(a) Phase of the FFT of the modulated carrier signal, $s(t)$

(b) Weighted phase of the FFT of the in-phase (I) component of the modulated carrier signal, $s_i(t)$

(c) Weighted phase of the FFT of the quadrature-phase (Q) component of the modulated carrier signal, $s_q(t)$

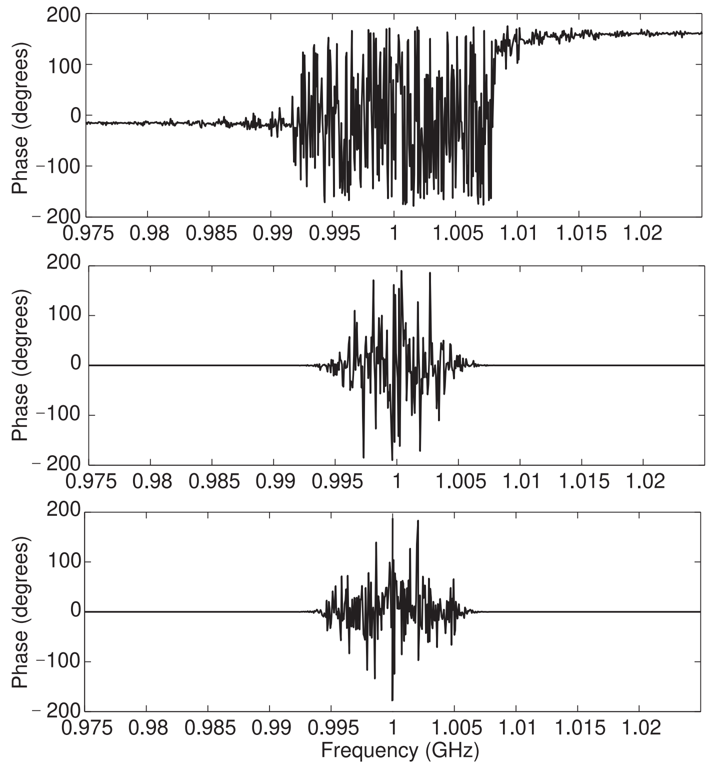


Figure 3-38: Phase of the FFT of the digitally modulated RF signal and its components. The phase of the $s_i(t)$ and $s_q(t)$ has been weighted by their spectrum amplitudes as otherwise meaningless numerical noise dominates the plots outside the passband. (Using 1024 symbols, 102.4 μs)

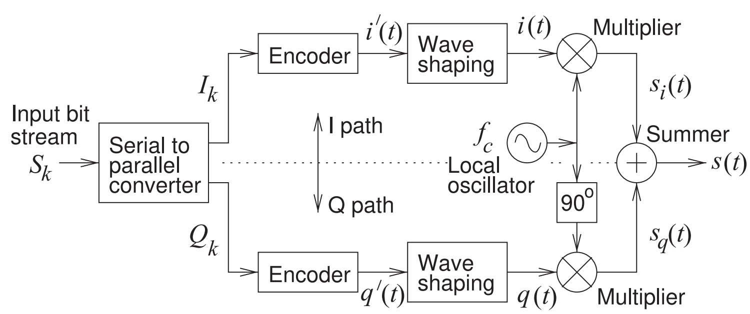


Figure 3-39: High-order digital quadrature modulator with independent I and Q paths yielding double sideband, suppressed carrier (DSB-SC) modulation.

3.10.4 QAM Digital Modulation

The QPSK digital modulator considered in the previous section had only four states. This is a consequence of the digital baseband bitstreams, I_k and Q_k, being used one bit at a time. If two or more bits are used at a time to address a constellation point, more data can be sent in the same RF bandwidth. Modulation that does this is called higher-order digital modulation. A higher-order quadrature modulator is shown in Figure 3-39 where now groups of two or more I_k bits and Q_k bits can be encoded as an analog signal. That is, if there are four possible values of $i(t)$ at the clock ticks, then pairs of I_k bits are encoded as one of four analog values. Such as the I_k

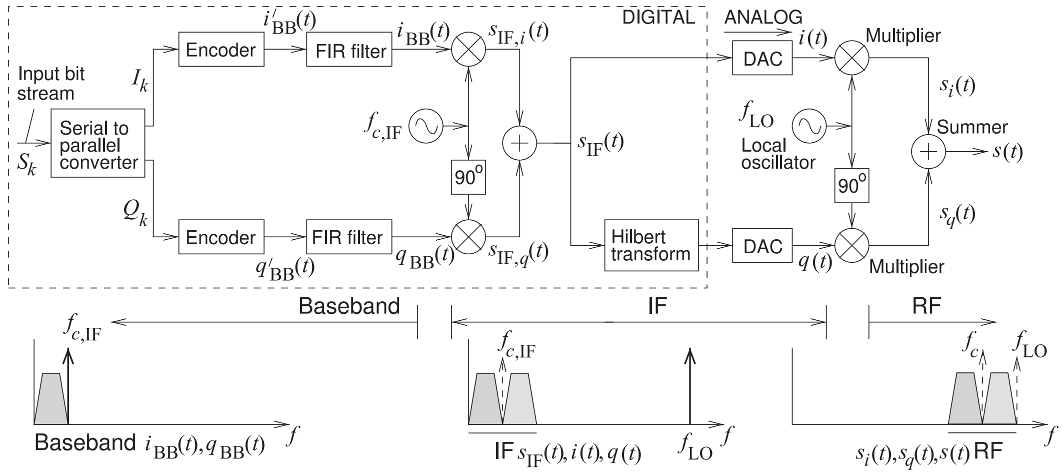


Figure 3-40: Complete SDR transmitter with bistream S_k input to an IF quadrature modulator producing a DSB-SC signal, s_{IF}, with IF carrier of frequency $f_{c,\text{IF}}$. Each finite-impulse response (FIR) filter implements raised cosine filtering. Then an analog SSB-SC quadrature modulator up-converts the IF as a DSB-SC RF signal centered at the RF carrier frequency $f_c = f_{\text{LO}} - f_{c,\text{IF}}$. The Hilbert transform implements a $90°$ phase shift.

bit pair 00 resulting in $i(t)$ before filtering, i.e. $i'(t)$, being encoded as -1 V, 01 being encoded as $-\frac{1}{3}$ V, 10 being encoded as $+\frac{1}{3}$ V, and 11 being encoded as $+1$ V. Q_k would be encoded the same way. This results in 16 possible states and this is called 16-QAM modulation. Operation of this quadrature modulator will be analyzed in the next section in the context of a complete SDR transmitter.

