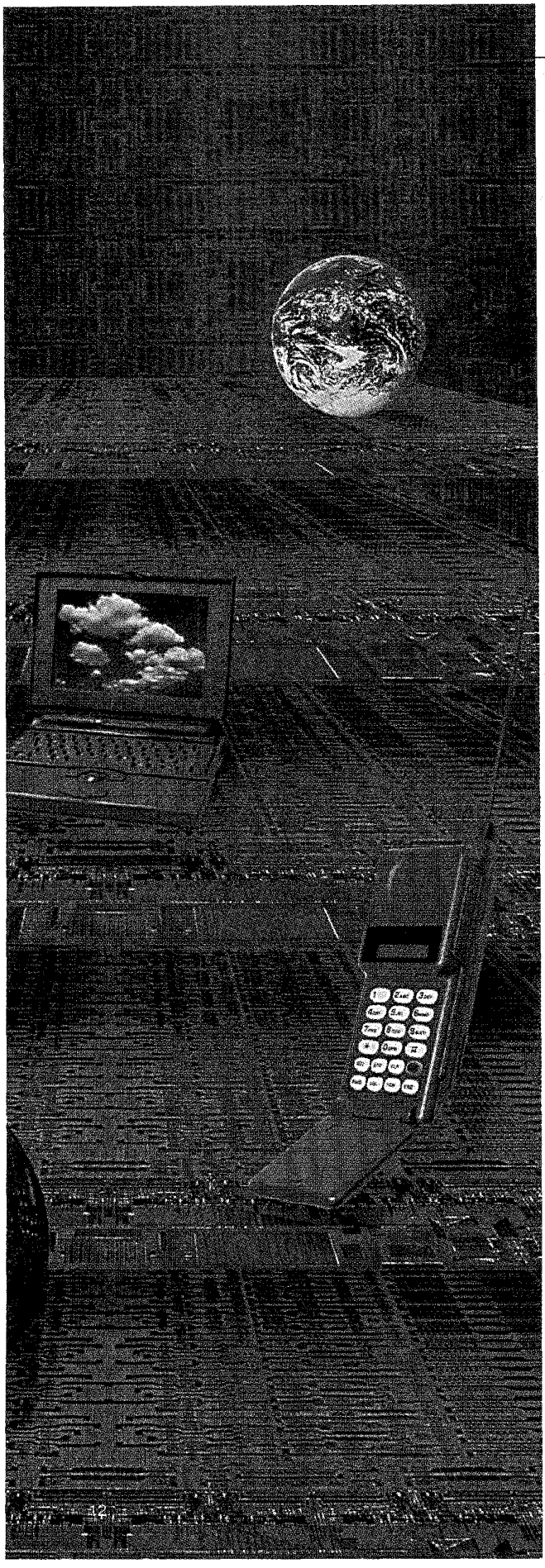


Challenges in Portable RF Transceiver Design

Behzad Razavi

As wireless products such as cellular phones become an everyday part of people's lives, the need for higher performance at lower costs becomes even more important. Overcoming the challenges involved in the design of radio-frequency (RF) transceivers can help meet this need. This article provides an overview of RF electronics in portable transceivers and describes design issues as well as current work toward achieving both high performance and low cost. To understand the implications in the design of RF integrated circuits (ICs) we look at the properties of the mobile communications environment. We then study receiver and transmitter architectures and their viability in present IC technologies. An example of an RF transceiver is given and the design of transceiver building blocks is discussed. We conclude by looking at future directions in RF design.

Wireless Communication Development

Wireless technology came to existence in 1901 when Guglielmo Marconi successfully transmitted radio signals across the Atlantic Ocean. The consequences and prospects of this demonstration were simply overwhelming; the possibility of replacing telegraph and telephone communications with wave transmission through the "ether" portrayed an exciting future. However, while two-way wireless communication did soon materialize in the military, wireless transmission in daily life remained limited to one-way radio and television broadcasting by large, expensive stations. Ordinary, two-way phone conversations would still go over wires for many decades. The invention of the transistor, the development of Shannon's information theory, and the conception of the cellular system — all at Bell Laboratories

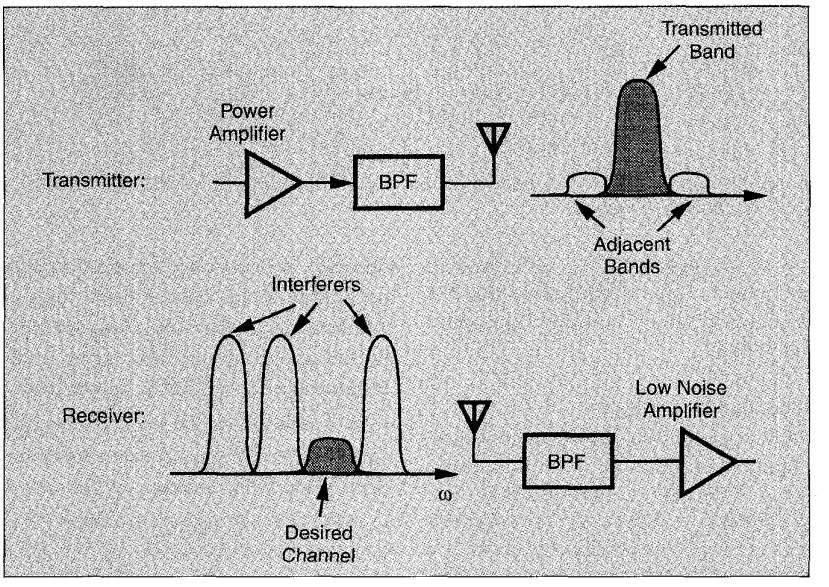
8755-3996/96/$5.00©1996IEEE

— paved the way for affordable mobile communications, as originally implemented in car phones and eventually realized in portable cellular phones (cell phones).

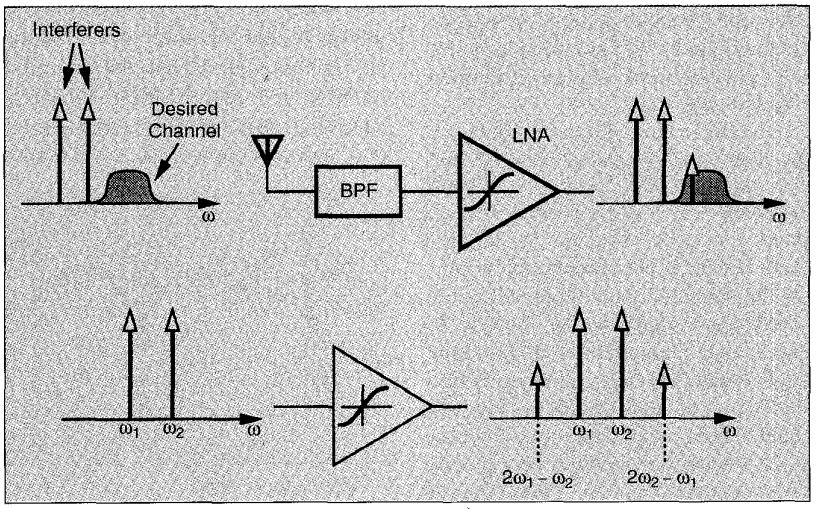
But, why the sudden surge in wireless electronics? Market surveys show that in the United States more than 20,000 people join the cellular phone system *every day*, motivating competitive manufacturers to provide phone sets with increasingly higher performance and lower cost. In fact, the present goal is to reduce both the power consumption and price of cell phones by 30% every year — although it is not clear for how long this rate can be sustained. A more glorious prospect, however, lies in the power of two-way wireless communication when it is introduced in other facets of our lives: home phones, computers, facsimile, and television.

While an immediate objective of the wireless industry is to combine cordless and cellular phones to allow seamless communications virtually everywhere, the long-term plan is to produce an "omnipotent" wireless terminal that can handle voice, data, and video as well as provide computing power. Other luxury items such as the global positioning system (GPS) are also likely to become available through this terminal sometime in the future. Personal communication services (PCS) are almost here.

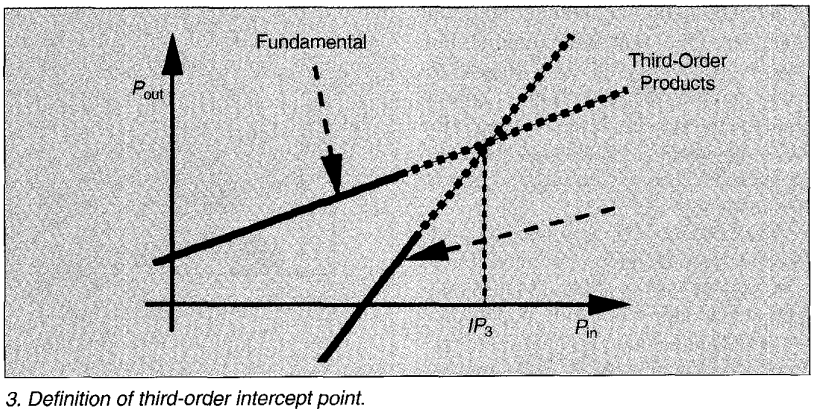
Today's pocket phones contain more than one million transistors, with only a very small fraction operating in the RF range and the rest performing low-frequency baseband signal processing. However, the RF section is still the design bottleneck of the entire system. This is primarily for three reasons. First, while digital circuits directly benefit from advances in integrated-circuit (IC) technologies, RF (analog) circuits do not benefit as much because they suffer from many more trade-offs and often require external components (such as inductors) that are difficult to bring onto the chip even in modern fabrication processes. Second, in contrast to other types of analog circuits, proper RF design demands a solid understanding of many areas that are not directly related to integrated circuits, e.g., microwave theory, communication theory, analog and digital modulation, transceiver architectures, etc. Each of these disciplines has been under development for many decades, making it difficult for an IC designer to acquire the necessary knowledge in a short time. Third, computer-aided analysis and synthesis tools for RF are still in their infancy,



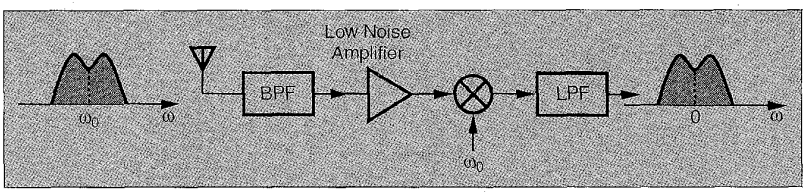
1. Simple RF front end.



