Amr Fahim

Radio Frequency Integrated Circuit Design for Cognitive Radio Systems

Springer

Radio Frequency Integrated Circuit Design for Cognitive Radio Systems

Amr Fahim

Radio Frequency Integrated Circuit Design for Cognitive Radio Systems

Springer

Amr Fahim
Intel Corporation
Newport Beach
California
USA

ISBN 978-3-319-11010-3 ISBN 978-3-319-11011-0 (eBook)
DOI 10.1007/978-3-319-11011-0

Library of Congress Control Number: 2014958307

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

This book is dedicated to Sophia, Hana, Sarah, my sister, my parents and the memory of my grandparents.

Preface

Cognitive radios have become a vital solution that allows sharing of the scarce frequency spectrum available for wireless systems. It has been demonstrated that it can be used for future wireless systems as well as integrated into 4G/5G wireless systems. Although there is a great amount of literature in the design of cognitive radios from a system and networking point of view, there has been very limited available literature detailing the circuit implementation of such systems. Our textbook, *Radio Frequency Integrated Circuit Design for Cognitive Radios*, is the first book to fill a disconnect in the literature between Cognitive Radio systems and a detailed account of the circuit implementation and architectures required to implement such systems. In addition, this book describes several novel concepts that advance state-of-the-art cognitive radio systems.

Contents

About the Author

Amr Fahim received his B.A.Sc, M.A.Sc, and Ph.D. degrees from the University of Waterloo in Computer Engineer-ing in 1996 and Electrical Engineering in 1997 and 2000, re-spectively.

He has nearly two decades of experience in the design and modeling RF/mixed-signal ICs SoC for wireless applica-tions. His is the author of over 25 papers and over 15 patents. He is also the author of the book *Clock Generators for SoC Processors*. He has given numerous talks worldwide on wireless communications and circuit design. He has also served as a reviewer for the IEEE Journal of Solid-State Circuits, IEEE Transactions on Circuits and Systems II, and IEEE Transactions on VLSI journals.

Chapter 1
Introduction

The demand for energy and spectrum-efficient, low-cost portable devices has been one of the central focuses of wireless communications development. One technology that promises to deliver new levels of spectral efficiency is cognitive radios. In essence, cognitive radio is a system whereby wireless devices are aware of the spectrum usage and environment around them and are able to make intelligent decisions on how to most efficiently use and share the spectrum with other devices in the wireless network. Cognitive radios were first applied in military applications as a means of providing a two-way jamming-proof radio communications. It has since found its way to commercial applications as a means of more efficiently utilizing the frequency spectrum.

1.1 What Is Cognitive Radio?

One of the barriers to achieve higher data rates in wireless devices is the scarce availability of frequency spectrum. Figure 1.1 shows frequency spectrum allocation in the USA from 30MHz to 3 GHz [1]. As the figure shows, the available unallocated spectrum is extremely scarce. Higher operating frequencies are usually avoided due to worse atmospheric attenuation of radio frequency (RF) signals.

The truth of the matter is that although much of the usable frequency spectrum is already allocated, it is very poorly utilized. If a technology can be devised where the frequency spectrum allocation is dynamic instead of static, it can more efficiently utilize the available spectrum. One such technology is called *cognitive radios*.

Cognitive radio was first defined by Mitola as a "system where users are capable of *intelligently* share spectrum with other users by *dynamically* allocating *resources"* [2]. The resource that is typically shared is the frequency spectrum itself. This class of cognitive radios is known as "spectrum-sensing cognitive radios" (SSCR). The term "intelligently" implies that one must be able to first sense the surrounding environment before taking a specific action to yield a certain desirable result. In the context of a cognitive radio system, it must be aware of the frequency spectrum

A. Fahim, *Radio Frequency Integrated Circuit Design for Cognitive Radio Systems,*
DOI 10.1007/978-3-319-11011-0_1

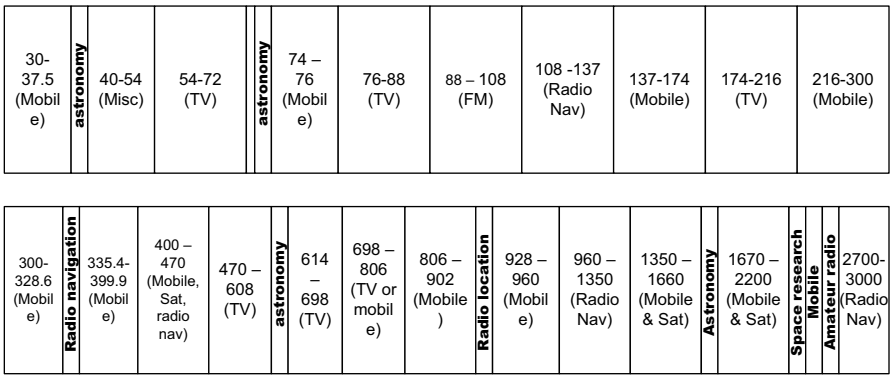


Fig. 1.1 Frequency spectrum allocation in the USA [1]

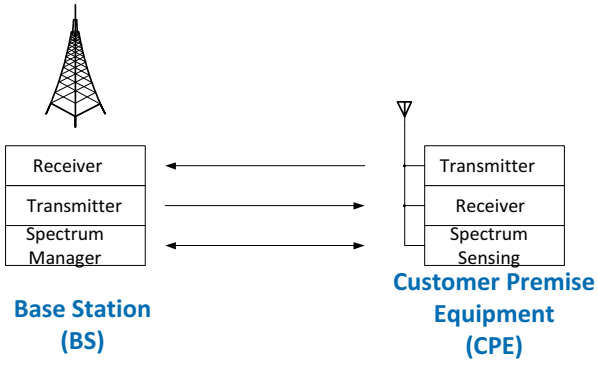


Fig. 1.2 Wireless cognitive radio device

usage and quality around it. This is usually accomplished by having a dedicated unit called a "spectrum-sensing unit." Once the spectrum is sensed, an intelligent decision would be made on what frequency to transmit the data on and at what amplitude level. Figure 1.2 shows the basic components of a wireless cognitive radio system. In this system, the customer premise equipment (CPE) is capable of sensing the environment around it. This information is sent to a base station (BS) where an intelligent decision is made regarding what frequency and amplitude the CPE can transmit data. The advantage of this approach is that the BS can collect data from several CPE and make a globally optimal decision for each CPE.

The cognitive radio must also be able to be sufficiently reconfigurable, especially in RF center frequency and channel bandwidth to accommodate the time-varying and dynamic nature of the wireless frequency spectrum. This quality is similar to another concept called *software-defined radio* (SDR) [3]. Although a necessary component of a cognitive radio, a cognitive radio encompasses the awareness, or cognition, of the wireless frequency spectrum it is immersed in. Figure 1.3 illustrates the relationship between a cognitive radio, SSCR, and an SDR. As the figure shows, an SDR is a component of a cognitive radio, but it can also be used in conventional wireless radios.

Table 1.1 Summary of radio definitions

Cognitive radio (CR)	Intelligently share spectrum with other users by dynamically allocating resources
Spectrum-sensing cognitive radio (SSCR)	Intelligently share the spectrum with other users by dynamically allocating frequency spectrum resources
Software-defined radio (SDR)	Radio device that has sufficient flexibility to set its radio parameters through software

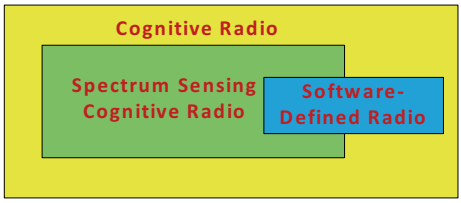


Fig. 1.3 Relationship between cognitive radio and software-defined radio (SDR)

In summary, there are three types of radios to be distinguished here: cognitive radio, SSCR, and SDR. The definition of each is summarized in Table 1.1. Throughout the rest of this textbook, the term "cognitive radio" will be used to describe a SSCR.

1.2 Brief History of Cognitive Radio

Cognitive radios were in use even before the term describing them was invented. As early as the 1980s, cognitive radios were used by the US military as a means of transmitting signals onto a secure frequency spectrum to avoid frequency jamming by the enemy.

The first studies for non-military use of cognitive radios appeared for disaster relief operations. For example, in 2000, a fireworks factory exploded killing 23 people in a small town in the Netherlands. This accident destroyed 400 homes and 15 streets and injured thousands of people [4]. Emergency relief services were not able to communicate with one another effectively due to limited frequency spectrum allocated to the emergency services, while spectrum in commercial cellular services was available. The ability to dynamically allocating such resource during emergency periods would have been advantageous.

It was not until 1999 when Mitola first published his paper where the term *cognitive radio* was coined. Since then, there has been significant research activity on how to build a cognitive radio system. The first major breakthrough for commercial cognitive radios came when analog television broadcast was being phased out from North America and much of Europe. This freed-up bandwidth was fought over vigorously by data transmission providers, while the television broadcast enti-

Table 1.2 Standardizing bodies for TVWS

IEEE 802.22	US standard that defines TVWS operation within the TV band
IEEE 802.11af	PHY and MAC layer for Wi-Fi operation in TVWS
ICogNeA/ECM392	European standard specifying PHY and MAC, mainly for home network applications in TVWS (HDTV streaming)
IEEE 802.15.4m	Amendments to the 802.15.4 to support TVWS in wireless personal area network (WPAN)
Infocomm Development Authority	TVWS task force in Singapore
OfCom	TVWS development in UK

IEEE Institute of Electrical and Electronics Engineers, *PHY* physical, *MAC* media access control, *TVWS* television white space, *HDTV* high-definition television

ties fought to keep control of this bandwidth for future use. This is nearly 1 GHz of bandwidth spanning from 48 to 860 MHz, or more in some parts of the world. The compromise solution reached is that data transmission would be allowed in the television band as long as it does not interfere with broadcast television transmission. The technology of choice to implement this system was cognitive radio. More specifically, the application of cognitive radio to the television band has been coined *television white space* (TVWS) technology [5].

Several standardizing bodies throughout the globe rushed to define a new standard to implement TVWS; some are listed in Table 1.2. There has been some limited deployment through much of the world using the different standards listed below. One development that is promising to take TVWS, or cognitive radios in general, to the next level is the widespread proliferation of Internet of things (IoT) devices [6]. Many IoT devices have the unique feature of having purely machine-to-machine (M2M) type of communication. M2M communications is characterized by short communication bursts that are very infrequent. There have already been several deployments of IoT devices equipped with cognitive radios in the market [7].

Another promising development is the use of cognitive radio concepts in mainstream cellular standards such as long-term evolution-advanced (LTE-A). Such schemes have recently been investigated and the results seem promising [8]–[10]. More specifically, the application of cognitive radio to such cellular systems is known as opportunistic spectrum access (OSA) [11]. In such systems, idle frequency bands are identified, thus allowing opportunistic and interference-free transmission with the primary system. In [12], a variant of LTE-A was proposed using OSA for the TVWS spectrum.

References

1. http://www.ntia.doc.gov/files/ntia/publications/2003-allochrt.pdf
2. J. Mitola III, G. Maguire, “Cognitive radio: making software radio more personal,” *IEEE Personal Communication Magazine*, vol. 6, no. 4, pp. 13–18, Aug 1999.
3. M. Dillinger, K. Madani, *Software defined radio: Architectures, Systems and Functions*, Wiley & Sons, 2003.
4. S. Castle, “After 400 homes and 15 streets are incinerated by fireworks blast, the Dutch ask: was it arson?,” *The Independent*, May 15, 2000.
5. C. Stevenson, et al., “IEEE 802.22: The first cognitive radio wireless regional area networks (WRANs) standard,” *IEEE Communications Magazine*, vol. 47, no. 1, pp. 130–1380.
6. A. McEwan and H. Cassimally, *Designing the Internet of Things*, West Sussex, UK, John Wiley & Sons, 2014.
7. “BT, Neul announce IoT Network in Milton Keyes,” *M2M Magazine*, May 23, 2014.
8. H. Cao, et al., “The design of an LTE-A system enhanced with cognitive radio,” *Proceedings of the 21st European Signal Processing Conference (EUSIPCO)*, 2013, pp. 1–5.
9. S. Lien, et al., “Cognitive radio resource management for future cellular networks,” *IEEE Wireless Communications*, vol. 21, no. 1, 2014, pp. 70–79.
10. D. Joshi, et al., “Dynamic spectrum shaping in LTE-Advanced cognitive radio systems,” *IEEE Radio and Wireless Symposium (RWS)*, 2013, pp. 19–21.
11. Q. Xiao, et al., “A unified approach to optimal Opportunistic Spectrum Access under collision probability constraint in cognitive radio,” *EURASIP Journal on Advances in Signal Processing*, 2010.
12. C. Herranz, et al., “Cognitive radio enabling opportunistic spectrum access in LTE-Advanced femtocells,” *IEEE Int'l Conference on Communications (ICC)*, 2012, pp. 5593–5597.

Chapter 2
Cognitive Radio Primer

In this chapter, wireless communication technologies relevant to cognitive radios are reviewed. This includes a discussion of some of the capacity limitations in wireless systems, nonideal effects in radio systems, and an overview of orthogonal frequency-division multiplexing (OFDM) encoding methods. Cognitive radios are detailed in this chapter. The architectural components as well as algorithms for frequency sensing and spectrum management are reviewed. Finally, some recent standardizations that are based on cognitive radios are reviewed along with a survey of some recent deployments of TV white space (TVWS) technologies, which is an implementation of cognitive radio.

2.1 Wireless Communication Technologies

One of the areas that has seen explosive growth over the past two decades is wireless communication technology. It took nearly a century of research and development to finally invent the radio in 1910. Research in this area started from the early works of electricity and magnetism that began in 1820 with the work of Hans Oersted [1]. It finally culminated in the practical invention of the wireless telegraphy in 1907 from the work of several people, including Guglielmo Marconi [2]. It took nearly another century to take this invention and provide true wireless connectivity between people where voice and data can be exchanged. The global launch of the Global System for Mobile Communications (GSM) in 1987 [3] marked the beginning of true wireless connectivity between people.

2.1.1 Wireless Connectivity Networks

A typical wireless connectivity network is shown in Fig. 2.1. This system is composed of two types of devices. A base station (BS) and a user premise equipment (PE). Both types of devices are capable of two-way communication. There is

A. Fahim, *Radio Frequency Integrated Circuit Design for Cognitive Radio Systems,*
DOI 10.1007/978-3-319-11011-0_2

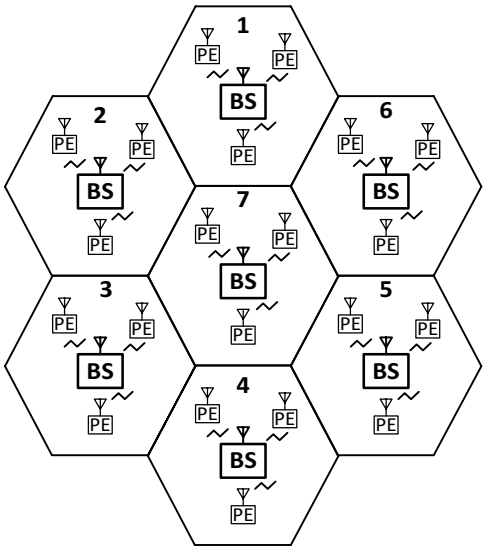


Fig. 2.1 Components of a basic wireless connectivity network

typically one BS per wireless communication area, called a cell unit. The PE typically only communicates with the BS, whereas the BS can communicate to any or all of the PEs in the network. The BS is responsible for allocating frequency spectrum to each PE. One way to increase coverage area is to increase the transmission power of the BS. Another way to increase coverage area is to add more BS units, each with a dedicated coverage area. To provide wireless communication to an entire area, the BS nodes are added to the network in such a way as to create adjacent but minimally overlapping wireless cell areas. Wireless communication in each cell area is regulated by the BS associated with that cell area. The size of the cell area depends on the physical terrain, maximum output transmission power of both the BS and the PE, and the receiver sensitivity of both the BS and PE.

In earlier cellular networks, a high-power BS would be used to cover a large cell area. It was noticed early on, however, that such networks suffer from several impairments. First, PE close to the BS can have their receivers saturated by the high output transmit power of the BS, thereby significantly reducing their sensitivity. The other major impairment is that the overall network capacity per unit area is limited since no wireless frequency spectrum can be reused within the cellular area. If the transmit power of the BS is reduced and the cell unit is broken up into several smaller cell units, then the wireless frequency spectrum can be reused and network capacity is significantly enhanced. This concept has been one of the driving technologies in the development of the next-generation wireless technologies, namely the migration of macro BS to small cells (pico BS and femto BS development) [4]. This concept is pictorially shown in Fig. 2.2.

There are other types of wireless networks that have received a significant amount of attention in both the academic and industrial circles. One such type of network is the unstructured wireless network. In this type of wireless network, the

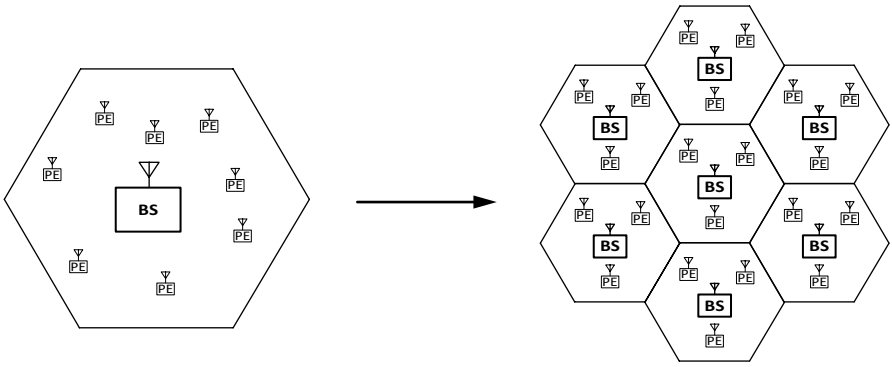


Fig. 2.2 Migration of macro base stations (*BS*) to small cells

location of a BS can be arbitrarily chosen or even mobile. An example of such a network is a wireless local area network (LAN, Wi-Fi) [5]. Another type of network is a distributed wireless network [6]. In this type of network, all nodes are typically identical and communicate with one another directly. In other words, they are self-configurable. Wireless sensor networks are an example of a distributed wireless network [7].

2.1.2 Wireless Channel Impairments

There are several impairments that a typical wireless channel suffers. A comprehensive treatment of all wireless channel impairments is beyond the scope of this book [8–11]. Only the more important impairments relevant to the discussion of cognitive radios are reviewed. The first such impairment is caused by the fact when a wireless signal is transmitted from an antenna; it is radiated outward in a spherical shape. This means that the received signal is a fraction of the transmitted signal and is proportional to the fraction of the surface area of the spherical wavefront that reaches the receiver. Assuming the receiver is represented as a point, the received power in mathematical terms is given as

$$P_r = P_t G_t G_r \frac{\lambda^2}{(4\pi d)^2} \tag{2.1}$$

where P_r is the received power, P_t is the transmitted power, G_t and G_r are the transmitted and received antenna gains, respectively, and λ is the wavelength of the electromagnetic signal, and d is the separation between the transmitter and receiver. Equation (2.1) is known as Friis formula, or the "path loss" formula, and is valid only for free space [12]. When atmospheric conditions are taken into account, the attenuation can become more severe and is frequency dependent. Figure 2.3 below shows the atmospheric attenuation as a function of frequency [13].

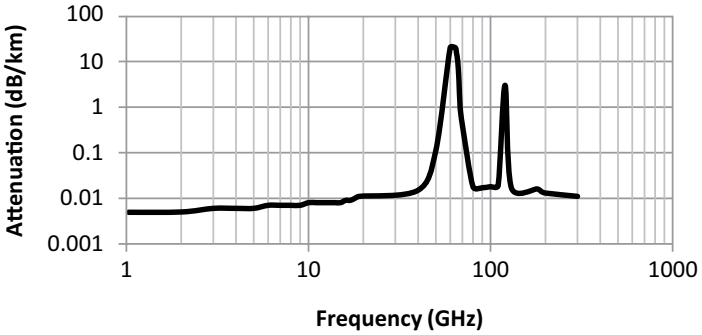


Fig. 2.3 Atmospheric attenuation as a function of frequency

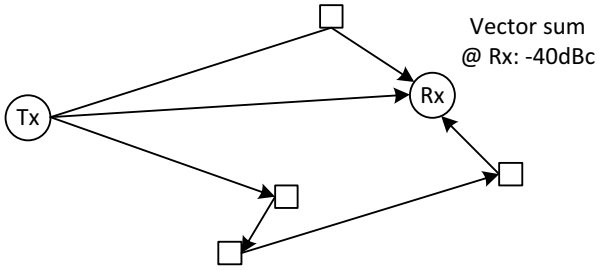


Fig. 2.4 Multipath fading in a wireless channel

Another important impairment is known as multipath fading. This is caused when the received signal at the antenna arrives from multiple paths, as shown in Fig. 2.4. The received signals arrive at the antenna at different times, and hence are offset from one another in phase, which can also be a time-varying phase change. In the worst case, two received signals are exactly 180° out of phase and hence perfectly cancel each other at the antenna. In practice, there may be a condition where several reflections of the same signal arrive at the antenna from different angles that cause the signal to be severely attenuated. This attenuation can be as much as 40 dB. Moreover, this attenuation can be time-varying, causing the signal to fade. This effect is known as multipath fading.

Another effect that can occur in a wireless channel is known as shadowing, shown in Fig. 2.5. This usually occurs in an urban environment, where there is a clear path of sight between the BS and the user, which is periodically blocked by moving objects, usually a truck or large vehicle. When the large object blocks the main path between the BS and the user, the user relies on secondary reflected paths, which are usually much less in amplitude, causing severe and sudden attenuation in received signal.

One important phenomenon that occurs in wireless communication is known as intersymbol interference (ISI). ISI can be caused by multipath propagation, where reflected versions of the signal undergo different phase distortions, causing certain symbols to smear together, as shown in Fig. 2.6. ISI can also be caused by band-

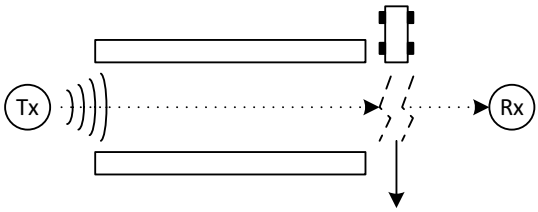


Fig. 2.5 Shadowing in a wireless channel

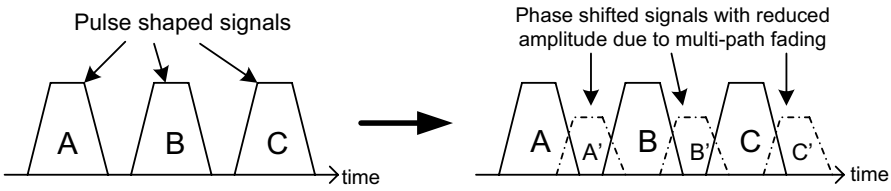


Fig. 2.6 Effect of ISI on received signal

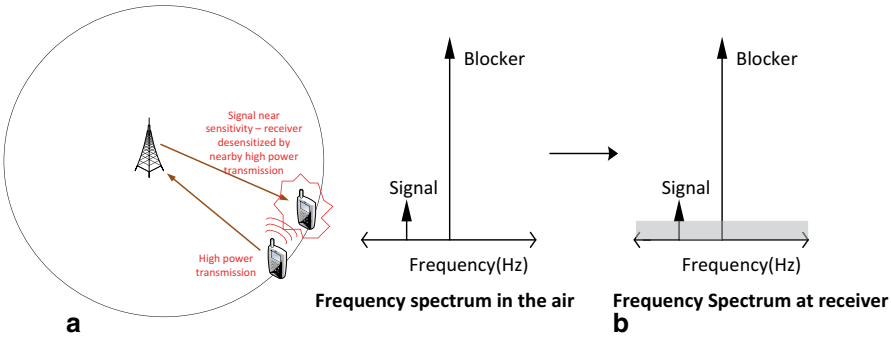


Fig. 2.7 Blocker desensitization in the receiver **a** scenario and **b** effect of receiver sensitivity

limited channels, whereby the high-end of the signal band receives more attenuation than lower-frequency components.

There are other impairments that result from sharing the frequency spectrum with other users. Blocker desensitization, shown in Fig. 2.7, is an important effect that occurs in wireless channels, especially in wireless cellular networks. It is caused when a nearby user is transmitting on a nearby channel with a large signal. The receiver detects the desired channel, but is also unable to completely filter out the high power transmission on the nearby channel. The undesired signal ends up saturating the receiver, reducing its sensitivity. Reduced receiver sensitivity reduction, in this scenario, is usually caused by the receiver, automatically reducing the gain in its front-end amplifiers to avoid the high power blocker from saturating the receiver. Figure 2.7a illustrates this scenario and Fig. 2.7b shows its effect on the receiver.

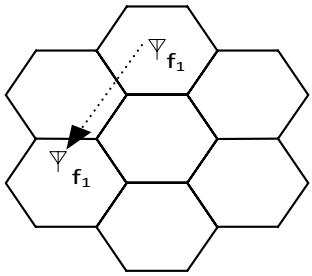


Fig. 2.8 Co-channel interference

Another impairment that is caused by sharing the frequency spectrum with other users is the co-channel interference, shown in Fig. 2.8. As stated earlier, one advantage of a cellular network is that a specific channel frequency can be reused by a different cell. This can result in a cellular network, where the same channel in the frequency spectrum is reused in another cell. Since, it can be seen from (2.1), the signal attenuation is finite, a small residual signal from one cell will leak into another cell. This signal leakage is worst if one cell is operating at maximum transmit power and the other cell is attempting to detect a very weak signal. Co-channel attenuation is usually minimized by disallowing two adjacent cellular areas to use the same channel. In other words, there is at least a two cellular area separation before the same channel is used, as illustrated in Fig. 2.8.

2.1.3 OFDM Primer

In order for two nodes to transmit digital data, the data must be encoded in a spectrally efficient method. One such method that has received widespread acceptance in many modern wireless communication networks is known as OFDM [14]. OFDM has been adopted by Digital Video Broadcasting (DVB)-T/T2 (terrestrial television), Institute of Electrical and Electronics Engineers (IEEE) 802.11a, g, n, ac, ad (wireless LAN), and Long-Term Evolution/Long-Term Evolution-Advanced (LTE/LTE-A, 4G cellular) standards, and many more [15–19]. OFDM has the advantages of high spectral efficiency, robustness to ISI, fading, and co-channel interference.

OFDM is a method of encoding digital data on multiple carrier frequencies, as shown in Fig. 2.9. The carriers are closely spaced and orthogonal to one another. Each carrier is modulated with the digital data to be transmitted. The encoding method on each carrier is quadrature amplitude modulated (QAM) for high data rates or phase-shift keying (PSK) for lower data rates. The close spacing of the carriers allows the transmitted signals to be treated as slowly modulated narrowband signals, which makes it possible to eliminate ISI [20].

In mathematical terms, the OFDM signal is given as

$$y(t) = \sum_{k=0}^{N-1} X_k e^{j2\pi kt/T} \tag{2.2}$$

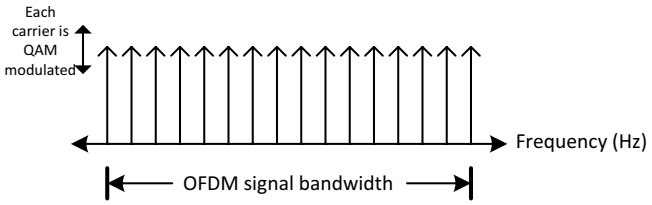
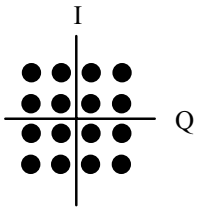


Fig. 2.9 A typical OFDM signal

Fig. 2.10 A 16-QAM signal



where X_k is the data to be transmitted, k is the number of carriers, and T is the period of the data symbol. It is important to note that $1/T$ is the frequency separation of the carriers in an OFDM signal. Recognizing that (2.2) is the in the form of a discrete inverse Fourier transform and restricting k to be a power of 2, an OFDM signal can be generated by simply running that data stream through an inverse fast Fourier transform (IFFT) operation. To guarantee orthogonality between all the carriers in an OFDM signal, the following condition must be satisfied:

$$\frac{1}{T}\int_0^T X_n e^{j2\pi nt/T} \bullet X_m e^{j2\pi mt/T} dt = 0, \quad m \neq n \tag{2.3}$$

which can be easily shown to be the case for any $m \neq n$.

Each carrier can be modulated by a QAM signal. A 16-QAM signal is shown in Fig. 2.10. A total of 16 symbols can be generated in the signal constellation, each symbol containing 4bits. In mathematical terms, a QAM signal is given as

$$y(t) = I(t) \bullet \cos(2\pi f_0 t) - Q(t) \bullet \cos(2\pi f_0 t) \tag{2.4}$$

where $I(t)$ and $Q(t)$ are amplitude modulated signals. For the 16-QAM example, $I(t)$ and $Q(t)$ would take one of four values, each.

One of the most important disadvantages of OFDM is the stringent linearity requirements on transceiver, due to the high peak-to-average power ratio (PAPR) in OFDM signals [21]. The gain of the receiver (and transmitter) is typically adjusted to track the average power of the signal. In order to accommodate the sudden peaks in amplitude, the transceiver must have high linearity. In order to understand why OFDM exhibits high PAPR, consider the two scenarios. In the first scenario, the data modulated on the carriers are entirely uncorrelated. Under this condition,

signal power at any given time is equal to the expected average power. In the other scenario, the data modulated on the carriers exhibit high correlation. Under the worst-case condition, all the carriers in the OFDM signal have the same phase and add up coherently. For this instance in time, the amplitude increases by 10 log *N,* where *N* is the number of carriers. Typically, a more statistical approach is taken to determining the PAPR of an OFDM signal. For example, if the data is considered random and has a Gaussian distribution, the 3σ peak number is usually used. The maximum peaking of the signal is known as the crest factor. In general, the crest factor of an OFDM signal is given as

$$CF = 10\log N + C_{fc} \quad \text{(in dB)} \tag{2.5}$$

where C_{fc} is the crest factor of each carrier individually, and N is the number of carriers in the OFDM signal.

2.2 Cognitive Radio Systems

Maximizing network capacity is one of the key goals of any network engineering exercise. In a cellular network, the cell size is reduced to maximize the frequency spectrum reuse. In a cognitive radio system, an alternative method of frequency spectrum reuse is used. As was shown earlier in this chapter, the frequency allocation spectrum is extremely congested, not allowing any new standards to emerge in the conventional radio frequency (RF) range of 30 MHz–3 GHz. The utilization of this allocated spectrum, however, is very poor. Better utilization of the frequency spectrum can be achieved if an incumbent user is allowed to temporarily use a spectrum that is already allocated to a different user. The incumbent user, however, must be able to continuously sense the frequency spectrum to detect when the primary user is present. At this point, the incumbent user must give priority to the original frequency spectrum owner and cease transmission. A wireless system whereby users have this *cognitive* ability is known as a cognitive radio system [22].

Cognitive radio systems are distinguished from conventional radio system in that they have the ability to sense the frequency spectrum around them and make intelligent decisions on how to best use the spectrum to maximize network capacity. Figure 2.11 shows the key components in a cognitive radio. Managing the spectrum in a dynamic fashion as well as the spectrum-sensing unit (SSU), shown in Fig. 2.11, are the distinguishing features of a cognitive radio. The figure shows that the spectrum managing is done by the BS [9]; although this is usually the case for most practical networks, in general, the customer PE can also be allowed to manage the spectrum.

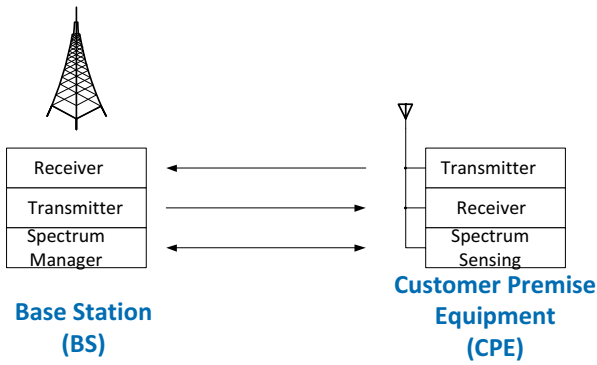


Fig. 2.11 Key components of a cognitive radio

2.3 Spectrum Management Techniques

A key issue in cognitive radio systems is how to reuse the frequency spectrum while not disturbing current transmissions. In other words, how can a cognitive radio system coexist with conventional static radio systems? There are three main methods that allow for this coexistence.

The first is known as a non-cooperative technique, shown in Fig. 2.12. Several types of PE are shown in this diagram, including fixed PE devices as well as portable PE devices. In this technique, each PE uses as much available spectrum as possible. Communication is only interrupted if the primary channel owner begins a transmission burst. The advantage of this approach is simplicity. The only type of channel analysis needed is if a channel is being used or not. Needless to say, this type of communication protocol is seldom used as it does not lead to the most efficient use of channel capacity.

Another more practical class of spectrum management method is known as a rule-based technique, shown in Fig. 2.13. In rule-based techniques, some centralized decision-making unit, usually the BS, is employed. Each PE would send its request for data bandwidth that is required as well as the PE's assessment of the frequency spectrum in its vicinity to the BS. Based on a set of rules, the BS would allocate frequency spectrum and instruct each PE to transmit at a certain level. Such rules may consist of static rules, such as maximum allowed bandwidth per user, limited transmit power levels, and checking with an online database to verify if a certain user is allowed to use the requested bandwidth (subscription fee-based). Other rule-based decisions may be dynamic, such as increasing the permitted power level when operating in a noisy channel, or decreasing the user's allowed bandwidth if several consecutive high bandwidth requests are made. Other rules are necessary to give the system a cognitive ability, such as following a listen-before-talk protocol. The advantage of rule-based techniques is that it provides relatively efficient use of the frequency spectrum with a small overhead in terms of decision making of how to efficiently use the spectrum. The disadvantage is that this system requires a centralized unit to allocate the frequency spectrum, namely the BS. Since the spectrum

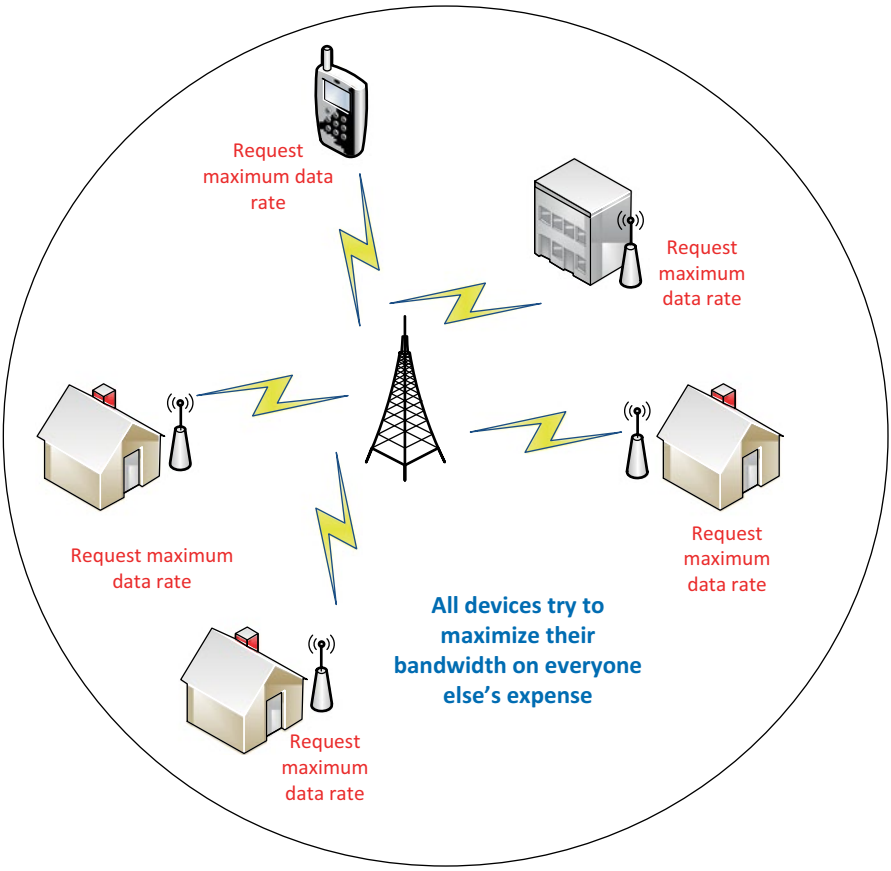


Fig. 2.12 Non-cooperative spectrum management technique

conditions near the BS may be different from the PE, the PE is required to send *its* analysis of the frequency spectrum. In other words, there is a centralized decision maker and all the PE can be thought of as frequency spectrum sensors without any decision-making capability.

The third category of spectrum management techniques is known as message-based techniques (Fig. 2.14). In message-based techniques, each PE is capable of making a decision on what frequency to transmit and with what power level. This decision, however, is made in coordination with nearby PE cells as well as the BS. This unique quality of the PE being able to make a decision independent of the BS can give a rise in spectrum usage efficiency (Fig. 2.14).

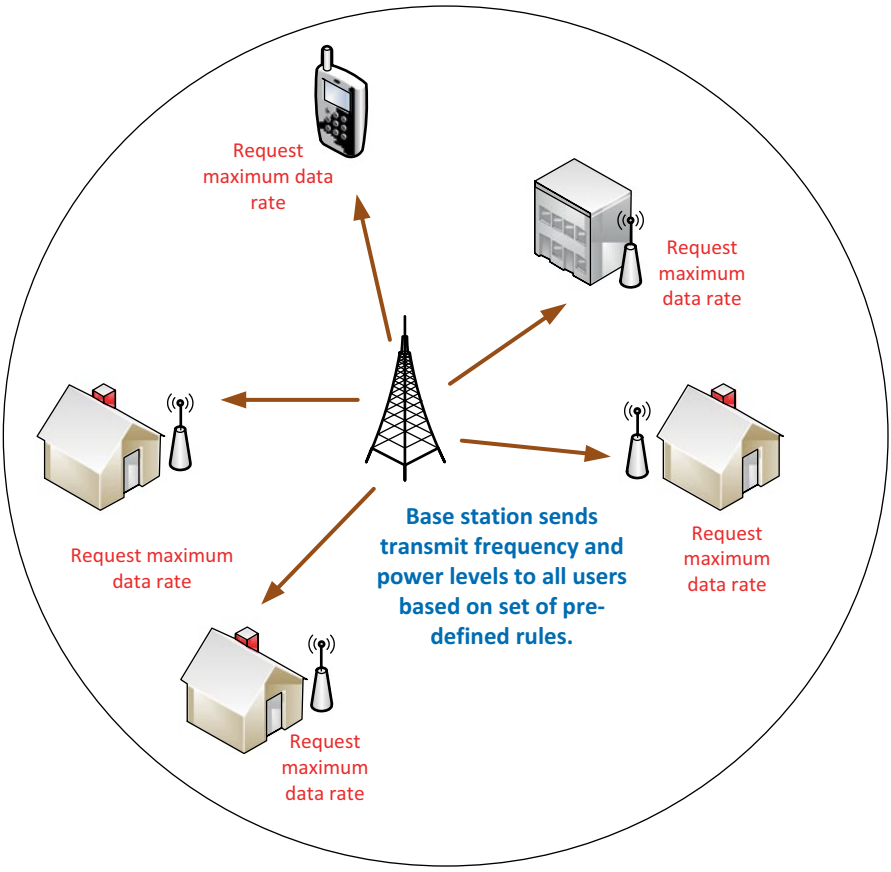


Fig. 2.13 Rule-based spectrum management technique

2.4 Spectrum Sensing

From a hardware perspective, the SSU is the distinguishing feature of a cognitive radio. As was discussed earlier, the PE should be able to sense its local frequency spectrum to detect if the channel owner is using its allocated frequency spectrum. The PE should also be able to detect if any other channels are available. To this end, the PE must be able to sense the entire frequency spectrum, also known as channel scanning, in addition to the channel that is currently used.

Sensing the current channel may be challenging. The obvious method of sensing the current channel is to require the PE unit to first cease transmission, then sense the current channel for the primary channel owner. This protocol is known as a blocking sensing method [23], as demonstrated in Fig. 2.15. In the first step, the transmission ceases on channel A, and instead the channel is sensed. The second step involves a decision process. If the primary channel owner is detected on channel A, then transmission is stopped and another channel is requested for transmission.

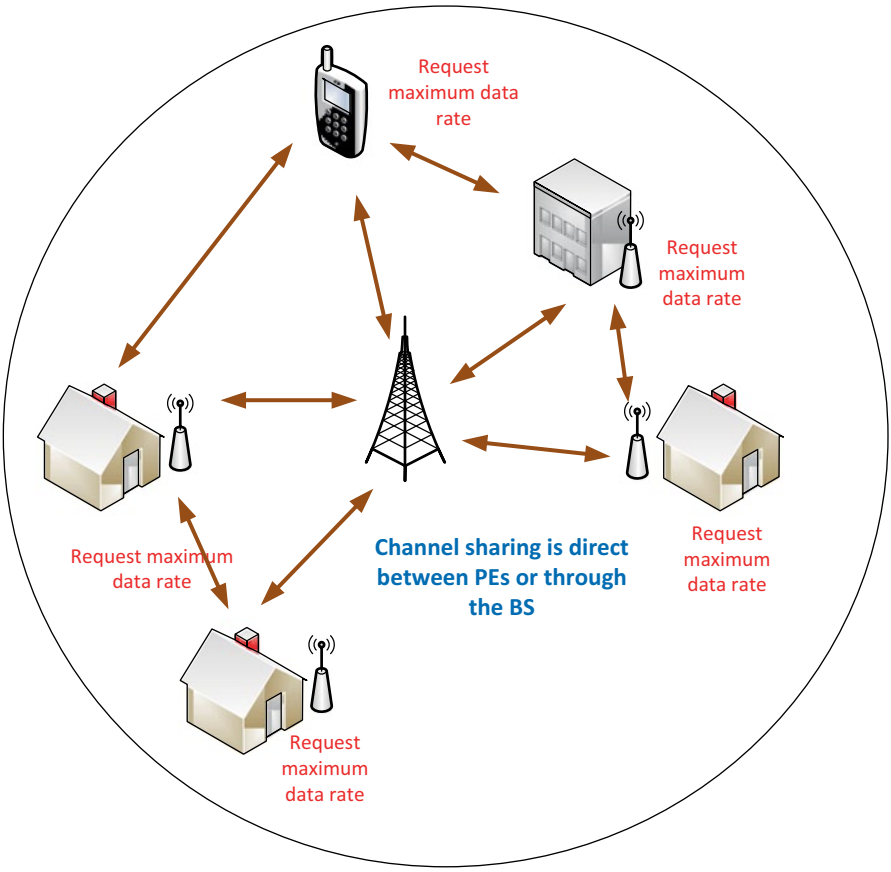


Fig. 2.14 Message-based spectrum management technique

Depending on the standard adopted, there is a minimum wait time where the channel must be vacant before the PE is allowed to transmit on the channel. For example, the IEEE802.22 standard (discussed in the next section) specifies a 30 s minimum time in which the channel must be vacant before used by the PE. Also, after the channel is used, it must stop periodically to ensure that the primary channel owner did not start transmitting on the channel.

The type of sensing algorithm used affects two important performance metrics: throughput penalty and latency penalty. Throughput penalty is the ratio of all the sensing times to the total transmission time plus the sense time. Latency penalty is the maximum duration of the sense operation. One way to minimize latency is to break up the sensing operation in to fast (and less accurate) sensing and slow (and more accurate) sensing [23]. If a signal is detected on the channel during the fast sense operation, then there is no need for an accurate sensing operation. If no signal is detected during a fast sense operation, then a more accurate, and slower, sense operation is required. This is illustrated in Fig. 2.16. To illustrate the effectiveness

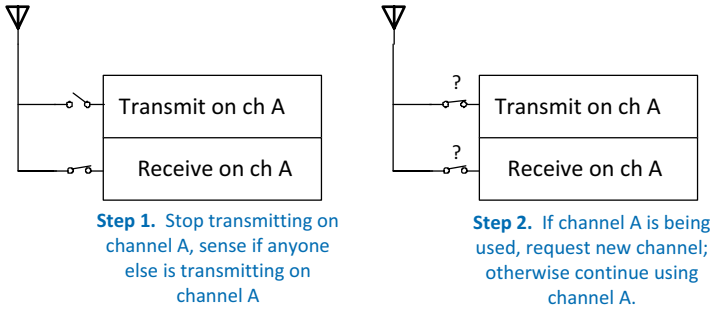


Fig. 2.15 Blocking sensing channel detection method

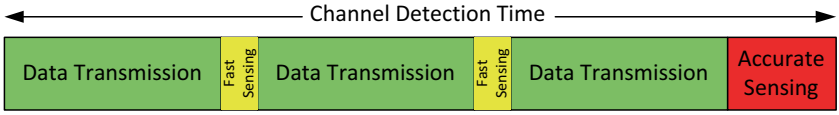
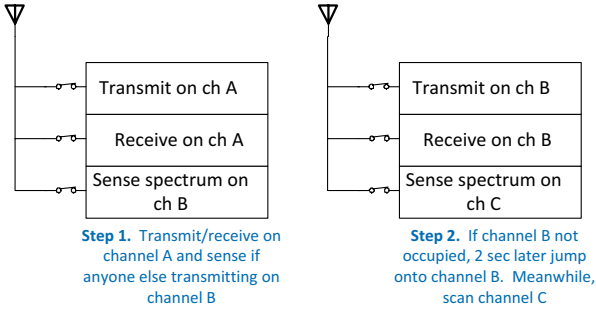


Fig. 2.16 Fast and slow sense operations to minimize the latency penalty

Fig. 2.17 Frequency-hopping channel detection technique



of this approach, consider a numerical example. If, for example, 500 ms of quiet time is required for the sense operation to complete, and the standard requires a channel sense every minute, the throughput penalty is calculated to be 0.5 %, but the latency penalty is 500 ms, a large number. If the sensing operation is broken up into 50 faster sense operations, the throughput penalty is unaffected (at 0.5 %). However, the latency penalty now is only 10 ms.

As was shown in Fig. 2.15, step 2 involves a decision process. If the primary channel owner is detected, then the PE must send a request for another channel to use. This means that the communication must be interrupted, thereby reducing the system throughput and increasing its latency. One method of avoiding this is by using a frequency-hopping technique [24, 25], shown in Fig. 2.17. This technique starts by transmitting and receiving on channel A, while simultaneously scanning for another available channel B. During the next sense operation, step 2, communication on channel A is stopped and resumed on channel B. While transmitting and

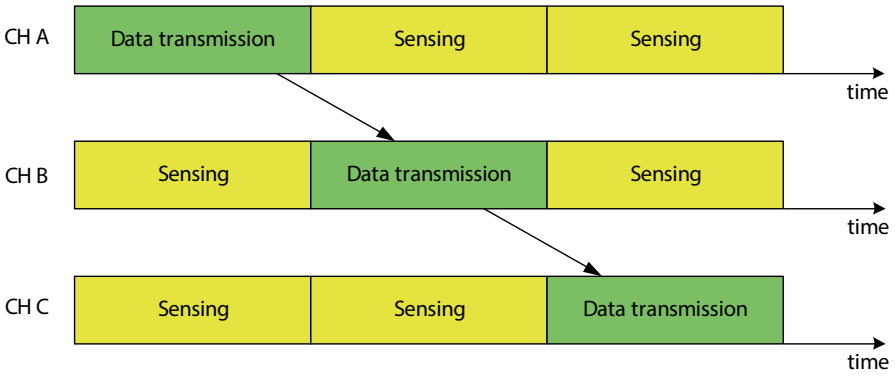


Fig. 2.18 Status of each channel over time for the frequency-hopping channel detection approach

receiving on channel B, yet another available channel C is detected, and prepared for use at the next sense time. This technique avoids requiring scanning the used channel for the primary channel owner, since the channel is given up at the next sense time iteration. For clarity, the status of each channel over time is shown in Fig. 2.18.

2.5 Signal Detection in Cognitive Radios

One of the primary functions of a SSU is to search for the primary user in the entire frequency spectrum [26, 27]. If the sensed signal is large, the detection operation is relatively simple. If the sensed signal is weak, however, the SSU must be able to distinguish a primary owner signal from random noise.

The simplest method of detecting whether a channel is used or not is to measure the energy in a given channel. This is particularly effective for signals that have signal-to-noise ratio (SNR) > 0dB. The energy of the channel is computed as

$$y(t) = \frac{1}{T}\int_0^T s(t) \bullet s^*(t) \bullet dt \tag{2.6}$$

where $s(t)$ is the signal over a channel, with bandwidth $1/T$, and $s^*(t)$ is the complex conjugate of the received signal. If the signal is very small, this operation must be repeated several times and averaged over several samples, in order to average out the random noise components. If the signal is weak, this operation can be very time consuming. In general, if N iterations of energy computations are performed, then 10 log N reduction of the signal detector's noise floor is possible [28].

Since the type of signal of the primary channel owner is known a priori, the signal detection operation can be made more intelligent. The class of techniques that

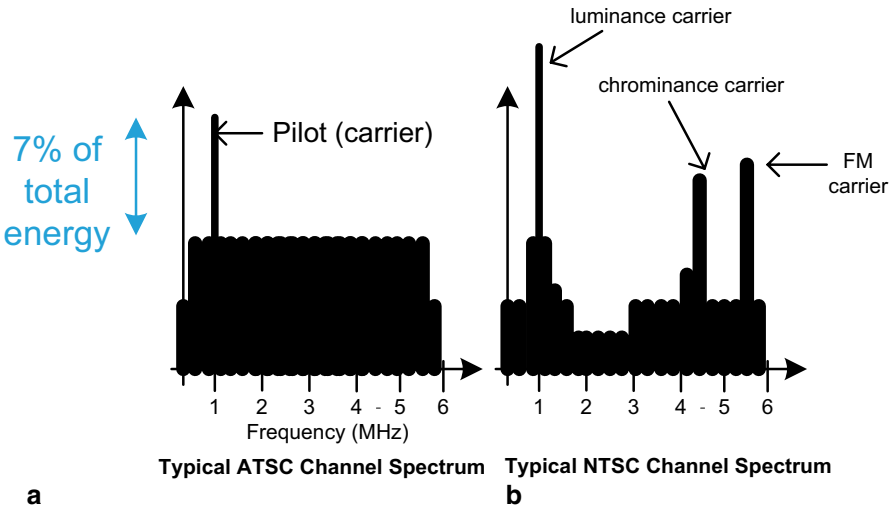


Fig. 2.19 Frequency spectrum of **a** ATSC and **b** NTSC TV signals. *ATSC* Advanced Television Systems Committee, *NTSC* National Television Systems Committee

use distinguishing features of the primary channel owner in order to make a decision as what whether a channel is used or not is known as feature-based techniques.

To illustrate the feature-based technique approach, consider television (TV) signals, shown in Fig. 2.19. A digital TV signal, Advanced Television Systems Committee (ATSC, Fig. 2.19a), has the distinguishing feature of having a pilot carrier at 310 KHz away from the band edge and 7% of the total power of the signal is contained in that carrier. The analog TV signal (Fig. 2.19b) has a distinguishing feature of having a large luminance carrier tone at 1.25 MHz away from the band edge. Since these features are narrowband, their power level can be measured accurately in a time-efficient manner.

A third type of signal detection method that has recently received widespread attention is based on measuring second-order signal statistics. The main assumption here is that any signals that are man-made are periodic and random signals are likely to be caused by natural phenomena. In other words, if the cross-correlation function of a given channel is computed, it should reveal any periodicity in the signal. If periodicity is detected, then it is assumed that the channel is being used; otherwise, the detected signal is random noise and the channel is available for use. As in energy-based detection, the noise floor of the detector can be improved by summing several cross-correlation measurements of the channel. The improvement in noise floor, however, is 20 log N, where N is the number of iterations [29]. A diagram of a cross-correlation-based signal detector is shown in Fig. 2.20.

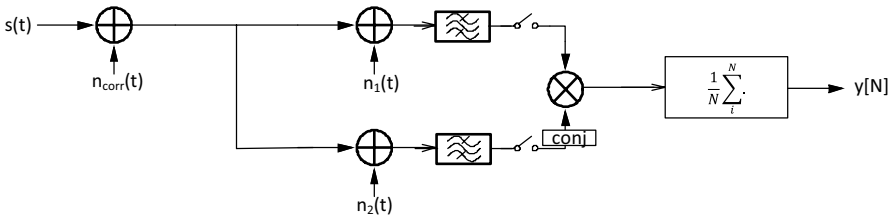


Fig. 2.20 Cross-correlation-based signal detector

2.6 Cognitive Radio Standardization

2.6.1 IEEE 802.22 and TVWS Technology

One of the first standardizations of cognitive radio was the IEEE 802.22 standard [30–33]. This standard describes how to apply a cognitive radio system to the TV broadcast band and is mainly applied in the USA. On June 12, 2009, high-power analog TV broadcasting was ceased in the USA and was replaced by digital TV broadcast. Several lower-power analog TV towers were allowed to operate until September 1, 2015. Wireless data providers, anxious to use this newly freed spectrum, attempted to lobby that two-way data communications is far more valuable to the public than digital TV broadcast. The result was a compromise solution, whereby data communication was allowed over the TV spectrum without any loss of TV signal quality and priority was given to TV broadcasters over the use of the spectrum. In addition to TV broadcasters, low-power wireless microphone users were allowed to transmit over the TV spectrum.

The relationship between cognitive radio, TVWS, and IEEE 802.22 is shown in Fig. 2.21. TVWS is an implementation of cognitive radio of the TV band, namely 40–860 MHz [34–43]. The IEEE 802.22 standard is not the only standardization of TVWS. There are others that have been adopted in Europe and Asia [44–47].

In terms of a data communication protocol, the IEEE 802.22 implementation of cognitive radio can be thought of as a long-range high data rate communication standard. The long range comes from the fact that the carrier frequencies in the TV

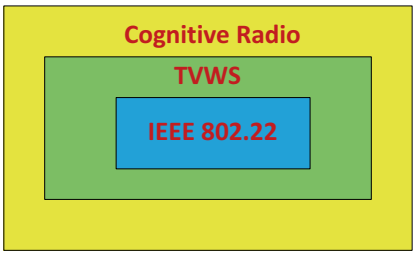


Fig. 2.21 Relationship between cognitive radio, TVWS, and IEEE 802.22. *TVWS* TV white space, *IEEE* Institute of Electrical and Electronics Engineers

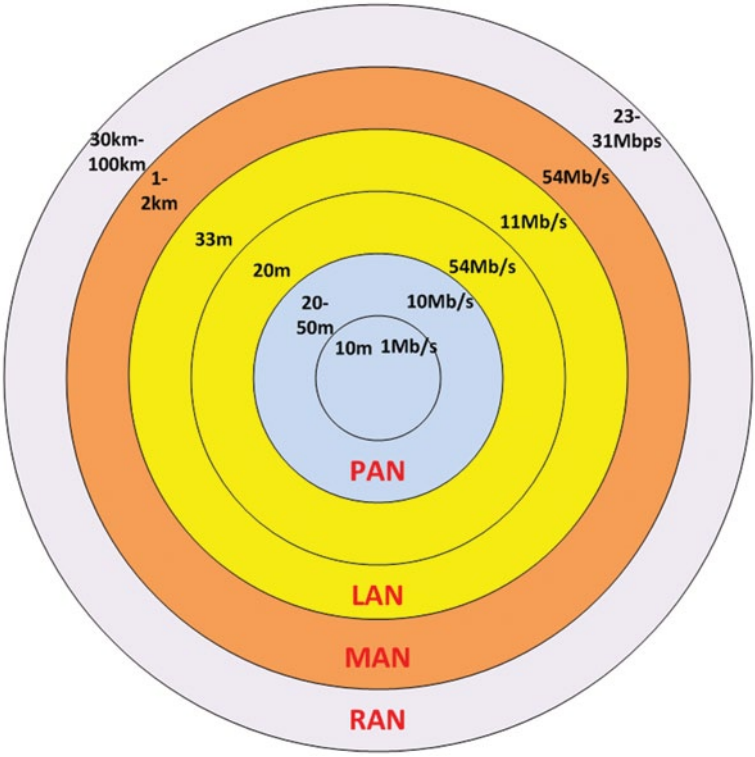


Fig. 2.22 IEEE wireless communication standards

band extend from 40 to 860 MHz. As was shown in (2.1), the received power is a function of the square of the carrier wavelength. Recognizing that $c=\lambda f$, where c is the speed of light constant and f is the carrier frequency, (2.1) can be rewritten as

$$P_r = P_t G_t G_r \frac{c^2}{(4\pi df)^2} \tag{2.7}$$

This demonstrates that the received power is inversely proportional to the carrier frequency. In other words, if the carrier frequency is reduced by a factor of 10, the received input power level is increased by a factor of 20, for the same antenna and receiver gain factors.

Figure 2.22 shows the IEEE 802.22 standard's relationship to other IEEE wireless standards. The ZigBee standard, IEEE 802.15, is a small-range low data rate standard, meant for applications such as wireless metering. The familiar Wi-Fi standard, IEEE802.11, is meant for higher data rates but limited range. The WiMAX standard, IEEE 802.16, which had limited success as a 4G wireless cellular standard, is meant for longer-range communication, up to 2 km, with high data rates of up to 54 Mb/s. Finally, the IEEE 802.22, TVWS, standard is meant for very high data rates over ranges up to 100 km.

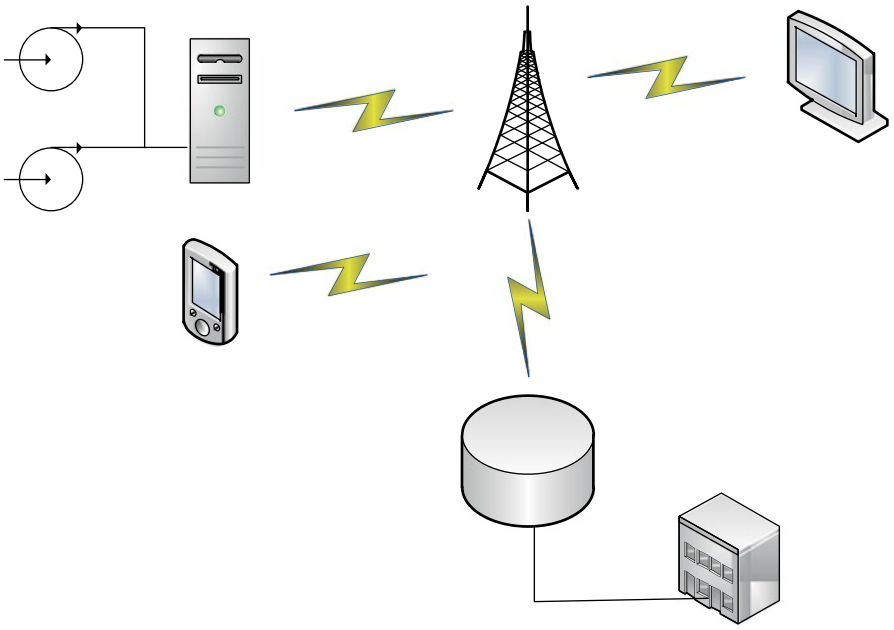


Fig. 2.23 IEEE 802.22 network applied to wireless metering

The IEEE 802.22 standard is meant to address several types of applications. The most obvious would be to use IEEE 802.22 as a form of long-range "Wi-Fi-like" for rural areas. Given the fact that TV broadcasting does not reach many rural areas in the USA, the unused frequency spectrum in the TV band can be shared between many cognitive PE. Data rates as high as 400–800 Mbps are possible.

Another application of IEEE 802.22 is for wireless metering [48]. The ZigBee standard (IEEE 802.15) is already being used for wireless metering. Wireless metering is characterized as having very low data rates and usually involves communication between wireless metering devices and an operator nearby. IEEE 802.22 standard would extend the range of wireless metering devices in such a way that the portable wireless metering devices can operate with a faraway BS directly, as shown in Fig. 2.23. The BS would then communicate to a database directly or an operator, if necessary. In other words, IEEE 802.22 applied to wireless metering offers the possibility of machine-to-machine (M2M) communication directly [49]. M2M communication is characterized as being very sporadic and bursty, ideal for TVWS technology, where cognitive PE devices transmit in bursts where the allocated frequency spectrum is not being used. TVWS devices optimized for wireless metering applications are currently available on the market have battery lifetimes measured in years [50].

Another application of IEEE 802.22 is for remote monitoring and supervisor control and data acquisition (SCADA), as shown in Fig. 2.24. This type of network would be applied to a limited urban area, such as a university campus or shopping center. The added value of an IEEE 802.22 standard is the direct connection of

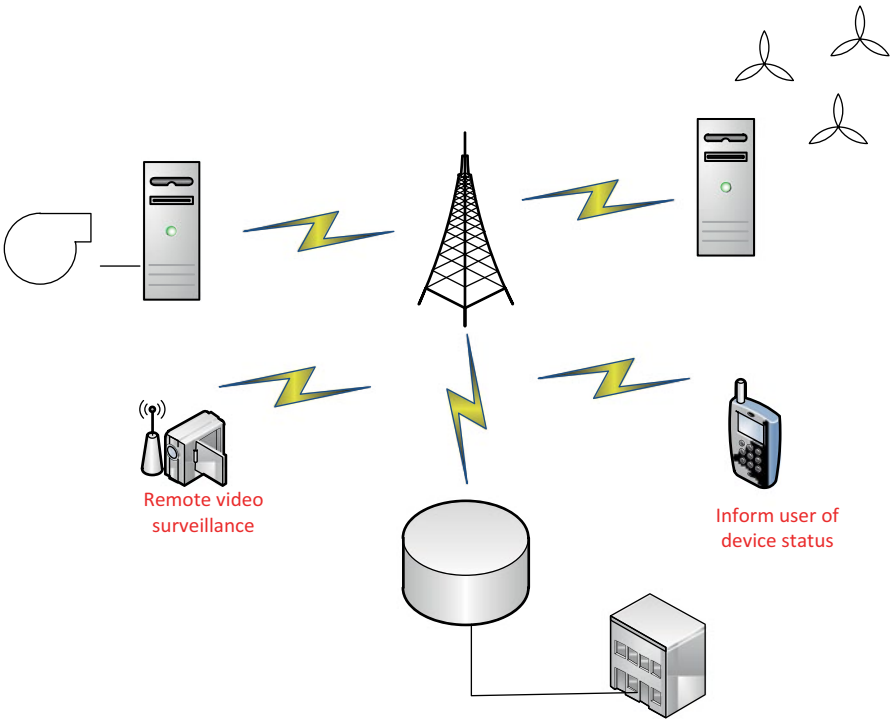


Fig. 2.24 IEEE 802.22 standard for SCADA application

SCADA devices to a BS device. As shown in Fig. 2.24, the communication in a SCADA network can be man to machine, machine to man, or machine to machine. In general, M2M type of communication is possible where remote machinery can be activated depending on analysis from remote video surveillance or based on data from a remote database.