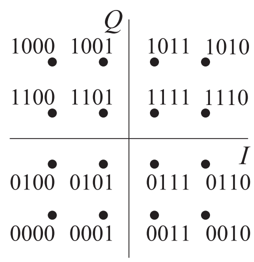
3.10.5 SDR Transmitter Using QAM Digital Modulation

This section describes a 16-QAM modulator as used in modern communications systems (4G and 5G). This reflects the capabilities of baseband DSPs which can perform many of the modulation operations. The architecture of a QAM quadrature-modulator is depicted in block diagram form in Figure 3-40. This modulator consists of two quadrature modulators, a DSP-based digital IF quadrature modulator on the left, and an analog RF quadrature modulator on the right. The digital IF quadrature modulator on the left is a DSB-SC quadrature modulator taking an input bitstream and partitioning it by converting from a serial input bit stream, S_k, into two parallel but independent bitstreams, I_k and Q_k. The output of the IF quadrature modulator, $s_{\text{IF}}(t)$, is a DSB-SC IF signal centered at the IF carrier frequency $f_{c,\text{IF}}$. This is then modulated by the analog RF quadrature modulator to create a SSB-SC RF signal which has the IF signal, $s_{\text{IF}}(t)$, as its modulating signal. The RF signal however retains the double sidebands of the IF signal so that $s(t)$ is a DSB-SC signal with carrier frequency $f_c = f_{\text{LO}} - f_{c,\text{IF}}$.

The 16-QAM system has the parameters given in Table 3-1. Since pairs of I_k bits and pairs of Q_k bits are used to encode the baseband analog signals i'_{BB} and q'_{BB} there are 16 possible states and 16-QAM modulation is the result. The constellation diagram of 16-QAM modulation is shown in

Input bitstream S_k	40 Mbit/s		
I_k bitstream	20 Mbit/s		
Q_k bitstream	20 Mbit/s		
Nominal bandwidth $i'_{\rm BB}$	10 MHz (discrete)	Bandwidth $i(t)$	20 MHz (analog)
Nominal bandwidth $q'_{\rm BB}$	10 MHz (discrete)	Bandwidth $q(t)$	20 MHz (analog)
Intermediate carrier, $f_{c,\rm IF}$	10 MHz (discrete)	Local oscillator, $f_{\rm LO}$	1 GHz
Nominal bandwidth $s_{{\rm IF},i}$	20 MHz (discrete)	Bandwidth $s_i(t)$	20 MHz (analog)
Nominal bandwidth $s_{{\rm IF},q}$	20 MHz (discrete)	Bandwidth $s_q(t)$	20 MHz (analog)
Nominal bandwidth $s_{\rm IF}$	20 MHz (discrete)	Bandwidth $s(t)$	20 MHz (analog)
IF carrier, $f_{c,\rm RF}$	10 MHz	RF carrier, $f_{c,\rm RF}$	990 MHz

Table 3-1: Parameters of the SDR transmitter considered in Section 3.10.4.



(a)

I_k or Q_k	$i(t)$ or $q(t)$
00	-1
01	$-\frac{1}{3}$
11	$+\frac{1}{3}$
10	$+1$

(b)

Figure 3-41: 16-QAM modulation: (a) constellation diagram; and (b) table translating binary value of I (or Q) to an absolute value $i(t)$ (or $q(t)$).

Figure 3-41(a) with the bit-wise addresses of the constellation points. There are 16 constellation points each of which indicates a symbol which provides four bits of information. In the case of QAM modulation the constellation diagram corresponds to a phasor diagram with each constellation point indicating the amplitude and phase of the RF signal at a point in time. The difference between the constellation diagram and a phasor diagram is that the constellation diagram does not change when the average power of the RF signal changes. Thus the constellation diagram of QAM corresponds to a phasor diagram that is continuously being re-normalized to the average RF power.

The table in Figure 3-41(b) shows the encoding that enables pairs of I_k bits and pairs of Q_k bits to address a symbol in the constellation diagram. That is, each of these bitstreams is encoded so that two bits are encoded as one analog value. Figure 3-42(a) shows the bits in the input bitstream, S_k, which are in groups of four bits with the first and third bits of the group becoming I bits and the second and fourth bits in the group becoming Q bits. The I_k and Q_k bitstreams are shown in Figures 3-42(b) and 3-42(c) respectively. A pair of I_k (Q_k) bits yields an encoded $i'_{\rm BB}$ ($q'_{\rm BB}$) value every 0.1 μs. Here $I_k = I_1I_2I_3I_4I_5I_6 = 011100$ so $i'(0.1\ \mu{\rm s}) = -\frac{1}{3}$, $i'(0.2\ \mu{\rm s}) = +\frac{1}{3}$, and $i'(0.3\ \mu{\rm s}) = -1$. Linearly interpolating these yields the $i'_{\rm BB}$ and $q'_{\rm BB}$ waveforms shown in Figure 3-42(d). In DSP of course these are discrete values but have been plotted as continuous waveforms which is more easily visualized.

FIR filtering of $i'_{\rm BB}$ and $q'_{\rm BB}$ yields the band-limited waveforms $i(t)$ and $q(t)$ shown in Figure 3-42(e). Filtering introduces a delay and here the delay is 0.2 μs. Thus the first encoded value before filtering is established at 0.1 μs, see

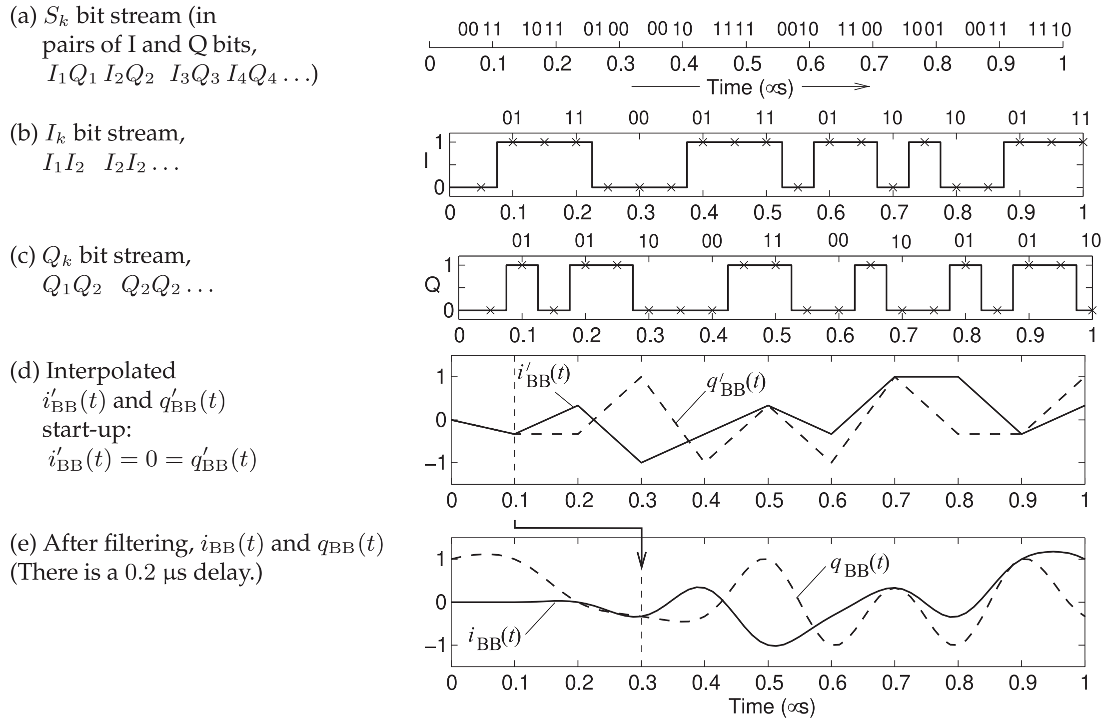


Figure 3-42: Baseband signals for 16-QAM modulation with markers showing the baseband bit impulse times. The bit pairs shown above each bit waveform 0.1 μs intervals.

the vertical dashed line in Figure 3-42(d), and the first symbol (represented by $i_{\rm BB}$ and $q_{\rm BB}$) is at 0.3 μs, see the vertical dashed line in Figure 3-42(e). The baseband values are actually discrete samples. With an oversampling factor of 8, the $i'_{\rm BB}$ samples, the interpolated encoded values, are as shown in Figure 3-43(b) and the interpolated $q'_{\rm BB}$ samples in Figure 3-43(d). In this SDR transmitter baseband filtering is a raised cosine filter implemented digitally as a finite impulse response (FIR) filter.