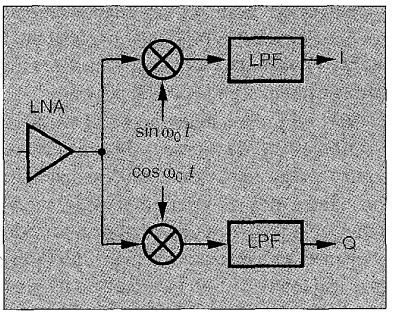
2. Effect of third-order nonlinearity in LNA.



3. Definition of third-order intercept point.



4. Simple homodyne receiver.



5. Homodyne receiver with quadrature down-conversion.

forcing the designer to rely on experience and intuition to predict the performance. For these reasons, RF IC designers have been a rare species.

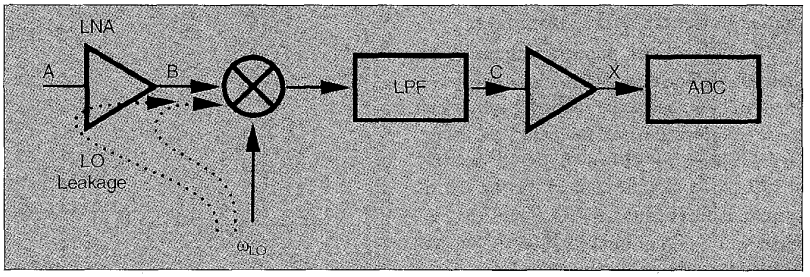
Wireless Environment

The wireless communications environment, especially in urban areas, is often called "hostile" because it imposes severe constraints upon the transceiver design. Perhaps the most important constraint is the limited spectrum allocated by regulatory organizations to wireless users. From Shannon's theorem, this translates to a limited rate of information, mandating the use of sophisticated techniques such as coding, compression, and bandwidth-efficient modulation, even for voice signals.

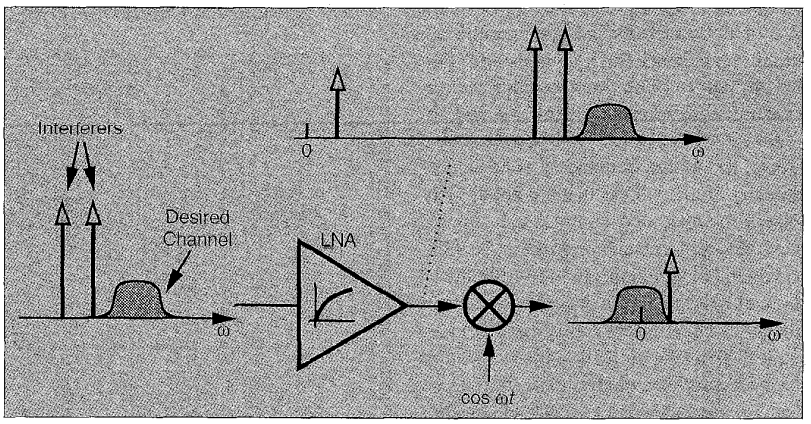
The narrow bandwidth available to each user also impacts the design of the RF front end. As depicted in Fig. 1, the transmitter must employ narrowband amplification and filtering to avoid "leakage" to adjacent bands, and the receiver must be able to process the desired channel while sufficiently rejecting strong neighboring channels. To gain a better feeling about the latter issue, we note that if the front-end bandpass filter (BPF) in a 900-MHz receiver is to provide 60 dB of rejection at 45 kHz from the center of the channel, then the equivalent Q of the filter is on the order of 10^7, a value difficult to achieve even in surface acoustic wave (SAW) filters. Since typical filters exhibit a trade-off between the loss and the Q and since in receiving very small signals the loss must be minimized, the out-of-channel rejection of the front-end filters is usually insufficient, requiring further filtering in the following stages (typically at lower center frequencies). This will be clarified later in this article.

The existence of large unwanted signals in the vicinity of the band of interest even after filtering creates difficulties in the design of the following circuits, in particular the front-end low-noise amplifier (LNA). As shown in Fig. 2, if the LNA exhibits nonlinearity, then the "intermodulation products" of two strong unwanted signals may appear in the desired band, thereby corrupting the reception. As a simple example, we note that if the input/output static characteristic of the LNA is approximated as $y(t) = \alpha_1 x(t) + \alpha_2 x^2(t) + \alpha_3 x^3(t)$ and $x(t) = A_1 \cos\omega_1 t + A_2 \cos\omega_2 t$, then the cubic term yields components at $2\omega_1 - \omega_2$ and $2\omega_2 - \omega_1$, either of which may fall in the band. The standard approach to quantifying this effect is to choose $A_1 = A_2$ and, using extrapolation, calculate the input power that results in equal magnitudes for the fundamental components and the intermodulation products (Fig. 3). Such value of input power is called the "third-order intercept point" (IP_3). It is interesting to note that this type of nonlinearity is important even if the signal carries information in its phase or frequency rather than in its amplitude.

Another important issue in the design of wireless receivers is the dynamic range of the input signal. Typically around 100 dB (a factor of 100,000 for voltage quantities), the dynamic rage is limited by a lower bound due to noise and an upper bound due to nonlinearities and saturation. The minimum detectable signal in today's handsets is in the vicinity of –110 dBm ($\approx$0.71 μV_{rms} in a 50-Ω system), thus demanding very low noise in the receive path. For the upper bound, the receiver must achieve a high



6. LO leakage to input.



7. Effect of second-order distortion.

linearity so as to minimize intermodulation products. Also, saturation effects at high input levels often mandate the use of gain control in various parts of receivers.

Receiver Architectures

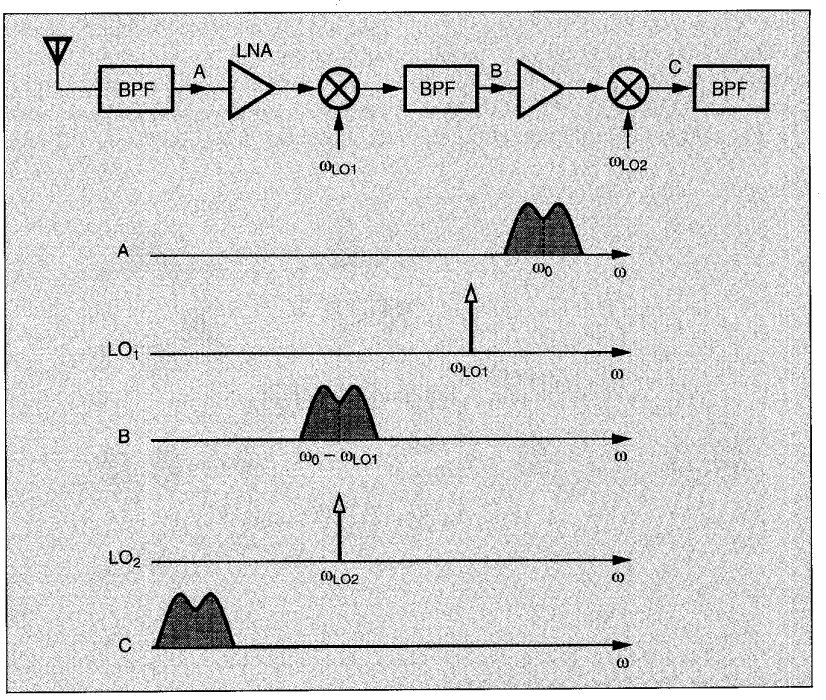
Complexity, cost, power dissipation, and the number of external components have been the primary criteria in selecting receiver architectures. As IC technologies evolve, architectures that once seemed impractical may return because, when they are implemented in today's advanced processes, their advantages outweigh their drawbacks.

Homodyne Architecture

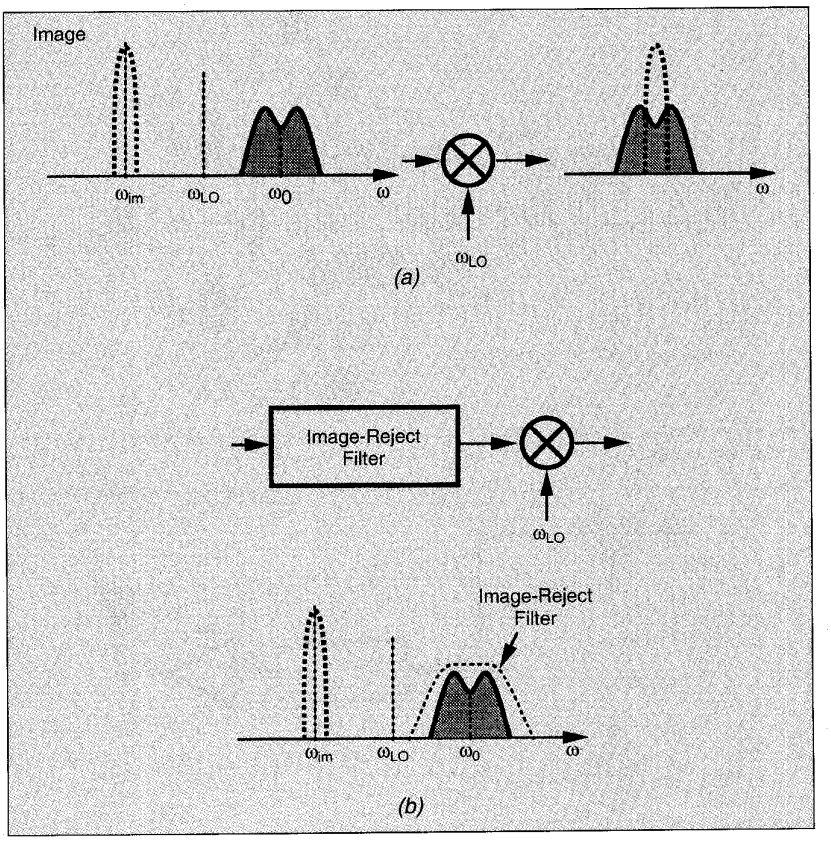
Also called "direct conversion" architecture, the homodyne receiver is the natural topology for downconverting a signal from RF to baseband. The idea is simply to mix the RF signal with a local oscillator (LO) output and low-pass filter the result such that the center of the band of interest is translated *directly* to zero frequency (Fig. 4). Because of its typically high noise, the mixer is usually preceded by an LNA. Also, in phase and frequency modulation schemes, the RF signal is mixed with both the LO output and its quadrature so as to provide phase information (Fig. 5).