Related to SCADA networks, remote medical patient monitor is possible with IEEE 802.22 networks, as shown in Fig. 2.25. Low-power patient monitoring equipment can transmit to the TV band directly. The BS has access to a patient database as well as the doctor or hospital monitoring equipment. In case of a medical emergency, the patient monitoring equipment would send a signal directly to the doctor/hospital monitoring equipment.

In general, it is possible to have a heterogeneous IEEE 802.22 network that serves multiple applications, as shown in Fig. 2.26. This figure shows that "Wi-Fi"-like users (both portable and fixed) can share a network along with long-range wireless metering and remote surveillance devices. The radius coverage of the IEEE 802.22 network is typically 30 km and can be as long as 100 km.

There are other competing standards for TVWS. Some of the most important are listed in Table 2.1 below. As the table shows, some of the standards are an extension of existing IEEE standards. Other standards, such as the European ECMA392 standard recognizes that cognitive radios are still in their infancy and limit the

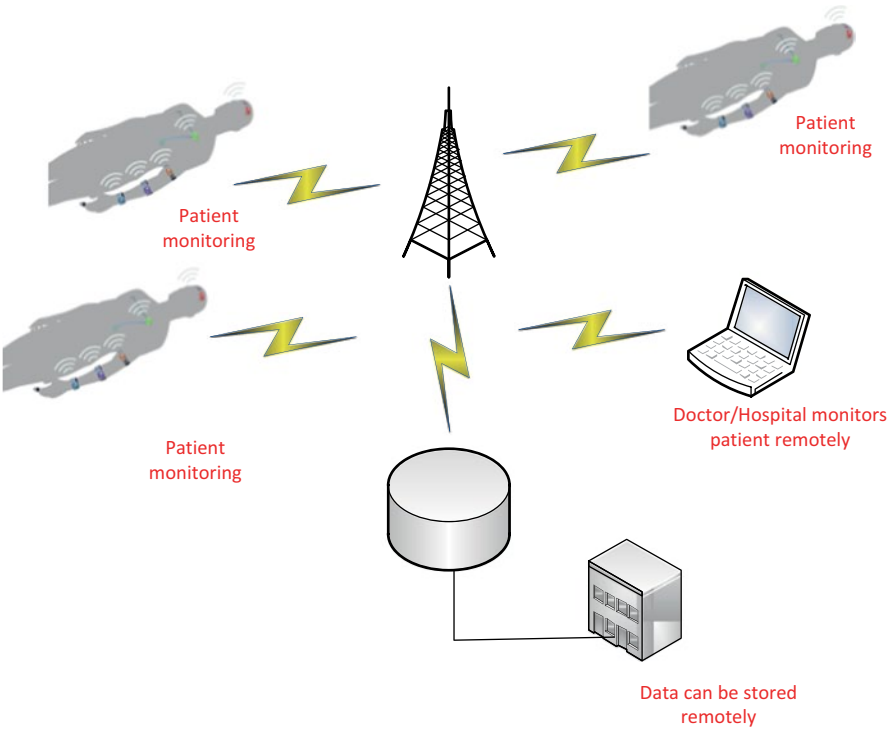


Fig. 2.25 IEEE 802.22 network for wireless medical sensing

application of cognitive radios to a very specific application. Yet others, such as the Infocomm Development Authority (IDA), target very specific applications in very specific regions.

2.6.2 TVWS Deployments

There have been many recent TVWS deployments worldwide. In February 2012, the first Federal Communications Commission (FCC)-approved IEEE 802.22 radio system was deployed in Wilmington, NC, in the USA. Wireless video cameras and remote sensors were installed for public utilities. Also, there were plans to provide public wireless access in public parks.

In September 2012, Singapore White Spaces Pilot Group (SWSPG), a consortium consisting of Microsoft, StarHub, and Institute for InfoComm Research, started a commercial pilot deployment of TVWS smart radio. The deployment was in the 700-MHz band and is marketed as a super-Wi-Fi access. The network extends to several nearby islands, providing a low-cost Wi-Fi access to the islands. The range per cell is 5 km, which is defined by the operating channel frequency.

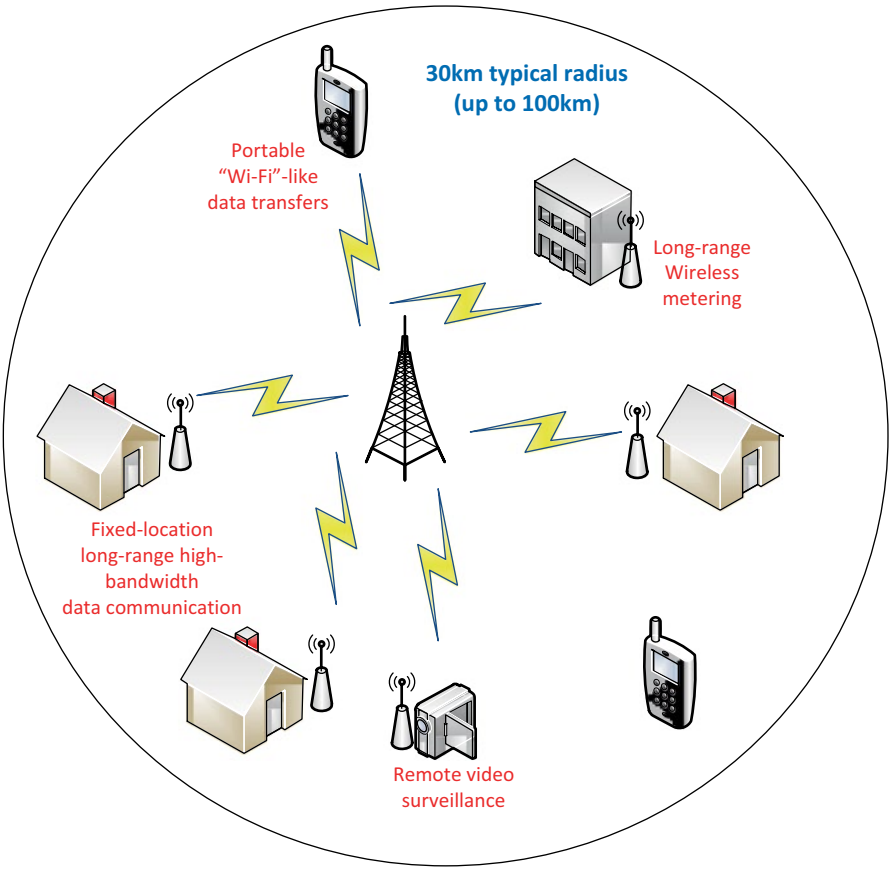


Fig. 2.26 Heterogeneous IEEE 802.22 network

Table 2.1 TVWS standards competing with IEEE 802.22

TVWS standard	Description
IEEE 802.11af	PHY and MAC for Wi-Fi operators in TVWS
ICogNeA/ECMA392	This is a European standard specifying the PHY and MAC, mainly for home network application in TVWS (HDTV streaming)
IEEE 802.15.4m	Amendments to the IEEE 802.15.4 support for TVWS in personal area network (WPAN)
Infocomm Development Authority (IDA)	TVWS task force in Singapore

TVWS TV white space, *IEEE* Institute of Electrical and Electronics Engineers, *PHY* physical, *MAC* media access control

In December 2012, a TVWS deployment was initiated in CapeTown, South Africa. Ten channels were made available for 2-Mbps internet access for schools over 10 km distances. Partners for this effort include Google, TENET, CSIR Meraka Institute, and e-Schools' Network and WAPA. It is also important to note that this is not an IEEE 802.22 compliant deployment.

Summary

In this chapter, wireless communication technologies relevant to cognitive radios were reviewed. This includes some of the capacity limitations in wireless systems, nonideal effects in radio systems, and an overview of OFDM encoding methods. Frequency reuse was identified as one key method of enhancing the network capacity of a wireless system. Cognitive radios were detailed in this chapter. It was shown how cognitive radios enhance network capacity by reusing time-sharing frequency spectrum with a priority rule-based algorithm. The architectural components as well as algorithms for frequency sensing and spectrum management were reviewed. Finally, some recent standardizations that are based on cognitive radios were reviewed. Although still in their infancy, the limited adoption these cognitive radio systems is creating valuable experimental data which will be used to further enhance cognitive radio systems.

References

1. Dibner, Bern, Oersted and the discovery of electromagnetism, New York, Blaisdell, 1962.
2. Bondyopadhyay, Prebir K., "Guglielmo Marconi—The father of long distance radio communication—An engineer's tribute". 25th European Microwave Conference, 1995. p. 879.
3. Redl, Siegmund M.; Weber, Matthias K.; Oliphant, Malcolm W. GSM and Personal Communications Handbook. Artech House Mobile Communications Library. Artech House, 1998.
4. Hunter, D.K., Mcguire, A., Parsons, G., "Mobile backhual for small cells," IEEE Communications Magazine, vol. 51, no. 9, 2013, pp. 60–61.
5. Molisch, A., "Wireless Local Area Network," Wireless Communications, pp. 731–750, 2011.
6. Phoha, S., Porta, T., Griffin, C., "Distributed Sensing and Data Gathering," Sensor Network Operations, pp. 421–508, 2006.
7. Zheng, J., Jamalipour, A., "Future Trends in Wireless Sensor Networks," Wireless Sensor Networks: A Networking Perspective, pp. 433–470, 2009.
8. A. Molisch, *Wireless Communications*, John Wiley & Sons: Hoboken, NJ, 2010.
9. D. Tse and P. Viswanath, *Fundamentals of Wireless Communications*, Cambridge University Press:Cambridge, UK, 2005.
10. V. Garg, *Wireless Communications and Networking*, Morgan Kaufmann Publishers: Boston, 2007.
11. T. Rappaport, *Wireless Communications:Principles and Practice*, Prentice-Hall: New York, 2002.
12. Friis, H.T., "A Note on a Simple Transmission Formula," *Proceedings of the IRE*, vol. 34, no. 5, 1946, pp. 254–256.

13. "Attenuation by Atmospheric gases," Recommendation ITU-R P.676–9, Feb 2012.
14. Li, Y., Stuber, G., *Orthogonal Frequency Division Multiplexing for Wireless Communications*, Springer: Boston, 2006.
15. E. Perahia and R. Stacey, *Next Generation Wireless LANs: 802.11n and 802.11ac*, Cambridge University Press:Cambridge, UK, 2013.
16. M. Gast, *802.11ac:A Survival Guide*, O'Reilly Media: Sebastopol, CA, 2013.
17. Y. Xiao and Y. Pan, *Emerging Wireless LANs, Wireless PANs, and Wireless MANs: IEEE 802.11, IEEE 802. 15, 802.16 Wireless Standard Family*, John Wiley & Sons:Hoboken, N.J., 2009.
18. M. Massel, *Digital Television: DVB-T COFDM and ATSC 8-VSB*, DigitalTVBooks.com, 2008.
19. C. Cox, *An Introduction to LTE:LTE-Advanced, SAE and 4G Mobile Communications*, John Wiley & Sons:Hoboken, N.J., 2012.
20. Montojo, J. and Milstein, L., "Effect of imperfection on the performance of OFDM systems," *IEEE Transactions on Communications*, vol. 57, no. 7, pp 2060–2070, 2009.
21. Wulich, D., "Definition of Efficient PAPR in OFDM," *IEEE Communications Letters*, vol. 9, no. 9, pp. 832–834, 2005.
22. C. Cordeiro, K. Challapali, D. Birru, and S. Shankar, "IEEE 802.22: The First Worldwide Wireless Standard based on Cognitive Radios," *DySPAN 2005*, pp. 328–337.
23. R. Balamurthi, H. Joshi, C. Nguyen, A. Sadek, S. Shelhammer, C. Shen, "A TV White Space Spectrum Sensing Prototype," *IEEE Int'l Symposium on Dynamic Spectrum Access Networks (DySPAN) 2011*, pp. 297–307.
24. J. Tong, H. Wu, C. Yin, Y. Ma and J. Li, "Dynamic Frequency Hopping vs. Non-Hopping in IEEE 802.22 Systems," *Proceedings of IC-NIDC 2009*, pp. 95–99.
25. R. Bizerra, A. Braga, and G. Carvalho, "A Spectrum Sensing Model for Continuous Transmission in Cognitive Radio Network," *Wireless Telecommunications Symposium (WTS) 2012*, pp. 1–7
26. Y. Hsieh, K. Wang, C. Chou, T. Hsu, T. Tsai and Y. Chen, "Quiet Period (QP) Scheduling Across Heterogeneous Dynamic Spectrum Access (DSA)-Based Systems," *IEEE Trans. On Wireless Communications*, vol. 11, no. 8, Aug 2012, pp. 2796–2805.
27. S. Lim, H. Jung, B. Jeong, "Efficient multi-channel signal detection algorithms for cognitive radio systems in TV white space," *IEEE Int'l Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM) 2012*, pp. 1–5.
28. Sequeira, S., Mahajan, R.R., Spasojevic, P., "On the noise power estimation in the presence of the signal for energy-based sensing," *IEEE 35th Sarnoff Symposium (SARNOFF)*, pp. 1–5, 2012.
29. Naraghi-Pour, M., Ikuma, T., "Autocorrelation-based spectrum sensing for cognitive radios," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 2, pp. 718–733, 2010.
30. C. Stevenson, G. Chouinard, Z. Lei, W. Hu, S. Shellhammer, W. Caldwell, "IEEE 802.22: The First Cognitive Radio Wireless Regional Area Network Standard," *IEEE Communications Magazine*, Jan 2009, pp. 130–138.
31. R. Al-Zubi, M. Siam, M. Krunz, "Coexistence Problem in IEEE 802.22 Wireless Regional Area Networks," *IEEE Global Telecommunications (GLOBECOM) 2009*, pp. 1–6.
32. M. Mollah, S. Islam, "Towards IEEE 802.22 Based SCADA System for Future Distributed System," *IEEE Int'l Conf on Informatics, Electronics & Vision (ICIEV) 2012*, pp. 1075–1080.
33. A. Fragkiadakis, E. Tragos, I. Askoxylakis, "A Survey on Security Threats and Detection Techniques in Cognitive Radio Networks," *IEEE Communications Surveys & Tutorials*, vol. PP, issue 99, 2012, pp. 1–18.
34. J. Kim, S. Lee, S. Kim, J. Ha, Y. Eo, and H. Shin, "A 54–862 MHz CMOS Transceiver for TV-Band White-Space Device Applications," *IEEE Trans. On Microwave Theory and Techniques*, vol. 59, no. 4, April 2011, pp. 966–977.
35. T. Baykas, M. Kasslin, M. Cummings, H. Kang, J. Kwak, R. Paine, A. Reznik, R. Saeed, S. Shellhammer, "Developing a Standard for TV White Space Coexistence: Technical Challenges and Solution Approaches," *IEEE Wireless Communications*, Feb. 2012, pp. 10–22.

36. S. Shellhammer, A. Sadek and W. Zhang, "Technical Challenges for Cognitive Radio in the TV White Space Spectrum," *Information Theory and Applications Workshop, 2009*, pp. 323–333.
37. T. Baykas, J. Wang, S. Filin, and H. Harada, "System Design to Enable Coexistence in TV White Space," *IEEE Wireless Communications and Network Conference (WCNCW) 2012*, pp. 436–441.
38. J. Choi, H. Chang, H. Choi, and W. Lee, "A Study on Interference Analysis Between DVB-T2 Broadcasting Service and TV White Space Device," *4th Int'l Conf. Ubiquitous and Future Networks (ICUFN) 2012*, pp. 234–235.
39. G. Villardi, C. Sun, Y. Alemseged, and H. Harada, "Coexistence of TV White Space Enabled Cognitive Wireless Access Points," *Workshop on Future Green Communications (WCNC) 2012*, pp. 18–23.
40. R. Dionisio, P. Marques, and J. Rodriguez, "Interference Study Between Wireless Microphone System and TV White Space Devices," *IEEE Cognitive Radio and Networks Symposium (ICC) 2012*, pp. 1874–1878.
41. C. McGuire, M. Brew, F. Darabi, S. Weiss, and R. Stewart, "Enabling Rural Broadband Via TV 'White Space'", *5th Int'l Symposium on Communications, Control and Signal Processing 2012*, pp. 1–4.
42. M. Kretschmer, C. Niephaus, G. Ghinea, "A Wireless Back-haul Architecture Supporting Dynamic Broadcast and White Space Coexistence," *21st Int'l Conf. on Computer Communications and Networks (ICCN) 2012*, pp. 1–6.
43. B. Gao, J. Park, Y. Yang, and S. Roy, "A Taxonomy of Coexistence Mechanisms for Heterogeneous Cognitive Radio Networks Operating in TV White Spaces," *IEEE Wireless Communications, Aug. 2012*, pp. 41–48.
44. J. Beek, J. Riihijarvi, A. Achtzehn, and P. Mahonen, "TV White Space in Europe," *IEEE Trans. On Mobile Computing*, vol. 11, no. 2, Feb 2012, pp. 178–188.
45. M. Rahman, C. Song and H. Harada, "Design Aspects of a Television White Space Prototype," *IEEE Vehicular Technology Conference (VTC Spring) 2012*, pp. 1–6.
46. R. Elliot, M. Enderwitz, K. He, F. Darabi, L. Crockett, S. Weiss, R. Stewart, "Partially Reconfigurable TVWS Transceiver for Use in UK and US Markets," *7th Int'l Conf. on Reconfigurable Comm.-Centric System-on-Chip (ReCoSoc) 2012*, pp. 1–6.
47. O. Herrera, A. Guiterez, A. Ospina, A. Galvis, "WRAN and LTE Comparison in Rural Environments," *IEEE Columbian Communications Conference (COLOM) 2012*, pp. 1–7.
48. S. Subramani, Z. Fan, S. Gormus, P. Kulkarni, M. Sooriyabandara, W. Chin, "WISEMEN: WhIte SpacE for smart MEteriNg," *IEEE Innovative Smart Grid Technologies (ISGT) 2012*, pp. 16.
49. Y. Zhang, R. Yu, M. Nekovee, Y. Liu, S. Xie, S. Gjessing, "Cognitive Machine-to-Machine Communications: Visions and Potentials for the Smart Grid," *IEEE Network, May/June 2012*, pp. 6–13.
50. Webb, W., "On using white space spectrum," *IEEE Communications Magazine*, vol. 50, no. 8, pp. 145–151, 2012.

Chapter 3
Wideband Receiver Design

An important element of cognitive radios is to be able to receive data signals from a wide frequency spectrum range. The wider the range, the more efficient use of spectrum becomes. In this chapter, techniques for wideband receiver design are discussed. The main components of a wideband receiver are detailed. This includes a discussion of wideband low-noise amplifiers (LNAs), high-performance radio frequency (RF) tracking filters, and image and harmonic reject mixers.

3.1 Receiver Metrics

Before approaching wideband receiver design, a few key metrics must first be discussed. The key metrics used to quantify the quality of a receiver design are signal dynamic range, frequency bandwidth, power consumption, and area. As Fig. 3.1 shows, all these metrics are interrelated and are usually traded-off with one another. For example, an increase in dynamic range (DR) may entail a reduction of bandwidth or an increase in power consumption, and vice versa. It is up to the radio frequency integrated circuits (RFIC) designer to select the optimal trade-off depending on the system being implemented.

To coexist with cellular and TV bands, the cognitive radio receiver must have a tuning range over the entire radio frequency (RF) range of such bands. Figure 3.2 shows the frequency allocation of the 44 bands that are used from universal mobile telecommunications system (UMTS) and long-term evolution (LTE) [31]. As the figure shows, the RF range is from nearly 450 MHz to 4.5 GHz. The TV bands around the world can range from 40 MHz to 1 GHz. This gives nearly two decades of frequency range of 40 MHz to 4.5 GHz, which is quite challenging. The receiver must have a wide-enough DR to be able to handle large blockers over a wideband, while still being able to receive a weak desired signal. Such requirements are usually specified by regulatory constraints such as spectral mask requirements and receiver sensitivity requirements as well as transmitter power constrains such as equivalent isotropically radiated power (EIRP) and other regulatory requirements for white

A. Fahim, *Radio Frequency Integrated Circuit Design for Cognitive Radio Systems*,
DOI 10.1007/978-3-319-11011-0_3

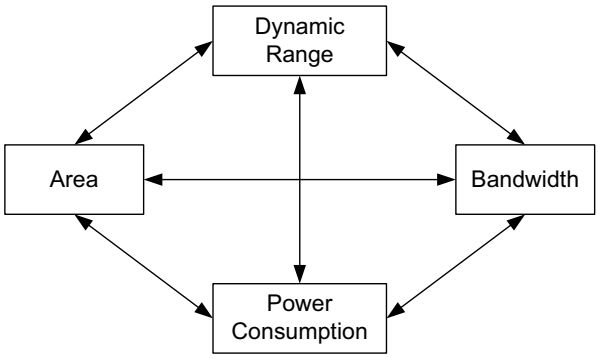


Fig. 3.1 Design metrics for wideband receivers

spaces devices that are specified by organizations such as the Federal Communications Commission (FCC) and the Office of Communication (OfCom) [27, 34].

Two important specifications that affect the receiver DR are the blocker profiles and sensitivity requirement. The sensitivity of an RF receiver is defined as the minimum detectable signal with an acceptable signal-to-noise ratio (SNR). A measure of noise of the receiver system can be given as a ratio of input-to-output SNR [64]:

$$F = \frac{SNR_{\text{in}}}{SNR_{\text{out}}} \tag{3.1}$$

where F is the noise factor. Noise figure (NF) is equal to $10 \cdot \log_{10}(F)$. If the receiver is impedance matched to a source resistance of R_s, (3.1) can be rewritten by expanding SNR_{in} to $P_{\text{sig}}/P_{\text{n,}Rs}$, where P_{sig} is the input signal power and $P_{\text{n,}Rs}$ is the noise power of a source resistor R_s. The available noise power of the source resistor R_s to the receiver can be given as:

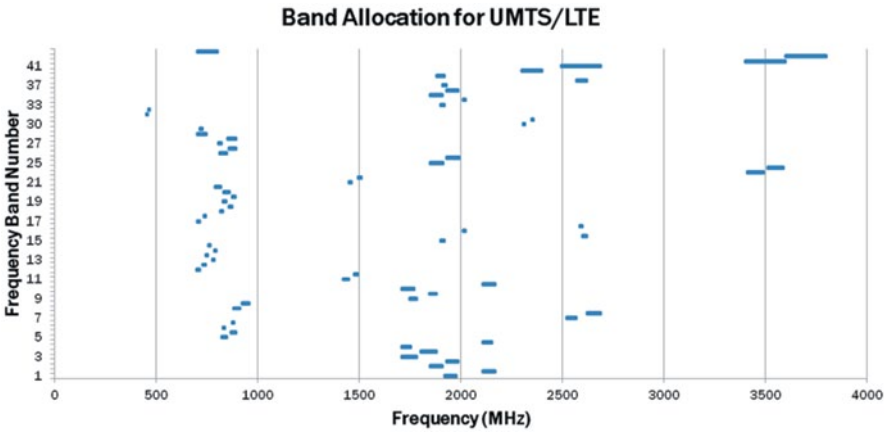


Fig. 3.2 Frequency band allocation for universal mobile telecommunications system (*UMTS*) and long-term evolution (*LTE*)

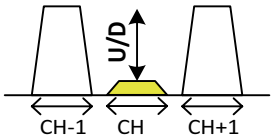


Fig. 3.3 Adjacent channel selection (*ACS*) specification

$$P_{Rs,av} = \frac{4kTR_s}{4} \cdot \frac{1}{R_{in}} = kT \tag{3.2}$$

where R_{in} is the input impedance ($=R_{\text{s}}$), k is Boltzmann's constant, T is temperature (in Kelvin). Using (3.2), (3.1) can be rewritten (in a decibel scale) as:

$$P_{sig,tot} = P_{Rs,av} + NF + SNR_{out} + 10 \cdot \log(BW) \tag{3.3}$$

where BW is the noise integration bandwidth, which in the context of a receiver system is one-channel bandwidth. Computing kT at room temperature, 290 K, (3.3) can be re-written as:

$$S_{min} = -174dBm / Hz + NF + SNR_{min} + 10 \cdot \log(BW) \tag{3.4}$$

where SNR_{min} is the minimum SNR to detect the signal with an acceptable bit error rate (BER) and S_{min} is the minimum sensitivity of the receiver. As (3.4) shows, the receiver sensitivity depends directly on the NF, which is a parameter under the control of the circuit designer.

In wireless systems, a variety of blocker profiles are specified, which then primarily determine the DR requirements of the receiver. One such requirement is the adjacent channel selection (ACS) [50], which is specified as shown in Fig. 3.3. In this scenario, strong undesired blocker signals are only one channel away from a weak desired signal, which is marked by "CH" in Fig. 3.3. The receiver signal is smaller than the adjacent channels by an amount specified as an undesired-to-desired (U/D) ratio, usually expressed in a decibel scale. In conventional receiver design, the adjacent channel cannot be sufficiently filtered at the RF. This means that any RF amplifiers in the receiver chain must be able to handle such large blockers without producing nonlinear intermodulation tones that can interfere with the desired signal. When a blocker results in intermodulation terms that rise above the desired signal, the blocker is said to *desensitize* the receiver. The noise floor must also be kept sufficiently low for proper reception of the desired signal, as specified by (3.4).

Another blocking profile, which is an extension of ACS, is known as narrowband blocking [14]. In this scenario, the U/D ratios of several channels near the desired signal are specified, as shown in Fig. 3.4. The further away the channel is, the higher the U/D ratio specification becomes. As before, the intermodulation terms resulting from the blocker must not interfere with the desired signal reception.

Another important blocker profile receiver system is known as receiver cross-modulation distortion (XMOD) [9]. This modulation scenario, shown in Fig. 3.5,

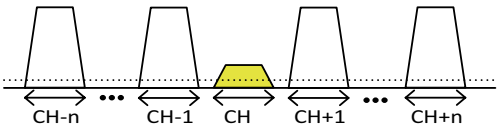


Fig. 3.4 Narrowband blocking specification

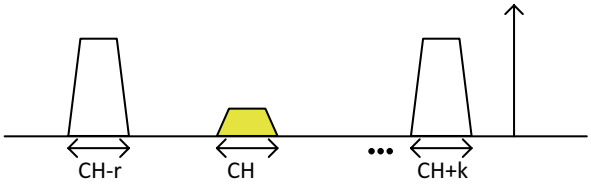


Fig. 3.5 Receiver cross-modulation scenario

involves a single-tone carrier wave (CW) blocker, along with one or two modulated blocker signals (denoted here as CH + k, CH − r). As with the other scenarios, the desired signal is the weak desired signal at frequency CH. Due to a third-order distortion, the CW can modulate with one or both of the blockers to produce tones within the band of interest, CH. This scenario is especially important in standards where duplex communication is allowed, such as code division multiple access (CDMA), UMTS, and some modes of LTE [29]. In duplex systems, one of the modulated blockers can be the transmitted signal from the same device where the receiver is located.

A typical front-end wireless receiver, a direct conversion receiver [61], is shown in Fig. 3.6. As the figure shows, the first block in the receiver is the low-noise amplifier (LNA) followed by an RF filter. This is followed by a pair of quadrature mixers

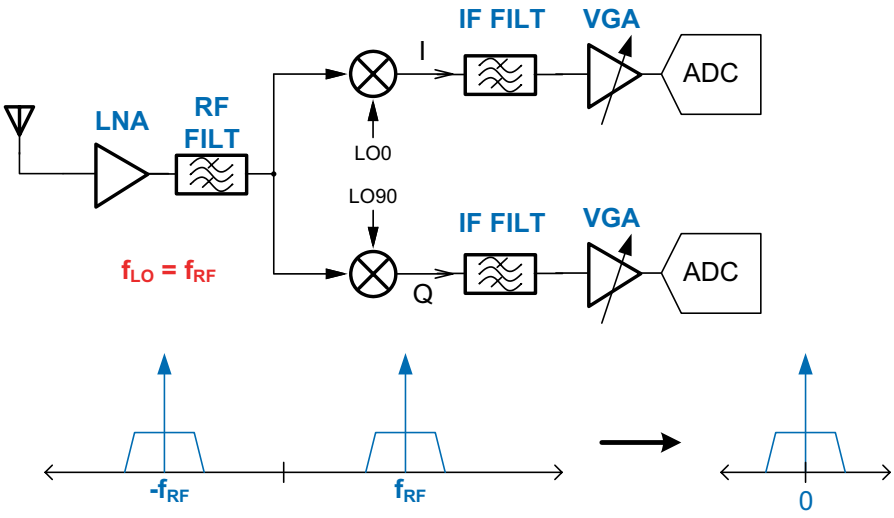


Fig. 3.6 Typical direct conversion front-end wireless receiver

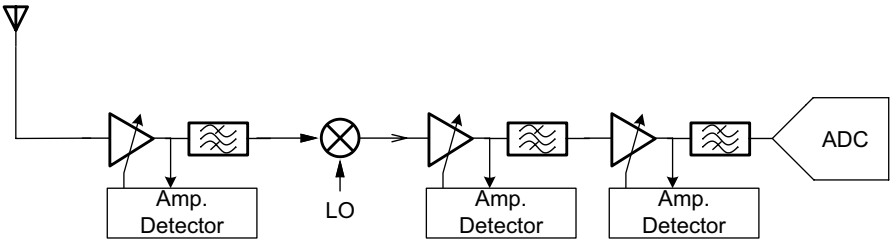


Fig. 3.7 Conceptual diagram of an receiver with optimal gain stages

that downconvert the desired signal into a complex (I/Q) signal. The baseband (BB) chain then further filters the undesired blockers and changes the gain of the received signal such as to fit the DR performance of the analog-to-digital converter (ADC).

The position of the LNA and RF filter may be interchanged depending on the system requirements of the receiver. The main trade-off is that an LNA first system results in a low-noise receiver, but may be susceptible to desensitization due to large blockers. An RF filter first system provides better immunity to blockers further away from the desired signal, but on the expense of the receiver's minimum sensitivity.

3.2 Receiver Gain Control

The concept of variable gain control in a receiver is central to maximizing the performance of the receiver. The receiver must be able to maximize its sensitivity for the weak faraway signals. In the presence of strong nearby transmitters, however, the sensitivity requirement of the receiver is somewhat relaxed. This means that the gain of each block in the receiver should be maximized in such a way that it does not desensitize the receiver. Consider a conceptual receiver shown in Fig. 3.7. There are three main components in each stage in the receiver: a variable gain amplifier (VGA), a filter, and an amplitude detector. Each VGA is followed by a filter block. The output of the VGA is fed into an amplitude detector, which then sets the optimal VGA gain in such a way that its output amplitude is maximized while avoiding in-band intermodulation terms. The bandwidth of the gain control loop is selected in such a way that it introduces no significant abrupt amplitude shift in the received signal. After the downconversion mixer, the BB filters have sufficiently sharp responses to filter out adjacent blockers.

The above analysis assumes a continuous gain range in the VGAs. As will be seen in the next section, the penalty of continuous gain control range is the added noise and nonlinearity to the amplifier stage (i.e., worse dynamic range). This issue is especially pronounced for RF amplifiers. For this reason, wireless receivers also make use of discretely controlled amplifier gain settings, known as programmable gain amplifiers (PGAs). Although PGAs have superior dynamic range, abrupt gain shifts can cause system-level issues in the receiver. To illustrate this point, consider,

for example, a system where the RF amplifier shown in Fig. 3.7 is a PGA with 6 dB step sizes. This means that if the gain of the amplifier is abruptly increased, subsequent amplifiers would be required to handle a 6-dB-higher signal level before the downstream VGAs have sufficient time to respond. Conversely, if the gain of the amplifier is abruptly lowered, the noise of the subsequent stages would be more pronounced. Ultimately, the step size in a PGA would determine the trade-off between the overall complexity of the PGA versus the linearity and noise requirement of subsequent stages. Also, depending on the modulation scheme, abrupt amplitude changes may not be tolerated [49]. For example, in a high-order quadrature amplitude modulation (QAM) signaling scheme (such as 64-QAM and higher), the step size must be limited to less than 1 dB. Moreover, analog television reception does not tolerate any abrupt amplitude changes, mandating the use of VGAs only [36].

Typical narrowband receivers have one or two automatic gain control loops: typically one for RF and another for the BB filters. In some receivers, only a BB AGC is used, where the detection is done after the ADC and the filtering of the adjacent blockers. The justification here is that the problematic blockers for the system are nearby blockers and cannot be filtered by an RF filter anyways. For these types of systems, the RF amplifier gain is adjusted depending only on the desired signal strength.

In wideband receivers for cognitive radios, however, multiple loop may be required, depending on the activity in the spectrum. If the user is in a metropolitan area, where the spectrum is reasonably used, the LNA gain must be lowered due to potentially problematic faraway blockers. In this case, the sense point for the automatic gain control has to be at the LNA and must control the LNA directly. LNA gain control is discussed in more detail in the next section.

The remainder of this chapter concentrates on detailed design trade-offs in a wideband receiver design as it pertains to cognitive radios. Wideband LNA topologies are discussed and compared. A wide variety of techniques for tunable RF filters are detailed. Finally, image rejection and harmonic rejection issues of downconversion mixers are reviewed.

3.3 Wideband LNA Design

The LNA is the central block in the receiver that has the strongest weight on the overall receiver NF. Before the wideband LNA design can be detailed, a few important RF design concepts must first be reviewed.

Input Matching and Noise One important assumption in deriving (3.4) was that the receiver is impedance matched to the source resistance. As can be seen in Fig. 3.6, the source impedance of the LNA is the antenna. In wireless systems, the antenna is assumed to have a 50-Ω impedance. In order to minimize the losses from the antenna to the LNA, the LNA input impedance must be tuned to facilitate the maximum power transfer from the antenna to the LNA. For maximum power transfer, the LNA input impedance must be matched to the antenna impedance for the

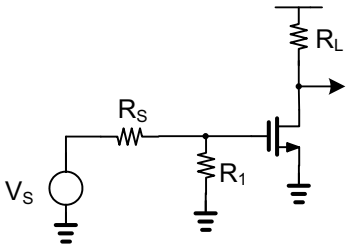


Fig. 3.8 Common-source amplifier with resistor input termination

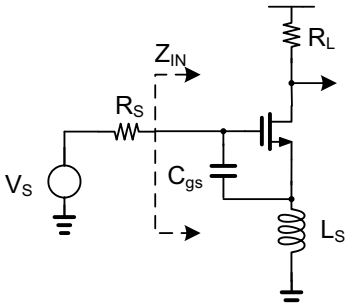


Fig. 3.9 Narrowband low-noise amplifier (*LNA*) matching using inductive degeneration

desired signal bandwidth [43]. Consider a simple common-source (CS) amplifier with a resistor load as shown in Fig. 3.8. One way to guarantee wideband matching is to resistively terminate the amplifier with resistor R_1 such that $R_1 = R_s$. Although, conceptually, this would provide proper termination for the amplifier, it is impractical for low-noise designs. To understand why this is the case, the noise factor of the amplifier in Fig. 3.8 can be computed as:

$$F = 1 + \frac{R_s}{R_1} + \frac{R_s}{g_m}\left(\frac{1}{R_1} + \frac{1}{R_s}\right)^2 \cdot \left[\gamma + \frac{1}{R_L g_m}\right] \tag{3.5}$$

As (3.5) shows, the minimum NF achievable is more than 3 dB if $R_1 = R_s$, which is too high for the LNA design for cellular and television systems.

One common method for termination is to use passive reactive components, such as a degeneration inductor as shown in Fig. 3.9 [16]. The input impedance can be shown to be:

$$Z_{in}(s) = sL_s + \frac{1}{sC_{gs}} + \frac{g_m L_s}{C_{gs}} \tag{3.6}$$

where C_{gs} is the negative-channel field-effect transistor (NFET) gate-to-source capacitance, L_s is the degeneration inductance, and g_m is the NFET transconductance.

For an operating frequency of $\omega_0 = \frac{1}{\sqrt{L_s C_{gs}}}$, the reactive components of the imped-

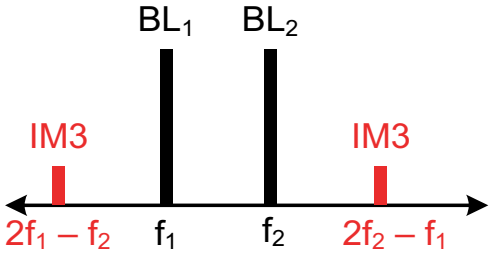


Fig. 3.10 Third-order intermodulation terms produced by 2 blocker tones

ance cancel out and only the third term in (3.6), the real component, remains. This means that with proper choice of the degeneration inductance and NFET device size, a 50-Ω match at resonance frequency is possible.

An analysis of the schematic shown in Fig. 3.9, shows that the NF can be given as:

$$F = 1 + \gamma g_m R_s \left(\frac{\omega_0}{\omega_T} \right)^2 \tag{3.7}$$

where $\omega_T = 2\pi f_T$ and f_T is the NFET unity gain frequency in Hertz (Hz). As (3.7) shows, a NF of less than 3 dB is now possible, and would improve as the device f_T is improved. The disadvantage of using inductor source degeneration is that the LNA bandwidth is now limited by the inductor–capacitor circuit (LC) tank formed by L_s and C_{gs}. Other parasitic effects such as the wire trace resistance and the parasitic NFET gate resistance must be included for a more accurate description of the NF.

Linearity There are four main metrics that are used to quantify the linearity of a LNA. The first of these metrics is known as input-referred intercept point for the third-order distortion (IIP3). As its name implies, this metric is used to measure the third-order intermodulation (IM3) terms produced by two large tones, as shown in Fig. 3.10. The IIP3 metric itself is given as [23]:

$$IIP3(dBm) = P_{BL}(dBm) + \frac{P_{BL}(dBm) - P_{IM3}(dBm)}{2} \tag{3.8}$$

where P_{BL} is the rms power of the blocker, P_{IM3} is the power level of the third-order intermodulation terms. As shown in Fig. 3.10, this type of distortion is problematic if there are blockers P_{BL1} and P_{BL2} located at center frequencies of f_1 and f_2, respectively, such that the desired signal is at a frequency $2f_1 - f_2$ or $2f_2 - f_1$. Typical values of IIP3 for a LNA are from −5 to +5 dBm.

The second of these metrics is known as input-referred intercept point for second-order distortion (IIP2). As its name implies, this metric is used to measure the second-order intermodulation (IM2) terms produced by two large tones, as shown in Fig. 3.11. The IIP2 metric is given as [23]:

$$IIP2(dBm) = 2P_{BL}(dBm) - P_{IM2}(dBm) \tag{3.9}$$

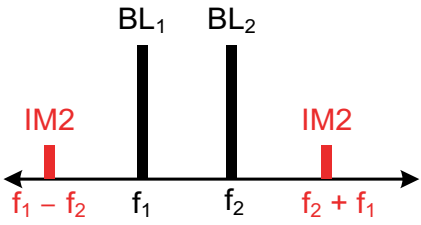


Fig. 3.11. Illustration of the second-order intermodulation terms produced by 2 t

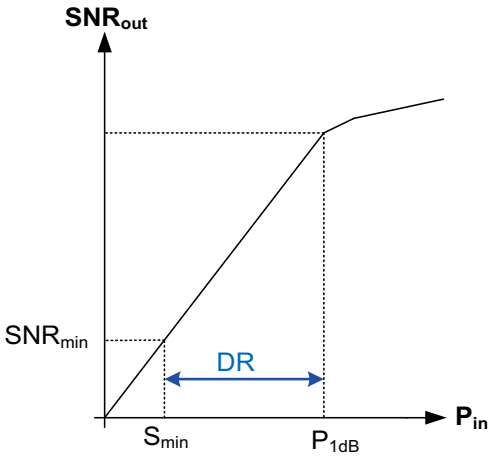


Fig. 3.12. Dynamic range defined by using the P_{1dB} metric

where P_{BL} is the rms power of the blocker, P_{IM2} is the power level of the second-order intermodulation terms. As Fig. 3.11 shows, if the two blockers are adjacent, the intermodulation terms are produced either at a very low frequency, or nearly double the blocker frequencies. This shows that blockers in the cellular band can result in distortion terms in the low-frequency television bands, for example. Alternately, two television signals and intermodulate at the receiver and produce a problematic IM2 term in the cellular band. For a conventional receiver, this is usually avoided by an external band-select filter, which filters blockers from other transmission standards. For a cognitive radio to make use of the maximum available spectrum, such external band-select filters should be avoided in favor of a tunable integrated filter.

The third common metric used to quantify the linearity of a receiver is known as the 1-dB compression point, or P_{1dB}. The P_{1dB} is defined as the input power level at which the gain of the amplifier drops by 1 dB due to nonlinear compression. This quantity is sometimes also used to define the DR of a receiver, as shown in Fig. 3.12. More specifically, DR is defined as

$$DR = P_{1dB} - S_{min} \tag{3.10}$$

where S_{min} is the minimum detectable signal and the units are in decibels.

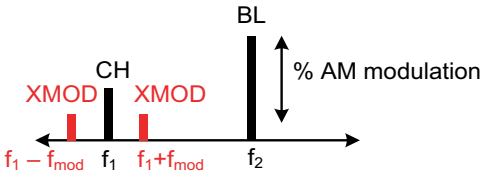


Fig. 3.13 Cross-modulation scenario with an AM-modulated blocker

The final metric that is commonly used to quantify linearity is known as XMOD, which is shown in Fig. 3.13. An amplitude-modulated (AM) blocker near the desired signal results in distortion such that the sidebands at the modulated frequency appear next to the desired signal. The AM modulation can reach up to 100%. This scenario emulates the condition when the blocker is a QAM-modulated signal. When a nearby blocker is AM modulated, it produces XMOD terms at frequency offsets equal to the AM modulation frequency. The amplitude of the cross-modulation term is given as [48]:

$$XMOD = \frac{m \cdot 4 \cdot P_{BL}}{IIP3 + 2P_{BL}} \tag{3.11}$$

where m is the modulation index.

3.3.1 Wideband Circuit Topologies

After introducing the critical metrics in LNA design, this section now focuses on wideband LNA circuit topologies. Figure 3.14 shows a survey of recent wideband (>1 GHz bandwidth) LNA implementations [6–8, 17, 35, 38, 42, 45, 47, 51, 67, 71, 72]. The NF and normalized IIP3 numbers are shown, where the normalization is with a unit gain value. As the figure shows, the NF varies from nearly 2–5 dB and IIP3 from −15 to +5 dBm.

The wideband LNA techniques can be broken into two categories. The first category is the common gate (CG) LNA topology [33]. A differential version of the standard CG LNA is shown in Fig. 3.15. Note that an inductive bias circuit is chosen here to minimize noise, but is not necessary. The resistor R_{in} shown is an equivalent resistance to show where the input resistance is measured, and is not a physical resistor.

A summary of the voltage gain, input resistance, and noise factor for the CG LNA is shown in Table 3.1. The main salient feature of this topology is that the input matching can be adjusted by choice of g_{m1}. The main drawback is high NF. For large channel lengths, the NF can reach a value of 3–4 dB. The NF significantly degrades as technology is scaled due to the significant increase of the noise parameter, γ. Linearity can also be problematic if the main source of nonlinearity is variation of the device V_{GS}.

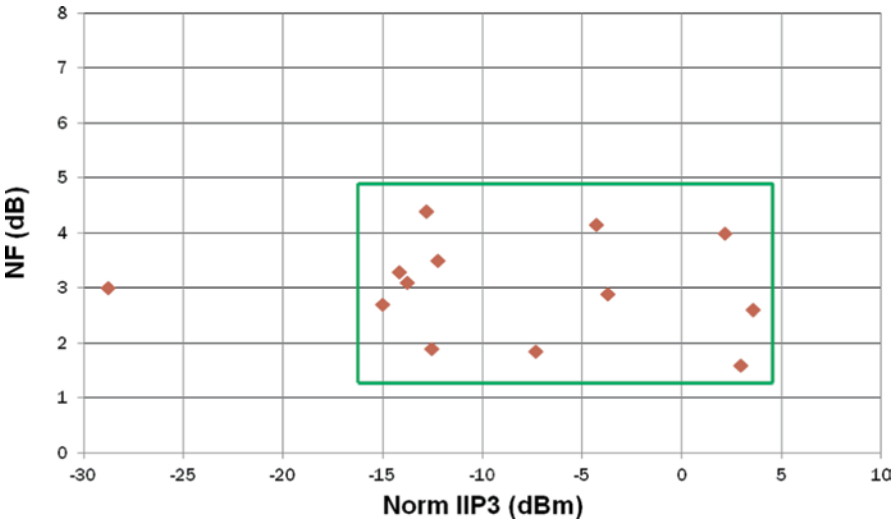


Fig. 3.14 Noise figure (*NF*) versus normalized input-referred intercept point for the third-order distortion (*IIP3*)

Fig. 3.15 Standard common gate (*CG*) low-noise amplifier (*LNA*) circuit topology

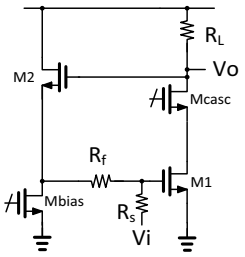


Table 3.1 Summary of performance of differential common gate (CG) low-noise amplifier (LNA)

Voltage gain	$A_v = g_{m1} \bullet R_L$
Input resistance	$R_{in} = \frac{2}{g_{m1}}$
Noise factor	$F = 1 + \frac{\gamma}{\alpha} + \frac{4R_s}{R_L}$

The CG LNA topology can be improved through the use of negative feedback, as shown in Fig. 3.16 [33]. The cross-coupled capacitors C_1 shown form a negative feedback value of −1. As with Fig. 3.15, the R_{in} resistor element is an equivalent resistance denoting how the equivalent input resistance is measured, and is not a physical resistor.

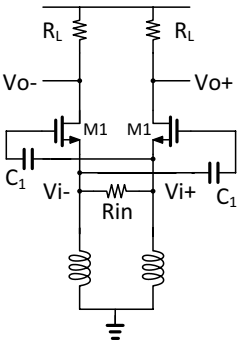


Fig. 3.16 Improved common gate (*CG*) low-noise amplifier (*LNA*) topology by capacitive negative feedback

Table 3.2 Summary of performance of differential common gate (CG) low-noise amplifier (LNA)

Voltage gain	$A_v = 2 \cdot g_{m1} \cdot R_L$
Input resistance	$R_{in} = \frac{1}{g_{m1}}$
Noise factor	$F = 1 + \frac{\gamma}{2\alpha} + \frac{4R_s}{R_L}$

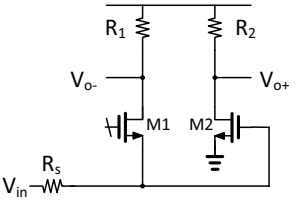


Fig. 3.17 Wideband common gate (*CG*) low-noise amplifier (*LNA*) topology with common-source (*CS*) stage for noise cancellation

A summary of the voltage gain, input resistance, and noise factor of the improved CG LNA is shown in Table 3.2. Comparing the performance with that of a conventional CG LNA, the gain is increased by a factor of 2, but more importantly the main noise component, γ/α, is reduced by a factor of 2 as well. NF values of around 3 dB are possible with this topology.

One interesting variant of the CG LNA topology is to merge it with a CS stage in order to cancel the LNA noise [18]. Such a configuration is shown in Fig. 3.17 below. A portion of the noise flowing through M1 to R1 is sensed as amplified by the M2–R2 branch. This produces an in-phase amplified noise component in the M2–R2 branch. By adjusting the M2 transconductance and the R2 resistor, the LNA can be tuned to cancel out the noise produced by the CG stage. This, of course, would result in an imbalance of common mode and swing between the two output differential nodes. This cancellation effect also excludes any kind of phase shifts that can result from parasitic capacitance along the CS and CG paths, which would be more pronounced at higher frequencies. There are several other variants on this topology that reduce the noise even further [2, 11, 15, 21, 59, 75]. A NF of at least 3 dB is possible with this configuration over a very wideband of operation.

Table 3.3 Performance summary of common gate (CG) low-noise amplifier (LNA) topology with common-source (CS) stage for noise cancellation

Voltage gain	$A_v = \dfrac{2(g_{m1}R_1 + g_{m2}R_2)}{1 + g_{m1}R_s}$
Input resistance	$R_{in} = \dfrac{1}{g_{m1}}$
Noise factor	$F = 1 + \dfrac{\gamma(g_{m1}R_1 - g_{m2}R_2)^2}{A_v^2} + \dfrac{4\gamma g_{m2}R_2^2 / R_s}{A_v^2} + \dfrac{4(R_1 + R_2)/R_2}{A_v^2}$

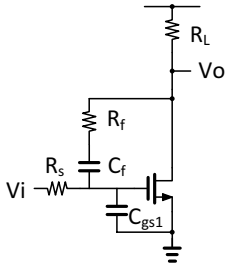


Fig. 3.18 Resistive shunt feedback low-noise amplifier (*LNA*) circuit topology

Table 3.3 shows the performance summary of the CG–CS wideband LNA. The first term shows the noise-cancellation effect in the CG stage. The second term shows the noise of the CS stage, and the third stage is the noise due to the load. The table also shows that the input match is the same as a conventional CG wideband LNA.

The other category of wideband LNA techniques relies on feedback principles. One such commonly used topology is the resistive shunt feedback LNA [19, 20], shown in Fig. 3.18. The capacitor C_f helps in separating the direct current (DC) input bias from the output DC bias. It also helps to stabilize the gain over a wider frequency band. The capacitor C_{gs} is the gate-to-source capacitance of the NFET device.

One interesting feature of this topology is that the feedback mechanism has the positive effect of partially cancelling the input noise. Consider an input voltage noise to the LNA field-effect transistor (FET) device. The polarity of this input voltage noise is flipped when it appears at current drain noise. A fraction of that noise then travels along the feedback network and is multiplied by the source impedance, R_s, to appear as voltage noise back into the FET input. Since the polarity of this noise is the opposite of the original voltage input noise source, it is partially canceled out depending on the transfer function along the feedback network back into the FET gate input. The degree of cancellation depends on the open loop gain of the amplifier as well as the feedback gain.

Table 3.4 Summary of performance of *R* shunt feedback low-noise amplifier (LNA) topology

Voltage gain	$A_v = \dfrac{R_L(1-g_{m1}R_f)}{R_f+R_L}$
Input resistance	$Z_{in} = \dfrac{R_f+R_L}{1+g_{m1}R_L} // \dfrac{1}{sC_{gs}}$
Noise factor	$F = 1+\dfrac{R_f}{R_s}\left(\dfrac{1+g_{m1}R_s}{1-g_{m1}R_f}\right)^2 + \dfrac{1}{R_sR_L}\left(\dfrac{R_f+R_s}{1-g_{m1}R_f}\right)^2 + \dfrac{\gamma g_{m1}}{\alpha R_s}\left(\dfrac{R_f+R_s}{1-g_{m1}R_f}\right)^2$

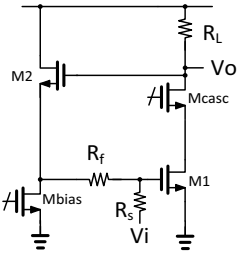


Fig. 3.19 Shunt–shunt feedback low-noise amplifier (*LNA*) circuit topology

A summary of the performance of the resistive shunt feedback LNA is summarized in Table 3.4. The gain is negative for a large $g_m \cdot R_f$ product, as expected. The input match is degraded by the C_{gs}, which is also expected. It is important to note that the input match is dependent on the feedback resistor as well as the load resistor. The NF improves for larger $g_m R_f$ products, but degrades as R_f is increased alone (second term of the expression dominates in that case). Sub-3 dB NF is possible with this topology.

Another common feedback LNA topology is the shunt–shunt feedback configuration shown in Fig. 3.19 [13]. In this configuration, the input impedance is set by the feedback loop formed by M2 and R_f. The matching network effectiveness is limited by the closed loop bandwidth of the loop. The dominant pole is at the output LNA node. Transistor M_{CASC} is a cascode transistor used to avoid the Miller effect of M1 [66]. The performance of the shunt–shunt feedback LNA is summarized in Table 3.5. One advantage of this topology over the resistive feedback LNA is that it offers a larger degree of freedom in that the matching is primarily determined by the feedback element M2 and R_f. The NF is inversely related to $1+g_{m2}R_f$, whereas the input matching is directly related to the same quantity, introducing an interdependence between the two.

Table 3.5 Shunt–shunt feedback low-noise amplifier (LNA) summary

Voltage gain	$A_v = -g_{m1}R_L$
Input resistance	$Z_{in} = \frac{1}{g_{m2}} \cdot \frac{1+g_{m2}R_f}{1+g_{m1}R_L}$
Noise factor	$F = F_{M1} + F_{Rf} + F_{M2} + F_{RL} + F_{M,bias}$ $F = 1 + \frac{\gamma R_s}{g_{m1}}\left(\frac{1}{R_s} + \frac{g_{m2}}{1+g_{m2}R_f}\right)^2 + R_s R_f \left(\frac{g_{m2}}{1+g_{m2}R_f}\right)^2$ $+ \frac{\gamma g_{m2} R_s}{\left(1+g_{m2}R_f\right)^2} + \frac{R_2}{g_{m1} \cdot A_v}\left(\frac{1}{R_s} + \frac{g_{m2}}{1+g_{m2}R_f}\right)^2 + \frac{\gamma_{bias} g_{m,bias} R_s}{\left(1+g_{m2}R_f\right)^2}$

3.3.2 LNA Gain Control

In order for a receiver to maximize performance given a spectrum environment, it is desirable to adjust the gain of the LNA to trade-off NF for linearity, as was illustrated in Sect. 3.2. There are two classes of LNA gain control: continuous (VGA) and discrete (PGA).

Discrete gain control can be applied to either CG LNA topologies or feedback type of topologies. Figure 3.20 shows a conceptual diagram of how a PGA can be applied to either type of wideband LNA topology. The first part of the PGA consists of a programmable attenuator, with 6 dB of attenuation per stage. The second part of the PGA consists of the wideband amplifier, which can be a CG LNA or a feedback type LNA. The PGA shown would have a gain range of A-6 (dB) down to A-24 (dB). An extra programmable stage (not shown) to bypass the attenuator altogether can be used to maximize the gain of the LNA. In theory, an arbitrary number attenuation stages can be used to maximize the attenuation as desired. In practice, however, the parasitic coupling between the input and output nodes of the LNA would limit the achievable attenuation to around 50–60 dB for a fully integrated solution. Such coupling can be due to parasitic capacitance due to wire traces or due to substrate coupling between the input and output terminals of the attenuator.

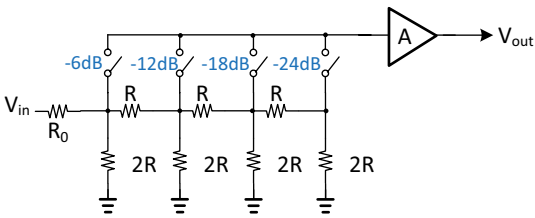


Fig. 3.20 Programmable gain amplifier (*PGA*) control applied to wideband LNA

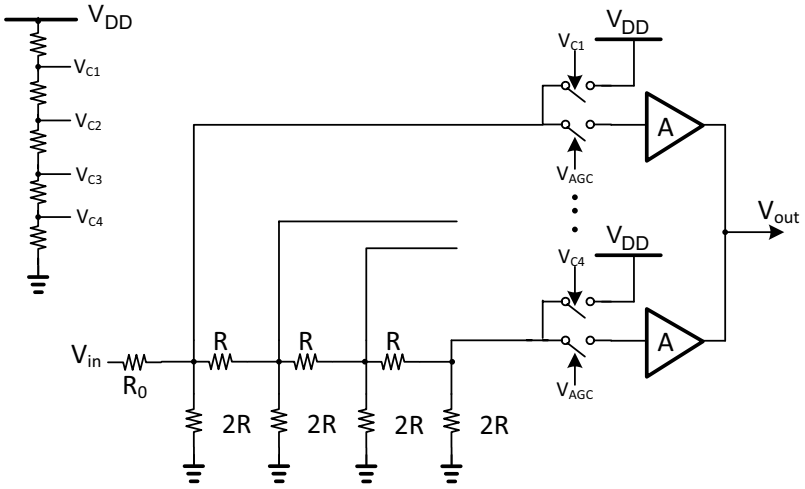


Fig. 3.21 Variable gain amplifier (*VGA*) control applied to a wideband low-noise amplifiers (*LNA*)

The programmable attenuator maintains a fixed input impedance of R_0 if $R=R_0$ for all gain values. Also note that in the case of a CG amplifier, an extra degree of the freedom in the design space is offered. The input impedance of the CG stage no longer needs to be R_0. Instead the parallel combination of the $2R$ resistor enabled and the input impedance of the CG has to be $2R$. This allows a reduction of the required g_m for matching, which can lead to lower power consumption. Similar optimizations can be done in the feedback LNA topologies when combining them with a variable attenuator.

The optimization of the switches in the PGA is also critical. On one hand, the on-resistance of the switches must be minimized in order to reduce the noise due to the resistors. This is typically accomplished by sizing up the switches. Increased switch size, however, reduces the bandwidth of the attenuator. Also, distortion of the attenuator may be introduced, especially for high blocker signals. This may be due to the body effect and the switch on-resistance variation with input signal swing. Techniques such as inserting a linearizing series resistor to the switch may be used on the expense of an increase in noise.

Continuous gain control is also possible. One such implementation is shown in Fig. 3.21 [10]. In this implementation, the wideband amplifier is broken into several "sub-amplifiers," each connected to an attenuation stage and a current steering differential pair is added to the input of each sub-amplifier, A. At any given gain control setting, only two sub-amplifiers are partially current steered. The rest of the cells are fully steered and hence the switches do not contribute to noise or linearity significantly. In this way, the sub-amplifier gain is effectively "soft switched" depending on the V_{AGC} control voltage.

Further optimizations to the LNA are possible depending on the topology of the amplifier. If a CG wideband LNA is considered for the amplifier implementation,

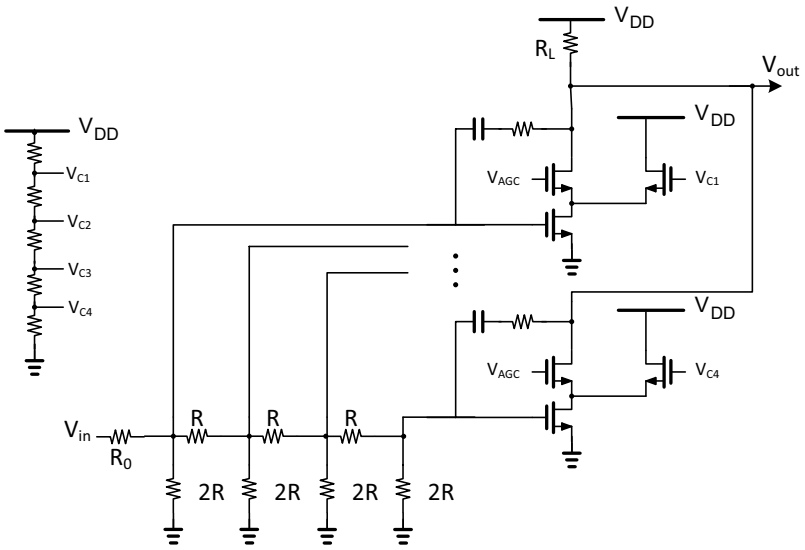


Fig. 3.22 Applying Variable gain amplifier (*VGA*) control to an *R* shunt feedback amplifier

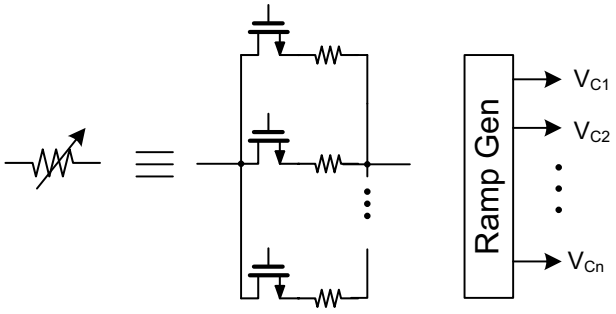


Fig. 3.23 Resistor soft switching and control for continuous gain control

the order of the CG stage and the current steering stage can be swapped, as is shown in [10]. This helps to reduce the parasitic load to the input of the LNA. Continuous variable gain control can be applied to a resistive shunt feedback amplifier using the approach shown in Fig. 3.21 by breaking up the NFET shown in Fig. 3.18 into "sub-amplifiers" with current steering pairs as cascode devices. This topology is shown in Fig. 3.22. This approach, however, also affects the input matching. A similar technique can be used in the active shunt–shunt feedback amplifier introduced in Fig. 3.19.

Although it is possible to use the approach shown in Fig. 3.21 on amplifier topologies other than CG (as shown in Fig. 3.22), alternate approaches may be more optimal. One such approach is based on the soft switching of resistors instead of transconductors [25], as shown in Fig. 3.23. In this way, the feedback resistor and load resistor in a resistive shunt feedback amplifier can be changed in order to affect the gain and input matching simultaneously.

