let us consider (5.3) which is derived for a negative resistance oscillator. From this expression it can be seen that if the series resistance of the capacitance branch of the parallel resonance circuit is equal to that of the inductance branch, the oscillation frequency becomes equal to the natural frequency of the resonance circuit.[5] Since the oscillation frequency depends only on L and C in this case, for which the losses of the two branches of the resonance circuit are equal, the frequency stability improves,[6] despite the decrease of the effective quality factor of the resonance circuit.

From these examples we understand that there are some measures to improve the stability of the oscillation frequency of integrated oscillators to some extent, but not sufficiently for applications that need very high frequency stability.

The classical method to obtain oscillations with very high frequency stability is to use a quartz crystal that electrically behaves as a very high-Q resonance circuit.

[4] This fact is applicable to the oscillator circuits constructed with discrete high-Q inductors and high-quality capacitors, and extensively used in the past.

[5] This phenomenon was investigated in 1934 by J. Groszkowski from a different and more basic point of view [46].

[6] This behavior was underlined in Section 4.1.1. Since the natural frequency, the frequency where the imaginary part of the impedance is zero and the frequency corresponding to the maximum of the magnitude of the impedance coincide, it is useful also for feedback oscillators.

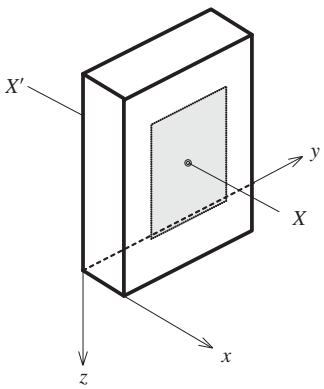


Figure 5.12 A quartz crystal cut, lapped and polished to the dimension corresponding to the targeted frequency, and metal electrodes plated on the two *y*–*z* faces for electrical connections.

5.3.1 Crystal oscillators

Figure 5.12 schematically shows a thin rectangular prism cut and lapped from a native quartz crystal. The *x, y* and *z* axes are called the electrical, mechanical and optical axes, respectively. Thin metal films are evaporated to make electrical connections on the faces parallel to the *x*–*y* plane. If a DC voltage is applied to the *X*–*X*′ port, the crystal expands or shrinks in the *y* direction, according to the polarity of the voltage. Correspondingly, if a pressure or tension is applied in the *y* direction, a voltage occurs between *X* and *X*′. This effect is called the "piezo-electric effect". It can be understood that under an alternating voltage the crystal starts to vibrate. If the frequency of the signal becomes equal to the mechanical resonance frequency of the crystal[7] the amplitude of the vibration becomes maximum, in other words, the crystal is in mechanical resonance. The losses of this resonance system are the internal friction losses of the crystal, the load corresponding to the connections and the friction of the air. With appropriate precautions (for example operating in vacuum), the quality factor of the system can be increased to the order of hundreds of thousands.

This electro-mechanical system behaves as an electrical circuit between its *X*–*X*′ connection ports, which can be represented with an equivalent circuit as shown in Fig. 5.13(a). In this figure the series branch represents the mechanical resonance. The electrical components can be derived from the value of the series resonance frequency (ω_S) and the value of the quality factor.[8] This branch becomes inductive above ω_S and resonates in parallel with C_p at ω_P, where $\omega_P > \omega_S$. The parallel resonance can be

[7] A solid prism has at least three resonance frequencies depending on the dimensions, the Young modulus and the density of the material. The resonance frequency mentioned above is the resonance frequency in the *y* direction. In addition to these fundamental resonance frequencies, the crystal can resonate at the harmonics (the overtones) of them.

[8] The inductance values are extremely high; usually in the range of 1 to 10 H, correspondingly *C* is very small. The resistance ranges from 100 ohm to several thousands of ohms. Parallel capacitance is in the range of picofarads.

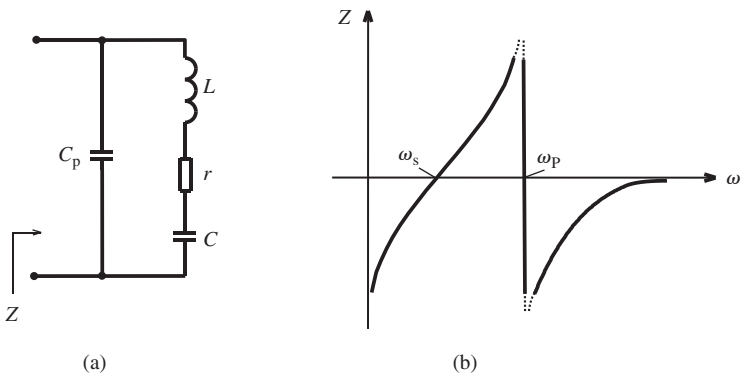


Figure 5.13 (a) The electrical equivalent of a quartz crystal. (b) The variation of the impedance. Note that the impedance axis is not to scale; in reality ω_S and ω_P are very close to each other.

adjusted – to some extent – with an additional parallel capacitance that increases the value of C_p.

Therefore, a quartz crystal can be used as a resonator in series resonance or in parallel resonance. Since the series resonance is mechanically based and depends only on the dimensions, it is very robust and can shift only owing to temperature variations. Cutting the prism with a specific angle that minimizes the temperature coefficient, and stabilizing the temperature of the crystal considerably help to improve the frequency stability. Hence the temperature coefficient can be well below 1 ppm/°C.

To build a crystal controlled oscillator, one way is to use a suitable negative resistance circuit to compensate for the losses of the crystal. Another way is to connect the crystal with an amplifier to provide a positive feedback at the resonance frequency of the crystal.

For a negative resistance crystal oscillator, the necessary transconductance values and the DC power consumption are considerably small since the losses of the crystal are very small compared to a conventional resonance circuit. Figure 5.14 shows an oscillator circuit based on the negative resistance circuit given in Fig. 4.7, with its source follower output buffer. This circuit can be also interpreted as a version of the Colpitts oscillator, where C_1 is totally or partially the input capacitance of M1.

Since the dimensions of crystals decrease with frequency, fabrication and precision problems increase. Usually for frequencies higher than 10 MHz it is useful to use the overtone resonances of crystals. The crystals originally intended to be used at a certain overtone are mounted into the case in such a manner to ease the vibration at this frequency.

The crystal oscillator circuit shown in Fig. 5.15 is a feedback oscillator. The output of M1 is connected to the input of M2 via a crystal. At the series resonance frequency of the crystal, its impedance is a resistance equal to r, which forms a positive feedback path together with R_1, to oscillate the circuit at the series resonance frequency of the crystal.

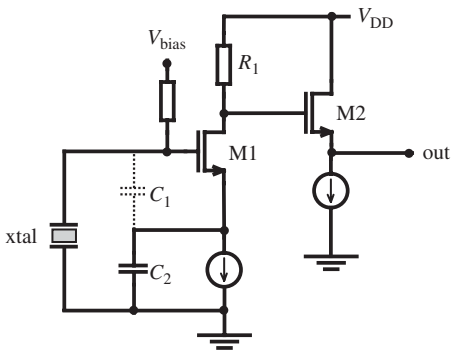


Figure 5.14 A negative resistance crystal controlled oscillator.

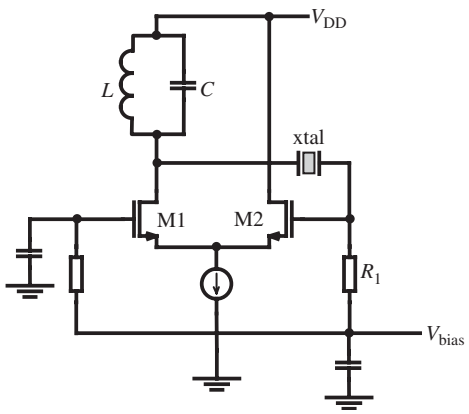


Figure 5.15 A crystal feedback oscillator that oscillates at the series resonance frequency of the crystal. This circuit is also suitable as an overtone oscillator, when the L-C circuit is tuned to a certain overtone of the crystal.

Another feature of this circuit is that it is suitable to make an overtone oscillator. If the resonance circuit is tuned to an overtone of the crystal, the gain becomes maximum at this frequency and the Barkhausen criterion can be fulfilled together with the crystal, to oscillate the circuit at this overtone frequency.

5.3.2 The phase-lock technique

We have seen that to obtain high-stability oscillations, crystal oscillators can serve at up to about 100 MHz. For higher frequencies, it is possible to control the frequency of a voltage controlled oscillator (VCO) with a high-stability crystal oscillator, in other words, to lock the phase of the VCO to the phase of the crystal oscillator.

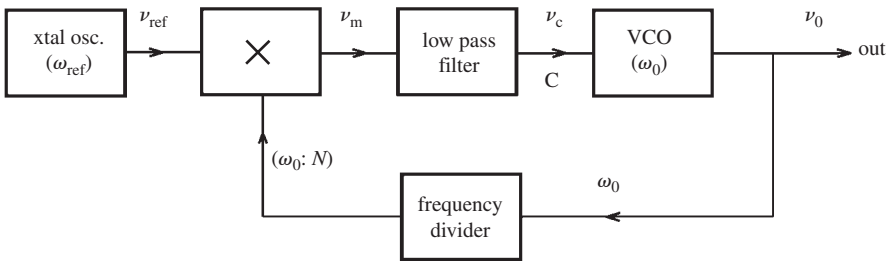


Figure 5.16 Block diagram of a phase-locked loop. Frequencies on the figure correspond to the locked operation.

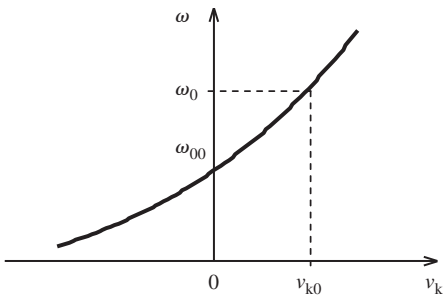


Figure 5.17 The presumed behavior of the VCO as a function of the frequency-control voltage.

The block diagram of a phase-locked loop (PLL) is shown in Fig. 5.16. The high-frequency output signal v_0, whose frequency is targeted as ω_0, will be generated by a VCO, which can be any type of L-C oscillator with a varactor as its tuning element. Assume that the frequency-control voltage characteristic of the VCO is as shown in Fig. 5.17. The target frequency is ω_0 and the frequency corresponding to $v_c = 0$ is ω_{00}, which is called as the "free-running frequency". A stable reference signal, v_{ref}, is being generated by a crystal oscillator, whose frequency must be a fraction of ω_0, such that $\omega_{ref} = \omega_0/N$, where N is an integer. The output frequency of the VCO is divided by N with a digital frequency divider and then applied to the input of an analog multiplier,[9] together with the output signal of the reference oscillator. The low-pass filter in the loop is usually a simple R-C filter.

To understand the behavior of the circuit, assume that in the beginning the loop is open at the C connection and the internal input bias voltage of the VCO is zero. For this case, if we show the output voltages of the reference oscillator, the VCO and the frequency divider as

[9] The analog multiplier acts as a phase comparator here. Therefore any other type of a phase comparator, for example an exclusive-OR circuit, can be used instead.

$$v_{\text{ref}} = V_{\text{ref}} \sin(\omega_{\text{ref}} t)$$
$$v_0 = V_0 \sin(\omega_{00} t + \varphi_0)$$
$$v_{\text{x}} = V_{\text{x}} \sin\left(\frac{\omega_{00}}{N} + \varphi_{\text{x}}\right)$$

the voltage at the output of the multiplier becomes

$$v_{\text{m}} = k[V_{\text{ref}} \sin(\omega_{\text{ref}} t)] \times \left[V_{\text{x}} \sin\left(\frac{\omega_{00}}{N} + \varphi_{\text{x}}\right)\right] \tag{5.18}$$

where k is the multiplication factor of the analog multiplier. Assuming that the multiplier is linear, (5.18) can be re-written as

$$v_{\text{m}} = \frac{1}{2} k V_{\text{ref}} V_{\text{x}} \cos\left[\left(\omega_{\text{ref}} - \frac{\omega_{00}}{N}\right)t - \varphi_{\text{x}}\right] - \frac{1}{2} k V_{\text{ref}} V_{\text{x}} \cos\left[\left(\omega_{\text{ref}} + \frac{\omega_{00}}{N}\right)t + \varphi_{\text{x}}\right]$$

and the signal at the output of the low-pass filter becomes

$$v_{\text{c}} = \frac{1}{2} k V_{\text{ref}} V_{\text{x}} \cos\left[\left(\omega_{\text{ref}} - \frac{\omega_{00}}{N}\right)t - \varphi_{\text{x}}\right] \tag{5.19}$$

This means that the control voltage v_{k} varies sinusoidally with a frequency that is equal to the difference between the frequency of the reference oscillator and that of the output of the frequency divider. Now, if we close the loop at C, the circuit locks at $v_{\text{k}} = v_{\text{k0}}$ which corresponds to the targeted output frequency, provided that v_{k0} is in the range of the amplitude of v_{c}.[10]

Hence, the output frequency becomes equal to ω_0 and maintains this value. But it must be noted that the reference oscillator checks the frequency of the VCO not at every period but once every N periods. Therefore, the control voltage at the output of the low-pass filter (that is a DC voltage when the circuit is locked) must keep its value for at least N periods.

5.3.3 Phase noise in oscillators

For an ideal sinusoidal oscillator the frequency spectrum contains only one component at the frequency of oscillation, f_0, which is usually called the "carrier". The components at the harmonics of f_0 indicate a nonlinear distortion on the waveform, which is an expected imperfection for an electronic circuit (see Fig. 5.18(a)). The unexpected reality is that the frequency spectrum of an oscillator contains an infinite number of side-frequencies on both sides of f_0, decreasing with the distance from f_0. The envelope of the side-frequencies exhibits a noise-like random character, as shown in Fig. 5.18(b). The power of a side-frequency in a 1 Hz bandwidth relative to the power of the carrier is called the "phase noise" of the oscillator for this frequency, and defined as

$$L(\Delta\omega) = 10 \log\left(\frac{P(f_0 + \Delta f)_{B=1\text{Hz}}}{P(f_0)}\right) \tag{5.20}$$

[10] For a detailed explanation of the locking mechanism refer to [47].

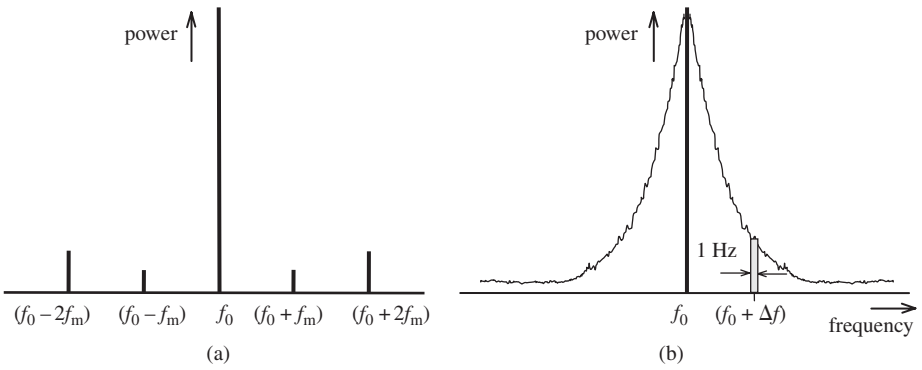


Figure 5.18 (a) The frequency spectrum of a "noiseless" oscillator. (b) The frequency spectrum of a real oscillator, for the vicinity of f_0.