The simplicity of the homodyne architecture makes it attractive for compact, efficient implementation of RF receivers [1, 2]. However, several issues have impeded its widespread use. We briefly describe these issues and their impact on the design of related ICs.

DC Offsets. Since in a homodyne receiver the downconverted band extends to the vicinity of the zero frequency, extraneous offset voltages can corrupt the signal and, more importantly, saturate the following stages. To understand the origin and impact of offsets, consider the more realistic circuit shown in Fig. 6. Here, the mixer is followed by a low-pass filter, a post-amplifier, and an analog-to-digital converter (ADC). We make two observations: (1) The isolation between the LO and RF ports of the mixer is not perfect; due to capacitive coupling and, if the LO signal is supplied externally, bond wire coupling, a finite amount of feedthrough exists from the LO port to points A and B. This effect is called "LO leakage." The leakage signal appearing at the input of the LNA is amplified and mixed with the LO signal, thus producing a DC component at point C. This phenomenon is



8. Heterodyne architecture.



9. (a) Problem of image, (b) image rejection by filtering.

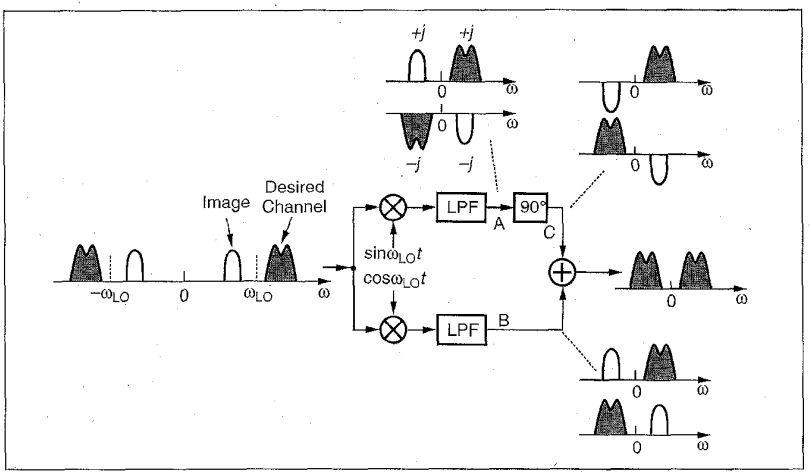
called "self-mixing." (2) The total gain from the antenna to point X is typically around 100 dB so that the microvolt input signal reaches a level that can be digitized by a low-cost ADC. Of this gain, approximately 25 to 30 dB is contributed by the LNA/mixer combination.

With the above observations and noting that the LO power is typically around 0 dBm (approximately 0.6 V_{pp}), and the LO leakage to point A on the order of –60 dB, we infer that the DC component at the output of the mixer due to self-mixing is roughly equal to 0 dBm –60 dB + 30 dB = –30 dBm, corresponding to a level of 10 mV. We also note that the signal level at this point can be as low as 25μV_{rms}. Thus, if directly amplified by the remaining gain of 70 dB, the DC component saturates the following circuits, prohibiting the amplification of the desired signal.

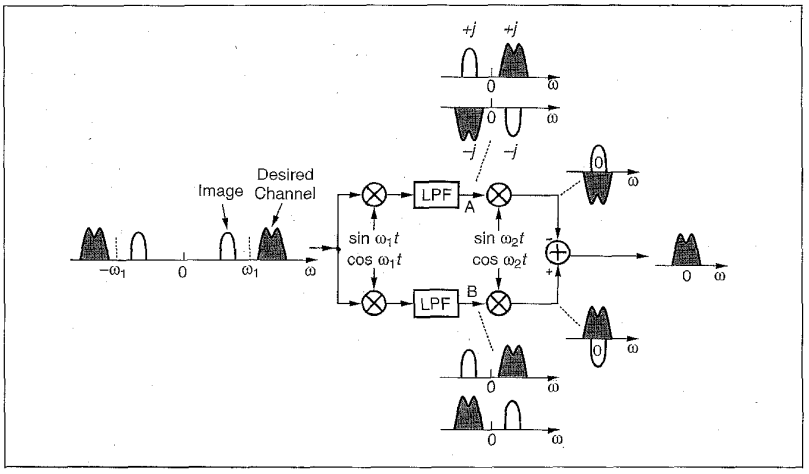
While high-pass filtering (i.e., AC coupling) may seem the solution here, in most of today's modulation schemes the spectrum contains information at frequencies as low as a few tens of hertz, mandating a very low corner frequency in the filter. In addition to difficulties in implementing such a filter in IC form, a more fundamental problem is its slow response, an important issue if the offset varies quickly. This occurs, for example, when a car moves at a high speed and the LO leakage reflections from the surrounding objects change the offset rapidly.

For these reasons, homodyne receivers require sophisticated offset-cancellation techniques. In [3], for example, the offset in the analog signal path is reduced by feeding information from the baseband digital signal processor (DSP). Alternatively, modulation schemes can be sought that contain negligible energy below a few kilohertz [4].

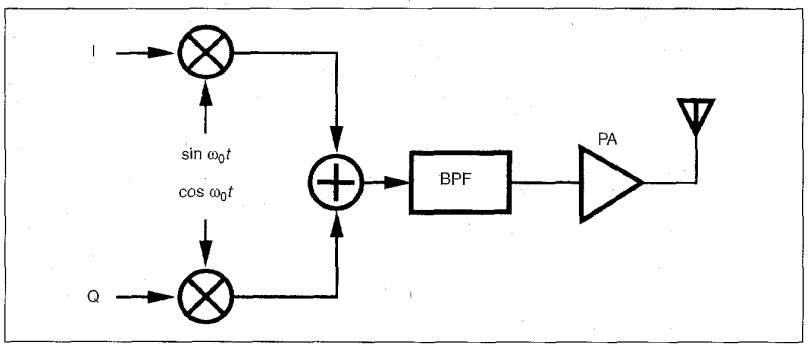
Even-Order Distortion. While third-order mixing was considered as a source of interference in Fig. 2, even-order distortion also becomes problematic in homodyne downconversion. As depicted in Fig. 7, if two strong interferers close to the channel of interest experience a nonlinearity such as $y(t) = \alpha_1 x(t) + \alpha_2 x^2(t)$, then they are translated to a *low* frequency before the mixing operation and the result passes through the mixer with finite attenuation. This is because, in the presence of mismatches that degrade the symmetry of the mixer, the mixing operation can be viewed as $x(t)(a + A \cos \omega t)$, indicating that a fraction of $x(t)$ appears at the output without frequency translation. A similar effect occurs if the LO output duty cycle deviates from 50%. Another issue is that the second harmonic of the input signal (due to the square term in the above equation) is mixed with the second harmonic of the LO output, thereby appearing in the baseband and interfering with the actual signal [5]. For these reasons, even-or-



10. Image rejection using single-sideband mixing.



11. Weaver architecture.



12. Direct conversion transmitter.

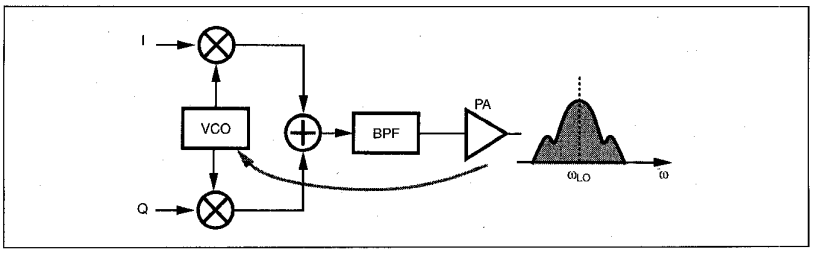
der intermodulation corrupts the baseband signal.

I-Q Mismatch. As mentioned above, in most phase and frequency modulation schemes the downconversion path must use quadrature mixing. The required I and Q phases of the LO raise an issue related to the mismatches between these two signals. If the amplitudes of the I and Q outputs are not equal or their phase difference deviates from 90°, the error rate in detecting the baseband signal rises. The task of generating I and Q phases with precise matching is discussed later.

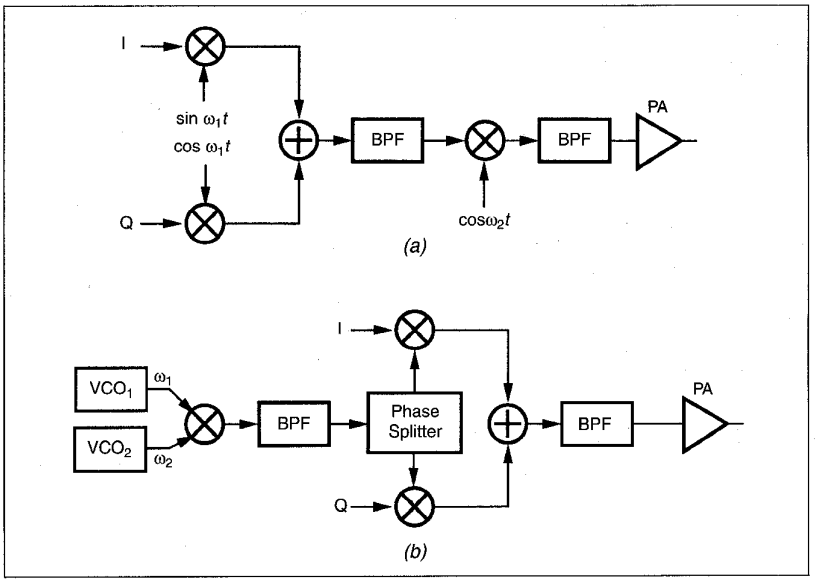
A second-order effect arises from the mismatches between the two mixers themselves. Since the mixers process high-frequency signals here, their phase and gain are sensitive to parasitics and hence susceptible to mismatches.

LO Leakage. In addition to introducing DC offsets, leakage of the LO signal to the antenna and radiation therefrom creates interference in the band of *other* receivers. The design of the wireless infrastructure and the regulations of the Federal Communications Commission (FCC) impose upper bounds on the amount of LO radiation, typically between –60 to –80 dBm.

