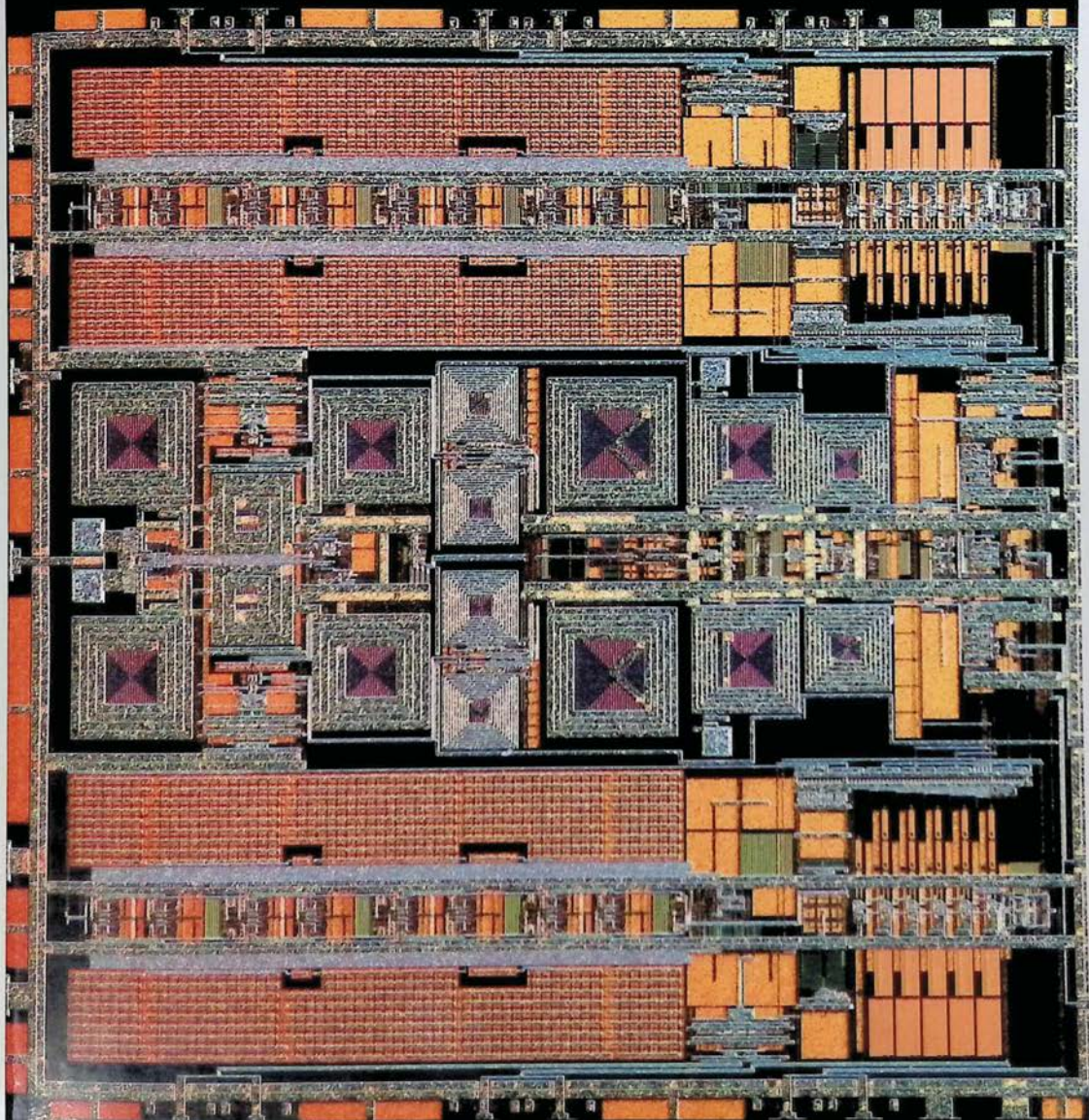


The Design of CMOS Radio-Frequency Integrated Circuits



Thomas H. Lee

THE DESIGN OF CMOS RADIO-FREQUENCY INTEGRATED CIRCUITS

THOMAS H. LEE

Stanford University



CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

<http://www.cambridge.org>

© Cambridge University Press 1998

This book is in copyright. Subject to statutory exception and
to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without
the written permission of Cambridge University Press.

First published 1998
Reprinted with corrections 1998
Reprinted 1999, 2000, 2001, 2002

Printed in the United States of America

Typeset in Times using AMS- $\text{\TeX}$

Library of Congress Cataloging-in-Publication Data

Lee, Thomas H., 1959–

The design of CMOS radio-frequency integrated circuits / Thomas H. Lee.

p. cm.

ISBN 0-521-63061-4 (hb) – ISBN 0-521-63922-0 (pb)

1. Metal oxide semiconductors, Complementary – Design and
construction. 2. Integrated circuits – Design and construction.
3. Radio – Transmitter-receivers. I. Title.

TK7871.99.M44L44 1998

621.39'732 – dc21

97-34158

CIP

A catalog record for this book is available from the British Library

ISBN 0 521 63061 4 hardback
ISBN 0 521 63922 0 paperback

CONTENTS

<i>Preface</i>	<i>page</i> xiii
1 A NONLINEAR HISTORY OF RADIO	1
1. Introduction	1
2. Maxwell and Hertz	1
3. Pre-Vacuum Tube Electronics	2
4. Birth of the Vacuum Tube	9
5. Armstrong and the Regenerative Amplifier/Detector/Oscillator	13
6. Other Radio Circuits	15
7. Armstrong and the Superregenerator	18
8. Oleg Losev and the First Solid-State Amplifier	20
9. Epilog	21
10. Appendix: A Vacuum Tube Primer	22
2 CHARACTERISTICS OF PASSIVE IC COMPONENTS	34
1. Introduction	34
2. Resistors	34
3. Capacitors	37
4. Inductors	47
5. Summary	57
6. Appendix: Summary of Capacitance Equations	57
Problem Set	58
3 A REVIEW OF MOS DEVICE PHYSICS	62
1. Introduction	62
2. A Little History	62
3. FETs: The Short Story	63
4. MOSFET Physics: The Long-Channel Approximation	64
5. Operation in Weak Inversion (Subthreshold)	72

6. MOS Device Physics in the Short-Channel Regime	73
7. Other Effects	77
8. Transit Time Effects	79
9. Summary	80
10. Appendix: 0.5- μm Level-3 SPICE Models	80
Problem Set	81
4 PASSIVE RLC NETWORKS	86
1. Introduction	86
2. Parallel <i>RLC</i> Tank	86
3. Series <i>RLC</i> Networks	91
4. Other Resonant <i>RLC</i> Networks	91
5. <i>RLC</i> Networks as Impedance Transformers	93
6. Examples	104
Problem Set	106
5 DISTRIBUTED SYSTEMS	114
1. Introduction	114
2. Link Between Lumped and Distributed Regimes	116
3. Driving-Point Impedance of Iterated Structures	117
4. Transmission Lines in More Detail	119
5. Behavior of Finite-Length Transmission Lines	124
6. Summary of Transmission-Line Equations	127
7. Artificial Lines	127
8. Summary	131
Problem Set	131
6 THE SMITH CHART AND S-PARAMETERS	134
1. Introduction	134
2. The Smith Chart	134
3. S-Parameters	138
4. Appendix: A Short Note on Units	140
5. Appendix: Why 50 (or 75) Ω ?	141
Problem Set	144
7 BANDWIDTH ESTIMATION TECHNIQUES	146
1. Introduction	146
2. The Method of Open-Circuit Time Constants	147
3. The Method of Short-Circuit Time Constants	161
4. Further Reading	166
5. Risetime, Delay, and Bandwidth	167
6. Summary	174
Problem Set	174

8 HIGH-FREQUENCY AMPLIFIER DESIGN	178
1. Introduction	178
2. Zeros as Bandwidth Enhancers	179
3. The Shunt-Series Amplifier	191
4. Bandwidth Enhancement with f_T Doublers	197
5. Tuned Amplifiers	199
6. Neutralization and Unilateralization	203
7. Cascaded Amplifiers	206
8. Summary	216
Problem Set	217
9 VOLTAGE REFERENCES AND BIASING	223
1. Introduction	223
2. Review of Diode Behavior	223
3. Diodes and Bipolar Transistors in CMOS Technology	225
4. Supply-Independent Bias Circuits	225
5. Bandgap Voltage Reference	227
6. Constant- g_m Bias	235
7. Summary	237
Problem Set	237
10 NOISE	243
1. Introduction	243
2. Thermal Noise	243
3. Shot Noise	250
4. Flicker Noise	252
5. Popcorn Noise	255
6. Classical Two-Port Noise Theory	256
7. Examples of Noise Calculations	260
8. A Handy Rule of Thumb	263
9. Typical Noise Performance	264
10. Appendix: Noise Models	265
Problem Set	266
11 LNA DESIGN	272
1. Introduction	272
2. Derivation of MOSFET Two-Port Noise Parameters	273
3. LNA Topologies: Power Match versus Noise Match	277
4. Power-Constrained Noise Optimization	284
5. Design Examples	288
6. Linearity and Large-Signal Performance	295
7. Spurious-Free Dynamic Range	302

	303
8. Summary	304
Problem Set	308
12 MIXERS	308
1. Introduction	309
2. Mixer Fundamentals	314
3. Nonlinear Systems as Linear Mixers	318
4. Multiplier-Based Mixers	335
5. Subsampling Mixers	337
6. Appendix: Diode-Ring Mixers	340
Problem Set	344
13 RF POWER AMPLIFIERS	344
1. Introduction	344
2. General Considerations	345
3. Class A, AB, B, and C Power Amplifiers	355
4. Class D Amplifiers	357
5. Class E Amplifiers	359
6. Class F Amplifiers	362
7. Modulation of Power Amplifiers	364
8. Summary of PA Characteristics	365
9. RF PA Design Examples	372
10. Additional Design Considerations	379
11. Design Summary	380
Problem Set	385
14 FEEDBACK SYSTEMS	385
1. Introduction	385
2. A Brief History of Modern Feedback	390
3. A Puzzle	391
4. Desensitivity of Negative Feedback Systems	395
5. Stability of Feedback Systems	396
6. Gain and Phase Margin as Stability Measures	398
7. Root-Locus Techniques	404
8. Summary of Stability Criteria	404
9. Modeling Feedback Systems	408
10. Errors in Feedback Systems	410
11. Frequency- and Time-Domain Characteristics of First- and Second-Order Systems	414
12. Useful Rules of Thumb	415
13. Root-Locus Examples and Compensation	422
14. Summary of Root-Locus Techniques	423
15. Compensation	

16. Compensation through Gain Reduction	423
17. Lag Compensation	426
18. Lead Compensation	429
19. Summary of Compensation	432
Problem Set	432
15 PHASE-LOCKED LOOPS	438
1. Introduction	438
2. A Short History of PLLs	438
3. Linearized PLL Models	441
4. Some Noise Properties of PLLs	447
5. Phase Detectors	450
6. Sequential Phase Detectors	455
7. Loop Filters and Charge Pumps	463
8. PLL Design Examples	470
9. Summary	478
Problem Set	478
16 OSCILLATORS AND SYNTHESIZERS	484
1. Introduction	484
2. The Problem with Purely Linear Oscillators	484
3. Describing Functions	485
4. Resonators	500
5. A Catalog of Tuned Oscillators	504
6. Negative Resistance Oscillators	509
7. Frequency Synthesis	514
8. Summary	524
Problem Set	525
17 PHASE NOISE	530
1. Introduction	530
2. General Considerations	530
3. Detailed Considerations: Phase Noise	532
4. The Hajimiri Model: A Time-Varying Phase Noise Theory	536
5. Summary	544
Problem Set	545
18 ARCHITECTURES	550
1. Introduction	550
2. Dynamic Range	551
3. Subsampling	565
4. Transmitter Architectures	566
5. Oscillator Stability	567