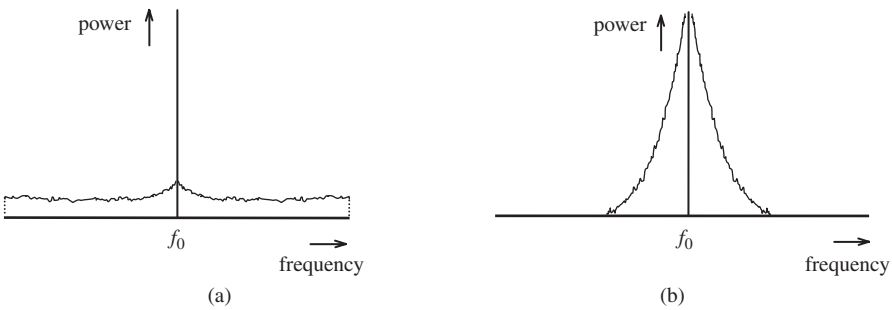


Figure 5.19 (a) Side-bands of a carrier, amplitude modulated with noise (the increase in the vicinity of the carrier corresponds to the $1/f$ noise). (b) Side-bands of a carrier, frequency modulated with white noise.

where $(f_0 + \Delta f)_{B=1\,Hz}$ is the power density (power for 1 Hz bandwidth) of a side-frequency of $f_0 + \Delta f$.

The physical reason of the noise side-frequencies can be explained by the modulation of f_0 with the noise in the circuit.

From modulation theory we know that if the amplitude of a carrier is modulated with a signal of f_m, two side-frequencies appear on the frequency spectrum, whose distances from f_0 are equal to f_m, and magnitudes are proportional to the amplitude of the modulating signal. Then we understand that, if an oscillator is amplitude modulated with "white" noise, two side-bands occur extending to infinity on both sides of f_0, whose envelope has an – almost – constant magnitude. If the noise in the system is not "white", for example if the dominant source of the noise is related to the transistors, owing to the flicker (or $1/f$ noise), then the magnitudes of the side-frequencies corresponding to low f_m values steadily increase towards the carrier, as shown in Fig. 5.19(a).

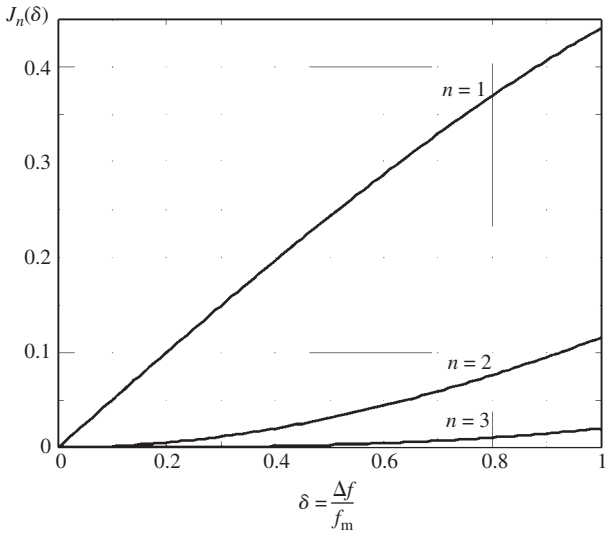


Figure 5.20 Bessel functions for very small frequency deviations as a function of the frequency modulation index.

The decreasing side-frequency components shown in Fig. 5.19(b) can be explained by the frequency modulation of f_0, i.e. the small Δf_0 fluctuations of the oscillation frequency around f_0. It is known from the modulation theory that if the frequency of a carrier is modulated with a modulation signal, side-frequencies occur on both sides of the carrier with a distance of f_m, $2f_m$, $3f_m$, etc. The magnitudes of these side-frequencies relative to the magnitude of the carrier depend on the frequency modulation index, $\delta = \Delta f_0/f_m$, and can be expressed using Bessel functions of the first kind, J_1, J_2, $\ldots$, J_n of δ:

$$\begin{aligned} v = {} & J_0(\delta) \times A \sin \omega_0 t \\ & + J_1(\delta) \times A[\sin(\omega_0 + \omega_m)t - \sin(\omega_0 - \omega_m)t] \\ & + J_2(\delta) \times A[\sin(\omega_0 + 2\omega_m)t + \sin(\omega_0 - 2\omega_m)t] \\ & + J_3(\delta) \times A[\sin(\omega_0 + 3\omega_m)t - \sin(\omega_0 - 3\omega_m)t] + \cdots \end{aligned}$$

where A is the amplitude and v is the instantaneous value of the frequency-modulated signal. Variations of the Bessel functions as a function of δ are given in the literature.[11]

In Fig. 5.20 the first three functions for small values of δ are shown, which correspond to a small Δf_0 fluctuation of the frequency around f_0. From these curves, it can be seen that the magnitudes of the side-frequencies corresponding to f_m, $2f_m$, $3f_m$, etc. decrease with f_m as shown in Fig. 5.19(b). From Fig. 5.21, the rate of decrease of the r.m.s. values of the side-frequencies with f_m can be calculated as 19.2 dB/decade, which is in good agreement with the ≈ 20 dB/decade slope obtained from measurements. Similar to the case of amplitude modulation, owing to contribution of flicker

[11] For example [48]. But the easiest way is to refer to MATLAB.

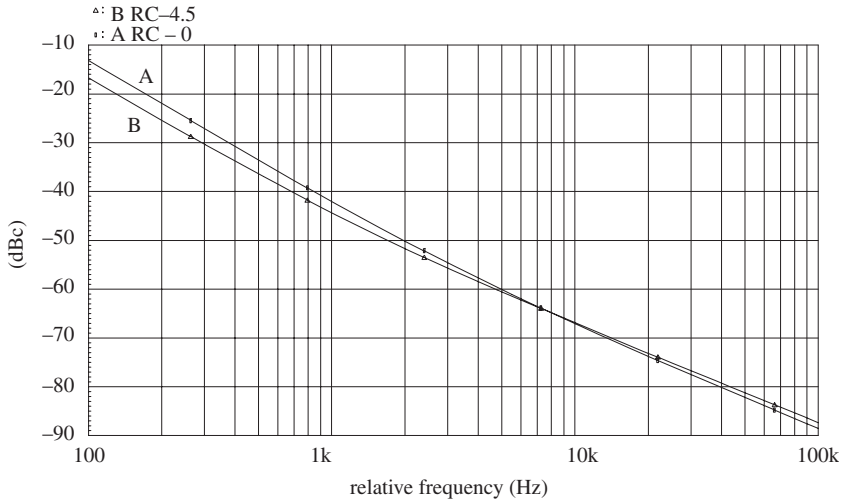


Figure 5.21 The phase noise characteristics of a differential negative resistance oscillator. Curve (A) and curve (B) correspond to $r_C = 0$ and $r_C = r_L$, respectively.

noise, the slope of the decrease of the magnitude of the envelope of the side-frequencies becomes higher in the vicinity of the carrier.

From the typical frequency spectrum of an oscillator shown in Fig. 5.18(b) we understand that the dominant contribution is related to the frequency modulation of the carrier with noise. There are several theories to explain the mechanism of this modulation. According to one of these theories, owing to the noise superimposed onto the oscillation signal, the zero-crossing points on the time axis may randomly shift back and forth, which corresponds to frequency modulation with a random and small frequency deviation [49].

Another theory is based on the random fluctuations of the limit-cycle of the oscillation that is related to the nonlinearity of the electronic part of the circuit [50]. In a negative resistance oscillator, for example, the amplitude for which the oscillation stabilizes depends on the value of the transconductance. The fluctuation of a the transconductance owing to the noise on the drain current of a transistor results in small variations of the amplitude and the phase of the oscillation, which changes the limit cycle.

The global frequency stability is certainly another factor that acts on the phase noise of the amplifier. To minimize the phase noise of an oscillator, it is necessary to try to minimize the effects of all these factors.

For example, to minimize the noise reaching the resonance circuit, a suitable filter can be used to by-pass the noise of the tail current source [51]. The amount of the distortion on the drain current that indicates where the amplitude of the oscillation will be self-stabilized can be adjusted to a level which minimizes the phase noise. Certainly, designing the circuit to fulfill the maximum frequency stability condition also helps to decrease the phase noise.

6 Analog–digital interface and system-level design considerations

In the earlier chapters of this book, we have introduced and examined the structure and operation of fundamental building blocks, or essential components, of high-frequency integrated circuits. The emphasis has been on transistor-level operation, the influence of device characteristics and parasitic effects, as well as the input–output behavior in time and frequency domains – with the intention to fill the gap between fundamental electronic circuits textbooks and more advanced RF IC design textbooks that mainly focus on the state-of-the-art. In this chapter, we will address a different domain and consider various aspects of "bringing together an entire system", especially for interfacing high-frequency analog components with corresponding digital processing blocks, using data converters. Instead of considering specific system architectures, the main aspects of system design will be presented with a generic approach in the following, as much as possible. The main philosophy is very similar to that of the earlier chapters: discussing design-oriented strategies and drawing attention to key issues that must be taken into account when combining analog and digital building blocks. For a detailed discussion of the system components and their circuit-level realizations, the reader is advised to consult any one of the excellent texts that already exist in this domain.[1]

6.1 General observations

The vast majority of integrated systems used in communication applications today consist of analog as well as digital building blocks, combined within one package in close proximity to each other, or fabricated on a single chip substrate. The majority of the algorithmic signal processing and modulation is handled by the digital system blocks, while the high-frequency communication is delegated to the analog HF front-end. The task of translation between these two domains, i.e. the analog–digital interface, is the responsibility of the data converters: the *analog-to-digital converter* (ADC) for converting analog signals into discrete-time, quantized data, and the *digital-to-analog converter* (DAC) for converting the output of digital processing into modulated, continuous-time, analog signals that can be transmitted over larger distances.

[1] For a detailed treatment of data converter architectures and the circuit-level realization of key building blocks, refer to Maloberti [61], Razavi [62], van de Plassche [63], or the Data Conversion Handbook [64].

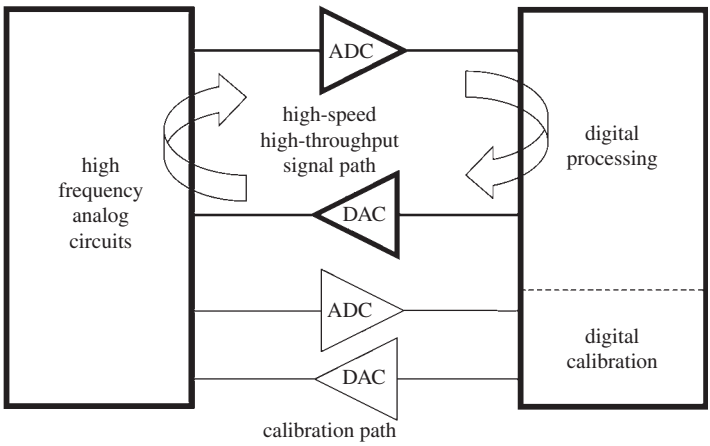


Figure 6.1 Simplified block diagram of a generic analog/digital system with data converters.

The role of digital circuitry in conjunction with high-frequency analog components is not only limited to algorithmic signal processing. Owing to their relatively smaller area overhead, digital circuits such as simple controllers, lookup tables and coefficient registers are becoming very convenient means for handling the off-line and on-line calibration of sensitive analog components. Digital gain and offset calibration of amplifiers, for example, have become increasingly commonplace. Note that in the case of calibration, circuit speed is a secondary concern – whereas accuracy, measured in number of bits, is of paramount importance. Thus, when considering the relation between the analog and digital parts of a system, it is also increasingly necessary to make a distinction between the high-speed *signal path*, which must handle a high data throughput (i.e. high processing speed) in real time, and the relatively low-speed *calibration path*, which must – by definition – have a higher bit resolution than the signal processing path (Fig. 6.1).

One particular observation is that in modern systems, the share of digital blocks is continuously increasing with respect to the overall system, and the boundary between the analog and digital modules is becoming more ambiguous. The increasing share of digital system blocks is mostly reflected in terms of increasing transistor count and functional complexity – but not necessarily in terms of actual silicon area – which we have to address in the context of system-level design. In fact, the *proportion* of the silicon area that must be reserved for high-frequency analog blocks on a typical mixed-signal chip is rising with each new technology generation, even though the functional complexity of the remaining digital system blocks is increasing. This is mainly owing to the fact that many of the active and passive components in high-frequency analog circuit blocks cannot be scaled down as easily as their digital counterparts, and this has to be taken into account in the overall system construction.

To give an example, an RF transceiver is usually defined as the entire system consisting of functional building blocks such as filters, amplifiers, frequency converters, modulator/demodulators, oscillators, synthesizers, data converters (ADC/DAC), switches, signal couplers, etc. The transceiver is not only formed by high-frequency (RF) components but

also by intermediate frequency (IF) and analog base-band circuitry and devices. The ADC and DAC are often seen as the boundary between the high-frequency analog parts of the transceiver and its digital counterpart. However, this boundary is getting increasingly blurred with state-of-the-art data converters operating at higher sampling rates – in this sense, the ADC/DAC and the corresponding digital signal processors now take over more and more functions of the IF and even the RF blocks. Thus, high-speed data converters are increasingly considered as part of the high-frequency systems, opening up new possibilities for so-called direct conversion systems, especially for mobile applications. The tendency in system-level design is to shift an increasing portion of high-frequency functions (including some of the filtering and signal conditioning, etc.) onto the digital section, as higher sampling and processing speeds become available. It is safe to assume that this trend will continue in the future. The shifting of data converters closer towards the analog front-end ultimately decreases the proportion of the analog circuitry in the overall system, and increases the functional complexity of the digital back-end. While this may increase the overall robustness and the flexibility of the system (using programmable and/or reconfigurable digital signal processing), it also increases the demand for higher conversion bandwidth in the ADC and DAC units (Fig. 6.2).