Table 3.6 Comparison of common gate (CG) and shunt feedback wideband low-noise amplifiers (LNAs)

	CG LNA	Shunt Fdbk LNA
Noise figure	3–5 dB	2–5 dB
Voltage Gain	10–20 dB	10–20 dB
Linearity (IIP3)	−7 to +7 dBm	−7 to +7 dBm
Power	1–5 mW	>10 mW
Matching	Input	Input/output

3.3.3 Comparison and Summary

In this section, we have seen two categories of topologies for wideband LNAs. Other topologies exist, including wideband LC matching network techniques [44]. These techniques, however, do not lend themselves well for integration. A summary comparison of the two categories of techniques discussed above (CG vs. feedback topologies) is found in Table 3.6. As the table shows, the main trade-off is power consumption requirement versus noise requirement. The shunt feedback LNA topologies offer lower NF, but at the expense of very large power consumption. The other trade-off is that the CG LNA only addresses input matching, whereas the shunt feedback LNA topology implicitly addresses both input matching as well as output matching.

Gain control is an important feature of any practical LNA design. Discrete gain control (resulting in PGAs) and continuous gain control (resulting in VGAs) were both covered. Design trade-offs in both techniques were discussed. Two topologies for continuous gain control were also covered.

3.4 RF Tracking Filter

Another important RF block in a wideband receiver is an RF tracking filter. Conventional narrowband wireless receivers do not use a tracking filter since they rely on an external surface acoustic wave (SAW) filter [58], as shown in Fig. 3.24. The external SAW filter guards the receiver from large external out-of-band (OOB) blockers.

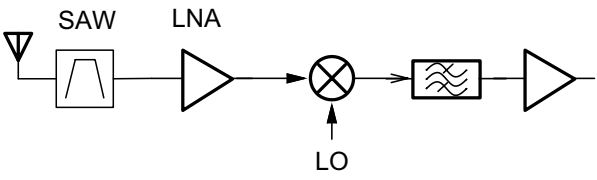


Fig. 3.24 Conventional narrowband wireless receiver

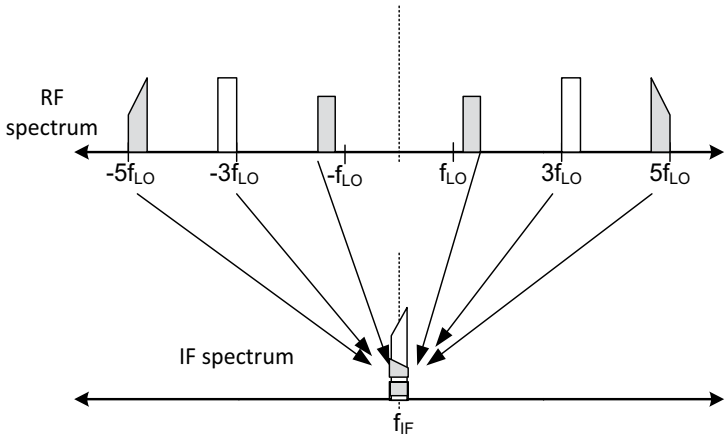


Fig. 3.25 Mixing of out-of-band (*OOB*) with higher-order harmonics of the local oscillator (*LO*)

SAW filters are especially important to guard against mixing OOB blockers due to higher-order harmonics present in the local oscillator (LO) signal. This scenario is illustrated in Fig. 3.25. The LO signal is usually a square wave signal, exhibiting strong odd-order harmonics. More specifically, a square wave LO signal can be expressed as:

$$sq(f) = \frac{4}{\pi} \sum_{n=1,3,5,\ldots} \frac{1}{n} sin\left(\frac{n\pi f_0}{f}\right) \tag{3.12}$$

From (3.12), it can be readily seen that the third-order harmonic is only 9.5 dB lower than the fundamental LO tone. This means that blockers at a frequency near 3LO would be downconverted by the mixer and possibly distort the desired signal at baseband (BB). Similar argument would be true of all the other odd-order harmonics of the LO (with decreasing magnitude with higher-order harmonic).

As stated earlier, narrowband wireless receivers avoid the issue of harmonic rejection through the use of an external SAW filter. SAW filters reject any OOB blockers, which may be located at the odd-order harmonics of the LO. Wideband wireless receivers, on the other hand, cannot make use of such SAW filters and must provide harmonic rejection through other means. One such method is through the use of an RF tracking filter. In this section, several types of RF tracking filters are detailed, each with its own set of trade-offs.

Another benefit of a tracking filter (or even an SAW filter) is that it limits the energy going into the LNA. Lower energy means that the LNA is not saturated with large blockers and the gain of the LNA can be increased, and hence increasing the sensitivity of the receiver.

The third benefit of filtering the RF spectrum before the downconversion mixer is enhanced image rejection. This is only important in systems where the signal is not converted directly to BB and, instead, is converted to an intermediate frequency

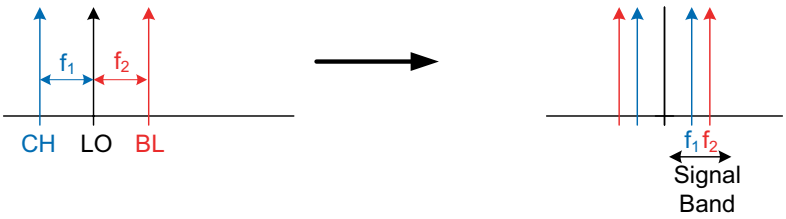


Fig. 3.26 Downconverter illustrating the concept of image folding onto the desired signal

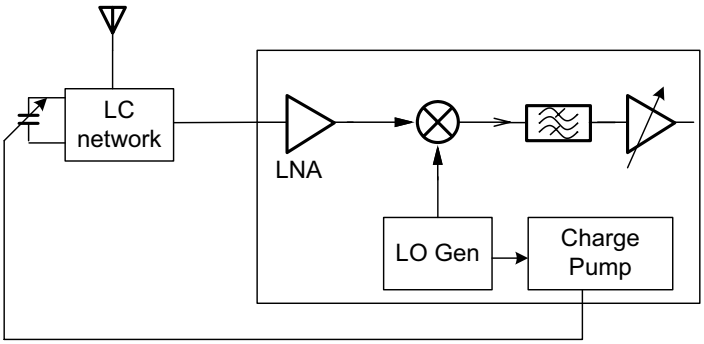


Fig. 3.27 Radio frequency (*RF*) tracking filter using high-Q tunable passive devices

(IF). The benefit of converting to an IF frequency includes avoiding low-frequency noise (namely flicker noise), which can be quite problematic, especially in CMOS implementation of receivers [12].

To illustrate the concept of image rejection, consider the downconversion operation shown in Fig. 3.26. The image channel is located at LO + IF and the desired channel is located at LO − IF. Since the LO contains both positive and negative frequency components, it would downconvert both the image channel at the desired channel to both IF and − IF, thus smearing the desired channel with the image channel once downconverted. More specifically, the mixing terms − LO + (LO − IF) and LO − (LO + IF) would smear into − IF and the mixing terms LO − (LO − IF) and − LO + (LO + IF) would smear into IF. It is important to note that it requires both the LO signal and the RF signal to contain both negative and positive frequency components (i.e., real signals). The importance and relevance of this point will become clear in the following sections. Furthermore, techniques to provide image rejection in the mixer are explored in Sect. 3.5.

3.4.1 High-Q Tunable Passive Discrete Filters

RF tracking filters have historically been built using external discrete inductors, capacitors, and varactors as shown in Fig. 3.27 [74]. As the figure shows, a bank of

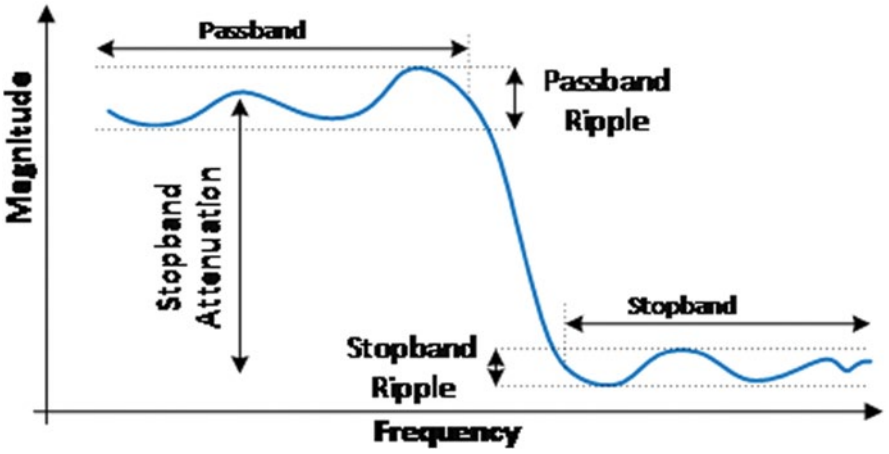


Fig. 3.28 Filter specifications

varactors is usually used to extend the range of the tracking filter. It is important to note that a charge pump is required to boost the control voltage on the varactor to above the power supply voltage to maximize its tuning range. It is common for the tuning voltage to exceed 10 V [41]. This means that the varactor is implemented in a specialized technology capable of handling such high voltages. The main design challenge of this type of tuning filter is to minimize the insertion loss and maintain equal input and output matching over a large frequency tuning range. The insertion loss in the filter is due to parasitic resistances in the LC components, varactor, and wire trace resistance. Equal input and output matching is desirable for maximum power transfer from the antenna to the LNA.

Although the details of filter design is beyond the scope of this book, an overview of the design procedure for such filters is presented here to get a glimpse of the engineering trade-offs involved in such filters. The filter design starts with specifying the passband frequency and ripple, stopband frequency and ripple, as shown in Fig. 3.28. The type of filter is then selected. The three types of filters are Butterworth filter, Chebyshev filter, and Elliptical filter [70], as listed in Table 3.7. $C_n^2(\omega)$ and $R_n^2(\omega)$ Chebyshev and Elliptical polynomials, respectively. Once the filtering profile is selected, the choice between the filter types then becomes a trade-off between the order of the filter (i.e., complexity and cost) and the group delay and ripple introduced by the filter. Transmission zeros may be added to tweak the filter design further. For example, adding zeros would alter the group delay of the filter, but on the expense of the filtering profile. Once the filter transfer function is determined, the LC component values and structures can be determined by partial fraction expansion of the filter transfer function [68].

Since it is desirable to have a tunable filter centered around the desired channel, a tunable band-pass filter is desired. Once the low-pass filter prototype's poles and

Table 3.7 Filter selection

Specification	Butterworth	Chebyshev	Elliptical
Transfer function (w/no zeros)	$\lvert T(j\omega)\rvert^2 = \frac{1}{1+\omega^{2n}}$	$\lvert T(j\omega)\rvert^2 = \frac{1}{1+\varepsilon^2 C_n^2(\omega)}$	$\lvert T(j\omega)\rvert^2 = \frac{1}{1+\varepsilon^2 R_n^2(\omega)}$
Passband ripple	+	O	−
Transition band	−	O	+
Filter order	−	O	+
Group delay	+	O	−

zeros have been selected, the filter is translated into a band-pass filter using the following frequency translation equation [68]:

$$s' = \frac{\omega_0}{B}\frac{s^2+1}{s} \tag{3.13}$$

where w_0 is the center frequency of the translated band-pass filter and is given by the geometric mean of the frequency passband boundaries of the band-pass response $(\sqrt{\omega_1\omega_2})$, and B is the bandwidth of the translated band-pass filter.

3.4.2 On-Chip Active Tunable RF Filters

Although very low insertion loss filters over a wide tuning range are possible with an external LC-based tracking filter, the cost of such a filter may be too prohibitive for low-cost portable devices. One possible alternative is to use on-chip active tunable RF filters.

The first type of such filters discussed here is a partially integrated filter shown in Fig. 3.29 [37, 40]. In this type of filter, only a single external inductor is used and a bank of programmable on-chip high-Q capacitors is used. As the figure illustrates, there are two methods of using the inductor. In the first method, shown in Fig. 3.29a, it can be used as an inductor load. In this case, the LC band-pass filtering occurs after the gain of the CS stage. This approach has two main drawbacks. First, the bulk of the nonlinearity of the CS stage is usually due to the V_{gs} variation of the amplifying FET device. Inserting the filter after the amplification stage does not protect this device from large blockers. The second drawback is that the inductor is inserted at a high impedance node. This means that since the inductor is external, there will be significant parasitic capacitance on that line, limiting the bandwidth at that node.

The second approach, shown in Fig. 3.29b, has the inductor placed at the source node of the FET device. In this configuration, it acts as a degeneration device. Moreover, computing the transfer function of the filter becomes:

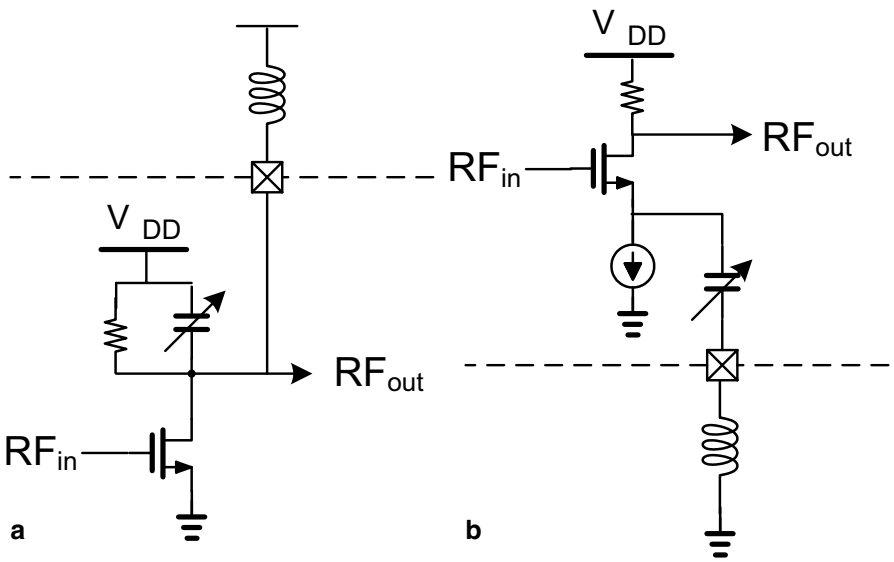


Fig. 3.29 Partially integrated radio frequency (*RF*) tracking filter with **a** inductor load and **b** inductor degeneration

$$H(s) = \frac{g_m R_L}{1 + g_m Z(s)} = \frac{sR_L / L}{s^2 + s\dfrac{1}{Lg_m} + \dfrac{1}{LC}} \tag{3.14}$$

Note that to realize a band-pass filter operation at the output, a notch filter must be realized for Z(s). This has significant advantages. Firstly, the desired channel would be at the center frequency of that notch filter, maximizing the gain of the amplifier. At frequency offsets away from the desired channel, the degeneration impedance is high, thereby reducing the gain and linearizing the FET device more and protecting it against harmful OOB blockers.

The other approach to on-chip active filters is to have fully integrated filters. Active filter design will be discussed in more detail in Sect. 3.6. In this section, only two topologies will be discussed that are suitable for operation at RF frequencies. The first is a transconductor-C resonator type of band-pass structure, shown in Fig. 3.30. The first transconductance ($g_{m,RF}$) stage converts the input RF voltage into a current. The back-to-back g_m-C stages forms an active LC band-pass filter, or a g_m-C resonator [1]. The transfer function of this filter is:

$$H(s) = \frac{g_{m,RF}}{C} \bullet \frac{s}{s^2 + \dfrac{s}{R_L C} + \omega_0^2} \tag{3.15}$$

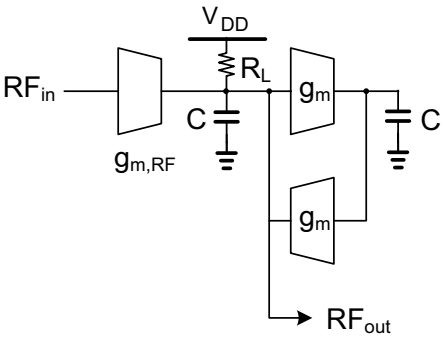


Fig. 3.30 g_m-C resonator as an active band-pass filter

where $\omega_0 = \frac{g_m}{C}$ is the center frequency of the band-pass filter. The back-to-back g_m devices effectively invert the impedance of the capacitor on the right-hand side into an active inductor, as seen at RF_{out}. The main drawback of such filters is their high noise performance [46]. The center frequency of this g_m-C band-pass filter can be tuned by tuning the capacitors, C, or the g_m of the transconductors.

One alternative to using a g_m-C resonator, is to use a source follower (SF)-based Sallen-Key (SK) filter biquad structure [63]. A low-pass second-order SK segment is shown in Fig. 3.31. This is a modified SK structure, where an additional SF stage is added to prevent a transmission zero that shorts the filter at high frequency and limits the attenuation that would otherwise be achievable with this filter [76]. Note that capacitors C_1 and C_2 are programmable in order to be able to tune the bandwidth of this low-pass filter. The Sallen–Key second-order transfer function is in the form of

$$H(s) = \frac{\omega_0^2}{s^2 + \frac{\omega_0}{Q}s + \omega_0^2} \tag{3.16}$$

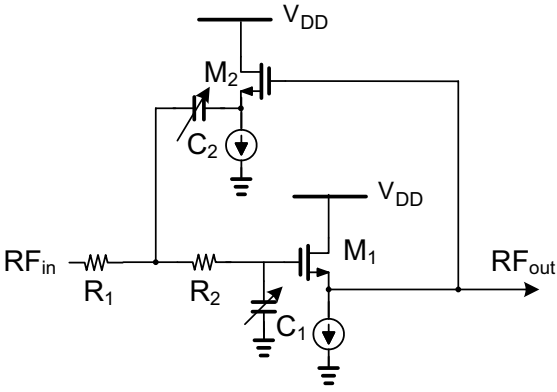


Fig. 3.31 Second-order source follower-based Sallen–Key filter

where $\omega_0 = 2\pi f_0 = \sqrt{R_1 R_2 C_1 C_2}$ and $\frac{\omega_0}{Q} = \frac{1}{C_2}\left(\frac{R_1 + R_2}{R_1 R_2}\right)$. The Q-factor, which control the peaking of the response of the filter is given as:

$$Q = \frac{\sqrt{R_1 R_2 C_1 C_2}}{C_1 (R_1 + R_2)} \tag{3.17}$$

It is important to stress the fact that this is a low-pass filter and not a band-pass filter. Although it is possible to implement a band-pass SK filter, a low-pass filter may be more desirable. The reason for this is that for the same filter complexity, a stronger roll-off frequency is possible with a low-pass filter, which aids in rejecting problematic high-frequency blockers near the odd-order harmonics of the LO.

3.4.3 Wideband Passive Sampled Filters

One class of filters that has recently been rediscovered relies on the sampling effect of a switch to upconvert a BB impedance. This technique was introduced many years ago in [32], but was recently rediscovered and introduced to a wireless receiver as was demonstrated in [3–5, 22, 35, 39, 52–54, 56, 59, 60]. The basic topology of the circuit proposed in [3] is shown in Fig. 3.32. It relies on quadrature mixers to upconvert a complex (I/Q) BB impedance into a real impedance at the RF node. The center frequency of the upconverted impedance can be simply adjusted by tuning the mixer clock (CLK) frequency. More specifically, the ideal RF impedance transfer function is

$$Z_{RF}(f) = Z_{BB}(f - f_0) \tag{3.18}$$

where f_0 is the fundamental frequency of the CLK signal.

The BB impedance shown in Fig. 3.32 consists of two capacitors with impedance of $Z_{BB,I} = \frac{1}{sC_{BB\_I}}$ and $Z_{BB,Q} = \frac{1}{sC_{BB\_Q}}$ for the I and Q quadrature BB channels, respectively. Taking into account that the CLK frequency is a square waveform, it can be shown that the upconverted impedance can be given as:

$$Zin = R_{sw} + \frac{1}{N}\left(sinc\left(\frac{\pi}{N}\right)\right)^2 Z_B // R_{sh} \tag{3.19}$$

where R_{sw} is the on-resistance of the mixer FET devices, Z_B is the BB impedance which is given by $Z_{B,I}$ or $Z_{B,Q}$ as shown above. R_{sw} is an equivalent switch resistance of higher-order harmonics of the CLK signal aliased to the fundamental CLK frequency, f_0. It can be shown [4] that R_{sw} can be given as:

$$R_{sh} = (R_s + R_{sw})\frac{sinc(\pi / N)^2}{1 - sinc(\pi / N)^2} \tag{3.20}$$

In this analysis, the quadrature CLK signals are assumed to be nonoverlapping, or have a 25 % duty cycle. The helps to minimize the mixer insertion loss, or NF.

As alluded in (3.18), the BB impedance is essentially translated into an arbitrary RF frequency. This means that the bandwidth of the filter is ideally preserved, regardless of the RF frequency chosen. This represents a breakthrough in filter selectivity compared to the continuous-time filters discussed in the previous section. To understand this, the quality factor (Q) of a filter is a measure of the sharpness of the filter and is given as

$$Q = \frac{f_0}{BW} \tag{3.21}$$

where f_0 is the same as before, the center frequency of the RF filter and BW is the bandwidth of the RF filter. Since the bandwidth is preserved, increasing the RF frequency results in higher Q, a feature not normally realizable by a continuous-time filter. For example, if a bandwidth of 10 MHz is selected with a center frequency of 1 GHz, the quality factor is 100. This level of performance is considered difficult even if external inductor and capacitor components are used, but can be easily realized by the sampled filter introduced in this section. Another important advantage that must be stressed is that the mixers used to upconvert the impedance are also used to downconvert the signal to BB, hence no additional downconverter mixers are necessary.

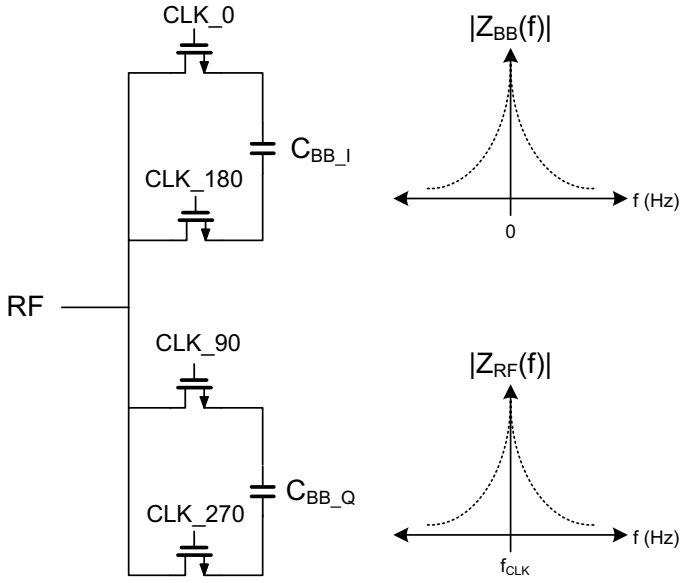


Fig. 3.32 Circuit topology and waveforms shown the *BB* and upconverted *RF* impedance

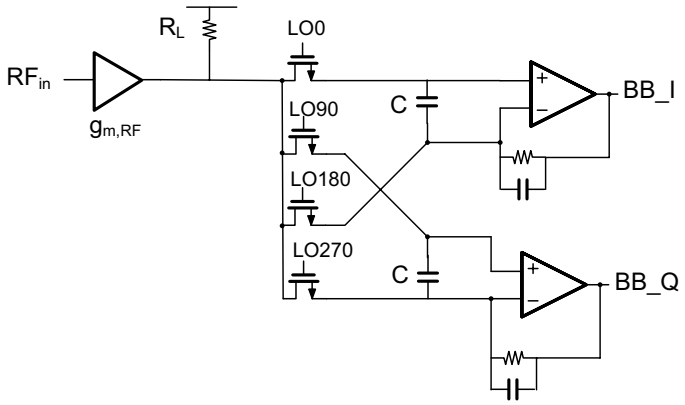


Fig. 3.33 A receiver employing a sampled filter with passive mixers

The sampled filter approach with a passive mixer is not without its drawbacks. The first drawback is the limitation of the stopband attenuation. To understand this, consider a receiver employing the proposed filter shown in Fig. 3.33. At frequencies at the center of the band of interest, the upconverted impedance may be higher than the load resistor, R_L. This means that the maximum gain of the RF amplifier is $g_{m,\,RF} \cdot R_L$. At frequencies away from the desired channel, the upconverted impedance shunts the load resistor R_L, lowering the RF gain. This continues up until the second term in (3.19) becomes much less than the mixer on resistance, R_{sw}, resulting in an RF gain of $g_{m,RF} \cdot R_{sw}$. This means that the stopband attenuation is limited to the ratio of R_L/R_{sw}, which typically does not exceed 15–20 dB.

Another point to make is that the noise and linearity of the BB impedance affects the RF signal directly. Since an RF filter is expected to be used to filter out large blockers that can otherwise saturate the receiver, the preferred BB impedance is a simple passive low-noise element, such as a capacitor. In theory, however, higher-order impedances can be realized by active BB filters. One such approach is to use a generalized impedance converter (GIC) element [73], as shown in Fig. 3.34a. To realize higher-order filters, the impedance Z may be substituted by another GIC filter. In order to realize a low-pass filter, Z_N is substituted by parallel resistor capacitor (RC) elements and Z_P by a resistor, R_1. The impedance of a single generic GIC is given as:

$$Z_{in} = -\frac{Z_N}{Z_P} Z \tag{3.22}$$

The input impedance of the schematic shown in Fig. 3.34b can be shown to be

$$Z_{in}(s) = -\left(\frac{R \bullet R_1}{1 + sRC}\right)^3 \bullet R_2 \tag{3.23}$$

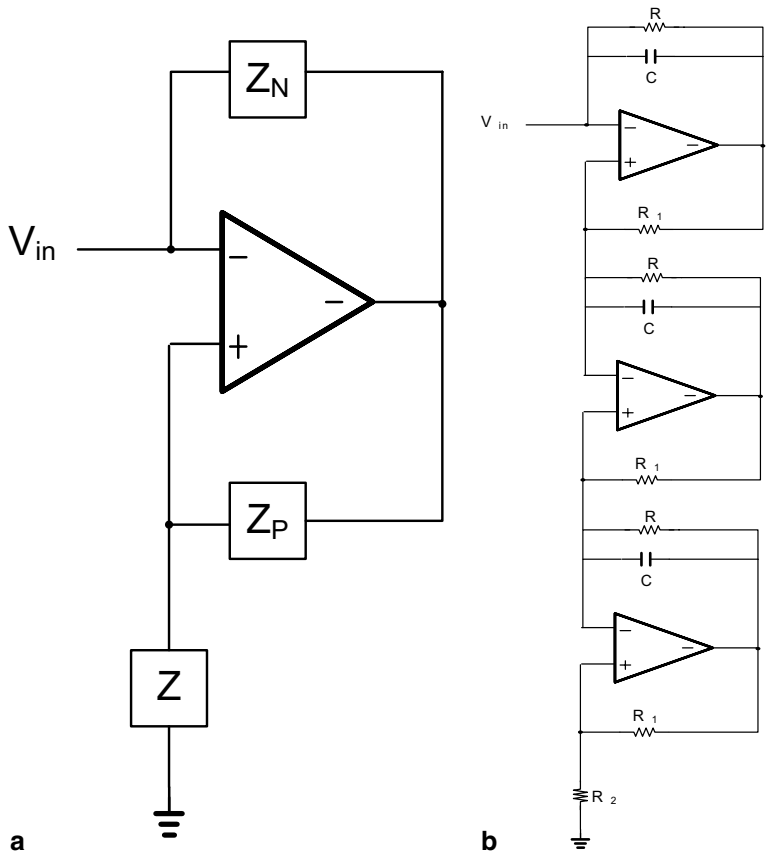


Fig. 3.34 **a** A generic generalized impedance converter (*GIC*) filter and **b** a third-order GIC filter realization by cascading three GIC stages

Higher-order filters may be realized, but a third-order response was found sufficient to provide significant filtering of adjacent channel blockers directly at an RF frequency.

Figure 3.35 shows the output RF impedance of the GIC filter clocked at 500 MHz frequency and compared to a simple capacitor load-based sampled filter. The switch on-resistance was near 60 Ω and the mixer on-resistance was nearly 25 Ω. The ratio of the resistances would improve as technology is scaled. As the figure shows, the second-order GIC filter (GIC2) provides a significant filtering advantage over that of a simple capacitor BB impedance (RC filter). Also, the overshoot found in transfer function actually translated into notches around the desired channel, further attenuating the adjacent blocker. This, of course, comes at the expense of added phase shift to the desired channel. The third-order GIC filter (GIC3) provide even tighter filtering, but not as dramatic as the difference between GIC2 and "RC filter."

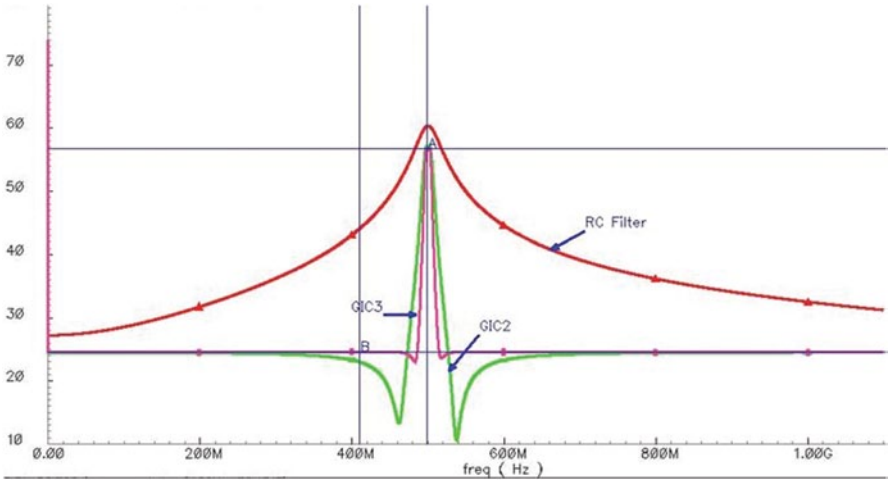


Fig. 3.35 Radio frequency (*RF*) filtering waveform of a sampling filter with passive mixers and a generalized impedance converter (*GIC*) baseband active filter

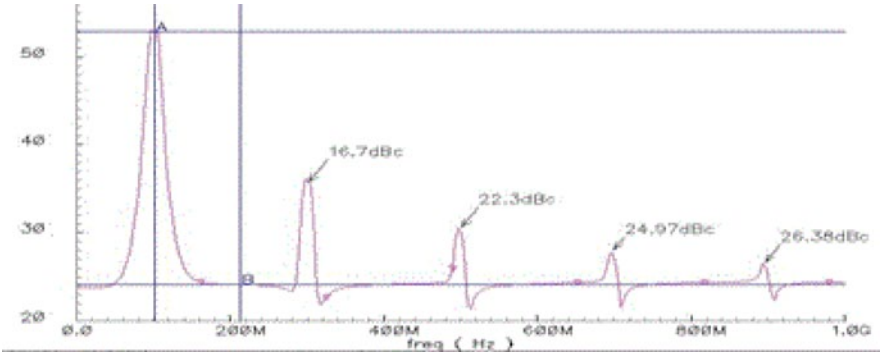


Fig. 3.36 Harmonic impedance upconversion in a sampled filter

Another limitation of the sampled filter using a passive mixer is the harmonic response of the filter itself. Although it forms a passband centered around the desired signal, it also has smaller passbands around harmonics of the CLK frequency. This is illustrated in Fig. 3.36. Notice that the harmonics are less than expected due to some attenuation of the harmonics by a passive RC filtering at the RF node.

3.4.4 Wideband Active Sampled Filters

Another approach to a sample filter is to use active filters. The basic topology of such a filter is shown in Fig. 3.37. The k_1 and k_2 blocks are active downconverter and upconverter mixers. Both mixers are assumed to be quadrature mixers. The *F*(s) block is a complex (I and Q) BB transfer function. This topology has some similari-

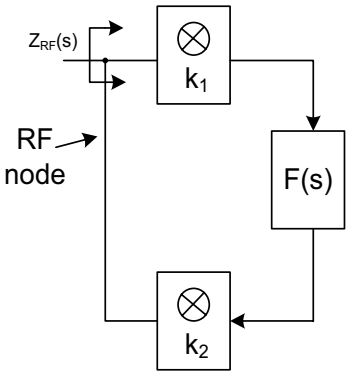


Fig. 3.37 Wideband active sampled filter schematic

ties and differences to the sampled filter with passive mixers. In both types of sampled filters, a BB impedance (or transfer function) is upconverted to RF. Since passive mixers are bidirectional, only one set of quadrature mixers is required. Active mixers, on the other hand, are unidirectional, thereby mandating a pair of quadrature mixers. The main advantage of the active sampled filter approach is that it does not suffer from the switch on-resistance limitation seen from the passive sampled filters. This means that very large stopband attenuation is, in theory, possible. To illustrate this point, note that the RF impedance, $Z_{RF}(s)$, in Fig. 3.37 can be given as:

$$Z_{RF}(\omega) = \frac{1}{k_1 k_2 F(\omega - \omega_0)} \tag{3.24}$$

This impedance will be in parallel with the R_L load impedance as was shown in Fig. 3.33. This means that the stopband attenuation is now limited by the loop gain of the active sampled filter, which is a design parameter, as opposed to the FET on-resistance, which is heavily technology dependent. The filter order and implementation is limited to guarantee stability of the filter. So for instance, if a high-order filter is chosen for $F(s)$, a sharp roll-off RF filter would result. However, the loop gain may have to be lowered in order to guarantee stability. This would result in reduced stopband attenuation. This presents a fundamental trade-off between filter sharpness and stopband attenuation.

Another difference that stands out is that instead of upconverting an impedance per se, an RF impedance is set depending on the transfer function of a BB filter transfer function. The BB filter transfer function, however, is inverted when translated into an RF impedance, as is implied by (3.24). This is due to the fact that the up- and downconverter mixers act to invert the transfer function of the BB filter when seen at the RF node. This is similar to the back-to-back g_m devices in the active continuous-time filter shown in Fig. 3.30. Table 3.8 shows the correspondence between the RF impedances to the $F(s)$-transfer function of an active sampled filter. As it can be seen, there is a wide variety of RF transfer functions that are realizable using this technique.

Table 3.8 Radio frequency (RF) impedance relationship with $F(s)$ for active sampled filter

$F(s)$	$Z_{RF}(s)$
Low-pass filter	Notch filter
High-pass filter	Band-pass filter
Real band-pass filter	Pair of frequency shifted notch filters
Complex band-pass filter	Frequency shifted single notch filter
Real bandstop filter	Pair of frequency shifted band-pass filters
Complex bandstop filter	A single frequency shifted band-pass filter

Despite its many advantages, there are some limitations to the active sampled filter approach. The first is related to the filtering profile of $F(s)$. For practical implementation, $F(s)$ cannot contain a passband at high frequency. In order to understand why, consider a high-pass filter implementation of $F(s)$. According to Table 3.8, this should result in a band-pass filter at RF. The filtering profile at the various stages of the filter is shown in Fig. 3.38. The first row shows the input RF signal, which is a symmetric real RF signal. The RF spectrum is first downconverted by multiply-

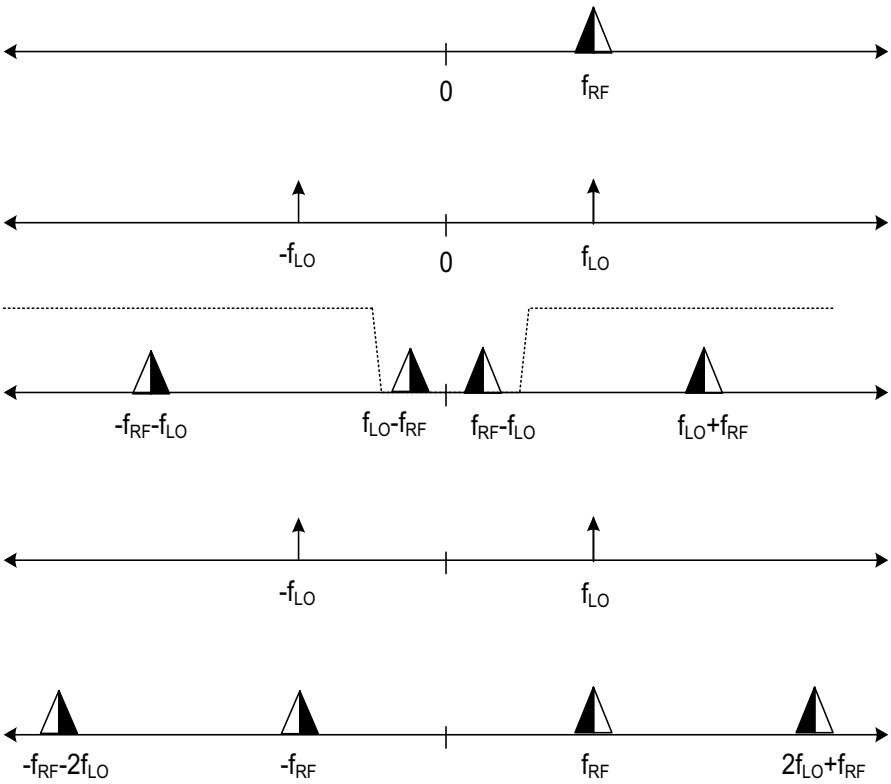


Fig. 3.38 Scenario where a high-pass filter is selected for $F(s)$ in an active sampled filter

ing it with the second row, the LO signal. The downconverted signal is then high-pass filtered, eliminating the low-frequency content, while preserving the content at LO+RF and −LO−RF. There lies the first difficulty, which is that the bandwidth of the high-pass filter must be twice that of the RF frequency band. This is impractical in most cases. The remaining high-frequency signal is then mixed again with the upconversion mixer. However, instead of upconverting, the second set of mixers mix the LO+RF and −LO−RF back to RF and −RF (in addition to 2LO+RF and −2LO−RF). This means that the desired signal will now be subtracted from the RF spectrum! This analysis shows that the only practical choices for $F(s)$ are low-pass, real band-pass and complex band-pass filters (i.e., only three of the six choices shown in Table 3.8).

Another important point to note is that a separate downconverter mixer is required to demodulate the desired signal to BB, since the BB and RF filters are inverses of one another. This would imply that the signal would be attenuated either at RF or BB, if we attempt to use the sampled active filter as a downconverter as well. This creates a more complex system, but depending on the overall filtering requirements, an active sampled filter may still be feasible from an area and power consumption point of view.

This above analysis does not mean that RF band-pass filter implementation based on the active sampled approach is not possible. One way to implement a band-pass filter is to use a hybrid of an active sampled filter and an active continuous-time filter, as shown in Fig. 3.39. The upconverter and downconverter mixers are broken

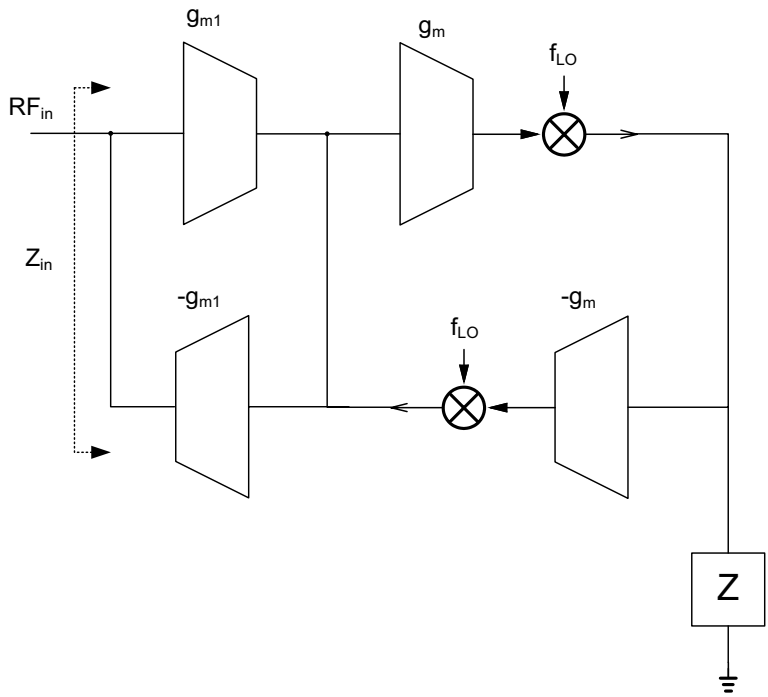


Fig. 3.39 Practical implementation of a band-pass active sampled filter

into g_m stages and mixing devices. The g_{m1} devices operate at RF frequency and work to invert the transfer function of the active sampled filter. In other words, if the impedance *Z,* is a low-pass filter, then the RF impedance Z_{in} is a band-pass response. This filter does not suffer from the impairment shown in Fig. 3.38. The transfer function of this filter then becomes

$$Z_{in}(\omega) = \left(\frac{g_m}{g_{m1}}\right)^2 F(\omega - \omega_0) \tag{3.25}$$

Another issue with the active sampled filter is that of LO feedthrough. If a low-pass filter is used to realize a notch filter at the output, the LO tone would appear in center of the notch filter, limiting the amount of rejection possible with a notch filter. One possible solution is to use a complex band-pass filter (CBPF) to offset the LO frequency away from the notch filtering area. This concept is shown in Fig. 3.40. The CBPF center frequency is $-f_c$ and translated to an RF notch filter centered at $f_{clk} - f_c$.

Figure 3.41 shows a sample of a notch filter implemented with a CBPF with a center frequency of 10 MHz. A center frequency of 800 MHz is shown. The filter was implemented in a 0.35-µm bipolar complementary metal–oxide–semiconductor (BiCMOS) technology to provide image rejection. The overshoot seen at the high-frequency end of the notch filter is due to the interaction of the high-frequency poles at the RF node with the poles introduced by the notch filter.

Another important issue in active sampled filter design is with regards to the shape of the LO signal. Normally, the LO signal is a square wave exhibiting a strong odd-order harmonics as seen before. As with the passive sampled filter approach, signal content at a higher-order harmonics can alias into the band of interest. When designing a notch filter, the alias terms would limit the notch depth. For example, the third-order harmonic is 9.5 dB lower than the fundamental. Given a loss of nearly 4 dB in the mixer quadrature switching network, this gives a maximum notch

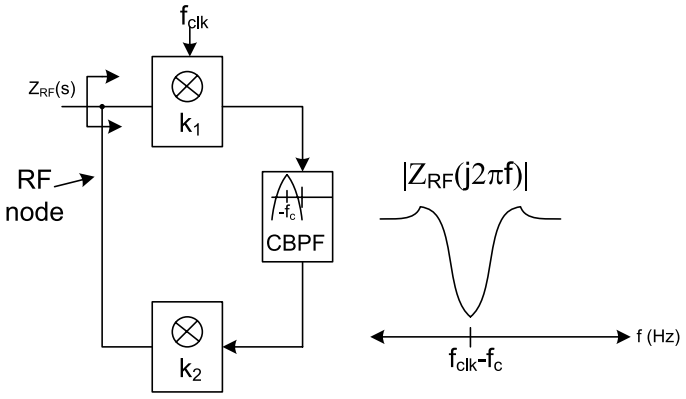


Fig. 3.40 Active sampled filter using a complex band-pass filter *F*(s)

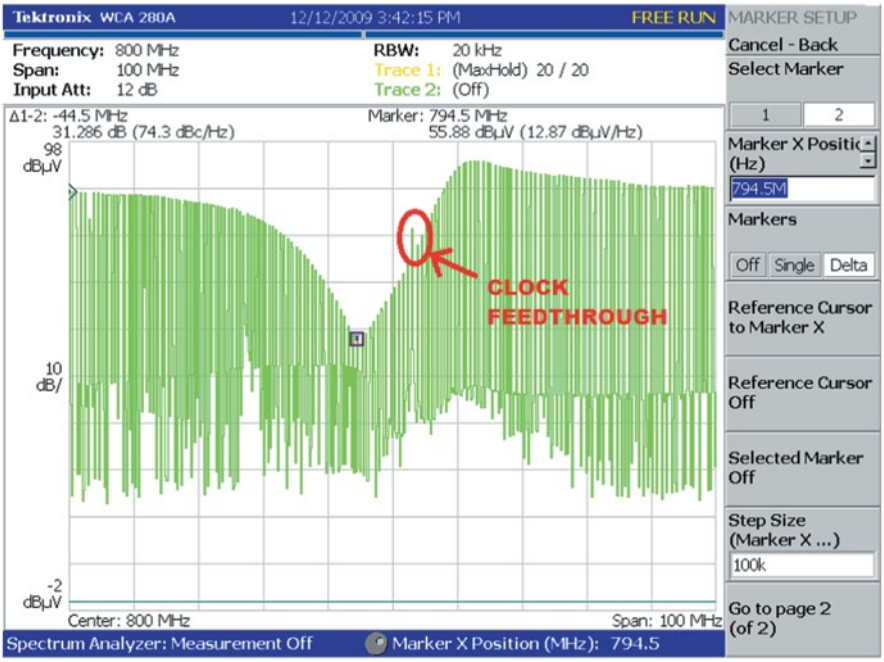


Fig. 3.41 Implementation of an active sampled notch filter using a complex band-pass filter (*CBPF*)

depth of around 13.5 dB. In order to reach notch depths higher than this value, either the LO harmonics need to be filtered or the RF signal needs to be filtered at the harmonics.

Both techniques listed in the previous paragraph of improving the notch depth can be used, as shown in Fig. 3.42. A simple tunable RC filter would provide some degree of rejecting RF content near the harmonics of the LO. A parallel path to the mixer is added clocked at 3LO with one third the amplitude (or 9.5 dB lower). The contents of this path subtracted from the output of the mixer. Both downconverter and upconverter mixers are equipped with this parallel path. This effectively nulls the third harmonic content of the LO. The phase and amplitude matching between the two phases is important, but somewhat relaxed since the maximum desired notch depth would hardly exceed 30 dB for most applications. The next problematic harmonic is the fifth harmonic, at a maximum strength of nearly 14 dB. Given that the RC filter can give at least an additional 6 dB, a notch depth of nearly 24 dB is readily achievable.

Another issue with the active harmonic filter is that it cannot be used to reject the image frequency if the same CLK signal is used for the LO downconverter. To understand this, consider the scenario shown in Fig. 3.43. When using the same LO, the CBPF is centered at the image frequency at a distance of IF = IM − LO. Since this is the same distance from the channel (LO − CH = IF), a component of the image blocker smears the desired signal depending of the I/Q mismatch in the downcon-

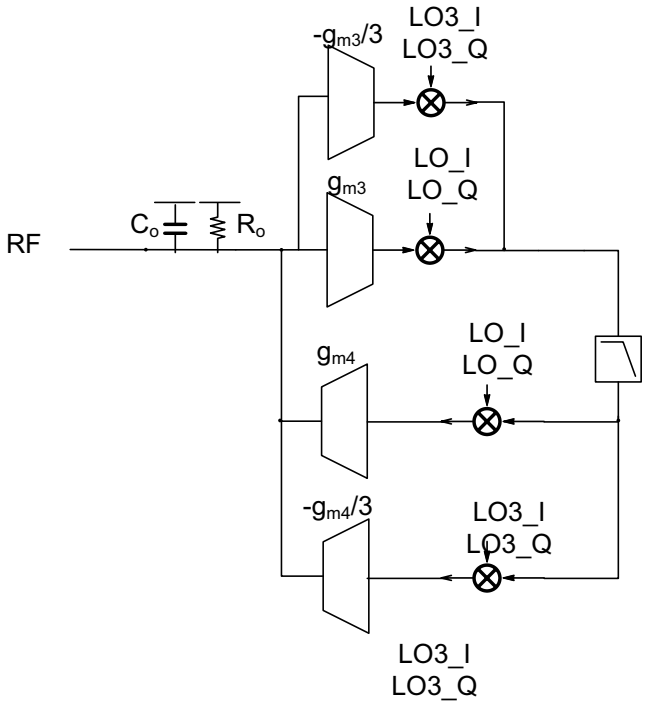


Fig. 3.42 Practical active sampled filter with third harmonic path

verter mixer. It is important to note that this smearing occurs before the CBPF is able to filter the image blocker. Assuming an I/Q matching of 30–40 dB, this means that a significant portion of the image may now smear the desired signal. After upconversion, the blocker may not be well filtered due to a portion of the signal appearing in the blocker; and worse yet, a portion of the blocker may now alias into the signal at RF before any downconversion. This shows that if filtering the image blocker is desired, then a different LO signal is required to avoid this issue.

The assumption in all the above analysis is that only one LO signal is used. As seen above, with the image rejection concern, a different LO may be required. This of course, would add complexity to the design not only from a hardware point of view but also from a frequency planning point of view. Since both CLK signals can contain harmonics, a blocker can now beat with any integer multiple of the filter CLK's harmonics and the main downconverter's LO harmonics to smear the down-converted desired signal. Although complex, this analysis is tractable and resembles that of frequency planning in heterodyne receivers.

3.4.5 Wideband Complex Sampled Filters

As we have seen in the last section, the use of complex filters can be advantageous. One interesting receiver architecture that has been proposed in the past is called a

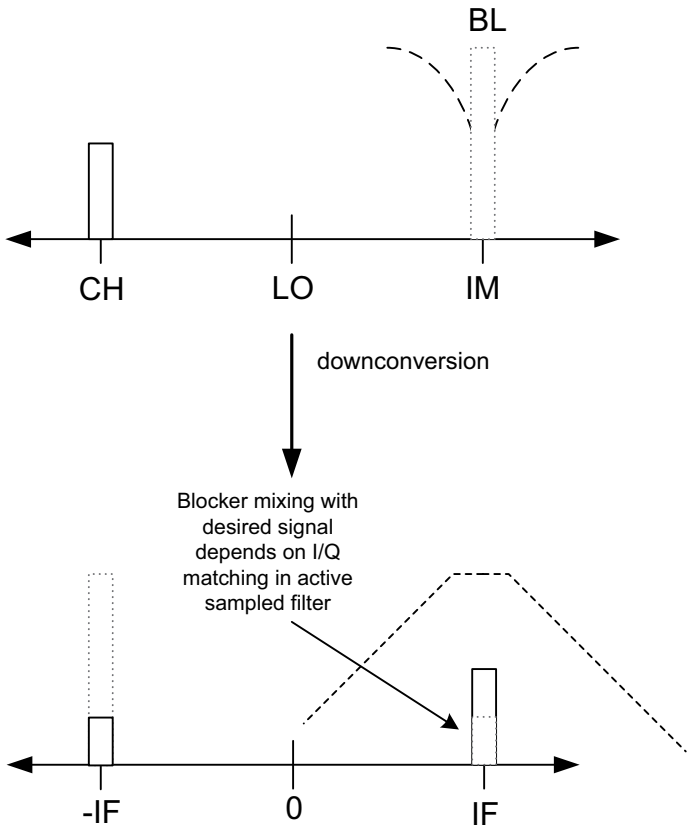


Fig. 3.43 Image rejection in active sampled filter

double quadrature downconverter, shown in Fig. 3.44 [62]. As the figure shows, it relies on generating a quadrature RF signal, then using two sets of quadrature downconverter mixers BB I and Q channels are formed. A polyphase filter then follows the mixers. The main purpose of such a receiver is to offer some image rejection at RF before downconversion, to augment the image rejection realized by the BB polyphase filter.

To understand why converting the RF signal into a complex signal aids in image rejection, recall the discussion in the beginning of Sect. 3.4 when image rejection was introduced. More specifically, it was mentioned that when an image is located at LO+IF and the desired channel is located at LO−IF, the mixing terms −LO+(LO−IF) and LO−(LO+IF) would smear together. Essentially, the signal from the positive frequency and the image from the negative frequency would smear (and vice versa). This means that if a complex RF signal is generated that contains only negative (or positive) frequency content, then image smearing is avoided altogether. This is exactly what a polyphase filter attempts to do.

The real RF signal is converted to a complex (I/Q) RF signal through the use of a passive polyphase filter, consisting of a network of resistors and capacitors,

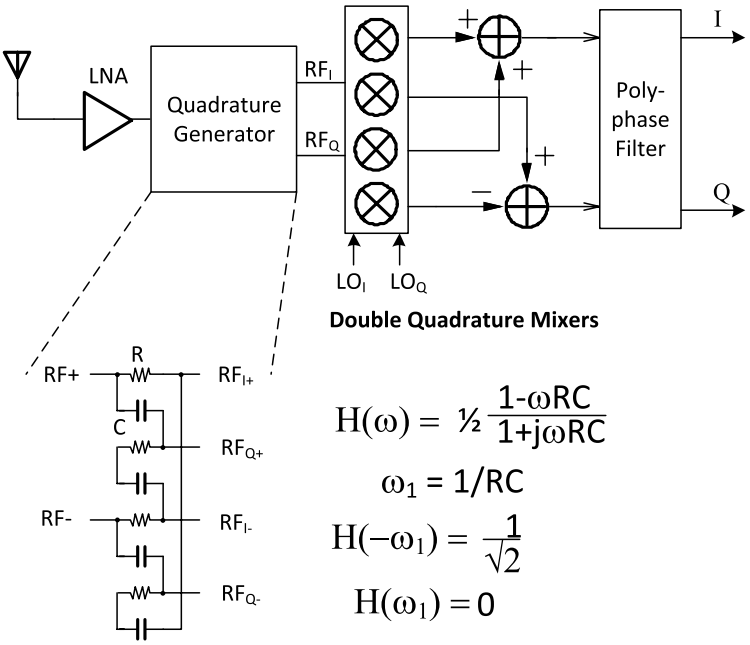


Fig. 3.44 Double quadrature downconverter

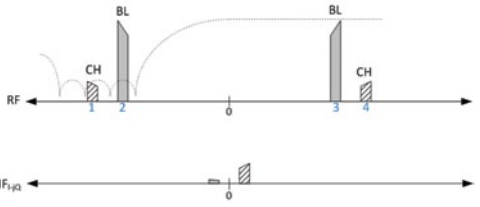


Fig. 3.45 Radio frequency (*RF*) and resulting intermediate frequency (*IF*) spectrum of a double quadrature receiver

as shown in Fig. 3.44. The frequency response of a polyphase filter is asymmetric around the frequency axis, rejecting negative frequency content while passing positive frequency content, or vice versa. This is done by inserting a complex notch at only one side of the frequency axis, at center frequency given by the product of RC. If wideband image rejection is desired, several polyphase filters can be cascaded, each with a different notch location. An illustration of the RF and the resulting IF frequency spectrum is shown in Fig. 3.45. Notice the four notches in the negative frequency of the RF spectrum, created by a fourth-order polyphase filter.

The main drawback of such a topology is that each polyphase filter stage introduces a loss, hence adding to the receiver NF. Another issue with a passive polyphase filter operating at RF frequencies is that tuning the RC center frequency to account for process and temperature drifts is difficult to implement without increasing the noise and nonlinearity significantly (by adding FET switches to trim the resistor, for example).

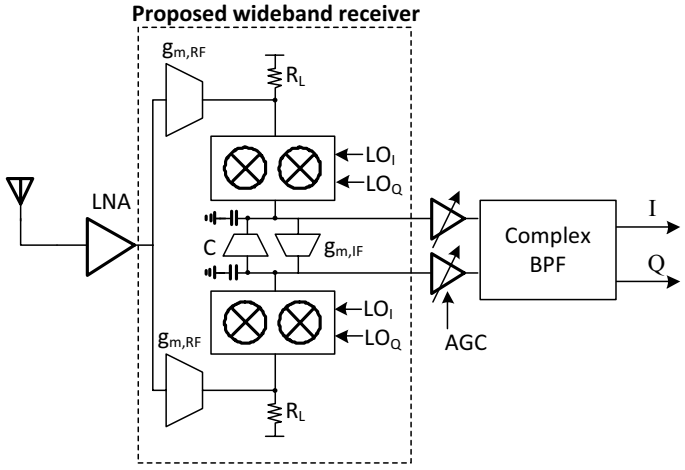


Fig. 3.46 Double quadrature receiver utilizing a complex sampled filter

An alternate method of implementing a double quadrature receiver is through the use of a complex sampled filter, as shown in Fig. 3.46 [26]. The proposed wideband receiver avoids the use of passive polyphase filters and instead upconverts a complex impedance to convert the RF signal into a complex signal. This can be accomplished through the use of a g_m-C resonator core operating at a desired IF frequency. The phase shift at the resonance frequency between the two resonator nodes is 90°. Using passive mixers, each resonator node can be upconverted to two independent RF nodes at a frequency selected by the LO, resulting in a quadrature RF signal.

There are a few advantages to this architecture. First, there is no tuning required for any RF filters, only the low-frequency g_m-C resonator frequency. Secondly, since passive mixers are used, the conversion is bidirectional, meaning that the g_m-C resonator and mixers can be used as downconveters as well. Thirdly, the resulting CBPF is sharp enough to provide some attenuation of blockers in addition to the image rejection.

The proposed double quadrature receiver utilizing a complex sampled filter was implemented in a 0.13-µm CMOS technology. The IF selected was 40 MHz. Figure 3.47 shows a frequency spectrum waveform demonstrating the complex filtering function with a center frequency of 360 MHz (and LO frequency of 400 MHz). The negative frequency filtering function was folded on this plot, as shown. An image rejection of nearly 20 dB is demonstrated with a fairly wide band-pass filter bandwidth of 15 MHz. In addition to image rejection, faraway blockers are also attenuated up to 17 dB.

3.4.6 Comparison and Summary

In this section, a variety of RF filtering techniques has been introduced. This ranged from LC-based RF tracking filters to integrated sampled filters. A comparison

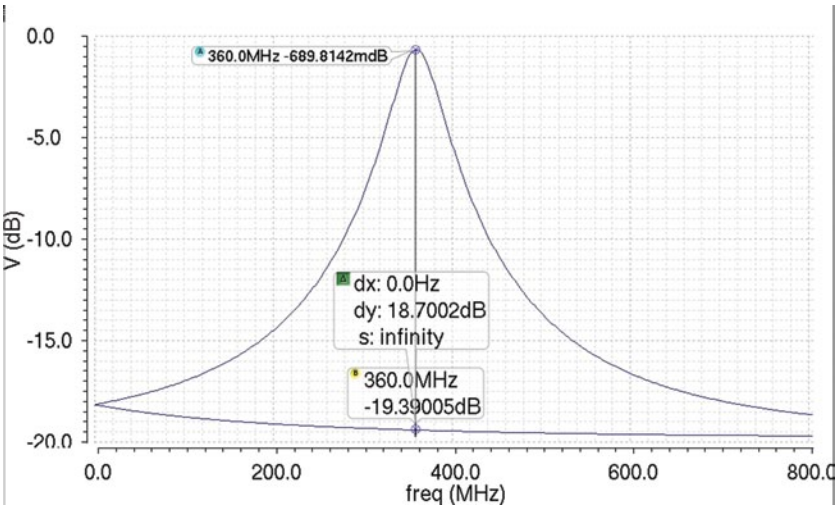


Fig. 3.47 Output frequency spectrum of proposed complex sampled filter

Table 3.9 Comparison summary of various radio frequency (RF) filtering techniques

	LC + varac	LC hybrid on-chip	Active RC	Passive sampled	Active sampled	Complex sampled
Area	Highest	High	Low/mid	Lowest	High	Low
On-chip	No	Partial	Yes	Yes	Yes	Yes
Power	Lowest	Low	Highest	Lowest	High	Low
Max atten	> 60 dB	> 40 dB	> 60 dB	15–20 dB	15–30 dB	15–20 dB
BW (MHz)	10–20	10–50	10–100	1–20 +	1–20 +	5–20 +
Q	5–20	5–20	< 4	> 100	> 100	> 100

BW **bandwidth,** *RC* **resistor capacitor**

summary of the different filtering techniques is shown in Table 3.9. As the table shows, each filter has its own trade-offs. Although sampled filters are powerful tools, there are two words of caution that must be said about them. Firstly, they are incapable of rejecting harmonics of the LO (or CLK signal). The reason for this should be obvious from their name. They become more powerful when combined with a continuous-time filter. Such an example was demonstrated when a continuous-time RC filter was used to augment the active sampled filter to improve its notch depth. The second note about sampled filters is that blockers may beat or mix with the LO signal to produce tones that smear the desired signal. Another possible scenario is that a blocker at a benign frequency can mix with the LO to a problematic frequency, which can saturate the receiver. Careful frequency planning is always advised when using sampled filters.

3.5 Downconversion Mixers

Once the RF spectrum has been sufficiently preconditioned (filtered and gained up/down), the RF spectrum frequency is translated by the downconversion mixer. The main purpose of the downconversion mixer is to simplify the requirements on the ADC. By translating the frequency down to BB or an IF, the required operating frequency of the ADC is lowered and as a result its power consumption and resolution are enhanced. As technology is scaled, this advantage diminishes. The main drawback of downconversion mixers is that the LO signal used to drive mixers is usually polluted with higher-order harmonics causing secondary undesirable frequency translations. One such translation that was introduced in the previous section is image blocker smearing. Another one that was also introduced earlier is harmonic blocker smearing. Adding image and harmonic blocker rejection into the downconversion mixer greatly enhances the performance of the receiver.

3.5.1 Image Reject Mixer

Image rejection in mixers was somewhat introduced in the previous section. A generic downconverter is shown in Fig. 3.48. The quadrature I/Q signals are combined to form a real signal. This is accomplished by first phase-shifting the Q signal by 90° then subtracted from the I signal.