6. Summary	568
Problem Set	568
19 RF CIRCUITS THROUGH THE AGES	571
1. Introduction	571
2. Armstrong	571
3. The "All-American" 5-Tube Superhet	575
4. The Regency TR-1 Transistor Radio	578
5. Three-Transistor Toy CB Walkie-Talkie	580

CHAPTER TWELVE

MIXERS

12.1 INTRODUCTION

Most circuit analysis proceeds with the assumptions of linearity and time invariance. Violations of those assumptions, if considered at all, are usually treated as undesirable. However, the high performance of modern communications equipment actually depends critically on the presence of at least one element that fails to satisfy linear time invariance. As noted in Chapter 1, the superheterodyne¹ receiver uses a mixer to perform an important frequency translation of signals. Armstrong's invention has been the dominant architecture for 70 years because this frequency translation solves many problems in one fell swoop (see Figure 12.1).²

In this architecture, the mixer translates an incoming RF signal to a lower frequency,³ known as the intermediate frequency (IF). Although Armstrong originally sought this frequency lowering simply to make it easier to obtain the requisite gain, other significant advantages accrue as well. As one example, tuning is now accomplished by varying the frequency of a local oscillator, rather than by varying the center frequency of a multipole bandpass filter. Thus, instead of adjusting several *LC* networks in tandem to tune to a desired signal, one simply varies a single *LC* combination to change the frequency of a local oscillator (LO). The intermediate frequency stages can then use *fixed* bandpass filters. Selectivity is therefore determined

¹ Why "super" heterodyne? The reason is that Fessenden had already invented something called the "heterodyne," and Armstrong had to distinguish his invention from Fessenden's.

² Proving once again that success has many fathers (while failure is an orphan, to complete J. F. Kennedy's saying), Lucien Levy and Walter Schottky, among others, laid claim to the superheterodyne. While it is certainly true that Armstrong was not the first to conceive of the heterodyne principle, he was the first to recognize how neatly it solves so many thorny problems, and was certainly the first to pursue its development vigorously.

³ Actually, one may also translate to a *higher* frequency, but we will defer a discussion of that case to Chapter 18.

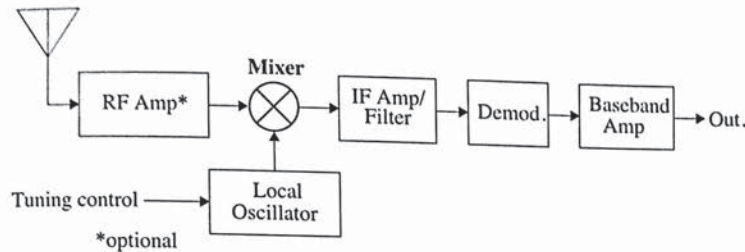


FIGURE 12.1. Superheterodyne receiver block diagram.

by these fixed-frequency IF filters, which are much easier to realize than variable-frequency filters. Additionally, the overall gain of the system is distributed over a number of different frequency bands (RF, IF, and baseband), so that the required receiver gain (typically 120–140 dB on a power basis) can be obtained without much worry about potential oscillations arising from parasitic feedback loops. These important attributes explain why the superheterodyne architecture still dominates, over 70 years after its invention.

12.2 MIXER FUNDAMENTALS

Since linear, time-invariant systems cannot produce outputs with spectral components not present at the input, mixers must be either nonlinear or time-varying elements in order to provide frequency translation. Historically, many devices (e.g., electrolytic cells, magnetic ribbons, brain tissue, and rusty scissors – in addition to more traditional devices such as vacuum tubes and transistors) operating on a host of diverse principles have been used, demonstrating that virtually *any* nonlinear element can be used as a mixer.⁴

At the core of all mixers presently in use is a multiplication of two signals in the time domain. The fundamental usefulness of multiplication may be understood from examination of the following trigonometric identity:

$$(A \cos \omega_1 t)(B \cos \omega_2 t) = \frac{AB}{2} [\cos(\omega_1 - \omega_2)t + \cos(\omega_1 + \omega_2)t]. \quad (1)$$

Multiplication thus results in output signals at the sum and difference frequencies of the input, signals whose amplitudes are proportional to the product of the RF and LO amplitudes. Hence, if the LO amplitude is constant (as it usually is), any amplitude

⁴ Of course, some nonlinearities work better than others, so we will focus on the more practical types.

modulation in the RF signal is transferred to the IF signal. By a similar mechanism, an undesired transfer of modulation from one signal to another can also occur through nonlinear interaction in both mixers and amplifiers. In that context the result is called *cross-modulation*, as mentioned in Chapter 11, and its suppression through improved linearity is an important design consideration.

Having recognized the fundamental role of multiplication, we now enumerate and define the most significant characteristics of mixers.

12.2.1 CONVERSION GAIN

One important mixer characteristic is conversion gain (or loss), which is defined as the ratio of the desired IF output to the value of the RF input. For the multiplier described by Eqn. 1, the conversion gain is therefore the IF output, $AB/2$, divided by A (if that is the amplitude of the RF input). Hence, the conversion gain in this example is $B/2$, or half the LO amplitude.

Conversion gain, if expressed as a power ratio, can be greater than unity in active mixers; passive mixers are generally capable only of voltage or current gain at best.⁵ Conversion gain in excess of unity is often convenient since the mixer then provides amplification along with the frequency translation. However, it does not necessarily follow that sensitivity improves, since noise figure must also be considered. For this reason, passive mixers may offer superior performance in some cases despite their conversion loss.

12.2.2 NOISE FIGURE: SSB VERSUS DSB

Noise figure is defined as one might expect: it's the signal-to-noise ratio (SNR) at the input (RF) port divided by the SNR at the output (IF) port. There's an important subtlety, however, that often trips up both the uninitiated and a substantial fraction of practicing engineers. To appreciate this difficulty, we first need to make an important observation: In a typical mixer, there are actually *two* input frequencies that will generate a given intermediate frequency. One is the desired RF signal, and the other is called the *image* signal. In the context of mixers, these two signals are frequently referred to collectively as *sidebands*.

The reason that two such frequencies exist is that the IF is simply the *magnitude* of the difference between the RF and LO frequencies. Hence, signals both above and below ω_{LO} by an amount equal to the IF will produce IF outputs of the same frequency. The two input frequencies are therefore separated by $2\omega_{IF}$. As a specific

⁵ An exception is a class of systems known as parametric converters or parametric amplifiers, in which power from the LO is transferred to the IF through reactive nonlinear interaction (typically with varactors), thus making power gain possible.

numerical example, suppose that our system's IF is 100 MHz and we wish to tune to a signal at 900 MHz by selecting an LO frequency of 1 GHz. Aside from the desired 900-MHz RF input, a 1.1-GHz image signal will also produce a difference frequency component at the IF of 100 MHz.

The existence of an image frequency complicates noise figure computations because noise originating in both the desired and image frequencies therefore becomes IF noise, yet there is generally no desired signal at the image frequency. In the usual case where the desired signal exists at only one frequency, the noise figure that one measures is called the *single-sideband* noise figure (SSB NF); the rarer case, where both the "main" RF and image signals contain useful information, leads to a double-sideband (DSB) noise figure.

Clearly, the SSB noise figure will be greater than for the DSB case, since both have the same IF noise but the former has signal power in only a single sideband. Hence, the SSB NF will normally be 3 dB higher than the DSB NF.⁶ Unfortunately, DSB NF is reported much more often because it is numerically smaller and thus falsely conveys the impression of better performance, even though there are few communications systems for which DSB NF is an appropriate figure of merit.⁷ Frequently, a noise figure is stated without any indication as to whether it is a DSB or SSB value. In such cases, one may usually assume that a DSB figure is being quoted.

Noise figures for mixers tend to be considerably higher than those for amplifiers because noise from frequencies other than at the desired RF can mix down to the IF. Representative values for SSB noise figures range from 10 dB to 15 dB or more. It is mainly because of this larger mixer noise that one uses LNAs in a receiver. If the LNA has sufficient gain then the signal will be amplified to levels well above the noise of the mixer and subsequent stages, so the overall receiver NF will be dominated by the LNA instead of the mixer. If mixers were not as noisy as they are, then the need for LNAs would diminish considerably. We will return to this theme in the chapter on receiver architectures (Chapter 18).

12.2.3 LINEARITY AND ISOLATION

Dynamic range requirements in modern, high-performance telecommunications systems are quite severe, frequently exceeding 80 dB and approaching 100 dB in many instances. As discussed in the previous chapter, the floor is established by the noise figure, which conveys something about how small a signal may be processed, whereas

⁶ This 3-dB difference assumes that the conversion gain to two equal sidebands is the same. Although this assumption is usually well satisfied, it need not be.

⁷ Two important exceptions in which both sidebands contain useful information are radio astronomy (as in the measurements of the echoes of the Big Bang) and direct-conversion receivers (see Chapter 18).

the ceiling is set by the onset of severe nonlinearities that accompany large input signals.

As with amplifiers, the compression point is one measure of this dynamic range ceiling, and is defined the same way. Ideally, we would like the IF output to be proportional to the RF input signal amplitude; this is the sense in which we interpret the term “linearity” in the context of mixers. However, as with amplifiers (and virtually any other physical system), real mixers have some limit beyond which the output has a sublinear dependence on the input. The compression point is the value of RF signal⁸ at which a calibrated departure from the ideal linear curve occurs. Usually, a 1-dB (or, more rarely, a 3-dB) compression value is specified. One may specify either the input or output signal strength at which this compression occurs, together with the conversion gain, to allow fair comparisons among different mixers.

The two-tone third-order intercept is also used to characterize mixer linearity. A two-tone intermodulation test is a relevant way to evaluate mixer performance because it mimics the real-world scenario in which both a desired signal and a potential interferer (perhaps at a frequency just one channel away) feed a mixer input. Ideally, each of two superposed RF inputs would be translated in frequency without interacting with each other. Of course, practical mixers will always exhibit some intermodulation effects, and the output of the mixer will thus contain frequency-translated versions of third-order IM components whose frequencies are $2\omega_{RF1} \pm \omega_{RF2}$ and $2\omega_{RF2} \pm \omega_{RF1}$. The difference frequency terms may heterodyne into components that lie within the IF passband and are therefore generally the troublesome ones, while the sum frequency signals can usually be filtered out.

As a measure of the degree of departure from linear mixing behavior, one can plot the desired output and the third-order IM output as a function of input RF level. The third-order intercept is the extrapolated intersection of these two curves. In general, the higher the intercept, the more linear the mixer. Again, one ought to specify whether the intercept is input- or output-referred, as well as the conversion gain, to permit fair comparisons among mixers. Additionally, it is customary to abbreviate the intercept as IP3, or perhaps IIP3 or OIP3 (for input and output third-order intercept point, respectively). These definitions are summarized in Figure 12.2.