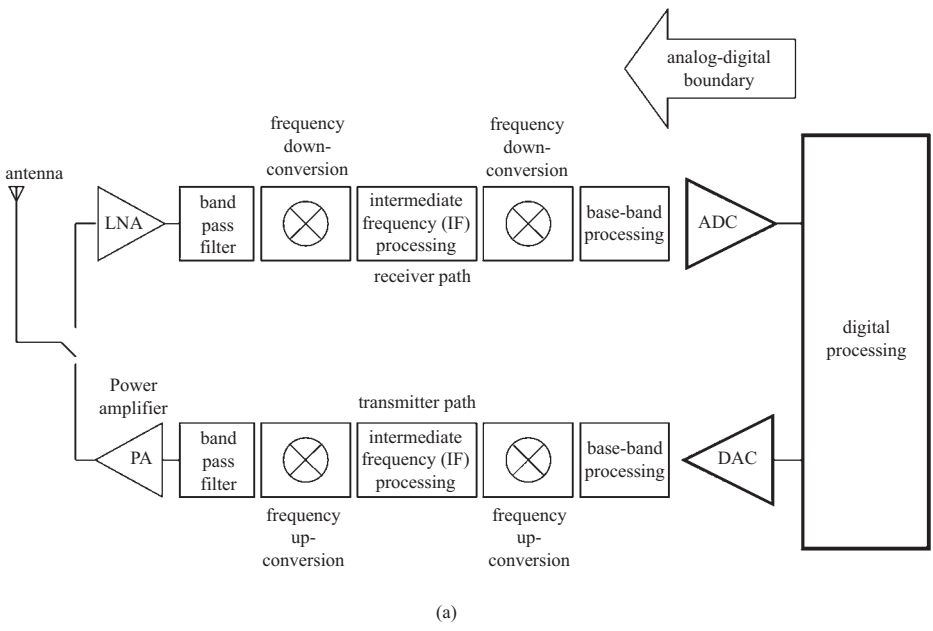
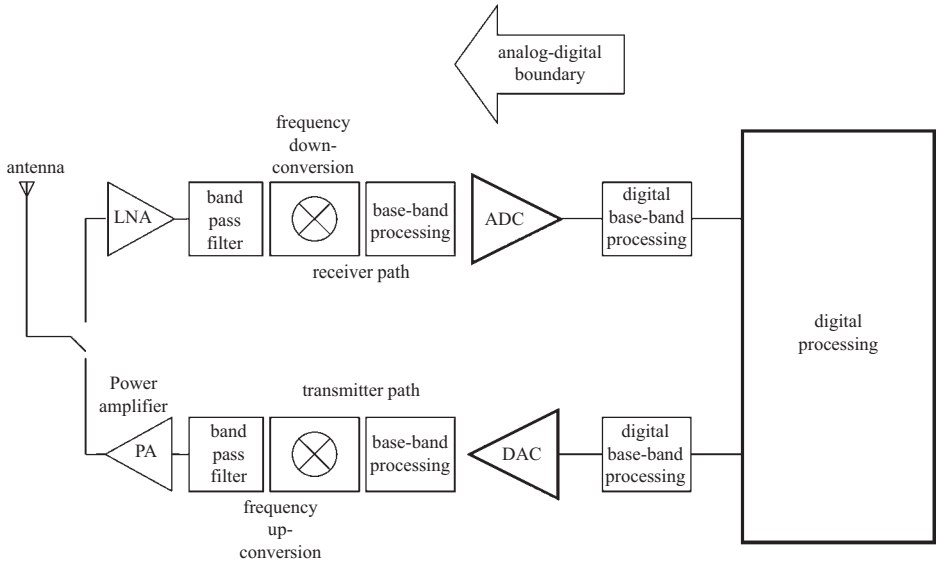
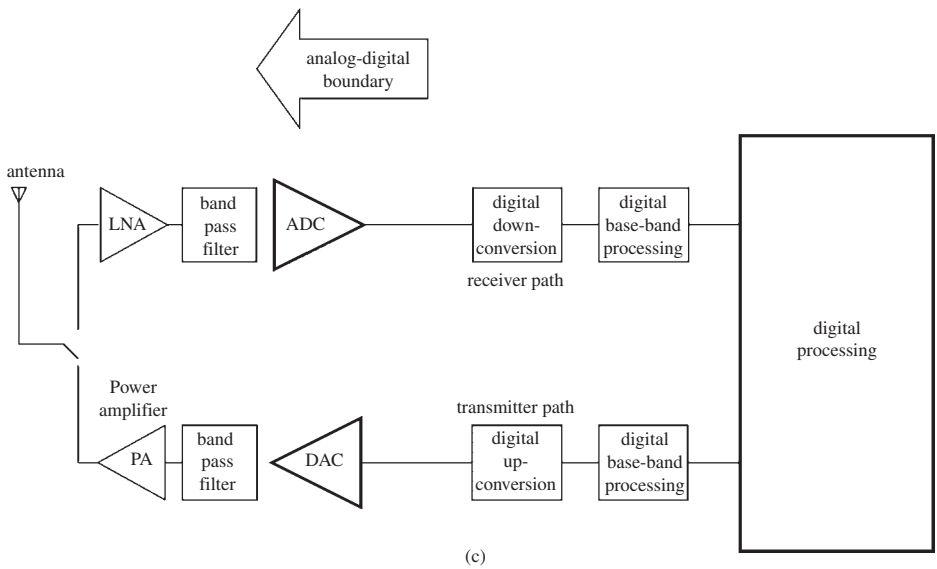


Figure 6.2 (a) Example of a conventional super-heterodyne transceiver architecture where the incoming high-frequency signals are first down-converted to the intermediate frequency (IF) range, and then to the base-band range, in two steps. The up-conversion is also handled in two steps. The sampling and conversion from analog into digital are done in the base-band domain. (b) The so-called "zero-IF" architecture relies on a single frequency down-conversion step, with the ADC and DAC operating at a higher sampling rate. (c) The direct sampling approach relies on very high-speed converters next to the analog front-end, and the down-conversion (as well as the corresponding up-conversion) takes place completely in the digital domain.



(b)



(c)

Figure 6.2 (cont.)

The interface between analog and digital domains requires particular attention in system design – especially when defining the specifications on both sides. One significant issue that has to be taken into account is that the operational characteristics of digital circuits and systems are almost exclusively described in the *time domain*. On the other hand, most – not all – of the key characteristics of analog circuits are

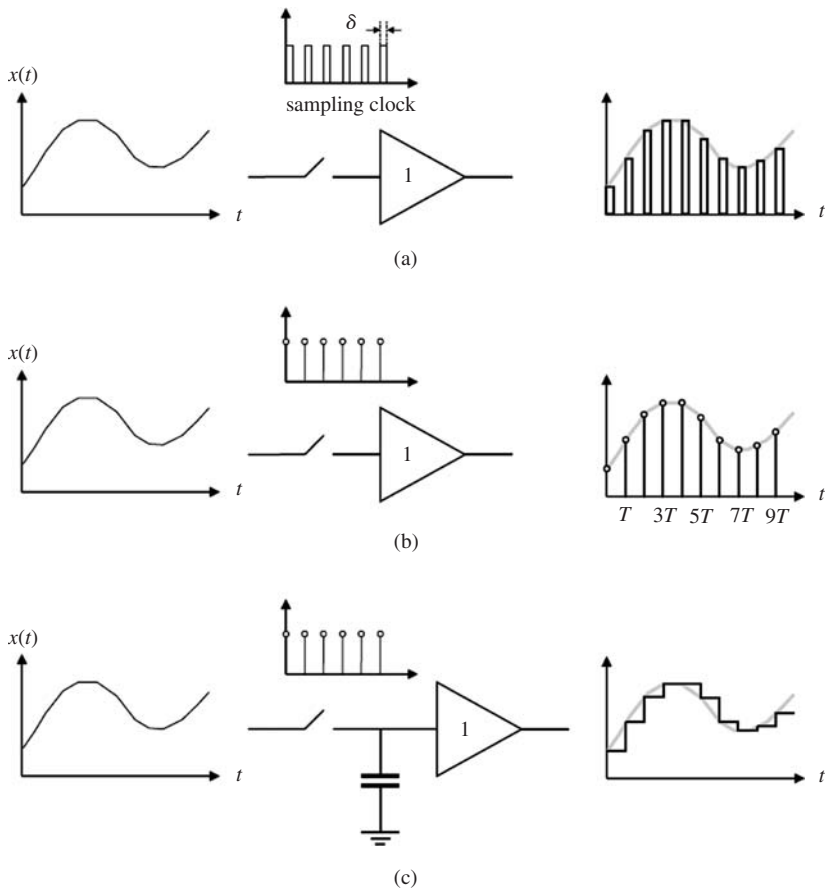


Figure 6.3 (a) Sampling of the analog input signal with an ideal sampler circuit. (b) As the sampling clock pulse width is reduced to an infinitesimally small duration, the sampled output will approach a sequence of modulated Dirac deltas. (c) Ideal sample-and-hold operation performed with a linear capacitor at the output of the sampling switch.

preferably described in the *frequency domain*, as we have seen extensively in the previous chapters of this book. Successful translation of key design specifications from one domain to the other requires a good understanding of the fundamental aspects of sampling theory, which provides the bridge between the continuous-time analog world and the discrete-time digital world.

6.2 Discrete-time sampling

Sampling is the key transformation that is required in all data converters. The sampling of a continuous-time analog signal can be performed by successive sample-and-hold operations using a periodic sampling clock, as seen in Fig. 6.3. The output of the

sampling process is a sequence of pulses whose amplitudes are derived from the input waveform samples. According to the *uniform sampling theorem*, a band-limited signal (i.e. a signal that does not have any spectral component beyond a certain frequency f_{max}) can be completely reconstructed from a set of uniformly spaced discrete-time samples, if these samples are obtained with a sampling rate (sampling frequency) of $f_s > 2\,f_{max}$. Note that, ideally, the output of the sampler is represented by a sequence of modulated Dirac deltas whose amplitude equals the amplitude of the sampled signal at the sampling times. Clearly, a practical sampling circuit does not generate a sequence of deltas but pulses with finite duration. The pulses are intended to represent the input waveform only at the exact sampling instances, nT, where $T = 1/f_s$. Figure 6.3(a) shows the block diagram of one such ideal sampler, which consists of a lossless, zero-delay switch, and an ideal unity-gain amplifier with infinite bandwidth. If the sampling clock pulse width δ is reduced to an infinitesimally short duration, the sampled output pulses will approach a sequence of Dirac deltas, as shown in Fig. 6.3(b). In practical applications, it is usually preferable to apply a temporary memory (hold) function at the output of the sampler, so that the amplitude of the sampled signal is preserved until the next sampling instant. For sampled voltage waveforms, a linear capacitor at the output of the sampling switch can fulfill this memory function, as shown in Fig. 6.3(c). This arrangement is commonly known as an ideal sample-and-hold circuit, with the assumption that the sampling pulse width δ can be made infinitesimally small. At the onset of each sampling clock pulse, the hold capacitor is instantaneously charged up to the level that corresponds to the input signal at that moment – and then preserves its charge until the next sampling instant. If the sampling clock pulse has a finite width, on the other hand, the same idealized circuit operates as a track-and-hold, where the output follows (tracks) the input during the period when the sampling clock is active, and then holds its value during the period when the sampling clock is inactive. To introduce some of the key aspects of sampling, we will assume ideal sampling with infinitesimally small pulse width, and no holding function, in the following.

In the following, we present some of the fundamental properties associated with signal sampling and reconstruction, in order to provide further insight concerning the operation of data converters. Consider a time-varying signal waveform which has a band-limited spectrum as shown in Fig. 6.4(a). If this waveform is sampled with an ideal sampler, at a sampling frequency that is larger than $2\,f_{max}$, the spectrum of the sampled signal can be reconstructed as depicted in Fig. 6.4(b), consisting of frequency-domain superposition of an infinite number of replicas of the input spectrum. These replicas are centered at integer multiples of the sampling frequency f_s, i.e. with each *image* spectrum being shifted along the frequency axis by $n\,f_s$, where n is an integer. Hence, the spectrum repeats itself periodically in the frequency domain. As long as the highest frequency component of the signal spectrum is smaller than one half of the sampling frequency, the original signal can be reconstructed by using a low-pass filter to isolate the band-limited signal. If the sampling frequency is less than $2\,f_{max}$, on the other hand, the replicas will partially overlap in the frequency domain (aliasing) and it will not be possible to reconstruct the original signal.

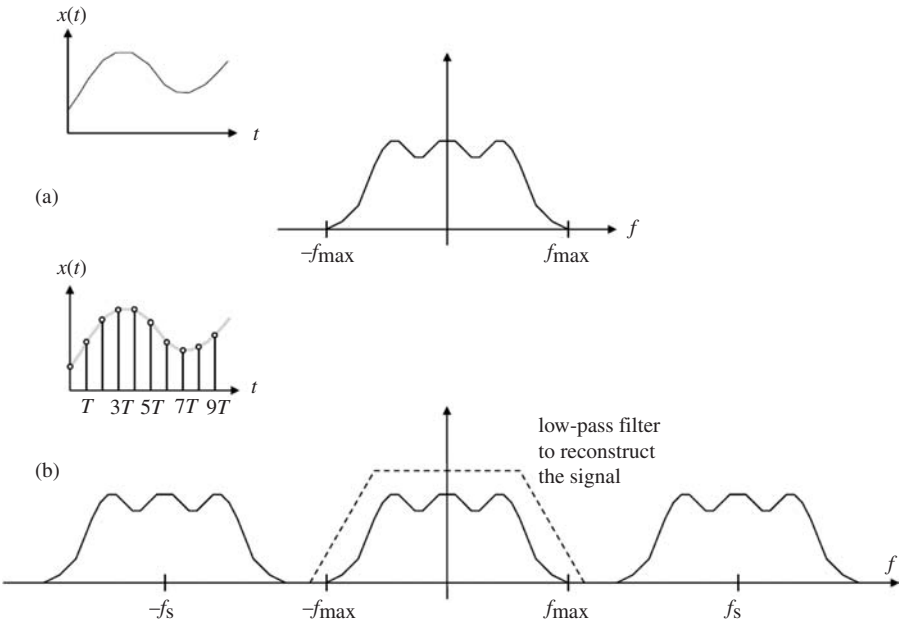


Figure 6.4 (a) Time-varying signal waveform and its band-limited frequency spectrum, with f_{max} showing the highest frequency component. (b) Reconstruction of the sampled signal produces an infinite number of image spectra, each shifted from the origin by an integer multiple of f_s.

It should also be kept in mind that the sampling condition described above must hold for all spectral components of the sampled signal, including unwanted noise and interferences. Note that noise usually has an unpredictable spectrum and it can produce spectral components at any frequency. If the input signal spectrum has an unexpected noise tail as shown in Fig. 6.5, the folded-shifted image (replica) of the same noise tail can corrupt the original signal in the sampled spectrum. To prevent this, it is advisable to filter out any unwanted spectral components that may lie beyond the intended limit frequency of f_{max} using an anti-aliasing filter, before sampling the signal.

6.3 Influence of sampling clock jitter

Up to this point, we have considered the conversion of continuous-time analog signals into discrete-time sampled signals, and their reconstruction back into continuous-time domain, assuming ideal components and ideal sampling conditions. Under realistic conditions, however, the sampling process is affected by the uncertainty of the clock edges in the time domain. The unpredictable variation of the clock edge, mainly owing to the thermal noise in the clock generator and the uncertainty of the logic delay, is called the *jitter* in the actual sampling instants. Sampling of a time-domain signal waveform in the presence of jitter is illustrated in Fig. 6.6. Note that the magnitude as

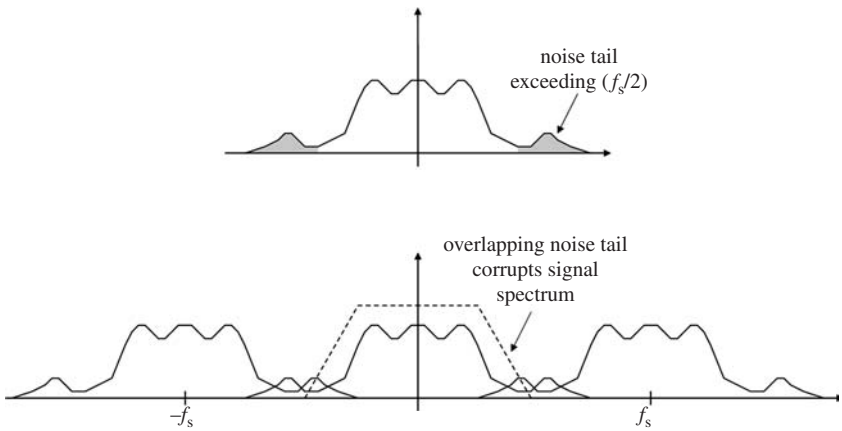


Figure 6.5 Aliasing of noise tails can corrupt the reconstructed signal spectrum. To avoid this, the bandwidth of the input signal must be limited using an anti-alias filter, prior to sampling.