The interpolated encoded samples $i'_{\rm BB}$ and $q'_{\rm BB}$ are filtered digitally using a finite impulse response (FIR) filter yielding the filtered samples $i_{\rm BB}$ and $q_{\rm BB}$ shown in Figures 3-43(c and e) respectively. The raised-cosine FIR filter has a span of 4 symbols (i.e. 0.4 μs) and since the oversampling factor is 8 the FIR filter's response is 33 samples long, see Figure 3-43(a). (With the 0th and 33rd bins being zero, it could be said to be 31 or 32 samples long.) (Specification of the raised cosine filter filter is completed with a roll-off factor of 0.7 which describes how quickly the impulse response rolls-off around its peak response.) The peak response of the filter is at the 16th sample, after 0.2 μs, and the impulse responses at 0 (-0.2 μs), 8 (-0.1 μs), 24 (0.1 μs), and 32 (0.2 μs) samples are zero. Thus the FIR filter introduces a 0.2 μs delay. There are other types of FIR filters including two-dimensional FIR filters that operate on the I and Q channels together and can provide even better management of the bandwidth of the modulated IF signal.

The next stage in the transmitter is multiplying $i_{\rm BB}$ by the IF LO having

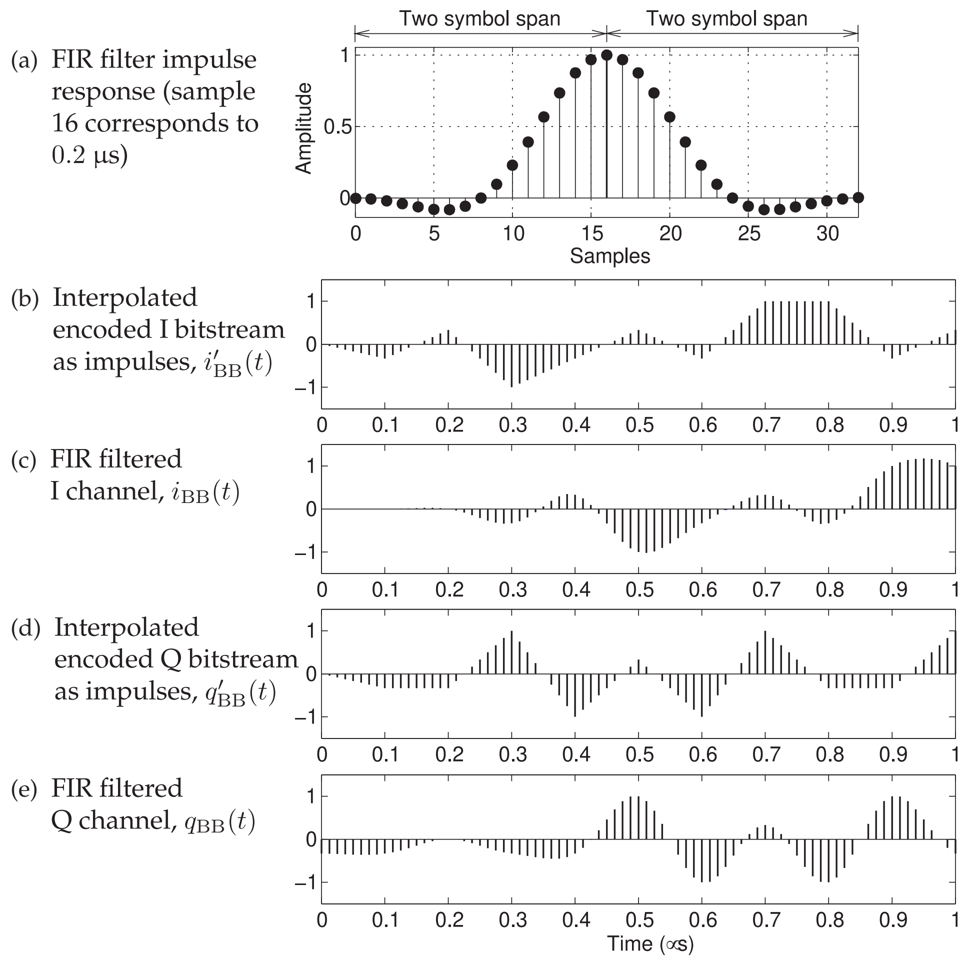


Figure 3-43: Sampled baseband signals for 16-QAM modulation corresponding to the analog baseband signals in Figure 3-42.

frequency $f_{c,\rm IF}$ and multiplying $q_{\rm BB}$ by a 90° phase-shifted IF LO. This results in the modulated IF signal, $s_{\rm IF}$, shown in Figure 3-44(a).

The procedure of producing a SSB-SC modulated RF signal from the IF signal is to drive the multipliers in the RF quadrature modulator of the second mixing stage with an in-phase IF signal and its quadrature, i.e. 90° phase-shifted version. In this case every frequency component of $s_{\rm IF}$ must be phase shifted by 90°. The mathematical procedure that does this is the Hilbert transform operating on a finite number of samples of $s_{\rm IF}$ and here 8192 symbols with an oversampling factor of 8 were considered so that the Hilbert transform operates on 65,536 samples. In DSP an FFT can be used so that an FFT of $s_{\rm IF}$ yields the amplitude and phase of discrete frequency components of $s_{\rm IF}$ and then the individual frequency samples are phase-shifted by 90°. Then an inverse FFT yields the required Hilbert transform in the time domain. A DAC then converts the digital samples into the analog signals $i(t)$ and $q(t)$. Their waveforms are shown in Figures 3-44(a and b).