Cubic nonlinearity can also cause trouble with a *single* RF input. As a specific example, consider building a low-cost AM radio. The standard IF for AM radios happens, unfortunately, to be 455 kHz (mainly for historical reasons). Tuning in a

⁸ Some manufacturers (and authors) report an *output* compression point. If the conversion gain at that point is known then the figure can be reflected back to the input point. Sadly, many insist on burying that bit of information, making it extremely difficult to perform fair comparisons of mixer performance. We will always state explicitly whether the figure is an input or output parameter.

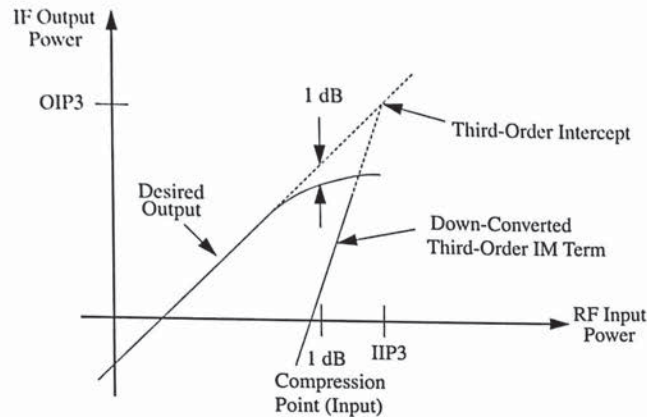


FIGURE 12.2. Definition of mixer linearity parameters.

station at 910 kHz (a legitimate AM radio frequency) requires that the LO be set to 1365 kHz.⁹ The cubic nonlinearity could generate a component at $2\omega_{\text{RF}} - \omega_{\text{LO}}$, which in this case happens to coincide with our IF of 455 kHz.

One might be tempted to assert that such a component is not a problem because it adds to the desired output. One therefore might even be tempted to consider this an asset. However, the third-order IM products have amplitudes that are no longer proportional to the input signal amplitude. Hence, they represent amplitude distortion that can corrupt the “correct” output (we’re talking about an amplitude-modulated signal, after all).

Even if the exact numerological coincidence of the foregoing example does not occur, various third-order IM terms can possess frequencies within the passband of the IF amplifier, ultimately degrading signal-to-noise or signal-to-distortion ratios.

Another parameter of great practical importance is isolation. It is generally desirable to minimize interaction among the RF, IF, and LO ports. For instance, since the LO signal power is generally quite large compared with that of the RF signal, any LO feedthrough to the IF output might cause problems at subsequent stages in the signal processing chain. This problem is exacerbated if the IF and LO frequencies are similar, so that filtering is ineffective. Even reverse isolation is important in many instances, since poor reverse isolation might permit the strong LO signal (or its harmonics) to work its way back to the antenna, where it can radiate and cause interference to other receivers.

⁹ A local oscillator frequency of 455 kHz also works, but is a less practical choice because such “low-side injection” requires the local oscillator to tune over a larger range than if the LO frequencies were above the desired RF.

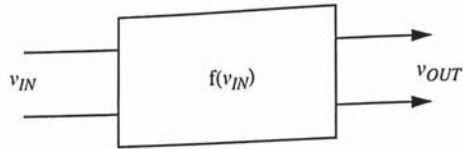


FIGURE 12.3. General two-port nonlinearity.

12.3 NONLINEAR SYSTEMS AS LINEAR MIXERS

We now consider how to implement the multiplication that is the heart of mixing action. Some mixers directly implement a multiplication, while others provide it incidentally through a nonlinearity. We follow an historical path and first examine a general two-port nonlinearity,¹⁰ since mixers of that type predate those designed specifically to behave as multipliers. See Figure 12.3.

If the nonlinearity is “well-behaved” (in the mathematical sense), we can describe the input–output relationship with a series expansion:

$$v_{OUT} = \sum_{n=0}^N c_n (v_{IN})^n. \quad (2)$$

Using an N th-order nonlinearity as a mixer requires the signal v_{IN} to be the sum of the RF input and the local oscillator signals. In general, the output will consist of three types of products: DC terms, harmonics of the inputs, and intermodulation products of those harmonics.¹¹ Not all of these spectral components are desirable, so part of the challenge in mixer design is to devise topologies that inherently generate few undesired terms.

Even-order nonlinear factors in Eqn. 2 contribute DC terms; these are readily filtered out by AC coupling, if desired. Harmonic terms, at $m\omega_{LO}$ and $m\omega_{RF}$, extend from the fundamental ($m = 1$) all the way up to the N th harmonic. As with the DC terms, they are also often relatively easy to filter out because their frequencies are usually well away from the desired IF.

The intermodulation (IM) products are the various sum and difference frequency terms. These have frequencies expressible as $p\omega_{RF} \pm q\omega_{LO}$, where integers p and q are greater than zero and sum to values up to N . Only the second-order intermodulation term ($p = q = 1$) is normally desired.¹² Unfortunately, other IM products might have frequencies close to the desired IF, making them difficult to remove, as we shall

¹⁰ We will shortly see the advantages of three-port mixers.

¹¹ Keep in mind that fundamentals are harmonics.

¹² The order of a given IM term is the sum of p and q , so a second-order IM product arises from the quadratic term in the series expansion.

see. Since it is generally true that high-order nonlinearities (i.e., large values of N in the power series expansion) tend to generate more of these undesirable terms,¹³ mixers should approximate square-law behavior (the lowest-order nonlinearity) if they only have one input port (as shown in Figure 12.3). We now consider the specific properties of a square-law mixer to identify its advantages over higher-order nonlinear mixers.

TWO-PORT EXAMPLE: SQUARE-LAW MIXER

To see explicitly where the desired multiplication arises in a square-law mixer, note that the only nonzero coefficients in the series expansion are the c_1 and c_2 terms.¹⁴ If we then assume that the input signal v_{IN} is the sum of two sinusoids,

$$v_{IN} = v_{RF} \cos(\omega_{RF}t) + v_{LO} \cos(\omega_{LO}t), \quad (3)$$

then the output of this mixer may be expressed as the sum of three distinct components:

$$v_{OUT} = v_{fund} + v_{square} + v_{cross}, \quad (4)$$

where

$$v_{fund} = c_1[v_{RF} \cos(\omega_{RF}t) + v_{LO} \cos(\omega_{LO}t)], \quad (5)$$

$$v_{square} = c_2\{[v_{RF} \cos(\omega_{RF}t)]^2 + [v_{LO} \cos(\omega_{LO}t)]^2\}, \quad (6)$$

$$v_{cross} = 2c_2v_{RF}v_{LO}[\cos(\omega_{RF}t)][\cos(\omega_{LO}t)]. \quad (7)$$

The fundamental terms are simply scaled versions of the original inputs, and therefore represent no useful mixer output; they must be removed by filtering. The v_{square} components similarly represent no useful mixer output, as is evident from the following special case of Eqn. 1:

$$[\cos \omega t]^2 = \frac{1}{2}[1 + \cos 2\omega t]. \quad (8)$$

Thus, we see that the v_{square} components contribute a DC offset as well as second harmonics of the input signals. These also must generally be removed by filtering.

The useful output comes from the v_{cross} components because of the multiplication evident in Eqn. 7. Using Eqn. 1, we may rewrite v_{cross} in a form that shows the mixing action more clearly:

$$v_{cross} = c_2v_{RF}v_{LO}[\cos(\omega_{RF} - \omega_{LO})t + \cos(\omega_{RF} + \omega_{LO})t]. \quad (9)$$

¹³ As with most sweeping generalities, there are exceptions to this one. In building frequency multipliers, for example, high-order harmonic nonlinearities are extremely useful. In mixer design, however, it is usually true that high-order nonlinearities are undesirable.

¹⁴ There may also be a nonzero DC term (i.e., c_0 may be nonzero), but this component is easily removed by filtering, so we will ignore it at the outset to reduce equation clutter.

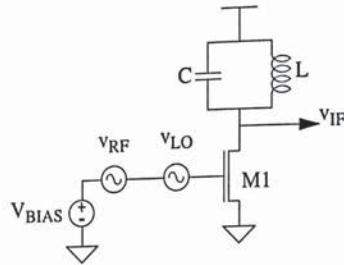


FIGURE 12.4. Square-law MOSFET mixer (simplified).

For a fixed LO amplitude, the IF output amplitude is linearly proportional to the RF input amplitude. That is, this nonlinearity implements a linear mixing, since the output is proportional to the input.

The conversion gain for this nonlinearity is readily found from Eqn. 9:

$$G_c = \frac{c_2 v_{RF} v_{LO}}{v_{RF}} = c_2 v_{LO}. \quad (10)$$

Just as any other gain parameter, conversion gain may be a dimensionless quantity (or a transconductance, transresistance, etc.). It is customary in discrete designs to express conversion gain as a power ratio (or its decibel equivalent), but the unequal input and output impedance levels in typical IC mixers makes a voltage or current conversion gain appropriate also. To avoid confusion, of course, it is essential to state explicitly the type of gain.¹⁵

As asserted earlier, the square-law mixer's advantages are that the undesired spectral components are usually at a frequency quite different from the intermediate frequency, and thus readily removed. For this reason, two-port mixers are often designed to conform to square-law behavior to the maximum practical extent.

Excellent square-law mixers may be realized with long-channel MOSFETs, or approximated by virtually any other type of nonlinearity in which the quadratic term dominates; see Figure 12.4. In this simplified schematic, the bias, RF, and LO terms are shown as driving the gate in series. The summation of RF and LO signals can be accomplished in practical circuits with resistive or reactive summers. Because the RF and LO signals are in series, there is poor isolation between them.

An alternative (but functionally equivalent) arrangement that reduces the effect of the relatively large LO signal on the RF port is shown in Figure 12.5. The RF signal

¹⁵ All too frequently, published "power" gain figures are essentially voltage gain measurements and are therefore grossly in error if the input and output impedance levels differ significantly, as they often do. It seems necessary to emphasize that watts and volts are not the same.

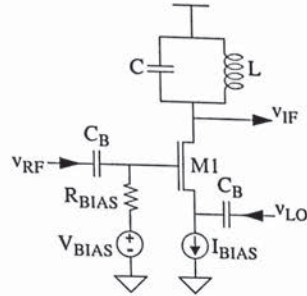


FIGURE 12.5. Square-law MOSFET mixer (alternative configuration).

drives the gate directly (through a DC-blocking capacitor), while the LO drives the source terminal. This way, the gate-to-source voltage is the sum of ground-referenced LO and RF signals. The bias current is set directly with a current source, and the DC gate voltage is determined by the value of V_{BIAS} . Resistor R_{BIAS} is chosen large enough to avoid excessive loading, and also to minimize its noise contribution.

In deriving an expression for conversion gain, we use our assumption that the device is long enough (and biased appropriately) to allow us to express the drain current as follows:

$$i_D = \frac{\mu C_{ox} W}{2L} (V_{gs} - V_T)^2. \quad (11)$$

Short-channel (high-field) devices are more linear as a result of velocity saturation, and thus are generally inferior to long devices as mixers.¹⁶

If the gate-source voltage V_{gs} is the sum of RF, LO, and bias terms, then we may write

$$i_D = \frac{\mu C_{ox} W}{2L} \{ V_{BIAS} + [v_{RF} \cos(\omega_{RF} t) - v_{LO} \cos(\omega_{LO} t)] - V_T \}^2, \quad (12)$$

from which one may readily find that the conversion gain (here, a transconductance) is simply

$$G_c = \frac{\mu C_{ox} W}{2L} \cdot v_{LO}. \quad (13)$$

This square-law device thus has a conversion transconductance that is independent of bias.¹⁷ It is still dependent on temperature (through mobility variation) and LO drive amplitude, however.