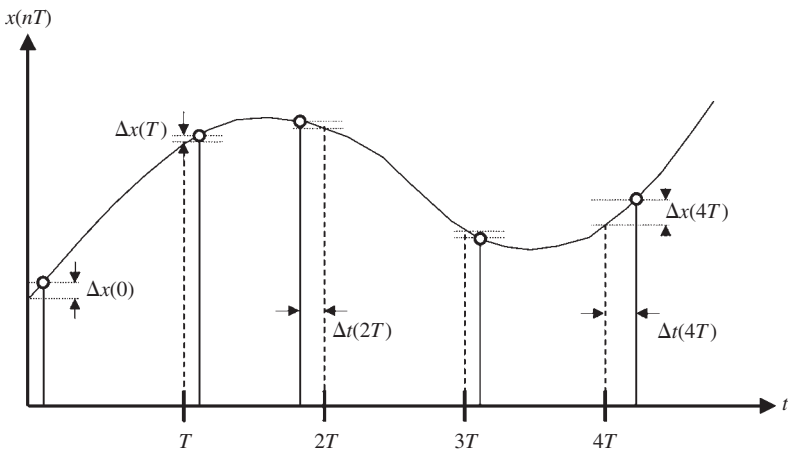


Figure 6.6 Timing jitter on the sampling clock will result in erroneous sample values, depending on the amount of timing error and the instantaneous slope of the signal.

well as the sign (direction) of jitter $\Delta t(nT)$ at each sampling instant will change according to a random distribution function, and independently of the previous sampling periods. The resulting sampling errors $\Delta x(nT)$ are also indicated in the figure. The amount of the sampling error depends on the magnitude and sign of the clock jitter, and also on the magnitude and sign of the time derivative of the sampled input signal. For a sinusoidal input signal with amplitude A and angular frequency ω_0, the sampling error at each sampling instant can be found to be

$$\Delta x(nT) = A\omega_0 \Delta t(nt)\cos(\omega_0 nT) \tag{6.1}$$

From the power of the jitter error, the resulting signal-to-noise ratio owing to clock jitter can be found as a function of input frequency and the mean clock jitter.

$$SNR_{\text{jitter}} = -20\log(\langle DJ(t)\rangle\omega_0)\text{dB} \tag{6.2}$$

Here, we assume that the instantaneous jitter amount $\Delta t(nT)$ is found by sampling the random variable $DJ(t)$, which is dictated by a white noise spectrum.

6.4 Quantization noise

Regarding the conversion of a continuous-time analog signal into a digital data stream, the discrete-time sampling process examined above is not the only source of systematic error. In order to process the sampled signal algorithmically by using binary arithmetic operators, the samples have to be quantized with a finite number of bits into discrete levels. This quantization process inevitably introduces a finite resolution for all sampled signals, which manifests itself as a *quantization error* ε_Q in the system – i.e. as an amplitude difference between the sampled signal and its level-quantized representation based on a limited number of bits. Under certain conditions, the quantization error can be represented in the form of *quantization noise*, which is reflected onto the signal-to-noise ratio of the data converter.

The main problems associated with the quantization process are illustrated with a simplified example in Fig. 6.7. Here, only a small section of the full signal range is shown, with discrete quantization levels described by four bits. It can be seen that for each discrete-time sampled signal, the quantizer has to perform a thresholding decision and assign the sample to the nearest discrete amplitude level. Unless the original sampled signal amplitude exactly corresponds to one of the quantization levels, the outcome of this thresholding decision is bound to produce an error term, at

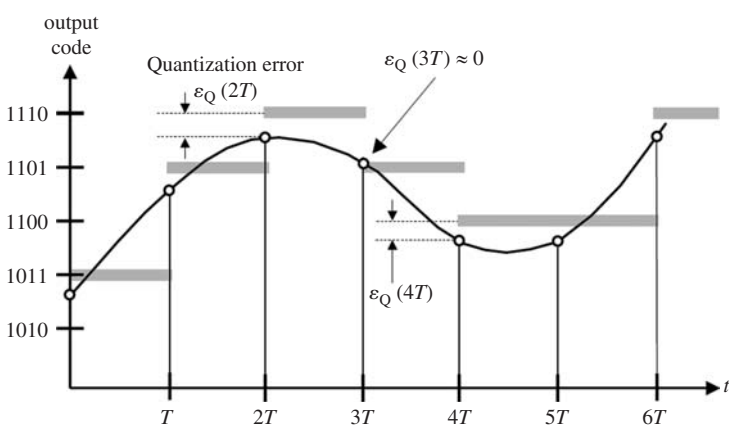


Figure 6.7 Quantization of samples with a finite number of bits which results in quantization error.

every sample point. If the dynamic range of the sampled signal can exercise all quantization levels with equal probability, and if the time-domain variation of the signal causes frequent code transitions to decorrelate successive samples, then the resulting quantization error can be treated as quantization noise with a widely spread spectrum. Assuming a sinusoidal input with the amplitude corresponding to the full dynamic range of the quantization interval, the time average power of the quantization noise can be calculated as a function of the maximum possible quantization error at each sample, i.e. the number of bits n used for quantization. Thus, the maximum achievable signal-to-noise ratio that is owing to quantization of samples with n bits can be expressed as:

$$SNR_{\text{quantization}} = (6.02n + 1.78)\text{dB} \tag{6.3}$$

This means that each additional bit of resolution can improve the SNR of the data converter by 6.02 dB. Note, however, that the expression (6.3) only accounts for the quantization noise, owing to a finite number of quantization steps assigned to the samples. In a real data converter, several other factors – such as thermal noise, bandwidth limitations of dynamic components, settling time limitations, etc. – bring about further errors that can also be viewed as additional noise components. Consequently, the expression given above can be re-written to define the *equivalent number of bits* (*ENOB*), with SNR_{total} representing the combined signal-to-noise ratio that accounts for all noise sources influencing the signal band of the data conversion system, including, but not only limited to, the quantization noise.

$$ENOB = (SNR_{\text{total}} - 1.78)/6.02 \tag{6.4}$$

This number is usually specified at a particular sampling frequency, and for a maximum bandwidth of the sampled signal. While the bit-resolution of a data converter is given as an integer number n, the effective number of bits derived from the expression above can be a real number, which is always less than n itself. The *ENOB* nevertheless provides a good indication of the overall accuracy of the data conversion system, under the specified operating conditions.

6.5 Converter specifications

This brief review of discrete-time sampling, and the limitations imposed by timing jitter and quantization noise, brings us to the fundamental specifications of data converters. The following discussion is kept at a sufficiently general level, with the aim of providing the reader with an introductory understanding of the key issues that relate to their functional characteristics. The specifications introduced here do not depend on the specific circuit realizations of the data converters. In the following, we will categorize the specifications into two broad classes, namely, as static and dynamic specifications.

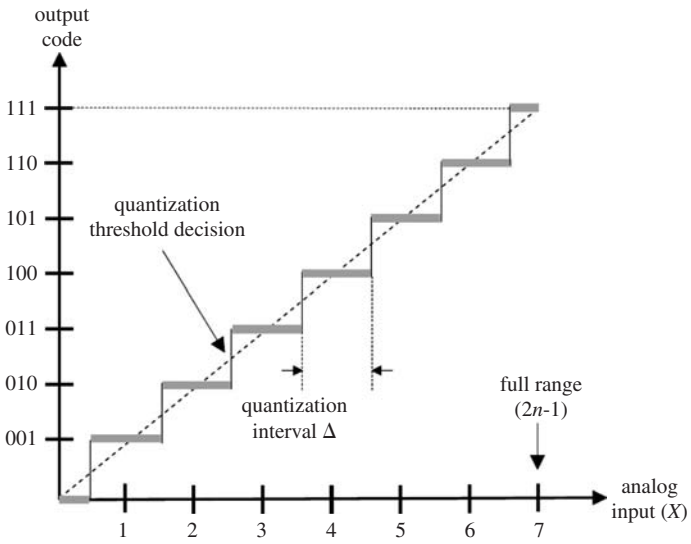


Figure 6.8 Ideal input–output characteristic of the analog-to-digital converter.

6.5.1 Static specifications

The static specifications are exclusively defined on the input–output characteristics of the data converter. The ideal input–output characteristic of an analog-to-digital converter (ADC) is a staircase, as depicted in Fig. 6.8, where the horizontal axis corresponds to the analog input level, and the vertical axis corresponds to the quantized output codes. Note that each step width is equal to the quantization interval $\Delta = X_{\text{full-range}}/(2^n-1)$, which corresponds to one *least-significant-bit* (LSB) in the output code and thereby defines the analog *resolution* of the converter. In an n-bit converter, the output codes range from 0 to (2^n-1) and, consequently, the input full scale range $X_{\text{full-range}}$ is divided into (2^n-1) quantization intervals, with the decision threshold for each quantization interval placed in the mid-point. Thus, the quantization error resulting from this thresholding decision ranges from $-(\Delta/2)$ to $+(\Delta/2)$, becoming equal to zero in the mid-point of each step. Deviations from this ideal input–output characteristic will manifest themselves in various ways.

Figure 6.9(a) shows a non-ideal input–output characteristic of an ADC with *offset error*, which causes a shift of all quantization steps by an equal amount. The offset error can also be defined for a DAC. In both cases, the input–output characteristic is shifted, while remaining parallel to the ideal one. A *gain error*, on the other hand, manifests itself as a change in the slope of the transfer characteristic, as shown in Fig. 6.9(b). In the presence of offset and/or gain error, it is expected that the validity range of all quantization intervals (and output codes) will change, and in extreme cases, some of the quantization intervals and the corresponding output codes may not be exercised at all. Gain and offset errors can occur simultaneously, in which case their

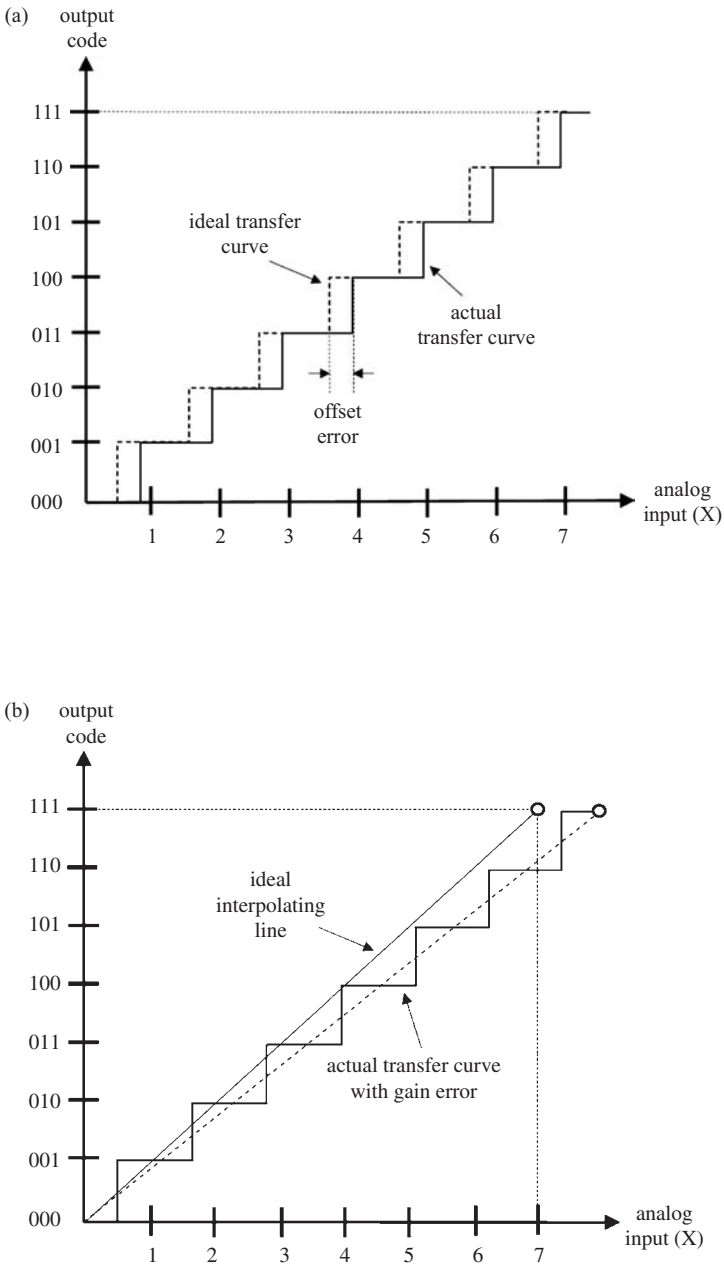


Figure 6.9 (a) Input–output characteristic with offset error. (b) Input–output characteristic with gain error.

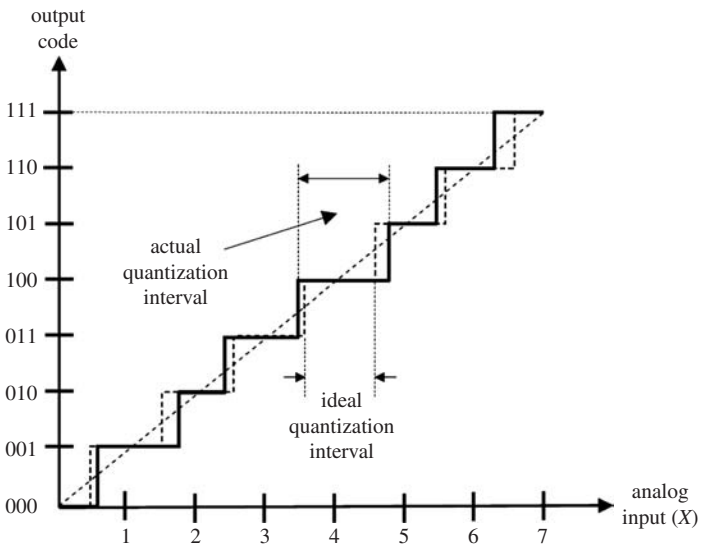


Figure 6.10 Deviation of the quantization interval from its ideal width leads to differential nonlinearity (DNL) error.

effects on the input–output characteristics will overlap. Nevertheless, since both gain and offset errors are linear non-idealities, they can usually be detected and corrected, using calibration techniques.