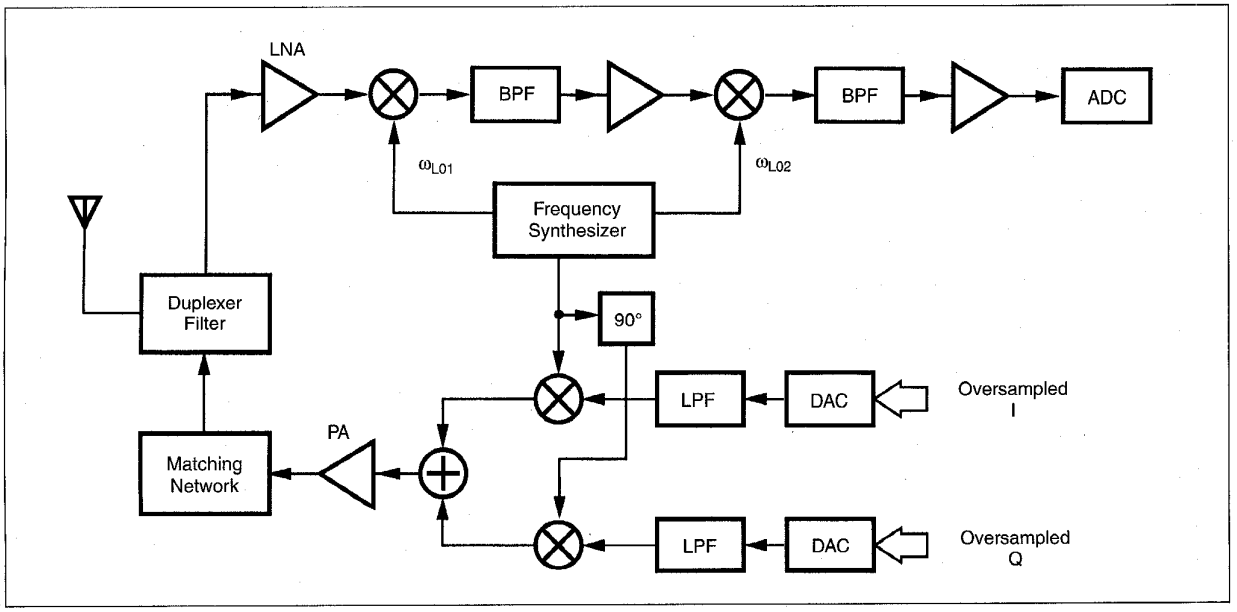
Flicker Noise. Owing to the limited gain provided by the LNA and the mixer, the downconverted signal is relatively small and quite sensitive to noise. Since device flicker noise becomes significant at low frequen-



13. Disturbance of VCO by PA in direct conversion transmitter.



14. Alternative transmitter architectures. (a) Two-step conversion, (b) Offset VCO.



15. Representative RF transceiver.

cies, amplification of the baseband signal with low noise is an important issue.

Heterodoyne Architectures

The design issues mentioned above for the homodyne receiver have motivated the invention of other architectures. Most commonly used is the heterodyne topology. (In this article, we do not make a distinction between "heterodyne" and "superheterodyne.") Illustrated in Fig. 8 in a simple form, a heterodyne receiver first downconverts the input to an "intermediate frequency" (IF). The resulting signal is subsequently bandpass filtered, amplified, and downconverted again. In the case of digital modulation, the last downconversion generates both I and Q phases of the signal.

The heterodyne architecture alleviates two of the homodyne reception issues by avoiding them at high frequencies or low signal levels. The effect of DC offsets of the first few stages is removed by bandpass filtering, and that of the last stage is suppressed by the total gain in the preceding stages. Also, I and Q mismatches occur at much lower frequencies and are therefore easier to control and correct. As for the LO leakage, since ω_{LO1} is out of the band of interest, it is suppressed by the front-end BPF and its radiation from the antenna is less objectionable.

Perhaps the most important feature of the heterodyne receiver is its selectivity, i.e., the capability to process and select small signals in the presence of strong interferers. While selecting a 30-kHz channel at a center frequency of 900 MHz requires prohibitively large Qs, in Fig. 8 bandpass filtering is performed at progressively lower center frequencies. For example, the third BPF may operate at a center frequency of 400 kHz, thereby providing high selectivity for a 30-kHz channel. In other words, the filters have much more relaxed requirements.

Despite the above merits, heterodyning entails a number of drawbacks. The most significant problem is the "image frequency." Since a simple mixer does not preserve the polarity of the difference between its input frequencies, it translates the bands both *above* and *below* the carrier to the same frequency [Fig. 9(a)]. Thus, the mixing operation must be preceded by an "image reject" filter [Fig. 9(b)], usually a passive one.

The issue of image rejection leads to an interesting trade-off among three parameters: the amount of image noise, the spacing between the band and the image (= 2 *IF*), and the *loss* of the filter. To minimize the image noise, we can either increase the IF (so that the filter provides more attenuation at the image frequency) or tolerate greater loss in the filter while increasing its Q. Since the LNA gain is typically less than 15 dB, the filter loss should not exceed a few dB, and the trade-off reduces to one between the image noise and the value of IF.

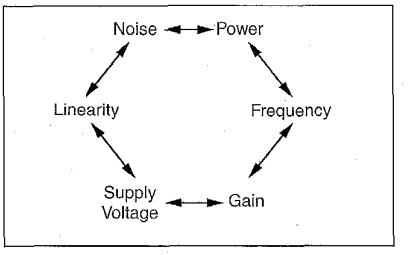
How high can the IF be? Recall from Fig. 8 that the filter following the first mixer must select the band. As the IF and, hence, the center frequency of this filter increase, so does the required Q, thereby imposing a fundamental trade-off between image rejection and channel selection. For the 900-MHz and 1.8-GHz bands, typical IFs range from 70 MHz to 200 MHz.

Another drawback of the heterodyne architecture is that the LNA must drive a 50-Ω impedance because the image-reject filter cannot be integrated and is therefore placed off-chip. This adds another dimension to the trade-offs among noise, linearity, gain, and power dissipation of the amplifier, further complicating the design. The image-reject and channel-select filters are typically expensive and bulky, making the heterodyne approach less attractive for small, low-cost wireless terminals. Nevertheless, heterodyning has been the dominant choice for many decades [6, 7].

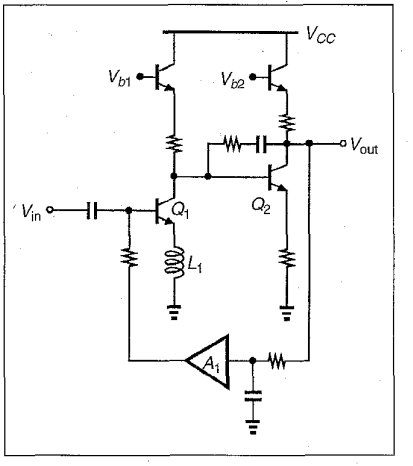
Image-Reject Architectures

The issues related to the image-reject filter have motivated RF designers to seek other techniques of rejecting the image in a heterodyne receiver. One such technique originates from a single-sideband modulator introduced by Ralph Hartley in 1928 [8]. Illustrated in Fig. 10, Hartley's circuit mixes the RF input with the quadrature outputs of the local oscillator, low-pass filters the resulting signals, and shifts one by 90° before adding them together. The reader can easily verify that if the input is equal to $A_{RF}\cos\omega_{RF}t+A_I\cos\omega_I t$, where ω_I is the image frequency, then the output is proportional to $A_{RF}\cos(\omega_{LO}-\omega_{RF})t$. As a more general case, we consider the input spectrum shown in Fig. 10 and note that mixing with $\sin\omega_{LO}t$ and $\cos\omega_{LO}t$ yields the spectra of Fig. 10 at nodes A and B, respectively (the factor $\pm j$ in these spectra is to indicate convolution with $\pm j\delta(\omega\pm\omega_{LO})/2$ (spectrum of $\sin\omega_{LO}t$)). Since a phase shift of +90° in the signal at A corresponds to multiplication by $+j$ and inverting the positive frequencies, we obtain the four spectra at nodes B and C as the inputs to the adder. The output is therefore free from the image.

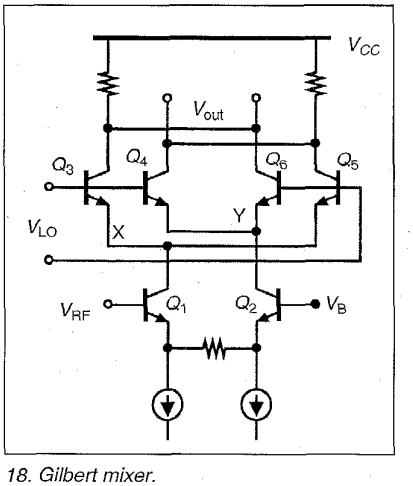
The principal drawback of image-reject mixers is their sensitivity to mismatches. For example, if the phase difference between the LO quadrature phases deviates from 90°, the cancellations shown in Fig. 10 are imperfect and some image noise corrupts the down-



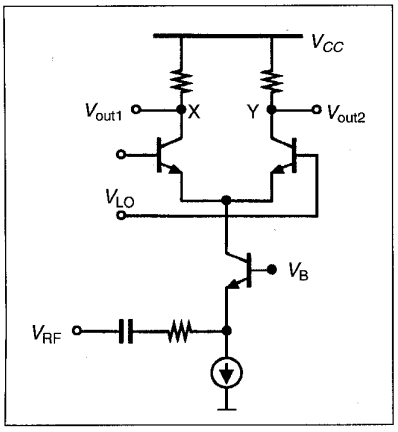
16. RF design hexagon.



17. Low-noise amplifier.



18. Gilbert mixer.



19. Single-balanced mixer.

converted signal [5]. For typical matching in IC technologies, the image is rejected by about 30 to 40 dB [9]. Another important issue is the higher power dissipation and/or noise due to the use of two high-frequency mixers. Also, circuits that shift the down-converted signal by 90° generally suffer from trade-offs among linearity, noise, and power dissipation.

Weaver Architecture

In our discussion of images, we noted that *any* frequency translation leads to corruption of the signal by the image, except when a symmetric band is brought down to zero frequency (homodyne). The Weaver technique allows an arbitrary translation of the signal band without image interference [10].

Illustrated in Fig. 11, this approach down-converts the signal in two steps. In the first step, the input is mixed with the quadrature phases of the first local oscillator and the result is low-pass filtered, yielding the spectra at nodes A and B. In the second step, these signals are translated to zero frequency and added together, thereby effecting image cancellation.

The important advantage of the Weaver architecture is that it does not require high-Q bandpass filters. Even though the LPFs shown in Fig. 11(a) must be preceded by capacitive coupling to eliminate DC offsets (similar to homodyne) and, as such, the combination is a bandpass filter, the out-of-band rejection of these filters is quite relaxed. Note that some amplification is necessary before the second set of mixers to reduce the effect of their noise. The Weaver method suffers from the same drawback as the image-reject mixer: incomplete cancellation of the image in the presence of mismatches.