¹⁶ The reader is reminded once again that "short-channel" actually means "high-field." Hence, even "short" devices may still behave quadratically for suitably small drain-source voltages.

¹⁷ This independence of bias holds only in the square-law regime. Enough bias must therefore be supplied to guarantee this condition. Hence, V_{BIAS} is not permitted to equal zero. In fact, it must be chosen large enough to guarantee that the gate-source voltage always exceeds the threshold voltage, since a MOSFET behaves exponentially in weak inversion.

Because perfect square-law behavior is not necessary to obtain mixing action, M_1 can be a bipolar transistor, for example, because the quadratic factor in the series expansion for the exponential $i_C - v_{BE}$ relationship dominates over a limited range of input amplitudes. Precisely because many nonlinearities are well approximated by a square-law shape over some suitably restricted interval, one can estimate the conversion gain for other nonlinear devices used as mixers once the value of the quadratic coefficient (c_2) is found. To underscore this point, let's estimate the conversion gain for one more nonlinear element, a bipolar transistor.

Conversion Gain of a Single Bipolar Transistor Mixer

To simplify the calculation, let us continue to ignore dynamic effects. Then we can use the exponential v_{BE} law:

$$i_C \approx I_S e^{v_{BE}/V_T}. \quad (14)$$

Expansion of this familiar relationship up to the second-order term yields¹⁸

$$i_C \approx I_C \left[1 + \frac{v_{IN}}{V_T} + \frac{1}{2} \left(\frac{v_{IN}}{V_T} \right)^2 \right]. \quad (15)$$

By inspection (well, almost),

$$c_2 = \frac{g_m}{2V_T}, \quad (16)$$

so that an estimate of the conversion gain is:

$$G_c = c_2 v_{LO} = g_m \cdot \frac{v_{LO}}{2V_T}. \quad (17)$$

The conversion gain here is a transconductance proportional both to the standard incremental transconductance, and to the ratio of the local oscillator drive amplitude to the thermal voltage. The conversion gain for a bipolar transistor is therefore dependent on bias current, LO amplitude, and temperature.

As in the corresponding derivation for a MOSFET, the foregoing computation ignores parasitic series base and emitter resistances. These resistances can linearize the transistor and therefore weaken mixer action. Thoughtful device layout is thus mandatory to minimize this effect.

12.4 MULTIPLIER-BASED MIXERS

We have seen that nonlinearities produce mixing incidentally through the multiplications they provide. Precisely because the multiplication is only incidental, these nonlinearities usually generate a host of undesired spectral components. Additionally,

¹⁸ We have implicitly assumed that the base-emitter drive contains a DC component as well as the RF and LO components, so that I_C is nonzero.

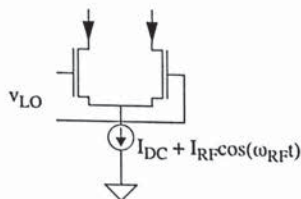


FIGURE 12.6. Single-balanced mixer.

since two-port mixers have only one input port, the RF and LO signals are generally not well isolated from each other. This lack of isolation can cause the problems mentioned earlier, such as overloading of IF amplifiers, as well as radiation of the LO signal (or its harmonics) back out through the antenna.

Mixers based directly on multiplication generally exhibit superior performance because they ideally generate only the desired intermodulation product. Furthermore, because the inputs to a multiplier enter at separate ports, there can be a high degree of isolation among all three signals (RF, LO, and IF). Finally, CMOS technology provides excellent switches, and one can implement outstanding multipliers with switches.

12.4.1 SINGLE-BALANCED MIXER

One extremely common family of multipliers first converts the incoming RF voltage into a current and then performs a multiplication in the current domain. The simplest multiplier cell of this type is sketched in Figure 12.6.¹⁹ In this mixer, v_{LO} is chosen large enough so that the transistors alternately switch (commutate) all of the tail current from one side to the other at the LO frequency.²⁰ The tail current is therefore effectively multiplied by a square wave whose frequency is that of the local oscillator:

$$i_{out}(t) = \text{sgn}[\cos \omega_{LO} t] \{I_{BIAS} + I_{RF} \cos \omega_{RF} t\}. \quad (18)$$

Because a square wave consists of odd harmonics of the fundamental, multiplication of the tail current by the square wave results in an output spectrum that appears

¹⁹ Mixers of this general kind are often lumped together and called Gilbert mixers, but only some actually are. True Gilbert multipliers function entirely in the current domain, deferring the problem of V - I conversion by assuming that all variables are already available in the form of currents. See Barrie Gilbert's landmark paper, "A Precise Four-Quadrant Multiplier with Subnanosecond Response," *IEEE J. Solid-State Circuits*, December 1968, pp. 365-73.

²⁰ One may also interchange the roles of LO and RF input, but the resulting mixer has lower conversion gain and worse noise performance, among other deficiencies. A more detailed discussion of this issue is deferred to a later section.

as shown in Figure 12.7 (ω_{RF} is here chosen atypically low compared with ω_{LO} to reduce clutter in the graph).

The output thus consists of sum and difference components, each the result of an odd harmonic of the LO mixing with the RF signal. In addition, odd harmonics of the LO appear directly in the output as a consequence of the DC bias current multiplying with the LO signal. Because of the presence of the LO in the output spectrum, this type of mixer is known as a *single-balanced* mixer. Double-balanced mixers, which we'll study shortly, exploit symmetry to remove the undesired output LO component through cancellation.

Although the current source of Figure 12.6 includes a component that is perfectly proportional to the RF input signal, V - I converters of all real mixers are imperfect. Hence, an important design challenge is to maximize the linearity of the RF transconductance. Linearity is most commonly enhanced through some type of source degeneration, in both common-gate and common-source transconductors; see Figure 12.8. The common-gate circuit uses the source resistance R_s to linearize the transfer characteristic. This linearization is most effective if the admittance looking into the source terminal of the transistor is much larger than the conductance of R_s . In that case, the transconductance of the stage approaches $1/R_s$.

Inductive degeneration is usually preferred over resistive degeneration for several reasons.²¹ An inductance has neither thermal noise to degrade noise figure nor DC voltage drop to diminish supply headroom. This last consideration is particularly relevant for low-voltage-low-power applications. Finally, the increasing reactance of an inductor with increasing frequency helps to attenuate high frequency harmonic and intermodulation components.

A more complete single-balanced mixer that incorporates a linearized transconductance is shown in Figure 12.9. The value of V_{BIAS} establishes the bias current of the cell, while R_B is chosen large enough not to load down the gate circuit (and also to reduce its noise contribution). The RF signal is applied to the gate through a DC blocking capacitor C_B . In practice, a filter would be used to remove the LO and other undesired spectral components from the output.

The conversion transconductance of this mixer can be estimated by assuming that the LO-driven transistors behave as perfect switches. Then the differential output current may be regarded as the result of multiplying the drain current of M_1 by a unit-amplitude square wave. Since the amplitude of the fundamental component of a square wave is $4/\pi$ times the amplitude of the square wave, we may write:

$$G_c = \frac{2}{\pi} g_m, \quad (19)$$

²¹ Capacitive degeneration has also been tried, but is markedly inferior to inductive degeneration because it increases noise and distortion at high frequencies.

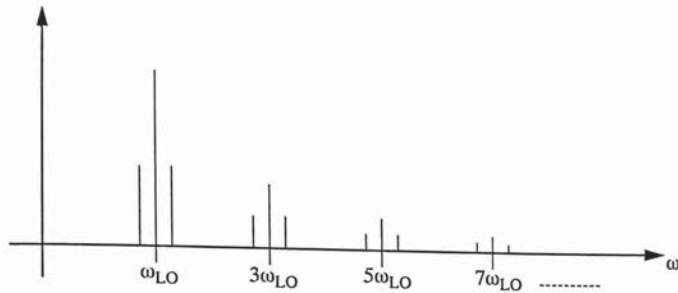


FIGURE 12.7. Representative output spectrum of single-balanced mixer.

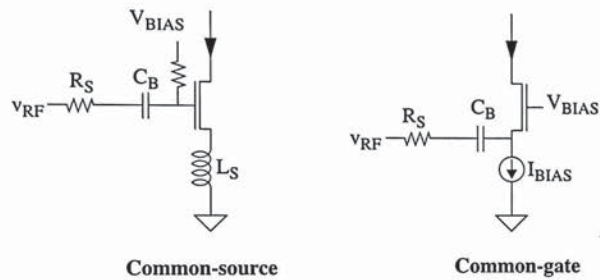


FIGURE 12.8. RF transconductors for mixers.

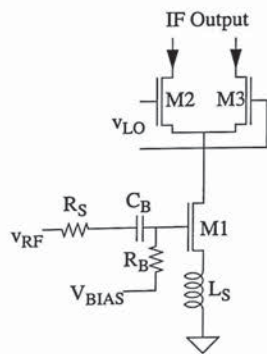


FIGURE 12.9. Single-balanced mixer with linearized transconductance.

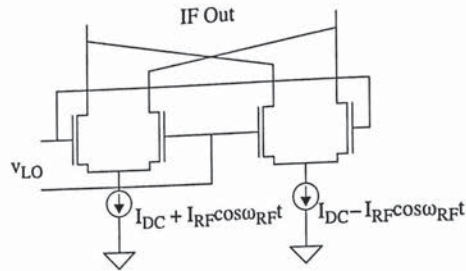


FIGURE 12.10. Active double-balanced mixer.

where g_m is the transconductance of the V - I converter and G_c is itself a transconductance. The coefficient is $2/\pi$ rather than $4/\pi$ because the IF signal is divided evenly between sum and difference components.

12.4.2 ACTIVE DOUBLE-BALANCED MIXER

To prevent the LO products from getting to the output in the first place, two single-balanced circuits may be combined to produce a double-balanced mixer; see Figure 12.10. We assume once again that the LO drive is large enough to make the differential pairs act like current-steering switches. Note that the two single-balanced mixers are connected in antiparallel as far as the LO is concerned but in parallel for the RF signal. Therefore, the LO terms sum to zero in the output, whereas the converted RF signal is doubled in the output. This mixer thus provides a high degree of LO-IF isolation, easing filtering requirements at the output. If care is taken in layout, IC realizations of this circuit routinely provide 40 dB of LO-IF isolation, with values in excess of 60 dB possible.

As in the single-balanced active mixer, the dynamic range is limited in part by the linearity of the V - I converter in the RF port of the mixer. So, most of the design effort is spent attempting to find better ways of providing this V - I conversion. The basic linearizing techniques used in the single-balanced mixer may be adapted to the double-balanced case, as shown in Figure 12.11

In low-voltage applications, the DC current source can be replaced by a parallel LC tank to create a zero-headroom AC current source. The resonant frequency of the tank should be chosen to provide rejection of whatever common-mode component is most objectionable. If several such components exist, one may use series combinations of parallel LC tanks. With such a choice, a complete double-balanced mixer appears as shown in Figure 12.12. The expression for the conversion transconductance is the same as for the single-balanced case.