The detailed steps of how the RF signal is downconverted is shown in Fig. 3.49. A sample real signal (labeled 1 and 4) and a real blocker (labeled 2 and 3) are shown. The blocker is located at the image frequency. The quadrature LO signals are shown in the second and third row. The fourth row is the downconverted signal in the I channel. As shown, both the image and the channel occupy the same frequency spectrum with the same polarity. The fifth row is the downconverted signal in the Q channel. The main difference here is that the negative frequency content of the blocker is inverted. Adding the fourth and fifth row together cancels the image blocker, but it also cancels the desired signal. For this reason the fifth row is multiplied by a 90°, or polyphase shifted first with respect to the I channel. The mathematical operation of a polyphase shift, in this case, is that it inverts all negative frequency content. So now, we can subtract a 90° phase-shifted Q signal from I and

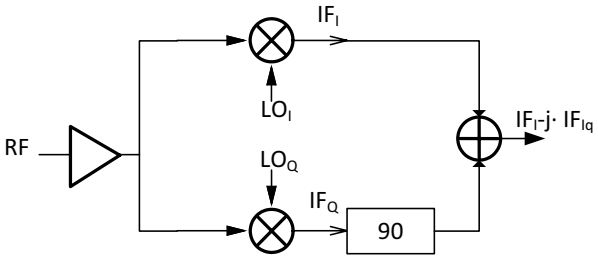


Fig. 3.48 Generic downconverter with real signal output

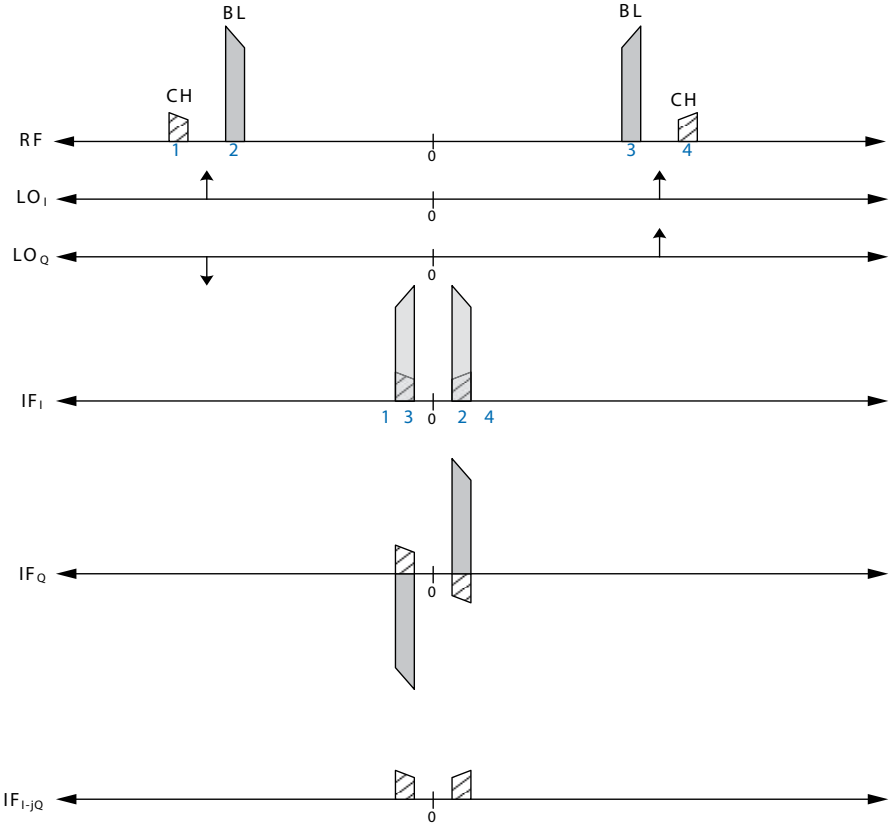


Fig. 3.49 Downconversion steps resulting in a real downconverted signal

the image blocker should cancel out and the desired signals add in phase, resulting in the spectrum shown in the last row of Fig. 3.49.

As is implied in the last step of Fig. 3.49, two signals need to cancel out perfectly in order to eliminate the image. In many applications, image rejection exceeding 60 dBc is required. The main cause of degradation of image rejection in receivers is mismatch in both phase and amplitude in the I and Q channels of the downconversion mixer. More specifically, the image rejection ratio (IRR) can be defined as [57]:

$$IRR = 20log_{10}\left(\frac{1+(\Delta A)^2-2(\Delta A)cos(\Delta\theta)}{1+(\Delta A)^2+2(\Delta A)cos(\Delta\theta)}\right) \tag{3.26}$$

where ΔA is the gain mismatch and Δθ is the phase mismatch between the I and Q channels. The gain mismatch results primarily from the mixer devices themselves. This mismatch is usually random device mismatch, or systematic mismatch due to layout asymmetries. The phase mismatch, however, results primarily from phase mismatches in the quadrature LO signals going into the mixer due to different parasitic paths. This mismatch component can be systematic and/or frequency dependent. IRR values of 30–50 dB are typically achievable without calibration [24].

3.5.2 Harmonic Reject Mixer

Another important feature in downconversion mixers is the ability to avoid blockers near the harmonics of the LO from smearing the desired signal at BB. The LO is always assumed to be either a square wave signal or a close approximation to a square wave, such that the even-order harmonics are small compared to the odd-order harmonics of the LO. The harmonic rejection issue is demonstrated in Fig. 3.50. The figure demonstrates blockers near the third and fifth harmonics of the LO. The desired signal is near LO. After downconversion, the desired signal is downconverted and smeared with both blockers.

One popular method of building harmonic rejection into the downconverter mixer is to "linearize" the LO signal [28]. This can be done by appropriately weighting eight phases of the LO signal as shown in Fig. 3.51. This effectively creates an LO waveform that looks more sinusoidal and eliminates the two nearest harmonics, namely the third and fifth harmonics.

This concept was extended in [30] by replacing the LO signal with a direct digital synthesizer (DDS). An input CLK frequency of 3 GHz was used (the maximum LO frequency synthesized was 1 GHz). The sine wave is synthesized by binary weighting the RF signal before multiplying (or mixing it) with the appropriate bit from the DDS. A 10-bit DDS was chosen to guarantee a 60-dB spurious-free dynamic range (SFDR). The mixer is now essentially a bank of well-matched and weighted transconductor cells with switching devices. The outputs of the switching devices is current and are summed by a resistor at BB Fig. 3.52.

Summary

In this chapter, the components of a wideband receiver have been detailed. First, a discussion on the requirements for wideband receivers was given. Wideband LNA topologies were then discussed and categorized into either CG techniques or feedback techniques. A detailed analysis and trade-off of each was given. The concept

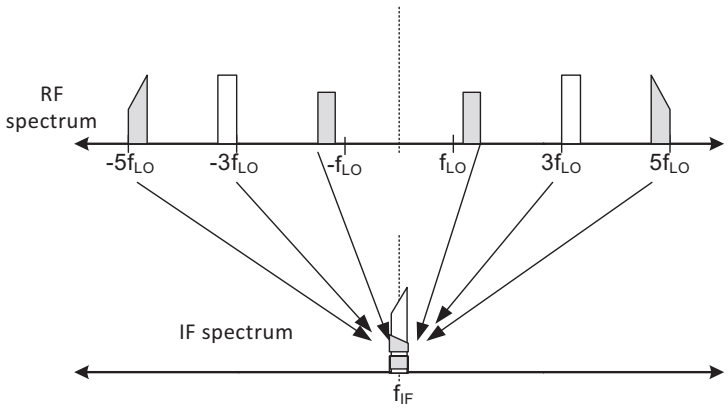


Fig. 3.50 Demonstration of harmonic blocker issue in receivers

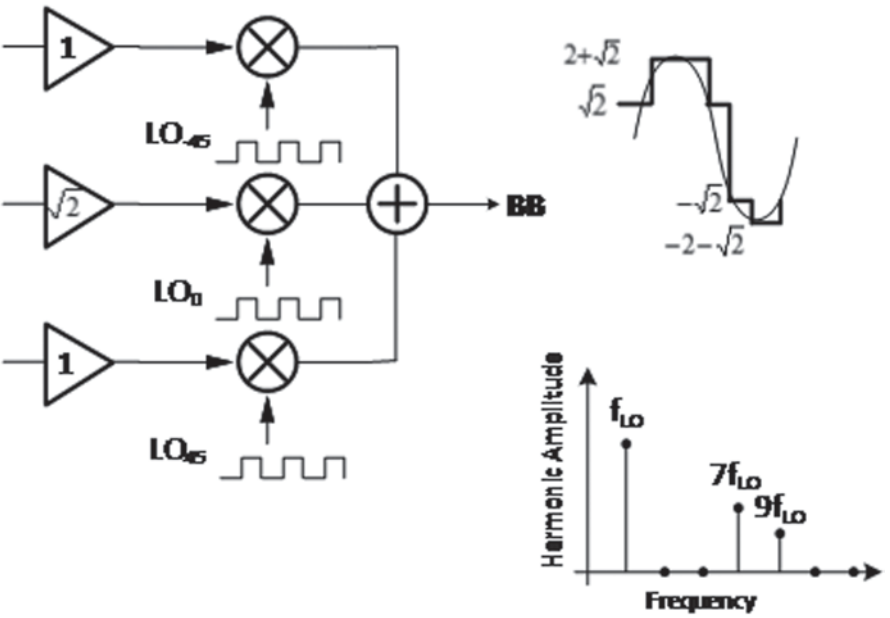


Fig. 3.51 Harmonic reject mixer by eight local oscillator (*LO*) phases

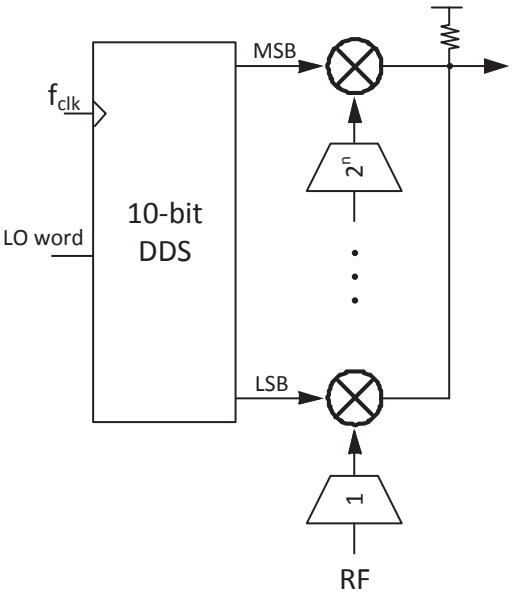


Fig. 3.52 A direct digital synthesizer (*DDS*) approach to harmonic rejection

of variable gain control was discussed. Different techniques for gain control were detailed. RF tracking filter techniques were then given. Each technique was detailed and the design trade-offs of each was given. Finally, a discussion of image and harmonic rejection techniques in downconversion mixers was given.

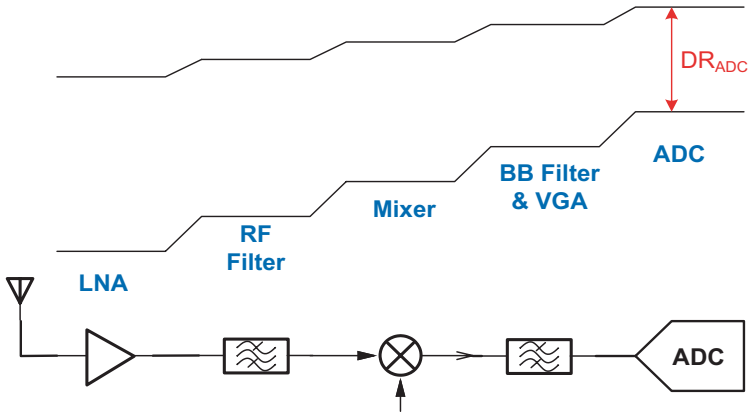


Fig. 3.53 Gain lineup example in a receiver

One important concept in receiver design is that of cascaded performance. As was discussed earlier, the DR of the receiver is maximized by adjusting the gain control of the LNA and possibly BB amplifiers. One possible scenario of a gain lineup is shown in Fig. 3.53. The optimization starting point is from the ADC. Given the ADC performance, the requirements for the BB filters, VGA, gain control in LNA, RF filter are all determined. If a high-resolution ADC is designed, then the receiver components can be greatly simplified as more of the filtering can be accomplished in the digital domain. The two lines shown represent the DR of each block. The lower line represents the noise floor of the block and the upper line is represents the P1dB of each block. As the figure shows, the LNA has the best noise floor, but the worst P1dB performance. On the other hand, the BB filters and VGA have the worst noise performance, but the best linearity numbers. This type of lineup is typical in receiver design.

The cascaded NF is dominated by the LNA, with each block contributing depending on the gain of the block preceding it. More specifically, the cascaded noise factor of a receiver is given by Friis' formula as [69]:

$$F = F_{LNA} + \frac{F_{TF} - 1}{G_{LNA}} + \frac{F_{mix} - 1}{G_{LNA}G_{TF}} + \frac{F_{BB} - 1}{G_{LNA}G_{TF}G_{mix}} + \frac{F_{ADC} - 1}{G_{LNA}G_{TF}G_{mix}G_{BB}} \quad (3.27)$$

where F_i is the noise factor of an individual block i and G_i is the power gain of an individual block i.

Conversely, the cascaded linearity performance is dominated by ADC, with each preceding block contributing depending on the gain of the previous block. More specifically, the cascaded IIP3 for a receiver is given by [69]:

$$\frac{1}{IIP3^2} = \frac{1}{IIP3^2_{LNA}} + \frac{G^2_{LNA}}{IIP3^2_{TF}} + \frac{G^2_{LNA}G^2_{TF}}{IIP3^2_{mix}} + \frac{G^2_{LNA}G^2_{TF}G^2_{mix}}{IIP3^2_{BB}} + \frac{G^2_{LNA}G^2_{TF}G^2_{mix}G^2_{BB}}{IIP3^2_{ADC}} \quad (3.28)$$

This is a linear scale value. To convert it to decibels, ten times logarithm of (3.28) must be taken.

References

1. A. Abrel, “Pure-mode Gm-C biquad filters,” IEEE Int'l Conference on Electonics, Circuits and Systems (ICECS) 1999, pp. 493–496.
2. A. Amer, E. Hegazi, H. Ragaie, “A 90-nm Wideband Merged CMOS LNA and Mixer Exploiting Noise Cancellation,” *IEEE J. of Solid-State Circuits (JSSC)*, vol. 42, no. 2, Feb 2007, pp. 323–328.
3. C. Andrews and A. Molnar, “A Passive Mixer-First Receiver with Digitally Controlled and Widely Tunable RF Interface,” *IEEE J. of Solid-State Circuits*, vol. 45, no. 12, Dec 2010, pp. 2696–2708.
4. C. Andews and A. Molnar, “Implications of Passive Mixer Transparency for Impedance Matching and Noise Figure in Passive Mixer-First Receivers,” *IEEE Trans on Circuits and Systems I*, vol. 57, no. 12, Dec 2010, pp. 3092–3103.
5. C. Andrews, et. al., “A Wideband Receiver with Resonant Multi-Phase LO and Current Reuse Harmonic Rejection Baseband,” *IEEE J. of Solid-State Circuits*, vol. 48, no. 5, May 2013, pp. 1188–1198.
6. A. Ansari, M. Yavari, “A Very Wideband Low Noise Amplifier for Cognitive Radios,” *IEEE Int'l Conf on Electronics, Circuits and Systems (ICECS)*, pp. 623–626, 2011.
7. J. Bae, S. Kim, H. Cho, I. Lee, D. Ha, S. Lee, “A CMOS Wideband Highly S. Woo, W. Kim, C. Lee, H. Kim, J. Laskar, “A Wideband Low-Power CMOS LNA with Positive-Negative Feedback for Noise, Gain, and Linearity Optimization,” *IEEE Trans on Microwave Theory and Techniques*, vol. 60, no. 10, Oct 2012, pp. 3169–3178.
8. J. Bae, et. al., “A CMOS Wideband Highly Linear Low-Noise Amplifier for Digital TV Applications,” *IEEE Trans on Microwave Theory and Techniques*, vol. 61, no. 10, Oct 2013, pp. 3700–3711.
9. A. Behzad, *Wireless LAN Radios*, Hoboken, New Jersey, John Wiley & Sons, Inc., 2008.
10. R. Berenguer, et. al., “Design of a highly integrated tuner suitable for analog and digital TV systems,” *European Solid-State Circuits Conference (ESSCIRC)*, 2004, pp. 335–338.
11. S. Blaakmeer, E. Klumperink, D. Leenaerts, B. Nauta, “Wideband Balun-LNA with Simultaneous Output Balancing, Noise-Canceling and Distortion Canceling,” *IEEE J. of Solid-State Circuits (JSSC)*, vol. 43, no. 6, June 2008, pp. 1341–1350.
12. S. Boppana, “Flicker noise mitigation in direct-conversion receivers for OFDM systems,” *ICASSP 2009*, pp. 2593–2596.
13. J. Borremans, P. Wambacq, C. Soens, Y. Rolain, M. Kuijk, “Low-Area Active-Feedback Low-Noise Amplifier Design in Scaled Digital CMOS,” *IEEE J. of Solid-State Circuits*, vol. 43, no. 11, Nov 2008, pp. 2422–2433.
14. M. Brandolini, et. al., “A+78 dBm IIP2 CMOS Direct Downconversion Mixer for Fully Integrated UMTS Receivers,” *IEEE J. of Solid-State Circuits*, vol. 41, no. 3, March 2006, pp. 552–559.J. Rogin, et. al, “A 1.5-V 45-mW Direct-Conversion WCDMA Receiver IC in 0.13um CMOS,” *IEEE J. of Solid-State Circuits*, vol. 38, no. 12, Dec 2003, pp. 2239–2248. duplex standards
15. F. Bruccoleri, E. Klumperink, B. Nauta, *Wideband Low Noise Amplifiers Exploiting Thermal Noise Cancellation*, Springer: Dordrecht, The Netherlands, 2005.
16. R. Caverly, *CMOS RFIC Design Principles*, Norwood, MA, Artech House, 2007.
17. P. Chang, S. Hsu, “A Compact 0.1-14 GHz Ultra-Wideband Low-Noise Amplifier in 0.13um CMOS,” *IEEE Trans. on Microwave Theory and Techniques*, vol. 58, no. 10, Oct 2010.
18. S. Chehrazi, A. Mirzaei, R. Bagheri, A. Abidi, “A 6.5 GHz Wideband CMOS Low Noise Amplifier for Multi-Band Use,” *IEEE Custom Integrated Circuits Conference (CICC)*, 2005, pp. 801–804.
19. M. Chen, J. Lin, “A 0.1-20 GHz Low-Power Self-Biased Resistive-Feedback LNA in 90 nm Digital CMOS,” *IEEE Microwave and Wireless Component Letters*, vol. 19, no. 5, May 2009, pp. 323–325.

20. H. Chen, D. Chang, Y. Juang, S. Lu, "A Compact Wideband CMOS Low-Noise Amplifier Using Shunt Resistor-Feedback and Series Inductive-Peaking Techniques," IEEE Microwave and Wireless Components Letters, vol. 17, no. 8, Aug 2007, pp. 616–618.
21. W. Chen, G. Liu, B. Zdravko, A. Niknejad, "A Highly Linear Broadband CMOS LNA Employing Noise and Distortion Cancellation," IEEE J. of Solid-State Circuits (JSSC), vol. 43, no. 5, May 2008, pp. 1164–1176.
22. M. Choi and S. Choi, "Self-mixed Interference Cancellation Method in Direct Conversion Receivers," IEEE Int'l Conference on Consumer Electronics (ICCE) 2013, pp. 411–412.
23. Y. Ding and R. Harjani, High-linearity CMOS RF Front-End Circuits, New York, Springer, 2005.
24. M. Elmala, and S. Embabi, "Calibration of Phase and Gain Mismatches in Weaver Image-Reject Receiver," IEEE J. of Solid-State Circuits, vol. 39, no. 2, Feb 2004, pp. 283–289.
25. H. Elwan, A. Tekin, K. Pedrotti, "A Differential ramp based 65dB-Linear VGA Technique in 65 nm CMOS," IEEE J. of Solid-State Circuits, vol. 44, no. 9, 2009, pp. 2503–2514.
26. A. Fahim, "A DC to 2 GHz downconverter with image rejection and high blocker tolerance for cognitive radios," IEEE Radio and Wireless Symposium (RWS) 2014, pp. 91–93.
27. Fitch, M., et. al., "Wireless service provision in TV white space with cognitive radio technology: A telecom operator's perspective and experience," IEEE Communications Magazine, vol. 49, no. 3, 2011, pp. 64–73.
28. T. Forbes, et. al, "Design and Analysis of Harmonic Rejection Mixers with Programmable LO Frequency," IEEE J. of Solid-State Circuits, vol. 48, no. 10, Oct 2013, pp. 2363–2374.
29. Goyal, S., et. al., "Analyzing a full-duplex cellular system," 47th Annual Conference on Information Sciences and Systems, 2013, pp. 1–6.
30. J. Greenberg, et. al., "A 40 MHz-to-1 GHz Fully Integrated Multistandard Silicon Tuner in 80 nm CMOS," IEEE Int'l Solid-State Circuits Conference (ISSCC) 2012, pp. 162–163
31. Gruet, C., et. al., "The LTE Evolution: Private Mobile Radio Networks," IEEE Vehicular Technology Magazine, vol. 8, no. 2, 2013, pp. 64–70.
32. D. Grunigen, et. al., "An Integrated CMOS Switched-Capacitor Bandpass Filter Based on N-Path and Frequency-Sampling Principals," IEEE J. of Solid-State Circuits, vol. 18, no. 6, Dec 1983, pp. 753–761.
33. H. Han, T. Kim, "A CMOS Programmable-Gain Amplifier for Digital TV with a + 9dBm IIP3 Cross-Coupled Common Gate LNA," IEEE Trans on Circuits and Systems II, vol. 59, no. 9, Sep 2012, pp. 543–547.
34. Huschke, J. et. al., "Spectrum requirements for TV broadcast services using cellular transmitters," IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN), 2011, pp. 22–31.
35. D. Im, "A + 9dBm Output P1dB Active Feedback CMOS Wideband LNA for SAW-less Receivers," IEEE Trans on Circuits and Systems II, vol. 60, no. 7, July 2013, pp. 377–381.
36. Jack, K., Video Demystified: A Handbook for the Digital Engineer, 5th Edition, Burlington, MA, Elsevier, 2007.
37. O. Jamin, et. al., "On-chip auto-calibrated RF Tracking filter for Cable Silicon Tuner," European Solid-State Circuits Conference (ESSCIRC), 2008, pp. 158–161.
38. P. Jamshidi, Sasan Naseh, "Wideband LNA with Reactive Feedback at the Input Matching Network," IEEE Int'l Conf on Electronics, Circuits and Systems (ICECS), pp. 330–333, 2011.
39. M. Kaltiokallio, et. al., "Wideband 2 to 6 GHz RF Front-End with Blocker Filtering," IEEE J. of Solid-State Circuits, vol. 47, no. 7, July 2012, pp. 1636–1645.
40. Y. Kanazawa, et. al., "A 130M to 1 GHz Digitally Tunable RF LC-Tracking Filter for CMOS RF receivers," IEEE Asian Solid-State Circuits Conference, 2008, pp. 469–472.
41. Y. Kang, "High-Voltage Analog System for Mobile NAND Flash," IEEE J. of Solid-State Cir. cuits, vol. 43, no. 2, Feb 2008, pp. 507–517.
42. J. Kim, J. Silva-Martinez, "Wideband Inductorless Balun-LNA Employing Feedback for Low-Power Low-Voltage Applications," IEEE Trans on Microwave Theory and Techniques, vol. 60, no. 9, Sep 2012, pp. 2833–2842.

43. T. Lee, *The Design of CMOS Radio-Frequency Integrated Circuits*, Cambridge, UK, Cambridge University Press, 2004.
44. L. Lee, "CMOS LNA for full-band ultra-wideband systems using a simple wide input matching network," *IET Microwaves, Antennas & Propagation*, vol. 4, no. 12, 2010, pp. 2155–2169.
45. H. Lee, C. Wang, C. Wang, "A 0.2-2.6 GHz Wideband Noise-Reduction Gm-Boosted LNA," *IEEE Microwave and Wireless Components Letters*, vol. 22, no. 5, May 2012, pp. 269–271.
46. S. Lerstaveesin, et. al, "A 48-860 MHz CMOS Low-IF Direct-Conversion DTV Tuner," *IEEE J. of Solid-State Circuits*, vol. 43, no. 9, Sep. 2008, pp. 2013–2024.
47. C. Li, Sh. Chou, G. Ke, P. Huang, "A Power-Efficient Noise Suppression Technique Using Signal-Nulled Feedback for Low-Noise Wideband Amplifiers," *IEEE Trans on Circuits and Systems II*, vol. 59, no. 1, Jan 2012, pp. 1–5.
48. Liebermann, D., "Cross-modulation figure of merit for transistor amplifier stages," *Proceedings of the IEEE*, vol. 58, no. 7, 1970, pp. 1063–1071.
49. Mao, B., et. al., "Investigation on the ENOB and clipping effect of real ADC and AGC in coherent QAM transmission system," *OFC/NFOEC*, 2012, pp. 1–3.
50. Matzek, T., "Television IF selectivity and adjacent channel interference," *IRE Transactions on Broadcast and Television Receivers*, vol. BTR-5, no. 1, 1959, pp. 18–21.
51. A. Meaamar, C. Boon, K. Yeo, M. Do, "A Wideband Low Power Low-Noise Amplifier in CMOS Technology," *IEEE Trans. On Circuits and Systems I*, vol. 57, no. 4, April 2010, pp. 773–782.
52. A. Mirzaei, "A 65 nm CMOS Quad-Band SAW-Less Receiver SoC for GSM/GPRS/EDGE," *IEEE J. of Solid-State Circuits*, vol. 46, no. 4, Apr 2011, pp. 950–964.
53. A. Mirzaei, et. al., "A Frequency Translation Technique for SAW-less 3G Receivers," *IEEE Symposium on VLSI Circuits 2009*, pp. 280–281.
54. A. Mirzaei, et. al., "A Low-Power Process-Scalable Super-Heterodyne Receiver with Integrated High-Q Filters," *IEEE J. of Solid-State Circuits*, vol. 46, no. 12, Dec 2011, pp. 2920–2932.
55. S. Moezzi, S. Bakhtiar, "Wideband LNA Using Active Inductor with Multiple Feed-Forward Noise Reduction Paths," *IEEE Trans on Microwave Theory and Techniques*, vol. 60, no. 4, April 2012, pp. 1069–1078.
56. A. Molnar and C. Andrews, "Impedance, Filtering and Noise in N-phase Passive CMOS Mixers," *IEEE Custom Integrated Circuits Conference (CICC) 2012*, pp. 1–8.
57. R. Montemayor and B. Razavi, "A Self-Calibrating 900-MHz CMOS Image-Reject Receiver," *IEEE European Solid-State Circuits Conference (ESSCIRC)*, 2000, pp. 320–323.
58. D. Morgan, *Surface Acoustic Wave Filters*, London, UK, Elsevier, 2007.
59. D. Murphy, et. al., "A Blocker-Tolerant Wideband Noise-Cancelling Receiver with a 2dB Noise Figure," *IEEE Int'l Solid-State Circuits Conference*, pp. 74–75, 2012.
60. D. Murphy, et. al., "A Blocker-Tolerant, Noise-Cancelling Receiver Suitable for Wideband Wireless Applications," *IEEE J. of Solid-State Circuits*, vol. 47, no. 12, Dec 2012, p. 2943–2963.
61. Namgoong, W. and Meng, T. H., "Direct-conversion RF receiver design," *IEEE Transactions on Communications*, vol. 49, no. 3, 2001, pp. 518–529.
62. M. Notten, et. al., "A low-IF CMOS double quadrature mixer exhibiting 58dB of image rejection for silicon TV tuners," *IEEE Radio Frequency Symposium (RFIC) 2005*, pp. 171–174.
63. S. Pactitis, *Active Filters: Theory and Design*, Boca Raton, Florida, CRC Press, 2008.
64. J. Pun, J. de Franca, and C. Azeredo-Leme, *Circuit Design for Wireless Communications*, Boston, Kluwer Academic Publishers, 2003.
65. B. Razavi, *RF Microelectronics*, Upper Saddle River, New Jersey, Prentice-Hall, 1998.
66. B. Razavi, Design of Analog CMOS Integrated Circuits, Boston, McGraw-Hill, 2001.
67. H. Rimentel and S. Bampi, "A 50 MHz - 1 GHz Wideband Low Noise Amplifier in 130 nm CMOS Technology," *IEEE SBCCI 2012*, pp. 1–6.
68. R. Schauman, M. Chausi, and K. Laker, *Design of Analog Filters: Passive, Active RC, and Switched Capacitor*, Englewood Cliffs, New Jersey, Prentice-Hall, 1990.

69. D. Shaeffer and T. Lee, *The Design and Implementation of Low-Power CMOS Radio Receivers*, Boston, Kluwer Academic Publishers, 1999.
70. R. Shaumann and M. Valkenburg, *Design of Analog Filters*, New York, Oxford University Press, 2001.
71. E. Sobhy, A. Helmy, S. Hoyos, K. Entesari, E. Sanchez-Sinencio, "A 2.8 mW Sub-2dB Noise-Figure Inductorless Wideband CMOS LNA Employing Multiple Feedback," *IEEE Trans. on Microwave Theory and Techniques*, vol. 59, no. 12, Dec 2011, pp. 3154–3161.
72. S. Song, D. Im, H. Kim, K. Lee, "A Highly Linear Wideband CMOS Low-Noise Amplifier Based on Current Amplification for Digital TV Tuner Applications," *IEEE Microwave and Wireless Components Letters*, vol. 18, no. 2, Feb 2008, pp. 118–120.
73. C. Toumazou, S. Moschytz, B. Gilbert, *Trade-offs in Analog Circuit Design: The Designer's Companion*, Dordrecht, The Netherlands, Kluwer Academic Publishers, 2002.
74. Tourret, J.R., "SiP Tuner with Integrated LC Tracking Filter for Both Cable and Terrestrial TV Reception," *IEEE J. of Solid-State Circuits*, vol. 42, no. 12, 2007, pp. 2809–2821.
75. X. Wang, J. Sturm, N. Yan, X. Tan, H Min, "A 0.6-3 GHz Wideband Receiver RF Front-End With a Feedforward Noise and Distortion Cancellation Resistive-Feedback LNA," *IEEE Trans. on Microwave Theory and Techniques*, vol. 60, no. 2, Feb 2012, pp. 387–392.
76. J. Weldon, "A 1.75 GHz Highly Integrated Narrow-band CMOS Transmitter with Harmonic-Rejection Mixers," *IEEE J. of Solid-State Circuits*, vol. 13, no. 12, Dec. 2001, pp. 2003–2015.

Chapter 4
Wideband Spectrum Sensing Techniques

From a physical standpoint, the distinguishing feature of a cognitive radio (CR) is the spectrum sensing unit (SSU). The SSU must be designed in such a way that it can scan the entire frequency spectrum and locate unused channels. The location of unused channels is not a trivial task for two main reasons. First, the frequency spectrum band may be large and hence sensing the entire spectrum may be impractically time consuming. Second, the detection of weak signals must be distinguished from noise generated by an unlicensed interferer. The second point indicates that a simple energy metric is not sufficient. In this chapter, sensing requirements for CR are listed. This is followed by a set of techniques that can enhance the performance of the SSU.

4.1 Requirements and Challenges

One central component in any cognitive radio (CR) is the ability to detect if channels in a frequency spectrum are being used by primary users. As was described in Chap. 2, the CR user equipment normally does not have a dedicated channel to transmit data. It shares a certain frequency spectrum with the primary user. A spectrum sensor must then be able to scan the entire spectrum and check whether there is an active user in a given channel or not. In Sect. 2.5, a brief introduction to spectrum sensing was given. The remainder of this chapter is dedicated to a more detailed discussion of spectrum sensing, its requirements and implementation.

To appreciate challenge of spectrum sensing compared to regular RF signal reception, one must remember the main purpose of spectrum sensing: to reliably determine the vacancy of a certain channel. This is challenging since the spectrum sensor must be able to distinguish a weak signal from the noise floor. In other words, it must measure the noise with sufficient accuracy to determine it is indeed noise! Consider, for example, the sensing requirements as dictated by the IEEE 802.22 standard, listed in Table 4.1. A receiver noise figure of 6 dB and 0 dBi antenna gain were assumed in the calculations of the receiver noise floor and signal-to-noise

A. Fahim, *Radio Frequency Integrated Circuit Design for Cognitive Radio Systems,*
DOI 10.1007/978-3-319-11011-0_4

Table 4.1 IEEE 802.22 signal sense requirements (FCC 2010 ruling) [1]

Signal type	Minimum level (dBm)	Bandwidth	R_x noise floor (dBm)	SNR, min (dB)
ATSC (digital TV)	−114	6 MHz	−100	−14
NTSC (analog TV)	−114	6 MHz (100 KHz)	−100	−14
Wireless Microphone	−107	200 KHz	−115	+8

IEEE Institute of Electrical and Electronics Engineers, *SNR* signal-to-noise ratio, *ATSC* Advanced Television Systems Committee, *NTSC* National Television System Committee

ratio $(SNR)_{min}$, respectively. It is important to note that the antenna gain of portable devices is usually –3 to –5 dBi. The fact that an SNR < 0 dB is required for detection illustrates on reason why the spectrum sensing requirements are even more stringent than that of a regular wireless receiver.

To understand why SNR < 0 dB must be detected, consider the scenario shown in Fig. 4.1, which is known as the *hidden incumbent problem* [2]. The CR is shown to be some distance from the primary transmitter, which is further from a primary receiver. The primary receiver is designed to receive the data signal from the transmitter with a specified minimum sensitivity level. By the time the signal from the primary transmitter reaches the CR, it will be below the noise floor. If the CR believes that this channel is not used, it will transmit on that same channel. Wireless standards regulate the maximum amplitude of this condition to be below the desired channel by a ratio called *signal-to-interference* (SIR) ratio. For example, analog TV specifies an SIR of 34 dB and digital TV an SIR of 23 dB [3]. This means that the CR must transmit at a level such that its signal, when received by the primary receiver, is still below the received primary signal by SIR dB. In mathematical terms, the minimum SNR required to be sensed by the CR is [4]:

$$SNR_{min} = \frac{P_p L(D,R)}{N} \tag{4.1}$$

where P_p minimum is the transmit power of the primary user, L(D, R) is the path loss of the transmitted signal from the primary transmitter (D) to the primary receiver (R), and N is the worst-case noise floor of the system as seen at the CR. Computing (4.1) would yield the last column shown in Table 4.1.

A sense requirements of SNR < 0 dB also implies a long detection time, where several samples are required to average out the noise floor. This is difficult due to two main reasons. Firstly, there is uncertainty in the channel spectrum. In a wireless environment, the channel experiences fading or shadowing as was demonstrated in Chap. 2. This can cause a channel to seem unused; when in reality the primary user is temporarily blocked or faded. Without using any advanced techniques, as will be shown in Sect. 4.3, this would mandate an additional 20–30 dB of sensitivity requirements on the spectrum sensor.

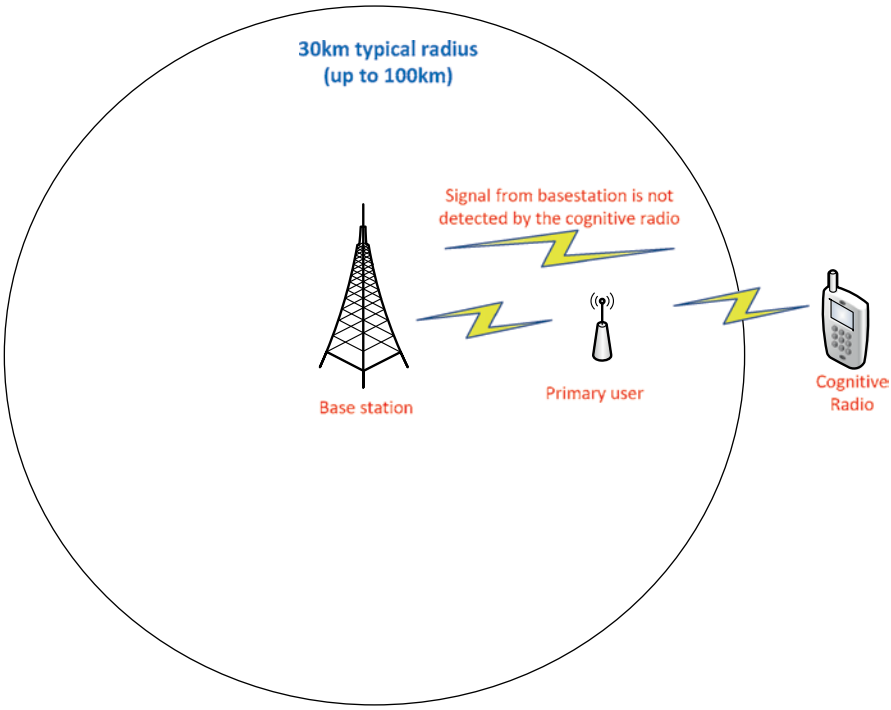


Fig. 4.1 Cognitive radio reception illustration

The second difficulty resulting from a long detection time is uncertainty in the noise floor. As will be shown in the following sections, an estimate of the noise floor in the system is necessary to make an accurate measure of whether a signal is present or not. The noise in the system is time varying due to, again, the time-varying nature of the channel. A nearby blocker, for instance, may be temporarily faded, but then appear at full strength a short time later and raise the noise floor of the channel being sensed.

Besides from stringent noise requirements, wideband receiver linearity may prevent the detection of unused channels. Consider, for example, the scenario shown in Fig. 4.2. In this scenario, large multiple blockers occupy the frequency spectrum. These blockers cause intermodulation and cross-modulation terms that may land in unused channels. These terms would cause the spectrum sensor to falsely report that a channel is used [5].

4.2 Spectrum Sensing Techniques

There are a wide variety of techniques for spectrum sensing, each with its own set of advantages and disadvantages. The issue of spectrum sensing in CRs is still an open research area as this issue has not been satisfactorily resolved. There are

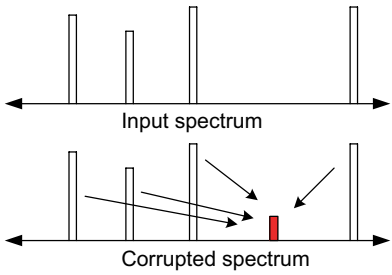


Fig. 4.2 Whitespace not detected due to wideband receiver nonlinearity

Table 4.2 Example of recent study of evaluation of the spectrum sensors of commercial cognitive radio implementations

TVWS Device	Detection rate of ATSC signal (%)	Detection rate of NTSC signal (%)	False alarm rate (%)
Adaptrum	91	89	25
I2R	94	25	19
Motorola	90	–	36
Philips	100	100	85

ATSC Advanced Television Systems Committee, *NTSC* National Television System Committee, *TVWS* television white space

many research papers detailing various methods of spectrum sensing showing excellent performance; however, significant performance degradation is exhibited in commercial settings. Consider, for example, a recent study done to evaluate the performance of various television white space (TVWS) devices from various vendors, shown in Table 4.2 [6]. As the table shows, it is difficult to balance proper detection of various television signals from the false alarm rate. The reason for this depends on where the threshold metric for detection is set. If the detection threshold is too loose, there will be a high probability of collisions, leading to low numbers in columns 2 and 3 of Table 4.2. On the other hand, if the threshold metric is set too tight, then there will be a high probability of false alarms and the spectrum is wasted. For example, the last row shows an implementation from Philips, which is able to have 100 % correct detection rate for both analog and digital TV signals. The same device, however, identified signature of television signals in 85 % of the unused channels!

The detection problem can be formulated as follows. A given channel is either unused or has a primary user active. Let H_0 denote the condition that the channel is unused and H_1 be the condition that the channel is used by a primary user, as shown in Fig. 4.3. The noise signal is denoted as $W[n]$, the primary user signal is denoted as $X[n]$, and the total received signal at the detector is $Y[n]$. The averaging window is N and the output of the detector is $\hat{H}$. The two measures that the detector must determine is the probability of a false alarm (P_{FA}) and the probability of missed detection (P_{MD}). The two metrics can be given as:

$$P_{FA} = P\{\widehat{H} = H_1 \mid H_0\} \tag{4.2}$$

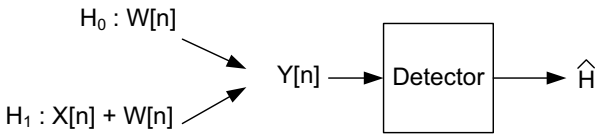


Fig. 4.3 Spectrum sensing problem formulation

and

$$P_{MD} = P\{\widehat{H} = H_0 \mid H_1\} \tag{4.3}$$

Assuming that W[n] is additive white noise with a Guassian distribution, (4.2) and (4.3) can be expressed as:

$$P_{FA} = Q\left(\frac{K - \mu_0}{\sigma_0}\right) \tag{4.4}$$

and

$$P_{MD} = 1 - P_D = 1 - Q\left(\frac{K - \mu_1}{\sigma_1}\right) \tag{4.5}$$

where μ_0 and μ_1 are the mean values of the noise and noise plus signal, respectively, σ_0 and σ_1 are the standard deviations of the noise and noise plus signal, respectively, K is fixed threshold value and Q is the Gaussian distribution given by

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-\frac{z^2}{2}} dz \tag{4.6}$$

The IEEE 802.22 standard specifies that both the probability of a false alarm (P_{FA}) and the probability of a missed detection (P_{MD}) must both be less than 10% for a detector sensitivity of better than −114 dBm (over a 6-MHz channel) [7]. There are several detection methods that will be reviewed next. Each detection method demonstrates a trade-off between hardware simplicity and detection sensitivity.

4.2.1 Energy-Based Sensing

There are three main categories of techniques for spectrum sensing. The first of these techniques is known as energy-based sensing. As its name implies, it simply senses the total energy in a channel. If the energy is small, then the channel is assumed to be vacant. In mathematical terms, the sense energy can be given as [8]:

$$E_{det} = \frac{1}{N} \sum_{m=1}^{N} y(m) \bullet y^*(m) \tag{4.7}$$

where N is the number of samples to be averaged and $y(m)$ is the sampled input signal from the spectrum. For every doubling of N, the detection noise floor is reduced by 3 dB. In theory, N can be increased until the noise floor is pushed below the signal to be detected. Another way to view this is that if there is some uncertainty in detecting the signal, and the uncertainty has a uniformly random distribution, the variance will decrease linearly with N, until the variance is 0 as $N \to \infty$ and the signal can be detected with absolute certainty.

For an energy-based sensing technique, $\hat{H}$ (as shown in Fig. 4.3) can be given as [9]:

$$\hat{H} = \frac{1}{N}\sum_{n=1}^{N}|Y[n]|^2 \underset{\geq}{\leq} \gamma \tag{4.8}$$

where γ is a fixed threshold. As stated earlier, N can be increased until the noise floor is pushed below the signal to be detected. In practice, however, there is a level of uncertainty in the noise floor. This can be due to several factors including time-varying channel conditions as well as variation in the noise figure of the receiver over time. The noise figure of the receiver varies due to temperature drifts or different modules near the receiver turning on or off. This level of uncertainty in noise can be expressed as:

$$NF^{'} = NF + x\ (dBm) \tag{4.9}$$

where x is the uncertainty in received noise level. This uncertainty places a lower bound beyond which the noise cannot be reduced reliably, limiting the usefulness of increasing the averaging window. Instead of having absolute mean values, the P_{MD} and P_{FA} probability distribution now have mean values m_0 and m_1, both of which are random variables. Two scenarios are illustrated in Fig. 4.4. In Fig. 4.4a, the variation in m_0 and m_1 produce a worst-case scenario where the two distributions, P_{FA} and P_{MD}, are close to one another. Tightening up the variances to achieve the desired P_{FA} and P_{MD} levels is possible by increasing the detection window interval, N. In Fig. 4.4b, the variation in m_0 and m_1 is so large that the mean of the two distributions overlap, making it impossible to improve the P_{FA} and P_{MD} numbers by simply increasing the length of the detection window, N. Note that the means values m_0 and m_1 are close together implies a low SNR scenario.

Figure 4.4 illustrates that there is a lower bound on the minimum SNR that is detectable using an energy-based detection caused by uncertainty in the noise level. This lower bound is known as the SNR wall [10]. Figure 4.5 illustrates the dependence of the SNR wall on the uncertainty in the noise floor of the receiver. In practice, a noise figure variation of less than 0.5 dB is difficult to achieve. As the figure shows, a variation of less than 0.1 dB is required to achieve a detection SNR level of −14 dB. For an energy-based detection method, it can be shown that the SNR wall is given by

$$SNR_{wall} = 10log_{10}\left[\frac{10^{2x/10}-1}{10^{x/10}}\right] \tag{4.10}$$

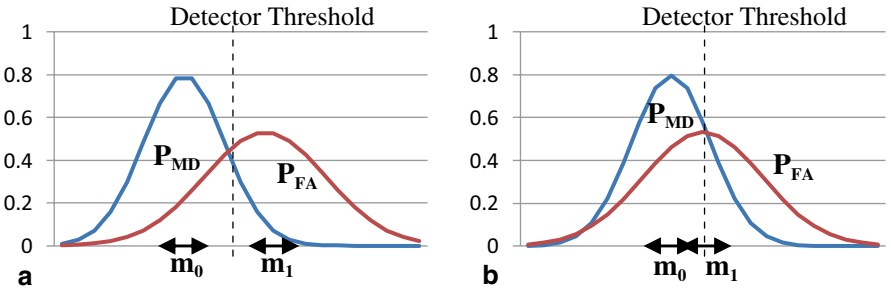


Fig. 4.4. P_{FA} and P_{MD} distributions for **a** signal-to-noise ratio (*SNR*) where the two distributions are far apart and **b** for lower SNR where distributions are closer together

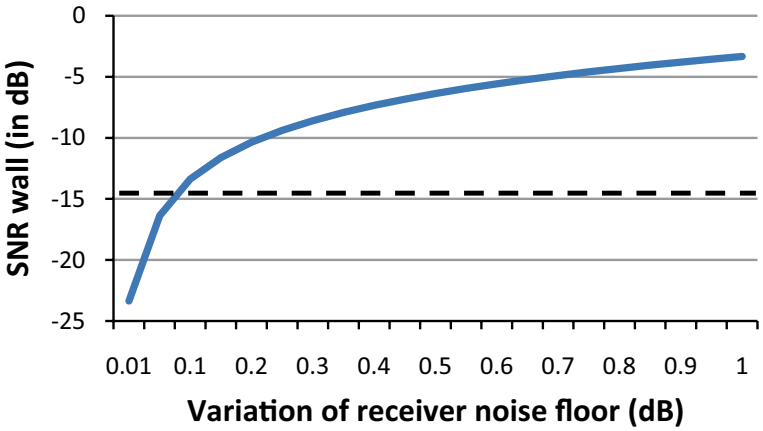


Fig. 4.5 Signal-to-noise ratio (*SNR*) wall as a function of noise floor variation, x

where x is the uncertainty in the noise floor (in decibels).

4.2.2 Feature-Based Sensing

As Fig. 4.5 shows, using an energy-based detection alone would not satisfy the sensitivity requirement of the IEEE 802.22 standard. Fortunately, there are other detection techniques that can reach minimum SNR levels lower than that of energy-based detection methods.

One such category of techniques is known as feature-based detection. In the feature-based detection, certain distinct features of the primary signal are exploited to ease the detection. For example, both analog and digital TV signals have pilot carriers at fixed frequency offsets from the band edge, as shown in Fig. 4.6 . For the Advanced Television Systems Committee (ATSC) (digital TV broadcast), the

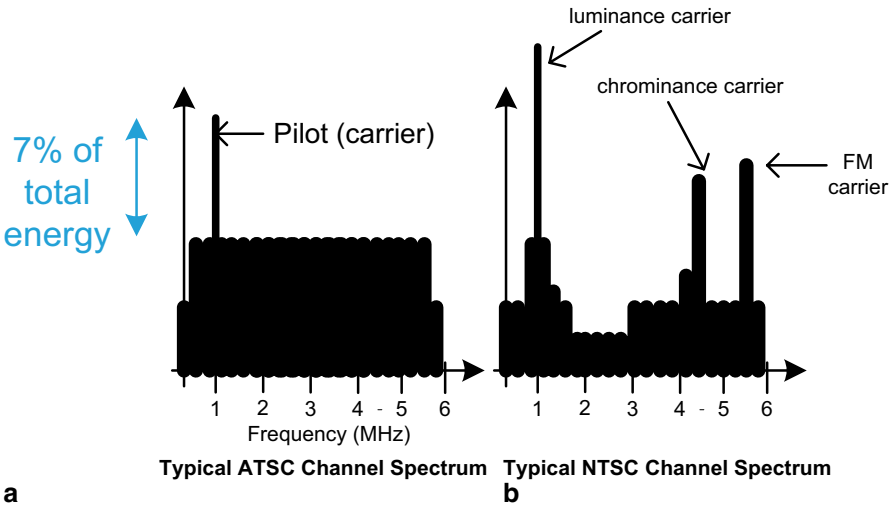


Fig. 4.6 Typical television broadcast signal **a** digital TV (*ATSC* Advanced Television Systems Committee) and **b** analog TV (*NTSC* National Television System Committee)

pilot carrier is at a 310-KHz offset from the band edge and 7% of the total energy of the signal is concentrated at the pilot carrier. The width of the pilot carrier is 19.4 KHz (actually, there are two very narrow pilot tones that are 19.4 KHz apart). For National Television System Committee (NTSC) (analog TV signal), the luminance carrier is 1.25 MHz offset from the band edge.

Another distinguishing feature in digital signals is a cyclic prefix code. A cyclic prefix code is a sequence that is repeated to help the primary receiver acquire phase and frequency lock to the primary transmitter. Cyclic code prefix is used in all practical digital transmission systems, including digital TV broadcast, WiFi, and cellular standards. A CR receiver can exploit this feature by attempting to detect a repeated code. This information is usually extracted by means of a spectral correlation function (SCF) [11].

In the previous section, it was shown that the SNR wall caused by noise floor uncertainty limited the minimum detectable signal. It is instructive to determine if any such limitation exists with pilot detection. In pilot detection, the pilot frequency is downconverted to zero center frequency, then a narrow and sharp low-pass filter is applied (the bandwidth of the low-pass filter would be slightly more than 19.4 KHz). In this case, the noise bandwidth is much smaller, significantly raising the SNR of the pilot signal. Moreover, since the pilot signal is a deterministic signal, with strong correlation between consecutive samples, the noise uncertainty can be averaged out over time. This means that unlike the energy-based detection, the SNR wall in a pilot-based detection can be lowered with an increased number of samples. More specifically, the SNR wall for coherent detection is given by

$$SNR_{wall}^{c} = SNR_{\text{wall}} + 10log\left(\frac{1}{\theta}\right) - 10log(N_c)(dB) \tag{4.11}$$

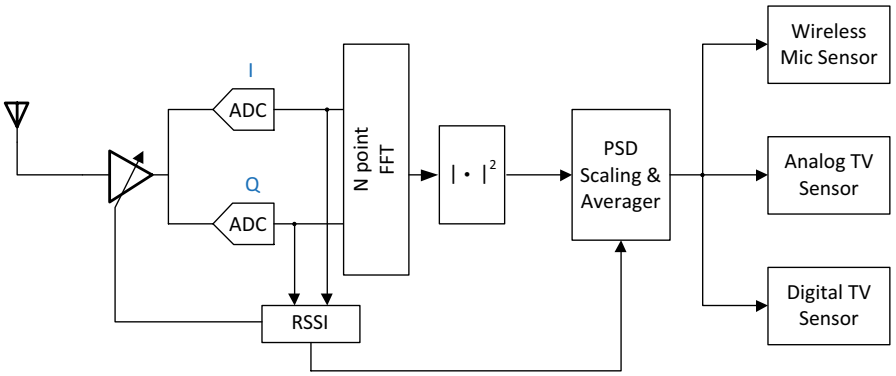


Fig. 4.7 Receiver employing feature-based detection

where the "c" superscript denotes pilot carrier, SNR_{wall} is given by (4.7), θ is the percentage of the energy of the carrier is a fraction of the total signal energy, and N_c is the number of samples of the carrier signal. As (4.8) clearly shows, the SNR wall for pilot carrier detection decreases indefinitely with increased number of samples.

There is, however, a different effect that limits the performance of the pilot-based detection. As was stated earlier, the pilot is downconverted to center frequency centered zero center frequency, then a very sharp and narrow low-pass filter is applied in order to determine the energy of the pilot signal. Typically, this filtering and downconverting is done without coherently demodulating the channel, since this would significantly increase the detection period. This means that frequency offsets in the local oscillator (LO) signal can exist and the downconverted pilot signal would be centered at an offset frequency away from DC. This, in turn, would mean that the low-pass filter would not be centered around the pilot resulting in inaccuracy of the estimated energy of the pilot signal. For a numerical example, consider a channel centered at 1 GHz and a desired energy estimate accuracy of 1%. This means that the frequency accuracy of the LO has to be 194 Hz (19.4 KHz/100), or 0.2 ppm (~194 Hz/1 GHz). This level of frequency accuracy is, in general, difficult to guarantee for a low-cost receiver design.

Another effect that will limit the achievable accuracy with pilot carrier detection is the time-varying nature of multipath fading. If the detection window, N_c, is too large, the pilot signal cannot be considered to have a fixed amplitude and will vary over time. This, undoubtedly, will limit the achievable accuracy of the pilot carrier detection accuracy.

Signal features other than the pilot carrier may be sensed [12–13]. For example, the ATSC signal (the US digital TV standard) contains specific pseudo-noise (PN) sequences in the Data Sync Field [14], a 511-bit PN sequence and three 63-bit PN sequences. The feature detection sensor can contain a correlator that tests the presence of these sequences as they are repeated over time. Also, as stated earlier, a cyclic prefix code, which exists in almost any digital communication standard, can be used in feature detection.

An example of a receiver employing feature-based detection is shown in Fig. 4.7. In this example, the power spectral density (PSD) of the received signal is first

computed. The PSD is then processed by dedicated units that attempt to correlate it to different types of primary users. In this example, the primary user can either be a wireless microphone sensor, analog TV signal, or digital TV signal. The TV sensors may implement a pilot detector. One important implementation detail in this receiver is that the PSD is scaled by a value that is set from the received signal strength indicator (RSSI) unit. The RSSI determines the broadband amplitude of the received signal, which may track any channel fading of the received signal.

4.2.3 Second-Order Statistics-Based Sensing

One popular category of sensing techniques are based on the second-order statistics. In this category of sensing techniques, the noise statistics are always assumed to be Gaussian. One measure that is commonly used is the autocorrelation metric. The autocorrelation of a signal is given by

$$R(\tau) = \frac{E[(x(t) - \mu) \bullet (x(t - \tau) - \mu)]}{\sigma^2} \tag{4.12}$$

where μ is the average value of $x(t)$ and σ^2 is the variance of $x(t)$. If a signal is completely random, the autocorrelation function can is equal to:

$$R(\tau) = \begin{cases} 0 & \tau \neq 0 \\ 1 & \tau = 0 \end{cases} \tag{4.13}$$

The discrete-time autocorrelation function is given as:

$$R_{xx}(j) = \lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N} x_n \bullet \overline{x_{n-j}} \tag{4.14}$$

If a signal is periodic, its autocorrelation function will also be periodic (i.e., it will contain nonzero values when $j \neq 0$). Figure 4.8 shows two cases of the autocorrelation function for a random input (Fig. 4.8a) and a square wave input with zero mean and random noise added to it (Fig. 4.8b). As shown in Fig. 4.8a, the autocorrelation has a peak at $j = 0$, which is expected from (4.14). Figure 4.8b, on the other hand, shows a periodic autocorrelation waveform that is a triangular wave, which is the integration of a square wave, as expected.

The limitation of the autocorrelation signal detection is similar to that of the energy detection. Uncertainty in the noise floor limits the achievable sensing and represents a minimum SNR wall below which signal detection is not possible. Using the autocorrelation approach with feature detection, however, can enhance its overall sensitivity. Figure 4.9 demonstrates a receiver using an autocorrelation unit for spectrum sensing. The delay element adjusts the integration window of the autocorrelation function.

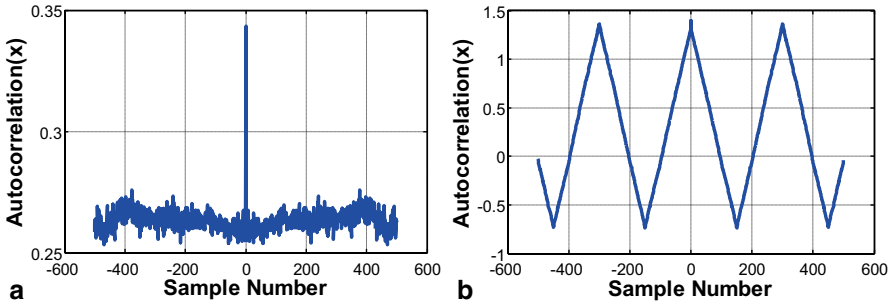


Fig. 4.8 Autocorrelation function for **a** a random signal and **b** a periodic plus random signal

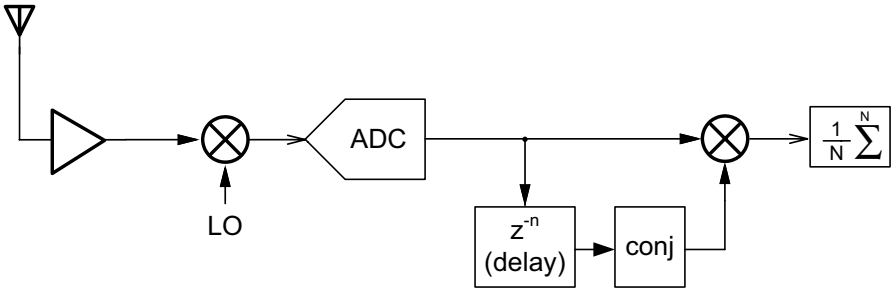


Fig. 4.9 Spectrum sensor with using an autocorrelation operator

Another approach to second-order statistical measure is through the use of the cross-correlation. In this approach, two identical receive channels are used, as shown in Fig. 4.10. The basic idea behind this approach is to cancel out the noise variations that result from the receiver noise floor uncertainty. Essentially, the signal is added constructively (correlated), with the noise of the two receiver paths averaging out since the noise is assumed uncorrelated. If all noise uncertainty is from the receiver unit, then this approach would completely cancel the SNR wall for a sufficiently large sample size. More specifically, the SNR wall for the cross-correlation-based detection is given as [15]:

$$SNR_{wall}^{x} = \lim_{N\to\infty} SNR_{min} = 10log_{10}\left[\rho\left(\frac{\epsilon_2 - \epsilon_1}{1-\epsilon_1}\right)\right]\text{dB} \tag{4.15}$$

where ε_2 and ε_1 are the upper and lower bounds on the noise uncertainty, respectively, and ρ is the correlation of the noise in the two receive paths. The superscript x in (4.15) denotes cross-correlation. N is the number of samples taken to average the cross-correlation estimate. If the noise in the two receive paths are completely uncorrelated, then the correlation coefficient ρ is equal to 0. In this case, the SNR wall is completely broken down and there is no limitation on the detector sensitivity. The result is still valid in the presence of LO phase uncertainty [15]. Compared to energy

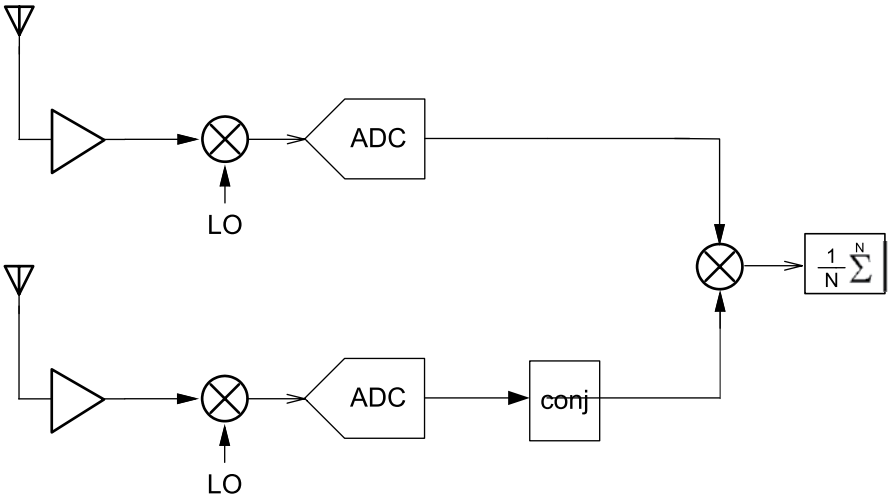


Fig. 4.10 Spectrum sensor using the cross-correlation operator

detection and autocorrelation detection, cross-correlation detection generally yields a 10–30 dB reduction in minimum detectable SNR in a practical implementation. This clearly can meet the IEEE 802.22 requirement of −14 dB SNR (assuming 6 dB noise figure). The main drawback of the cross-correlation detection technique is the extra hardware complexity and power dissipation of the extra receiver path. This complexity can be reduced by sharing hardware between the two paths, which comes at the expense of increased correlation coefficient, ρ.

4.2.4 Summary and Comparison

In this section, the various methods of spectrum sensing are summarized and compared as shown in Table 4.3 below. The simplest form is the energy-based techniques, where the average energy of the signal is sensed. This energy estimate is enhanced by repeating the measurement several times. The theory behind doing so relies on that the noise is Gaussian uniformly distributed additive noise. It also assumes that the noise distribution (mean and variance) are fixed. In practice, however, uncertainty in the noise floor arises due to changes in the receiver circuitry operating conditions (such as temperature) or the channel conditions (such as channel fading). This uncertainty limits the detectable signal to levels that are insufficient for most standards, including the IEEE 802.22 standard.

Another category of techniques is known as the feature-based detection. In these techniques, a distinguishing feature of a signal is searched for to determine if a primary user is present. For broadcast TV standards (both ATSC and NTSC, for example), a pilot signal is used to synchronize the receiver to the incoming signal. Such a signal is located at a fixed frequency offset from the band edge and occupies a specific fraction of the total signal energy. Techniques that exploit this feature to

Table 4.3 Comparison of different spectrum sensing techniques

	Energy detector	Pilot detector	Feature detector	Autocorrelator	Cross-correlator
N (samples)	$1/SNR^2$	$1/\theta \cdot SNR$	k/SNR^2	$1/SNR^2$	$1/SNR^2$
Theoretical limit	SNR wall	LO offset (0.1 ppm)	LO offset (1 ppm)	SNR wall	None
Main Computation	Noise estimator	Pilot energy estimator	FFT, correlator	Auto-correlation	Cross-correlation
Hardware complexity	Low	Mid	High	Mid	Highest

SNR signal-to-noise ratio, *LO* local oscillator, *FFT* fast Fourier transform

detect the presence of a primary user are called pilot detection techniques. Its primary drawback is its sensitivity to LO frequency offsets.