Transmitter Architectures

In contrast to the variety of approaches invented for RF reception, transmitter architectures are found in only a few forms. This is because issues such as image rejection and band selectivity are more relaxed in transmitters, leaving the output power amplifier (PA) design as the primary challenge.

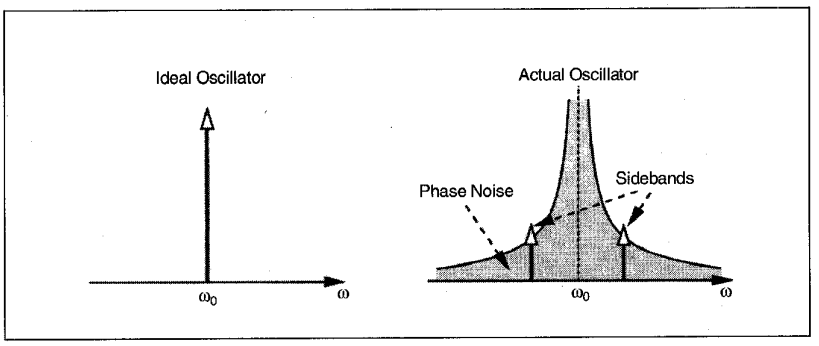
A simple direct conversion transmitter is shown in Fig. 12. Here, the baseband signal is mixed with the LO output and the result is bandpass filtered and applied to the PA. A matching network is usually interposed between the PA and the antenna to allow maximum power transfer and filter out-of-band components that result from nonlinearities in the amplifier. Note that since the baseband signal is produced in the transmitter and is therefore sufficiently strong, the noise of the mixers is much less critical here than in receivers.

Direct conversion architectures suffer from an important drawback: disturbance of the transmit local oscillator by the output PA. Illustrated in Fig. 13, this issue arises because the PA output is a modulated waveform with high power and a spectrum centered around the voltage-controlled oscillator VCO frequency. Thus, despite various shielding techniques that attempt to isolate the VCO, the "noisy" output of the PA still corrupts the oscillator spectrum. (The actual mechanism of this corruption is called "injection pulling" or "injection locking." When disturbed by a close interferer at frequency ω_i, an oscillator operating at ω_0 tends to shift to ω_i.) This problem worsens if the PA is turned on and off periodically to save power.

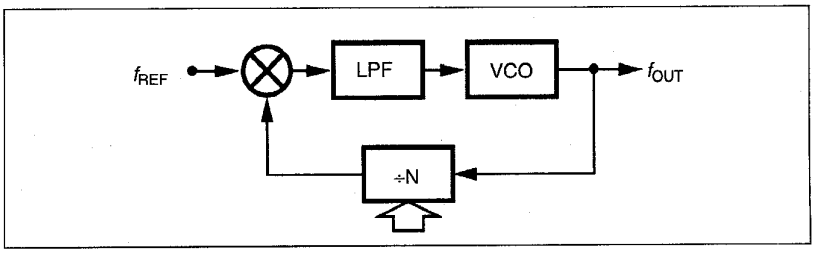
The above difficulty is alleviated if the PA output spectrum is sufficiently higher or lower than the VCO frequency. For example, as shown in Fig. 14(a), the upconversion can be performed in two steps, generating a final spectrum that differs from ω_2 by ω_1 [11]. Alternatively, the VCO frequency can be "offset" by adding or subtracting the output frequency of another oscillator (Fig. 14(b)) [7]. Note that in both cases, some filtering is required to reject unwanted parts of the spectrum.

The most difficult part of transmitters to design is the PA, mainly because of severe trade-offs among its efficiency, linearity, and supply voltage. In typical PA topologies, the efficiency drops as the circuit is designed for higher linearity or lower supply voltage. For a typical peak output power of 1 W, an efficiency of 50% means that an additional 1 W is wasted, which is a substantial amount with respect to the power dissipation of the rest of a portable phone.

The reader may wonder why the linearity of the PA is important if only the phase of the carrier is modulated. Indeed in analog



20. Phase noise and sidebands in the output of oscillators.



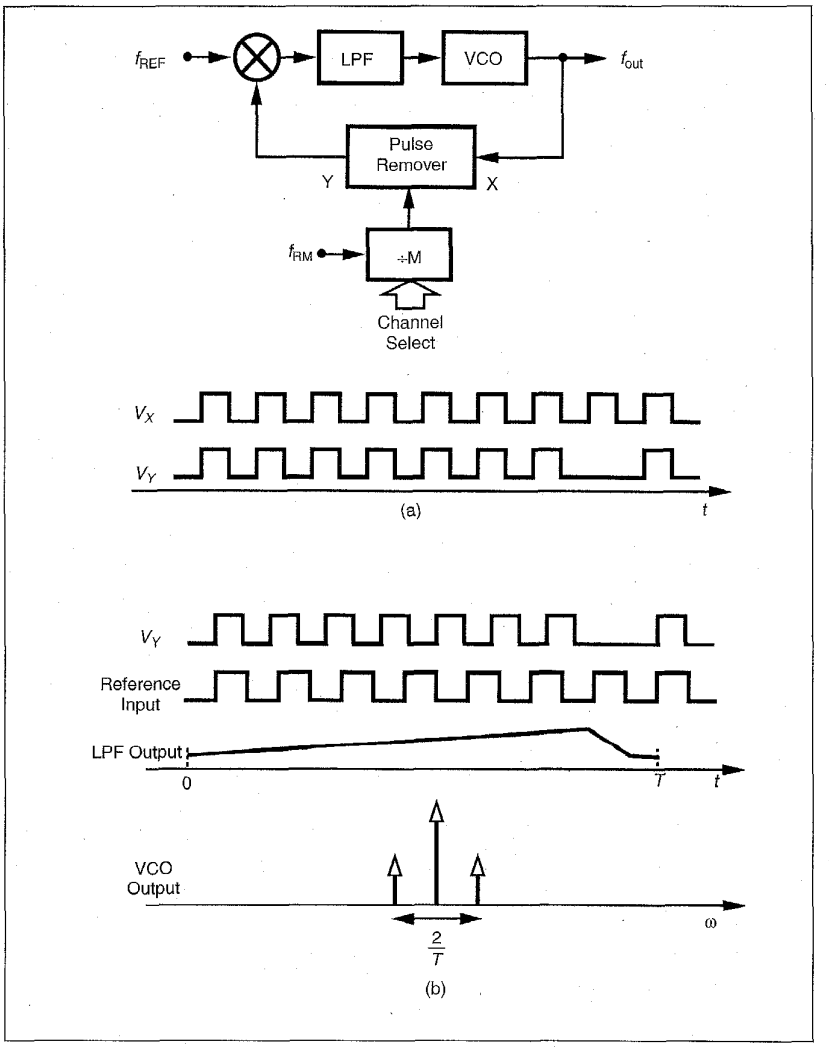
21. Pulse swallow synthesizer.

FM systems, the linearity is not critical and the efficiency trades only with the supply voltage, usually approaching 60% at the peak output power. On the other hand, in digital modulation schemes such as quadrature phase shift keying (QPSK) the situation is more complicated. Since a QPSK signal has a relatively wide spectrum, it usually undergoes bandpass filtering to limit its bandwidth to that of one channel. The resulting signal, however, does not have a constant envelope, i.e., it exhibits some amplitude modulation. Now, if this signal experiences nonlinear amplification, its spectrum widens, spilling into adjacent channels and defeating the purpose of bandpass filtering.

In order to resolve this issue, RF system designers have employed two different strategies. First, they have found digital modulation schemes in which the envelope of the signal remains constant after filtering and, hence, the spectrum does not widen in the presence of PA nonlinearities. These schemes are known as "continuous phase modulation," where the phase of the carrier varies smoothly from one bit to the next. Second, they have devised feedback and feedforward circuit techniques to improve the linearity of PAs with negligible degradation in efficiency [13, 14, 21].

Overall System

With the above discussion of transceiver architectures, we can now consider a more complete system. Shown in Fig. 15 is a transceiver with heterodyning in the receive path and direct conversion in the transmit path. The transmit VCO may employ the offset technique of Fig. 14(b) to avoid injection pulling.



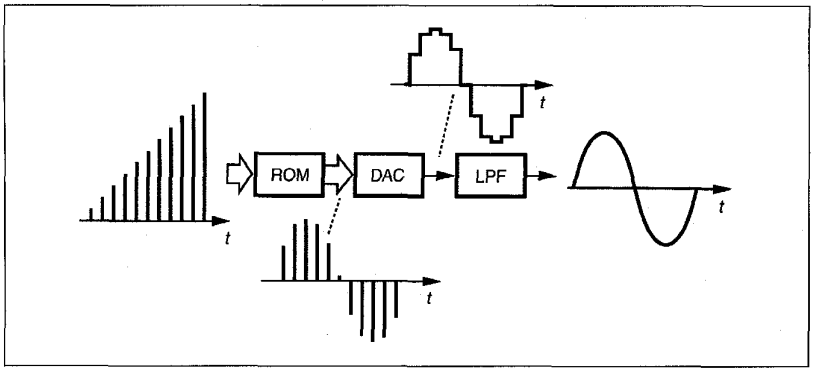
22. (a) Fractional-N synthesizer, (b) problem of reference sidebands.

In most mobile phone systems, the transmit and receive bands are different, with the translation performed at the base station. In a full-duplex system (where reception and transmission occur simultaneously through a single antenna), this is necessary because the two paths must be somehow separated. With two different bands, this is accomplished by a narrowband front-end filter, called the "duplexer." This filter also suppresses out-of-band noise and interference in the receive path.

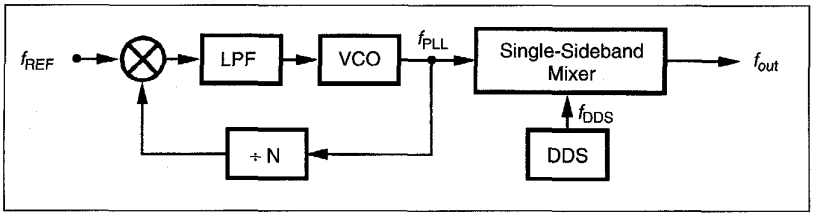
In Fig. 15, the receive and transmit LOs are embedded in a frequency synthesizer. When initiating a call, a mobile unit is assigned two communication channels (for receive and transmit) by the base station. The synthesizer selects the proper carrier frequency for each channel according to a digital input. The important issues here are how "pure" the synthesizer output is, and how fast can it can switch the LO frequency from one channel to another. We return to these issues in the section on frequency synthesizers.