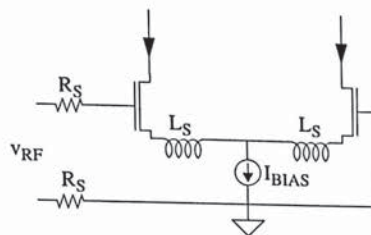


FIGURE 12.11. Linearized differential RF transconductor for double-balanced mixer.

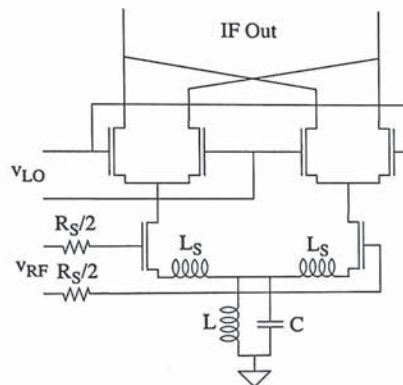


FIGURE 12.12. Minimum supply-headroom double-balanced mixer.

Noise Figure of Gilbert-Type Mixers

Computing the noise figure of mixers is difficult because of the cyclostationary nature of the noise sources. One technique involves characterization of the time-varying impulse response, arguing that a mixer is at least linear, if not time-invariant.²² Although the method is accurate and quite suitable for analysis, it conveys limited design insight. Nonetheless, we can identify several important noise sources and make general recommendations about how to minimize noise figure.

One noise source is certainly the transconductor itself, so that its noise figure establishes a lower bound on the mixer noise figure. The same approach used in computing LNA noise figure may be used to compute the transconductor noise figure.

²² C. D. Hull and R. G. Meyer, "A Systematic Approach to the Analysis of Noise in Mixers," *IEEE Trans. Circuits and Systems I*, v. 40, no. 12, December 1993, pp. 909–19.

The differential pair also degrades noise performance in a number of ways. One noise figure contribution arises from imperfect switching, which causes attenuation of the signal current. Hence, one challenge in such mixers is to design the switches (and associated LO drive) to provide as little attenuation as possible.

Another NF contribution of the switching transistors arises from the interval of time in which both transistors conduct current and hence generate noise. Additionally, any noise in the LO is also magnified during this active gain interval. Minimizing the simultaneous conduction interval reduces this degradation, so sufficient LO drive must be supplied to make the differential pair approximate ideal, infinitely fast switches to the maximum practical extent. Finally, the 3-dB attenuation inherent in ignoring either the sum or difference signal automatically degrades noise figure (by 3 dB) since the noise cannot be discarded so readily. As a result, practical current-mode mixers typically exhibit SSB noise figures of at least 10 dB, with values more frequently in the neighborhood of 15 dB.

Linearity of Gilbert-Type Mixers

The IP3 of this type of mixer is bounded by that of the transconductor, so the three-point method used to estimate the IP3 of ordinary amplifiers may also be used here to estimate the IP3 of the transconductor. If the LO-driven transistors act as good switches then the overall mixer IP3 generally differs little from that of the transconductor. To guarantee good switching, it is important to note that – although sufficient LO drive is necessary – excessive LO drive is to be avoided. To understand the reason that excessive LO drive is a liability rather than an asset, consider the effect of ever-present capacitive parasitic loading on the common-source connection of a differential pair. As each gate is driven far beyond what's necessary for good switching, the common-source voltage is similarly overdriven. A spike in current results. In extreme cases, this spike can cause transistors to leave the saturation region. Even if that does not occur, the output spectrum can become dominated by the components arising from the spikes, rather than the downconverted RF. Hence, one should use only enough LO drive to guarantee reliable switching, and no more.

A Short Note on Simulation of Mixer IP3 with Time-Domain Simulators

Common circuit simulators, such as SPICE, provide accurate mixer simulations only reluctantly, if at all. The problem stems from two fundamental sources: The wide dynamic range of signals in a mixer forces the use of far tighter numerical tolerances than are adequate for “normal” circuit simulations; and the large span of frequencies of important spectral components forces long simulation times. Hence, obtaining an accurate value for IP3 from a transient simulation, for example, is usually quite challenging. Furthermore, a correct noise figure simulation for CMOS mixers is not possible at all with commercially available tools, because the device noise models

presently in use are incorrect. The reader is therefore cautioned to treat mixer simulation results with great skepticism.

Because even the "accurate" options available in some simulation tools are orders of magnitude too loose to be useful for IP3 simulations, one specific action that mitigates some of these problems is to tighten tolerances progressively until the simulation results stop changing significantly. In particular, the behavior of the IM3 component in an IP3 simulation is an extremely sensitive indicator of whether the tolerances are sufficiently tight. If the IM3 terms do not exhibit a +3 slope (on a dB scale), chances are high that the tolerances are too loose. One must also make sure that the amplitudes of the two input tones are chosen small enough (i.e., well below either the compression or intercept points) to guarantee quasilinear operation of the mixer; otherwise, higher-order terms in the nonlinearity will contribute significantly to the output and confound the results. In the early phases of design, the three-point method may be applied to the transconductor to estimate its IP3 without having to suffer the agony of a transient simulation.

Another subtle consideration is to guarantee equal time spacing in the transient simulation, since FFT algorithms generally assume uniform sampling. Because some simulators use adaptive time stepping to speed up convergence, significant spectral artefacts can arise when computing the FFT. One may set the time step to a tiny fraction of the fastest time interval of interest to assure convergence without resort to adaptive time stepping. As an example, one might have to use a time step (parameter "delmax" in HSPICE²³) that is three orders of magnitude smaller than the period of the RF signal. Hence, for a 1-GHz RF input, one might need to use a 1-ps time step. It is this combination of iteration, tight time step, and numerical tolerance problems that causes IP3 simulations to execute so slowly.²⁴

Additional Linearization Techniques

Because the linearity of these current-mode mixers is controlled primarily by the quality of the transconductance, it is worthwhile to consider additional ways to extend linearity. Philosophically, there are four methods for doing so: predistortion, feedback, feedforward, and piecewise approximation. These techniques can be used alone or in combination. What follows is a representative (but hardly exhaustive) set of examples of these methods.

Predistortion cascades two nonlinearities that are inverses of each other, and it shares with feedforward the need for careful matching. Predistortion is actually nearly

²³ HSPICE is a trademark of the Meta-Software Corporation.

²⁴ Measuring these quantities in the laboratory also requires some care. As with the simulation, the amplitudes of the two input tones must be low enough to avoid excitation of higher-order nonlinearities (which would cause a slope of other than +3) but still sufficiently larger than the noise floor.

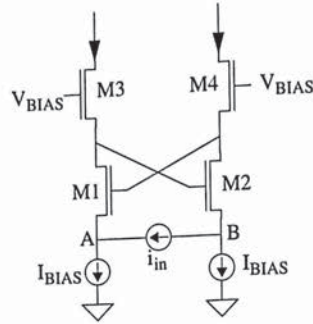


FIGURE 12.13. MOSFET cross-quad.

ubiquitous, as it is the principle underlying the operation of current mirrors. In a mirror, an input current is converted to a gate-to-source voltage through some nonlinear function that is then undone to produce an output current exactly proportional to the input. Predistortion is also fundamental to the operation of true Gilbert mixers, where a pair of junctions computes the inverse hyperbolic tangent of an input differential current, and a differential pair subsequently undoes that nonlinearity.

Negative feedback computes an estimate of error, inverts it, and adds it back to the input, thereby helping to cancel the errors that distortion represents. The reduction in distortion is large as long as the loop transmission magnitude is large. Because a negative feedback system computes the error a posteriori, the overall closed-loop bandwidth must be kept a small fraction of the inherent bandwidth capabilities of the elements comprising the system; otherwise, the a posteriori estimate will be irrelevant at best and destabilizing at worst. The series feedback examples of this chapter are popular methods for linearizing high-frequency transconductors.

Contrary to common prejudice, positive feedback cannot be precluded as a linearizing technique. Furthermore, since loop transmission magnitudes must be less than unity to guarantee stability, the bandwidth penalty is much less severe than for negative feedback. As an illustrative example, the *cross-quad*, adapted from its bipolar progenitor, uses positive feedback to synthesize a virtual short-circuit; see Figure 12.13.

To show that this connection presents a short circuit to an applied current, i_{in} , consider how the voltages at the sources of M_1 and M_2 change as i_{in} changes. As i_{in} increases, the gate-source voltages of M_2 and M_4 increase by an equal amount, while those of M_1 and M_3 similarly decrease. The voltage at node A is:

$$V_{BIAS} - (V_{gs4} + V_{gs1}). \quad (20)$$

Similarly, the voltage at node B is:

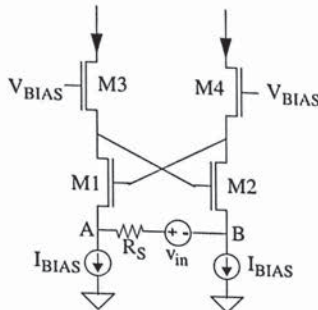


FIGURE 12.14. Cross-quadrant transconductor.

$$V_{\text{BIAS}} - (V_{gs3} + V_{gs2}). \quad (21)$$

That is, the voltage at each source terminal is below V_{BIAS} by an amount equal to the sum of a high V_{gs} and a low V_{gs} . Hence, the two source voltages are always equal; the circuit synthesizes a virtual short circuit.²⁵

Such a short circuit can be used to shift the burden of linearity away from active elements to a passive element, such as a resistance; this is shown in Figure 12.14. Because nodes A and B are at the same potential, the current injected into A is equal to v_{in}/R_s . This injected current is thus perfectly proportional to the input voltage, and is recovered as a differential output current at the drains of M_3 and M_4 .

A variation of the cross-quadrant applies the input voltage across the gates of the top pair, as seen in Figure 12.15. The value of the transconductance is still equal to the conductance of R_s .

Feedforward is another linearization technique; it computes an estimate of the error at the same time the system processes the signal, thereby evading the bandwidth and stability problems of negative feedback. However, the error computation and cancellation then depend on matching, so the maximum practical distortion reduction tends to be substantially less than generally attainable with negative feedback. Feedforward is most attractive at high frequencies, where negative feedback becomes less effective owing to the insufficiency of loop transmission.

An example of feedforward correction applied to a transconductor is an adaptation of Pat Quinn's bipolar "cascomp" circuit²⁶ (Figure 12.16). As can be seen, this transconductor consists of a cascoded differential pair to which an additional differential pair has been added. Some linearization is provided by the source degeneration

²⁵ This analysis neglects body effect. Practical implementations do not work quite ideally as a result.

²⁶ "Feedforward Amplifier," U.S. Patent #4,146,844, issued 27 March 1979, reissued 1984.

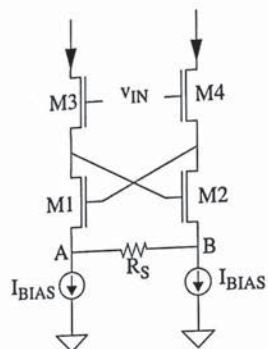


FIGURE 12.15. Alternate connection of cross-quadrant transconductor.