(a) In-phase IF waveform, $i(t)$

(b) Quadrature-phase IF waveform, $q(t)$

(c) RF waveform, $s(t)$

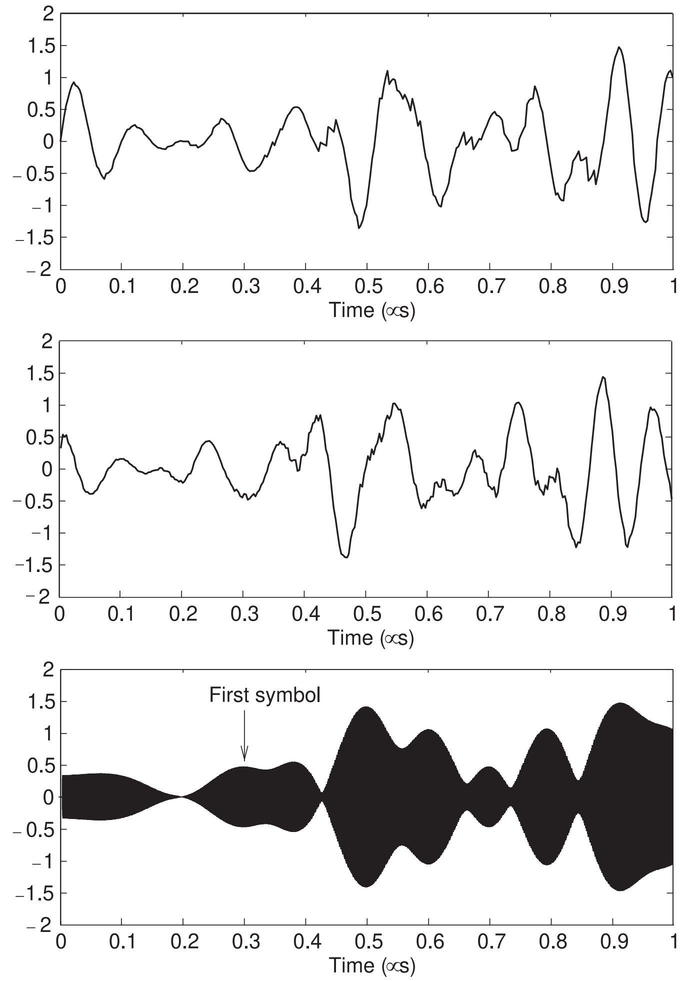


Figure 3-44: Waveforms in the RF quadrature modulator where the in-phase IF signal $i(t)$ is just the output, $s_{\mathrm{IF}}(t)$ of the IF quadrature multiplier. The first 0.2 μs can be ignored as the FIR filter output is settling. (The IF sampling interval is 3.125 ns. The Hilbert transform operates on 2^{16} =65,537 IF samples or 204.8 μs of data corresponding to 2048 symbols or 8192 bits. The RF time step is 31.25 ps. The RF signal spectrum was calculated using a 2^{22} point FFT.

Then these signals are quadrature modulated using an RF LO of 1 GHz yielding the RF waveform shown in Figure 3-44(c). The SSB-SC RF analog modulator has translated the DSB-SC IF signal to RF to produce a DSB-SC RF signal. Plots of $i(t), q(t)$ and $s(t)$ over a longer time are shown in Figure 3-45.

The spectra of the IF and RF signals are shown in Figure 3-46. It is seen that the bandwidth of the IF modulated signal is just under 20 MHz and the bandwidth of the RF signal is the same. The carrier of the IF signal is 10 MHz, the center of the DSB-SC modulated IF. The RF signal is centered around 990 MHz and this is the carrier frequency of the radio signal. That is the 10 MHz carrier frequency, $f_{c,\mathrm{IF}}$ has been up-converted to the lower

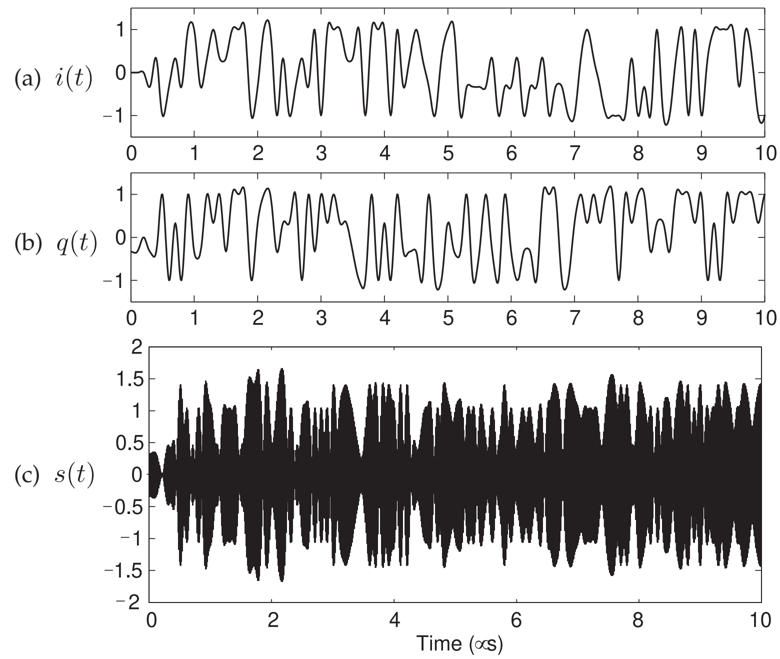


Figure 3-45: In-phase and quadrature-phase waveforms and RF waveforms in the RF quadrature modulator.

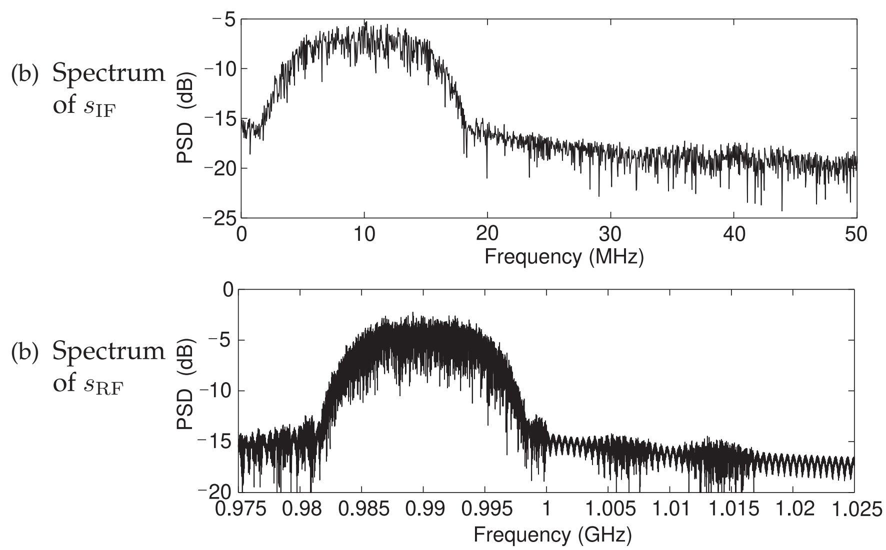
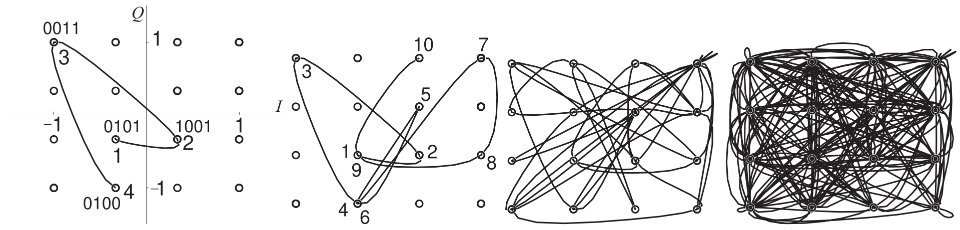


Figure 3-46: Spectrum of the IF and RF modulated carriers.

sideband of the 1 GHz LO.