In the receive path, the downconverted signal is applied to an ADC. The ADC is necessary even if the information lies in the phase (or frequency), because baseband operations such as equalization, matched filtering, and despreading are performed with higher precision in the digital domain than in the analog domain. Digital signal processors have thus become an integral part of wireless transceivers.

In the transmitter, the digitized voice undergoes compression and coding. The resulting stream of ONEs and ZEROs is subsequently oversampled and subdivided into multi-bit words, which are then applied to two digital-to-analog converters (DACs) (Fig. 15). This operation takes place for an interesting reason. In digital modulation schemes, the ideal pulse shape for each bit produced in the baseband is quite different from a rectangular function. For example, as mentioned earlier, the modulated carrier may need to be such that its envelope remains constant after filtering. Thus, it is usually necessary to convert the rectangular pulses to another shape. Furthermore, the bandpass filtering required after modulation



23. Direct digital synthesizer.



24. Phase-locked synthesizer with DDS offset mixing.

can be equivalently performed as low-pass filtering on the baseband signal.

Building Blocks

RF architectures impose severe requirements upon the performance of their constituent circuits. The very small signal amplitude received by the antenna in the presence of large interferers mandates both careful allocation of noise and linearity to various stages and sufficient suppression of spurious components generated in the frequency synthesizers and the PA.

As with most analog systems, RF circuits suffer from trade-offs among various parameters. We illustrate such trade-offs in an "RF design hexagon" (Fig. 16), where six important circuit parameters are shown to trade each other. It is interesting to note that in some cases, e.g., in power amplifiers, if the supply voltage is reduced, the power dissipation may *increase*. For this reason, supply scaling in RF circuits lags behind that in digital circuits. The RF design hexagon also indicates that simple figures of merit such as the transit frequency, f_T, unity power gain frequency, f_{max}, and gate delay cannot be easily used to predict RF performance because they do not reflect many of the trade-offs.

In addition to the six parameters shown in Fig. 16, another factor influences the choice of RF circuits and technologies: cost. The design of today's wireless phones begins with a cost (and weight) target, placing very strict limitations on the choice of each building block. While higher levels of integration can reduce the overall cost, issues such as substrate and supply coupling and the ever-existing need for external components (e.g., inductors) have made the progress difficult. In this section, we describe some of the important building blocks of radio systems to familiarize the reader with their design issues and their role in the overall system.

LNAs and Downconversion Mixers

As the first stages to handle the received band, LNAs and downconversion mixers carry the heaviest burden in terms of noise and linearity. In a typical heterodyne front end (Fig. 8), the duplexer introduces a loss of 2 to 3 dB, in effect "magnifying" the noise of the LNA by the same amount when it is referred to the antenna port. Furthermore, the LNA must drive the 50-Ω input impedance of the image-reject filter while providing a reasonable gain. It is important to note that the LNA gain must be chosen according to the noise *and* linearity of the mixer. If this gain is too low, the mixer noise dominates the overall noise figure, and if it is too high, the input signal to the mixer creates large intermodulation products. For these reasons, the design of the LNA is very critical.

The very low noise required of the LNA usually mandates the use of only one active device at the input without any (high-frequency) resistive feedback. In order to provide sufficient gain while driving 50 Ω, LNAs typically employ more than one stage. An interesting example is shown in Fig. 17 [15], where the first stage utilizes a bondwire inductance of 1.5 nH to degenerate the common-emitter amplifier without introducing additional noise. This technique both linearizes the LNA and makes it possible to achieve a 50-Ω input impedance. Bias voltages V_{b1} and V_{b2} and the low-frequency feedback amplifier A_1 are chosen so as to stabilize the gain against temperature and supply variations. The circuit exhibits a noise figure of 2.2 dB, an IP_3 of −10 dBm, and a gain of 16 dB at 900 MHz.

The issue of linearity becomes more significant in mixers because they must handle signals that are amplified by the LNA. While it may seem that the issue of noise is relaxed by the same factor, in practice, (active) mixers exhibit much higher noise simply because they employ more devices in the signal path than do LNAs and suffer from various noise frequency folding effects. As an example, consider the Gilbert cell mixer shown in Fig. 18. Since the LNA output is usually single-ended, the base of Q_2 is connected to a reference voltage. The RF input stage is resistively degenerated to provide sufficient linearity, but at the cost of higher input noise. Now consider the four switching devices Q_3-Q_6. During switching, all of these devices are on for part of the period (if the LO waveform is not an ideal rectangular signal), thereby contributing both shot noise and base resistance thermal noise to the output. Furthermore, even when the switching is complete, the devices that are on (e.g., Q_3 and Q_4) continue to introduce noise because the capacitance at nodes X and Y is quite large. For these reasons, the noise figure of a Gilbert cell with reasonable linearity usually exceeds 10 dB.

Shown in Fig. 19 is a simpler mixer with single-ended RF input [15]. This circuit achieves a noise figure of 15.8 dB with an IP_3 of +6 dBm. An interesting point should be mentioned regarding the noise behavior of this circuit with differential or single-ended outputs. We note that if the output is taken from X with respect to ground, then the mixer operation can be viewed as multiplication of the input signal by a square wave toggling between 0 and A, where A is the

gain of the circuit. Such a square wave has a DC component equal to $A/2$, thereby amplifying *low-frequency* noise components in the signal path by roughly the same factor. Thus, the output noise originates from both downconverted high-frequency noise and amplified "feedthrough" low-frequency noise. In contrast, if the output is taken from X with respect to Y, the equivalent multiplicative square wave toggles between $-A$ and $+A$, yielding zero gain for low-frequency noise in the RF path. In practice, mismatches between Q_2 and Q_3 or deviation of the LO duty cycle from 50% yield a finite feedthrough of low-frequency noise.

Frequency Synthesizers

The RF synthesizer in a transceiver generates precisely spaced carrier frequencies according to a digital input. For example, in the NADC standard, the receive channels are 30 kHz apart and range from 869 MHz to 894 MHz, indicating the very high precision required in defining each channel frequency. If the frequency error is 10 parts per million, then each channel is offset by 9 kHz, an appreciable value with respect to the bandwidth of 30 kHz.

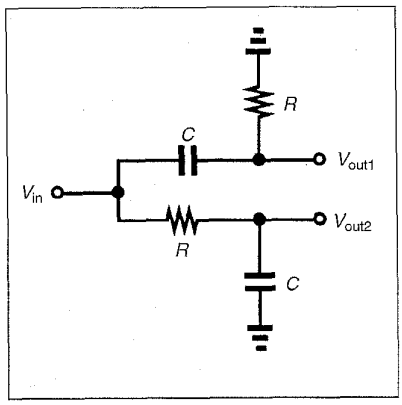
In addition to precision, the "purity" of the carrier signal is also critical. For an ideal oscillator operating at ω_0, the spectrum assumes the shape of an impulse whereas for an actual oscillator, the spectrum exhibits "skirts" as well as "sidebands" around the center or "carrier" frequency (Fig. 20). A simple unlocked oscillator usually exhibits no sidebands, but when embedded in a synthesizer, it may. In typical transceivers, the sidebands must be about 60 dB below the carrier.

A common synthesizer architecture is depicted in Fig. 21, where the output frequency of a VCO is divided by N and locked to an accurate, crystal-based reference frequency, f_{REF}, thereby generating $f_{out} = Nf_{REF}$ [16, 17]. To vary the output frequency, the division factor is changed by means of the channel-select bits. For example, if f_{REF} = 30 kHz and N varies from 29,000 to 29,800, the output frequency covers the range 870 MHz to 894 MHz. In practice, the mixer is replaced with a phase/frequency detector to ensure lock despite process and temperature variations in the VCO center frequency. The primary drawback of this architecture is its slow lock behavior. This is because the loop bandwidth is typically an order of magnitude less than f_{REF}, yielding a long transient response when the channel is changed.

Shown in Fig. 22(a) is a fast-locking architecture, called the "fractional-N" topology. The idea here is that if one pulse is removed from a periodic waveform every T seconds, then the "average" frequency is reduced by $1/T$ hertz. Thus, the frequency at node Y is lower than f_{out} by Mf_{RM}, where M is the number of pulses removed every $1/f_{RM}$ seconds, and hence $f_{out} = f_{REF} + Mf_{RM}$. Since f_{REF} is now independent of channel spacing, the loop bandwidth can be relatively large and the lock time small. In reality, the pulse remover is followed by a fixed-modulus divider so that the required f_{REF} is below approximately 50 MHz, a range for which low-noise crystal oscillators are available.

The fractional-N architecture suffers from an important drawback: sidebands near the carrier frequency. To understand the issue, note that the two waveforms applied to the mixer in Fig. 22(a) have different shapes even though their "average" frequencies are equal (Fig. 22(b)). While the reference input is strictly periodic with a frequency f_{REF}, the output of the pulse remover has an *instantaneous frequency* of $f_{REF} + 1/T$ followed by a missing pulse (for one pulse removed every T seconds). The mixer/LPF combination produces the averaged phase difference between these waveforms, yielding a ramp that returns to zero at the end of each missing pulse. Modulating the frequency of the VCO, the ramp introduces sidebands that are spaced at $1/T$ with respect to the main carrier. To overcome this issue, another ramp with the same amplitude but opposite polarity can be added to this waveform, but the residual effects due to mismatches usually require external adjustments.

Another fast-locking architecture is based on "direct digital synthesis" (DDS).