The *differential nonlinearity* (DNL) error is defined as the deviation of the quantization step (interval) width from that of an ideal step width (Fig. 6.10). This deviation can occur owing to systematic non-idealities in circuit elements such as the nonlinear characteristics of transistors, and also as a result of random variations in the size and value of replicated components such as the resistors in a voltage-divider chain. The DNL error is usually specified in terms of percentage of the full scale, or in terms of LSB. Since the error of each quantization interval is measured separately, the stated DNL value is usually the maximum error among all quantization intervals of the converter. Regardless of its origin, the DNL error cannot be corrected completely by calibration, owing to its inherently nonlinear nature. In a data converter with no other non-idealities, a DNL error with a maximum magnitude of less than 0.5 LSB is usually tolerable, since the resulting loss of accuracy in this case will always remain less than one LSB. Figure 6.11 shows the DNL variation of a typical 12-bit ADC, over 4096 individual quantization levels. The maximum deviation remains less than 0.5 LSB, which is (1/8192) of the full range.

The *integral nonlinearity* (INL) error is defined as the deviation of the input–output characteristic from the endpoint-fit line, which practically connects the starting and ending points of the transfer function (Fig. 6.12). As in the DNL case, the error is calculated individually for each quantization interval (or for each output code) and the maximum error among all quantization intervals is cited as the DNL

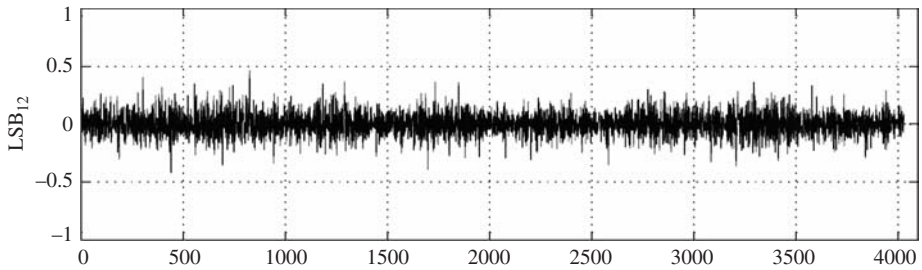


Figure 6.11 DNL variation of a 12-bit ADC, over the entire quantization range.

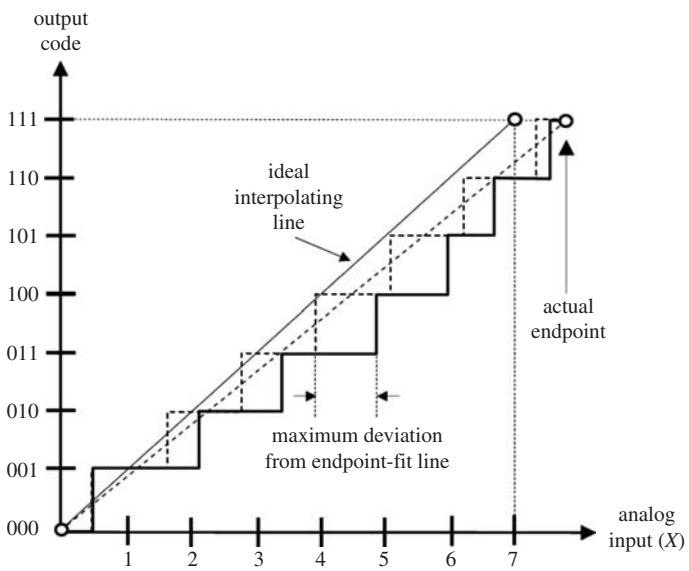


Figure 6.12 Deviation of the input–output characteristic from the endpoint-fit line.

value, in terms of percentage of the full scale or in terms of LSB. The definition of the error as the deviation from the endpoint-fit line instead of the ideal interpolating line effectively decouples the nonlinearity measure from possible offset and gain errors. A large deviation from the endpoint-fit line corresponds to harmonic distortion, which also manifests itself on the dynamic specifications. Similarly, a large DNL term is seen as an additional noise term that impacts the overall SNR, together with quantization noise. In the example shown in Fig. 6.13 for a 12-bit ADC, the maximum INL error is seen to be less than 4 LSB. Note the INL error on the two endpoints is equal to zero since, by definition, the INL is calculated as the deviation from the endpoint-fit line.

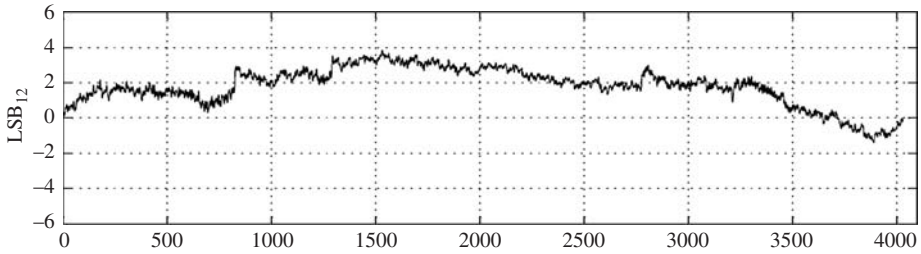


Figure 6.13 INL variation of a 12-bit ADC, over the entire quantization range.

6.5.2 Frequency-domain dynamic specifications

The *signal-to-noise ratio* (SNR) is one of the key dynamic characteristics of data converters. As we have already seen, the upper limit of the SNR of a data converter is primarily determined by its bit-resolution, i.e. the number of bits used to quantize and represent the sampled signals. In addition to this quantization noise, the SNR of a data converter accounts for the influence of all noise sources, within the entire signal frequency interval. The nonlinear distortion terms generated by the sinusoidal input, however, are *not* included in this calculation.

The *total harmonic distortion* (THD) is the ratio of the root-mean-square value of the fundamental signal to the mean value of the root-sum-square of its harmonics. Note that, generally, only the first five harmonics are considered to be significant. The THD of an ADC is specified with the input signal at full scale.

The *signal-to-noise-and-distortion ratio* (SNDR or SINAD) is defined as the ratio between the root-mean-square of the fundamental signal and the root-sum-square of all harmonic components, plus all noise components. In the literature, this measure is sometimes also called the *total harmonic distortion plus noise* (THD+N). Being one of the quantitative specifications that takes into account all possible noise and distortion components, the SNDR is typically used to estimate the *equivalent number of bits* (ENOB) of a data converter – by replacing SNR_{total} in Expression (6.4). The SNDR is usually specified at a given sampling rate, as a function of the input signal frequency, and also at a given input signal frequency, as a function of the sampling frequency. Since the linear non-idealities of circuit components (such as bandwidth and settling time limitations) tend to manifest themselves at higher operating frequencies, the SNDR usually degrades with frequency. The sampling frequency at which the SNDR drops by 3 dB determines the maximum sampling rate of the converter.

The *spurious-free-dynamic range* (SFDR) is the ratio of the root-mean-square value of the signal to the root-mean-square value of the worst spurious signal, regardless of where this spur is located in the frequency spectrum. The worst spur may or may not be a harmonic of the original signal. SFDR is an important specification in communications systems because it represents the smallest value of signal that can be distinguished from a large interfering signal. Figure 6.14 shows the calculation of the SFDR based on a sample frequency spectrum, where the distance between the signal peak and the largest

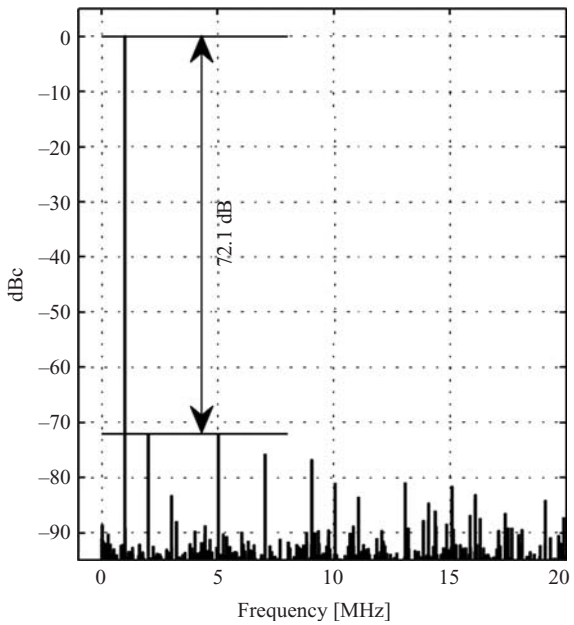


Figure 6.14 Calculation of SFDR based on a sample frequency spectrum.

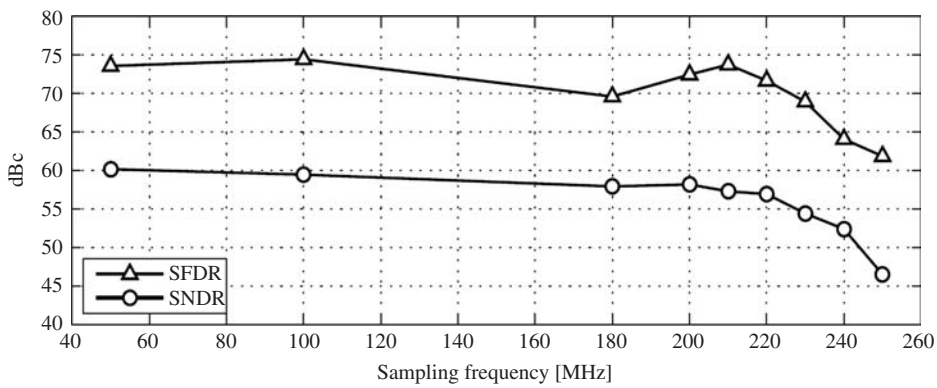


Figure 6.15 Variation of SNDR and SFDR as a function of the sampling clock frequency for a high-speed pipelined analog-to-digital converter.

spur is shown to be 72.1 dB. As in the case of SNDR, the SFDR values are usually calculated for the entire input frequency spectrum at a given sampling clock frequency, and also for a range of sampling clock frequencies at a given input frequency. Figure 6.15 shows the variation of the SNDR and SFDR of a pipelined ADC as a function of the sampling frequency, at an input frequency of 20 MHz. It can be seen that both SNDR and SFDR curves tend to roll off as the sampling frequency is increased, and they both drop by more than 3 dB beyond the usable sampling speed limit of the ADC.

It is interesting to note that the three important frequency-domain specifications discussed here, namely, SNR, THD and SNDR, are linked in a very straightforward manner. These three measures are defined as the numerical ratios of (S/N), (S/D) and ($S/(N+D)$), respectively:

$$\begin{aligned} \text{SNR} &= 20\log(S/N) \\ \text{THD} &= 20\log(S/D) \\ \text{SNDR} &= 20\log[S/(N+D)] \end{aligned} \tag{6.5}$$

With simple manipulation, it can be shown that any one of these three specifications can be derived using the following expressions, as long as the remaining two are known:

$$\begin{aligned} \text{SNR} &= -10\log(10^{-SNDR/10} + 10^{-THD/10}) \\ \text{THD} &= -10\log(10^{-SNDR/10} + 10^{-SNR/10}) \\ \text{SNDR} &= -10\log(10^{-SNR/10} + 10^{-THD/10}) \end{aligned} \tag{6.6}$$

6.6 Additional observations on noise in high-frequency ICs

We have already seen that the continuing trends in technology are enabling the integration of complete systems on a single die, which may include a combination of RF transceivers, analog processing, A/D and D/A conversion as well as complex digital functions and memory on a single chip. However, when sensitive analog parts are combined with complex digital blocks operating at very high switching frequencies, the noise generated by the digital parts is inevitably transmitted to the analog blocks, predominantly through the common substrate, resulting in a reduction of the dynamic range, or reduction of the accuracy of the analog circuits. While discussing the co-existence of high-frequency, high-sensitivity analog blocks with high-speed digital blocks, this fact also has to be taken into account.

Noise in digital CMOS circuits is mainly generated by the rapid voltage variations caused by the switching of logic states, and the related charge-up/charge-down currents. In a conventional CMOS logic gate, the rapid change of voltage in internal nodes is coupled to the substrate through junction or wiring capacitances, causing charges to be injected into the substrate. Eventually, these substrate currents cause voltage drops that can perturb analog circuits through capacitive coupling and through variation of the threshold voltage owing to a body effect. Additionally, the high instantaneous currents needed to rapidly charge or discharge parasitic capacitances add up to large current spikes in the supply and ground distribution networks, a phenomenon known as simultaneous switching noise (SSN). These current spikes cause voltage noise primarily through the inductance of off-chip bond-wires and on-chip power-supply rails. Ground supply networks are usually connected to the substrate, resulting in a direct coupling of the noise, and power networks are typically connected to very large N-well areas, resulting in a consequently very large parasitic coupling capacitance to the

substrate. Therefore, power and ground distribution networks are very noisy in CMOS circuits, and at the same time ideal mediums for the noise coupling to the substrate. Signal nets can also couple to the substrate, through diffusion and wiring capacitances, and signals with high energy and switching activity are thus critical from a noise perspective. This is the case especially for clock networks, which carry the most active signals and dissipate large amounts of power.

Two effective techniques to reduce the noise generation in digital circuits are the reduction of the voltage swings and the cancellation of transient currents during switching events. Specialized logic circuit families (single-ended and differential) that generate less noise than classical CMOS logic can be implemented, and are thus suitable for integration in a mixed-mode environment as a replacement or as a complement of CMOS logic.

Experimental studies have shown that single-ended logic families achieve only marginal improvement over regular CMOS in terms of noise. While differential logic families are the most promising candidates that offer improved noise reduction, traditional automation tools and design flows fail to accommodate many aspects associated with their differential nature. For this reason, large-scale implementation of digital circuits with low-noise differential logic families and their integration with high-sensitivity analog circuits remains a difficult task, which poses significant challenges for future systems design.

Appendix A
Mobility degradation due to the transversal field

In Chapter 1 it was mentioned that, especially for small-geometry (thin gate-oxide) MOS transistors, the electron and hole mobilities dramatically decrease due to the transversal electric field on the channel region. This effect is investigated in several publications and modeled in all advanced transistor models.

In this appendix the degradation curves of electron and hole mobilities of surface channel NMOS and PMOS transistors are given for certain typical technologies.

The curves were based on the expressions given in Refs. [3] and [4].

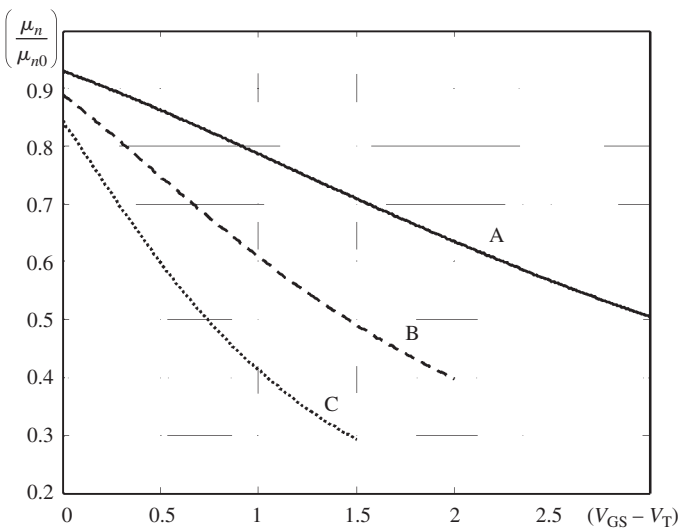


Figure A.1 Relative variation of the channel electron mobility of typical NMOS transistors as a function of the gate overdrive voltage.
Curve A: for a typical 0.35 micron technology ($V_{TN} = 0.5$ V, $T_{ox} = 7.5$ nm).
Curve B: for a typical 0.18 micron technology ($V_{TN} = 0.35$ V, $T_{ox} = 4$ nm).
Curve C: for a typical 0.13 micron technology ($V_{TN} = 0.25$ V, $T_{ox} = 2.3$ nm).