The constellation diagram together with trajectories of the modulated carrier signal are shown in Figure 3-47 for various numbers of symbols. This also corresponds to the phasor diagram of the QAM modulated signal except that the constellation diagram is effectively a phasor diagram that is re-normalized to the average power level so that the constellation diagram does not change with the average RF power changes. A phasor diagram would change of course as the average RF power changed. The trajectories in Figure 3-47 go through a constellation point every 0.1 µs. So provided that



(a) 4 symbols (0.4 µs) (b) 10 symbols (1 µs)(c) 40 symbols (4 µs)(d) 200 symbols (20 µs)

Figure 3-47: Constellation diagram with trajectories of the phasor of f_c for various symbol lengths with the first symbol at 0.3 µs. Each symbol has 4 bits. In (a) and (b) the first symbol is labeled '1', etc.

(a) IF, $s_{\mathrm{IF}}(t)$

(b) RF, $s(t)$

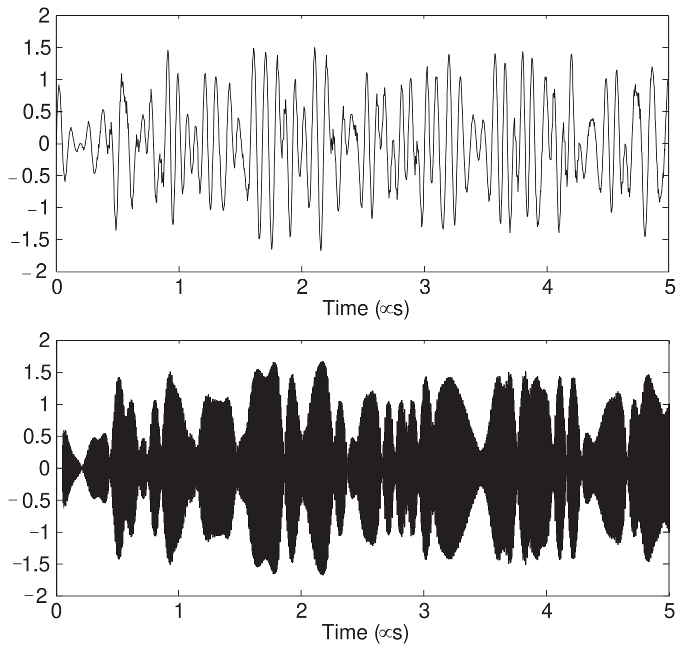
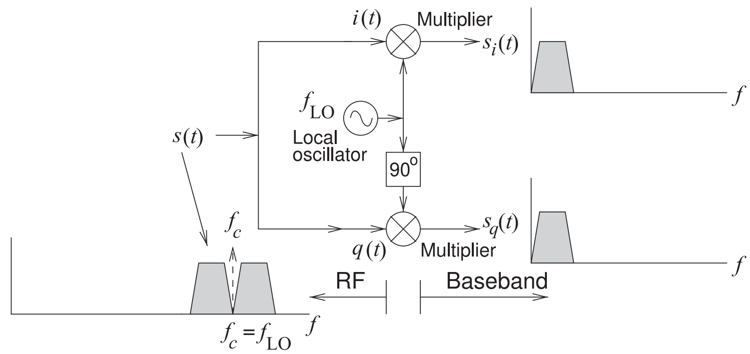


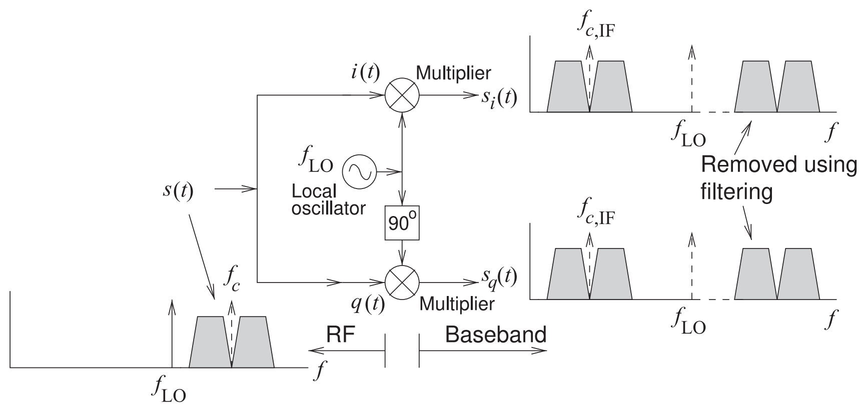
Figure 3-48: Comparison of the IF and RF modulated signals.

the clock in a receiver is correctly aligned and samples the RF signal every 0.1 µs, the transmitted symbols will be precisely recovered. This of course is in the absence of noise and circuit non-idealities.

The transmitter described here has two modulation stages. The IF quadrature modulator modulates the baseband signal as a DSB-SC IF signal with a bandwidth of around 20 MHz on a 10 MHz carrier. The second stage, the RF quadrature modulator, takes the IF waveform and SSB-SC modulates it on a 990 MHz carrier and maintains the IF bandwidth. The relationship between the IF and RF waveforms is easier to see by viewing the waveforms over a 5 µs span corresponding to 50 cycles of the 10 MHz IF carrier and approximately 5000 cycles of the RF LO (and 5051 cycles of the RF carrier), see Figure 3-48.



(a) Quadrature demodulator with $f_{\rm LO} = f_c$



(b) Quadrature demodulator with $f_{\rm LO} < f_c$, $f_{c,\rm IF} = f_c - f_{\rm LO}$.

Figure 3-49: Quadrature demodulators used to down-convert an I/Q modulated RF signal.

Summary

The SDR transmitter involves many technologies. This interplay was best demonstrated using the specific parameters of the SDR transmitter considered in this section. The main limitation is on the bandwidth of the IF modulated signal as this relates directly to the performance of the DSP unit which in turn impacts battery drain. The advantages of implementing a first stage using DSP are tremendous and nearly any modulation scheme can be supported simply by changing the parameters in the DSP unit. The analog RF unit can support very wide bandwidths as fixed frequency filters are not needed. The most critical performance parameter is balancing the I and Q paths through the entire transmitter chain. With most of the modulation occurring in the DSP unit, balancing of the digital portions of the I and Q paths is achieved precisely. The challenge then falls to balancing the I and Q paths in the analog RF modulator.

3.11 SDR Quadrature Demodulator

The SDR circuit used to demodulate an I/Q modulated RF signal is very similar to the standard quadrature demodulator. Complete demodulation and separation of the I channel and the Q channel is obtained when the LO of the demodulator coincides with the RF carrier frequency, see Figure 3-49(a). If the LO is not equal to the carrier frequency the RF modulated signal will be down-converted to an intermediate frequency signal which corresponds to a band around the difference frequency of the carrier and the LO, and a band around the sum frequency of the carrier and the LO, see Figure 3-49(a). Then lowpass filtering would be used to eliminate the higher frequency band. The signal centered at the intermediate frequency can then be I/Q demodulated by another quadrature demodulator.

3.12 SDR Receiver

An SDR receiver implements the demodulation shown in Figure 3-49(a) in two stages with analog RF quadrature demodulation in the first stage similar to that shown in Figure 3-49(b). There is a lot of detail in this section but this is required to understand the implementation of an SDR receiver at the level of design.