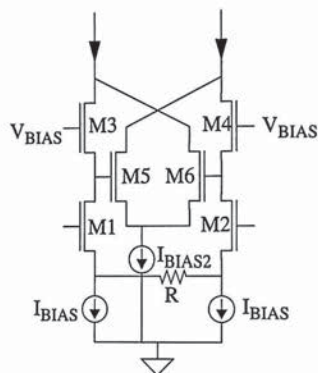
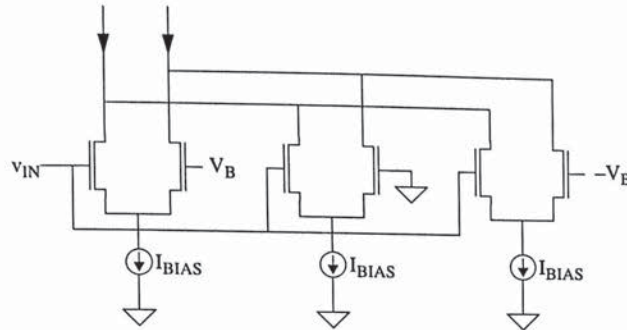


FIGURE 12.16. MOSFET cascomp.

resistor R , but significant nonlinearity remains in the transconductance of inner differential pair M_1 – M_2 . To see this explicitly, consider that the voltage across the resistor is the input voltage minus the difference in gate-to-source voltages of M_1 and M_2 :

$$V_R = v_{in} - (v_{gs1} - v_{gs2}) = v_{in} - \Delta v_{gs1}. \quad (22)$$

The goal is to have a differential output current precisely proportional to v_{in} , so any Δv_{gs} represents an error. The cascoding pair possesses the same Δv_{gs} as the input pair, which is measured by the inner differential pair. A current proportional to this error is subtracted from the main current to linearize the transconductance. The name “cascomp” derives from this combination of a cascode and error compensation. Although the inner pair is shown as an ordinary differential pair for simplicity, it is frequently advantageous to linearize it to increase the error correction range.

FIGURE 12.17. CMOS g_m cell.

Another nonfeedback approach is piecewise approximation, which exploits the observation that virtually any system is linear over some sufficiently small range. It divides responsibility for linearity among several systems, each of which is active only over a small enough range so that the composite exhibits linearity over an extended range.

Gilbert's bipolar "multi-tanh"²⁷ arrangement is an example of piecewise approximation. In MOS form, it appears as shown in Figure 12.17. Each of the three differential pairs behaves as a reasonably linear transconductance over an input voltage range centered about V_B , 0, and $-V_B$, respectively. For input voltages near zero, the transconductance is provided by the middle pair, and is roughly constant for small enough v_{IN} . As the input voltage deviates significantly from zero, the tail current eventually steers almost completely to one side of the middle pair. However, with an appropriate selection of bias voltage V_B , one of the outer pairs takes over and continues to contribute an increase in output current; see Figure 12.18.

The overall transconductance is the sum of the individual offset transconductances, and can be made roughly constant over an almost arbitrarily large range by using a sufficient number of additional differential pairs, each offset appropriately. The trade-off is an increase in power dissipation and input capacitance.

12.4.3 POTENTIOMETRIC MIXERS

Gilbert-type mixers first convert an incoming RF voltage into a current through a transconductor, whose linearity and noise figure set a firm bound on the overall mixer linearity and noise figure. An alternative to using voltage-controlled current

²⁷ The name derives from the fact that the transfer characteristic of a bipolar differential pair is a hyperbolic tangent.

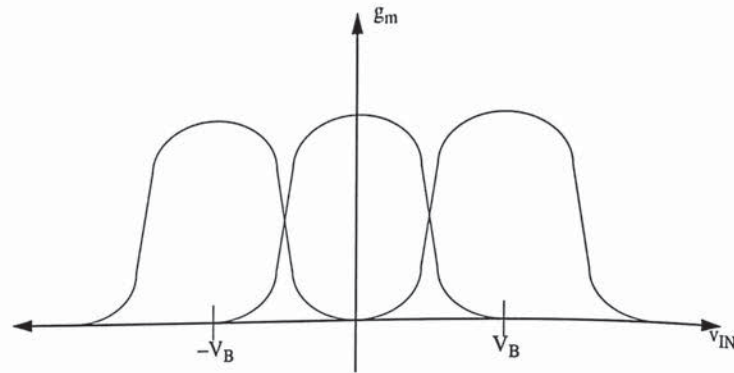


FIGURE 12.18. Illustration of linearization by piecewise approximation.

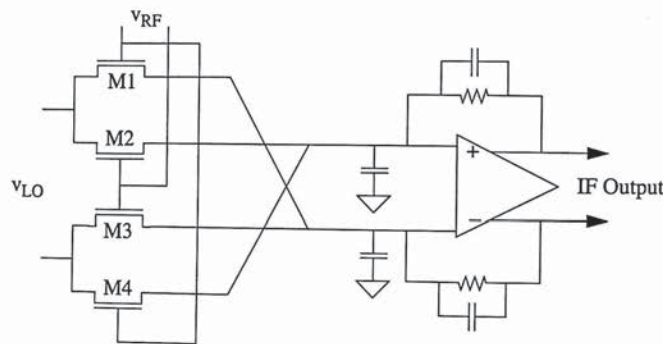


FIGURE 12.19. Potentiometric mixer.

sources in V - I converters is to use voltage-controlled resistances. For example, consider varying the resistance of a triode-region MOSFET in a manner inversely proportional to the incoming RF signal. If the voltage between drain and source is maintained at a fixed value, the current flowing through the device will be a faithful replica of the RF voltage, and if v_{ds} varies with the LO then the current will be proportional to the product of the LO and RF signals. One possible implementation of this idea is sketched in Figure 12.19.²⁸ The four MOSFETs perform the mixing, while the capacitors remove the sum frequency component as well as higher-order products.

²⁸ J. Crols and M. Steyaert, "A 1.5GHz Highly Linear CMOS Downconversion Mixer," *IEEE J. Solid-State Circuits*, v. 30, no. 7, July 1995, pp. 736-42.

The RF input drives the gates of the transistors, while the LO drives the sources. A simplified analysis assumes that the resistances of the transistors are inversely proportional to the RF signal. In that case, the current through the devices is

$$i_{in} = \frac{v_{LO}}{r_{ds}} \approx v_{LO} \cdot \mu C_{ox} \frac{W}{L} [(v_{RF} - V_T) - v_{LO}] \approx K \cdot v_{LO} \cdot v_{rf}. \quad (23)$$

Because the current is then the result of a multiplication of the RF and LO signals, there are components at the sum and difference frequencies, as desired. This current flows through the feedback resistors so that the IF signal is available as an output voltage. The op-amp need only have enough bandwidth to handle the difference frequency component, since the sum component is filtered out by the four capacitances.

Note that, for good linearity, the gate overdrive must greatly exceed v_{LO} . Hence, v_{RF} must possess a sufficiently large DC component to satisfy this inequality for as large a value of v_{rf} that must be accommodated.

Practical mixers of this type may exhibit good linearity (e.g., 40 dBm IIP3) but high noise figures (e.g., 30 dB). The high noise figures are the result of the resistive thermal noise of the input FETs (which is worst when the signal levels are small) and the difficulty of providing a good noise match with the broadband op-amp. As a consequence, the overall dynamic range of this type of mixer is typically about the same as conventional Gilbert-type current-mode mixers.

12.4.4 PASSIVE DOUBLE-BALANCED MIXER

So far, we've examined active mixers only, with their attendant need for linear trans-conduction. However, passive mixers have some attractive properties, such as the potential for extremely low-power operation. Considering that CMOS technology offers excellent switches, high-performance multipliers based on switching are naturally realized in CMOS form.

In the active mixers considered so far, representations of the RF signal in the form of currents, rather than the RF voltages themselves, are effectively multiplied by a square-wave version of the local oscillator. An alternative that avoids the $V-I$ conversion problem is to switch the RF signal directly in the voltage domain. This option is considerably easier to exercise in CMOS than bipolar form, which is why bipolar mixers are almost exclusively of the active, current-mode type.

The simplest passive commutating CMOS mixer consists of four switches in a bridge configuration (see Figure 12.20). The switches are driven by local oscillator signals in antiphase, so that only one diagonal pair of transistors is conducting at any given time. When M_1 and M_4 are on, v_{IF} equals v_{RF} , and when M_2 and M_3 are conducting, v_{IF} equals $-v_{RF}$. A fully equivalent description is that this mixer multiplies the incoming RF signal by a unit-amplitude square wave whose frequency is

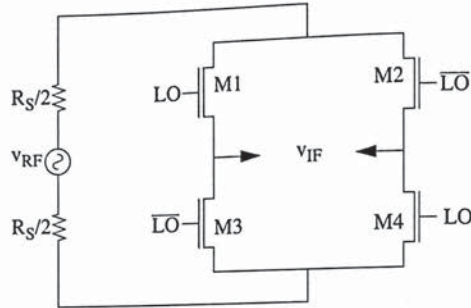


FIGURE 12.20. Simple double-balanced passive CMOS mixer.

that of the local oscillator. Hence, the output contains many mixing products that result from the odd-harmonic Fourier components of the square wave.²⁹ Luckily, these are often readily filtered out, as discussed previously.

The voltage conversion gain of this basic cell is easy to compute from the foregoing description. Assuming multiplication by a unit-amplitude square wave, we may immediately write

$$G_C = 2/\pi. \quad (24)$$

Here, the $2/\pi$ factor again results from splitting the IF energy evenly between the sum and difference components.³⁰

In practice, the actual voltage conversion gain may differ somewhat from $2/\pi$ because real transistors do not switch in zero time. Hence, the incoming RF signal is not multiplied by a pure square-wave signal in general. Perhaps contrary to intuition, however, the effect of this departure from ideal assumptions is usually to *increase* the voltage conversion gain above $2/\pi$.

A more general expression for the voltage conversion gain is somewhat cumbersome to derive, so we will state only the relevant insights here.³¹ The output of the mixer may be treated as the product of three time-varying components and a scaling factor:

$$v_{IF}(t) = v_{RF}(t) \cdot \left[\frac{g_T(t)}{g_{T \max}} \cdot m(t) \right] \cdot \left[\frac{g_{T \max}}{g_T} \right]. \quad (25)$$

²⁹ This situation is the same as with the current-mode mixers, however. Also, even harmonics of the LO terms may be nonzero if the duty cycle of the square wave is not exactly 50%.

³⁰ If we assume equal source and load terminations, then this gain corresponds to a 3.92-dB voltage and power loss. Many practical implementations, such as the discrete passive mixers discussed in Section 12.6, typically exhibit a somewhat greater conversion loss than this theoretical limit because of additional sources of attenuation (e.g. nonzero switch drop, skin effect loss, etc.). Common conversion losses for mixers of this type are in the neighborhood of 5 dB to 6 dB.

³¹ For a detailed derivation, see A. Shahani et. al., "A 12mW Wide Dynamic Range CMOS Front-End for a Portable GPS Receiver," *IEEE J. Solid-State Circuits*, December 1997.

The function $g_T(t)$ is the time-varying Thévenin-equivalent conductance as viewed from the IF port, while $g_{T\max}$ and $\overline{g_T}$ are the maximum and average values, respectively, of $g_T(t)$. The *mixing function*, $m(t)$, is defined by

$$m(t) = \frac{g(t) - g(t - T_{LO}/2)}{g(t) + g(t - T_{LO}/2)}, \quad (26)$$

where $g(t)$ is conductance of each switch and T_{LO} is the period of the LO drive. The mixing function has no DC component, is periodic in T_{LO} , and has only odd harmonic content because of its half-wave symmetry.