Other feature detection techniques rely on the cyclic nature of certain attributes of the primary signal. Such signals include a cyclic prefix code or fixed preamble signal that occurs periodically. Such periodic signals are known as cyclostationary signals and feature-based detection techniques search for these periodic signals. Such detection techniques are also sensitive to LO frequency offsets, but less so than pilot detection techniques.

Another class of techniques rely on second-order statistics of signals. One such technique relies on the autocorrelation function. The advantage of the autocorrelation detection techniques is that its output is quite distinct in the presence of any periodic signal, which is an indication of the presence of a signal. Its main drawback is that is suffers the same limitations as energy-based detection techniques.

An improvement over the autocorrelation detection technique is the cross-correlation detection technique. It relies on computing the cross-correlation of two identical receive paths with the same input signal. The main assumption here is that the uncertainty in the noise is primarily caused by the receiver circuitry itself. The cross-correlation cancel out these noise components, assuming that the noise is independent between the two receive paths. This noise cancellation effect is most effective if all the noise variation uncertainty occurs in the receiver itself and the noise in the two receive paths are entirely independent. Theoretically, the SNR wall for cross-correlation is completely removed. This technique is also robust to phase shifts in the LO.

4.3 Energy-Efficient Spectrum Sensing Techniques

One of the main bottlenecks in spectrum sensors today is their power dissipation. The spectrum sensor is required to scan the entire frequency spectrum for available channels. As was seen in the previous section, multiple samples per channel are required. This can easily increase the detection time per channel to be a significant fraction of 100 ms. Given that there may be hundreds of channels in the frequency

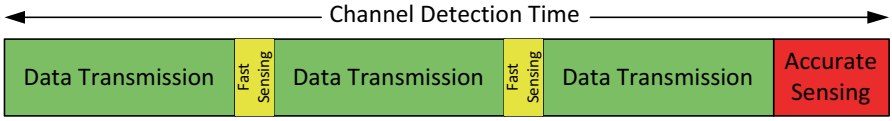


Fig. 4.11 Two-step sensing technique to reduce latency of transmitted signal

bands of interest, the entire spectrum sensing operation may take up to a minute. This would mean that the spectrum sensor would be nearly continuously operating, reducing transmission throughput and increasing the receiver's overall energy consumption.

4.3.1 Adaptive Two-Step Sensing Technique

One method of improving the energy efficiency of spectrum sensors is to employ a two-step sensing technique [16]. In many cases, the used channel is occupied by a primary user that is easy to detect. In this case, a simple energy-based detection technique can be used to detect large primary user signals. If such signal is detected, then no further detection is necessary, significantly cutting down on the detection time and drastically reducing the number of computations required for the spectrum sensing operation for that channel.

If the same channel that is used for transmission is scanned for primary user, then using this two-step sensing technique would also reduce the latency in transmission time. In this case, the fast sensing operations would be simple energy detection operations. After several transmissions, if no primary user was detected, a longer and more accurate sensing operation can be employed. This concept is shown graphically in Fig. 4.11.

4.3.2 Cooperative Spectrum Sensing Techniques

Another approach to reducing energy consumption associated with spectrum sensing is to distribute the task over several CR users within a region. Information about the availability of the spectrum is shared among several nearby users. This enables more accurate spectrum usage information and distributes the heavy burden of sensing the entire frequency spectrum among several users. Techniques based on these principles are known as *cooperative spectrum sensing* [17–19].

Cooperative spectrum sensing techniques have been shown to provide robust detection in the face of multipath fading, shadowing, and receiver noise uncertainty. This is achieved primarily through spatial diversity between the different nodes in the network. Cooperative techniques, however, do incur a penalty in the form communication overhead between the different users. In this section, two cooperative sensing techniques are detailed.

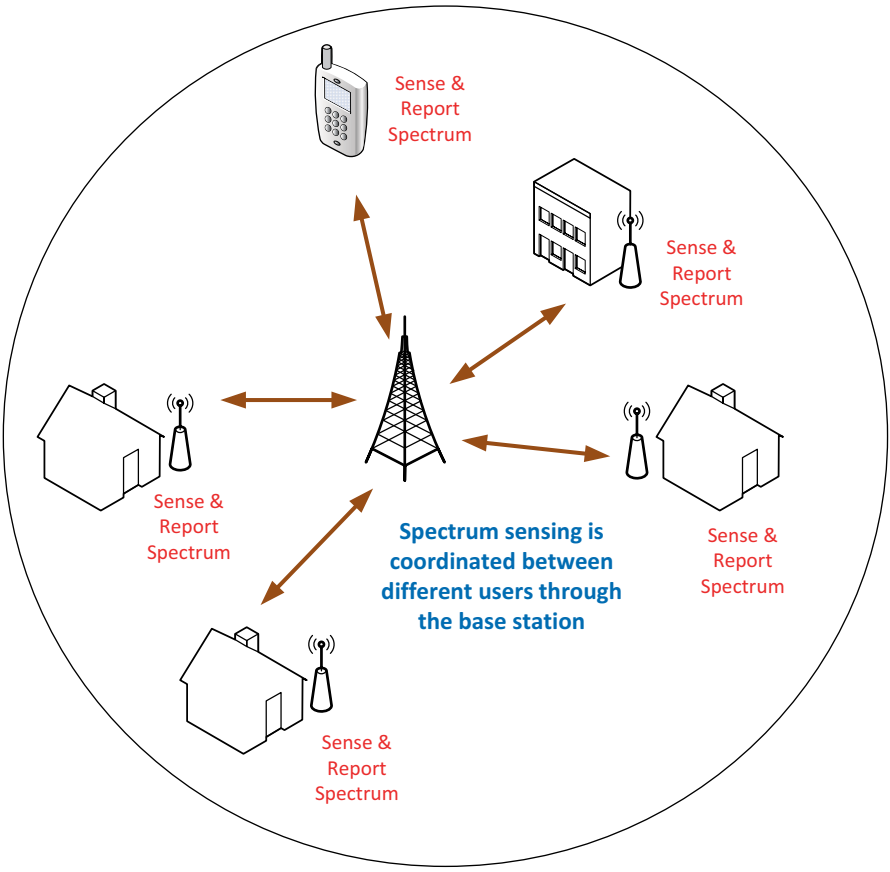


Fig. 4.12 Centralized cooperative spectrum

The first of these techniques is known as centralized cooperative spectrum sensing, illustrated in Fig. 4.12. In this illustration, the centralized element is shown to be the base station. In general, any network element can be the centralized element, not necessarily the base station. The spectrum sensing operation is a three-step process. First, the centralized element selects the channel for sensing. This is a critical and important step. The criterion used to select which elements to invoke for this task greatly affects the trade-off between the energy efficiency and performance of the spectrum sensing operation in the network as a whole. If too few elements or too many elements are selected, then this can lead to either inefficient spectrum sensing or loss of energy efficiency, respectively. Also, suboptimal choice of which elements to choose can lead to loss of performance and degraded energy efficiency. Once the network elements are chosen, the network elements sense the desired frequency channel and report their results back to the centralized network element. In order to reduce the overhead associated with cooperative spectrum sensing, the results reported back to the centralized element is usually in a compressed form,

such as spectrum statistics and a binary decision of whether the channel is used by a primary user or not. This communication is done via a dedicated control channel. The last step involved in the centralized cooperative spectrum is the centralized element processing the information from all the network elements, then making a decision of whether or not the channel is indeed available for use. This decision is then sent to the network elements via a dedicated reporting channel.

The decision-making process is a critical one. Since the results reported from all the other network nodes may, in the simplest case, be a simple binary decision, the options of processing of this information are also limited. One decision would be an AND operation, whereby all network nodes must agree that a channel does indeed have a presence of a primary user to declare that the channel unavailable for the CR. Another decision would be an "OR"-based decision, whereby, only one spectrum sensor needs to report that the primary user has been detected to declare the channel unusable by the CR. The third possible decision would be a weighted decision from all users, where a decision of at least k out of N spectrum sensors, declaring a primary user detected, would end up in a decision that a primary user is indeed using the spectrum. If the probability of false alarm (P_{FA}) and the probability of correctly detecting a primary user (P_{D}) are equal for all spectrum sensors, then the joint probability of false detection and true detection are given as:

$$P_{FA,joint}(k) = prob\{H_1 \mid H_0\} = \sum_{i=k}^{N} \binom{N}{i} p_{FA}^{i} (1 - p_{FA})^{N-i} \tag{4.16}$$

and

$$P_{D,joint}(k) = prob\{H_1 \mid H_1\} = \sum_{i=k}^{N} \binom{N}{i} p_{D}^{i} (1 - p_{D})^{N-i} \tag{4.17}$$

respectively.

The other cooperative spectrum sensing technique is known as distributed spectrum sensing, illustrated in Fig. 4.13. As its name implies, there is no centralized network element coordinating the spectrum sensing operation. In this spectrum sensing technique, each network node begins by sensing a channel independently. Once sensed, the results are then broadcast to other network elements using the dedicated control channel. The results are in the form of channel statistics or a simple binary decision of whether the channel is used or not. If differing results are reported, each network element reexamines its decision given the decision of the other network elements and the new decision is reevaluated. This process iterates until consensus is reached among all network elements.

In comparison to the centralized scheme listed above, the distributed scheme exhibits a much higher communication overhead. Moreover, since there is no centralized element coordinating the spectrum sensing operation, all network elements are involved, which entails high energy consumption overhead. On the other hand, since a global solution has been reached by consensus of all network elements, the decision is regarded as more robust than the centralized scheme.

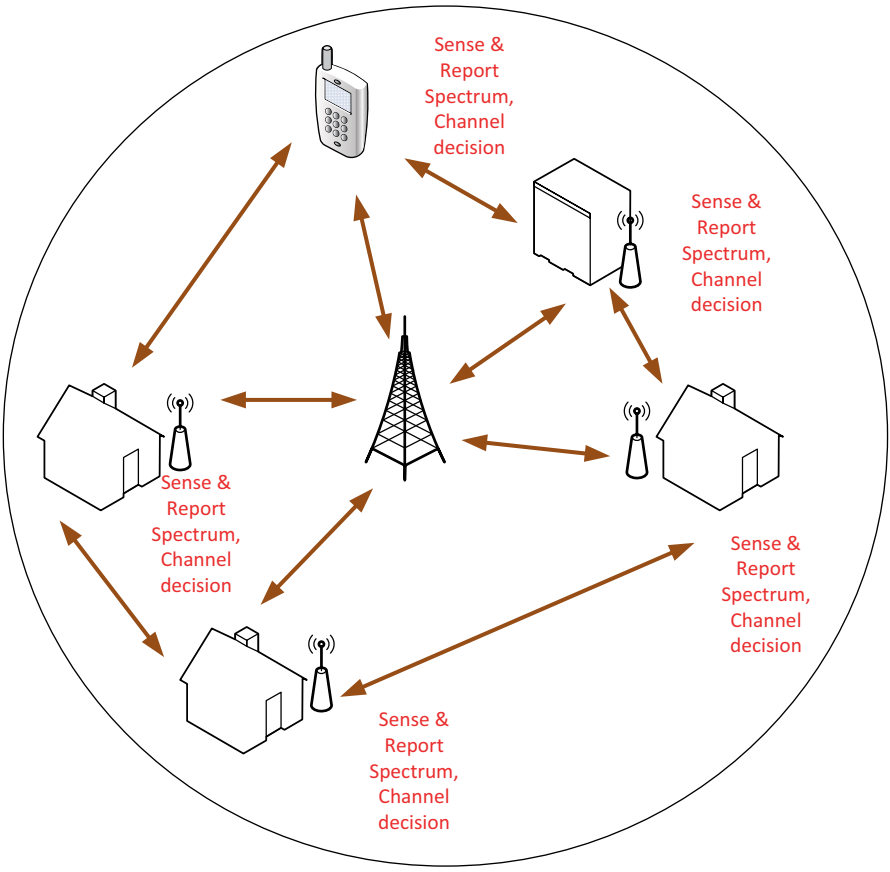


Fig. 4.13 Distributed spectrum sensing

Table 4.4 Energy efficiency of cooperative spectrum sensing techniques

Cooperative method	Cooperative gain	Cooperative overhead
Centralized	High	Low
Distributed	Higher	High

The performance of the two cooperative spectrum sensing techniques are summarized in Table 4.4. As the table shows, centralized spectrum sensing offers a low communication overhead, resulting in a higher energy efficiency. Distributed spectrum sensing, on the other hand, results in a higher accuracy determination of channel availability.

Summary

In this chapter, spectrum sensing techniques and implementations have been presented. The main challenge of spectrum sensing lies in the very low SNR detection levels that are required in order to prevent collision with a weak transmitter that is outside the reception range of the cognitive user, but not that of the primary receiver. Different spectrum sensing techniques have been reviewed. The sensitivity of energy-based detection has been found to be insufficient and must be supplemented with other techniques. Lastly, cooperative spectrum sensing techniques have been explored as a method of improving the performance and energy efficiency of the spectrum sensors.

References

1. C. Cordeiro, K. Challapali, D. Birru, and S. Shankar, "IEEE 802.22: The First Worldwide Wireless Standard based on Cognitive Radios," *DySPAN 2005*, pp. 328–337.
2. D. Cabric, S. Mishra and R. Boderson, "Implementation issues in spectrum sensing for cognitive radio," *38th Asilomar Conference on Signals, Systems and Computers*, 2004, pp. 772–776.
3. S. Shellhammer, A. Sadek and W. Zhang, "Technical challenges for cognitive radio in the TV white space spectrum," *Information Theory and Applications Workshop*, 2009, pp. 323–333.
4. A. Ghasemi and E. Sousa, "Spectrum sensing in cognitive radio networks: requirements challenges and design trade-offs," *IEEE Communications Magazine*, April 2008, pp. 32–39. equation requirement for sensitivity
5. D. Mahrof, et. al., "On the effect of spectral location of interferers on linearity requirements for wideband cognitive radio receivers," *IEEE Symposium of New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, 2010, pp. 1–9.
6. R. Balamurthi, H. Joshi, C. Nguyen, A. Sadek, S. Shelhammer, C. Shen, "A TV White Space Spectrum Sensing Prototype," IEEE Int'l Symposium on Dynamic Spectrum Access Networks (DySPAN) 2011, pp. 297–307.
7. S. Shellhammer, "Spectrum Sensing in IEEE 802.22," EURASIP, 2008, pp. 1–6.
8. A. Sahai, et. al., "Fundamental design trade-offs in cognitive radio systems," TAPAS 2006, pp. 1–9.
9. T. Yucek, H. Arslan, "A survey of spectrum sensing algorithms for cognitive radios," IEEE Journal of Communications Surveys and Tutorials, vol. 11, no. 1, 2009, pp. 116–130.
10. S. Kalmkar, A. Banerjee, A. Gupta, "SNR wall for generalized energy detection under noise uncertainty in cognitive radio," Asia-Pacific Conference on Communications (APCC), 2013, pp. 375–380.
11. D. Cabric, S. Mishra, R. Broderson, "Implementation issues in spectrum sensing for cognitive radios," *Proc. 38th Asilomar Conference Signals, Systems and Computers*, 2004, pp. 772–776.
12. L. Ma, Y. Li, A. Demir, "Matched filter assisted energy detection for sensing weak primary user signals," *IEEE Int'l Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2012, pp. 3149–3152.
13. S. Shobana, et. al., "Matched Filter Based spectrum sensing on cognitive radio of OFDM WLANs," *Int'l Journal of Engineering & Technology (IJET)*, vol. 5, no. 1, pp. 142–146.

14. Y. Cong, L. Xing, J. Bi, “Spectrum Sensing Algorithm Based on ATSC DTV Signal Structure,” *Int’l Journal on Smart Sensing and Intelligent Systems*, vol. 6, no. 5, Dec 2013, pp. 2136–2154.
15. O. Alink, et. al., “Lower the SNR wall for Energy Detection Using Cross-Correlation,” *IEEE Trans on Vehicular Technology*, vol. 60, no. 8, pp. 3748–3757.
16. Y. Hsieh, K. Wang, C. Chou, T. Hsu, T. Tsai and Y. Chen, “Quiet Period (QP) Scheduling Across Heterogeneous Dynamic Spectrum Access (DSA)-Based Systems,” *IEEE Trans. On Wireless Communications*, vol. 11, no. 8, Aug 2012, pp. 2796–2805.
17. I. Akyildiz, B. Lo, R. Balakrishnan, “Cooperative spectrum sensing in cognitive radio networks: A survey,” *Physical Communication*, Apr 2011, pp. 40–62.
18. S. Cheng, Foundation of Cognitive Radio Systems, Croatia, Intech, 2012.
19. F. Ye, H. Zhao and Y. Li, “An improved cooperative spectrum sensing in cognitive radio,” *Journal of Computational Information Systems*, vol. 9, no. 5, 2013, pp. 1719–1726.

Chapter 5
High-Linearity Wideband Transmitter

One of the main components of a cognitive radio is the radio frequency (RF) wireless transmitter. Similar to the wireless receiver, the main challenge in RF transmitters is wideband, linear operation. In this chapter, various wideband linear transmitter architectures are reviewed, including direct conversion, polar modulator, and direct digital upconversion transmitter architectures. Digital predistortion techniques to correct for analog impairments in the RF transmitter are addressed. The issue of high linearity and power efficiency in output driver amplifiers is discussed.

5.1 Requirements

There are three main specifications on the transmitter portion of a wireless transceiver. The first of these specifications is the *adjacent channel power rejection* (ACPR) [38]. The purpose of any transmitter is to transmit at a certain channel frequency with a specified bandwidth. Due to transmitter nonlinearity and noise levels, the transmitter may corrupt adjacent channels. The ACPR is a measure of how much unwanted transmission occurs on adjacent channels. This effect is shown in Fig. 5.1. The undesired tones appearing on adjacent channels are known as spectral regrowth. As the figure implies, the level of spectral regrowth is expected to decrease as you move further away from the transmit channel.

One common method in specifying ACPR is to use a spectral mask. A spectral mask is usually specified as shown in Fig. 5.2. In this case, the spectral mask requirement as specified by the IEEE 802.22 standard in the ultrahigh frequency (UHF) band (for portable devices) is compared to that of the IEEE 802.11 (Wi-Fi) spectral mask [7]. The television white space transmission (TVWS) channel bandwidth is 6 MHz. As shown, the IEEE 802.22 ACPR requirement for the adjacent channel is −55dBr (where the 'r' denotes reference level, i.e., referred to the transmitted signal). As the figure also shows, a floor requirement of −69dBr is specified. This stringent requirement is due to the fact that some channels of the UHF frequency band are used for astronomical purposes. Noise injected into these

A. Fahim, *Radio Frequency Integrated Circuit Design for Cognitive Radio Systems,*
DOI 10.1007/978-3-319-11011-0_5

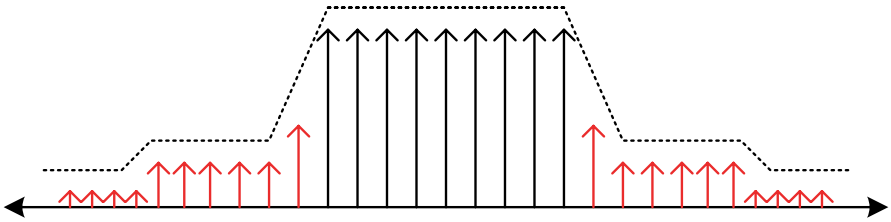


Fig. 5.1 Spectral regrowth on adjacent channels

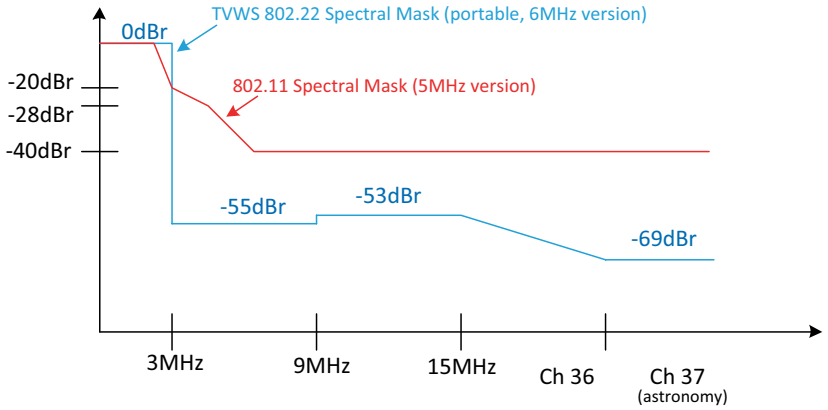


Fig. 5.2 Spectral mask specification of IEEE 802.22 (UHF band) portable devices compared to IEEE 802.11 (Wi-Fi)

frequencies from cognitive radios must be below the noise floor dictated by astronomical research facilities. For completeness, the IEEE 802.22 spectral mask for fixed devices is shown in Fig. 5.3. As the figure shows, the specifications are more stringent, especially for channels further away from the transmit channel. The main reason for this is that the output power from fixed location cognitive radios is higher than that of their portable counterparts.

The second of the transmitter specifications of a wireless transceiver is the *error vector magnitude* (EVM) [9]. The EVM is a measure used to quantify the degradation of the transmitted signal constellation by measuring how far the points in a signal constellation deviate from the ideal location. Mathematically, the EVM can be expressed in decibels by

$$EVM = 10log_{10}\left(\frac{P_{error}}{P_{reference}}\right)dB \tag{5.1}$$

or as a percentage from the ideal constellation, given by

$$EVM = \sqrt{\frac{P_{error}}{P_{reference}}} \times 100\% \tag{5.2}$$

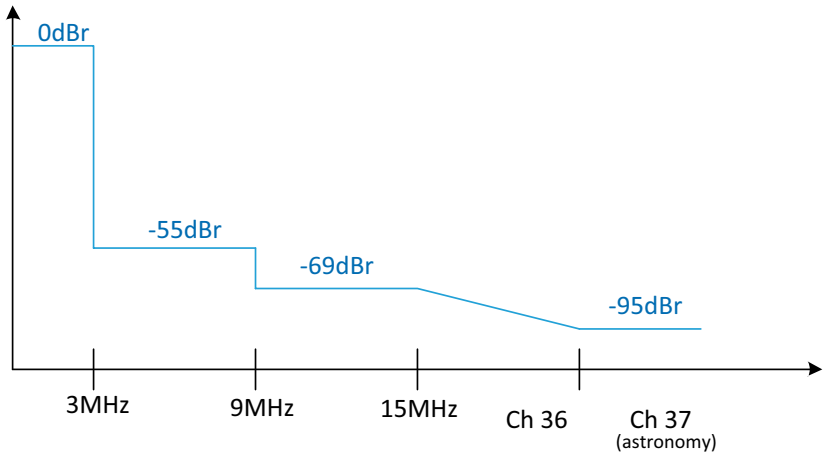
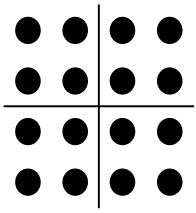


Fig. 5.3 Spectral mask specification of the IEEE 802.22 (UHF band) fixed devices

Fig. 5.4 Signal constellation of a 16-QAM signal



where P_{error} is the root-mean-square (rms) power of the error vector and $P_{reference}$ is the reference constellation average power. An example of a signal constellation is shown in Fig. 5.4, which is a signal constellation of a 16-quadrature amplitude modulation (QAM) signal. The two-dimensional plane is the complex plane. In the presence of signal degradation, the location of the black circles (each representing a data codeword, or symbol) moves in a random fashion. This means that there will be less effective distance between the symbols in the signal constellation.

There are several sources for degradation in the signal constellation. One such degradation is due to the phase noise in the local oscillator (LO) signal, which, as we will see in the following sections, is used to upconvert a baseband signal to the desired RF channel frequency. This source of degradation manifests itself as random rotation of the entire signal constellation.

Another source of degradation is caused by LO feed through in the upconversion mixers. This type of degradation would cause a DC shift (either vertically or horizontally) in the signal constellation. This is due to the DC term produced at the output of the mixer as a result of mixing the LO with a component of the LO that is fed into the signal port of the mixer. This DC shift can be time-varying and dependent on the input data signal.

Other sources of signal constellation degradation can be due to lack of sufficient image rejection and insufficient linearity in the transmitter. Lack of sufficient image

Table 5.1 IEEE 802.22 specification on maximum EIRP under various conditions

Device class	Mobility	EIRP (dBm)	Sensing (dBm)	Transmitallowed on adjacent TV channels (dBm)
Fixed	No	+36	N/A	No
Portable only	Yes	+20	N/A	Yes, but <+16
Sensing only	Yes	+17	−114	Yes, but <+16

rejection would create a "shadow" signal that smears the desired signal. Lack of sufficient linearity would cause intermodulation terms to rise up and distort the signal itself. This is especially problematic in the case of an orthogonal frequency-division multiplexing (OFDM) signal, where the signal is composed of several closely spaced modulated tones. The distortion terms of two of these tones, for example, can intermodulate to produce IM2 and IM3 terms that corrupt two other tones within the OFDM signal itself.

Another important specification is the *equivalent isotropically radiated power* (EIRP) [54]. This metric quantifies the amount of power that is radiated at the output of the antenna of a wireless transmitter device. Mathematically, the EIRP is quantified as

$$EIRP = 10log_{10}\left(\frac{P_T \cdot G_a}{P_L}\right)(dBm) \tag{5.3}$$

where P_T is the output power of power amplifier (PA; in mW), G_a is the power gain of the antenna, and P_L is the power loss in the path between the PA and the antenna. The EIRP is usually expressed in the decibel scale with units of dBm. Table 5.1 lists the maximum EIRP allowed under the IEEE 802.22 standard for various conditions and type of devices.

5.2 Direct UpconversionTransmitter

There are several techniques that can be used to convert the baseband digital signal into a modulated RF signal. One such technique that is popular for low-power applications is the direct-conversion transmitter, shown in Fig. 5.5. The digital baseband signal is split into I and Q channels. This is done since most commercial communication standards rely on quadrature signaling schemes [7]. Two identical digital-to-analog converters (DACs) are used to convert the pair of digital signals to analog form. The DACs are followed by a pair of low-pass anti-alias filters, which are necessary to band limit the signal and avoid alias terms. The band-limited analog signals are then frequency translated into an RF signal using a pair of upconversion mixers clocked by quadrature LO signals. The upconverted RF signal is given as

$$RF_{out} = BB_I \bullet LO_I + BB_Q \bullet LO_Q. \tag{5.4}$$

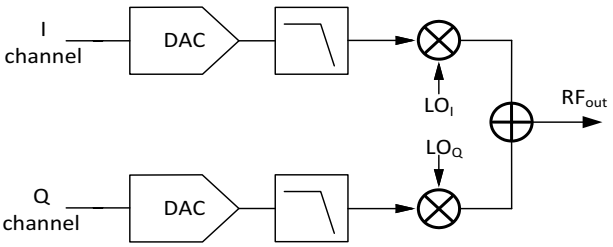


Fig. 5.5 Direct upconversion transmitter

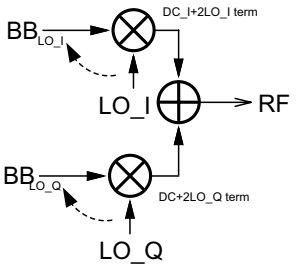


Fig. 5.6 LO leakage in upconversion mixer

The main advantage of this topology is low power consumption and low area. Its main disadvantages are a result of the nonidealities between the two quadrature paths. The first of these nonidealities is the random mismatches between the I and Q channel. The mismatches between the DACs can result in an amplitude mismatch, $\frac{\Delta A}{A}$. The quadrature LO signals are assumed to be exactly 90° apart. Mismatches between the quadrature LO signals can result in phase deviation away from 90°, given as $\frac{\Delta\theta}{\theta}$. If the mismatches between I and Q channels are referred to the Q channel, Eq. (5.4) can be rewritten as

$$RF_{out} = BB_I \bullet LO_I + BB_Q\left(1+\frac{\Delta A}{A}\right)\bullet LO_Q\left(1+\frac{\Delta\theta}{\theta}\right), \tag{5.5}$$

which can be rewritten as

$$RF_{out} = BB_I \bullet LO_I + BB_Q \bullet LO_Q\left(1+\frac{\Delta A}{A}+\frac{\Delta\theta}{\theta}+\frac{\Delta A}{A}\bullet\frac{\Delta\theta}{\theta}\right). \tag{5.6}$$

The term in parenthesis causes EVM distortion.

Another effect that degrades the performance of direct up conversion transmitters is LO leakage into the baseband signal port. Due to the finite isolation between the LO port and the baseband signal port in the upconversion mixer, a small component of the LO port will appear in the baseband signal port, as shown in Fig. 5.6. This causes the LO signal to mix with itself. This causes a DC shift in the output RF signal constellation. The DC shift is in the horizontal or vertical direction, for the I and Q channel self-mixing, respectively.

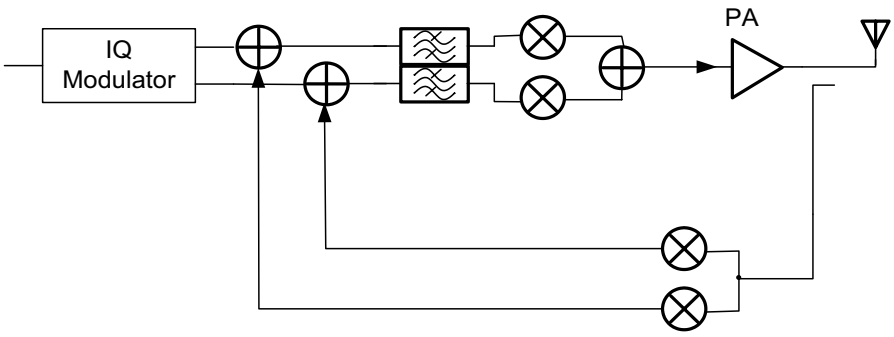


Fig. 5.7 Cartesian loop transmitter

Another nonideal effect is the noise upconversion. Since the baseband signal is centered around DC, the low-frequency flicker noise is upconverted around the center of the channel. This can be minimized through careful optimization of the noise in the baseband circuitry as well as having symmetric LO signals [15]. Also, the noise of the mixer and PA following the baseband filters must satisfy the spectral mask requirement of the standard being implemented. If the far-out noise requirement cannot be met, then a band-pass filter (BPF) around the RF signal may be required, increasing the system complexity.

5.3 Cartesian Loop Transmitter

One way to meet the spectral mask requirement is to insert the transmitter in a feedback loop. One such feedback technique is the Cartesian loop transmitter [20], shown in Fig. 5.7. The basic idea of a Cartesian loop is to use the concept of feedback in order to linearize the transmitter. This is accomplished by sending a fraction of the output power of the PA back into a pair of quadrature downconverters (the LO signals are not shown in Fig. 5.7). The downconverted signals are now subtracted from the baseband I and Q signals. Mathematically, the RF output signal going into the antenna becomes

$$RF_{out} = \frac{BB_I LO_I + BB_Q LO_Q}{1 + A_F (BB_I LO_I + BB_Q LO_Q)} \tag{5.7}$$

where A_F is a gain factor inserted into the feedback loop. It should be stated that in any feedback system, it is assumed that the nonlinearity is caused by the forward path transfer function. In this context of the RF transmitter, this is given by the numerator of Eq. (5.7). The linearity of the feedback path is assumed to be much higher than that of the open loop transmitter.

Another important feature to mention is that Eq. (5.7) does not explicitly take into account the frequency translation of the transmitter. When doing so, the low-pass

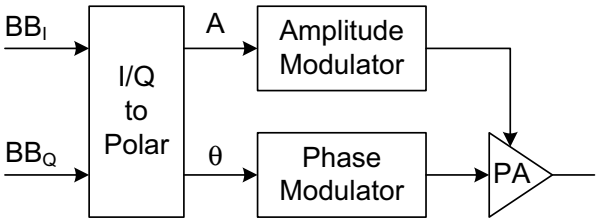


Fig. 5.8 Conventional polar modulator transmitter

filter (LPF) immediately following the subtractor is translated into a BPF at the output of the PA by the upconversion mixers. This means that there is an inherent BPF centered around the transmit channel. This greatly assists in meeting the spectral mask requirement of the transmitter. The stopband attenuation of the filter is determined by the open loop gain of the transmitter.

The Cartesian feedback loop transmitter is not without its disadvantages. The most obvious is that of stability. Since this is a feedback system that may contain several poles, guaranteeing stability without the introduction of zeros or lowering the open loop gain is difficult. Lowering the open loop gain both reduces the stopband attenuation of the BPF response and degrades the linearity of the transmitter. Another issue with the feedback structure is that the signal bandwidth is now limited to the bandwidth of the Cartesian feedback loop, making it difficult to be used for channel bonding or channel aggregation (see Chap. 2).

As shown in Fig. 5.7, the feedback structure consists of quadrature downconverters as well as quadrature analog subtractors to close the feedback loop. It can rarely be guaranteed that the linearity of such components is sufficiently below that of the open loop transmitter without burning an excessive amount of current. For this reason, the complexity (area) and power consumption of a conventional Cartesian feedback loop transmitters are considered to be excessive for low-power portable applications.

5.4 Polar Modulator Transmitter

Another popular transmitter topology that has been used for RF wireless applications is polar modulator transmitter [29, 33, 41]. A polar modulator aims to increase the efficiency of a PA. As will be seen in Sect. 5.8, the PA power efficiency can be increased by saturating its input, i.e., operating it in a nonlinear fashion. This allows the use of highly nonlinear PAs, which are far more power efficient. A conventional polar modulator transmitter is shown in Fig. 5.8. Unlike the other transmitters shown earlier where I and Q baseband signals were used, the polar modulator splits the baseband modulated signal into its phase and amplitude components. This can be easily derived from the I and Q signals by the following pair of equations:

$$A_{polar} = \sqrt{BB_I^2 + BB_Q^2} \tag{5.8}$$

$$\theta_{polar} = \tan^{-1}\frac{BB_Q}{BB_I} \tag{5.9}$$

The phase component can be upconverted by combining the I and Q signals into a real signal, then passing the signal through a limiter. A limiter is a device that extracts the phase information of a signal by maintaining the zero crossings, but saturating the amplitude to either the minimum or maximum supply voltage. The phase information from both I and Q channels are then upconverted to RF using a pair of mixers, similar to the direct upconversion transmitter. The amplitude path is first filtered, then used to modulate the envelope of the PA output by modulating its power supply directly. The LPF along the amplitude path may be necessary to band limit the signal and to meet the spectral mask specification of the transmitter.

Although the polar modulator presented does help greatly in reducing the overall power consumption of the wireless RF transmitter by allowing the use of a power efficient PA, it does present some challenges [19]. The first challenge is in balancing the phase relationship between the amplitude and phase paths. When the signal is recombined into a real signal from its amplitude and phase components, the components must undergo the same phase shift. If not, it introduces a noticeable error in the output spectrum and performance is significantly degraded. To understand why this is the case, consider a signal with amplitude A and phase θ. If these components are separated, as is the case with a polar modulator, then the phase component is added to a time-delayed component of the amplitude, and the resulting signal can be expressed as

$$s(t) = A(t-\tau)\bullet[\cos\theta(t) + j\sin\theta(t)] \tag{5.10}$$

where τ is the difference in the propagation delay between the amplitude and the phase paths. If τ is as large as one symbol duration, the amplitude of one symbol data would be added to the phase of next phase symbol, resulting in a large error in the output signal.

Another nonideality that must be dealt with for a polar modulator transmitter is the gain and phase variation resulting from process, voltage, and temperature (PVT) variation. Namely, the phase path can experience phase variation (PM–PM) as well as gain variation translated into phase (AM–PM) noise. The analog path can result in AM–AM and AM–PM noise. This can be compensated for by predistorting the digital input signal, as will be detailed in Sect. 5.7. This is a one-time calibration operation during production test in addition to power supply and temperature sensors to allow the distortion terms to track the amplitude and phase variation due to voltage and temperature, respectively.

There are other possible implementations for the polar modulator transmitters. In one such implementation, the phase information can be upconverted to an RF frequency by means of a phase-locked loop (PLL), as shown in Fig. 5.9. After splitting the signal into an amplitude and phase component, the phase information is

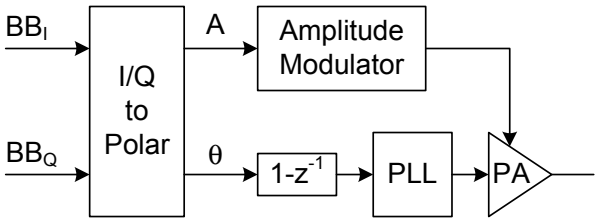


Fig. 5.9 Polar modulator with PLL for phase upconversion

translated into frequency information through a digital differentiator, $1-z^{-1}$. This frequency information is then fed to the PLL's feedback division ratio, which is a digital word representing the desired output frequency. This frequency information is then translated back into phase and upconverted to the desired RF frequency by the PLL's linear-phase response. The advantage of this technique is that the pair of DACs and quadrature upconversion mixers associated with the phase path of the polar modulator can be eliminated. Another advantage of this approach is that the phase upconversion has an inherent BPF around it, which assists in meeting the spectral mask requirement of the system. The disadvantage of this approach is that the PLL has to be well calibrated precisely to obtain a constant digital-to-phase transfer function across all operating conditions. The PLL is an analog system and its parameters will vary with PVT variations. Another disadvantage of this approach is that the data bandwidth is limited by the PLL bandwidth. Frequency preemphasis can be used to enhance the bandwidth, but this requires precise matching of the digital preemphasis filter to the analog filter response of the PLL, which is difficult in practice [4, 32].

Another method of phase upconversion is through the use of an offset PLL [29], shown in Fig. 5.10. The phase component of the baseband signal is extracted after upsampling it to an intermediate frequency (IF) passing the signal through a limiter, as shown in Fig. 5.10. The phase information of the baseband signal is then compared to a constant phase signal, derived from the PLL. The offset PLL upconverts the phase portion of the signal from IF to RF. The choice of the IF and RF frequency can be chosen by the divider ratios M and N. More specifically,

$$f_{RF} = f_{PLL}\left(\frac{N+1}{M \bullet N}\right) \tag{5.11}$$

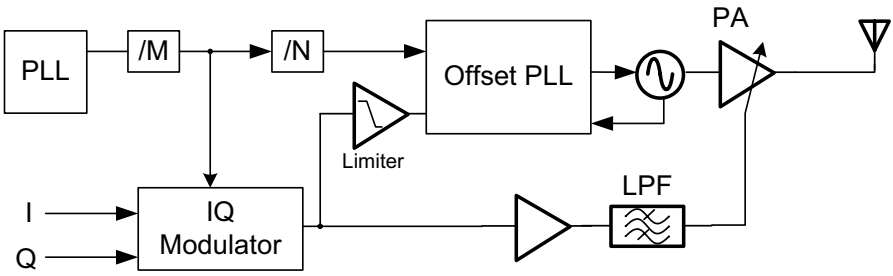


Fig. 5.10 Polar modulator using an offset PLL

and

$$f_{IF} = \frac{f_{PLL}}{M \bullet N} \tag{5.12}$$

One major advantage of this technique, in comparison to the PLL technique mentioned earlier, is that it does not suffer from the 20logN degradation of noise. This greatly enhances the phase noise floor at the desired output RF frequency, which reduces the overall EVM. One disadvantage of the offset PLL technique is that it requires two voltage-controlled oscillators (VCOs)—one for the PLL and another for the offset PLL. This complicates the frequency planning in such a way that one must avoid the beating of the two PLL output frequencies, and their divided down frequencies (by M and N) in such a way that it corrupts the output RF frequency or causes spurious content on other user channels.

The details of the offset PLL are shown in Fig. 5.11. As shown, the offset PLL is similar to a conventional PLL, except that the feedback divider is substituted with a mixer to translate the RF frequency into an IF frequency. A high-speed phase detector can be used, such as an XOR gate. The choice of the IF frequency is dictated by the frequency planning in order to avoid unwanted spurs as a direct result of coupling between the offset PLL and the main PLL. The F_{in} input is the modulated data upconverted into an IF frequency.

The PLL used for phase upconversion can be digitized as shown in Fig. 5.12 [10, 55]. In this variation, the PLL is replaced with an all-digital PLL (ADPLL), thereby easing calibration of the PLL. The VCO is shown as a separate block and replaced by a digitally controlled oscillator (DCO). The digital amplitude signal is upsampled RF directly, as will be shown in the Sect. 5.5. The phase modulator path can be replaced with an upsampler followed by a calibrated delay chain [34]. Since the phase path still relies on a PLL, the phase modulation path benefits from an implicit BPF.

One last important variation of a polar modulator is the closed-loop polar modulator [44], shown in Fig. 5.13. This particular closed-loop polar modulator uses an offset PLL in order to translate the phase information up to the transmit RF

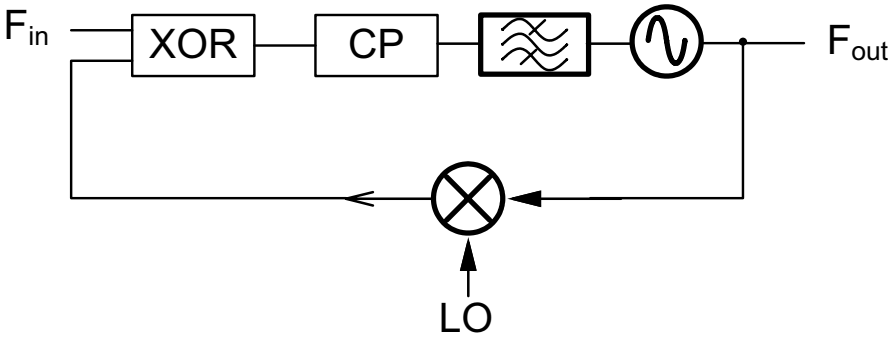


Fig. 5.11 An offset PLL used in a polar modulator transmitter

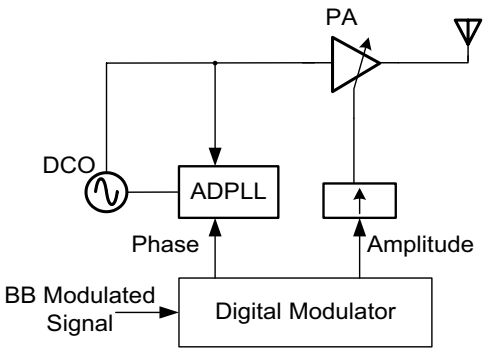


Fig. 5.12 All-digital polar modulator transmitter

frequency. In general, any type of phase upconversion can be used. Just as in a Cartesian feedback loop transmitter, a fraction of the PA signal is fed back through the use of a downconversion mixer. The dowconverter mixer translates the RF signal into an IF frequency. Both phase and amplitude information are extracted from the IF feedback signal and are subtracted from the input IF phase and amplitude signals, respectively. The phase information is fed as input to the offset PLL and the amplitude information is fed back into the amplitude control loop. An automatic gain control (AGC) loop is shown here, which enables control of the gain of both the forward and feedback paths of the amplitude loop. In essence, a closed-loop polar modulator is a hybrid between a Cartesian feedback loop and a polar modulator transmitter. The advantage of this technique is that it combines the linearity advantages of a Cartesian feedback loop transmitter with the power efficiency of the closed-loop polar modulator. The added linearity comes from the fact that the entire transmitter is connected in a closed-loop feedback system. Since the PA is part of the closed-loop response, the PA is also linearized in this system. The power efficiency comes from the fact that the input to the PA is from the phase path, which

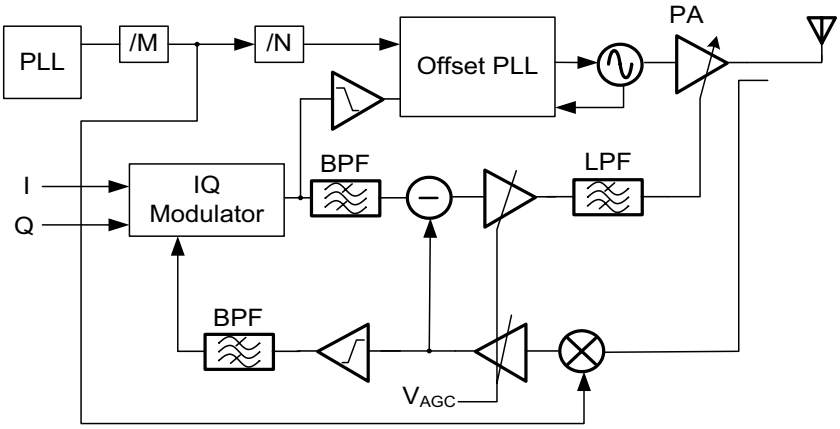


Fig. 5.13 Closed-loop polar modulator

hard switches the PA, thereby increasing its power efficiency (as will be seen in Sect. 5.8). As with other polar modulators, the amplitude path controls the power supply of the PA, thereby enabling envelope modulation of the upconverted phase signal.

Although the closed loop nature of both the amplitude and phase loops is able to track PVT variations, the overhead needed for this closed loop response is considered to be very high for low-power handheld applications. Moreover, the calibration technology in open-loop polar modulators has significantly advanced to the point that its performance is close to that of a well-optimized closed-loop modulator.

5.5 Direct Digital Upconverter Transmitter

As technology scales, the performance and bandwidth of DACs improve. Figure 5.14 shows a survey of some recent high-performance DAC designs [2, 3, 8, 25, 46, 48, 49]. As the figure shows, the bandwidth of modern DAC designs has exceeded the 1 GHz mark, enabling the digitization of the transmitter up to the required RF output frequency. The resolution (number of bits) of the DACs has also improved dramatically. Since the resolution of the DAC determines the quantization noise floor, it is important that this level be below the ACPR required by the standard implemented.

There are many methods to digitize the RF transmitter design. One such method is illustrated in Fig. 5.15 [6]. The sampling rate of the digital baseband signal is upconverted to a higher frequency by a process known as upsampling, which will be detailed in the following paragraphs. It is important to note that the upsampling process does not frequently translate the baseband signal. It merely increases the sampling rate. This has the effect of spacing out the digital alias terms further out, easing the requirements of the analog anti-alias filter following the DAC [40]. Once upsampled, the digital data are converted into an analog signal by a pair of high-speed DACs operating at the desired RF frequency. The outputs of the DACs are then summed and sent to the PA. An RF BPF may be required to suppress the alias terms from the DAC output. Such high-speed DACs are often referred to as RFDACs in the literature [18, 23]. Unfortunately, this term is used loosely in the

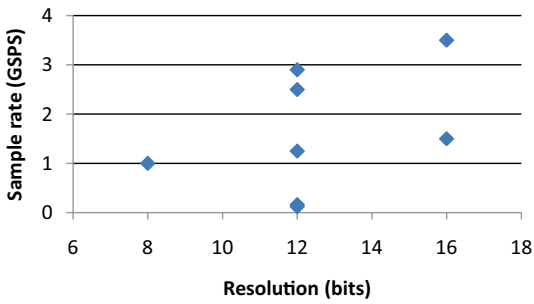


Fig. 5.14 Survey of recent high-performance DAC designs

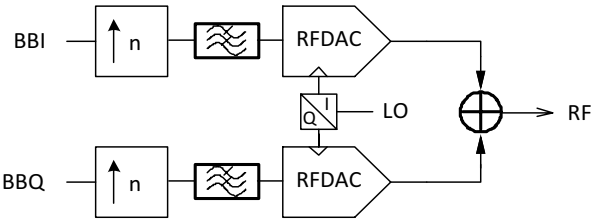


Fig. 5.15 Digital upconverter transmitter architecture

literature to refer to two different structures. These structures are reviewed and differentiated in the following paragraphs.

A high-speed RFDAC is used to simultaneously convert the pair of quadrature digital signals into an analog waveform and upconvert the signal to the desired RF frequency. When used in this form, the RFDAC is implemented by a structure shown in Fig. 5.16. As shown in the figure, the DAC is merged with a Gilbert cell mixer. To differentiate from another form of an RFDAC shown later, it is referred to here as RFDAC1. It is sometimes referred to in the literature as a mixer-DAC, which is more descriptive. The load Z_L shown above is usually a resistive load, but it can also be an inductive load, or an RF choke. A resistive load is easily integrated onchip, but can require a significant amount of headroom, reducing the dynamic range of the RFDAC1. An RF choke, on the other hand, has a DC impedance of nearly 0 ohms, and hence allows for the maximum possible headroom. Disadvantages of an RF choke include requiring an external load and reliability concerns. An external load not only increases cost but also limits the bandwidth of the output node of the RFDAC1. Reliability concerns can develop if the output swing of the RFDAC1 goes above the supply voltage by an excessive amount (since the DC nominal bias of the output node is already at the supply voltage).

Another form of digitizing the transmit path is to use a digital upconverter (DUC), shown in Fig. 5.17 [23]. In this case, the frequency translation is performed in the digital domain. There are several advantages to this. First, only one RFDAC

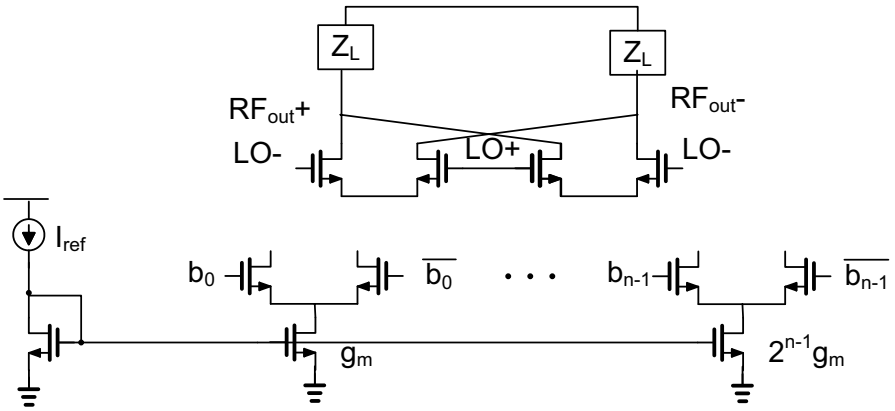


Fig. 5.16 Circuit-level implementation of the RFDAC1 structure

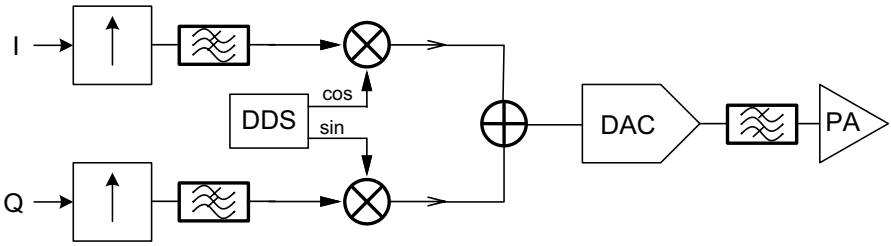


Fig. 5.17 Digital upconverter (DUC) with digital quadrature upconverter

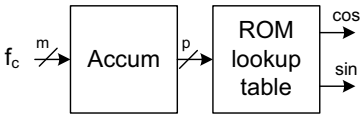


Fig. 5.18 Direct digital synthesizer (DDS)

is now required. Unlike RFDAC1, shown in Fig. 5.16, this RFDAC does not require a built-in mixer operation, alleviating much needed headroom. Instead, a digital pair of quadrature mixers (i.e., multipliers) is used. The LO is substituted with a direct digital synthesizer (DDS), shown in Fig. 5.18. The direct digital synthesizer produces a digital word representing cosine and sine functions given a desired output frequency word. In order to be able to synthesize any desired output frequency within the band of interest, the DDS must be clocked at least twice the maximum required output frequency. This is done in order to satisfy the Nyquist criterion in avoiding smearing caused by digital alias term. In reality, a higher oversampling ratio is usually required. This type of RFDAC will be referred to as RFDAC2.

One common choice of upsampled frequency, IF, of the DUC, shown in Fig. 5.17, is to choose the IF frequency to be one fourth that of the required RF frequency. In this configuration, the output of the DDS is now reduced to $\cos\left(\frac{\pi}{2}k\right)$ and $\sin\left(\frac{\pi}{2}k\right)$, where k is an integer. The output of the DDS can be reduced to two levels by phase shifting both the sine and cosine terms by $\frac{\pi}{4}$. In this case, the output of the DDS is consistently either $\frac{1}{\sqrt{2}}$ or $-\frac{1}{\sqrt{2}}$. This means that the normalized output of the DDS for the sine and cosine functions can be $\{+1, +1, -1, -1,\ldots\}$ and $\{+1, -1, -1, +1,\ldots\}$, respectively. This allows for significant reduction of hardware by eliminating the need for a pair of digital multipliers and a full DDS implementation. One disadvantage of this approach is that it can potentially downconvert third-order harmonic distortion terms down to the desired output channel. This third-order distortion harmonic can potentially mix with the $4f_{IF}$ term and downconverted directly to f_{IF}, thereby corrupting the output signal. One potential advantage that a DUC offers, which is missed by this simplification in the DDS output signal, is that of implicit harmonic rejection. The output spectrum of a pure sine wave is a single pair of

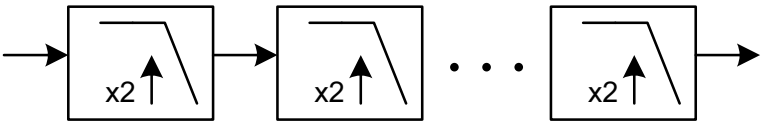


Fig. 5.19 Typical upsampler used in an RF transmitter

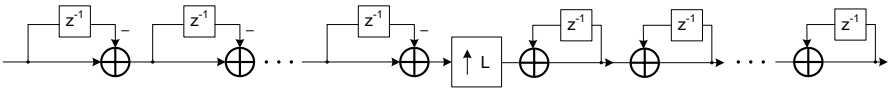


Fig. 5.20 CIC filter for upsampler

tones, one at f_c and another at $-f_c$. When the DDS approximates this sine wave by using only two levels, it approximates the sine wave by a square wave, which, as was shown in Chap. 3, contains significant odd-order harmonics. This means that any tones at near higher order harmonics may be potentially downconverted to the output channel, similar to what was seen in Chap. 3.

One critical component in digitizing the RF transmitter is the upsampler. A typical upsampler used in an RF transmitter is shown in Fig. 5.19. The interpolation function is a two-step process. In the first step, a data sample is followed by L−1 zeros, where L is the integer upsampling factor. Second, an LPF is used to smooth out the zeros which were placed between the data samples. Inserting zeros, as opposed to holding the previous value, has two major advantages. First, holding the previous value implies a zero-order hold (ZOH) operation, which introduces an undesired droop to the data signal, due to the implicit sinc filtering associated with the ZOH operation. Second, having nonzero values, instead of zero padding between data samples, increases the hardware complexity of the interpolator.

The reduction in hardware complexity by zero padding can be seen by analyzing the filtering operation of the LPF. In general, a finite impulse response (FIR) filter operation can be given by

$$y[r+nL] = \sum_{k=0}^{M} x[n-k] \cdot h[r+kL] \tag{5.13}$$

where r is an integer, L is the interpolation factor, h[·] is the LPF, and x[·] is the input data signal. If only every Lth data sample is nonzero, then the filtering hardware can be reduced by a factor of approximately L. A common implementation for h[·] is a half-band filter [50]. A half-band filter has the characteristic that every other filter coefficient is zero, leading to a very efficient filter implementation. This comes at the expense of poor aliasing performance near the Nyquist frequency. To address this, an interpolation factor of 2 is associated with every half-band filter segment.

Another popular implementation of the LPF is a cascaded integrator-comb (CIC) filter [51]. The structure of the CIC filter is shown in Fig. 5.20. The filter consists of

N differentiator segments in cascade, followed by the upsampler and then followed by N integrators. The overall resulting transfer function is

$$H(z) = \left[\sum_{k=0}^{L-1} z^{-k}\right]^N = \left(\frac{1-z^{-L}}{1-z^{-1}}\right)^N \tag{5.14}$$

where N is the number of differentiator and integrator stages. Equation (5.14) implements a sinc filter, an LPF. The order of the sinc is determined by the number of stages, N.

One important advantage of DUC is that multiple carrier frequencies can be supported simultaneously [43, 53]. This is accomplished by having a bank of direct digital synthesizers and digital quadrature upconverters. The output of all the digital quadrature upconverters is summed digitally. The final digital word that represented the sum of the different channels is then converted to analog form by a single RFDAC2. This is an important feature since supporting channel bonding and channel aggregation is now straightforward with a DUC. Another advantage of the RFDAC2 approach is that the same DUC can be used as for multiple wireless standards by simple digital manipulation of the digital upsamplers, filters, and digital quadrature mixers [27, 57].

5.6 RF Digital-to-Analog Converter

Since a high-speed DAC is the central component of a digital RF transmitter, this section concentrates on the design details and trade-offs of such DACs. In order to quantify the quality of the DAC design, certain standard metrics are used. One such metric is the signal-to-noise ratio (SNR). The SNR is defined as the ratio of the rms value of the desired signal, P_s, in comparison to all other spectral components integrated over a band of interest [24]. The band of interest is usually defined as the Nyquist frequency. The DC term is excluded from this integration band.

Another metric, which is closely related to SNR, is the spurious free dynamic range (SFDR). The SFDR is defined as the ratio of the rms value of the desired signal, P_s, to the second highest spurious signal within a frequency band defined by the Nyquist rate. The tones considered part of this test can include spurious signals that are injected from external environments as well as harmonic and intermodulation distortion caused by the DAC itself. The DC term is excluded from this spurious tone search.

The linearity of the DAC is also important. Since most DAC designs are differential, the second-order distortion term is much smaller than third-order distortion terms. Higher-order distortion terms decay rapidly. For this reason, the third-order intermodulation distortion (IM3) is an important metric in DAC design.

Static linearity of the DAC is also important. There are two main metrics: the integral nonlinearity (INL) and the differential nonlinearity (DNL). The INL is de-

fined as the maximum deviation of the output of the DAC from an ideal straightline as the codewords are incremented sequentially. Since the minimum and maximum values of the DAC are considered to be the two end points of the ideal straight line representing the range of the DAC, the worst-case deviation usually occurs near the mid-codeword. This means that INL is worst near the mid-codeword. This metric is usually reported in units of LSB, i.e., the output voltage or current is normalized to the ideal size of an LSB voltage or current.

DNL is defined as the ratio of the variation in step size between two adjacent codewords to the ideal step size of the DAC. In essence, it is a measure of the change in slope between adjacent codewords. This is usually worst when there is a structural change as the next most significant bit (MSB) is toggled. For example, if a different structure is used to implement the fourth bit in a DAC than the first three least significant bits (LSBs), the worst-case DNL may be between codewords 0111 and 1000. The DNL metric is usually reported in units of LSB, i.e., normalized to the ideal size of the LSB.

5.6.1 DC Matching Requirements

The first requirement of a DAC is to ensure that low-frequency matching requirements are satisfied. This usually results from random and systematic mismatches in the elements used to implement the DAC. Random mismatches can be a result of device matching between the different elements. If a current steering type of DAC is employed, for example, then the current matching is given by [36]

$$\left(\frac{\Delta I}{I}\right)^2 = \frac{4A_{VT}^2}{V_{GS}-V_T}\cdot\frac{1}{WL}+\frac{A_\beta^2}{WL} \tag{5.15}$$

where $\frac{\Delta I}{I}$ is the current mismatch between different current mirror elements in the DAC, V_{GS}–V_T is the overdrive voltage of the current mirror, WL is the area of the current mirror device (width and length), A_{VT} is a process-dependent coefficient describing the V_T variation, and A. is a process-dependent coefficient describing the random area variation. Clearly, Eq. (5.15) demonstrates that the best way to reduce the current mismatch is to use larger devices and larger overdrive voltages. There are other sources of random mismatches, such as the variation of the value of the load device. In practice, however, the dominant source of random static mismatch is usually current mismatch due to the V_T variation of the current mirror devices.

For modern deep submicron technologies, Eq. (5.15) is accurate only for devices with large values of L. In FinFET technologies, for example, the choice in the value of L may be limited [47]. The saturation current in a deep submicron technology is given by

$$I_{ds} = WC_{OX}(V_{GS}-V_T)v_{sat} \tag{5.16}$$

where v_{sat} is the saturation velocity, which is equal to 10^7 cm/s. The variation in current is then given by

$$\frac{\Delta I}{I} = \frac{\Delta(C_{OX} v_{sat})}{C_{OX} v_{sat}} + \frac{\Delta W}{W} - \frac{\Delta V_T}{V_{GS} - V_T} \tag{5.17}$$

The specification of the mismatch in the current source determines the maximum resolution, in bits, that the DAC can support. More specifically, if a current-based N-bit DAC is implemented, the variance of the current due to random mismatches of the jth element in the DAC is given as [45]:

$$\sigma_j(I^2) = \left(\frac{j \cdot \left(1 - \frac{j}{2^N}\right)}{2^{2N}} \right) \left(\frac{\Delta I}{I} \right)^2 I_{REF} \tag{5.18}$$

where I_{REF} is the size of an LSB current and $\frac{\Delta I}{I}$ is given by Eq. (5.15) or (5.17). This means that the maximum INL (derived from the worse-case accumulated current variance of Eq. (5.18)) that can be tolerated in the DAC is given by [5]

$$INL_{MAX} = \sqrt{2^{N-1}} \left(\frac{\Delta I}{I} \right) LSB \tag{5.19}$$

One might believe that increasing the resolution of the DAC by a factor of 2 implies better matching by a factor of 2 as well; however, Eq. (5.19) implies that if the required resolution of the DAC increases by 2 bits (4× resolution), the current matching only needs to improve by 2×. Here, I_{REF} is assumed to be equal to one LSB. This result is true if the mismatch between the DAC cells is truly random, i.e., uncorrelated. If there is correlation in the mismatch between the different cells, such as a static offset present in only one row of cells of the DAC array, then this error is correlated error. Correlated error results in worse INL error than predicted by Eq. (5.19).