The block diagram of a particular implementation of an SDR receiver is shown in Figure 3-50 which has a first stage in the analog domain implemented in an RF modem chip, and a second digital stage that is implements DSP in the baseband chip. While these are separate chips at the time of this writing they will probably be combined in a single chip someday.

The analog portion of the SDR receiver is shown in Figure 3-50. This first stage separates the I and Q channels and outputs baseband frequency signals that are sampled and input to a second stage of quadrature demodulation to produce the final baseband signals. This second stage is implemented digitally. The signals at the output of the first analog stage are also called intermediate frequencies. Digital demodulation, on the right in Figure 3-50, is performed in a DSP unit commonly referred to as a baseband modem chip. A key attribute is that the LO frequency of the first demodulation stage, f_{LO} can be set at relatively few discrete values while the LO in the second stage, $f_{\mathrm{LO,IF}}$ is set finely for full carrier recovery. Together $f_{\mathrm{LO}} + f_{\mathrm{LO,IF}} = f_c$, the frequency of the carrier of the received signal $s(t)$. The DSP unit can implement decoding such as CDMA- or OFDM-decoding.

The rest of this section traces the signal flow through the SDR receiver. The signal presented to the demodulator is $G_L s(t)$ where $s(t)$ is the transmitted DSB-SC signal in Equation (3.13). G_L, which will be very small, is the link gain accounting for loss in transmission and receiver gain and, in the absence

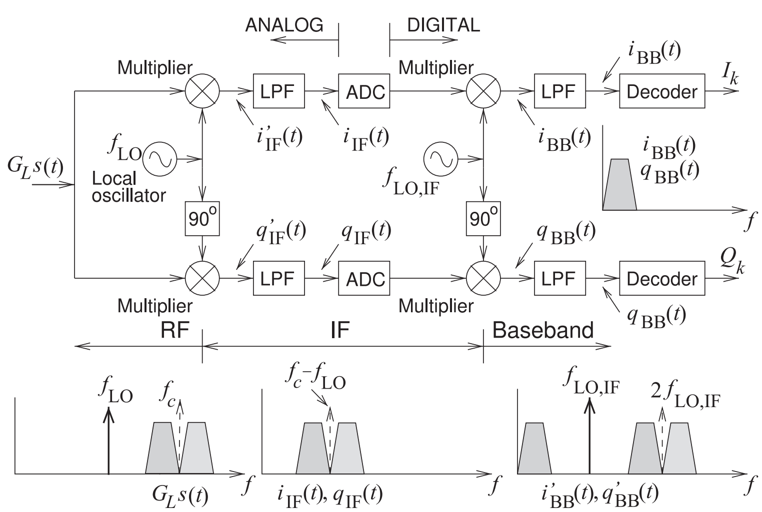


Figure 3-50: SDR receiver with RF input signal $G_L s(t)$ with the first LO frequency, f_{LO}, less than the carrier signal f_c. Note that $2f_{\mathrm{LO,IF}} = f_c - f_{\mathrm{LO}}$.

of interference, $s(t)$ is the transmitted radio signal.

3.12.1 Demodulation of the I component

If the transmitted signal is the DSB-SC signal $s(t)$ with carrier f_c in Equationver (3.13) (replacing $\omega_{\rm LO}$ by f_c) then the signal presented to the demodulator in the receiver is $G_L s(t)$ where G_L, which will be very small, is the link gain accounting for loss in transmission and receiver gain. The spectrum of the received DSB-SC signal, $G_L s(t)$, is shown on the bottom left of Figure 3-50 with the carrier frequency f_c identified by the dash arrow. The LO frequency of the first demodulator at $f_{\rm LO}$ is also shown. The in-phase component of the demodulated signal after the first quadrature demodulator is, see Figure 3-50,

$$\begin{aligned}
i'_{\rm IF}(t) &= G_L s(t)\sin(\omega_{\rm LO}t)\\
&= \tfrac{1}{2}\left[A_i(f_i)\cos(\omega_c-\omega_i)t + A_q(f_q)\sin(\omega_c-\omega_q)t\right.\\
&\quad \left.-\ A_i(f_i)\cos(\omega_c+\omega_i)t - A_q(f_q)\sin(\omega_c+\omega_q)t\right]\sin(\omega_{\rm LO}t)\\
&= \tfrac{1}{2}\left[A_i(f_i)\sin(\omega_c+\omega_{\rm LO}-\omega_i)t - A_i(f_i)\sin(\omega_c-\omega_{\rm LO}-\omega_i)t\right.\\
&\quad + A_q(f_q)\cos(\omega_c-\omega_{\rm LO}-\omega_q)t - A_q(f_q)\cos(\omega_c+\omega_{\rm LO}-\omega_q)t\\
&\quad - A_i(f_i)\sin(\omega_c+\omega_{\rm LO}+\omega_i)t + A_i(f_i)\sin(\omega_c-\omega_{\rm LO}+\omega_i)t\\
&\quad \left.-A_q(f_q)\cos(\omega_c-\omega_{\rm LO}+\omega_q)t + A_q(f_q)\cos(\omega_c+\omega_{\rm LO}+\omega_q)t\right]. \qquad (3.15)
\end{aligned}$$

Equation (3.15) includes intermediate frequency terms centered at $\omega_c - \omega_{\rm LO}$ and high frequency terms centered at $\omega_c + \omega_{\rm LO}$. These high frequency terms can be eliminated through lowpass filtering leaving the lowpass filtered received signal

$$\begin{aligned}
i_{\rm IF}(t) &= \tfrac{1}{2}\left[-A_i(f_i)\sin(\omega_c-\omega_{\rm LO}-\omega_i)t + A_q(f_q)\cos(\omega_c-\omega_{\rm LO}-\omega_q)t\right.\\
&\quad \left.+\ A_i(f_i)\sin(\omega_c-\omega_{\rm LO}+\omega_i)t - A_q(f_q)\cos(\omega_c-\omega_{\rm LO}+\omega_q)t\right]\\
&= \tfrac{1}{2}A_i(f_i)\left[-\sin(\omega_c-\omega_{\rm LO}-\omega_i)t + \sin(\omega_c-\omega_{\rm LO}+\omega_i)t\right]\\
&\quad + \tfrac{1}{2}A_q(f_q)\left[\cos(\omega_c-\omega_{\rm LO}-\omega_q)t - \cos(\omega_c-\omega_{\rm LO}+\omega_q)t\right]\\
&= \tfrac{1}{2}A_i(f_i)\left[\sin(\omega_i-\omega_c+\omega_{\rm LO})t + \sin(\omega_i+\omega_c-\omega_{\rm LO})t\right]\\
&\quad + \tfrac{1}{2}A_q(f_q)\left[\cos(\omega_q-\omega_c+\omega_{\rm LO})t - \cos(\omega_q+\omega_c-\omega_{\rm LO})t\right] \qquad (3.16)
\end{aligned}$$

and this is a DSB-SC signal with a carrier frequency $f_c - f_{\rm LO}$. The location of $i_{\rm IF}$ is shown in Figure 3-50 and its spectrum is the middle spectrum at the bottom of the figure. Of course Equation (3.16) is a discrete signal and the spectrum shows a finite bandwidth signal describing $i_{\rm IF}(t)$ over the range of f_i components. (That is, the linear sum of $i_{\rm IF}(t)$ in Equation (3.16) for all f_i components.)