The Fourier transform of the first bracketed term in Eqn. 25 has a value of $2/\pi$ at the LO frequency for a square-wave drive (as asserted earlier), and a value of $1/2$ for a sinusoidal drive, so the effective mixing function indeed contributes a higher conversion gain for a square-wave drive. However, the second bracketed term is unity for a square-wave drive (because the peak and average conductances are equal) but $\pi/2$ for a sinusoidal drive. The overall conversion gain is greater with a sinusoidal drive because the second term more than compensates for the smaller contribution by the (effective) mixing function. The difference is not particularly large, however. With a sinusoidal drive, the conversion gain is $\pi/4$ (−2.1 dB), compared with the $2/\pi$ gain (−3.92 dB) obtained with the square-wave drive.

Because of the spectrum of the (effective) mixing function, undesirable products can appear at the IF port of this type of mixer. The subject of filtering therefore deserves careful consideration, especially in connection with the issue of input and output terminations. In discrete designs, the source and load impedances are usually real and well-defined (50 Ω , for example), but the sources and loads for IC mixers are usually on-chip and not at all standardized. Far from a liability, this lack of standardization is a degree of freedom that the IC engineer can exploit to improve performance. As a specific example, reactive source and load terminations might be preferable because they do not generate noise. Because it is difficult to obtain broadband operation with reactances, narrowband operation is implied for most practical mixers with reactive terminations. Fortunately, there are many applications for which this restriction is not a serious limitation.

A somewhat more elaborate passive mixer that exploits this freedom of source and load terminations appears as Figure 12.21.³² Note that this mixer assumes a capacitive load, represented as C_L in the schematic. This assumption reflects the typical situation in fully integrated CMOS circuits, and stands in contrast with the resistive terminations common in discrete designs. A capacitive load generates no thermal noise of its own, and also helps filter out high-frequency noise and distortion.

³² This example is adapted from A. Shahani et al., "A 12mW Wide Dynamic Range CMOS Front-End for a Portable GPS Receiver," *ISSCC Digest of Technical Papers*, February 1997, pp. 368–9.

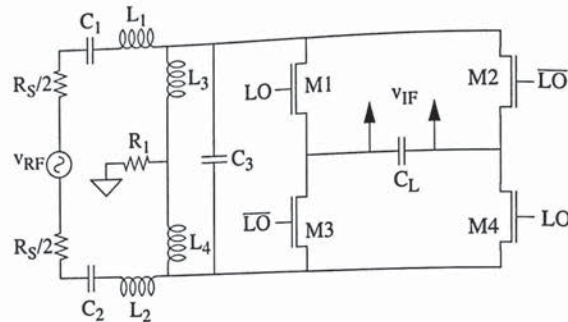


FIGURE 12.21. Low-noise, narrowband passive mixer.

The input network consists of an L-match in cascade with a parallel tank. The L-match, comprising L_1 and part of the tank capacitance, provides an impedance transformation that moderately boosts the RF signal voltage to help reduce the voltage conversion loss. The parallel tank, formed by L_3 and $C_3 + C_L$, filters out-of-band noise and distortion components present at the input and generated by the mixer itself. Resistor R_1 sets the common-mode potential for the input circuit. Because any nonlinearity in the tank capacitance reduces IP3, C_3 is best implemented as a metal-metal capacitor. To reduce the area consumed, a good choice is to use a lateral flux or fractal capacitor.

Because of the small voltage boost provided by the L-match, the voltage conversion loss can be somewhat better than the 3.92 dB that a simple switch bridge would exhibit ideally. As an example, one implementation in a 0.35- μm technology exhibits a 3.6-dB voltage conversion loss with a 1.6-GHz RF and 1.4-GHz LO.³³

Both noise figure and IP3 are strong functions of the LO drive, since the resistance of the switches in the “on” state must be kept low and constant to optimize both parameters. The IP3 is also a function of the amount of voltage boost provided by the L-match. This boost may be adjusted downward to trade conversion gain for improved IP3 and, in some cases, it may be appropriate to remove the L-match altogether. Typical SSB noise figures of 10 dB and input IP3 of 10 dBm are readily achievable with an LO drive amplitude of 300 mV.³⁴ As a crude estimate, the DSB noise figure of this type of mixer is approximately equal to the power attenuation.

As a final note on the noise performance of this type of mixer, it is perhaps useful to observe that the absence of DC bias current implies the absence of $1/f$ noise. This

³³ A. Shahani et al. (February 1997), op. cit.

³⁴ This value applies to a sine-wave LO.

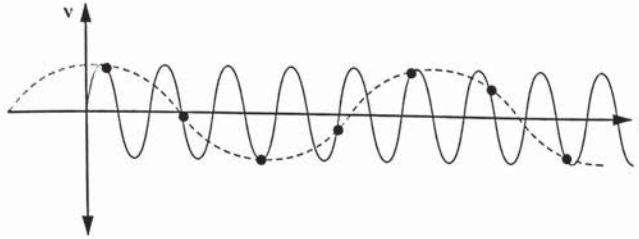


FIGURE 12.22. Illustration of subsampling.

property is particularly valuable in direct conversion receiver architectures, where such noise is often dominant.

To reduce the power consumed by the LO drivers, the gate capacitance of the switches may be resonated with an inductor (for narrowband applications), resulting in a power reduction by a factor of Q^2 . It is trivial to reduce the power to the order of a milliwatt or less, even at gigahertz frequencies.

12.5 SUBSAMPLING MIXERS

The high quality of CMOS switches has also been exploited to realize what are sometimes called *subsampling* mixers. This type of mixer exploits the observation that the information bandwidth of modulation is necessarily lower than the carrier frequency. Hence, one may satisfy the Nyquist criterion with a sampling rate that is also lower than the carrier frequency, effecting downconversion in the process.

As can be seen in Figure 12.22, the higher-frequency signal is sampled at the instants indicated by the dots, while the downconverted signal is shown as the lower-frequency reconstruction. The theoretical advantage of this approach is that it may be easier to realize samplers that operate at a frequency well below that of the incoming RF signal. From the figure, it should be clear that a properly designed track-and-hold circuit (Figure 12.23) serves as a subsampling mixer.³⁵

In the sample (track) mode, transistors M_1 through M_5 are turned on while transistors M_6 and M_7 are placed in the “off” state. Devices M_3 , M_4 , and M_5 put a voltage equal to the common-mode voltage level V_{CM} on the right-hand terminals of the sampling capacitors, while input switches M_1 and M_2 connect the capacitors to the RF input signal. Because M_6 and M_7 are open, the op-amp is irrelevant in this tracking mode, and the tracking bandwidth is simply set by the RC time constant formed by the total switch resistance and sampling (and parasitic) capacitance. Because the

³⁵ This example is adapted from P. Chan et al., “A Highly Linear 1-GHz CMOS Downconversion Mixer,” *IEEE J. Solid-State Circuits*, December 1993.

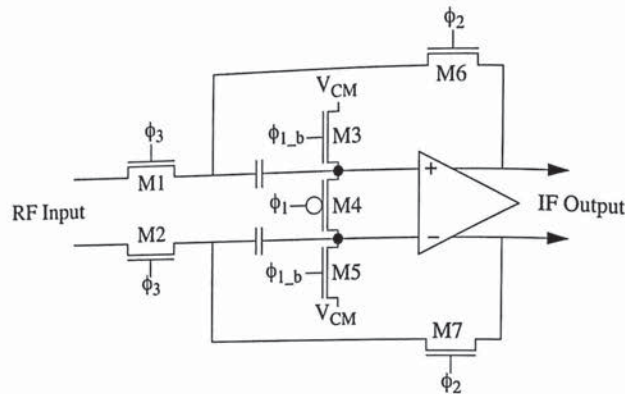


FIGURE 12.23. Track-and-hold subsampling mixer (simplified).

system operates open-loop in this mode, it is easy to obtain tracking bandwidths far in excess of what can be achieved with a feedback structure. For example, it is trivial to obtain tracking bandwidths greater than 1 GHz in a 1- μm technology.

In the hold mode, all switch states are reversed, so that the only conducting transistors are the two feedback devices M_6 and M_7 . In this mode, the circuit degenerates to a pair of charged capacitors feeding back around the op-amp. The settling time of this system need only be fast relative to the (slow) sampling period, rather than to the RF signal period. Thus, the bandwidth penalty associated with feedback is not serious.

Although a subsampler is clocked at a relatively low frequency, the sampler must still possess good time resolution or else sampling errors result. Therefore, beyond an adequate tracking bandwidth, one must also have low aperture jitter (i.e., low uncertainty in the sampling instants), and this requirement places extraordinary demands on the phase noise of the sampling clock. Hence, even though the frequency of the sampling clock need only satisfy the Nyquist criterion applied to the *modulation* bandwidth, its absolute time jitter must be a tiny fraction of the *carrier* period.

Another problem is that the sampling operation converts more than just the signal. Noise at the input to the sampler undergoes folding into the IF band, resulting in an unfortunate noise boost roughly equal to the ratio of RF and IF bandwidths. Because the RF bandwidth typically exceeds the IF bandwidth by large amounts, subsampling mixers can exhibit large noise figures (e.g., 25-dB SSB NF). The large linearity implied by the high third-order intercepts often exhibited by these types of mixers is offset by their poor noise performance, so that the dynamic range of the mixer is frequently no better (or even worse) than what one may achieve with conventional architectures. In fact, the noise and IP3 performance of many subsampling mixers can be replicated by preceding a conventional mixer with a resistive divider. In principle, an LNA with sufficient gain may be used to overcome the mixer's noise, but it

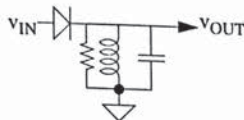


FIGURE 12.24. Simple diode mixer.

is difficult in practice to realize LNAs that provide simultaneously high gain and high linearity, so again overall (system) dynamic range may actually suffer. As a result of these problems, one must take great care in applying subsampling.

12.6 APPENDIX: DIODE-RING MIXERS

This appendix considers a number of passive mixers that are common in discrete implementations. The four-diode double-balanced mixer has particularly good characteristics, and is nearly ubiquitous in high-performance discrete equipment.

12.6.1 SINGLE-DIODE MIXER

The simplest and oldest passive mixer uses a single diode, as seen in Figure 12.24. In this circuit, the output RLC tank is tuned to the desired IF, and v_{IN} is the sum of RF, LO, and DC bias components. The nonlinear $V-I$ characteristic of the diode provides diode currents at a number of harmonic and intermodulation frequencies, and the tank selects only those at the IF.

It is tempting to reject this circuit as hopelessly unsophisticated. It does not provide any isolation, and it doesn't provide any conversion gain, for example. However, at the highest frequencies, it may be difficult to exploit other types of nonlinearities, and such simple mixers may be suitable. In fact, all of the detectors³⁶ for radar sets developed in WWII were single-diode circuits.³⁷ Additionally, many early UHF television tuners also used mixers of this type. Much of the modern work in the millimeter-wave bands simply would not be possible without such mixers.