The matching between cells in a multibit DAC is very layout dependent. In order to enhance the matching between the different components, the DAC is composed of identical cells that are laid out symmetrically, as shown in Fig. 5.21. As the figure shows, the cells of the DAC occupy a symmetric two-dimensional area. Dummy devices are inserted around all four sides of the two-dimensional structure. A thermometer decoder is used in order to activate the individual cells sequentially.

Although the layout in Fig. 5.21 is symmetric, a few nonidealities can arise. Consider a finite resistance on the power supply and ground lines powering the DAC cells. If the main power and ground connections are from the top left-hand corner, for example, a gradient voltage would result along the power and ground lines, with the worst gradient being the cell furthest away from the main power and ground connections. Figure 5.22 shows the effect of this gradient on the linearity of

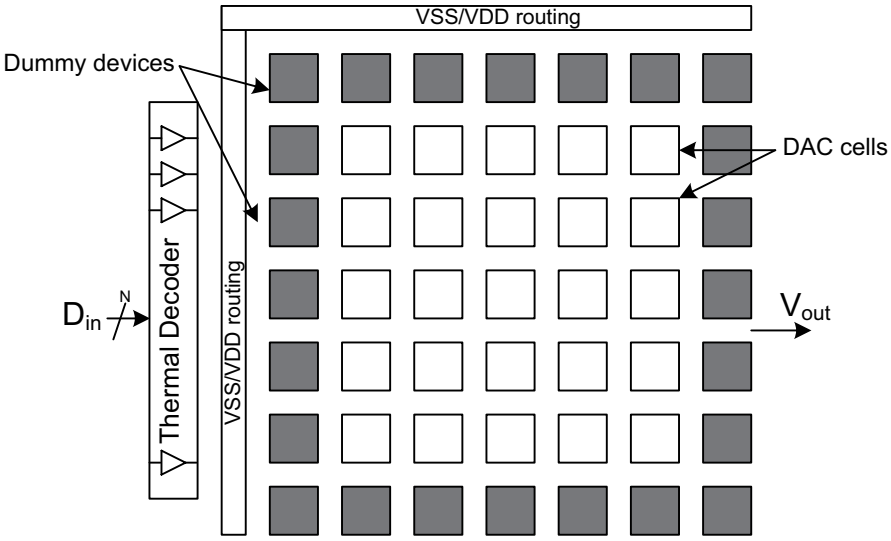


Fig. 5.21 Symmetric layout of cells implementing a DAC

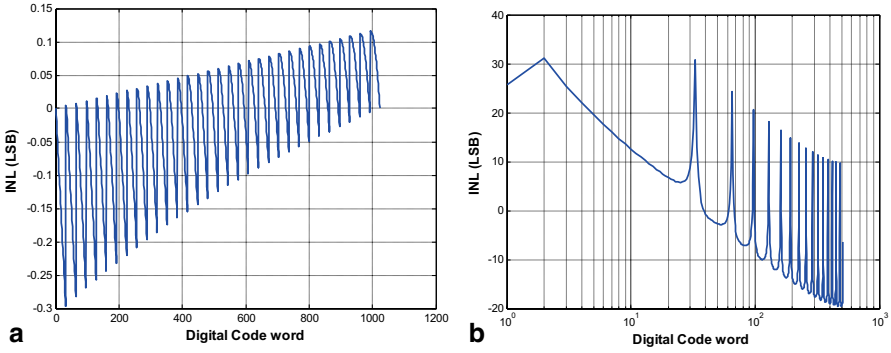


Fig. 5.22 Effect of power and ground gradients on DAC **a** INL **b** and FFT of INL

a 10-bit DAC arranged as a 32×32 array of identical cells. Figure 5.22a shows the INL degradation as the digital codeword is incremented. It is assumed that the cell associated with the minimum codeword is closest to the main power and ground connection, and the cell associated with the maximum code word is farthest from the main power and ground connections. The periodic nature of the INL distortion is due to the change in supply and ground gradients as the codewords toggle cells in the next row. The periodic nature of the error introduced by gradients results in a data-dependent spur in the frequency domain. This is illustrated in Fig. 5.22b as a tone due to the supply and ground gradient decreases the SFDR of the DAC.

One method of reducing the effect of gradients is to exploit common centroid layout techniques. One such technique is known as symmetrical switching [26]. In symmetrical switching, the order of the columns of the DAC elements is randomized

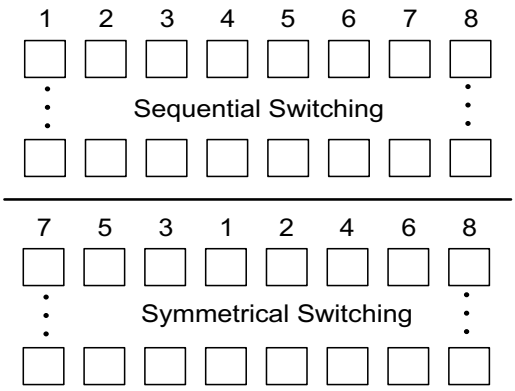


Fig. 5.23 DAC using symmetrical switching

as shown in Fig. 5.23. The resulting INL distortion is shown in Fig. 5.24. Although not obvious from the time–domain plot, the fast Fourier transform (FFT) plot exhibits much smaller high-frequency tones.

In all the above analysis, it was assumed that the main power supply and ground connections are located at a corner. If the connections are near the midpoint of any of the four borders of the array, as opposed to one of the four corners, the analysis becomes different. In this case, instead of a gradient error, the error becomes symmetric. Since the error is symmetric, the symmetrical switching technique has no effect on correcting this error.

In order to counteract symmetrical errors, a hierarchical symmetrical switching technique was developed [30]. In this technique, a different column ordering is used, as shown in Fig. 5.25. This ordering is done in such a way that it counteracts the effects of both gradient and symmetrical errors. Figure 5.26 shows the resulting INL distortion as well as the frequency domain plot showing the resulting tonal response with hierarchical symmetrical switching.

Thus far, it has been assumed that rows are decoded sequentially. In reality, however, there may be gradients running vertically that would not be corrected for by any of the abovementioned techniques. One popular method of accounting for

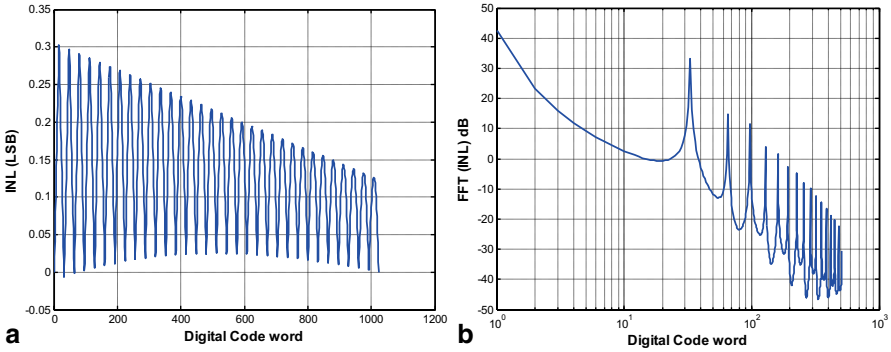


Fig. 5.24 Effect of gradient error on DAC with symmetrical switching **a** INL and **b** FFT of INL

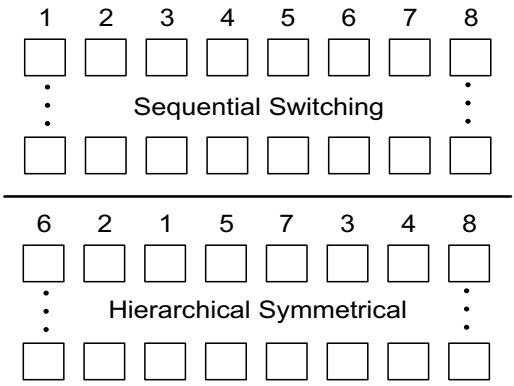


Fig. 5.25 Illustration of the hierarchical symmetrical switching technique

gradient and symmetrical errors in any direction is known as the Q^2 random walk decoding [52]. In this technique, the elements in a 4×4 array are tiled in a random fashion, as shown in Fig. 5.27. The disadvantage of this technique is higher decoding logic and wiring complexity. Note that each cell must now receive its own decoded bit to activate it. In all the previous techniques, a common row decoded bit was shared for all cells in a row. Using the Q^2 random walk technique, 14-bit resolution DAC is possible [52]. It is important to note that there are many variations of this sequencing reported in the literature that also attempt to reduce gradient errors [13, 16, 17, 21, 56].

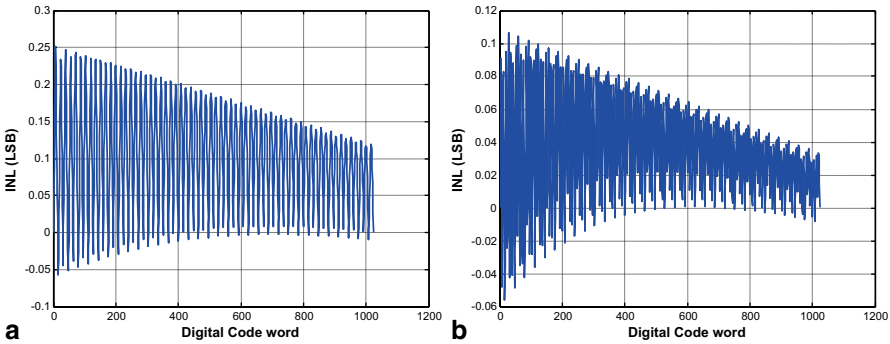


Fig. 5.26 INL of hierarchical symmetrical decoding with **a** gradient and **b** symmetrical errors

Fig. 5.27 Q^2 random walk decoding of a 4×4 tile

12	10	6	2
8	1	14	4
5	15	0	9
3	7	11	13

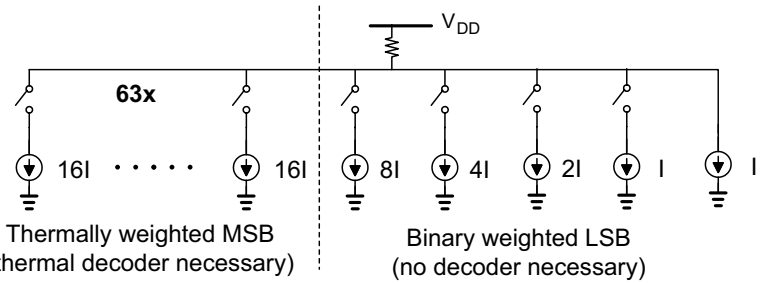


Fig. 5.28 Example of a 10-bit segmented DAC

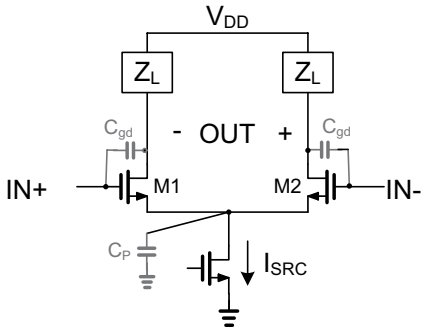


Fig. 5.29 High-speed current-steering DAC element

A common approach to DAC design is to split the DAC array into two (or more) segments. Such DACs are called segmented DACs [42]. The main idea behind a segmented DAC is to provide a trade-off between linearity and area. If all the DAC elements are binary-weighted elements, as opposed to equally sized elements, as shown earlier, large area savings are possible. In this case, the worst-case DNL would involve the switching of 2^{N-1} elements. This means that the worst-case DNL is now , where σ is the standard deviation of an LSB. Furthermore, no decoding logic is required. For the thermally weighted element arrays seen previously, the DNL is always equal to σ. The worst-case INL is the same regardless of the segmentation used. A compromise between area and linearity can be reached by allowing a small number of LSBs to be binary weighted, thereby limiting the linearity degradation while saving area. An example of a segmented 10-bit DAC is shown in Fig. 5.28. A 6-bit MSB and 4-bit LSB segmentation is shown.

5.6.2 Transient Effect and Glitches

An important source of nonlinearity for high-speed RFDACs is the glitch energy. There are many sources of glitch energy in an RFDAC. Consider a high-speed current steering-based DAC element shown in Fig. 5.29. Transistors M1 and M2 are used as switches to steer the current, I_{src}, to either OUT+ or OUT−. The C_{gd}

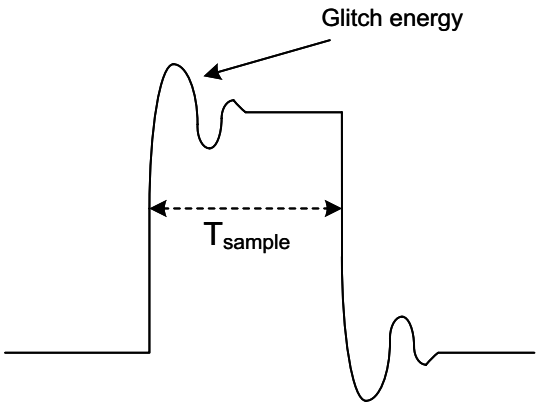


Fig. 5.30 Waveform demonstrating glitch energy at output of a DAC during switching

capacitance shown in the figure can couple the clock signal to the output, causing a glitch. Steering the current from one side to another also causes the current source I_{src} to glitch due to the fact that the capacitor C_p needs to charge or discharge as its voltage shifts from OUT+ to OUT−, or vice versa. This transient current is summed with the DC current from I_{src}, causing a glitch on the output.

The glitch energy can be quantified by examining the AC energy in the signal over a sampling period. Figure 5.30 shows the DAC output waveform when switching from one code word to the adjacent code word. The initial spike in the beginning is due to the glitching in the DAC cell. When calculating the glitch energy, this spike is averaged out over a sampling period. In some cases, only the area outside a ½ LSB error is integrated when calculating the glitch energy. The glitch energy is usually expressed in units of ps·V.

5.6.3 Sampling Effects in DACs

The time-sampled nature of DACs creates effects that are important to understand. Figure 5.31 shows a typical output of a DAC with a sine wave input. The output is a quantized sine wave. There are two important effects observed in the figure. First,

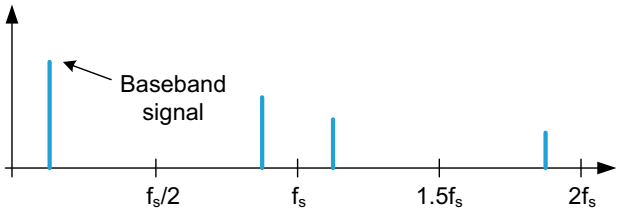


Fig. 5.31 Typical DAC output

alias terms arise around harmonics of f_s, the sampling frequency. To understand why this is the case, consider the case where only discrete samples are taken of the ideal analog signal, x(t), at locations $x(nT_s)$. The resulting waveform can be taken as an infinite summation of the $x(nT_s)$ terms. This is equivalent to convoluting the signal x(t) with an impulse train function with period T_s. Since the Fourier transform of x(t) is given as

$$X(f) \triangleq \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} dt \tag{5.20}$$

the output spectrum of the sampled terms can be given as

$$X_s(f) = \sum_{k=-\infty}^{\infty} X(f - kf_s) = \sum_{k=-\infty}^{\infty} T_s \bullet x(nT_s) e^{-j2\pi nTf} \tag{5.21}$$

Clearly, the output spectrum of Eq. (5.21) is periodic and symmetric around f_s.

The second effect seen in Fig. 5.31 is the decrease in amplitude of the alias terms. To understand the second term, it is important to realize that the quantized values at the output of the DAC arise by sampling the digital word and holding the analog-converted value for an entire sampling period. This creates a zero-order hold (ZOH) function, which is a rectangular pulse of width T_s, where $T_s = 1/F_s$. The ZOH function is effectively multiplied by the ideal analog signal. Since the frequency domain function of a ZOH is a sinc function, the DAC output spectrum is a convolution of a sinc function with the time-sampled analog signal. This is different from the previous case, wherein sampling function was an impulse function. Since the ZOH function is given by [35]

$$x_{ZOH}(t) = \sum_{k=-\infty}^{\infty} x(nT_s) \bullet rect\left(\frac{t - \frac{T_s}{2} - nT_s}{T_s} \right) \tag{5.22}$$

where rect is a rectangular function, the Fourier transform of the ZOH function is

$$X_{ZOH}(f) = \frac{1 - e^{-j2\pi fT_s}}{j2\pi fT_s} = e^{-j2\pi fT_s} \bullet sinc(fT_s) \tag{5.23}$$

This shows that the frequency spectrum of the sampled signal $x(nT_s)$ is convoluted with a sinc function in the frequency domain. This accounts for the decreasing amplitude of the alias terms observed in Fig. 5.31.

5.7 Digital Predistortion in RF Transmitters

The concept of digital predistortion (DPD) in RF transmitters covers a very broad set of techniques, which all rely on an estimate of the analog-distorted signal, which is then used to develop a predistortion correction digital term to compensate for

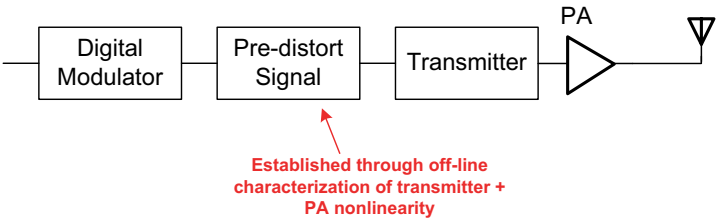


Fig. 5.32 Digital predistortion (DPD) in RF transmitters

this distortion [28, 31, 58]. Figure 5.32 shows the general concept of DPD. The nonlinearity of the PA and the transmitter is characterized off-line. A predistortion transfer function is then developed in such a way that it counteracts the distortion in the transmitter and PA. This is done to cover the different operating frequencies and gain settings of the transmitter, as well as the variation in the nonlinearity under temperature and power supply variation. On-chip temperature and power supply sensors are used as part of this predistortion transfer function. The DPD function is usually stored in the form of a look-up table.

There are several types of distortion that can occur in the RF transmitter path. The first of kind can be a nonlinear distortion causing intermodulation terms to arise in the output spectrum. This can be a result of AM–AM distortion or AM–PM distortion. It then becomes the objective of the predistortion block to counteract both these types of distortion. Mathematically, this can be expressed as [12]

$$y_n = A(r_n)e^{j(\varphi_n+\varphi(r_n)} \tag{5.24}$$

where r_n and φ_n are the transmitter digital input amplitude and phase, respectively, $A(\cdot)$ and $\varphi(\cdot)$ represent the AM–AM and AM–PM distortion, respectively. The simplest form of predistortion would be to multiply Eq. (5.24) with the inverse amplitude and phase transfer functions; however, not all nonlinear functions are invertible. Moreover, those that are invertible do not always have a one-to-one correspondence between input and output [59]. Direct measurement techniques would also require precise phase alignment between the digitized output PA signal sample and the digital input signal to estimate the AM–AM and AM–PM distortion.

Another method of obtaining the predistortion function is to use what is known as a semi-blind estimation technique [22]. The advantage of such a technique is that precise phase alignment between the PA output and the digital input is not required. In this technique, the statistics of the digital input signal and the output analog signal of the PA are examined. After obtaining the joint probability distribution function (PDF) of both the input signal, $f_X(r,\phi)$ and the PA output signal $f_Y(r,\phi)$, the AM–AM distortion function is given as [59]

$$A(r) = F_Y^{-1} \bullet F_X(r) \tag{5.25}$$

where $F_X(r)$ and $F_Y(r)$ are given by

$$F_X(r) = \int_0^r \int_{-\pi}^{\pi} f_X(r,\varphi) d\varphi \tag{5.26}$$

and

$$F_Y(r) = \int_0^r \int_{-\pi}^{\pi} f_Y(r,\varphi) d\varphi \tag{5.27}$$

The AM–PM distortion term can now be given as

$$\varphi(r) = \arg\min_{\varphi} \left[f_Y\left(\varphi \mid A^{-1}(r)\right) - f_X(\varphi + \varphi \mid r) \right]^2 \tag{5.28}$$

where $f_X(\phi \mid r)$ and $f_Y(\phi \mid r)$ are the conditional PDFs of phase ϕ given amplitude r associated with $f_X(\phi \mid r)$ and $f_Y(\phi \mid r)$, respectively.

Another popular method for DPD is to express Eq. (5.24) as a polynomial. More specifically, Eq. (5.24) can be rewritten as

$$y_n = x_n \sum_{i=0}^{K} a_i \left| x_n \right|^i \tag{5.29}$$

where a_i are polynomial coefficients. Specific known distortions in the transmitter can then be simplified to eliminate certain terms from Eq. (5.29). For example, if the transmitter is perfectly balanced, the even-order distortion terms can be eliminated. Also, higher-order terms may not have to be well approximated for sufficient level of distortion cancellation. Another strong advantage of the polynomial representation shown in Eq. (5.29) is that undesired cross-modulation between multicarrier outputs (in the case of carrier aggregation) can be digitally compensated for as well [14, 39].

In addition to nonlinear distortion cancellation, DPD can also be used to enhance harmonic and image rejection. This is done by digitally calibrating and adjusting for mismatch in the LO signal and imbalance between I and Q paths. For a fully digital transmitter, there is no imbalance between I and Q paths since these two paths only exist in the digital domain. In a direct upconverter architecture, as was shown in Sect. 5.2, there may be I and Q amplitude and phase imbalances, which may degrade the transmitter performance. Such imbalances can be digitally calibrated off-line by downcoverting the RF signal back into the baseband signal, digitizing the signal and defining the inverse transfer function to correct for image rejection.

A popular method of correcting for image blockers is to use a technique known as the Churchill's method [11]. Amplitude and phase errors between I and Q paths can be given as

$$I_1 = (1+\alpha) A \cos(2\pi f_i t) + A_{DC,I} \tag{5.30}$$

$$Q_1 = A\sin(2\pi f_i t + \epsilon) + A_{DC,Q} \tag{5.31}$$

where I_1 is the in-phase analog component at RF, Q_1 is the quadrature phase analog component at RF, α is the gain imbalance between the I and Q paths, ε is the phase imbalance between I and Q paths, $A_{DC,\,I}$ is the DC offset of the I path, $A_{DC,\,Q}$ is the DC offset of the Q path. In order to eliminate the gain and phase imbalance, the complex input signal, $s_1 = I_1 + jQ_1$, must be modified by a complex correction factor yielding s_2 given by

$$s_2 = s_1 + (E + jP)I_1 \tag{5.32}$$

where E and P are the real and imaginary components of the image-correcting factor. It can be shown that if *E* and *P* are given by

$$E = \frac{\cos\epsilon}{1+\alpha} - 1 \tag{5.33}$$

$$P = -\frac{\sin\epsilon}{1+\alpha} \tag{5.34}$$

the resulting s_2 value from Eq. (5.32) is given by

$$s_2 = A\cos(2\pi f_i t)\cos\epsilon + jA\sin(2\pi f_i t)\cos\epsilon \tag{5.35}$$

Note that the real and imaginary components of s_2 have no amplitude α or phase E imbalance.

Generally speaking, the phase and amplitude offset between the I and Q channels is unknown and must be estimated. Churchill's method describes a four-point discrete Fourier transform (DFT) method to be performed on the data with an input at exactly one fourth the sampling rate. More specifically,

$$\hat{\vec{s}}_1 = DFT\{I_1 + jQ_1\} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -j & -1 & j \\ 1 & -1 & 1 & -1 \\ 1 & j & -1 & -j \end{bmatrix} \vec{s}_1 \tag{5.36}$$

The result is four equally spaced samples over one input period. The first component of $\hat{\vec{s}}_1$ is the average of the four samples, or the DC estimate of the input signal. The second component of $\hat{\vec{s}}_1$ is the desired signal and the fourth component is the image signal. The third component is also a measure of the DC component. Since it is desirable to have no image component, it follows from Eqs. (5.32) and (5.36) that

$$\hat{\vec{s}}_1(4) + (E + jP)\cdot I_1(4) = 0 \tag{5.37}$$

In other words, the image correction terms E and P are given as

$$E = -Re\left\{\frac{2\hat{\tilde{s}}_1(4)}{\hat{\tilde{s}}_1^*(2) + \hat{\tilde{s}}_1(4)}\right\} \tag{5.38}$$

$$P = -Im\left\{\frac{2\hat{\tilde{s}}_1(4)}{\hat{\tilde{s}}_1^*(2) + \hat{\tilde{s}}_1(4)}\right\} \tag{5.39}$$

The DC correction terms can be given as

$$A_{DC,I} = \frac{1}{4} Re\left\{\hat{\tilde{s}}_1(1)\right\} \tag{5.40}$$

$$A_{DC,Q} = \frac{1}{4} Im\left\{\hat{\tilde{s}}_1(1)\right\} \tag{5.41}$$

The correction estimates given by E and P can exactly cancel out the image component resulting from mismatch between the I and Q channels. Factors such as the quantization noise resulting from bit width limitations and noise in the input signal itself would limit the degree of possible image rejection. Bit width limitations can be easily avoided by design (increasing the bit width such that it is not a limiting factor in achieving the desired image rejection). Noise in the signal itself can be a result of device noise of any analog component in the transmitter, which includes the DAC, the anti-alias filter, and the up-conversion mixer. This noise is usually random in nature and difficult to avoid. One method of avoiding this noise component is to increase the number of points in the DFT sample. Assuming the noise sources are independent, the variance in the error of the image correction terms for an N point DFT is given as

$$\sigma_{img,N}^2 = \frac{2\sigma_{img,1}^2}{N} \tag{5.42}$$

where N is an integer multiple of 4 and $\sigma_{img,1}^2$ is the image correction estimate of a four-point DFT. Calibration with a test frequency other than one fourth the sampling rate can be used with a simple modification of Churchill's method [37].

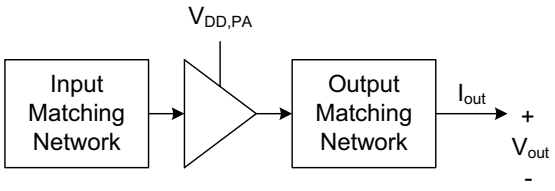


Fig. 5.33 A generic power amplifier

5.8 High-Linearity and High-Efficiency PAs

One important block in the transmitter path is the PA. Both the linearity and power-efficiency of the PA can be the limiting factor for both performance and cost of the entire transceiver. In this section, different configurations and classes of PAs are detailed.

Before providing a detailed discussion of different classes of PAs, a few basic concepts that are used to quantify PA performance are first given. A generic PA is shown in Fig. 5.33. As the figure shows, both input and output matching networks are provided. This is to ensure the maximum power transfer between the transmitter and the PA at the input node, as well as between the PA and the antenna or T/R switch at the output node. In the context of PAs, the output power refers to the total available power at the circuit following the PA. The total power gain of the PA in Fig. 5.33 can be shown to be

$$G_T = \underbrace{\frac{1}{1-|s_{11}|^2}}_{G_{in}} \underbrace{|s_{21}|^2}_{G_1} \underbrace{\frac{1}{1-|s_{22}|^2}}_{G_{out}} \tag{5.43}$$

where s_{11} is the input reflection coefficient of the PA, s_{22} is the output reflection coefficient, and s_{21} is the forward transmission coefficient. Equation (5.43) assumes that the input and output are matched for maximum power transfer. In the literature, the term power gain of a PA usually refers to the $G_{PA}=G_1 \cdot G_{out}$ product.

The transfer of the input power to the output with a certain power gain, G, is performed by powering the PA from a supply source, $V_{D,\,PA}$, shown in Fig. 5.33. The amount of power dissipated from $V_{D,\,PA}$ is important since this determines, to a large extent, the power dissipation of the entire RF front-end transceiver. A metric commonly used to quantify this is the *amplifier efficiency, η,* which is given as

$$\eta = \frac{P_{OUT\_PA}}{P_{VDD\_PA}} \tag{5.44}$$

where P_{OUT_PA} is the output power of the PA and $P_{VD,\,PA}$ is the power drawn from $V_{D,\,PA}$. The drawback of this metric is that it does not take into account the input power level. Another metric that incorporates the input power level is known as the *power-added efficiency (η_{PAE})*, which is given by

$$\eta_{PAE} = \frac{P_{PA,OUT} - P_{PA,IN}}{P_{VDD,PA}} = \eta\left(1 - \frac{1}{G_{PA}}\right) \tag{5.45}$$

The PAE can be interpreted as the efficiency of the PA as a ratio of the difference between the input and output power of the PA to the power dissipated by the power supply of the PA.

The power efficiency and PAE provide a method of comparing the overall efficiency of the PA. These metrics, however, do not take into account the relative output swing (either voltage or current) when expressing efficiency. An alternative method of quantifying the output power of a PA is the *power output capability,* C_P, which is given as

$$C_P = \frac{P_{PA\_OUT}}{I_{out,pk} V_{out,pk}} \tag{5.46}$$

This metric is useful in comparing different types of PAs. Note that the peak output current and voltages are measured at the amplifier's output before the termination network, whereas the output power is measured after the termination network. Other metrics that express linearity, such as P_{1dB}, IIP3, and ACPR are important for PAs; however, these metrics have been discussed at length previously in Chap. 3 and earlier sections of this chapter.

Another important concept for PAs is the concept of conduction angle. The conduction angle is defined as the duration (in phase) where the current drawn from $V_{D,PA}$ is nonzero. For example, if the current drawn from $V_{D,PA}$ is continuous, then the conduction angle, $2\theta_c$, is said to be 360°. The '2' term emphasizes the assumption that the current waveform is symmetric; meaning that the current waveform from $[-\theta_c, 0]$ is a mirror image of the current from $[0, \theta_c]$. Higher power efficiency in a PA is achieved by lowering the conduction angle. Lowering the conduction angle, however, reduces the linearity of the PA. This effect gives rise to the established trade-off between linearity and power efficiency in PA design.

Class-A PA The simplest PA is a class-A PA. A class-A PA is defined as a PA with conduction angle, $2\theta_c$, of 360°. Figure 5.34 shows the typical waveforms of a class-A amplifier for a sine wave input. The current through $V_{D,PA}$ is given as

$$i_A(\theta) = I_{dc} - I_{out} \sin\theta \tag{5.47}$$

where I_{dc} is the DC component of current drawn from $V_{D,PA}$, I_{out} is the peak output current, and θ is the angular time given as $\theta = 2\pi ft$. The output current and voltage are then expressed as

$$i_{out}(\theta) = I_{dc} - i(\theta) = I_{out} \sin\theta \tag{5.48}$$

$$v(\theta) = V_{dc} + V_{out} \sin\theta = V_{dc} + RI_{out} \sin\theta \tag{5.49}$$

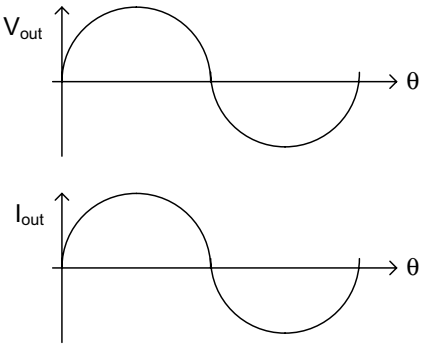


Fig. 5.34 Typical current and voltage waveforms of a class-A PA

where R is the amplifier's load resistance. Note that the DC power consumption, or current drawn from $V_{D,\,PA}$, is given as

$$P_{dc} = V_{DD,PA} \bullet I_{dc} = \frac{V_{DD,PA}^2}{R} \tag{5.50}$$

Since by definition, a class-A amplifier is always conducting current, this can only occur if both the output current and the voltage are less than the DC current and voltage, respectively. This means that the maximum output power dissipation can now be given as

$$\max P_{out} = \max \frac{V_{out}^2}{2R} = \frac{V_{dc}^2}{2R} \tag{5.51}$$

The power efficiency, η can now be given as

$$\eta = \frac{P_{out}}{P_{dc}} \leq 50\% \tag{5.52}$$

This means that if 1W output is required, the power draw of more than 2W is needed from the power supply. Although Eq. (5.51) indicates that a higher power output level would enhance the efficiency of the PA, a certain back-off ratio may be required in order to accommodate the bursty nature of the signal amplitude. The peak-to-average ratio (PAPR) can typically vary from 7 to 14dB. This means that higher PAPR usually implies lower power efficiency. The efficiency of class-A amplifiers can be as low as 10–20 % for low output power levels. Figure 5.35 shows the power efficiency of a class-A amplifier as a function of the output voltage level, V_{out}.

The power output capability, C_P, of a class-A amplifier is given as

$$C_P = \frac{P_{PA,OUT,max}}{(2V_{dc})(2I_{dc})} = \frac{V_{DD,PA}^2}{2R} \frac{R}{4V_{DD,PA}^2} = 0.125 \tag{5.53}$$

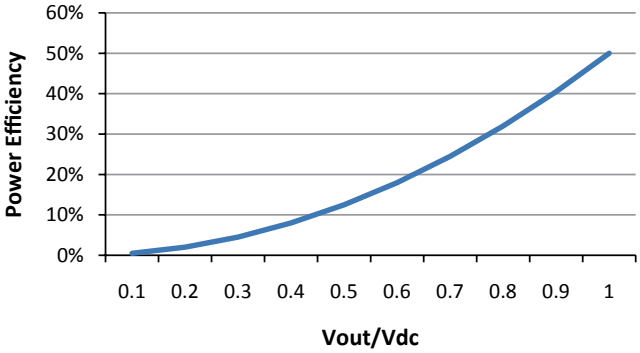


Fig. 5.35 Class-A PA η versus V_{out}

The peak output voltage and current at the amplifier's output are twice that of the PA power supply voltage.

Class-B Amplifier Power efficiencies of 10–20 % are impractical for portable low-power applications. One way to enhance the efficiency of the PA is to saturate the PA in such a way that it does not conduct the current for half the duration of the input period. In other words, the conduction angle can be lowered to $2\theta_c = 180°$. Figure 5.36 shows the typical current and voltage waveforms of a class-B amplifier for a sine wave input. As the figure shows, class-B operation is enabled by eliminating the negative half of the current swing. Two class-B amplifiers connected in parallel are assumed. One amplifier conducts current between 0° and 180° and the other conducts current between 180° and 360°.

The output power of the class-B amplifier described above is the same as that of a class-A amplifier. The power drawn from $V_{D,\,PA}$, on the other hand is now given as

$$P_{VDD,PA} = V_{DD,PA} \bullet \int_0^{2\pi} I \left| \sin\theta \right| d\theta = \frac{2}{\pi} \frac{V_{DD,PA}^2}{R} \tag{5.54}$$

This means that the peak efficiency occurs when the output swing is equal to the power supply voltage of the PA, $V_{D,\,PA}$. The power efficiency of the class-B amplifier is given as

$$\eta = \frac{P_{out}}{P_{dc}} = \frac{\pi}{4} \frac{V_{out}}{V_{DD,PA}} \le \eta_{max} = \frac{\pi}{4} = 78.5\% \tag{5.55}$$

The peak power efficiency of a class-B amplifier is a major improvement over a class-A amplifier. The linearity, however, is worse. To understand why, consider the waveforms that were shown in Fig. 5.36. It is assumed that there are two independent branches that conduct at different times. First, if there is any mismatch resulting in a gap where no current is conducted, then this introduces a strong nonlinearity near the zero crossing of the amplifier. Second, the absolute value of the current is used, which itself also introduces a strong nonlinearity at the zero crossing of the amplifier. One way to mitigate this effect is to add a constant current in parallel to

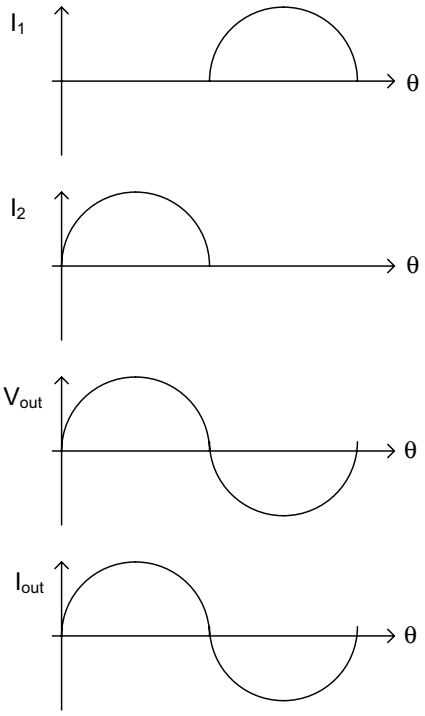


Fig. 5.36 Typical current and voltage waveforms of a class-B PA

the class-B amplifier. Since now less of a fraction of the current comes from the nonlinear component, the amplifier is effectively linearized. Such an amplifier is called a class-AB amplifier. The power output capability of a class-B amplifier is given as

$$C_P = \frac{P_{PA,out,max}}{2(2V_{DD,PA})I_{VDD,PA}} = \frac{V_{DD,PA}^2}{2R}\frac{R}{4V_{DD,PA}^2} = 0.125 \tag{5.56}$$

The output current is half of that of the class-A amplifier case, since the conduction angle is half; however, the assumption here is that there are two class-B amplifiers working in different conduction phases. This makes the class-B amplifier have the same output power capability as a class-A amplifier.

Class-C Amplifier In a class-B amplifier, it has been demonstrated that limiting the conduction angle can have a large impact on the power efficiency of the amplifier. As was demonstrated, this came at the expense of increased nonlinearity. Increased nonlinearity implies higher harmonic content. In the context of cognitive radios, a wideband operation is required of the transmitter. This implies that harmonic frequencies of the PA can fall onto other users and thereby jamming the user's transmitted signal. This would necessitate the need of an RF BPF or to lower the output power to attenuate the harmonic content.

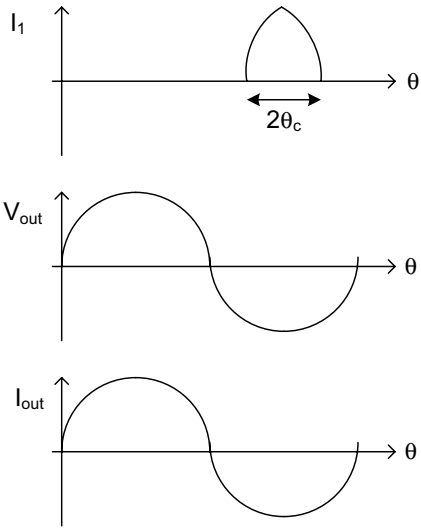


Fig. 5.37 Typical current and voltage waveforms of a class-B PA

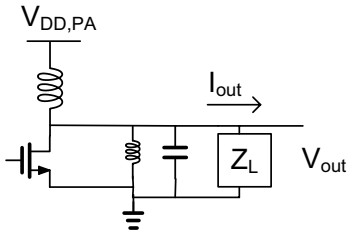


Fig. 5.38 Generic class-C amplifier

In a class-C amplifier, higher power efficiencies and better linearity than class-B amplifiers are attempted. Higher efficiency is achieved by lowering the conduction angle even further. This is accomplished by controlling the DC bias of the main PA device such that it only conducts current during a short duration of a period. Figure 5.37 shows the typical current and voltage waveforms of a class-C amplifier. Although a sinusoidal waveform for the current is shown in Fig. 5.37, it can take on any shape. The linearity of a class-C amplifier is improved by providing a shunt LC network, as shown in Fig. 5.38. This shunt LC network would pass the signal at the operating frequency, but attenuate it at harmonic frequencies of the output RF frequency.

One popular method of modeling the current with the small conduction angle shown in Fig. 5.37 is to use the following model [1]:

$$i(\theta) = \begin{cases} \dfrac{I_M(\cos\theta - \cos\theta_c)}{1-\cos\theta_c} & , -\theta_c \le \theta \le \theta_c \\ 0 & , \quad otherwise \end{cases} \tag{5.57}$$

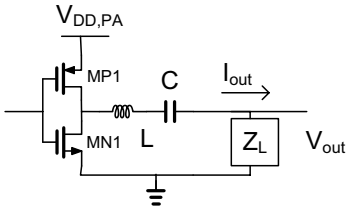


Fig. 5.39 A generic class-D power amplifier

This would yield an average, or DC, current of

$$I_{dc} = I_M \frac{\sin\theta_c - \theta_c \cos\theta_c}{\pi(1-\cos\theta_c)} \tag{5.58}$$

and the current flowing to the load of the PA is given as

$$I_{out} = I_M \frac{\theta_c - \sin\theta_c \cos\theta_c}{\pi(1-\cos\theta_c)} \tag{5.59}$$

which give an average power draw from $V_{DD,\,PA}$ of

$$P_{VDD,PA} = V_{DD,PA} I_{dc} = V_{DD,PA} \bullet I_M \frac{\sin\theta_c - \theta_c \cos\theta_c}{\pi(1-\cos\theta_c)} \tag{5.60}$$

The power efficiency is then given as

$$\eta = \frac{P_{out}}{P_{dc}} = \frac{1}{2}\frac{V_{out}}{V_{DD,PA}} \frac{\theta_c - \sin\theta_c \cos\theta_c}{\sin\theta_c - \theta_c \cos\theta_c} \le \eta_{max} = \frac{\theta_c - \sin\theta_c \cos\theta_c}{\sin\theta_c - \theta_c \cos\theta_c} \tag{5.61}$$

The output power capability of the class-C PA is given as

$$C_P = \frac{P_{out}}{V_{DD,PA} I_M} = \frac{\theta_c - \sin\theta_c \cos\theta_c}{4(\sin\theta_c - \theta_c \cos\theta_c)} \tag{5.62}$$

which has a maximum value of $C_p = 0.1341$ at a conduction angle of $2\theta_c = 245°$. According to Eq. (5.59), the power efficiency at this conduction angle is >60 %. The main disadvantage of a class-C amplifier is that it relies on an external LC tank in order to boost the linearity of the amplifier. In the context of a wideband transmitter, the BPF must be tunable, which increases complexity and degrades the power efficiency of the PA (since the tunable filter would introduce more loss than a fixed filter).

Class-D Amplifier Higher power efficiencies than any of the previous classes of PAs is possible by using what is known as a *zero-voltage switched (ZVS) amplifier,* or simply a switched amplifier. A generic class-D PA is shown in Fig. 5.39. Similar to a class-C amplifier, a class-D amplifier relies on an LC tank load. The main difference, however, is how the active devices are used. A pair of high-voltage field effect transistor (FET) output devices, MP1 and MN1, is hard switched by the RF

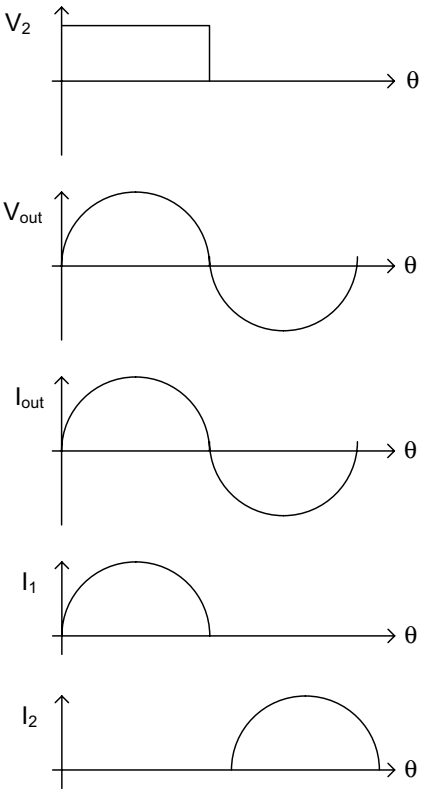


Fig. 5.40 Typical current and voltage waveforms of a class-D PA

input, meaning they act as switches. In the analysis that follows, it is assumed that the FETs have zero ON resistance, infinite OFF resistance, zero parasitic capacitance, and the LC tank is a lossless ideal tank.

The typical current and voltage waveforms of a class-D amplifier are shown in Fig. 5.40. The voltage at the drain of the NFET, v_2, is a square wave, as it follows the input voltage. The distinguishing feature of a class-D amplifier is the fact that when the v_2 voltage is nonzero, the i_2 current is zero, and vice versa. In this case, the power dissipated by the PA that is not sent to the output is zero. The series resonant circuit is tuned to the RF output frequency, and as a result, the output current is sinusoidal. Since a square wave can be represented as an infinite sum of odd-order harmonic sinusoidal signals, v_2 can be given as

$$v_2(\theta) = V_{DD,PA}\left(\frac{1}{2} + \frac{2}{\pi}\sum_{n=1}^{\infty}\frac{\sin\left[(2n-1)\theta\right]}{2n-1}\right) \tag{5.63}$$

The output current contains no higher order harmonics as is given as

$$i_{out}(\theta) = \frac{v_2\,|_{n=1}}{R} = \frac{2}{\pi}\frac{V_{DD,PA}}{R}\sin\theta \tag{5.64}$$

Table 5.2 Summary of PA classes and their performance

Class	η_{MAX}(%)	η_{TYP}(%)	C_p	Main idea
A	50	30–45	0.125	$2\theta_c = 360°$
B	78	60–70	0.125	$2\theta_c = 180°$
C	100	70–90	0.134	$2\theta_c < 180°$; V_{out}min when $I_{out} > 0$
D	100	80–90	0.159	V_0 min when $I_0 > 0$ Min time overlap when $I_0 > 0$ and $V_0 > 0$
E	100	80–90	0.09	Combine both elements of class-D in LC tank + switch

The output voltage is

$$v_{out}(\theta) = i_{out}(\theta) \bullet R = \frac{2}{\pi} V_{DD,PA} \sin\theta \tag{5.65}$$

From Eqs. (5.64) and (5.65), the output power is

$$P_{out}(\theta) = \frac{2}{\pi^2} \frac{V_{DD,PA}^2}{R} \tag{5.66}$$

which is the same as the power dissipated by the PA, yielding a power efficiency $\eta = 100\,\%$. The maximum power output capability of the class-D amplifier is given as

$$C_P = \frac{P_{out}}{2V_{DD,PA} I_{out}} = \frac{2V_{DD,PA}^2}{2\pi^2 R} \frac{\pi R}{V_{DD,PA}^2 2} = \frac{1}{2\pi} \tag{5.67}$$

Practically, losses due to finite FET ON and OFF resistances, losses in the LC tank, and parasitic capacitances associated with the FETs will lower the efficiency of the amplifier. Class-E and class-F amplifiers are two other variants of switched amplifiers that have the same basic operating principal as a class-D amplifier, but with some efficiency enhancements over class-D amplifiers.

Table 5.2 summarizes the different classes of PAs discussed above. As the table shows, the amplifier efficiency is well correlated to the conduction angle, $2\theta_c$. The typical efficiencies depend heavily on the back off required on the PA to accommodate the required PAPR. For example, a WCDMA signal may need a PAPR of 7dB, while a digital video broadcasting-terrestrial (DVB-T) signal would require a PAPR of 14dB, resulting in less power efficiency in the PA.

Summary

In this chapter, design issues in wideband transmitters have been explored. The requirements of transmitters including EVM, ACPR, and EIRP have been detailed. The simplest transmit architecture, the direct upconverter, has been detailed. The

Cartesian loop transmitter has been shown to provide better linearity on the expense of complexity and data bandwidth. Polar modulator architecture, which has gained recent popularity due to its ability to support nonlinear power efficient PAs, has also been detailed.

Special attention has been devoted to direct digital upconverter transmit architectures. This included the digital upconverter filters and upsamplers. It also included an in-depth analysis of DAC design. The concept of DPD as a means to correct for analog impairments has been introduced. Various calibration techniques to correct for nonlinearity as well as image rejection have been given. Finally, various PA topologies have been explored. This included transconductance-based amplifiers, such as class-A, class-B, class-AB, and class-C amplifiers. Switched amplifiers have also been explored, which include class-D, class-E, and class-F amplifiers.

References

1. M. Albulet, *RF Power Amplifiers*, Atlanta, GA, Noble Publishing, 2001.
2. Analog Devices. 2.5GSPS Direct Digital Synthesizer with 12-bit DAC, AD9915 Datasheet Rev. B, 2012.
3. Analog Devices, 3.5GSPS Direct Digital Synthesizer with 12-bit DAC, AD9914 Datasheet Rev. B, 2012.
4. G. Avitabile and N. Lofu, "EVM Degradation in EDGE Two Point Modulation Scheme due to Quantization Effects," *IEEE Radio and Wireless Conference*, 2004, pp. 299–302.
5. Bastos, J. et. al., "A 12-bit intrinsic accuracy high-speed CMOS DAC," *IEEE J. of Solid-State Circuits*, vol. 33, no. 12, Dec 1998, pp. 1959–1969.
6. E. Bechthum, et. al., "Systematic analysis of the impact of mixing locality on mixing-DAC linearity for multicarrier GSM," *IEEE Int'l Symposium on Circuits and Systems*, 2012, pp. 241–244.
7. A. Behzad, Wireless LAN Radios:System Definition to Transistor Design, Hoboken, New Jersey:John Wiley & Sons, 2008.
8. Z. Boos et al. "A Fully Digital Multimode Polar Transmitter Employing 17b RF DAC in 3G Mode," *IEEE Int'l Solid-State Circuits Conf (ISSCC)*, 2011, pp. 376–377.
9. R. Caverly, CMOS RFIC Design Principles, Norwood, MA, Artech House, 2007.
10. J. Chen, et. al., "The Design of All-Digital Polar Transmitter Based on ADPLL and Phase Synchronized ΔΣ Modulator," *IEEE J. of Solid-State Circuits (JSSC)*, vol. 47, no. 5, May 2012, pp. 1154–1164.
11. F. E. Churchill, G. W. Ogar, and B. J. Thompson, "The correction of I and Q errors in a coherent processor," *IEEE Trans on Aerospace and Electronic Systems*, vol. AES-17, no. 1, pp. 131–137, Jan 1981.
12. M. Collados, et. al., "A Low-Current Digitally Predistorted 3G-4G Transmitter in 40 nm CMOS," *IEEE Radio Frequency Integrated Circuits (RFIC) Symposium 2013*, pp. 141–144.
13. Y. Cong and R. Geiger, "Switching sequence optimization for gradient error compensation in thermometer-decoded DAC arrays," *IEEE Trans on Circuits and Systems II*, vol. 47,no. 7, July 2000, pp. 585–595.
14. H. Dabag, et. al., "All-Digital Cancellation Technique to Mitigate Receiver Desensitization in Uplink Carrier Aggregation in Cellular Handsets," *IEEE Trans on Microwave Theory and Techniques*, vol. 61, no. 12, Part 2, 2013, pp. 4754–4765.
15. H. Darabi and A. Abidi, "Noise in RF-CMOS mixers:a simple physical model," *IEEE J. of Solid-State Circuits*, vol. 35, no. 1, Jan 2000, pp. 15–25.

16. J. Deveugele, et. al., "A Gradient-Error and Edge-Effect Tolerant Switching Scheme for High-Accuracy DAC," *IEEE Trans on Circuits and Systems I*, vol. 51, no. 1, Jan 2004, pp. 191–195.
17. J. Deveugele, et. al., "A gradient-error and edge-effect tolerant switching scheme for a high-accuracy DAC," *IEEE Trans on Circuits and Systems I*, vol. 50, no. 1, Jan 2004, pp. 191–195.
18. G. Hueber and R. Staszewski, *Multi-Mode/Multi-Band RF Transceivers for Wireless Communications*, Hoboken, New Jersey, John Wiley & Sons, 2011.
19. I. Jung and Y. Kim, "A CMOS low-power digital polar modulator system integration for WCDMA transmitter," *IEEE Trans on Industrial Electronics*, vol. 59, no. 2, Feb 2012, pp. 1154–1160.
20. P. Kennington, R. Wilkinson, K. Parsons, "Noise performance of a Cartesian loop transmitter," *IEEE Trans on Vehicular Technology*, vol. 46, no. 2, Feb 1997, pp. 467–476.
21. K. Krishna, et. al., "Spatial averaging and ordering in matched element arrays," *IEEE Custom Integrated Circuits Conference(CICC)*, 2002, pp. 453–456.
22. M. Li, et. al., "Signal processing challenges for emerging digital intensive and digitally assisted transceivers with deeply scaled technology (invited)," *IEEE Workshop on Signal Processing Systems*, 2013, pp. 324–329.
23. F. Luo, Digital Front-End in Wireless Communications and Broadcasting, Cambridge, UK:Cambridge University Press, 2011.
24. F. Maloberti, *Data Converters*, Dorchrecht, Netherlands:Springer, 2007.
25. Maxim Integrated, 14-Bit, 2.3Gsps Direct RF Synthesis DAC with Selectable Frequency Response, MAX 5879, 2012.
26. T. Miki, et. al., "An 80-Mhz 8-bit CMOS D/A Converter," *IEEE J. of Solid-State Circuits*, vol. SC-21, no. 12, Dec 1986, pp. 983–988.
27. B. Mohr, et. al., "An RFDAC based reconfigurable multistandard transmitter in 65 nm CMOS," *IEEE Radio Frequency Integrated Circuits (RFIC) Symposium*, 2012, pp. 109–112.
28. B. Mohr, et. al., "Analysis of digital predistortion architectures for direct digital-to-RF transmitter systems," *IEEE Int'l Midwest Symposium on Circuits and Systems(MWSCAS)*, 2012, pp. 650–653.
29. A. Montalvo, "Polar Modulators for Linear Wireless Transmitters," *IEEE Int'l Solid-State Circuits Conference (ISSCC) 2005*, T7 tutorial.
30. Y. Nakamura, et. al., "A 10-b 70-MS/s CMOS D/A Converter," *IEEE J. of Solid-State Circuits*, vol. 26, no. 4, Apr 1991, pp. 637–642.
31. H. Nelson and S. Ferguson, "Digital predistortion techniques for mobile PA test," *High Frequency Electronics*, vol. 13, no. 6, June 2014, pp. 32–42.
32. B. Neurauter, et. al., "GSM 900/DCS 1800 fractional-N modulator with two-point modulation," *Proc of the Microwave Symposium Digest*, vol. 1, 2002, pp. 425–428.
33. M. Norris, "Transmitter Architectures [GSM Handsets]," *IEE Colloquium on The Design of Digital Cellular Handsets*, 1998, pp. 4/1–4/6.
34. P. Nuyts, et. al., "A fully digital delay line based GHz range multimode transmitter front-end in 65-nm CMOS," *IEEE J. of Solid-State Circuits*, vol. 47, no. 7, July 2012, pp. 1681–1692.
35. A. Oppenheimer and R. Schafer, *Discrete-Time Signal Processing*, Upper Saddle River, NJ:Prentice-Hall, 2009.
36. M. Pelgrom, et. al., "Matching properties of MOS transistors," *IEEE J. of Solid-State Circuits*, vol. 24, no. 10, Oct 1989, pp. 1433–1439.
37. K. Pun, J. da Franca and C. Azeredo-Leme, Circuit Design for Wireless Communications: Improved Techniques for Image Rejection in Wideband Quadrature Receivers, Boston:Kluwer Academic Publishers, 2003.
38. B. Razavi, *RF Microelectronics*, Upper Saddle River, New Jersey: Prentice-Hall, 1998.
39. P. Roblin, et. al., "Concurrent Linearization," *IEEE Microwave Magazine*, vol. 14, no. 7, Nov/Dec 2013, pp. 75–91.
40. M. Sadeghifar and J. Wikner, "Modeling and analysis of aliasing image spurs problem in digital-RF-converter-based IQ modulators," *IEEE Int'l Symposium on Circuits and Systems (ISCAS)*, 2013, pp. 578–581.

41. W. Sander, S. Schell, B. Sander, "Polar Modulator for Multi-mode Cell Phones," *IEEE Custom Integrated Circuits Conf (CICC)*, 2003, pp. 439–445.
42. J. Schoeff, "An inherently monotonic 12 bit DAC," *IEEE J. of Solid-State Circuits*, vol. SC-14, no. 6, Dec 1979, pp. 940–911.
43. N. Silva, et. al., "Novel fine tunable multichannel all-digital transmitter," *Proc of the Microwave Symposium Digest*, 2013, pp. 1–3.
44. T. Sowlati, et. al., "Quad-Band GSM/GPRS/EDGE Polar Loop Transmitter," *IEEE J. of Solid-State Circuits*, vol. 39, no. 12, Dec 2004, pp. 2179–2189.
45. J. Staryzk and R. Mohn, "Cost-oriented design of a 14-bit current steering DAC macrocell," *IEEE Int'l Circuits and Systems Conference (ISCAS)*, vol. 1, 2003, pp. 965–968.
46. S. Talaeie, et. al., "A 0.18um CMOS fully integrated RFDAC and VGA for WCDMA transmitters," *IEEE Radio Frequency Integrated Circuits (RFIC) Symposium*, 2008, pp. 157–160.
47. S. Tang, et. al., "FinFET—a quasi planar double-gate MOSFET," *IEEE Int'l Solid-State Circuits Conference (ISSCC)*, 2001, pp. 118–119.
48. Texas Instruments, Quad-Channel, 16-Bit, 1.5GSPS Digital-to-Analog Converter (DAC), DAC34SH84 Datasheet, 2013.
49. W. Tseng, C. Fan, and J. Wu, "A 12b 1.25GS/s D/A in 90 nm CMOS with > 70 dB SFDR up to 500 MHz," *IEEE Int'l Solid-State Circuits Conf (ISSCC)*, 2011, pp. 192–193.
50. P. Vadiyanathan and T. Nguyen, "A TRICK for the design of FIR half-band filters," *IEEE Trans on Circuits and Systems*, vol. CAS-34, no. 3, March 1987, pp. 297–300.
51. R. Vaishnavi and V. Elamaran, "Implementation of CIC filter for DUC/DDC," *Int'l Journal of Engineering and Technology (IJET)*, vol. 5, no. 1, Feb 2013, pp. 357–365.
52. G. Van der Plas, et. al., "A 14-bit accuracy intrinsic accuracy Q^2 random walk CMOS DAC," *IEEE J. of Solid-State Circuits*, vol. 34, no. 12, Dec 1999, pp. 1708–1718.
53. J. Vankka, et. al., "A Multicarrier GMSK Modulator with On-Chip D/A Converter for Base Stations," *IEEE J. of Solid-State Circuits*, vol. 37, no. 10, Oct 2002, pp. 1226–1234
54. J. Volakis, *Antenna Engineering Handbook*, New York, NY:Mc Graw Hill, 2007.
55. A. Werquin, A. Frappe, A. Kaiser, "A multi-path multi-rate CMOS polar DPA for wideband multi-standard RF transmitters," *IEEE Radio Frequency Integrated Circuits (RFIC) Symposium*, 2013, pp. 327–330.
56. Z. Yu, D. Chen, R. Geiger, "1-D and 2-D switching strategies achieving near optimal INL for thermometer-coded current steering DACs," *IEEE Int'l Conference on Circuits and Systems (ISCAS)*, vol. 1, 2003, pp. 909–912.
57. Y. Yu, H. Shi, W. Ni, "An I/Q Channel 12-bit 200MS/s CMOS DAC with Three Stage Decoders for Wireless Communication," *Int'l Conf. on Wireless Communications and Signal Processing*, pp. 1–4, 2009.
58. Z. Zhu, H. Leung and X. Huang, "Challenges in reconfigurable radio transceivers and application of nonlinear signal processing for RF impairment mitigation," *IEEE Circuits and Systems Magazine*, pp. 44–63, 2013.
59. Z. Zhu, et. al., "Challenges in Reconfigurable Radio Transceivers and Application of Nonlinear Signal Processing for RF Impairment Mitigation," *IEEE Circuits and Systems Magazine*, vol. 13, no. 1, 2013, pp. 44–65, 2013.

Chapter 6
Wideband Phase-Locked-Loop-Based Frequency Synthesis

Frequency synthesizers in cognitive radio play a central role in providing the local oscillator (LO) signal for frequency translation and channel scanning. At the heart of the frequency synthesizer is the phase-locked loop (PLL). In this chapter, frequency synthesizer design requirements including frequency range and phase noise performance are first given. This is followed by detailed analysis of both integer-N PLL and ΣΔ fractional-N PLL. Sources of phase noise in PLL design are detailed along with a methodology to minimize the phase noise. Design details of important PLL components such as the charge pump, the voltage-controlled oscillator (VCO) and the feedback frequency divider are given.

6.1 Jitter and Phase Noise Primer

Before starting a detailed discussion on PLL design and how to optimize its performance, a few terms must first be defined. The two most prevalent terms describing the performance of PLLs are phase noise and jitter. Jitter can be defined as the *statistical* measure of the deviation of a periodic signal's actual edges corrupted by noise from the ideal periodic signal. The corruption of the edges is due to amplitude noise (either external or intrinsic to the PLL circuitry) conversion into phase noise (AM–PM conversion of noise), the distinction between amplitude and phase noise is shown graphically in Fig. 6.1.

Jitter can be deterministic or random. Deterministic jitter occurs in a predictable and periodic fashion. Random jitter at any given time is uncorrelated to any other sample. Phase noise is the frequency representation of phase fluctuations of a signal. Figure 6.2 illustrates both deterministic and random jitter in both the time and frequency domains. As the figure shows, deterministic jitter manifests itself as spurious response in the frequency domain. Random jitter manifests itself as a continuous function. The exact shape of this response is typical of a PLL phase noise plot as will be explained in Sect. 6.3.