Single Stage Demodulation with $f_{\rm LO} = f_c$

The IF signal, $i_{\rm IF}(t)$ is then mixed with the IF LO with radian frequency $\omega_{\rm LO,IF}t = \omega_c t - \omega_{\rm LO}t + \phi_i$ producing the recovered I channel baseband signal

$$\begin{aligned}
i_{\rm BB}(t) &= \tfrac{1}{2}A_i(f_i)\left[\sin(\omega_i t+\phi_i) + \sin(\omega_i t-\phi_i)\right\}\\
&\quad + \tfrac{1}{2}A_q(f_q)\left[\cos(\omega_q t+\phi_i) - \cos(\omega_q t-\phi_i)\right]. \qquad (3.17)
\end{aligned}$$

Then if $\phi_i = 0$, which is when the phase of the signal at $f_{\mathrm{LO,IF}}$ has been recovered correctly, the demodulated baseband signal for the I channel is

$$i_{\mathrm{BB}}(t) = A_i(f_i)\sin(\omega_i t). \tag{3.18}$$

Two stage demodulation $f_{\mathrm{LO}} < f_c$

In an SDR demodulator there is two-stage demodulation with an analog stage followed by a digital stage and this is what is shown in Figure 3-50. The analog stage, producing $i_{\mathrm{IF}}(t)$ in Equation (3.16), is in a chip commonly referred to as the RF modem chip. The second stage is usually in a separate chip called the baseband processing chip. One implementation of an SDR receiver uses an analog LO adjusted in discrete frequency steps and, with $f_{\mathrm{LO}} < f_c$ the resulting down-converted signal $i_{\mathrm{IF}}(t)$ is a DSB-SC signal with a carrier at $f_c - f_{\mathrm{LO}}$, see the middle spectrum at the bottom of Figure 3-50. A second stage of quadrature demodulation operates on i_{IF}. The i_{IF} signal is sometimes called the baseband signal if the focus is on the RF modem chip.

The intermediate signal $i_{\mathrm{IF}}(t)$ in Equation (3.17) is mixed with an IF carrier signal at frequency $f_{\mathrm{LO,IF}}$ as follows

$$\begin{aligned} i'_{\mathrm{BB}}(t) = \big\{ &\tfrac{1}{2}A_i(f_i)\left[\sin(\omega_i - \omega_c + \omega_{\mathrm{LO}})t + \sin(\omega_i + \omega_c - \omega_{\mathrm{LO}})t\right] \\ &+ \tfrac{1}{2}A_q(f_q)\left[\cos(\omega_q - \omega_c + \omega_{\mathrm{LO}})t - \cos(\omega_q + \omega_c - \omega_{\mathrm{LO}})t\right]\big\} \\ &\times \cos(\omega_{\mathrm{LO,IF}})t. \end{aligned} \tag{3.19}$$

The spectrum of this signal is shown in the bottom right of Figure 3-50 as the spectrum immediately adjacent to DC.

The intermediate LO frequency $f_{\mathrm{LO,IF}}$ can be set with fine adjustment so that $f_{\mathrm{LO,IF}} = f_c - f_{\mathrm{LO}}$ which produces a baseband signal $i'_{\mathrm{BB}}(t)$. Carrier recovery then becomes the process of determining the phase-synchronized $f_{\mathrm{LO,IF}}$ which is just f_c frequency offset by the numerical value of f_{LO}. (Note that $\omega_c - \omega_{\mathrm{LO}} - \omega_{\mathrm{LO,IF}} = 0$, $-\omega_c + \omega_{\mathrm{LO}} + \omega_{\mathrm{LO,IF}} = 0$, $\omega_c - \omega_{\mathrm{LO}} + \omega_{\mathrm{LO,IF}} = 2\omega_{\mathrm{LO,IF}}$, and $-\omega_c + \omega_{\mathrm{LO}} - \omega_{\mathrm{LO,IF}} = -2\omega_{c,\mathrm{IF}}$. The I channel baseband signal is then

$$\begin{aligned} i'_{\mathrm{BB}}(t) &= i'_{\mathrm{IF}}(t)\cos(\omega_{\mathrm{LO,IF}}t) \\ &= G_L\big\{\tfrac{1}{2}A_i(f_i)\left[\sin(\omega_i - \omega_c + \omega_{\mathrm{LO}})t + \sin(\omega_i + \omega_c - \omega_{\mathrm{LO}})t\right] \\ &\quad + \tfrac{1}{2}A_q(f_q)\left[\cos(\omega_q - \omega_c + \omega_{\mathrm{LO}})t - \cos(\omega_q + \omega_c - \omega_{\mathrm{LO}})t\right]\big\} \\ &\quad \times \cos(\omega_{\mathrm{LO,IF}}t) \\ &= \tfrac{1}{2}G_L A_i(f_i)\left[\sin(\omega_i + 0)t + \sin(\omega_i - 2\omega_{\mathrm{LO,IF}})t\right. \\ &\quad \left. + \sin(\omega_i + 2\omega_{\mathrm{LO,IF}})t + \sin(\omega_i + 0)t\right] \\ &\quad + \tfrac{1}{2}G_L A_q(f_q)\left[+\cos(\omega_q - 2\omega_{\mathrm{LO,IF}})t + \cos(\omega_q + 0)t\right. \\ &\quad \left. - \cos(\omega_q + 0)t - \cos(\omega_q + 2\omega_{\mathrm{LO,IF}})t\right]. \end{aligned} \tag{3.20}$$

After lowpass filtering (in the DSP) to eliminate components centered at $\pm 2\omega_{\mathrm{LO,IF}}$

$$i_{\mathrm{BB}}(t) = G_L A_i(f_i)\sin(\omega_i t) \tag{3.21}$$

This is the final desired signal and its spectrum is shown as the inset on the middle right of Figure 3-50 (between the two decoders). This analysis has been undertaken with a single tone for the I channel but by extension this holds for the actual I channel signal. Sampling $i_{\rm BB}(t)$ at the clock ticks yields the sequence of symbols that were transmitted which after decoding yields the bitstream I_k.