As another note on this circuit, it can be used as a crude demodulator for AM signals if the input signal is the AM signal (at either RF or IF). When used in this manner, the output inductor is removed entirely, no LO is used, and a simple RC network provides the output filtering. Millions of "crystal" radio sets used this type of detector (known in this context as an envelope detector), and even most AM superheterodyne radios built today use a single-diode demodulator.

³⁶ We will use the terms "detector" and "demodulator" interchangeably.

³⁷ The birth of modern semiconductor technology can be traced directly to the development of microwave diodes for radar. By the end of WWII, point-contact microwave diodes capable of operation well into the gigahertz range became widely available.

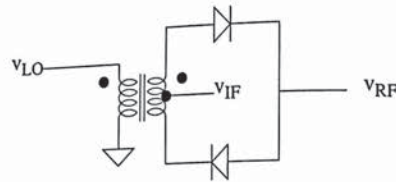


FIGURE 12.25. Single-balanced diode mixer.

12.6.2 TWO-DIODE MIXERS

There are several other ways to use diodes as mixers. As we'll see, it will appear that a diode bridge can be used as just about anything, depending on which terminals are defined as input and output and which way the diodes point.³⁸

With two diodes, it's possible to construct a single-balanced mixer. In this case, one may obtain isolation between LO and IF, but there is poor RF–IF isolation; see Figure 12.25. Assume that the LO drive is sufficient to make the diodes act as switches, regardless of the magnitude of the RF input. With a positive value for v_{LO} , both diodes will be on (note the reference dots on the transformer windings), effectively connecting v_{RF} to the IF output. When v_{LO} goes negative, the diodes open-circuit and disconnect v_{RF} . Hence, this mixer acts the same as the active commutating mixer studied previously.

The poor RF–IF isolation should be self-evident from the comment that the diodes connect the RF and IF ports together whenever the diodes are on. Similarly, it should be evident that symmetry guarantees excellent RF–LO isolation. Whenever the diodes are on, the RF voltage can only develop a common-mode voltage across the transformer windings, so no voltage can be induced at the LO port.

12.6.3 DOUBLE-BALANCED DIODE MIXER

By adding two more diodes and one more transformer, we can construct a double-balanced mixer to provide isolation among all ports (see Figure 12.26). Once again, assume that the LO drive is sufficient to cause the diodes to act as switches. In the circuit shown, the left pair of diodes is on whenever the LO drive is negative, whereas the right pair of diodes is on whenever the LO drive is positive.

³⁸ Diodes can even be used to provide *gain* by exploiting the nonlinear junction capacitance to make a thing known as a *parametric amplifier*. The nonlinearity can be used to transfer energy from a local oscillator (known as the *pump* in par-amp parlance) to the signal, instead of the more conventional transfer of power from a DC source to the signal frequency. Parametric amplifiers can be extremely low-noise devices, since only pure reactances are needed to make them work.

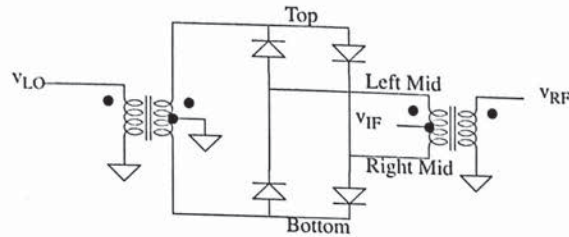


FIGURE 12.26. Double-balanced diode mixer.

With the LO drive positive, the voltage at “Right Mid” must be zero by symmetry, since the center tap of the input transformer is tied to ground. Thus, v_{IF} equals v_{RF} (again, note the polarity dots). With the LO drive negative, it is “Left Mid” that has a zero potential, and v_{IF} equals $-v_{RF}$. Hence, this mixer effectively multiplies v_{RF} by a unit-amplitude square wave whose frequency is that of the LO.

Isolation is guaranteed by the symmetry of the circuit. The LO drive forces a zero potential at either the top or bottom terminal of the output transformer, as noted previously. If the RF input is zero, there will be no IF output. Hence, this configuration provides LO–IF isolation. Similarly, we can show LO–RF isolation by considering a zero IF input. Since, again, there is a zero potential at either the top or bottom terminal of the output transformer, there will be no primary voltage and therefore no secondary voltage.

These passive mixers are available in discrete form, and perform exceptionally well. The upper limit on the dynamic range is typically constrained by diode breakdown, and isolation is a function of the matching levels achieved.

With a single quad of diodes, typical double-balanced mixers routinely achieve conversion losses in the neighborhood of 6 dB and isolation of at least 30 dB, and can accommodate RF inputs of up to 1 dBm at the 1-dB compression point while requiring an LO drive of 7 dBm. Higher RF levels can be accommodated if series connections of diodes are used in place of each diode of Figure 12.26, the drawback being an increased LO drive requirement to guarantee switching operation of the diodes. Using a total of 16 diodes, for example, extends the RF input range to around 9 dBm but also requires a whopping 13 dBm of LO drive.

12.6.4 FINAL NOTE ON DIODE MIXERS

When actually using such mixers, one should be aware that it is critically important to terminate all ports in the proper characteristic impedance – not only at the RF, IF, and desired LO frequencies, but at the image frequencies as well. If only narrowband terminations are used, it is possible for reflections of various intermodulation products

to degrade performance seriously. Hence, it is generally insufficient merely to use a standard RLC tank as an output bandpass filter without an intermediate buffering stage to guarantee a broadband resistive termination. Failure to satisfy this condition can be the source of many perplexing phenomena.

PROBLEM SET FOR MIXERS

PROBLEM 1 Using the device models from Chapter 3, design a single-balanced mixer with inductive source degeneration to achieve an IIP3 of +6 dBm. What is the conversion transconductance?

PROBLEM 2 We've seen the utility of synthesizing a virtual short-circuit for linearizing transconductances. Suppose that someone were to propose the alternative circuit of Figure 12.27.

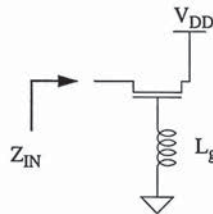


FIGURE 12.27. Common-gate transconductor with gate inductance (biasing not shown).

- First derive an expression for the incremental impedance looking into the source terminal.
- Select a value of inductance that makes the real part of the input impedance zero. Are additional elements needed at the input port to make the imaginary part also zero? If so, then what are they, and what are the expressions for their values?
- Having synthesized the virtual short, is this current recoverable as a drain current signal? If so, sketch how this circuit would be used to make a mixer.

PROBLEM 3 Simulate the version of the cross-quad transconductor in which the input signal drives the gates. Use minimum-length devices with a width of $100\ \mu\text{m}$, a total bias current of 4 mA (supplied from ideal current sources), and a resistance of $200\ \Omega$. Use the model parameters for the process described in Chapter 3.

- Initially ignore body effect, and measure the low-frequency transconductance with the output terminated in 4-V DC sources. Is it what you expect? Explain.

- (b) Now include body effect and repeat. Comment on any differences, and also determine the input common-mode voltage range over which the transconductance is roughly constant.
- (c) Find the input-referred third-order intercept from simulations using two equal-amplitude tones, at 95 MHz and 105 MHz, that possess a common-mode value in the center of the range found in (b). Note that this value must be reported as a voltage, since the input impedance is not specified. Be sure to observe the simulation caveats discussed in the chapter to avoid incorrect values.
- (d) Compare your answer to (c) with an estimate obtained from the three-point method.

PROBLEM 4 In this problem, we assess the relative merits of square-wave and sinusoidal local oscillator waveforms. Consider the specific case of an ideal multiplier used as a mixer in which the RF signal is $A \sin(\omega_{RF}t)$.

- (a) Which type of LO yields a higher conversion gain if they both have the same amplitude?
- (b) Compare the relative requirements on postmixing filters for the two LO signals.

PROBLEM 5 Consider the double-balanced passive mixer shown in Figure 12.28.

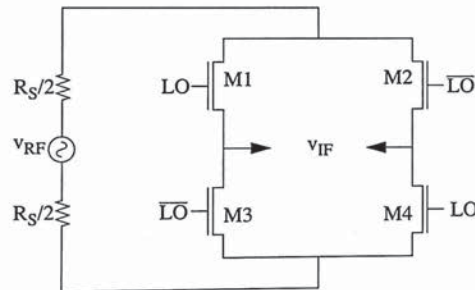


FIGURE 12.28. Simple double-balanced passive CMOS mixer.

- (a) If the IF port is terminated in a resistance equal to R_S , what is the conversion gain? Assume infinitely fast switching, and neglect switch resistance.
- (b) Because of the effective multiplication by a square wave, the IF output contains components at frequencies in addition to the desired one. Sketch the approximate output spectrum if the RF input is a single frequency sinusoid, and discuss filtering requirements.
- (c) How would your answer to (b) change if the LO drive did not possess a perfect 50% duty cycle? Put your answer on a more quantitative basis by explicitly

deriving an expression for the spectrum of a non-50% duty cycle square wave. Express your answer in terms of D , the duty ratio (here, ideally 0.5). What does your answer tell you about the need for symmetry in device switching?

PROBLEM 6 In the previous problem, switch resistance was neglected, a deficiency we now repair.

- Derive an expression for the maximum acceptable switch resistance if the degradation in conversion gain (relative to the ideal case) is not to exceed 1 dB.
- Provide an expression for device width corresponding to your answer in (a). For simplicity, you may assume square-law behavior. Express your formula in terms of the LO gate overdrive.
- Provide an expression for the total power consumed by the LO in driving the capacitance of all the switches as computed from your answer to part (b).

PROBLEM 7 The conversion gain equations and three-point method for estimating mixer IIP3 from transconductor IIP3 assume that the current-mode, LO-driven switches are perfect. However, as mentioned in the text, it is possible (perhaps even easy) for improper LO drive to cause degradation of conversion gain and significant distortion.

To explore the issue of imperfect switching in more detail, consider the simple differential MOS current switch used in a single-balanced mixer (see Figure 12.29). Simulate this circuit with an ideal current source drive, as shown in the figure. Let the DC bias current be 1 mA, and select an RF current amplitude of 100 μ A. With devices (use the level-3 models from Chapter 3) that are each 100- μ m wide, plot the amplitude of the 10-MHz IF component (measured as a differential output current into DC voltage sources of a value larger than the common mode of the LO drive) as the LO drive amplitude increases from 0 V to 1 V in 100-mV steps. How does the conversion gain vary? What does this experiment tell you about how one should select the LO amplitude?

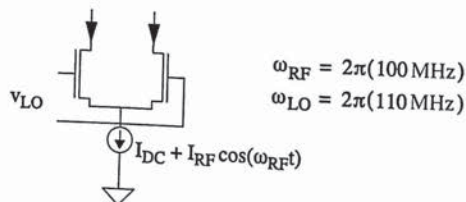


FIGURE 12.29. Single-balanced mixer.

PROBLEM 8 Repeat the previous problem, but instead of focusing on conversion gain, look at the output spectrum at other than the IF. What conclusions can you

draw? In particular, explore the effect of the source–bulk capacitance of the transistors. Set them to zero and resimulate. Compare and comment.

PROBLEM 9 Design a three-pair “multi-tanh” transconductor using the device models from Chapter 3. The specifications are a total allowed bias current of 5 mA and a low-frequency transconductance of 20 mS, with a maximum ripple of $\pm 10\%$. Verify your design with simulations.