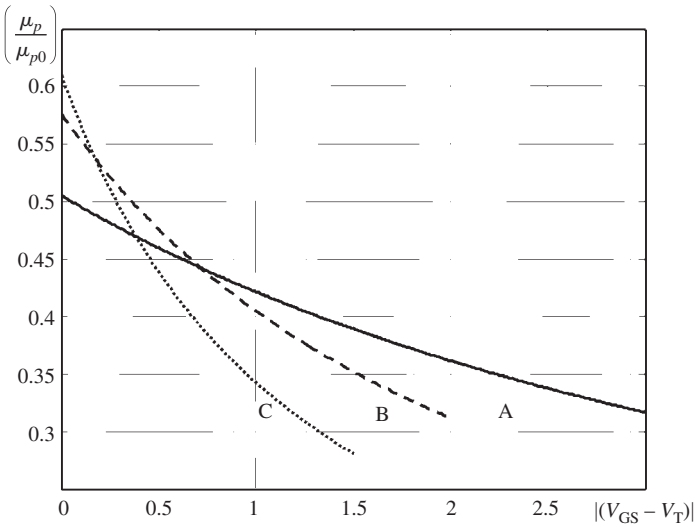


Figure A.2 Relative variation of the surface channel hole mobility of typical PMOS transistors as a function of the gate overdrive voltage.
Curve A: for a typical 0.35 micron technology ($V_{TP} = -1$ V, $T_{ox} = 7.5$ nm).
Curve B: for a typical 0.18 micron technology ($V_{TP} = -0.35$ V, $T_{ox} = 4$ nm).
Curve C: for a typical 0.13 micron technology ($V_{TP} = -0.20$ V, $T_{ox} = 2.3$ nm).

Appendix B
Characteristic curves and parameters of AMS 0.35 micron NMOS and PMOS transistors

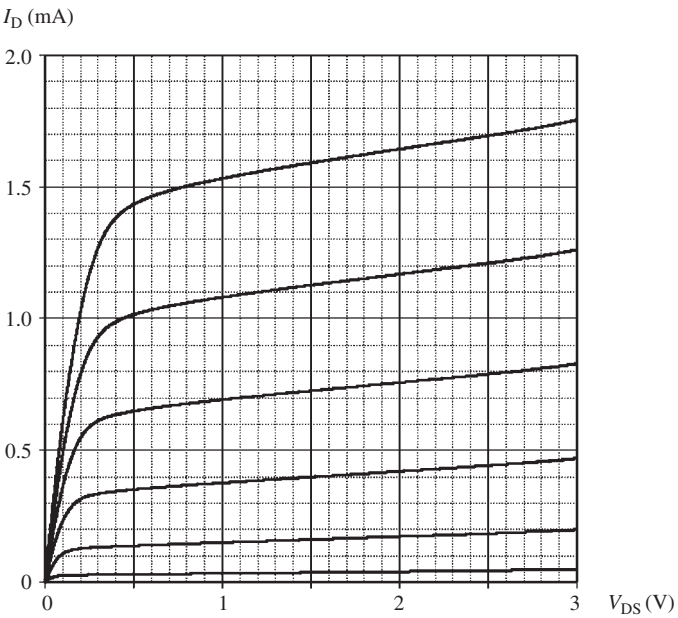


Figure B.1 The output characteristics of an AMS 0.35 NMOS transistor (L=0.35 μm, W=35 μm, V_{GS} = 0.6 V–1.1 V).

NMOS basic parameters for hand calculations (obtained from BSIM3-v3 parameters and rounded)

$V_T(0) = 0.5\,[\mathrm{V}],\ \mu_{n0} = 475\left[\mathrm{cm}^2/\mathrm{V.s}\right]$

$T_{ox} = 7.6\,[\mathrm{nm}], C_{ox} = 4.54 \times 10^{-7}\left[\mathrm{F/cm}^2\right] = 4.54\left[\mathrm{fF}/\mu\mathrm{m}^2\right]$

$C_{DGO} = C_{SGO} = 1.2 \times 10^{-10}[\mathrm{F/m}] = 0.12[\mathrm{fF}/\mu\mathrm{m}],\ C_{GBO} = 1.1 \times 10^{-10}[\mathrm{F/m}] = 0.11[\mathrm{fF}/\mu\mathrm{m}]$

$C_j(0) = 9.4 \times 10^{-4}\left[\mathrm{F/m}^2\right] = 0.94\left[\mathrm{fF}/\mu\mathrm{m}^2\right],\ C_{jsw}(0) = 2.5 \times 10^{-10}[\mathrm{F/m}] = 0.25[\mathrm{fF}/\mu\mathrm{m}]$

$\mathrm{RDSW} = 345[\mathrm{ohm}/\mu\mathrm{m}],\ \mathrm{RSH} = 75[\mathrm{ohm/square}]$

$\lambda = 0.073\left[\mathrm{V}^{-1}\right]$ (derived from output curves)

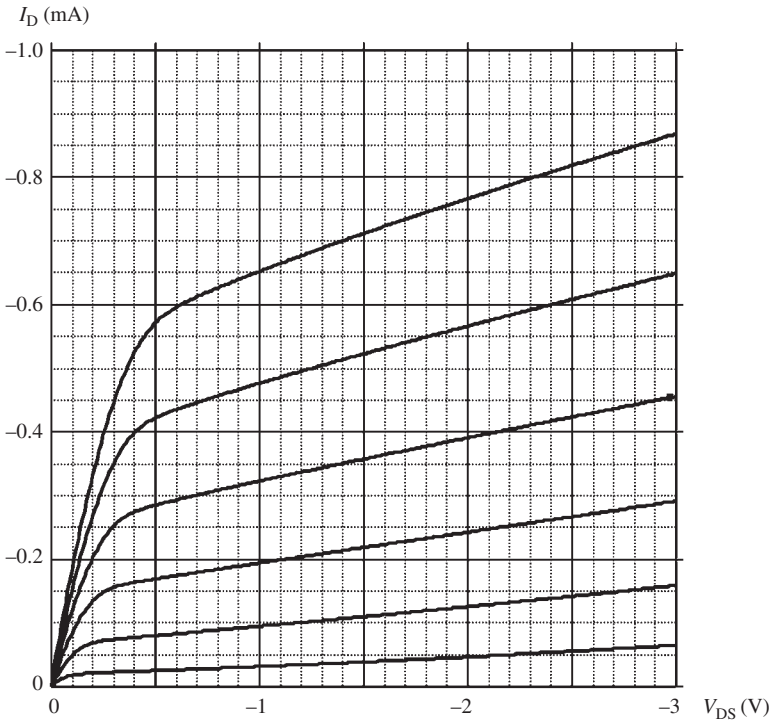


Figure B.2 The output characteristics of an AMS 0.35 PMOS transistor (L=0.35 V_{DS} (V) μm, W=35 μm, V_{GS} = 0.6 V–1.3 V).

PMOS basic parameters for hand calculations (obtained from BSIM3-v3 parameters and rounded)

$V_T(0) = -0.7[\text{V}],\ \mu_{p0} = 148\left[\text{cm}^2/\text{V.s}\right]$

$T_{ox} = 7.6[\text{nm}],\ C_{ox} = 4.54\times10^{-7}\left[\text{F}/\text{cm}^2\right] = 4.54\left[\text{fF}/\mu\text{m}^2\right]$

$C_{DGO} = C_{SGO} = 8.6\times10^{-11}[\text{F}/\text{m}] = 0.086[\text{fF}/\mu\text{m}],\ C_j(0) = 1.36\times10^{-3}\left[\text{F}/\text{m}^2\right] = 1.36\left[\text{fF}/\mu\text{m}^2\right]$

$C_j(0) = 1.36\times10^{-3}\left[\text{F}/\text{m}^2\right] = 1.36\left[\text{fF}/\mu\text{m}^2\right],\ C_{jsw}(0) = 3.2\times10^{-10}[\text{F}/\text{m}] = 0.25[\text{fF}/\mu\text{m}]$

$\text{RDSW} = 1033[\text{ohm}/\mu\text{m}],\ \text{RSH} = 130[\text{ohm/square}]$

$\lambda = 0.2\left[\text{V}^{-1}\right]$ (derived from output curves)

Appendix C
BSIM3-v3 parameters of AMS 0.35 micron NMOS and PMOS transistors

BSIM3-v3 parameters of AMS 035 micron NMOS and PMOS transistors[1]

```
.MODEL MODN NMOS LEVEL=7

* ____________________________________________________
********* AMS 0.35 SIMULATION PARAMETERS *********
* ____________________________________________________
* format : PSPICE
* model : MOS BSIM3v3
* process : C35
* revision : 3.1;
* extracted : B10866 ; 2002–12; ese(487)
* doc# : ENG-182 REV_3
* ____________________________________________________
*                TYPICAL MEAN CONDITION
* ____________________________________________________
*
*      *** Flags ***
+MOBMOD = 1.000e+00        CAPMOD = 2.000e+00
+NOIMOD = 3.000e+00

*      *** Threshold-voltage related model parameters ***
+K1 =5.0296e−01
+K2 =3.3985e−02        K3 = −1.136e+00        K3B = −4.399e−01
+NCH =2.611e+17        VTH0 = 4.979e−01
+VOFF=−8.925e−02       DVT0 = 5.000e+01       DVT1 = 1.039e+00
+DVT2 =−8.375e−03      KETA =2.032e−02
```

[1] Note that usually there are several versions of device parameters for transistors with the same lithographic dimensions. The parameters given in this appendix are related to one of the conventional 0.35 μm devices fabricated by Austria Micro Systems AG (abreviated as AMS). Other parameter sets (e.g. high voltage version), detailed technology data, as well as layout design rules can be requested directly from the foundry and/or from academic service organizations such as EUROPRACTICE.

```
+PSCBE1 =3.518e+08      PSCBE2 =7.491e−05
+DVT0W = 1.089e−01      DVT1W =6.671e+04      DVT2W =−1.352e−02

*       *** Mobility related model parameters ***
+UA =4.705e−12          UB =2.137e−18         UC =1.000e−20
+U0 =4.758e+02

*       *** Subthreshold related parameters ***
+DSUB =5.000e−01        ETA0 =1.415e−02       ETAB =−1.221e−01
+NFACTOR=4.136e−01

*       *** Saturation related parameters ***
+EM = 4.100e+07         PCLM = 6.948e−01
+PDIBLC1 = 571e−01      PDIBLC2 = 2.065e−03   DROUT =5.000e−01
+A0 =2.541e+00          A1 = 0.000e+00        A2 = 1.000e+00
+PVAG = 0.000e+00       VSAT =1.338e+05       AGS = 2.408e−01
+B0 =4.301e−09          B1 =0.000e+00         DELTA =1.442e−02
+PDIBLCB  =3.222e−01

*       *** Geometry modulation related parameters ***
+W0 = 2.673e−07         DLC =3.0000e−08
+DWB = 0.000e+00        DWG =0.000e+00
+LL = 0.000e+00         LW = 0.000e+00        LWL =0.000e+00
+LLN =1.000e+00         LWN =1.000e+00        WL =0.000e+00
+WW =−1.297e−14         WWL =−9.411e−21       WLN =1.000e+00
+WWN =1.000e+00

*       *** Temperature effect parameters ***
+AT =3.300e+04          UTE =−1.800e+00
+KT1 =−3.302e−01        KT2 =2.200e−02        KT1L =0.000e+00
+UA1 =0.000e+00         UB1 =0.000e+00        UC1 =0.000e+00
+PRT =0.000e+00

*       *** Overlap capacitance related and dynamic model parameters ***
+CGDO =1.200e−10        CGSO =1.200e−10       CGBO =1.100e−10
+CGDL =1.310e−10        CGSL =1.310e−10       CKAPPA =6.000e−01
+CF =0.000e+00          ELM =5.000e+00
+XPART =1.000e+00       CLC =1.000e−15        CLE =6.000e−01

*       *** Parasitic resistance and capacitance related model parameters ***
+RDSW =3.449e+02
+CDSC =0.000e+00        CDSCB = 1.500e−03     CDSCD =1.000e−03
+PRWB = −2.416e−01      PRWG =0.000e+00       CIT =4.441e−04

*       *** Process and parameters extraction related model parameters ***
+TOX = 7.575e−09        NGATE =0.000e+00
+NLX =1.888e−07
```

```
*       *** Substrate current related model parameters ***
+ALPHA0 =0.000e+00          BETA0 = 3.000e+01

*       *** Noise-effect related model parameters ***
+AF = 1.507e+00             KF = 2.170e−26          EF = 1.000e+00
+NOIA =1.121e+19            NOIB =5.336e+04         NOIC =−5.892e−13

*       *** Common extrinsic model parameters ***
+LINT = −5.005e−08          WINT = 9.4030e−08       XJ = 3.000e−07
+RSH =7.000e+01             JS =1.000e−05
+CJ =9.400e−04              CJSW =2.500e−10
+CBD =0.000e+00             CBS =0.000e+00          IS =0.000e+00
+MJ =3.400e−01              N =1.000e+00            MJSW =2.300e−01
+PB =6.900e−01              TT =0.000e+00
+PBSW =6.900e−01

MODEL MODP PMOS LEVEL=7

* format : PSPICE
* model : MOS BSIM3v3
* process : C35
* revision : 3.1;
* extracted : C64685 ; 2002–12; ese(487)
* doc# : ENG-182 REV_3
* ____________________________________________________
*                   TYPICAL MEAN CONDITION
* ____________________________________________________
*
*   *** Flags ***
+MOBMOD =1.000e+00          CAPMOD = 2.000e+00
+NOIMOD =3.000e+00

*   *** Threshold-voltage related model parameters ***
+K1 =5.9959e−01
+K2 =−6.038e−02             K3 =1.103e+01           K3B =−7.580e−01
+NCH =9.240e+16             VTH0 =−6.915e−01
+VOFF =−1.170e−01           DVT0 =1.650e+00         DVT1 =3.868e−01
+DVT2 =1.659e−02           KETA =−1.440e−02
+PSCBE1 =5.000e+09          PSCBE2 =1.000e−04
+DVT0W =1.879e−01           DVT1W =7.335e+04        DVT2W =−6.312e−03

*       *** Mobility related model parameters ***
+UA = 5.394e−10             UB = 1.053e−18          UC =1.000e−20
+U0 = 1.482e+02
```

```
*       *** Subthreshold related parameters ***
+DSUB = 5.000e-01        ETA0 = 2.480e-01        ETAB = -3.917e-03
+NFACTOR=1.214e+00

*       *** Saturation related parameters ***
+EM = 4.100e+07          PCLM = 3.184e+00
+PDIBLC1=1.000e-04       PDIBLC2=1.000e-20       DROUT =5.000e-01
+A0 = 5.850e-01          A1 = 0.000e+00          A2 = 1.000e+00
+PVAG =0.000e+00         VSAT =1.158e+05         AGS =2.468e-01
+B0 =8.832e-08           B1 =0.000e+00           DELTA=1.000e-02
+PDIBLCB=1.000e+00

*       *** Geometry modulation related parameters ***
+W0 =1.000e-10           DLC =2.4500e-08
+DWB =0.000e+00          DWG =0.000e+00
+LL =0.000e+00           LW =0.000e+00           LWL =0.000e+00
+LLN =1.000e+00          LWN =1.000e+00          WL =0.000e+00
+WW =1.894e-16           WWL =-1.981e-21         WLN =1.000e+00
+WWN =1.040e+00

*       *** Temperature effect parameters ***
+AT = 3.300e+04          UTE = -1.300e+00
+KT1 =-5.403e-01         KT2 =2.200e-02          KT1L = 0.000e+00
+UA1 = 0.000e+00         UB1 =0.000e+00          UC1 = 0.000e+00
+PRT = 0.000e+00

*       *** Overlap capacitance related and dynamic model parameters ***
+CGDO =8.600e-11         CGSO =8.600e-11         CGBO =1.100e-10
+CGDL =1.080e-10         CGSL =1.080e-10         CKAPPA =6.000e-01
+CF = 0.000e+00          ELM = 5.000e+00
+XPART = 1.000e+00       CLC = 1.000e-15         CLE = 6.000e-01

*       *** Parasitic resistance and capacitance related model parameters ***
+RDSW =1.033e+03
+CDSC =2.589e-03         CDSCB =2.943e-04        CDSCD =4.370e-04
+PRWB =-9.731e-02        PRWG =1.477e-01         CIT = 0.000e+00

*       *** Process and parameters extraction related model parameters ***
+TOX=7.754e-09           NGATE=0.000e+00
+NLX=1.770e-07

*       *** Substrate current related model parameters ***
+ALPHA0=0.000e+00        BETA0=3.000e+01

*       *** Noise-effect related model parameters ***
+AF=1.461e+00            KF=1.191e-26            EF=1.000e+00
+NOIA  =5.245e+17        NOIB =4.816e+03         NOIC =8.036e-13
```

```
*       *** Common extrinsic model parameters ***
+LINT =−7.130e−08          WINT =3.4490e−08        XJ =3.000e−07
+RSH =1.290e+02            JS =9.000e−05
+CJ =1.360e−03             CJSW =3.200e−10
+CBD=0.000e+00             CBS =0.000e+00          IS=0.000e+00
+MJ =5.600e−01             N =1.000e+00            MJSW=4.300e−01
+PB =1.020e+00             TT = 0.000e+00
+PBSW=1.020e+00
* ____________________________________________________________________
```