A. Fahim, *Radio Frequency Integrated Circuit Design for Cognitive Radio Systems,*
DOI 10.1007/978-3-319-11011-0_6

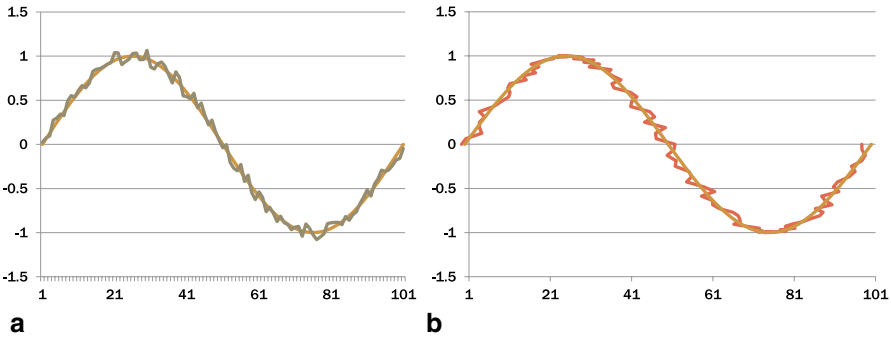


Fig. 6.1 Illustration of sine wave corrupted by **a** amplitude noise and **b** phase noise

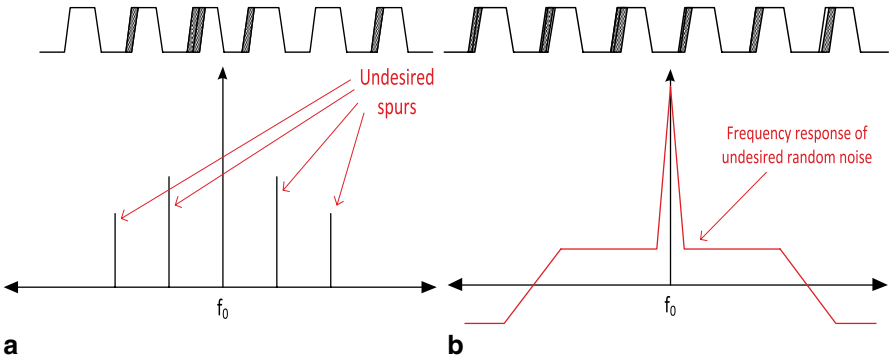


Fig. 6.2 Time domain and frequency domain representation of **a** deterministic jitter and **b** random jitter

There are several definitions of jitter [1]. *Absolute jitter* is the difference between the minimum clock period corrupted by noise from the ideal clock period. This definition is useful for digital application that requires timing closure. The *RMS jitter* is the root mean square summation of all the phase deviations that occur in each clock cycle. This measure is useful for quantifying random jitter. *Peak-to-peak jitter* is the difference between the maximum clock period and minimum clock period; useful when quantifying deterministic jitter. *Cycle-to-cycle jitter* is the clock variation between two clock edges. This metric is useful in serial links applications.

There are many sources of jitter. Some noise sources are generated the by PLL itself. This category of noise sources is referred to as *intrinsic* noise sources. The voltage and current noise sources of the individual components of the PLL (transistors and resistors, for example) give rise to amplitude noise. This amplitude noise undergoes an amplitude-to-phase (AM–PM) conversion of noise. One such example of AM–PM conversion of noise in digital logic (such as frequency dividers) is the altering of zero crossings by amplitude shifts as illustrated in Fig. 6.3. In this case, amplitude noise occurring during the waveform transitions is translated into jitter. There are other mechanisms that translate amplitude noise into jitter as will be demonstrated later.

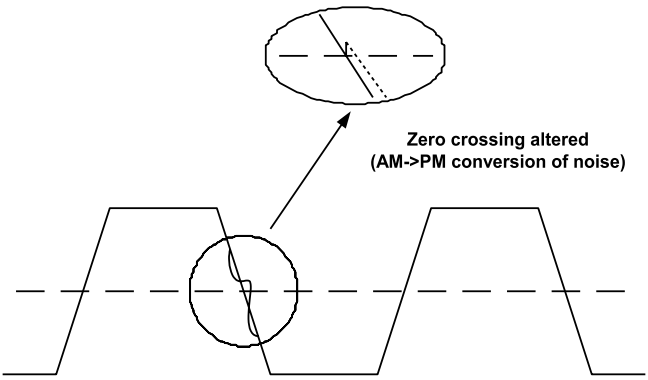


Fig. 6.3 Conversion of AM–PM noise by altering zero crossing time

The other category of noise is due to environmental disturbances around the PLL. This is known as *extrinsic* noise. Power supply and ground bounces that shift the internal bias operating points or digital circuit's zero crossing point can cause phase noise. Crosstalk and noise pickup from adjacent circuitry can induce jitter. As with the case of intrinsic sources of noise, external noise source also undergo an AM–PM conversion of noise which results in jitter. External noise sources can be mitigated by improving the power supply rejection of the PLL power supplies as well as proper layout shielding.

The relationship between jitter and phase noise is important to understand. A sine wave corrupted by jitter can be given as

$$v(t) = V \bullet \cos[\omega_0 t + n_2(t)] \tag{6.1}$$

where ω_0 is the angular frequency of the LO signal. Now, consider the special case where $n_2(t)$ is a single tone expressed as

$$n_2(t) = \frac{\Delta f}{f_m} \sin \omega_m t = \theta_p \bullet \sin \omega_m t \tag{6.2}$$

where θ_p is the frequency modulation index and ω_m is the angular frequency of the tone considered for this analysis. Substituting (6.2) into (6.1) and expanding yields

$$v(t) = V \cdot \left\{ \cos \omega_0 t \cos\left[\theta_p \sin \omega_m t\right] - \sin \omega_0 t \sin\left[\theta_p \sin \omega_m t\right] \right\} \tag{6.3}$$

If it is assumed that $\theta_p << 1$ (i.e., small phase noise), then

$$\cos\left(\theta_p \sin \omega_m t\right) \approx 1 \tag{6.4}$$

and

$$\sin[\theta_p \sin \omega_m t] \approx \theta_p \sin \omega_m t \tag{6.5}$$

therefore, substituting (6.4) and (6.5) into (6.3) and expanding yields

$$v(t) = V \cdot \left\{ \cos\omega_0 t - \frac{\theta_p}{2}\left[\cos(\omega_0 - \omega_m)t - \cos(\omega_0 + \omega_m)t\right]\right\} \tag{6.6}$$

The analysis resulting in (6.6) shows that a frequency-modulated single-tone jitter source results in a pair of tones in the frequency domain centered around the LO frequency with amplitude $^1/_2 \cdot V \cdot \theta_p$. The single sideband phase noise can now be expressed as

$$L(f_m) = \left(\frac{v_n}{V}\right)^2 = \frac{\theta_p^2}{4} = \frac{\theta_{rms}^2}{2} \tag{6.7}$$

The above analysis shows the relationship of a single tone in the phase noise plot to the jitter. In reality, the phase noise plot is a continuous spectrum, as shown in Fig. 6.2b. The tonal analysis resulting in (6.6) can be considered to be the impact of phase noise over a 1-Hz bandwidth. The phase noise plot can be integrated over the desired bandwidth to yield value for the jitter. Since, by definition, jitter is a statistical value and it is assumed that jitter from a phase noise plot is the result of random noise, the summation must be done in a root-mean-square fashion. This means that the rms jitter (in units of seconds), $t_{j,\,rms}$, is given as

$$t_{j,rms} = \frac{1}{2\pi f_0}\sqrt{2\int_{f_1}^{f_2} L(f)\,df} \tag{6.8}$$

6.2 Requirements

The phase noise of the PLL is an important parameter in any wireless transceiver. It serves as the LO port in both the downconverter mixer or upconverter mixers, affecting the performance of both the transmit and receive paths. The PLL is also used as a sampling clock to both the analog-to-digital converter (ADC) in the receive path and the digital-to-analog converter (DAC) in the transmit path, affecting the performance of both paths.

There are two phenomena that determine the phase noise requirement of the PLL when it is used as an LO signal. The first of these is known as *reciprocal mixing* [2]. It has been demonstrated in Chap. 3 that it is possible for two blockers to cross-modulate and produce an undesired signal that lands on the receiver channel. Since there is a finite amount of phase noise at any frequency offset away from the LO frequency, there will be a component of the LO phase noise that mixes with a large blocker back into the desired receive channel. This noise translation may limit the noise floor of the receiver if the LO phase noise at a certain offset frequency is too high.

The reciprocal mixing effect is shown graphically in Fig. 6.4. The phase noise profile is simplified as a triangular shape centered around LO. An IF receiver is assumed in this plot, meaning that the desired channel is offset away from the LO

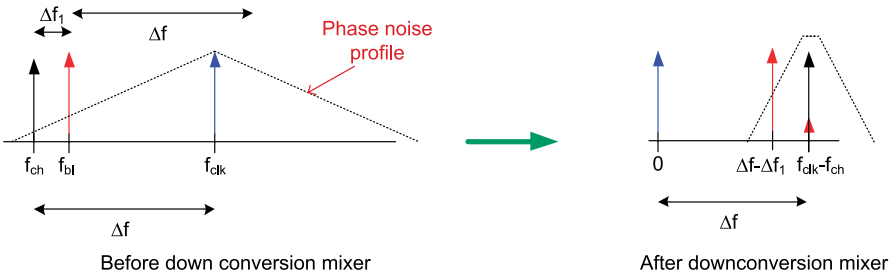


Fig. 6.4 The reciprocal mixing effect in a receiver

frequency by an amount equal to IF, or Δf as shown in Fig. 6.4. An undesired blocker, located at f_{bl} mixes with the LO phase noise located at Δf away from the blocker to downconvert a portion of the blocker set by the phase noise level. In order for the phase noise to not cause any degradation, the maximum phase noise level must be

$$PN < P_{ch} - P_{bl} - SNR_{desired} - margin - 10\log_{10} BW \text{ (dB)} \tag{6.9}$$

where P_{ch} is the channel power level, P_{bl} is the blocker power level, $SNR_{desired}$ is the SNR required by the standard, the margin is usually set to 10 dB, and BW is the bandwidth (in Hertz). The phase noise of (6.9) is expressed in dBc, relative to the amplitude of the LO signal. The phase noise is specified at a frequency offset, Δf_1, away from the LO center frequency. All blocker profiles must be evaluated to ensure that the worst case phase noise is specified and met.

Another transceiver requirement that places a specification on the phase noise requirement on the PLL is that of the error vector magnitude (EVM) [3]. As stated in Chap. 5, the EVM can be stated as a percentage. The EVM can be a result of many sources, one being the phase noise of the LO. As a result, the EVM specification determines the rms jitter contribution of the PLL, with the phase noise integration band being the bandwidth of the output signal to be transmitted.

The PLL can also be used as a sampling clock for the data converters in the wireless transceiver. The jitter requirement in this case is called *aperture jitter* [4]. This jitter is also an rms jitter than is integrated over the bandwidth of the output signal to be transmitted. The jitter requirement is given as

$$t_{j,rms} = \frac{10^{-\frac{SNR}{20}}}{2\pi f_{BW}} \tag{6.10}$$

where f_{BW} is the signal bandwidth, SNR is the desired signal-to-noise ratio in dB. The aperture jitter requirement can be derived from analyzing how much amplitude variation is due to clock phase variation (PM–AM conversion), as shown in Fig. 6.5.

In the case of a radio frequency digital-to-analog converter (RFDAC), the SNR requirement is driven by the adjacent channel power ratio (ACPR) specification.

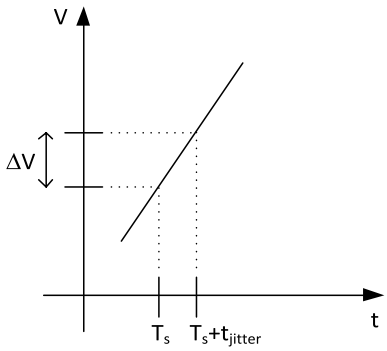


Fig. 6.5 Jitter resulting in voltage variation in a data converter

The SNR requirement for RFDAC may be in excess of 70 dB. Figure 6.6 shows the rms jitter requirement on the sampling clock. As the signal bandwidth approaches 100 MHz, the jitter requirement approaches the 100 fs barrier. The aperture jitter value, however, is a worst case analysis. In reality, the jitter number can be substantially relaxed. Although the worst case bandwidth is used in (6.10), in reality small amount of signal power may actually be present at higher frequencies of the data signal. This fact has been used extensively in wide data bandwidth applications to ease the jitter specification [5]. Realistic jitter specifications are usually obtained through simulations with typical data patterns.

6.3 Phase-Locked Loops (PLL) Primer

The most popular method of generating an on-chip programmable frequency is through the use of a PLL. The PLL frequency synthesis is a form of indirect frequency synthesis, where a low-frequency spectrally pure clock source is used to generate an intermediate voltage than a higher-frequency clock source. The generated clock source is frequency and phase locked to the low-frequency clock source to maintain its spectral purity.

6.3.1 Integer PLL

The most popular form of PLLs used today is the single-loop charge-pump-based PLL [6], shown in Fig. 6.7. The on-chip clock source is generated by a voltage-controlled oscillator (VCO). A VCO produces an output frequency that is proportional to its input voltage level. The VCO output signal is frequency divided by a feedback divider, L, and brought down to the same frequency as the external low-frequency spectrally pure clock source. The phases of the two signals are compared by a phase-frequency detector (PFD) that generates a pair of signals, UP and DN. The UP and DN signals then turn on either a positive or negative current, respec-

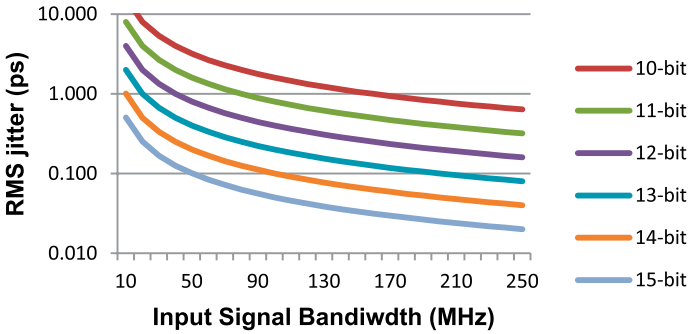


Fig. 6.6 RMS jitter requirement versus input signal bandwidth and SNR

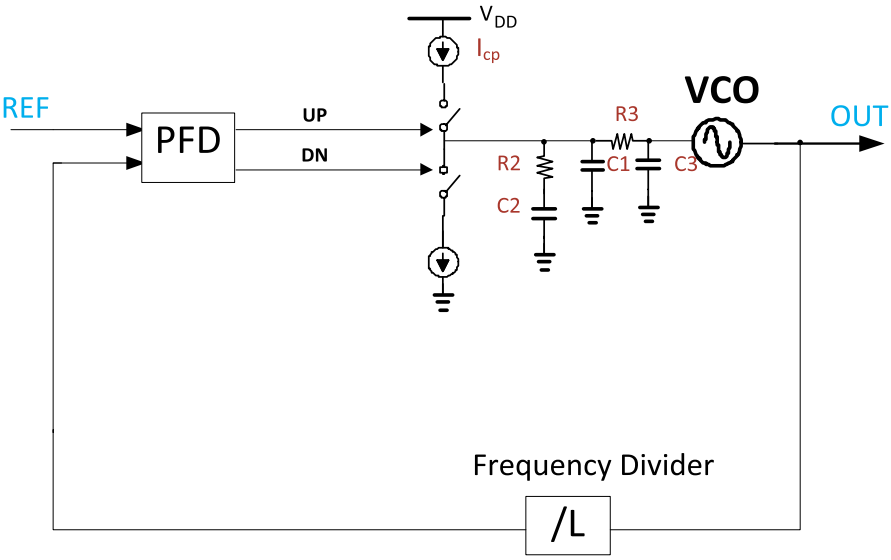


Fig. 6.7 An charge-pump integer-N PLL

tively, for a brief amount of time to adjust the passive loop filter's voltage. The loop filter filters the voltage excursions caused by abruptly turning the current sources on and off and loop filter output voltage is then used to control the frequency of the VCO directly.

When the PLL is near lock, the phase deviations detected by the PFD are small and the loop can be linearized. Such a linear model is called the linear phase model of the PLL, where input and output phases, not frequency, are observed. The phase transfer function from input REF to the PLL output can be given as

$$H(s) = \frac{\left(\frac{I_p}{2\pi}\right)(2\pi K_{vco} / s)F(s)}{1 + \left(\frac{I_p}{2\pi}\right)(2\pi K_{vco} / s)F(s) / L} \tag{6.11}$$

where the linear model for the PFD and charge pump is $K_d = \frac{I_p}{2\pi}$ in units of A/rad. The linear model for the VCO gain is $K_v = 2\pi K_{vco} / s$ in units of rad/V. Note that an integrating 1/s in K_v is necessary to translate the output frequency of the VCO into phase. F(s) is the linear transfer function of the loop filter, which is given as

$$F(s) = \frac{s + \frac{1}{R_2 C_2}}{C_1 R_3 C_3 s \left[s + \frac{1}{R_2 C_1}\left(1 + \frac{C_1}{C_2}\right)\right]\left[s + \frac{1}{R_3 C_3}\right]} \tag{6.12}$$

As implied by Fig. 6.7 and (6.11) and (6.12), the loop contains two integrators: one from the VCO and another from the loop filter. This means that the loop is able to correct for any phase or frequency steps with no residual error. The H(s) transfer function as shown in (6.11) is a low-pass filter response.

The additional poles introduced by C1 and the R3–C3 pair serve to remove any voltage ripples, as will be seen in Sect. 6.4. These poles are usually spaced a decade apart in order to avoid stability issues [7]. Without these branches, (6.11) becomes a second-order system. Reducing the system down to a second-order system has the advantage of providing closed-form design equations that can be used to optimize the PLL parameters. Two parameters that are used to describe a second-order system are the natural frequency and damping ratio, which for a second-order linear PLL are given as

$$\omega_n = \sqrt{\frac{K_{vco} I_p}{NC}} \tag{6.13}$$

and

$$\zeta = \frac{\omega_n RC}{2}, \tag{6.14}$$

respectively. The resulting bandwidth of the PLL (ignoring the higher order poles) becomes

$$f_{3dB} = \frac{\omega_n}{2\pi}\sqrt{2\zeta^2 + 1 + \sqrt{\left(2\zeta^2 + 1\right)^2 + 1}} \text{ (Hz)} \tag{6.15}$$

Using (6.14), (6.15) can be approximated as

$$f_{3dB} \cong \frac{1}{2\pi}\frac{K_{vco} I_p R}{N} \text{ (Hz)} \tag{6.16}$$

Another important point to note is that the PLL is a time-sampled system. More specifically, the PFD updates the loop once every reference cycle. The PLL, however, can still be treated as a continuous time system if the PFD update rate far exceeds

to the loop bandwidth of the PLL. A reference frequency to PLL bandwidth ratio of greater than 10 is usually recommended.

6.3.2 ΣΔ Fractional-N PLL

The integer PLL described above is capable of producing any output frequency that is a multiple of the input frequency. The ratio of output frequency to input frequency is L. If L is a programmable parameter, then the PLL can, in theory, produce any integer multiple of the frequency of the input clock source, REF. In direct conversion receivers, if the LO is set to the desired RF frequency, then minimum step size of the LO is equal to one channel frequency. This presents two challenges. First, the input frequency to the PLL is restricted to be one channel wide. This can severely limit the performance of the PLL especially for narrow bandwidth transmissions, as is shown in Sect. 6.4. Second, the RF transceiver may be required to support multiple standards, making it nearly impossible to synthesize any output frequency from a single crystal source.

One solution to providing LO support to multiple standards is to construct a PLL that is capable of generating any output frequency, irrespective of the input frequency. One method of achieving this goal is to replace the integer-L feedback divider with a fractional divider. A fractional divider is actually an integer divider, which has a dynamic division ratio that toggles between two (or more) integer values. The toggling is performed in such a way that the average division ratio is equal to the desired fractional ratio. Consider, for example, a case where a fractional division ratio of L = 63.25 is desired. The feedback divider value will be set to {64,63,63,63} for four consecutive cycles and the sequence is repeated. The averaging operation is approximated by the low-pass response of the PLL. In general, if the denominator of the fractional ratio is equal to 2^N, then the PLL closed loop bandwidth must be set to less than $F_{REF}/2^N$ to filter out the spurious response of the repeated sequence. Figure 6.8 shows a simulation plot of the output spectrum of PLL centered around 2.5075 GHz. The fractional ratio is equal to $3/2^3 = 3/8$ and the reference frequency

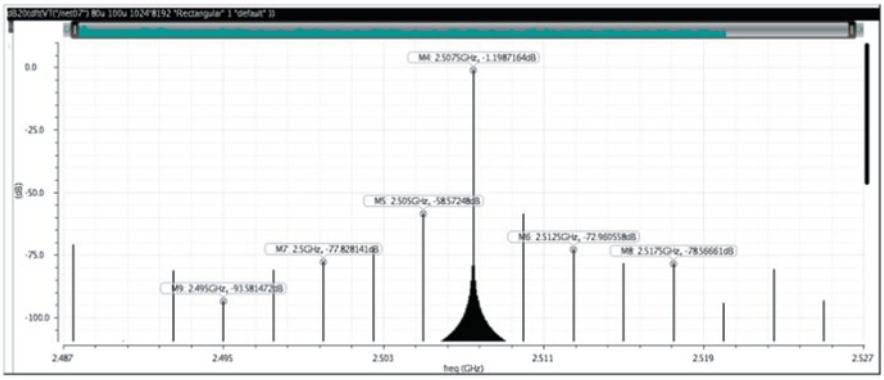


Fig. 6.8 Output spectrum of a fractional-N PLL

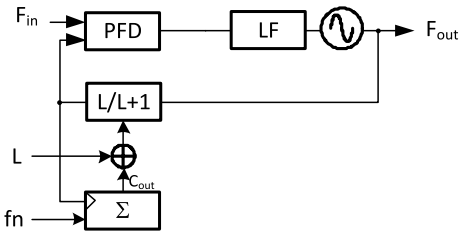


Fig. 6.9 Basic fractional-N PLL architecture

is 20 MHz. As the figure shows fractional spurs occur at 2.5 MHz intervals apart, it corresponds to $1/2^3$ times the reference frequency. In this case, the denominator of the fractional ratio is small. In a more practical scenario, the value of N can easily be 20–30. This results in fractional spurs occurring much closer in frequency to the LO carrier signal.

The implementation of a basic fractional-N divider is fairly straightforward. Figure 6.9 shows a fractional-N PLL with a fractional divider [8]. The division ratio is dithered between two values: L and L+1. The control signal to change the division ratio comes from the "carry out" port of an accumulator. If the bit width of the accumulator is N, the denominator of the fractional value is 2^N. The numerator is set by an external control word, fn. The input clock signal for the accumulator is from the output of the feedback divider itself (as opposed to the REF signal directly). This is done in order to synchronize the modification of the feedback division ratio with the feedback divider itself. This is crucial in order to avoid metastability errors in the feedback divider.

An example of how the accumulator control the feedback division ratio is shown in Fig. 6.10. In this example, 3-bit accumulator is assumed and fn=3. In this case, the desired fractional value is 3/8. As the figure shows, as the clock is toggled, the contents of the accumulator are incremented in a modulo fashion. When the accumulator overflows, the Carry Out signal is asserted incrementing the feedback division ratio by 1. The sequence shown is repeated indefinitely with repetition length of 8 cycles. This gives rise to the fractional spurs appearing at intervals of $f_{REF}/8$. The average division ratio then becomes L+3/8.

As stated earlier, the fractional spurs can occur at very low offset frequencies from the carrier. Lowering the bandwidth to filter out the fractional spurs would be impractical and the resulting jitter performance of the PLL would be poor. An alternative method of generating a fractional divider is to use a sigma–delta (ΣΔ) modulator instead of an accumulator, as shown in Fig. 6.11 [9]. Comparing Fig. 6.9

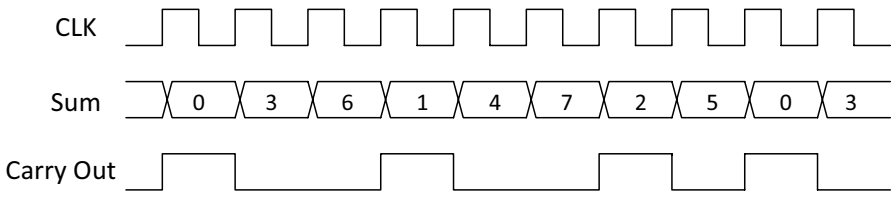


Fig. 6.10 Typical waveforms for a fractional ratio of fn=3/8

to Fig. 6.11 shows that the L divider has been replaced by a multi-modulus divider (MMD). From a functional point of view, the main difference between an L divider and an MMD is that the MMD division ratio is dithered over a wider range. From an implementation point of view, the dividers are identical.

A sigma–delta modulator randomizes the fractional sequence shown in Fig. 6.10. This randomization is done in such a way that the spurs are translated into high-frequency noise. The low-pass filter closed-loop response of the PLL then filters the high-frequency noise. To understand how this operates, consider a generic first-order sigma–delta modulator, shown in Fig. 6.12. The quantizer would be equivalent to the Carry Out signal shown in Fig. 6.10. Obtaining a closed-form expression for this loop is difficult due to the nonlinear quantizer. If it is assumed that there is sufficient activity at the output of the integrator, the toggling of the quantizer output can be approximated as a random process. In this case, the quantization error can be assumed to be a random additive error to the integrator output. This results in the linear model shown in Fig. 6.13. The term K is a gain factor associated with the integrator.

The transfer function from x(t) to y(t) is called the signal-transfer function (STF) and is given as

$$STF = \frac{K}{s+K} \tag{6.17}$$

and the transfer function from the quantization noise, $n_q(t)$ to the output is known as the noise-transfer function (NTF) and is given by

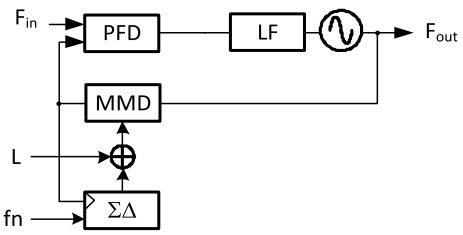


Fig. 6.11 ΣΔ fractional-N PLL architecture

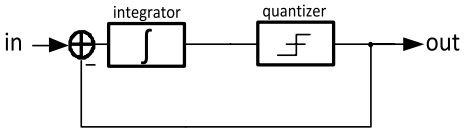


Fig. 6.12 First-order sigma–delta modulator

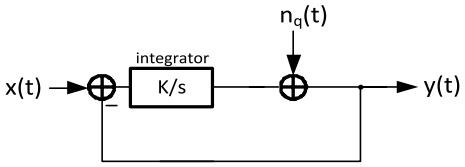


Fig. 6.13 Linear model of a sigma–delta modulator

$$NTF = \frac{s}{s+K} \tag{6.18}$$

Analysis of (6.17) and (6.18) reveals an interesting fact. The frequency response of the STF is a low-pass filter, with a passband gain of 0 dB and cutoff frequency of K rad/s. The frequency response of the NTF is a high-pass filter with a passband gain of 0 dB and cutoff frequency of K rad/s. Since the final output y(t) is the sum of the STF and NTF, the high-frequency content of y(t) is the undesired quantization noise; whereas the low-frequency content of y(t) is the desired signal. For this reason, a sigma–delta modulator is usually followed by a low-pass filter in order to filter out the undesired quantization noise.

As stated earlier, the quantization noise is considered to be a random process that is uniformly distributed. Since the probability density function is uniformly distributed from $-\Delta/2$ to $\Delta/2$ (where Δ is the quantizer step size), summing the pdf over this interval yields the quantization noise level, which is given as

$$q_n^2 = \frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2} x^2 dx = \frac{\Delta^2}{12} \tag{6.19}$$

A popular method of generating a high-resolution sigma–delta modulator is to cascade several first-order modulators together, as shown in Fig. 6.14. The configuration in Fig. 6.14 is known as a MASH111 sigma–delta modulator [10] and is a cascade of three first-order modulators. Several stages are cascaded together to

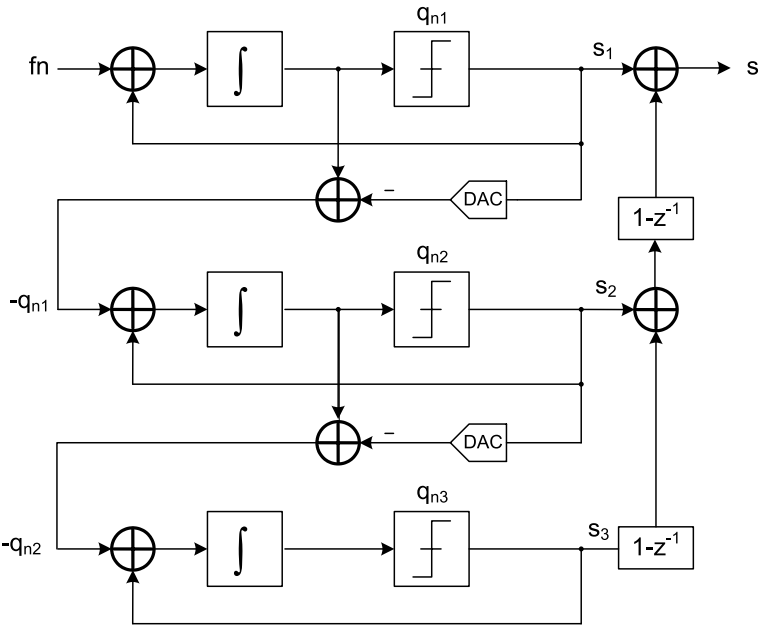


Fig. 6.14 A MASH111 sigma–delta modulator

improve the quantization noise floor of the sigma–delta modulator. The input *fn* is the input analog signal to the ΣΔ modulator. s_1 is the digital output of the first stage modulator and can be given as

$$s_1 = fn + q_{n1}\left(1 - z^{-1}\right) \tag{6.20}$$

where q_{n1} is the quantization noise and is shaped by a digital high-pass filter $1-z^{-1}$. The quantization error, q_{n1}, can be extracted by subtracting the output of the quantizer from its input. Since the output is in digital form, it must first be translated into an analog signal using a 1-bit digital-to-analog converter (DAC), as shown in Fig. 6.14. The quantization error from the first stage is now the input to the second stage modulator. The output of the second stage is given as

$$s_2 = q_{n1} + q_{n2}\left(1 - z^{-1}\right) \tag{6.21}$$

A similar expression is derived for s_3, the output of the third modulator stage. The output of each modulator stage is first differentiated then added to the stage preceding it. The output can then be expressed as

$$\begin{aligned} s = {} & fn + q_{n1}\left(1 - z^{-1}\right) + \left(1 - z^{-1}\right) \\ & \left[-q_{n1} + q_{n2}\left(1 - z^{-1}\right) + \left(1 - z^{-1}\right)\left[-q_{n2} + q_{n3}\left(1 - z^{-1}\right)\right]\right] \end{aligned} \tag{6.22}$$

which can be simplified to

$$s = fn + q_{n3}\left(1 - z^{-1}\right)^3 \tag{6.23}$$

As (6.23) shows, what effectively was done with a MASH111 sigma–delta modulator is that the quantization noise is now shaped by a third-order digital high-pass filter. This is a general result, where an nth order sigma–delta modulator would have n^{th} order shaping of the quantization noise. The quantization noise of a second-order modulator is compared to a third order modulator in Fig. 6.15. As the figure shows, the low-frequency noise is significantly less for a third order modulator.

In the context of a sigma–delta PLL, the input signal *fn* is a DC signal representing the desired fractional frequency value. The low-pass filter following the sigma–delta modulator is the low-pass closed loop response of the PLL. The quantization noise is the dithering of the feedback division ratio. The step size of the quantizer is equal to a VCO period.

Figure 6.16 shows a schematic diagram of a digital sigma–delta modulator. Each accumulator represents an integrator stage. The "Carry Out" is equivalent to the output of the quantizer. The quantization error is then the difference between the accumulated value minus the "Carry Out" bit, if the "Carry Out" bit is considered to be the MSB of the accumulated value. In other words, if the accumulated value is represented as ACC, the quantization error is

$$e_{q,i} = ACC_i - y_i = -SUM_i \tag{6.24}$$

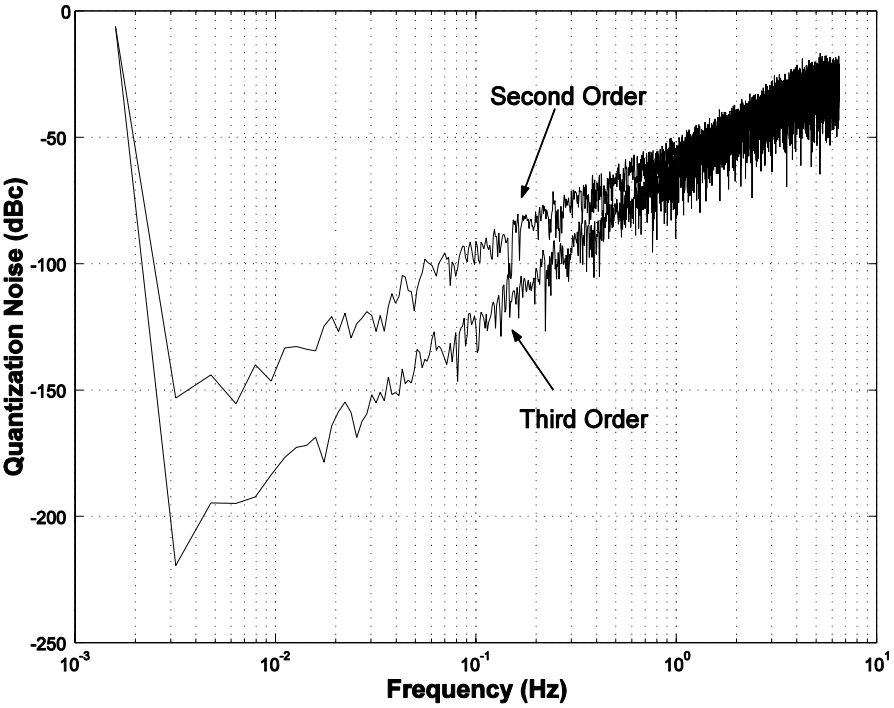


Fig. 6.15 Comparison of noise shaping between a second-order and third-order sigma–delta modulator.

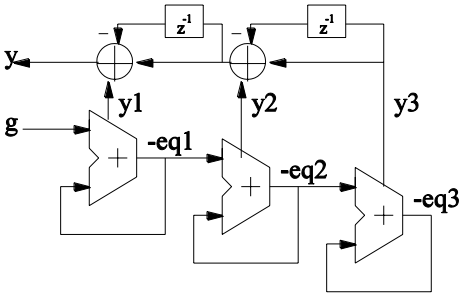


Fig. 6.16 A digital sigma–delta modulator

where SUM_i is the sum value of the accumulator of the ith accumulator. In other words, a digital sigma–delta modulator does not require an explicit DAC or analog subtractor as was needed in the analog equivalent of Fig. 6.14. Instead, the SUM output of one accumulator (which is equal to $-e_q$ as per (6.24)) can be simply fed into the input of the next accumulator. As in the analog equivalent of Fig. 6.14, the "Carry Out" terminals of each accumulator stage are first digitally differentiated by a $1-z^{-1}$ operator before adding to the previous stage, resulting in the expression given by (6.23).

One consequence of using a MASH type of modulator is that the output dithering range of the modulator increases with each cascaded stage. More specifically, this can be seen by analyzing Fig. 6.16 more carefully. If the range of the binary y_i outputs is [0,1], the output of the first differentiator associated with the y_3 output is [−1,0,1]. When added to y_2, the dynamic range becomes [−1,0,1,2]. When this output is differentiated, the output dynamic range becomes [−3,−2,−1,0,1,2,3]. Finally, when adding the y_1 output, the final dynamic range is the integer range of [−3,+4], requiring a 3-bit output.

There are wide variety of other higher-order sigma–delta topologies that employ feedforward and feedback techniques. One motivation of using such techniques is to reduce the output dithering range of the sigma–delta modulator to one bit. Table 6.1 summarizes the trade-off between single bit and multibit MASH sigma–delta modulators.

As stated earlier, a sigma–delta modulator dithers the division ratio of the feedback divider. The frequency divider is usually modeled as a 1/L scaling factor in the phase domain model of the PLL. This is, however, true if the input is from the VCO. The phase domain model from the sigma–delta modulator to the frequency divider output is different. The phase deviation at the output of the frequency divider depends on the sum of all previous phase deviations that the frequency divider experienced. This is because the new L count of the feedback frequency divider does not start until the previous count has completed. In other words, the phase output of the frequency divider depends on the sum of all the phases of all the previous divider counts plus the VCO phase. In the case of an integer PLL, the phase deviation at the output of the divider is proportional to $k \bullet L + \theta_{vco}$, or $k + \theta_{vco} / L$ when referred to the PLL output. In the case of a sigma–delta modulated frequency divider, the phase error (normalized to the PLL output) is equal to [11].

$$\theta_{div}[k] = \frac{1}{L_{nom}}\left(-2\pi\sum\nolimits_{m-1}^{k} n[m-1] + \theta_{vco}[k]\right) \tag{6.25}$$

where n[m−1] is the frequency divider value at time step m−1, which is equal to L plus the output of the sigma–delta modulator, and L_{nom} is the average division ratio which is L plus fn. What (6.25) reveals is that the output of the sigma–delta modulator is summed, or integrated, before appearing as phase at the frequency divider output. In other words, in the z-domain, the phase output of the frequency divider (referred to the PLL output) is given as

Table 6.1 Comparison of multibit and single bit ΣΔ modulators

Multibit ΣΔ	Single bit ΣΔ
Unconditionally stable	Stable for only a range of inputs
Better in-band performance	Lower out-of-band performance
Modular topology	Hardware coefficients hand crafted

$$\theta_{div}[k] = \frac{2\pi}{L_{nom}} \frac{z^{-1}}{1 - z^{-1}} \tag{6.26}$$

Another way of validating the presence of an integration term as shown in (6.25) and (6.26) is to examine the units at each point. The input to the sigma–delta modulator is the desired fractional frequency. Since the PLL model is a phase domain model, the output of the sigma–delta modulator must undergo an integration operation in order to be compatible with the unit at the output of the frequency divider, which is phase.

One final word regarding this point is with regard to using a PLL as a phase path in a polar modulator as was shown in Fig. 5.9. The I and Q signals at baseband were converted into polar form: amplitude and phase. The digital phase word produced by the I/Q-to-polar converter would then be used to modulate the feedback division ratio in a PLL. As was shown by (6.26), the phase would undergo an integration operation and would lead to incorrect phase modulation of the signal. To counteract this effect, the digital phase word must undergo a digital differentiation operation, $1-z^{-1}$, as shown in Fig. 5.9. After the integration operation of the feedback divider, this would reproduce the phase signal back into the PLL loop, and the phase is correctly modulated.

6.4 PLL Phase Noise Optimization

One critical metric in PLLs is its phase noise performance, or jitter. The jitter performance of the PLL is the weighted sum of the jitter performance of all the various components of the PLL. Figure 6.17 shows a linear model of a PLL showing the jitter contribution of each PLL. As the figure shows, the jitter component of each block is modeled as additive jitter added to the output of each block. Transfer function from each noise source to the output of the PLL is known as the *jitter transfer function*. Note that the noise contribution of the crystal oscillator (XO) buffer is the same as the linear transfer function of the PLL, which was given by (6.11). The only noise contribution of the sigma–delta modulator is assumed to be the frequency-

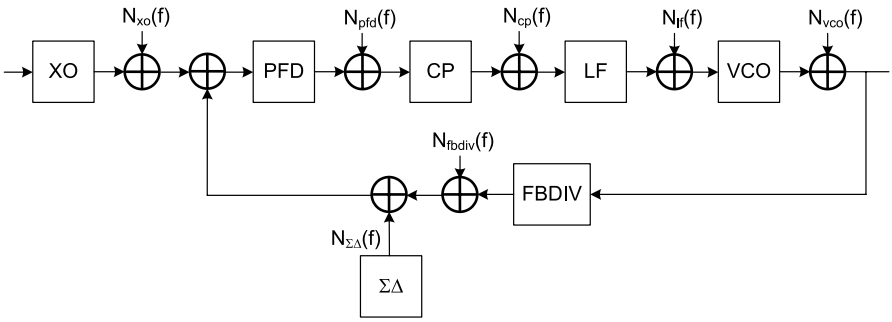


Fig. 6.17 Linear model of PLL showing jitter contributions

shaped quantization noise, the rest are assumed to be composed of thermal and flicker noise components.

The noise-transfer functions illustrated in Fig. 6.17 are given by (6.27)–(6.33).

$$H_1(s) = H_{xo}(s) = \frac{I_p K_{vco} F(s)}{s + I_p K_{vco} F(s)/L} \tag{6.27}$$

$$H_2(s) = H_{pfd}(s) = \frac{2\pi I_p K_{vco} F(s)}{s + I_p K_{vco} F(s)/L} \tag{6.28}$$

$$H_3(s) = H_{cp}(s) = \frac{2\pi K_{vco} F(s)}{s + I_p K_{vco} F(s)/L} \tag{6.29}$$

$$H_4(s) = H_{lf}(s) = \frac{I_p K_{vco}}{s + I_p K_{vco} F(s)/L} \tag{6.30}$$

$$H_5(s) = H_{vco}(s) = \frac{s}{s + I_p K_{vco} F(s)/L} \tag{6.31}$$

$$H_6(s) = H_{fbdiv}(s) = \frac{I_p K_{vco} F(s)}{s + I_p K_{vco} F(s)/L} \tag{6.32}$$

$$H_7(s) = H_{\Sigma\Delta}(s) = \frac{I_p K_{vco} F(s)}{s + I_p K_{vco} F(s)/L} \tag{6.33}$$

As (6.27)–(6.33) show, the filtering response of the PLL to noise injection differs depending on where the noise is injected. Noise generated by the sigma–delta modulator, feedback divider, crystal oscillator buffer, and PFD and charge pump are all low-pass filtered. These noise contributors are often referred to as in-band noise contributors. Note that the magnitude of the various low-pass filters vary. Noise injected into the loop filter is bandpass filtered, whereas noise generated by the VCO is high-pass filtered. Also, note that the noise-transfer function of the feedback divider and the crystal oscillator buffer are the same.

Special attention must be paid to the noise-transfer function of the sigma–delta modulator. Equation (6.33) shows the transfer function of the quantization noise shaped by the sigma–delta modulator and the feedback divider integration operation as it appears at the final PLL output. This means that (6.33) can be expanded into

$$H_{\Sigma\Delta}(s) = \frac{I_p K_{vco} F(s)}{s + I_p K_{vco} F(s)/L} \cdot \frac{z^{-1}}{1 - z^{-1}} \cdot \left(1 - z^{-1}\right)^m \tag{6.34}$$

Considering the magnitude of (6.34), it can be simplified to

$$H_{\Sigma\Delta}(f) = (2\pi)\left|\frac{I_p K_{vco} F(j2\pi f)}{j2\pi f + I_p K_{vco} F(j2\pi f)/L}\right| \cdot \left|2\sin\left(\frac{\pi f}{F_{ref}}\right)\right|^{m-1} \tag{6.35}$$

where F_{ref} is the reference frequency of the PLL. Since the quantization noise of the sigma–delta modulator is assumed to be uniformly distributed over F_{ref}, its phase noise contribution as seen at the output is given as

$$\theta^2_{n,\Sigma\Delta}(f) = \frac{(2\pi)^2}{12F_{ref}} \cdot \left|\frac{I_p K_{vco} F(j2\pi f)}{j2\pi f + I_p K_{vco} F(j2\pi f)/L}\right|^2 \cdot \left|2\sin\left(\frac{\pi f}{F_{ref}}\right)\right|^{2(m-1)} \tag{6.36}$$

The varying filtering characteristics of the PLL noise-transfer functions leads to an optimal point for choosing a closed-loop bandwidth for minimum PLL integrated jitter. More specifically, a bandwidth is chosen such that the noise contributors of the in-band noise sources are equal to that of the VCO noise (and the ΣΔ modulator in the case of a fractional-N PLL). The noise contribution of the resistor noise in the loop filter is assumed to be small in this analysis. Mathematically stated, the PLL bandwidth optimization problem then becomes selecting the PLL loop parameters (namely charge pump current, VCO gain and loop filter components) such that the PLL integrated noise is minimized. The PLL integrated phase noise function is given as

$$\theta^2_{rms}(f) = \sum_{i=1}^{7} \int_{f_1}^{f_2} N_i^2(f) \cdot |H_i(f)|^2 \, df \tag{6.37}$$

where $H_i(f)$ is given by (6.27)–(6.33) and $N_i^2(f)$ is the power spectral density of the noise of the ith component in the loop, f_1 and f_2 are the integration filter ranges (f_1 is the minimum frequency offset from the carrier and f_2 is the maximum frequency offset over which phase noise is a concern).

The choice of PLL bandwidth can now be viewed as an optimization problem where (6.37) is minimized given constraints on the ranges of the PLL parameters: K_v, I_{cp} and loop filter values. Figure 6.18 shows the normalized optimal PLL bandwidth

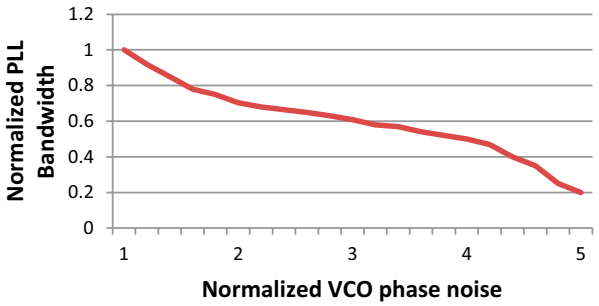


Fig. 6.18 Optimal PLL bandwidth as in-band and VCO noise are varied

as the normalized VCO phase noise is increased for an integer PLL. As the figure shows, the optimal PLL bandwidth shrinks as the VCO phase noise is increased.

6.5 Charge Pump Circuit Implementation

Phase-Frequency Detector (PFD) Two important components in an analog PLL is the phase-frequency detector (PFD) and charge pump. A conventional PFD is shown in Fig. 6.19 [12]. The PFD operates as a three-state machine. Although there are two digital outputs (UP and DN) and there are theoretically four states, the fourth state (UP = DN = 1) is disallowed by the feedback AND gate. The feedback AND gate detects this condition and resets both flip-flops to the zero state. The R signal is the signal from the clock reference source (typically a crystal oscillator buffer) and the V signal is the divided down VCO signal from the feedback divider. If the R signal leads the V signal, this means that the VCO phase is too slow and the PFD instructs the VCO to advance its phase by asserting the UP signal. On the other hand, if the V signal leads the R signal, this means that the VCO phase is faster than the reference phase and the PFD instructs the VCO to retard its phase by asserting the DN signal. When both the R and V signals are in phase, the UP and DN signals remain in the zero-state.

One issue with the three-state phase detector is that it suffers from what is known as the dead zone issue [13]. Consider the locked condition where the R signal is equal in phase and frequency to the V signal. It is possible, in this case, that the path formed by the feedback AND gate along with the reset delay in the flip-flops is faster than the time required to fully settle the phase detector, causing partial toggling of the UP and DN signals. In this case, the effective charge pump current may be significantly less than what was predicted by from DC simulations. This results in a drastic reduction in closed loop bandwidth and can seriously alter the predicted overall phase noise performance. This singularity, or *dead zone,* near zero phase of the charge pump versus input phase error is illustrated graphically in Fig. 6.20.

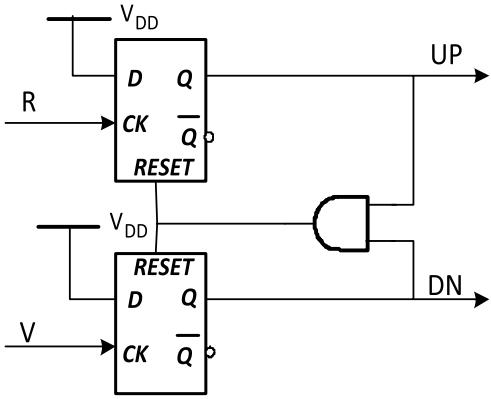


Fig. 6.19 A conventional 3-state phase-frequency detector (PFD)

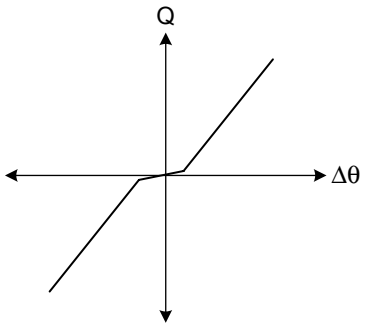


Fig. 6.20 Charge pump versus input phase error using a 3-state PFD

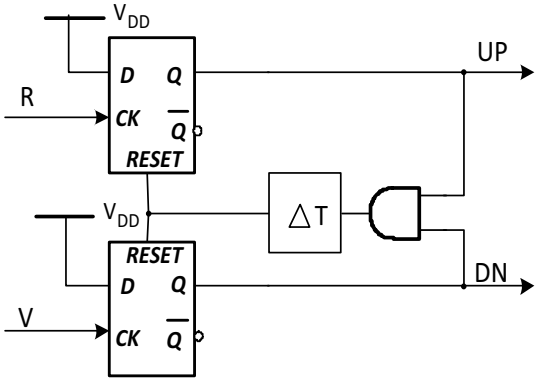


Fig. 6.21 A 4-state phase-frequency detector (PFD)

One common method of resolving the dead zone issue is to use what is known as a 4-state PFD, shown in Fig. 6.21. A fourth state is now allowed in a PFD by inserting a delay after the feedback AND gate. This fourth state allows for the UP = DN = 1 state. This state allows sufficient time for both the UP and DN currents in the charge pump to fully settle to their required values, avoiding the deadzone issue discussed above. One on hand, the delay through the AND gate should be sufficiently long to allow proper settling of the UP and DN charge pump currents. On the other hand, the delay through the AND gate should be minimized in order to reduce noise due to the charge pump current (as will be seen later). The remaining three states of the 4-state PFD shown in Fig. 6.21 are the same as the 3-state PFD shown in Fig. 6.19. The valid states of a 4-state PFD are summarized in Table 6.2.

Table 6.2 Valid states of a 4-state PFD

State	State condition	Operation
−1	DN = 1, UP = 0	VCO phase too early
0	DN = 0, UP = 0	PLL is in phase lock
1	DN = 0, UP = 1	VCO phase too late
Z	DN = 1, UP = 1	PLL phase held

Charge-Pump A basic current steering charge pump is shown in Fig. 6.22. In this type of charge pump, two current sources are used. The I_{UP} current source is used to convert positive phase error into charge integrated on the loop filter. I_{DN} current source is used to convert negative phase error into charge integrated on the loop filter. The amount of time that either I_{UP} or I_{DN} currents are integrated onto the loop filter is controlled by the M1 and M2 devices, which act as switches. Transistors M3 and M4 steer the current through an alternate branch to prevent the current sources from completely turning off. The operation of the unity gain buffer will be explained shortly.

There are a number of nonidealities that arise from the charge pump shown in Fig. 6.22. First, there may be a static current mismatch between the I_{UP} and I_{DN} currents. In order for the PLL to lock, the average charge introduced by the I_{UP} current source into the loop filter must be equal to the amount of charge introduced by the I_{DN} current source into the loop filter. In mathematical terms,

$$Q_{up} = Q_{dn} \quad \Rightarrow \quad I_{UP} \cdot \Delta t_{up} = I_{DN} \cdot \Delta t_{dn} \tag{6.38}$$

where Q_{up} is the total charge introduced by the I_{UP} current source over one reference period, Q_{dn} is the total charge introduced by the I_{DN} current source over one reference period, Δt_{up} is the pulse width of the UP signal, and Δt_{dn} is the pulse width of the DN signal. In the presence of current mismatch between I_{UP} and I_{DN} currents, a static phase offset develops in the loop to compensate for this mismatch. If the charge pump mismatch is merged into the I_{DN} current such that $I_{DN}=I+\Delta I$ and $I_{UP}=I$, the resulting static phase error between the input and output of the PLL is given as

$$\Delta t_{offset} = \frac{\Delta I}{I} t_{on} \tag{6.39}$$

where t_{on} is the minimum pulse width set the delay element in the feedback path of the PFD.

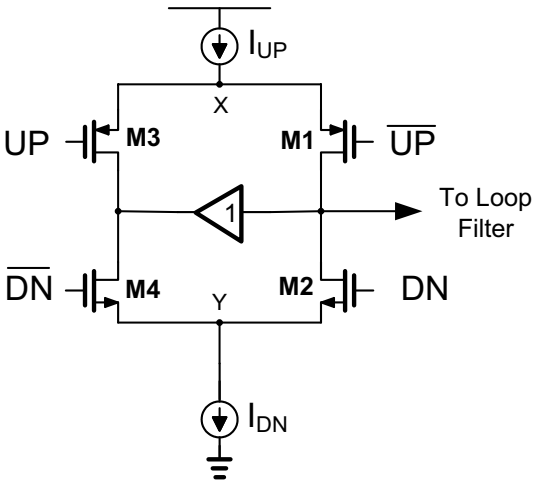


Fig. 6.22 A current steering charge pump

Another impairment of the charge pump shown in Fig. 6.22 is known as charge sharing. When either the UP or DN signal undergoes a low-to-high transition I_{UP} or I_{DN} current, respectively, it is pumped into the loop filter. At the moment before this occurs, the voltage potential at node X or Y matches that of node D (ignoring the "on" resistance of the M3 and M4 switches); whereas, the loop filter voltage at this time instant can be different from the voltage potential at node D. At the moment, when the current source is switched over, the X or Y nodes are shorted with the loop filter. Since these nodes have different voltages, charge transfer or sharing, occurs between the nodes. More specifically, if we consider charge sharing between node X and the loop filter node, F, the transferred charge is equal to

$$q_{err,cs} = C_x V_x - C_x V_f = \frac{C_x C_{LF}\left(V_x - V_{LF}\right)}{C_x C_{LF}} \tag{6.40}$$

where V_X is the voltage at node X, V_f is the final settling voltage after charge sharing between node X and the loop filter node, C_X is the parasitic capacitance at node X, C_{LF} is the total loop filter capacitance. The charge can be regarded as an input phase error independent charge error term that is integrated at every reference cycle.

One way to reduce the effect of charge sharing is to ensure that the voltage at node X or Y is equal to the loop filter voltage. This is the primary function of the unity gain buffer. The loop filter is sensed and the other node in the current steering differential pair, typically called the dummy node, is forced to be equal to the loop filter node through the unity gain buffer. The current source and sink of the buffer should be large enough to absorb the I_{DN} and I_{UP} currents, respectively.

An opamp that is typically used to implement the unity gain buffer is shown in Fig. 6.23. A rail-to-rail input folded cascode opamp is shown. The speed of the opamp should be sufficiently large that it is much larger than the closed-loop bandwidth of the PLL and can recover reasonably well within one reference cycle from any glitches that can occur on the dummy node as a result of switching activity. Also, the rail-to-rail input stage is useful for low-voltage operation. The output stage of the folded cascode structure is single-ended, as the figure shows.

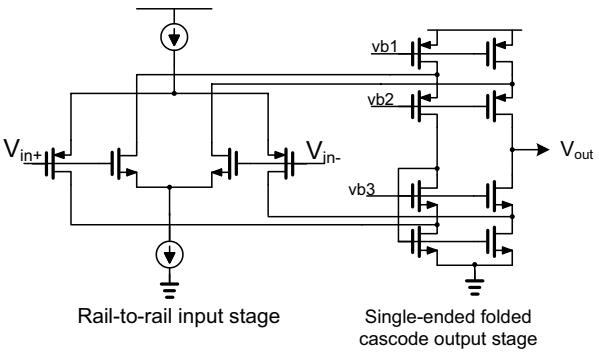


Fig. 6.23 Opamp used in charge pump

The second charge impairment in the current steering charge pump shown in Fig. 6.22 is charge injection. Charge injection is caused by the mobile charge in the metal–oxide–semiconductor field-effect transistor's (MOSFET) inversion layer, which is forced to leave the channel when the gate voltage changes. Any inversion charge that escapes to the loop filter node is an additional charge error term. When the M1 and/or the M2 transistor turns off, half of the charge in the channel goes to the X or Y node, respectively, and the other half goes to the loop filter capacitors. The charge in the channel is given as [14]

$$Q_{ch} = C_{ox}\left[V_{GS} - V_T\right] \tag{6.41}$$

where C_{ox} is the gate oxide capacitance, V_{GS} is the gate to source voltage, and V_T is the MOSFET device threshold voltage. Since it is assumed that the device V_{DS} is nearly zero (i.e., no horizontal electric field across the MOSFET channel), half the charge escapes through the drain, while the other half escapes through the source. This is somewhat of an approximation since the resistive nature of the channel forces a voltage gradient to develop within the channel as the charge attempts to escape. If this voltage becomes lower than the substrate voltage, it can cause charge to escape through the substrate. This effect is known as *charge pumping* and was first discovered by Brugler and Jesper [15]. From a circuit design perspective, this effect is desirable since it minimizes the charge error term appearing at the loop filter terminal. Also note that there are two devices at the loop filter node injecting charge in different directions (the p-type field effect transistor (PFET) injects hole whereas the n-type field effect transistor (NFET) inject electrons). The total charge error introduced by charge injection at the loop filter node then becomes

$$q_{ci,err} = \frac{C_{ox}\left[-V_{DD} + V_T\right]}{2} \tag{6.42}$$

where V_{DD} is the supply voltage, V_T is the (MOSFET) threshold voltage, and it was assumed that the PFET switch is sized twice that of the NFET switch. This indicates that the circuit design parameters affecting charge injection are the area of the device (C_{ox}) and the power supply voltage.

The third type of charge impairment is clock feedthrough [16]. Clock feedthrough is caused by the sharp rising and falling edges of the UP and DN signals coupling through the gate parasitics capacitors of the switching onto the loop filter. This scenario is illustrated in Fig. 6.24. The sharp rise and fall time through the last inverter stage in a PFD causes a current flow through to the output capacitance seen at that inverter stage. The output load capacitance can be split into two capacitors, C_{gd}, the gate-to-drain capacitance of the switch in the charge pump, and C_p which is all other parasitics capacitances at the output load of the inverter. The current flowing through the C_{gd} capacitor gives rise to the charge error term due to clock feedthrough. More specifically, the charge error appearing at the loop filter can be given as

$$q_{err,clk} = \int_0^{T_{ref}} \left[\frac{C_{gd}}{C_{gd} + C_p}\left(C_{gd} + C_p\right)\frac{dv(t)}{dt} \right] dt \tag{6.43}$$

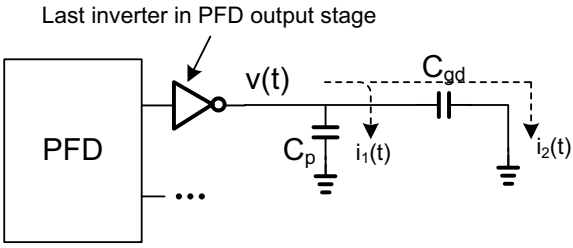


Fig. 6.24 Clock feedthrough scenario in a charge pump switch

where v(t) is the output voltage of the last inverter stage in the PFD and T_{ref} is the reference period (equal to $1/F_{ref}$). As Fig. 6.24 shows, the simplifiying assumption is that the loop filter can be treated as short. Since in reality the impedance of the loop filter is higher, (6.43) can be treated as an upper bound on charge error due to clock feedthrough.

The optimization of the output noise of a charge pump and PFD are also important. The AC noise analysis of a charge pump is fairly straight forward. The current noise density of all the devices can be referred to the output, then multiplied by the square of the impedance seen at the output. Since the noise is low-pass filtered by the loop filter, minimizing noise only at low frequencies would be important. The noise of a PFD + charge pump, however, is complicated by the fact that the operating of PFD is periodic at steady state. This causes a finite amount of sampling of the charge pump noise. This sampling causes alias terms to appear. For example, if the PFD is operating at 100 MHz and there is a noise source at 1 MHz, due to the sampled and periodic nature of the PFD during locked condition, an alias noise term will appear at 99 MHz, 101 MHz and at ± 1 MHz around all harmonics of the PFD sampling rate (100 MHz). Moreover, the amplitude of the harmonic samples is shaped by the sinc function introduced by the t_{on} pulse width of the PFD output signals UP and DN, as shown in Fig. 6.25. The implication of the phenomenon is that the wider the UP and DN pulse widths, the tighter the sinc function resulting in more noise filtering prior to the charge pump (i.e., more filtering of PFD or XO noise). Wider UP and DN pulse widths, however, result in more noise due to the charge pump itself (since noise is integrated over a longer period of time).

It can be shown that the power spectral density of device noise from the charge pump shaped by this sampling response is given by

$$S_Y(f) = \sum_{n=-\infty}^{\infty} \left| \frac{T_{on}}{T_{ref}} sinc\left(\frac{n \bullet T_{on}}{T_{ref}} \right) \right|^2 \bullet S_x(f) \tag{6.44}$$

where $S_x(f)$ is the power spectral density of the device noise sources in the charge pump. This shows that the power spectral noise density is reduced for smaller T_{on} pulse widths of the UP and DN signals.

Another important issue is the effect of nonlinear charge pump output characteristic. As seen in Sect. 6.3.2, the quantization noise in a sigma–delta modulator can be modeled as uniformly distributed additive noise source. This is under two

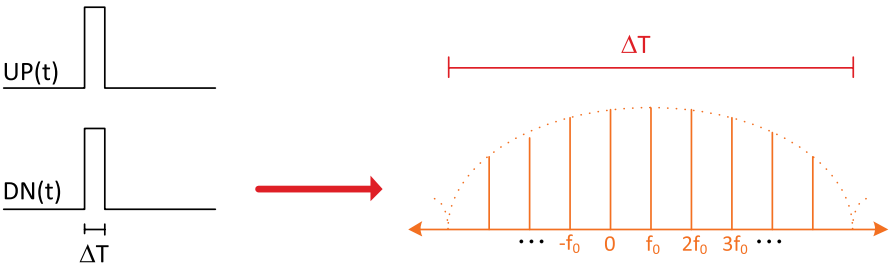


Fig. 6.25 Periodic UP and DN signals lead to a sampled sinc function to appear in the frequency domain

assumptions. First, there is sufficient activity at the output of the integrator preceding the quantizer such that it can be treated as a random process. Second, the step sizes in a multibit quantizer are equal to one another. This second assumption is necessary to maintain an equiprobable, and hence uniform, noise density of the quantization noise. The step size in a ΣΔ PLL is ultimately equal to the unit amount of charge dumped to the loop filter for unit step size in the ΣΔ modulator.