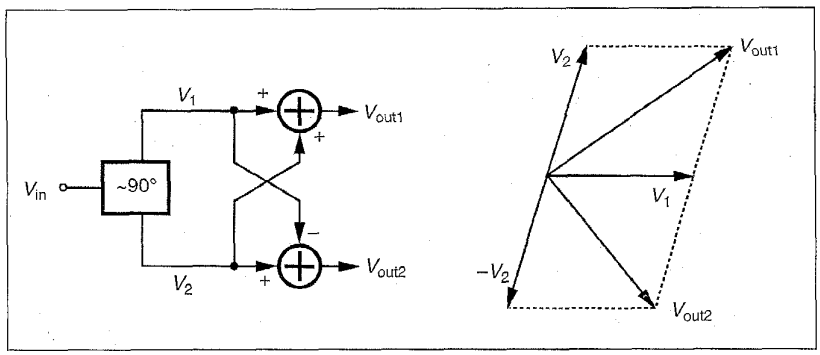


25. RC-CR quadrature network.

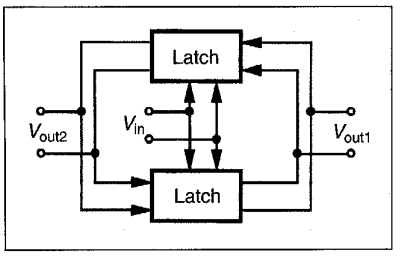
Illustrated in Fig. 23 in a simple form, this technique first generates a *digital* sinusoid by reading consecutive amplitude values from a ROM and subsequently converts the signal to analog by means of a DAC and a low-pass filter. Frequency variation is accomplished by changing the increment in the address word of the ROM (i.e., the increment in phase of the sinusoid). Since DDS employs no analog feedback, it achieves much faster settling than phase-locked topologies. Furthermore, its output phase noise arises primarily from that of the clock, which can be derived from a high-quality fixed-frequency crystal oscillator.

Despite these merits, the DDS architecture requires high-speed operations, thereby consuming significant power in both the digital section and the DAC, and trading precision for speed in the latter. Note that, from Nyquist theorem, for a 900-MHz output, the DDS must operate at no less than 1.8 GHz. Another concern is the substrate noise produced by the complex digital circuits and the resulting corruption of the DAC output.

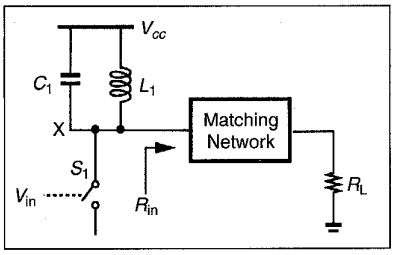
To alleviate some of the above issues, DDS can be combined with phase locking.



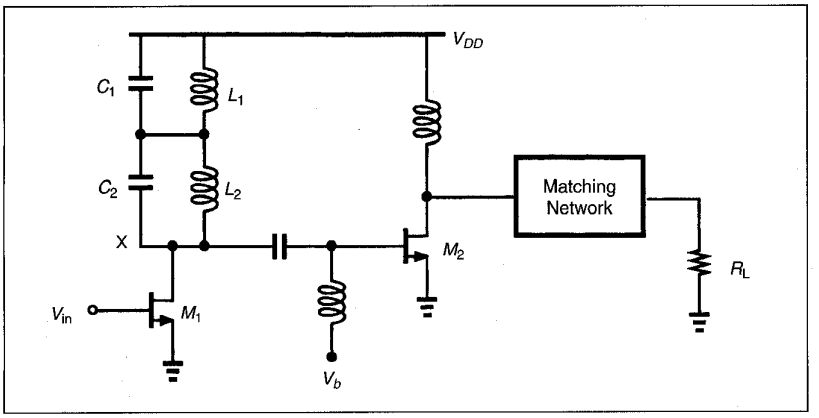
26. Quadrature generation by subtraction and addition.



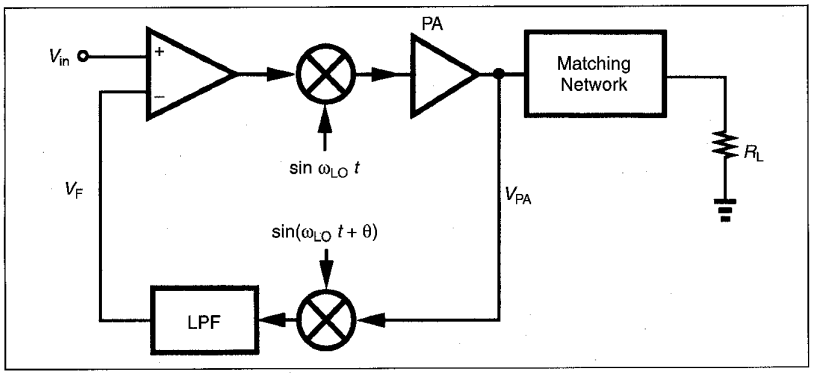
27. Quadrature generation by frequency division.



28. Simple nonlinear power amplifier.



29. Nonlinear power amplifier in [19].



30. PA linearization by frequency-translated feedback.

An example is shown in Fig. 24, where a 900-MHz carrier is generated by phase-locking and the required band is covered by adding a variable DDS-based frequency ranging from 0 to 25 MHz. For "frequency addition," a simple mixer is not adequate because it generates both the difference and the sum frequencies, components that are very closely spaced when $f_{DDS} \rightarrow 0$ and hence difficult to filter selectively. Thus, a "single-sideband mixer" is employed. The operation is based on the identity: $\cos(\omega_1+\omega_2)t = \cos\omega_1 t \cos\omega_2 t - \sin\omega_1 t \sin\omega_2 t$, necessitating quadrature signals at both frequencies. Note that the PLL bandwidth can be quite large because f_{PLL} need not be variable.

The architecture of Fig. 24 faces two difficulties. First, if the DDS DAC output exhibits harmonic distortion (components at $2f_{DDS}$, $3f_{DDS}$, etc.), then the SSB output is corrupted by unwanted sidebands. Second, phase and amplitude mismatches in the SSB mixer yield a component at $f_{PLL}-f_{DDS}$, again an unwanted sideband. For all the above reasons, frequency synthesizers are the topic of active research today.

Quadrature Signal Generation

In our discussion thus far, we have repeatedly seen the need for quadrature signals, e.g., in Figs. 5, 10, 11, 12. Circuit techniques must therefore be sought that can generate (narrowband) signals that are 90° out of phase.

A simple approach is to shift the signal by ±45° using an *RC-CR* network (Fig. 25). It can be easily shown that the phase difference between V_{out1} and V_{out2} is equal to 90° for all frequencies, but the output amplitudes are equal only at $\omega=1/RC$. Thus, if the absolute value of RC varies with temperature and process, so does the frequency at which equal-amplitude quadrature signals exist. While this is a problem with the RF/IF path, the technique can nevertheless be used with LO signals if the amplitudes are made equal by clipping.

Another technique that can be used with "digital" signals (i.e., carriers in downconversion and upconversion) is illustrated in Fig. 26 [7]. Here, the input is first split into two signals, V_1 and V_2, that are approximately 90° out of phase. Subsequently, V_1 and V_2 are added to and subtracted from each other, generating V_{out1} and V_{out2}. It can be seen from the phasor diagram that if V_1 and V_2 have equal amplitudes, V_{out1} and V_{out2} have a phase difference of exactly 90° (but also some amplitude mismatch, which can be corrected by limiting).

A third technique of generating quadrature phases of a digital signal with frequency f_1 is to use a master-slave flip-flop to divide a signal at $2f_1$ by a factor two (Fig. 27). The outputs of the master and slave latches are 90° out of phase if the input has a 50% duty cycle. The difficulty here is that the generation and division of the signal at $2f_1$ may consume substantial power or simply be impossible due to technology limitations.

Power Amplifiers

Power amplifiers are among the most power-hungry building blocks of RF transceivers, and their design is especially difficult because of supply-efficiency-linearity trade-offs [18].

To understand some of these issues, note that to deliver 1 W of sinusoidal power to a 50-Ω antenna, the peak-to-peak voltage swing at the antenna is approximately 20 V. If the peak-to-peak swing provided by the

PA is limited to 5 V — either constrained by the supply voltage or device breakdown voltage — then a "matching network" (e.g., a transformer) is required to interface the PA with the antenna. This network, in essence, transforms the 50-Ω impedance of the antenna to 50 Ω/16 seen by the PA so that a 5-V swing generates 1 W of power.

The high current levels in the PA and the matching network can introduce considerable resistive loss, thereby degrading the efficiency. At lower supply voltages, the required current levels and the loss are higher. As an example, consider the simple nonlinear PA shown in Fig. 28, where L_1 and C_1 resonate at the carrier frequency and switch S_1 pumps energy into the tank on every other transition of V_{in}. Usually implemented with a silicon or GaAs field-effect transistor (FET), S_1 exhibits a finite on-resistance, dissipating power if carrying current. (Note that for the 5-V example above, the peak current drawn by the switch is approximately equal to 1.6 A.) Thus, the ideal phase relationship between V_{in} and V_{out} is such that S_1 is on only when V_{out} is close to zero. Even in this case, however, S_1 turns on for a greater fraction of the period because of the finite transition time of V_{in}. Now suppose the supply voltage is halved and the matching network is modified so that R_{in} decreases by a factor of four. Then, to maintain the same output power, the current provided by S_1 must double, and to keep the power dissipation in S_1 the same, its on-resistance must be lowered by a factor of four. Since the on-resistance of FETs depends on both their gate-source overdrive voltage (i.e., the supply voltage) and their channel width, this can be accomplished by increasing the device width by a factor of eight but at the cost of introducing higher capacitance at node X.

Shown in Fig. 29 is the simplified circuit of a nonlinear metal semiconductor FET (MESFET) PA operating at 835 MHz [19]. The first stage incorporates two tanks in series, one tuned to the first harmonic and the other to the third, so as to provide an approximation of a square wave at node X. This technique reduces the transition times at the gate of M_2, minimizing the power loss in the output transistor. Operating from a 2.5-V supply, the circuit delivers 250 mW with 50% efficiency. Note that the low substrate loss in MESFET technologies makes it possible to integrate the entire PA on one chip.

As mentioned earlier, certain types of phase-modulated signals experience significant band spreading as they are amplified by a nonlinear PA. The π/4-DQPSK scheme used in NADC is an example. In these cases, either linear PAs such as class A configurations or nonlinear topologies employing linearizing techniques are utilized. As the efficiency of class A amplifiers is relatively low (theoretically limited to 50% and in practice rarely exceeding 40%), the second approach has been studied extensively. Shown in Fig. 30 is a simplified example, where a negative feedback path consisting of a downconversion mixer, an LPF, and an error amplifier serves to improve the linearity of the PA. As with ordinary negative feedback systems, this topology attempts to minimize the difference between V_{in} and V_F. Thus, if the nonlinearity of the downconversion mixer and LPF is negligible, V_{PA} is a close replica of V_{in} but at the carrier frequency. It is important to note that the PA required here cannot be arbitrarily nonlinear. For example, in the circuit of Fig. 28, the waveform at node X has very little dependence on the shape of V_{in} if the switch turns on and off abruptly, and, therefore, no amount of feedback can improve the linearity. In other words, the PA can be more nonlinear than class AB but must be less nonlinear than hard switching schemes such as classes C through G [20].