Appendix D
Current sources and current mirrors

Current sources and mirrors are among the most important building blocks of analog CMOS circuits and are being extensively used as DC current sources, DC current mirrors and AC current mirrors. For DC and AC current mirrors, the mirroring factor can be unity, or more than unity; in other words a current mirror can be considered and used as a current amplifier.

In this appendix, first the DC behavior of the DC current sources and current mirrors will be analyzed. Afterwards, the frequency characteristics of AC current mirrors, which have severe effects on the overall performance of many circuits, will be investigated in some detail.

D.1 DC current sources

The simplest MOS current source is a MOS transistor operating in the saturation region, i.e. $V_{\mathrm{DS}} \geq (V_{GS} - \mathrm{V}_T)$, *(Fig. D.1(a)). Here "load" represents the circuit to be fed by this current source. The current is mainly determined by the gate–source voltage* V_{GS}*, but owing to the channel length modulation effect, also has a weak dependence on* V_{DS}*. We know from (1.14b) that the load current can be expressed as*

$$I_{\mathrm{L}} = \frac{1}{2}\mu.C_{\mathrm{ox}}\frac{W}{L}(V_{\mathrm{GS}} - V_{\mathrm{T}})^2\left(\frac{1+\lambda.V_{\mathrm{DS}}}{1+\lambda.(V_{\mathrm{GS}} - V_{\mathrm{T}})}\right) \tag{D.1}$$

In this circuit, owing to the quadratic character of the load current–control voltage relation, the sensitivity is high. Therefore, to obtain a certain desired load current it is necessary to adjust and maintain V_{GS} at an appropriate value with precision.

Another way to bias the current-source transistor is shown in Fig. D.1(b). In this circuit the gate bias of the current-source transistor M2 is determined as the gate–source voltage of the diode connected transistor M1, biased with a constant drain bias current, or "reference current", I_{r}. Figure D.1(c) shows the multiple output version of the circuit, sharing the same reference transistor.

From (1.14b), the drain currents of M1 and M2 in Fig. D.1(b) can be written as

$$I_{\mathrm{r}} = \frac{1}{2}\mu_{\mathrm{n}}C_{\mathrm{ox}}\frac{W_1}{L}(V_{\mathrm{GS}} - V_{\mathrm{T}})^2\left(\frac{1+\lambda V_{\mathrm{GS}}}{1+\lambda(V_{\mathrm{GS}} - V_{\mathrm{T}})}\right) \tag{D.2a}$$

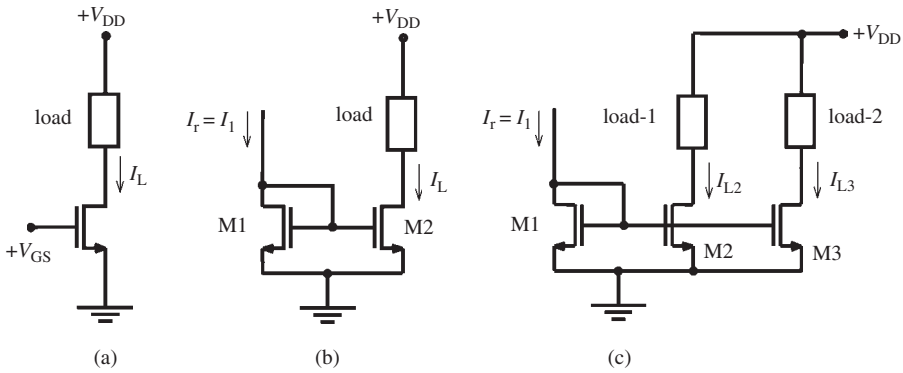


Figure D.1 (a) A MOS transistor as a DC current source. (b) Single output, (c) multiple output basic MOS current mirrors.

$$I_L = \frac{1}{2}\mu_n C_{ox}\frac{W_2}{L}(V_{GS} - V_T)^2\left(\frac{1 + \lambda V_{DS2}}{1 + \lambda(V_{GS} - V_T)}\right) \tag{D.2b}$$

From these expressions I_L can be solved in terms of the reference current I_r:

$$I_L = \frac{W_2}{W_1}\frac{1 + \lambda V_{DS2}}{1 + \lambda V_{GS}} I_r = BI_r \tag{D.3}$$

Expression (D.3) shows that for this circuit the load current is proportional to the reference current. The mirroring coefficient is

$$B = \frac{I_L}{I_r} = \frac{W_2}{W_1}\frac{1 + \lambda.V_{DS2}}{1 + \lambda.V_{GS}} \tag{D.4}$$

and mainly depends on the ratio of the widths. The secondary effects of λ and V_{DS2} on B are negligible only for long-channel-length transistors. The sensitivity of the load current with respect to the variations of V_{DS} (in other words the variations of the DC resistance of the load circuit) can be calculated as

$$\frac{dI_L}{dV_{DS2}} = \frac{W_2}{W_1} I_r \frac{\lambda}{1 + \lambda.V_{GS}} \tag{D.5}$$

which is also equal to the small-signal output conductance of M2, in other words the internal conductance of the current source.

To approach an ideal current source, which has a zero internal conductance, λ must be as small as possible, in other words long-channel transistors must be used, to the expense of the increase of the parasitics. There are several alternative solutions to obtain a very low internal conductance. One of them, the cascode current mirror, is shown in Fig. D.2. Although the low-frequency small-signal output conductance of this circuit is considerably lower than that of the basic circuit given in Fig. D.1(b), it has a severe drawback. In this circuit there is more than one transistor on the load current path and each additional transistor "steals" from the supply voltage budget,

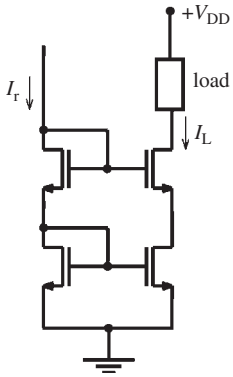


Figure D.2 Cascode current mirror.

which is getting increasingly limited with diminishing supply voltages. Consequently, for small-geometry–low-voltage circuits, using the basic circuit with sufficiently larger channel lengths is more advantageous.

D.2 Frequency characteristics of basic current mirrors

Current mirrors are being extensively used to mirror the AC current of a branch to another branch in many analog circuits. The mirroring coefficient can be unity as in the active loads of a differential pair or can have a value greater than unity, as in some OTA circuits, as investigated in Chapter 3.

The small-signal equivalent circuit of a current mirror with its main parasitics is given in Fig D.3(a). The parasitic capacitances in this circuit are as follows.

- The total gate–source capacitance of M1, which is the sum of the gate capacitance and the gate–source overlap capacitance:

$$C_{gs1} = C_{ox} W_1 L \left(\frac{2}{3} + \frac{CGSO}{C_{ox} L} \right) = k_{ol} C_{ox} W_1 L$$

- The total gate–source capacitance of M2, which is the sum of the gate capacitance and the gate–source overlap capacitance:

$$C_{gs2} = C_{ox} W_2 L \left(\frac{2}{3} + \frac{CGSO}{C_{ox} L} \right) = k_{ol} C_{ox} W_2 L$$

- The gate–drain capacitance of M2:

$$C_{dg2} = CGDO \times W_2 \cong CGSO \times W_2$$

- The total drain junction capacitance of M1:

$$C_{jd1} \cong W_1 X_1 C_j \left[1 + \frac{2}{X_1} \frac{C_{jsw}}{C_j} \right] = W_1 X_1 C'_j$$

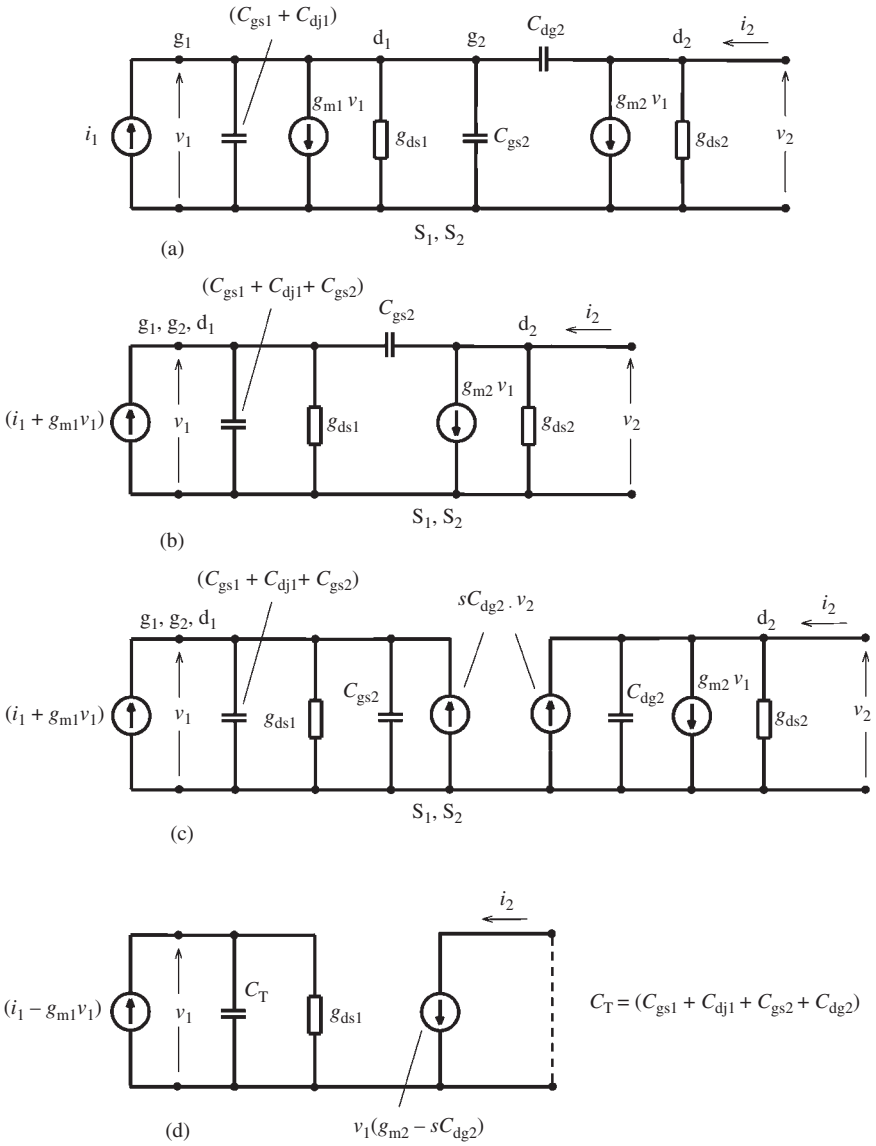


Figure D.3 (a) The small-signal equivalent circuit of the basic current source. (b) and (c) Simplification and application of the modified Miller transformation. (d) Simplified circuit for $|Y_L| \gg |g_{ds2} + j\omega C_{dg2}|$ to calculate B.

where X_1 denotes the second dimension of the drain area, C_j the bottom junction capacitance per unit area and C_{jsw} the side-wall capacitance per unit length (see Chapter 1).

The current transfer ratio of the mirror can be found from Fig. D.3(d):

$$B = \frac{i_2}{i_1} = B_0 \frac{s_p}{s_0} \frac{(s - s_0)}{(s - s_p)} \tag{D.6}$$

where

$$s_0 = +\frac{g_{m2}}{C_{dg2}} \qquad s_p = -\frac{g_{m1} + g_{ds1}}{C_T} \cong -\frac{g_{m1}}{C_T} \tag{D.7}$$

and the low-frequency value of B,

$$B_0 = \frac{g_{m2}}{(g_{m1} + g_{ds1})} \cong \frac{g_{m2}}{g_{m1}} \cong \frac{W_2}{W_1} \tag{D.8}$$

The magnitude of the zero frequency is obviously very high compared to that of the pole frequency, the pole strongly dominates and the 3 dB frequency is determined by the pole:

$$f_{(3dB)} \cong \frac{1}{2\pi} \frac{g_{m1}}{C_T} \tag{D.9}$$

In this expression g_{m1} depends on the geometry, the DC operating conditions and the mode of operation. It is known that the transistors can operate either in the normal saturation region (the pre-velocity saturation region) or in the velocity saturation region, depending on the channel length and the DC operating point. The value of the gate–source capacitance that is the dominant component of C_T is also different for non-velocity saturated and velocity saturated regimes. Therefore the 3 dB frequency of a current mirror must be investigated separately for these two cases.