Due to the three charge impairments listed earlier, the charge error near zero phase error has the largest deviation from the ideal linear charge pump characteristic. Figure 6.26 shows an example of how the average current integrated over one reference cycle per phase error changes as the input phase error to the PFD is swept from −1 ns to +1 ns. The reference period was 50 MHz in this simulation. The phase error is expressed as a percentage deviation from the ideal linear charge pump. Note that at larger phase error, the static phase error settles to a nonzero value. This is due to both DC current mismatch between the I_{UP} and I_{DN} currents as well as a residual net charge error from the three charge impairments listed earlier. As the sigma–delta modulator dithers the phase from the feedback divider, the phase detector is dithered

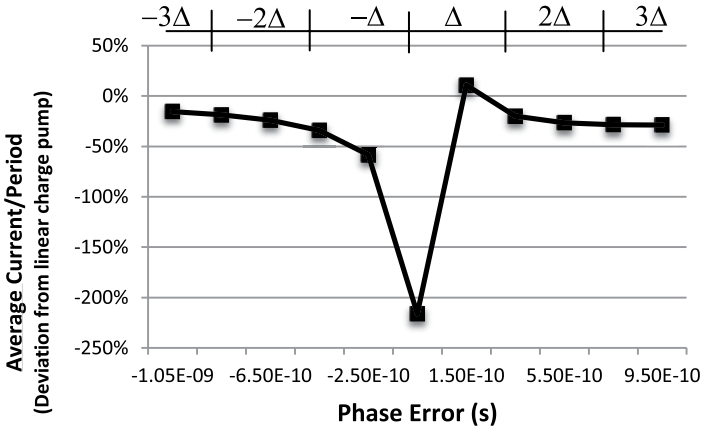


Fig. 6.26 Average current integrated over one reference period per phase error versus phase error

around the nonlinear region shown in Fig. 6.26. Each phase dither step is equal to a sigma–delta quantizer phase step size. Since the phase step size is converted to charge by integrating the curve in Fig. 6.26 over the phase step size, nonlinear charge quantization step sizes result at the output of the charge pump.

The effect of nonlinear step sizes is twofold. First, it causes noise folding. Noise folding is a phenomenon where high-frequency quantization noise is folded back to lower frequency. This can be explained by considering the sigma–delta quantization frequency spectrum as individual closely spaced tones. These tones are then multiplied by the charge pump transfer function, which is a nonlinear function. The charge pump nonlinear function can be approximated by a polynomial containing second- and third-order terms. This means that the sigma–delta quantization noise spectrum undergoes both second-order and third-order distortion. Considering two high-frequency tones within the sigma–delta noise spectrum, f_1 and f_2, the second-order intermodulation terms land on $f_1 - f_2$ and $f_1 + f_2$ frequencies. The $f_1 - f_2$ component results in low-frequency noise. Now, considering the full range of the quantization noise spectrum, there are several combinations of f_1 and f_2 that can potentially land within the PLL closed-loop bandwidth. These intermodulation terms are then aggregated to produce an increase in in-band phase noise. As similar argument holds for third-order distortion. Although higher-order distortion terms are possible, they usually have less effect than second- and third-order distortion. An example of noise folding in a sigma–delta PLL is shown in Fig. 6.27. Two plots are overlaid. One with a nonlinearity in the charge pump (the curve with the higher in-band phase noise) and the other with the nonlinearity in the charge pump corrected. It is important to note that this noise folding mechanism occurs *before* the filtering action of the PLL's low-pass filter following the charge pump.

6.6 Voltage-Controlled Oscillator Implementation

The VCO is a central component of the PLL. Its performance determines the tuning range of the PLL and to a large degree its overall phase noise, or jitter performance. Most integrated VCOs for wireless applications rely on an LC tank that is excited by a negative transconductance element to provide oscillation. In this section, VCO phase noise theory is first reviewed, followed by design trade-offs involved in determining the frequency tuning range and phase noise of the oscillator.

6.6.1 VCO Phase Noise Theory

In its most general form, an oscillator may be represented by the positive feedback system shown in Fig. 6.28. The transfer function of this system is

$$H(j\omega) = \frac{F(j\omega)}{1 - F(j\omega)} \tag{6.45}$$

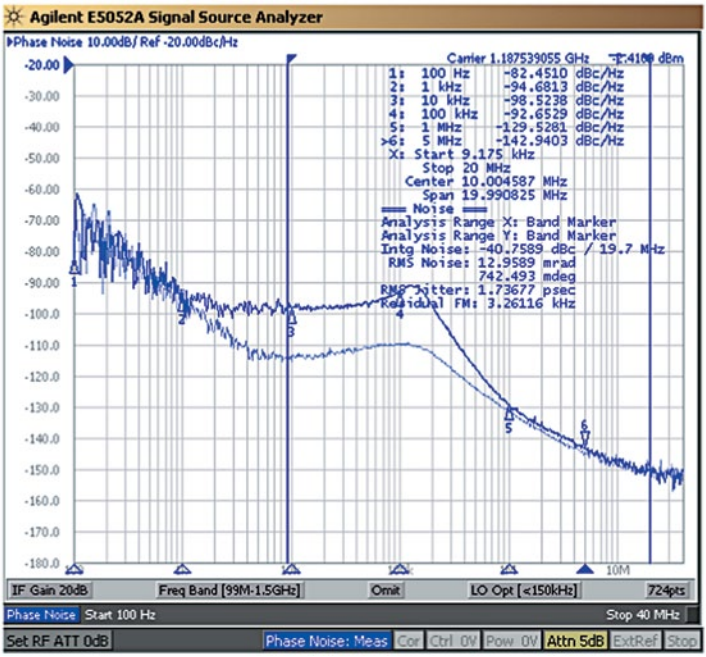


Fig. 6.27 Noise folding in a sigma–delta PLL

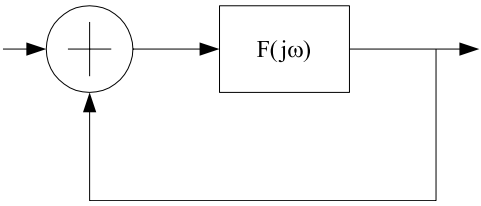


Fig. 6.28 A generic positive feedback system

where F(jω) is the open-loop transfer function of the oscillator, and H(jω) is the closed-loop transfer function of the oscillator. Since it is a positive feedback system, the amplitude of oscillation will grow until a nonlinearity in the oscillator causes its amplitude to saturate. Such nonlinearities include power supply limitations or transistor saturation. At large amplitudes, circuit nonlinearities become so severe that the gain of F(jω) becomes 1 and the total phase shift around the feedback loop is $0° \pm 360 \cdot n°$, where n is a positive integer. These two requirements constitute what is known as the *Barkhausen's* criteria for oscillation [17]. When $F(j\omega_o) = 1$, the oscillator's frequency is ω_o rad/s.

Another way of viewing the system representing an oscillator is by looking at the root locus of the system in Fig. 6.28 while varying F(jω). Initially, F(jω) is much greater than one. As the voltage swing grows, circuit nonlinearities cause the magnitude of F(jω) to quickly decrease up until the poles of the system are on the imaginary axis. At this point, stable oscillation is sustained at a certain frequency ω_o.

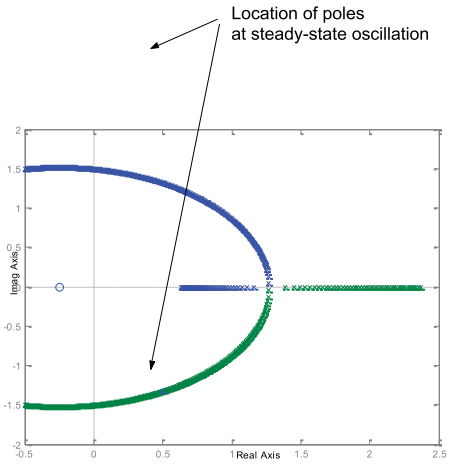


Fig. 6.29 Root locus of positive feedback system

At this point, the steady-state magnitude of the system is exactly one. Any perturbation in the magnitude would cause the poles to shift to the right or left causing a frequency or phase shift. This translation of magnitude error to frequency or phase error is what causes jitter. The root locus plot of the system in Fig. 6.28 is shown in Fig. 6.29.

An important parameter of an oscillator's performance is its *open-loop Q factor*. Simply stated, the open-loop Q factor is a measure of how much the closed-loop feedback system opposes variation in oscillation frequency. One commonly used equation for open-loop Q factor is

$$Q = \frac{\omega_0}{\Delta\omega} \tag{6.46}$$

where ω_o is the oscillation frequency and $\Delta\omega$ is the double sided frequency offset where the spectral density of the transfer function is one-half its peak value at ω_o as shown in Fig. 6.30. Using this definition and considering only one-side of the energy of the H(jω) transfer function, the magnitude of H(jω) at $\omega_o+\Delta\omega$ is

$$\left|H(j\omega)\right| = \frac{R_p}{2} = \frac{R_p}{2}\left(\frac{\omega_0}{Q \bullet \Delta\omega}\right) \tag{6.47}$$

where R_p is the maximum amplitude of H(jω).

Assuming that the dominant component of noise is thermal noise of R_p, the power spectral density of the thermal noise shaped by the oscillator transfer function is

$$\frac{\overline{v_n^2}}{\Delta f} = \frac{\overline{i_n^2}}{\Delta f} \bullet \left|H(j\omega)\right|^2 = 4kTR_p \bullet \left(\frac{\omega_0}{2Q\Delta\omega}\right)^2 \tag{6.48}$$

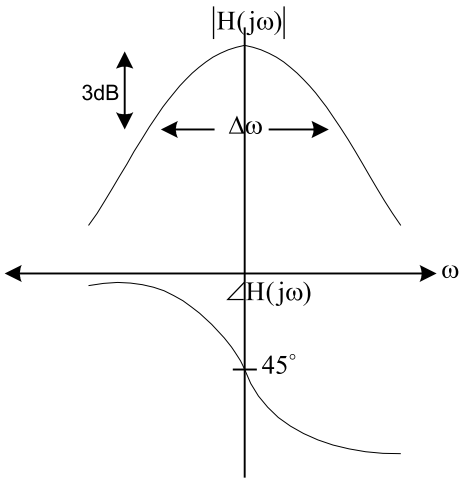


Fig. 6.30 Bode plot of closed-loop transfer function of oscillator

Equation (6.48) gives the total output spectrum including the amplitude as well as phase spectrum. Using the equipartition theorem of thermodynamics [18], the total spectrum is split evenly between phase noise spectrum and amplitude noise spectrum. However, since the oscillator nonlinearities provide an indirect form of amplitude control, the total output spectrum of the oscillator is given as only half of equation (6.48). Also, phase noise is normally reported as relative to the carrier signal at ω_o. Using these two facts, the phase noise spectrum of an oscillator dominated by thermal noise is given as

$$L\{\Delta\omega\} = \frac{2kT}{P_c}\bullet\left(\frac{\omega_0}{2Q\Delta\omega}\right)^2 \tag{6.49}$$

where P_c is the amplitude power of the oscillator. This formula is known as *Leeson's* formula [19]. Given this expression, the rms phase error can be found by summing (6.49) over the entire noise bandwidth. Since much of the power of the phase noise lies within a frequency bandwidth of $\Delta\omega$ around ω_o, the rms phase error can be found as

$$\theta_{e,rms} = \int_{-\frac{\omega_0}{2Q}}^{\frac{\omega_0}{2Q}} \frac{2kT}{P_c}\bullet\left(\frac{\omega_0}{2Q\Delta\omega}\right)^2 d(\Delta\omega) \tag{6.50}$$

Evaluating this integral yields the following result:

$$\theta_{e,rms} = \frac{kT}{P_c}\bullet\left(\frac{\omega_0}{Q}\right)\text{rads} \tag{6.51}$$

Converting radians to time, the rms jitter $T_{j,\,rms}$ is given as

$$T_{j,rms} = \frac{kT}{2P_c}\left(\frac{1}{Q}\right)\text{sec} \tag{6.52}$$

This is a very significant result. It clearly shows that for a given VCO topology (fixed Q), the rms jitter varies only with the power level of the oscillator output signal. *This means that there is a direct trade-off between power consumption and jitter in oscillators.*

The above analysis assumes that the noise of the oscillator depends only on the thermal noise of the devices used to build the oscillator. Such an assumption leads to the conclusion that the phase noise decreases at a rate of 20 dB/decade indefinitely. In practice, the phase noise floor is limited by the VCO output buffers, where the device noise is not shaped by the VCO phase noise shaping function. Also, at small frequency offsets Δω, flicker noise is more dominant. A more accurate plot of phase noise spectrum of an oscillator is shown in Fig. 6.31.

An equation representing the graph in Fig. 6.31 can be given as

$$L\{\Delta\omega\} = \frac{2kTF}{P_c}\left[1+\left(\frac{\omega_0}{2Q\Delta\omega}\right)^2\right]\left[1+\frac{\Delta\omega_{1/f}^3}{|\Delta\omega|}\right] \tag{6.53}$$

where F is a empirical parameter (or "fudge" factor) and $\Delta\omega_{1/f}^3$ is the corner frequency between the $1/f^2$ and $1/f^3$ regions of the phase noise spectrum. Equation (6.53) is known as the *Leeson–Cutler* formula [20]. The F parameter accounts for other noise sources other than the losses in the F(jω) system. Such noise sources can include intrinsic and extrinsic noise sources. The "1" factor accounts for the fact that there is a phase noise floor that defines the absolute minimum phase noise achievable at all offset frequencies. The $1/f^3$ region corresponds to upconversion of 1/f noise (mainly device flicker noise) to near the oscillation frequency. The Leeson–Cutler formula suggests that the boundary between the $1/f^3$ and $1/f^2$ regions is

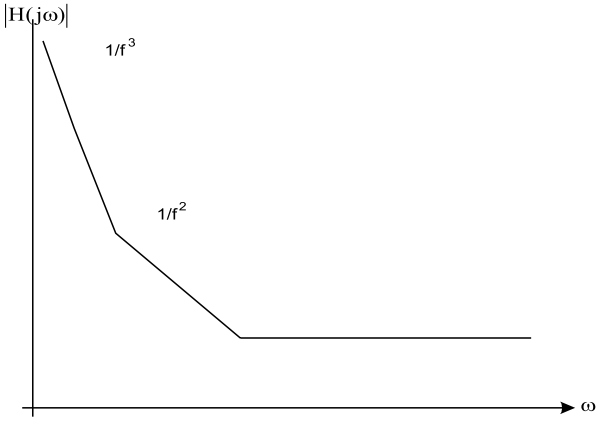


Fig. 6.31 More realistic phase noise spectrum of an oscillator

exactly the boundary between the 1/f and thermal noise regions. However, empirical data suggests otherwise for reasons that are explained in the next section.

6.6.2 Cyclostationary Analysis of VCO Phase Noise

The "fudge" factor F in the Leeson–Cutler formula given by (6.53) has been an unsatisfying factor to many designers for a number of years. One theory that has helped account for much of this "fudge" factor is the cyclostationary analysis of VCO phase noise. In this section, the periodic nature of the VCO output and its effect on phase noise is analyzed in more detail.

Figure 6.32 shows the translation of voltage noise into phase error at various regions of operation given a periodic signal, such as the VCO output voltage. As the figure shows, the amount of jitter resulting can differ drastically depending on the time instant the amplitude noise is injected into the VCO. When noise is injected during the voltage transition of the VCO output, it can potentially alter the zero crossings of the VCO, inducing phase noise. For this reason, the phase noise expression for VCOs is said to be a time-varying function. Moreover, since the output of a VCO is periodic, the noise sources that induce jitter vary periodically with time. Such noise sources are called *cyclostationary* noise sources [21].

The output of an oscillator may be given as

$$V_{out}(t) = A(t) \cdot f\left[\omega_0 t + \phi(t)\right] \tag{6.54}$$

where the function f is periodic in 2π and $\phi(t)$ and $A(t)$ are the phase and amplitude variations due to noise, respectively. Amplitude variations may alter the zero-crossing of the oscillator's output, and hence translate into phase variation. If the amplitude variation is small enough, the amplitude-to-phase translation may be assumed to be linear. An instantaneous change in voltage due to noise injection would cause an instantaneous change in charge, which is given by

$$\Delta V = \frac{\Delta q}{C_{node}} \tag{6.55}$$

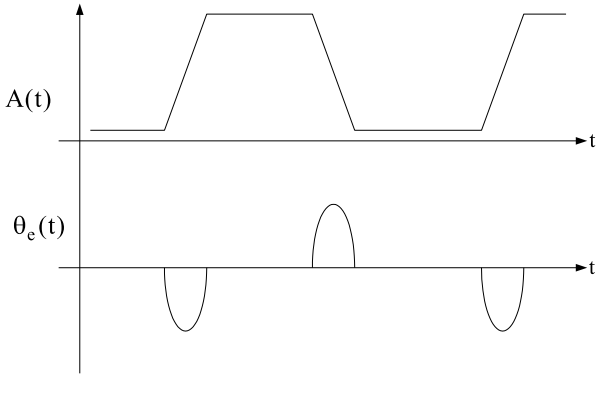


Fig. 6.32 Jitter transfer function for various noise injection times

where C_{node} is the capacitance of the node which experienced charge injection due to noise. The phase variation can be given by [22]

$$\Delta\phi = \Gamma(\omega_0 t)\frac{\Delta q}{q_{swing}} \tag{6.56}$$

where $q_{swing} = C_{node} \cdot V_{swing}$, and V_{swing} is the voltage swing of the node which experienced the noise charge injection. $\Gamma(\omega_0 t)$ is a unitless time-varying function, called the *impulse sensitivity function (ISF)* [23]. When an internal signal is high or low, amplitude noise will have little or no effect on phase error on the output of the oscillator. This means that the ISF is a small or zero value during that interval. On the other hand, the ISF is maximized at the time of a transition switching of an internal node. Note that once a phase error has occurred, it is not corrected for. Therefore, the phase error accumulates indefinitely as time increases. This type of analysis assumes that the oscillator is a linear time-varying (LTV) system.

Assuming that the total noise in the oscillator can be represented as current noise i(t), the output phase error of the oscillator is given as

$$\phi(t) = \frac{1}{q_{max}}\int_{-\infty}^{t} \Gamma(\omega_0)\cdot i(\tau)d\tau \tag{6.57}$$

where q_{max} is the maximum charge injected by the noise source. Since ISF is a periodic function, it can be represented as a Fourier series

$$\Gamma(\omega_0\tau) = \frac{c_0}{2} + \sum_{n=1}^{\infty} c_n \cos(n\omega_0\tau) \tag{6.58}$$

Substituting (6.58) back into (6.57) gives

$$\phi(t) = \frac{1}{q_{max}}\left[\frac{c_0}{2}\int_{-\infty}^{t} i(\tau)\,d\tau + \sum_{n=1}^{\infty} c_n \int_{-\infty}^{t} i(\tau)\cos(n\omega_0\tau)\,d\tau\right] \tag{6.59}$$

Consider two cases for i(t): (a) when it is a low-frequency sinusoidal signal with frequency $\Delta\omega$, and (b) when it is a sinusoidal signal with a frequency near the carrier ω_o with frequency $\omega_o \pm \Delta\omega$. In the first case, only the first integral in (6.59) contains a significant signal, and the resulting output phase error is given as

$$\phi(t) = \frac{I_0 c_0 \sin(\Delta\omega t)}{2q_{max}\Delta\omega} \tag{6.60}$$

where I_0 is the maximum amplitude of the input current. A similar result can be shown for case (b), but with Fourier coefficient c_1. More generally, it can be shown that for a current sinusoidal input with frequency $\omega_o \pm n\Delta\omega$, the output phase error will be a sinusoid with frequency $\Delta\omega$ and magnitude proportional to c_n. These phase

errors are then upconverted to ω_o (for c_0) and downconverted to ω_o (for c_2, c_3, …, c_n) to the oscillation frequency by the nonlinear oscillator transfer function. Noise sources near ω_o remain the same. This means that noise injected into the oscillator is only significant if it is near dc, near the oscillator frequency, or a harmonic of the oscillator frequency. A similar result was reached in [24].

The statistics of the timing jitter depends on the correlation of the noise sources involved. In the case of thermal noise, the noise sources are considered to be random and uncorrelated. Therefore, it follows that considering only the thermal noise of an oscillator and jitter measured over a time interval ΔT, the standard deviation of the jitter of the oscillator is given as [25]

$$\sigma_{\Delta T} = \kappa\sqrt{\Delta T} \tag{6.61}$$

where κ is a proportionality constant which can be shown to be equal to

$$\kappa = \frac{\Gamma_{rms}}{q_{max}\omega_0}\sqrt{\frac{1}{2}\frac{\overline{i_n^2}}{\Delta f}} \tag{6.62}$$

where ω_0 is the output target frequency in rads/sec, q_{max} is equal to $C_L \cdot V_{SW}$, where C_L is the parasitic capacitance at the output of the oscillator and V_{SW} is the voltage swing of the oscillator, and $\frac{\overline{i_n^2}}{\Delta f}$ is the power spectral density of the thermal noise of the active devices in the oscillator. For CMOS transistors, the drain current noise spectral density is given by [26]

$$\frac{\overline{i_n^2}}{\Delta f} = 4kT\gamma\mu C_{ox}\frac{W}{L}\left(V_{GS} - V_T\right)^2 \tag{6.63}$$

where γ is a coefficient equal to 2/3 for long-channel transistors. As equations (6.62) and (6.63) show, jitter due to thermal noise can be minimized by reducing the ISF, increasing the voltage swing (at the expense of power consumption), increasing the operating frequency, or reducing the power spectral density of the device thermal noise.

Correlated noise sources are usually a result of low-frequency noise, such as flicker noise, as well as power supply bounces. Flicker noise can be minimized by using large transistors. One interesting result from ISF analysis is that the corner frequency of $1/f^2$ and $1/f^3$ can be accurately determined to be [27]

$$\omega_{1/f^3} = \omega_{1/f}\left(\frac{\Gamma_{dc}}{\Gamma_{rms}}\right)^2 \approx \omega_{1/f}\left(\frac{c_0}{c_1}\right)^2 \tag{6.64}$$

where $\omega_{1/f}$ is the corner frequency of 1/f noise and thermal white noise. Therefore, the upconversion of flicker noise can be minimized by having a more symmetric

waveform around the x-axis (for zero dc level) and by maximizing the oscillator's voltage swing. This is contrary to the original assumption by Leeson when deriving his formula [19].

One important parameter in the design of a VCO is its power supply rejection. The analysis injection of a tone from the power supply near the carrier frequency is equivalent to modulating the amplitude of the VCO. If the modulated envelope is expressed as a sinusoidal signal, the resultant AM modulated signal can be expressed as

$$y_{vco}(t) = \left[1 + m \cdot \cos\left(2\pi f_m t + \phi\right)\right] \cdot V_{swing} \cdot \sin\left(2\pi f_c t\right) \tag{6.65}$$

where f_m is the frequency of the modulation of the VCO amplitude envelope, f_c is the VCO center frequency, and m is the AM modulation index, which is equal to the ratio of the amplitude of the noise tone to the VCO amplitude. Expanding the expression given by (6.65) leads to

$$y_{vco}(t) = V_{swing} \cdot \sin\left(2\pi f_c t\right) + \frac{m \cdot V_{swing}}{2}\left[\sin\left(2\pi\left(f_c + f_m\right)t + \phi\right) + \sin\left(2\pi\left(f_c - f_m\right)t - \phi\right)\right] \tag{6.66}$$

which shows that AM modulation results in a pair of side tones around the VCO carrier.

These two side tones are further shaped by two characteristics of the VCO. The first is the bandpass nature of the VCO, as was illustrated in Fig. 6.30. The higher the Q of the VCO, the more filtering is achieved. The filtering improves at a rate of 20 dB per decade of frequency offset, f_m, from the VCO center frequency, f_c. The other characteristic is the shaping of the VCO phase noise within the closed loop response of the PLL. As was shown by (6.31), the VCO phase noise is high-pass filtered when enclosed in a PLL. This means that phase noise at low-frequency offsets from the carrier are well suppressed, whereas, high-frequency noise is passed through. Aligning the bandpass filtering response dependent on the Q of the VCO (which appears as a low-pass filter when viewed on a phase noise plot with the y-axis centered at the VCO center frequency) along with the high-pass nature of the VCO noise shaping in a PLL, leads to a power supply to output transfer function which appears to be bandpass shaped, as shown in Fig. 6.33. This shows that the worst case noise would appear at the PLL closed-loop bandwidth. Note that both x and y axes are logarithmic scales.

6.6.3 LC VCO Design

The most prevalent VCO topology is an LC VCO [28]. An LC VCO relies on an inductor–capacitor resonator tank to produce an oscillator. For a lossless LC tank, the tank would simply need an initial excitation signal to start oscillation. Since the

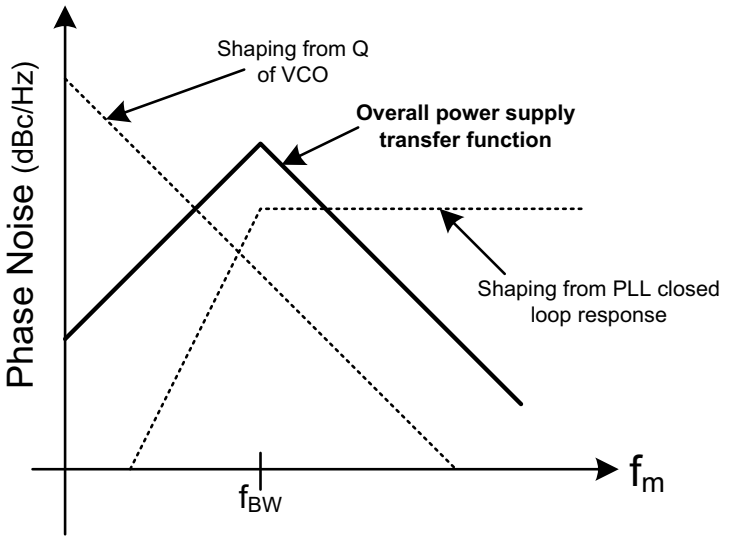


Fig. 6.33 Power supply noise-transfer function in VCOs

tank is lossless, the output voltage could theoretically reach infinity. In any practical LC VCO implementation, there are resistive losses in both the inductor and capacitor. A negative conductance generator would be used to cancel out the losses in the LC tank.

Before discussing the various topologies of LC VCOs, RLC tanks are first reviewed. Consider a parallel RLC resonator shown in Fig. 6.34. The admittance of the tank is given as

$$I(\omega) = \left\{ G + j\left[\omega C - \frac{1}{\omega L} \right] \right\} \bullet V_{tank} \tag{6.67}$$

where $G = 1/R_p$, L and C are the inductance and capacitance components of the resonator, and I(ω) is the current through the LC tank. At resonance frequency, the imaginary components cancel out. The frequency is calculated by setting the imaginary part of (6.67) to zero, yielding

$$f_{vco} = \frac{1}{2\pi\sqrt{LC}} \tag{6.68}$$

where $\omega = 2\pi f_{vco}$. Since the tank is not ideal, the amplitude is bounded by R_p. More specifically, the amplitude of the tank is given by

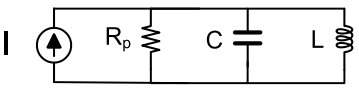


Fig. 6.34 Parallel RLC resonator

$$V_{swing} = R_p \bullet I \tag{6.69}$$

An alternative, yet equivalent to (6.46), definition to the quality factor, Q, of the tank can be defined as

$$Q = \omega_{vco} \frac{energy\ stored}{avg\ power\ dissipated} \tag{6.70}$$

Considering a peak current in the tank, I_{pk}, and using (6.68)–(6.70) can be expanded as

$$Q = \frac{\frac{1}{2} C \left[I_{pk} R_p \right]^2}{\frac{1}{2} I_{pk}^2 R_p} = \frac{R_p C}{\sqrt{LC}} = \frac{R_p}{\sqrt{L/C}} \tag{6.71}$$

Note that equivalent expressions for Q can be given as

$$Q = \frac{R_p}{\omega_{vco} L} = \omega_{vco} R_p C \tag{6.72}$$

Similar expressions for Q can be given for a series RLC resonator, where

$$Q = \frac{\sqrt{L/C}}{R_s} = \frac{\omega_{vco} L}{R_s} = \frac{1}{\omega_{vco} R_s C} \tag{6.73}$$

where R_s is the resistance associated with series resonance. In a practical LC tank, both the inductor and capacitor have a series parasitic resistance component, as shown in Fig. 6.35. As shown, the series equivalent resistance has been converted into a parallel resistance. It can be shown [29] that the R_p and R_s are related by the following equation

$$R_p = R_s \left(Q^2 + 1 \right)^2 \tag{6.74}$$

This relationship hold only for narrow frequency ranges around the resonance frequency and for sufficiently high values of Q (Q > 10). The resulting total Q of the tank is given as

$$Q^{-1} = Q_C^{-1} + Q_L^{-1} \tag{6.75}$$

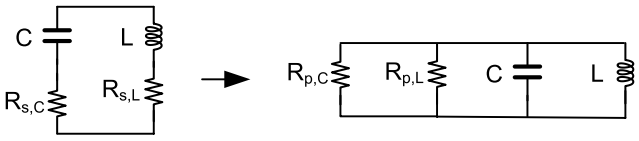


Fig. 6.35 Practical LC tank and its equivalent circuit

Where Q_C is the tank Q assuming degradation from the capacitor series resistance only and Q_L is the tank Q assuming degradation from the inductor series resistance only. As will be shown later, Q_L dominates at high VCO frequencies, whereas Q_C dominates at low VCO frequencies.

VCO Topologies The most popular method of creating a negative transconductance element is to use a cross-coupled MOSFET pair. There are three possible topologies for the LC VCO each with a different cross-coupled MOSFET topology, as shown in Fig. 6.36. In all three topologies, a current source is used to provide current to the VCO. The first of these topologies is the CMOS $-g_m$ based LC VCO. The main advantage of this type of topology is that it can provide symmetric rise and fall waveforms, lowering the flicker noise corner as was explained by (6.64). Symmetric rise and fall times can be adjusted to the first order by properly adjusting the ratio of PFET to NFET devices. This adjustment, of course, would misalign as the process is shifted. The disadvantage of a CMOS $-g_m$ based LC VCO is that it requires extra headroom due to the additional $v_{ds,\,sat}$ required in cascading PFET and NFET devices along with a current source, when compared to an NMOS or PMOS only $-g_m$ based LC VCO. This limits the maximum voltage swing allowed in the VCO before saturating the VCO signal, which causes the far out noise of the LC VCO to be inferior to that of NMOS or PMOS only $-g_m$ based LC VCOs for the same current consumption in the tank. Note that the center tapped inductor in the PFET only or NFET only LC VCOs can be substituted by a low-impedance bias reference to avoid the voltage swing across the LC tank from going above V_{DD} or to a negative voltage, respectively. Also note that the current source shown in Fig. 6.36a can be inserted on top, connected to V_{DD}, as shown in the figure or alternatively at the bottom, connected to ground.

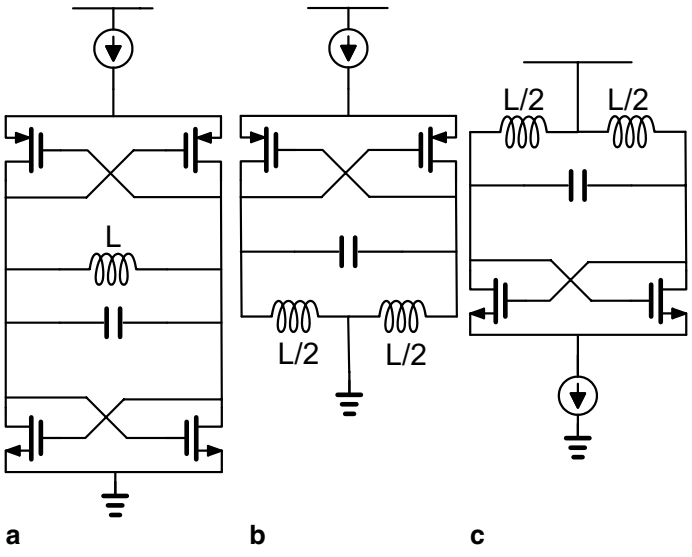


Fig. 6.36 LC VCO topologies based on **a** CMOS $-g_m$, **b** NMOS $-g_m$ and **c** PMOS $-g_m$

In low-voltage applications, there may not be enough headroom to add a current source. This is especially true of a CMOS LC VCO, shown in Fig. 6.36a. In order to reduce the required headroom, the current source may be substituted with a resistor. The current across the resistor may be regulated by a feedback amplifier controlling the power supply of the VCO, as shown in Fig. 6.37. Care must be taken in the choice of the resistor. Too large of a resistor value would increase the required headroom, perhaps more than that of a current source. Too small of a resistor value would load the LC tank reducing the effecting swing across the tank, as will be shown later. The reference voltage at the input of the comparator is set by a replica circuit that forces the current through the VCO to match that of the current source of the replica circuit (if the resistor in the VCO is equal to that of the replica bias reference). The design of the error amplifier is critical as it determines the overall power supply rejection and can be a source of phase noise in the VCO.

The design of the VCO involves choosing the center frequency, frequency tuning range, and phase noise performance of the VCO. Since an LC tank is used for the VCO, the center frequency of the VCO was given in (6.68). The total inductance is the designed inductor plus any parasitic inductance in the LC tank. The parasitic inductance becomes important for VCOs with center frequency greater than 5 GHz, necessitating a 3-D EM solver for proper inductance extraction. The total capacitance includes the designed capacitor as well as parasitic capacitances of the MOSFET devices as well as the wire trace connecting the LC tank components.

Varactors The designed capacitance of the VCO consists of a varactor and a tuning capacitor. A varactor is a device whose capacitance varies with the applied voltage. One terminal of the varactor is the loop filter voltage and the other is the VCO tank voltage. This is the device which controls the VCO gain (MHz/V). Since the total capacitance in the LC tank would consist of the varactor and other tank capacitances, the varactor area would have to be increased if a larger VCO gain is required. Hence higher VCO gain is problematic for two main reasons: more noise due to the varactors and the maximum frequency of the VCO is reduced. Separating the capacitance in (6.68) into a fixed and voltage-dependent terms, then differentiat-

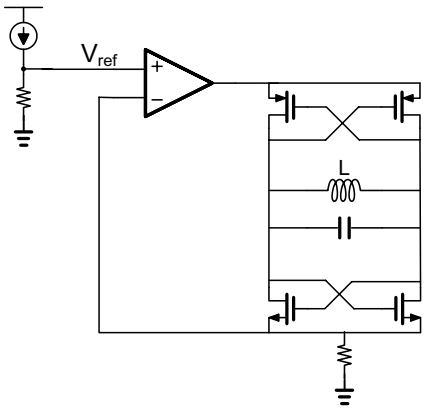


Fig. 6.37 Low-voltage LC VCO

ing the equation with respect to voltage applied to the varactor yields an expression for the VCO gain that is given as

$$K_v = -\frac{1}{2} \bullet L \bullet \frac{dC(V)}{dV} \bullet f_c \tag{6.76}$$

where f_c is the center frequency of the VCO given by (6.68), L is the inductor of the LC tank, and $\frac{dc(V)}{dV}$ is the large signal gain of the varactor given by the slope of its C–V curve.

There are two types of MOS-based varactors that can be implemented. The first, inversion mode varactor, is shown in Fig. 6.38 [30]. Shown is a PFET device in an n-well. Note that the source and drain are shorted together and form the positive terminal of the varactor. The negative terminal of the varactor is the gate node. As the source-to-gate voltage is increased, an inversion layer forms under the channel. This increases the capacitance seen across the varactor terminals to be equal to the gate oxide capacitance, C_{ox}. This is equal to C_{max} shown in Fig. 6.38. As the source-to-gate voltage is reduced to a voltage below the absolute value of the PFET threshold voltage, V_T, the inversion layer disappears the capacitance now consists of the gate-oxide capacitance in series with the depletion capacitance in the N-well region. The presence of the depletion region is guaranteed since the N-well is biased to the highest possible potential, V_{DD}. The total capacitance is nearly half of the gate-oxide capacitance, and is shown in Fig. 6.38 as C_{min}. The channel resistance in both inversion and depletion region determine the overall noise performance of the varactor.

Although the graph shown in Fig. 6.38 shows a steep C–V curve, one must keep in mind that this is an AC C–V curve. When using a varactor in a VCO, the actual capacitance seen by the tank is a time averaged capacitance as the V_{SG} is changed from minimum VCO voltage swing to its maximum relative to the loop filter voltage. As a result, the large signal C–V curve is much shallower than the AC C–V curve, resulting in much more linear tuning voltage characteristics of the VCO.

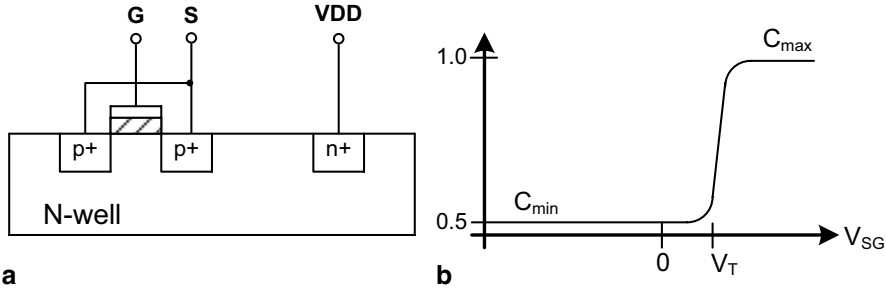


Fig. 6.38 Inversion mode varactor **a** device cross-section and **b** C–V curve

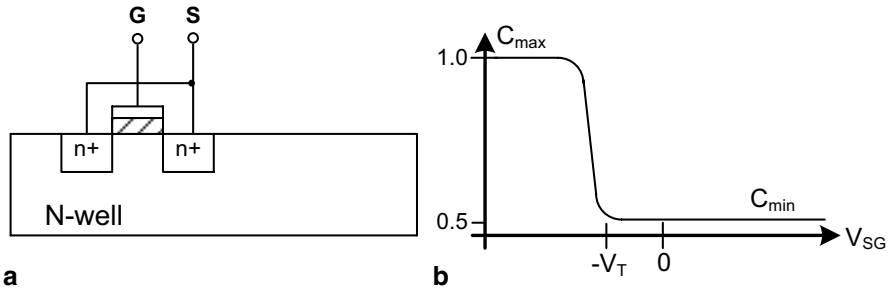


Fig. 6.39 Accumulation mode varactor **a** device cross-section and **b** C–V curve

Another popular implementation of a MOS varactor is the accumulation mode varactor, sometimes known as an nfet-in-nwell varactor, shown in Fig. 6.39 [31]. Note that since the diffusion materials are n + materials inserted in an N-well, it is not possible for an inversion layer to develop. This fact gives an advantage of an accumulation mode varactor over an inversion mode varactor in terms of its noise performance. As the gate voltage is increased beyond the V_T of the device, like particles, namely electrons, start to accumulate across the channel, creating a conduction band between the source and drain. The capacitance measured from the gate to the diffusion regions is now given as C_{ox}, or C_{max} as shown in Fig. 6.39b. When the gate voltage is reduced by the source voltage such that $V_{GS} < V_T$, then the device enters depletion region of operations, where a depletion region develops across the channel. This creates a series capacitance with C_{ox}. The effective capacitance of the varactor is now reduced by half.

Varactors used in VCOs depend on the topology of the loop filter. If a single-ended loop filter is used, then the varactors are arranged as shown in Fig. 6.40. Shown in the diagram is a parallel bank of varactors. If larger K_V (VCO gain) is required, more varactors are used. The voltage level of V_{ref} is chosen in such a way that the varactor C–V curve is centered for a desired loop filter voltage, usually $V_{DD}/2$. Also note that the reference voltage is heavily filtered to avoid any noise degradation due to the reference voltage circuitry. The varactors are then AC coupled from the LC tank to allow independent choice of the DC operating points of the tank itself, which

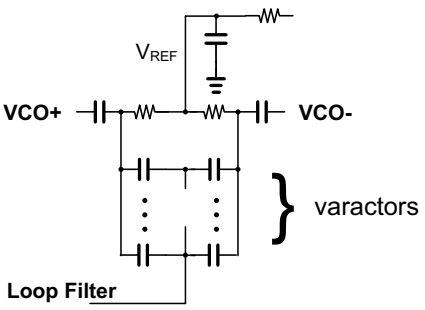


Fig. 6.40 Practical use of varactors in a an LC VCO

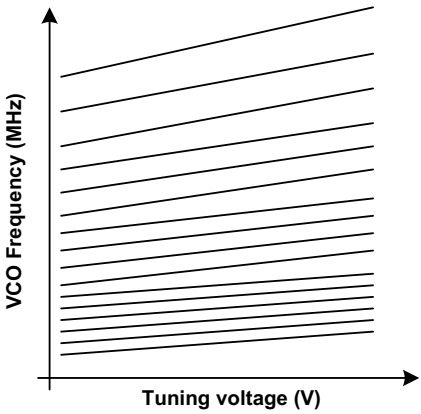


Fig. 6.41 Typical VCO tuning curves

are chosen to minimize phase noise. In differential PLL designs, the reference voltage may be substituted for the second polarity of the differential loop filter.

Capacitor Tune Array Once the VCO center frequency is chosen, the VCO frequency range is determined by the capacitor tune array implementation. The capacitor tuning array consists of digitally tunable capacitor cells. The capacitance of each cell varies between a C_{min} and a C_{max}. The ratio between the C_{max}/C_{min} determines the tuning efficiency of each cell. The absolute value of C_{min} and C_{max} determine the resolution of the capacitor array. The capacitor array can be viewed as a capacitor digital-to-analog converter (DAC), where the output is capacitance, instead of a voltage or current. In this case, the partitioning and sizing of capacitor elements is similar to the DAC design discussion in Sect. 5.6.

The VCO gain (K_v) and the C_{min} and C_{max} are chosen such that there are no frequency gaps in the VCO frequency range. Frequency gaps arise when the maximum frequency in one tuning array setting is less than the minimum frequency in the next tuning array setting. This can result from process, voltage, and temperature variation altering both the C–V curve as well as the C_{min} and C_{max} values. One way to avoid frequency gaps is to overlap the frequency tuning curves as shown in Fig. 6.41. The end points of each line represent the C_{min} and C_{max} of a capacitor tune setting corresponding to an f_{max} and f_{min}, respectively. The continuous set of points along of each line represent the set of frequencies corresponding to tuning the varactor capacitance by the PLL's loop filter voltage (x-axis). As the curve shows, there is significant overlap in frequency between adjacent capacitor tune settings. This is done in order to avoid any frequency gaps that can arise over process, voltage and temperature variations. Another point to note is that both the spacing between adjacent capacitor tune settings and VCO gain increase as the frequency increases. This is due to the increased sensitivity of the LC tank frequency to variations in the capacitor as the LC product is reduced.

The typical capacitor cell element is shown in Fig. 6.42. Typically, the sizes of the cell elements are binary weighted by the bit setting it controls. For example, a 5-bit capacitor cell array would contain 5 elements of the schematic shown in

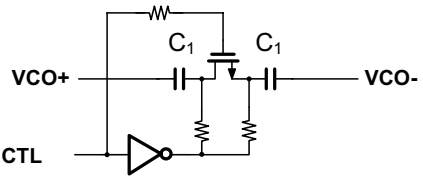


Fig. 6.42 Capacitor cell element

Fig. 6.42, with sizes of 1x, 2x, 4x, 8x, 16x, and 32x of a unit size element corresponding to bits 0, 1, 2, 3, 4, and 5, respectively. Due to varying dimensions for the capacitors and MOSFET elements, the contribution of boundary capacitances such as the fringe and sidewall capacitances to the overall capacitance of the capacitor cell element differ. These differences can create a nonmonotonic behavior in the output frequency of the VCO, which is most pronounced near the mid-codeword when codeword changes from $01\cdots1$ to $10\cdots0$ or vice versa. As was shown in Sect. 5.6, the DNL in this case is $\sqrt{2^{N-1}}\sigma$ where σ is the standard deviation in the capacitance of an LSB capacitor cell element. Designing for the worst case DNL would necessitate large overlap between the capacitor step sizes, which would reduce the overall frequency tuning range.

As was also shown in Sect. 5.6, this nonmonotonicity can be reduced by using thermally weighted unit elements of the schematic shown in Fig. 6.42. The disadvantage of this approach is that it incurs a large parasitic interconnect capacitance overhead, which limits the maximum attainable VCO frequency. A compromise approach is to use binary weighted elements for the LSBs and thermally weighted elements for the MSBs, similar to a segmented DAC.

The limitations of C_{min} and C_{max} values are important to understand. Figure 6.42 can be redrawn showing the parasitic capacitances as shown in Fig. 6.43. When the CTL signal is low, the NFET transistor is off and the C_{ds} is now in series with the two C_1 capacitances, resulting in a C_{min} capacitance. The overall differential capacitance is effectively lowered by the series capacitance between the source and drain of the NFET transistor. This capacitance is given by the series capacitance of C_{gs} and C_{gd} in parallel with C_{ds}. The equivalent lumped capacitance, C_x, can be given as

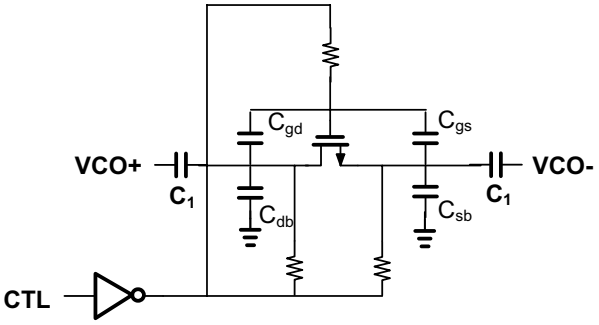


Fig. 6.43 The capacitor tuning cell with parasitic capacitances shown

$$C_x = C_{ds} + \frac{C_{gs}C_{gd}}{C_{gs} + C_{gd}} \tag{6.77}$$

The resistor in series with the gate of the NFET device is necessary in order to prevent signal losses to ground from the C_{gd} and C_{gs} capacitances to the preceding inverter. The source-to-bulk capacitance, C_{sb}, and the source-to-drain capacitance, C_{db}, form impedances to ground. These two capacitors, C_{sb} and C_{db}, (along with any other parallel parasitic capacitance to ground) is referred to as C_{p1} and C_{p2}, respectively. The purpose of biasing the source and drain through the two resistors is to maximize the capacitance difference between the C_{min} and C_{max} by utilizing the fact that the FET capacitances are voltage dependent. The C_{min} capacitance can then be given by

$$C_{min} = \frac{C_1}{1 + \dfrac{C_1^2}{C_x\left[C_1 + C_{p1}\right]} + \dfrac{C_x C_1}{\left[C_1 + C_{p1}\right]\left[C_x + C_{p1}\right]}} \tag{6.78}$$

It can be shown that for the ideal case of $C_{p1}=C_{p2}=0$ and $C_x \to \infty$, (6.78) reduces to $C_1/2$.

When the CTL signal is high, the NFET transistor is high and the R_{ds}, the channel resistance, becomes important. In this state, the differential capacitance of the capacitor tuning cell is maximized, since the series C_{ds} capacitor is now shunted by a low-impedance R_{ds}, resulting in a C_{max} capacitance. Using Fig. 6.43, the impedance across the tuning capacitor cell at the CTL high state can be shown to be

$$Z_{tune}(s)\big|_{CTL=1} = \frac{s + \dfrac{1}{R\left[C_1 + C_{p2}\right]}}{sC_{p1}\left[s + \dfrac{C_1 + 2C_{p2}}{C_{p1}R\left[C_1 + C_{p2}\right]}\right]} \tag{6.79}$$

which is a two pole, one zero system. Figure 6.44 shows how $Z_{tune}(s)$ varies over frequency. For most practical applications, the oscillating frequency is beyond the second pole shown in the figure. Considering an "equivalent" capacitance, C_{eq}, that matches the curve at high frequencies (frequencies beyond the second pole), the equivalent C_{max} can be computed. Equating (6.79) to $\dfrac{1}{sC_{eq}}$ and noting the C_{eq} is C_{max} results in

$$C_{max} = \frac{C_{p1}\left[s + \dfrac{C_1 + 2C_{p2}}{C_{p1}R\left[C_1 + C_{p2}\right]}\right]}{s + \dfrac{1}{R\left[C_1 + C_{p2}\right]}} \tag{6.80}$$

$$C_{max} = \frac{C_{p1}\left[s+\frac{C_1+2C_{p2}}{C_{p1}R[C_1+C_{p2}]}\right]}{s+\frac{1}{R[C_1+C_{p2}]}}$$

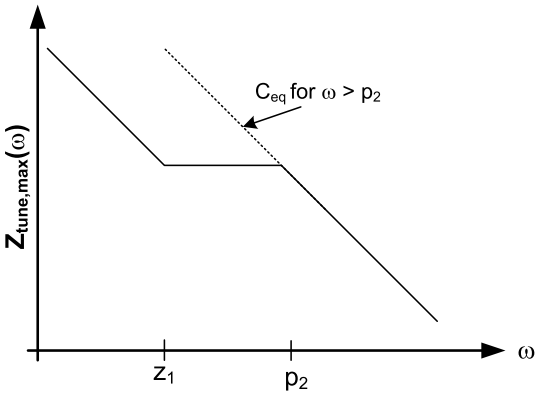


Fig. 6.44 Impedance of tuning capacitor cell on the C_{max} state

It is important to note that the value of equivalent C_{max} is frequency dependent.

The noise limitations of the capacitor tuning array element are also important to understand. As was shown in Fig. 6.35 and (6.75), the total R_p, the equivalent resistance in a parallel LC tank, is a parallel combination of the equivalent series resistance of the inductor and the capacitor. The following analysis assumes that the R_p degradation due to the capacitor tuning cells is dominant. When the capacitor array is set to its minimum capacitor setting, the series resistance is equal to R_{ds}. If the NFET and capacitor sizes are scaled with the bit weighting, it can be assumed that the equivalent series resistance scales down with increasing parallel capacitor tuning cells turned on (resistors added in parallel). If N is the number of cells turned on and R_p is the equivalent parallel resistance of the LC tank, the resulting output voltage noise can be given as

$$v_n^2 = \frac{4kT}{R_{ds}} N \bullet R_p^2 \tag{6.81}$$

which is the sum of all the current noise density terms from the capacitor tuning elements whose NFETs are turned on times the square of the equivalent output resistance. Note that R_{ds} is series resistance of a unit capacitor (LSB size). The output swing of the VCO was given by (6.69). Using (6.81) and (6.69) an expression for the output noise can be shown to be

$$\frac{v_n^2}{V_{swing}^2} = \frac{\frac{4kT}{R_{ds}} N \bullet R_p^2}{I^2 \bullet R_p^2} = \frac{4kTN}{I^2 R_{ds}} \tag{6.82}$$

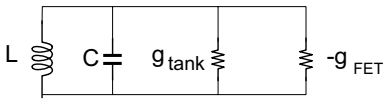


Fig. 6.45 Equivalent model of cross-coupled FET LC oscillator

which shows that the phase noise degrades as more tuning elements are turned on. This means that phase noise is expected to be worse at lower frequencies, due to more elements turned on (assuming the narrow band tuning range of the VCO). One method, as will be seen shortly, to improve phase noise at lower frequencies is to increase the bias current of the VCO.

Negative Transconductance Optimization The phase noise optimization of an LC VCO depends on several parameters. One important parameter is the choice of negative transconductance value. An equivalent model of the LC VCO is shown in Fig. 6.45, where the losses in the tank are represented by g_{tank} ($g_{tank} = 1/R_p$) and the negative $-g_m$ due to the cross-coupled FETs is represented by $-g_{FET}$. In order to sustain oscillation, the tank losses, g_{tank} must be canceled by ensuring that $g_{FET} > g_{tank}$. The oscillation amplitude will continue to increase until a nonlinearity is reached that causes the large signal gain to be one at the desired oscillation frequency. This nonlinearity may be supply voltage headroom or bias current limitation across the real impedance component of the tank. The oscillation frequency was given by (6.68).

As long as the amplitude of oscillation is not limited by the circuit's voltage headroom, increasing the energy in the tank will lead to reduction in phase noise and jitter. Energy stored in the LC tank is

$$E_{tank} = \frac{1}{2} C V_{tank}^2 \tag{6.83}$$

Substituting (6.66) into (6.81) and solving for the tank voltage swing yields

$$V_{swing}^2 = 2E_{tank}\omega^2 L \tag{6.84}$$

Intuitively, this equation would lead one to believe that increasing the inductance alone would result in lower-phase noise. However, in an LC oscillator, the tank's voltage and thermal noise voltage increase at the same rate. Under the simplifying assumption that the noise is dominated by upconverted thermal noise ($1/f^2$ noise), the ratio of the tank voltage to thermal noise voltage can be shows to be [32]

$$\frac{V_{swing}^2}{\overline{v_n^2}} = \frac{2E_{tank}}{kT} \tag{6.85}$$

where k is Boltzmann's constant. Using (6.83) and (6.84) the ratio of (6.85) can be rewritten as

$$\frac{V_{swing}^2}{\overline{v_n^2}} = \frac{I_{bias}^2}{\omega^2\left(g_{tank}^2 L\right)kT} \tag{6.86}$$

The expression in (6.86) shows that in order to maximize the amplitude to noise ratio, the product of $g_{\tan k}^2 L$ (L/R_p^2) must be minimized for the same current consumption. Increasing the bias current also helps to improve the phase noise performance. This equation assumes that the amplitude of oscillation is not limited by supply voltage.

The above analysis is based on the assumption that the tank voltage swing can be increased by increasing the bias current of the LC tank. This region of operation is called the *current mode region* [33]. If the voltage swing is limited by the power supply (or any other amplitude limiting nonlinearity), the VCO is said to operate in the *voltage mode region* [33]. Increasing the current further in the voltage mode region only serves to increase the noise with no corresponding increase in tank voltage swing. Furthermore, since saturating the VCO in this manner leads to a distorted sinusoidal output voltage waveform, higher-order harmonic content would grow, which would lead to more noise folding and worse phase noise performance, as was demonstrated in Sect. 6.3.2. The power spectral density of a distorted VCO output waveform is shown in Fig. 6.46.

Improving the noise performance once the voltage mode of operation is reached is possible. If the RLC parallel tank resistance, R_p, is dominated by the inductor, then reducing the inductor value would lead to less tank R_p. Consider an LC VCO with inductance L and bias current I. If the swing is maximized such that it is on the edge of the voltage mode of operation and the inductance L is reduced to L/2, the voltage swing is now one-half the original amplitude since R_p has been reduced by a factor of 2. Moreover, the tank is now operating well into the current mode of operation. Increasing the current by a factor of 2 would increase the swing back to the original value. Since the noise (assuming it is dominated by thermal noise) increases with the $\sqrt{I}$, the noise increases by $\sqrt{2}$; however, the voltage noise is obtained by multiplying by R_p^2, which has been reduced by a factor of 2. This leads to

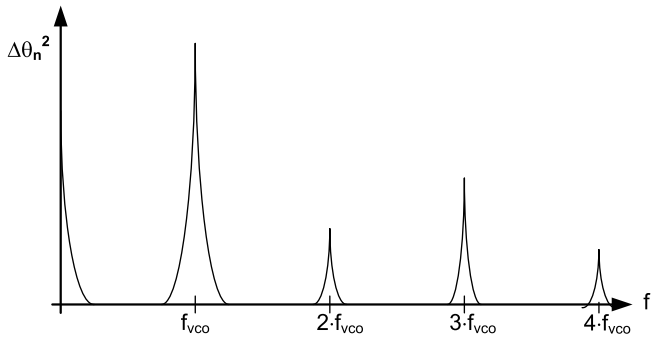


Fig. 6.46 Noise folding effect in a distorted VCO output waveform

$\sqrt{2}$ reduction in noise compared to the original case. This means that the original LC VCO with inductance L and bias current I has 3 dB higher phase noise thank an LC VCO with half the inductance and double the current (assuming the original LC VCO has been optimized to be at the edge of voltage mode of operation).

Summary

In this chapter, wideband frequency synthesizer design has been reviewed. At the heart of a frequency synthesizer is a PLL. The requirements for PLLs in a wireless transceiver were given. This mainly consists of requirements stemming from reciprocal mixing effect as well as aperture jitter in both the ADC and the DAC. Linear analysis of the PLL was given, including full analysis of the effect of the various noise components in a PLL. This was followed by a more in-depth nonlinear noise analysis of both the charge pump and the VCO components in a PLL.

References

1. A. Fahim, Clock Generators for SOC Processors: Circuits and Architectures, Boston: Kluwer Academic Publishers, 2005.
2. D. Shaeffer and T. Lee, The Design and Implementation of Low-Power CMOS Radio Receivers, Boston, Kluwer Academic Publishers, 1999.
3. R. Caverly, CMOS RFIC Design Principles, Norwood, MA, Artech House, 2007.
4. T. Riley, M. Copeland, and T. Kwasniewski, "Delta-sigma modulation in fractional-N frequency synthesis," IEEE J. Solid-State Circuits, vol. 28, no. 5, pp. 553–559, May 1993.
5. N. Dalt, et. al., "On the Jitter Requirements of the Sampling Clock for Analog-to-Digital Converters," IEEE Trans. Circuits and Systems I, vol. 49, no. 9, pp. 1354–1360, September 2002.
6. F. Gardner, "Charge-pump phase-lock loops," IEEE Transactions on Communications, vol. COM –28, pp. 1849–1858, Nov. 1980.
7. R. Best, Phase-Locked Loops: Theory, Design and Applications, New York: Mc Graw-Hill, 1993.
8. V. Manassewitsch, Frequency Synthesizers, New York: John-Wiley, 1987.
9. B. Miller and R. Conley, "A Multiple Modulator Fractional Divider," IEEE Trans. on Instrumentation and Measurement., vol. 40, pp. 578–583, June 1991.
10. R. M. Gray, "Oversampled sigma-delta modulation," IEEE Trans. on Communications, vol. COM-35, no. 4, pp. 481–489, April 1987.
11. Perrott, M. H., Trott, M. D., Sodini, C. G., "A modeling approach for S-D fractional-N frequency synthesizers allowing straightforward noise analysis," IEEE J. of Solid-State Circuits, vol. 37, no. 8, pp. 1028–1038, Aug 2002.
12. B. Wolaver, Phase-Locked Loop Circuit Design, Upper Saddle River, NJ: Prentice-Hall, 1991.
13. W. Egan, Frequency Synthesis by Phase Lock, New York: John-Wiley, 2000.
14. B. Razavi, Design of Analog CMOS Integrated Circuits, McGraw-Hill: Boston, 2001.
15. J. Brugler and P. Jespers, "Charge pumping in MOS devices," IEEE Trans on Electron Devices, vol. ED-16, no. 3, pp. 297–302, March 1969.

16. Yang, H. K. and El-Masry, E. I., "Clock feedthrough analysis and cancellation in current sample/hold circuits," IEE Proceedings of Circuits, Devices and Systems, vol. 141, no. 6, pp. 510–516, June 1994.
17. Bongiorno, J., Graham, D., "An extension of the Nyquist-Barkhausen stability criterion to linear lumped-parameter systems with time-varying elements," IEEE Trans on Automatic Control, vol. 8, no. 2, pp. 166–170, Feb 1963.
18. K. Wark and D. Richards, Thermodynamics, 6th Ed., McGraw-Hill: New York, 1999.
19. D. B. Leeson, "A Simple Model of Feedback Oscillator Noises Spectrum," Proceedings of the IEEE, vol. 54, pp. 329–154, Feb. 1966.
20. L. S. Cutler and C. L. Searle, "Some aspects of the theory and measurement of frequency fluctuations in frequency standards," Proceedings of the IEEE, vol. 54, pp. 136–154, Feb. 1966.
21. S. Haykin, Digital Communications, Wiley: New York, 1988.
22. A. Hajimiri and T. Lee, "A General Theory of Phase Noise in Electrical Oscillators," IEEE J. of Solid-State Circuits, vol. 3, no. 2, pp. 179–194, February 1998.
23. A. Hajimiri and T. Lee, The Design of Low Noise Oscillators, Boston: Kluwer Academic Publishers, 1999.
24. B. Razavi, "A Study of Phase Noise in CMOS Oscillators," IEEE J. of Solid-State Circuits, vol. 31, no. 3, pp. 331–343, March 1996.
25. J. McNeill, "Jitter in Ring Oscillators," IEEE J. of Solid-State Circuits, vol. 32, no. 6, pp. 870–879, June 1997.
26. W. Liu, MOSFET Models for SPICE Simulation, New York: John Wiley & Sons, Inc. 2001.
27. A. Hajimiri, "Noise in Phase-Locked Loops," IEEE Custom Integrated Circuits Conference, pp. 1–6, 2001.
28. T. Lee and A. Hajimiri, "Oscillator phase noise: A tutorial," IEEE J. of Solid-State Circuits, vol. 35, no. 3, March 2000, pp. 326–335.
29. T. Lee, The Design of CMOS Radio-Frequency Integrated Circuits, 2nd edition, Cambridge, UK, Cambridge University Press, 2004.
30. Andreani, P., Mattisson, S., "On the use of MOS varactors in RF VCOs," IEEE J. of Solid-State Circuits, vol. 35, no. 6, pp. 905–910, June 2000.
31. Hanafi, H., et. al., "0.5um CMOS device design and characterization," European Solid-State Device Research Conference, pp. 91–94, 1987.
32. D. Ham and A. Hajimiri, "Concepts and Methods in Optimization of Integrated LC VCOs," IEEE J. of Solid-State Circuits, vol. 36, no. 6, pp. 896–909, June 2001.
33. J. Rael and A. Abidi, "Physical Processes of phase noise in differential LC oscillators," IEEE Custom Integrated Circuits Conference (CICC), 2000, pp. 569–572.

Index

A. Fahim, *Radio Frequency Integrated Circuit Design for Cognitive Radio Systems,*
DOI 10.1007/978-3-319-11011-0