For upconversion circuits that use I/Q demodulation in the last stage (e.g., Fig. 12), the feedback path includes I/Q downconversion mixers, providing two feedback signals. In this form, the technique is called "Cartesian feedback" [21].

A critical issue in the architecture of Fig. 30 is the proper choice of θ. Since each stage in the loop introduces a finite phase shift, θ must be chosen to guarantee stability. However, temperature and process variations require that θ vary accordingly. Furthermore, the PA output power is varied in most systems, and its phase shift usually depends on the output power.

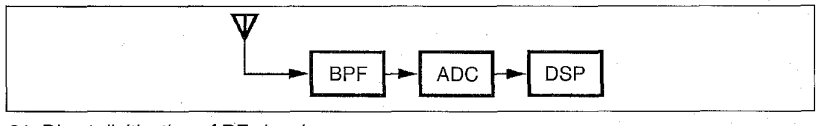
Future Directions

Wireless fever continues to spread — even many companies that used to have only digital products have come aboard the RF ship. At the same time, the wireless infrastructure is evolving to support increasingly more services and types of communication. While the standard cellular system will persist for some time, other methods continue to emerge to alleviate the power and cost issues of cellular phones.

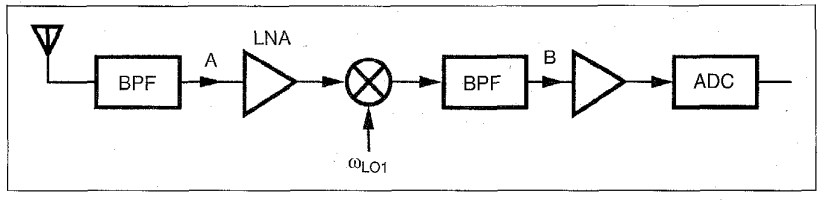
Let us say that the goal of industry is to lower the cost, power dissipation, weight, and size of the portable cellular phone so that it eventually reduces to the size of a credit card. This translates to a minimum number of (off-chip) components; low-power dissipation, especially in the PA and baseband DSP(s); and fewer batteries. In this section, we consider some general approaches to achieving this goal.

Microcell and Picocell Structures

Each cell in the present cellular system has a diameter of roughly 20 km. Thus, the portable phone must produce enough RF power to allow communication with the base station over a 10-km distance. Cost and efficiency issues of PAs, especially as the supply voltage decreases to minimize the number of batteries, make it desirable to reduce the size of each cell. This applies particularly to the cases where most of the communication occurs within a small, but densely populated area. Reduction of cell



31. Direct digitization of RF signal.



32. IF digitization.

size can lead to "microcell" (less than 1 km in radius) and even "picocell" (less than 100 m in radius) structures, with the required RF power going down to only a few tens of milliwatts.

The principal barrier in reducing the cell size is the need for additional base stations. Complexity, cost, and *real estate* issues make it difficult to divide a crowded metropolitan area into smaller cells, each of which must be served by a new base station. Nevertheless, the concept is particularly well-suited to communication within large buildings and is also actively pursued for a wider range of applications.

More Digital, Less Analog

The complexity of the typical heterodyne receiver in Fig. 8, especially the need for several off-chip, expensive, and bulky filters, has created a dream for RF system designers: discard all the analog signal processing, directly digitize the signal received by the antenna, and perform all the operations in the digital domain (Fig. 31). Since the ADC would need a dynamic range of more than 100 dB and an input bandwidth of greater than 1 GHz, this dream is not practical in today's technology. However, the idea of moving the A/D interface closer to the antenna seems promising.

A step toward this direction is "IF sampling," wherein the last mixer in the heterodyne chain is replaced with an A/D converter (Fig. 32). Since typical ADCs perform sampling before quantization, they can operate as downconversion mixers as well. Another possibility is to simply digitize the signal and carry out the quadrature multiplication digitally.

As the A/D interface comes closer to the antenna, two of its parameters must simultaneously improve: the dynamic range and the input bandwidth. Since A/D converters suffer from the same trade-offs as those depicted in Fig. 16, increasing both the resolution and the input bandwidth requires substantial power dissipation penalty. As a consequence, with present technology this interface may remain at the second IF in Fig. 8, but further advances in ADC design will bring it closer to the front end [22, 23].

CMOS Radio

While MOS (metallic oxide semiconductor) devices were considered noisy, slow devices up to about a decade ago, scaling has dramatically improved their performance. The lower cost and faster advance of complementary MOS (CMOS) with respect to silicon bipolar and III-V technologies has motivated great efforts in designing RF CMOS circuits [24, 25, 26, 27]. Even though the reported performance of these circuits may not be adequate for stringent applications such as cellular phones, the potential has created a promising picture.

One-Chip Radio

Another dream of RF designers is to place the entire receive and transmit paths on one chip, reducing the overall cost, power dissipation, and number of external components. The feasibility of this idea will depend on various factors such as the wireless infrastructure (e.g., picocell environment), modulation schemes, RF architectures, and IC technologies. Furthermore, the problem of crosstalk between signals whose amplitudes differ by 100 dB remains to be studied and quantified. In silicon technology the substrate noise generated by large voltage swings can corrupt the signals at various points in the system.

Behzad Razavi is a member of the technical staff at Hewlett-Packard Laboratories in Palo Alto, California. **CD**

References

1. I.A.W. Vance, "Fully Integrated Radio Paging Receiver," *IEE Proc.*, vol. 129, part F, pp. 2-6, 1982.

2. J. Sevenhans et al, "An Integrated Si bipolar RF transceiver for a zero IF 900 MHz GSM digital radio front end of hand portable phones," *Proc. CICC*, pp. 7.7.1-7.7.4, 1991.

3. D. Haspeslagh et al., "BBTRX: A Baseband Transceiver for a Zero IF GSM Portable Station," Proc. CICC, pp. 10.7.1-10.7.4, 1992.

4. C.D. Hull, R.R. Chu, and J.L. Tham, "A direct-conversion receiver for 900 MHz (ISM band) spread-spectrum digital cordless telephone," *ISSCC Dig. Tech. Papers*, pp. 344-345, Feb. 1996.

5. A.A. Abidi, "Direct-conversion radio transceivers for digital communications," *IEEE Journal of Solid-State Circuits*, vol. 30, pp. 1399-1410, Dec. 1995.

6. V. Thomas, J. Fenk, and S. Beyer, "A One-chip 2 GHz Single Superhet Receiver for 2 Mb/s FSK Radio Communications," *ISSCC Dig. Tech. Papers*, pp. 42-43, Feb. 1994.

7. T.D. Stetzler et al., "A 2.7-4.5 V Single chip GSM Transceiver RF Integrated Circuit," *IEEE Journal of Solid-State Circuits*, vol. 30, pp. 1421-1429, Dec. 1995.

8. R. Hartley, "Single-sideband Modulator," US Patent: 1,666,206, April 1928.

9. M.D. McDonald, "A 2.5 GHz BiCMOS Image-reject Front End," *ISSCC Dig. Tech. Papers*, pp. 144-145, Feb. 1993.

10. D.K. Weaver, "A Third Method of Generation and Detection of Single-sideband Signals," *Proc. IRE*, vol. 44, pp. 1703-1705, 1956.

11. C. Marshall et al., "2.7 V GSM Transceiver ICs with On-chip Filtering," *ISSCC Dig. Tech.* Papers, pp. 148-149, Feb. 1995.

12. S. Kumar, "Power Amplifier Linearization Using MMICs," *Microwave J.*, pp. 96-104, April 1992.

13. D. Myer, "Design Linear Feedforward Amps for PCN Systems," *Microwaves & RF*, pp. 121-133, Sept. 1994.

14. J.C. Pedro and J. Perez, "An MMIC Linearized Amplifier Using Active Feedback," *Proc. IEEE Microwave and Millimeter-Wave Monolithic Circuits Symp.*, pp. 113-116, 1993.

15. R.G. Meyer and W.D. Mack, "A 1-GHz BiCMOS RF Front-end Integrated Circuit," *IEEE Journal of Solid-State Circuits*, vol. 29, pp. 350-355, March 1994.

16. J.A. Crawford, *Frequency Synthesizer Design Handbook*, Artech House, 1994.

17. W.F. Egan, *Frequency Synthesis by Phase Lock*, New York, Wiley & Sons, 1981.

18. S.L. Wong et al., "A 1 W 830 MHz Monolithic BiCMOS Power Amplifier," *ISSCC Dig. Tech. Papers*, pp. 52-53, Feb. 1996.

19. T. Sowlati et al., "Low Voltage, High Efficiency GaAs Cass E Power Amplifiers for Wireless Transmitters," *IEEE Journal of Solid-State Circuits*, vol. 30, pp. 1074-1080, Oct. 1995.

20. H.C. Kraus, C.W. Bostian, F.H. Raab, *Solid State Radio Engineering*, New York: Wiley & Sons, 1980.

21. M. Johansson and T. Mattsson, "Transmitter Linearization Using Cartesian Feedback for Linear TDMA Modulation," Proc. IEEE Veh. Tech. Conf., pp. 542-546, 1990.

22. A. Hairapetian, "An 81 MHz IF receiver in CMOS," ISSCC Dig. Tech. Papers, pp. 56-57, Feb. 1996.

23. J.E. Eklund and R. Arvidsson, "A 10 b 120 MS/s Multiple Sampling Single Conversion CMOS A/D Converter for I/Q Demodulation," ISSCC Dig. Tech. Papers, pp. 294-295, Feb. 1996.

24. A. Rofougaran et al., "A 1-GHz CMOS RF front-end IC with Wide Dynamic Range," Proc. ESSCIRC, pp. 250-253, 1995.

25. A. Karanicolas, "A 2.7 V 900 MHz CMOS LNA and Mixer," ISSCC Dig. Tech. Papers, pp. 50-51, Feb. 1996.

26. J. Crols and M.S.J. Steyaert, "A Single-chip 900 MHz CMOS Receiver Front end with a High Performance Low-IF topologies," IEEE Journal of Solid-State Circuits, vol. 31, pp. 1483-1492, Dec. 1995.

27. S. Sheng et al, "A low-power CMOS chipset for Spread Spectrum Communications," ISSCC Dig. Tech. Papers, pp. 346-347, Feb. 1996.