D.2.1 Frequency characteristics for normal saturation

If the transistors of the current mirror are operating in the pre-velocity saturation mode, the transconductance of the reference transistor and the total capacitance of the input node are

$$g_{m1} \cong \mu C_{ox} \frac{W_1}{L} (V_{GS} - V_T)$$

and

$$\begin{aligned} C_T &= C_{gs1} + C_{gs2} + C_{dg2} + C_{dj1} \\ &= k_{ol} C_{ox} L W_1 + k_{ol} C_{ox} L W_2 + CGDO \cdot W_2 + X_1 C_j' W_1 \end{aligned}$$

which can be arranged as

$$C_T = W_1 \left[k_{ol} C_{ox} L (1 + B_0) + B_0 \cdot CGDO + X_1 C_j' \right] \tag{D.10}$$

The 3 dB frequency of the gain function can be arranged assuming $g_{m1} \gg g_{dg1}$ and $C_{GSO} = C_{GDO}$, which is reasonable for most cases, as

$$f_{(3dB)} = \frac{\mu(V_{GS} - V_T)}{\left[k_{ol}L^2(1 + B_0) + B_0 \frac{L \cdot CGDO}{C_{ox}} + LX_1 \frac{C'_j}{C_{ox}}\right]}$$

From this expression it is possible to examine the effects of any one of the parameters, and to reach valuable design hints. It can be easily seen that the 3 dB frequency

- decreases with B_0 (or with the number of outputs),
- strongly decreases with the channel length,
- is independent of the gate width,
- increases with mobility (NMOS is better),
- increases with the gate overdrive, or the DC current of the source transistor.

The parameters in these expressions are directly given in the model parameter lists, except μ. Especially for gate oxide thicknesses smaller than 10 nm, the value of μ must be calculated in terms of the gate voltage (see Chapter 1). The C_j and C_{jsw} values given in the model parameter lists correspond to zero junction voltage. The actual values of these capacitances are smaller owing to the reverse bias of the junction.

D.2.2 Frequency characteristics under velocity saturation

We know that for small gate lengths – and especially for NMOS transistors – the operating mode must be checked to see if the transistor is in the velocity saturation mode, or not. If the transistor is operating in the velocity saturation mode, the transconductance must be calculated from (1.34) as $g_m = C_{ox}Wv_{sat}$. In addition, the gate–source capacitance of a velocity saturated transistor is $C_{gs} = C_{ox}WL + (CGSO \times W)$, as derived in Problem 1.1.

Using these parameters the 3 dB frequency can be found as

$$f_{(3dB)} = \frac{kv_{sat}}{\left[L(1 + B_0) + B_0 \frac{L \cdot CGDO}{C_{ox}} + X_1 \frac{C'_j}{C_{ox}}\right]}$$

This expression can be interpreted as follows.

- The 3 dB frequency of a transistor does not change with the operating point, provided that the transistor remains in velocity saturation.
- The gate-length dependence of the 3 dB frequency of a velocity saturated transistor is not as severe as that of a non-velocity saturated transistor.
- The 3 dB frequencies of NMOS and PMOS transistors are approximately equal, provided that both are in velocity saturation.

References

[1] S. Wolf, *Silicon Processing for the VLSI Era,* vol. 3 – *The Submicron MOSFET*, Lattice Press, 1995.

[2] N. Shigyo, T. Shimane, M. Suda, T. Enda, and S. Fukuda, "Verification of saturation velocity lowering in MOSFET's inversion layer", *IEEE Trans. on Electron Devices*, vol. **45**, no. 2, pp. 460–464, Feb. 1998.

[3] K. Chen, J. Duster, H. C. Wann, P. K. Ko, and C. Hu, "MOSFET carrier mobility model based on gate oxide thickness, threshold and gate voltages", *J. Solid State Electron*., vol. **39**, pp. 1515–1518, Oct. 1996.

[4] K. Chen and C. Hu, "Performance and Vdd scaling in deep submicrometer CMOS", *IEEE J. of Solid-State Circuits*, vol. **33**, pp. 1586–1589, Oct. 1998.

[5] R. S. Muller, T. I. Kamins, and M. Chan, *Device Electronics for Integrated Circuits*, 3rd. ed., Wiley, 2003.

[6] S. Wolf, *Silicon Processing for the VLSI Era,* vol. 3 – *The Submicron MOSFET*, Lattice Press, 1995.

[7] Y. P. Tsividis, *Operation and Modeling of the MOS Transistor*, 2nd. ed., McGraw-Hill, 1999.

[8] U. Cilingiroglu, *Systematic Analysis of Bipolar and MOS Transistors*, Artech House, 1993.

[9] K. K. Ng and W. T. Lynch, "Analysis of the gate-voltage dependent series resistance of MOSFETs", *IEEE Trans. on Electr. Dev.*, vol. ED-**33**, p. 965, July 1986.

[10] K. K. Ng and W. T. Lynch, "The impact of the intrinsic series resistance on MOSFET scaling", *IEEE Trans. on Electr. Dev.*, vol. ED-**34**, p. 503, March 1987.

[11] E. Gondro, "An improved bias dependent series resistance description for MOS models", Compact Model Council Meeting, Sauta-Clara, USA, May 1998.

[12] D. Foty, *MOSFET Modeling with SPICE, Principles and Practice*, Prentice-Hall, 1997.

[13] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge University Press, 1998.

[14] A. B. Glaser and G. E. Subak-Sharpe, *Integrated Circuit Engineering; Design, Fabrication and Applications*, Addison-Wesley, 1979.

[15] H. Samavati, A. Hajimiri, A. R. Shahani, G. N. Nasserbakht, and T. H. Lee, "Fractal capacitors", *IEEE Journal of Solid State Circuits*, vol. **33**, no. 12, pp. 2035–2041, Dec. 1998.

[16] O. E. Akcasu, "High capacitance structure in a semiconductor device", U.S. Patent No. 5,208,725, May 4, 1993.

[17] R. L. Bunch and S. Raman, "Large signal analysis of MOS varactors in CMOS-GmLC VCOs", *IEEE Journal of Solid State Circuits*, vol. **38**, no. 8, pp. 1325–1332, Aug. 2003.

[18] P. Andreani and S. Mattisson, "On the use of MOS varactors in RF VCOs", *IEEE Journal of Solid State Circuits*, vol. **35**, no. 6, pp. 905–910, June 2000.

[19] S. S. Mohan, M. M. Hershenson, S. P. Boyd, and T. H. Lee, "Simple accurate expressions for planar spiral inductances", *IEEE Journal of Solid State Circuits*, vol. **34**, no. 10, pp. 1419–1424, Oct. 1999

[20] "ASITIC": Analysis and Simulation of Spiral Inductors and Transformers for ICs. EECE, University of California, Berkeley.

[21] "SPIRAL", OEA International, Inc., Morgan Hill, CA, USA.

[22] K. B. Ashby, I. A. Koullias, W. C. Finley, J. J. Bustek, and S. Moinian, "High Q inductors for wireless applications in complementary silicon bipolar process", *IEEE Journal of Solid State Circuits*, vol. **31**, no. 1, pp. 4–9, Jan. 1996.

[23] C. P. Yue and S. S. Wong, "On-chip spiral inductors with patterned ground shields for Si-based RFICs", *IEEE Journal of Solid State Circuits*, vol. **33**, no. 5, pp. 743–752, May 1998.

[24] S. M. Kang and Y. Leblebici, *CMOS Digital Integrated Circuits, Analysis and Design*, 3rd ed., McGraw-Hill, 2003.

[25] A. B. Grebene, *Bipolar and MOS Analog Integrated Design*, John Wiley & Sons, 1984.

[26] J. M. Miller, "Dependence of the input impedance of a three-electrode vacuum tube upon the load in the plate circuit", *Natl. Bur. Std. (U.S.) Research Papers*, vol. **15**, pp. 367–385, 1919.

[27] R. W. Hickman and F. V. Hunt, "On electronic voltage stabilizers", *Review of Scientific Instruments*, pp. 6–21, Jan. 1939.

[28] Y. P. Tsividis, *Operation and Modeling of the MOS Transistor*, McGraw-Hill, 1987.

[29] W. S. Percival, "Thermionic valve circuits", British Patent Spec., no. 460,562, Granted Jan. 1937.

[30] D. Leblebici, "The distributed amplifier using transistors as delay elements", Ph.D. Thesis, Istanbul Technical University, 1966.

[31] R. C. Liu, K. L. Deng, and H. Wang, "A 0.6–22 GHz broadband CMOS distributed amplifier", *IEEE Radio Frequency Integrated Circuit Symposium*, pp. 103–106, June 2003.

[32] Y. Ayasli, R. L. Mozzi, J. L. Vorhaus, L. D. Reynolds, and R. A. Pucel, "A monolithic GaAs 1–13-GHz traveling-wave amplifier," *IEEE Trans. Microwave Theory and Techniques*, vol. MTT-**30**, no. 7, pp. 976–981, July 1982.

[33] E. M. Cherry and D. E. Hooper, "The design of wide-band transistor feedback amplifiers", *Proc. IEE*, pp. 375–389, Feb. 1963.

[34] R. P. Feynman, *Lectures on Physics*, vol. 1, pp. 23–1 to 23–9, Addison Wesley Publishing Co., 1963.

[35] W. B. Kuhn, F. W. Stephenson, and A. Elshabini-Riad, "A 200 MHz CMOS Q-enhanced LC bandpass filter", *IEEE J. Solid State Circ.*, vol. **31**, no. 8, pp. 1112–1122, Aug. 1996.

[36] T. Soorapanth and S. S. Wong, "A 0-dB IL 2140 $\pm$ 30 MHz bandpass filter utilizing Q-enhanced spiral inductors in standard CMOS", *IEEE J. Solid State Circ.*, vol. **37**, no. 5, pp. 579–586, May 2002.

[37] M. E. Van Valkenburg, *Analog Filter Design* (The Oxford Series in Electrical & Computer Engineering), Oxford University Press Inc., USA, 1996.

[38] *Reference Data for Radio Engineers*, 4th ed., International Telephone and Telegraph Corp., 1956.

[39] B. D. H. Tellegen, "The gyrator, a new electric network element", Philips Research Rept., vol. 3, no. 2, pp. 81–101, 1948.

[40] L. W. Huelsman, *Active Filters: Lumped, Distributed, Integrated, Digital and Parametric*, McGraw-Hill, 1970.

[41] A. G. J. Holt and J. Taylor, "Method of replacing ungrounded inductances by grounded gyrators", *Electron. Letters*, vol. **1**, p. 105, 1965.

[42] Z. Du, K. Gong, J. S. Fu, B. Gao, and Z. Feng, "A compact planar inverted-F antenna with a PBG-type ground plane for mobile communications", *IEEE Transactions on Vehicular Technology*, vol. **52**, no. 3, pp. 483–489, May 2003.
[43] A. van der Ziel, *Noise in Solid State Devices and Circuits*, John Wiley and Sons, 1986.
[44] R. P. Jindal, "Noise associated with distributed resistance of gate structures in integrated circuits", *IEEE Trans. on Electron Devices*, vol. **31**, no. 10, pp. 1505–1509, 1984.
[45] A. J. Scholten, L. F. Tiemeijer, and R. van Langevelde, "Noise modeling for RF CMOS circuit simulation", Invited Paper, *IEEE Trans. on Electron Devices*, vol. **50**, no. 3, pp. 618–632, 2003.
[46] G. Knoblinger, P. Klein, and M. Tiebout, "A new model for thermal channel noise of deep submicron MOSFETs and its application in RF-CMOS design", *IEEE J. Solid State Circuits*, vol. **36**, no. 5, pp. 831–837, May 2001.
[47] S. Asgaran and M. J. Deen, "Analytical modeling of MOSFET's channel noise and noise parameters", *IEEE Trans. on Electron Devices*, vol. **51**, no. 12, pp. 2109–2114, Dec. 2004.
[48] Z. Q. Lu and Y. Z. Ye, "A simple model for channel noise of deep submicron MOSFETs", *IEEE Conference on Electron Devices and Solid-State Circuits*, pp. 309–312, 2005.
[49] C. H. Chen and M. J. Deen, "Channel noise modeling of deep submicron MOSFETs", *IEEE Trans. on Electron Devices*, vol. **49**, no. 8, pp. 1484–1487, Aug. 2002.
[50] M. J. Deen, C. H. Chen, S. Asgaran, G. A. Rezvani, J. Tao, and Y. Kiyota, "High frequency noise of modern MOSFETs: compact modeling and measurement issues", Invited Paper, *IEEE Trans. on Electron Devices*, vol. **53**, no. 9, pp. 2062–2081, Sept. 2006.
[51] V. Szekely, M. Rencz, and B. Courtois, "Tracing the thermal behavior of ICs", *IEEE Design and Test of Computers*, vol. **15**, no. 2, pp. 14–21, 1998.
[52] K. Shadron, M. Stan, M. Barcella, *et al.*, "Hotspot: techniques for modeling thermal effects at the processor-architecture level", THERMINIC Workshop, Madrid, 1–4 Oct. 2002.
[53] Y. Kim, H. Morishita, Y. Koyanagi, and K. Fujimoto, "A folded loop antenna system for handsets developed and based on the advanced design concept", *IEICE Trans. Commun.*, vol. **E84**-B, no. 9, Sept. 2001.
[54] E. S. Yang, *Fundamentals of Semiconductor Devices*, McGraw-Hill, pp. 100–103, 1978.
[55] J. Groszkowski, "The interdependence of frequency variation and harmonic content, and the problem of constant-frequency oscillators", *Proc. IRE*, vol. **21**, no. 7, pp. 958–981, 1934.
[56] A. B. Grebene, *Bipolar and MOS Analog Integrated Circuits*, John Wiley & Sons, pp. 627–652, 1984.
[57] H. B. Dwight, *Mathematical Tables*, Dover Books, pp. 144–167, 1958.
[58] A. Hajimiri and T. H. Lee, "A general theory of phase noise in electrical oscillators", *IEEE J. of Solid State Circuits*, vol. **33**, no. 2, pp. 179–194, Feb. 1997.
[59] A. Demir, A. Mehrorta, and J. Roychowdhury, "Phase noise in oscillators: a unifying theory and numerical methods for characterization", *IEEE Trans. on Circ. and Sys. – I, Fundamental Theory and Applications*, vol. **47**, no. 5, pp. 655–674, May 2000.
[60] P. Andreani and H. Sjöland, "Tail current noise suppression in RF CMOS VCOs", *IEEE J. of Solid State Circuits*, vol. **37**, no. 3, pp. 342–348, March 2002.
[61] F. Maloberti, *Data Converters*, Springer, 2007.
[62] B. Razavi, *Principles of Data Conversion System Design*, Wiley-IEEE Press, 1994.
[63] R. J. van de Plassche, *CMOS Integrated Analog-to-Digital and Digital-to-Analog Converters*, Springer, 2007.
[64] *Data Conversion Handbook (Analog Devices)*, Engineering Staff Analog Devices Inc., Newnes, 2004.

Index