3.12.2 Demodulation of the Q component

Demodulation of the Q channel proceeds similarly. The quadrature-phase component is demodulated in a similar way except that the phase of the LO signals differ by $90°$. If the transmitted signal is the DSB-SC signal $s(t)$ with carrier f_c in Equation (3.13) (and replacing $\omega_{\rm LO}$ by f_c) and with a receiver LO frequency $f_{\rm LO}$, the in-phase demodulated signal is

$$\begin{aligned}
q'_{\rm IF}(t) &= G_L s(t)\sin(\omega_{\rm LO}t - \pi/2)\\
&= -s(t)\cos(\omega_{\rm LO}t)\\
&= -\tfrac{1}{2}\left[A_i(f_i)\cos(\omega_c-\omega_i)t + A_q(f_q)\sin(\omega_c-\omega_q)t\right.\\
&\quad \left. -\ A_i(f_i)\cos(\omega_c+\omega_i)t - A_q(f_q)\sin(\omega_c+\omega_q)t\right]\cos(\omega_{\rm LO}t)\\
&= -\tfrac{1}{2}\left\{A_i(f_i)\cos(\omega_c+\omega_{\rm LO}-\omega_i)t + A_i(f_i)\cos(\omega_c-\omega_{\rm LO}-\omega_i)t\right.\\
&\quad + A_q(f_q)\sin(\omega_c-\omega_{\rm LO}-\omega_q)t + A_q(f_q)\sin(\omega_c+\omega_{\rm LO}-\omega_q)t\\
&\quad - A_i(f_i)\cos(\omega_c+\omega_{\rm LO}+\omega_i)t - A_i(f_i)\cos(\omega_c-\omega_{\rm LO}+\omega_i)t\\
&\quad \left. -A_q(f_q)\sin(\omega_c-\omega_{\rm LO}+\omega_q)t + A_q(f_q)\sin(\omega_c+\omega_{\rm LO}+\omega_q)t.\right\}
\end{aligned} \tag{3.22}$$

This includes intermediate frequency terms centered at $\omega_c - \omega_{\rm LO}$ and high frequency terms centered at $\omega_c + \omega_{\rm LO}$. These high frequency terms can be eliminated through lowpass filtering leaving the lowpass filtered receiver signal

$$\begin{aligned}
q_{\rm IF}(t) &= -\tfrac{1}{2}\left[A_i(f_i)\cos(\omega_c-\omega_{\rm LO}-\omega_i)t + A_q(f_q)\sin(\omega_c-\omega_{\rm LO}-\omega_q)t\right.\\
&\quad \left. -\ A_i(f_i)\cos(\omega_c-\omega_{\rm LO}+\omega_i)t - A_q(f_q)\sin(\omega_c-\omega_{\rm LO}+\omega_q)t\right]\\
&= \tfrac{1}{2}A_i(f_i)\left[-\cos(\omega_c-\omega_{\rm LO}-\omega_i)t + \cos(\omega_c-\omega_{\rm LO}+\omega_i)t\right]\\
&\quad + \tfrac{1}{2}A_q(f_q)\left[-\sin(\omega_c-\omega_{\rm LO}-\omega_q)t + \sin(\omega_c-\omega_{\rm LO}+\omega_q)t\right]\\
&= \tfrac{1}{2}A_i(f_i)\left[-\cos(\omega_i-\omega_c+\omega_{\rm LO})t + \cos(\omega_i+\omega_c-\omega_{\rm LO})t\right]\\
&\quad + \tfrac{1}{2}A_q(f_q)\left[\sin(\omega_q-\omega_c+\omega_{\rm LO})t + \sin(\omega_q+\omega_c-\omega_{\rm LO})t\right].
\end{aligned} \tag{3.23}$$

Single-Stage Demodulation With $f_{\rm LO} = f_c$

If the LO frequency equals the carrier frequency but there is an offset of ϕ_i so that $\omega_{\rm LO}t = \omega_c t + \phi_i$, then the recovered baseband signal is

$$\begin{aligned}
q_{\rm BB}(t) &= \tfrac{1}{2}A_i(f_i)\left[-\cos(\omega_i t+\phi_i) + \cos(\omega_i t-\phi_i)\right\}\\
&\quad + \tfrac{1}{2}A_q(f_q)\left\{\sin(\omega_q t+\phi_i) + \sin(\omega_q t-\phi_i)\right].
\end{aligned} \tag{3.24}$$

Then if $\phi_i = 0$ the demodulated baseband signal for the I channel is

$$q'_{\rm BB}(t) = A_q(f_q)\sin(\omega_q t) \tag{3.25}$$

Two stage demodulation $f_{\mathrm{LO}} < f_c$

When $q_{\mathrm{IF}}(t)$ in Equation (3.17) is mixed with an IF LO signal at frequency $f_{\mathrm{LO,IF}}$ the resulting signal is

$$\begin{aligned} q'_{\mathrm{BB}}(t) = \{ & \tfrac{1}{2}A_i(f_i)\,[\sin(\omega_i - \omega_c + \omega_{\mathrm{LO}})t + \sin(\omega_i + \omega_c - \omega_{\mathrm{LO}})t] \\ & + \tfrac{1}{2}A_q(f_q)\,[\cos(\omega_q - \omega_c + \omega_{\mathrm{LO}})t - \cos(\omega_q + \omega_c - \omega_{\mathrm{LO}})t]\} \\ & \times \cos(\omega_{\mathrm{LO,IF}})t. \end{aligned} \tag{3.26}$$

The intermediate LO frequency $f_{\mathrm{LO,IF}}$ can be set with fine adjustment so that $f_{\mathrm{LO,IF}} = f_c - f_{\mathrm{LO}}$ which produces a baseband signal $q_{\mathrm{BB}}(t)$. (Note that $\omega_c - \omega_{\mathrm{LO}} - \omega_{LO,\mathrm{IF}} = 0$, $-\omega_c + \omega_{\mathrm{LO}} + \omega_{LO,\mathrm{IF}} = 0$, $\omega_c - \omega_{\mathrm{LO}} + \omega_{\mathrm{LO,IF}} = 2\omega_{\mathrm{LO,IF}}$, and $-\omega_c + \omega_{\mathrm{LO}} - \omega_{\mathrm{LO,IF}} = -2\omega_{\mathrm{LO,IF}}$. The Q channel baseband signal is then

$$\begin{aligned} q'_{\mathrm{BB}}(t) &= q_{\mathrm{IF}}(t)\cos(\omega_{\mathrm{LO,IF}}t) \\ &= G_L \left\{ \tfrac{1}{2}A_i(f_i)\,[-\cos(\omega_i - \omega_c + \omega_{\mathrm{LO}})t + \cos(\omega_i + \omega_c - \omega_{\mathrm{LO}})t] \right. \\ &\quad \left. + \tfrac{1}{2}A_q(f_q)\,[\sin(\omega_q - \omega_c + \omega_{\mathrm{LO}})t + \sin(\omega_q + \omega_c - \omega_{\mathrm{LO}})t] \right\} \\ &\quad \times \cos(\omega_{\mathrm{LO,IF}}t) \\ &= \tfrac{1}{2}G_L A_i(f_i)\,[-\cos(\omega_i + 0)t + \cos(\omega_i - 2\omega_{\mathrm{LO,IF}})t \\ &\quad + \cos(\omega_i + 2\omega_{\mathrm{LO,IF}})t + \cos(\omega_i + 0)t] \\ &\quad + \tfrac{1}{2}G_L A_q(f_q)\,[\sin(\omega_q - 2\omega_{\mathrm{LO,IF}})t + \sin(\omega_q + 0)t \\ &\quad - \sin(\omega_q + +2\omega_{\mathrm{LO,IF}})t - \cos(\omega_q + 0)t]\,. \end{aligned} \tag{3.27}$$

After lowpass filtering (in the DSP) to eliminate components centered at $\pm 2\omega_{\mathrm{LO,IF}}$

$$q_{\mathrm{BB}}(t) = G_L A_q(f_q)\cos(\omega_q t). \tag{3.28}$$

This can be compared to the original quadrature signal which led to the DSB-SC modulated signal $s(t)$. The f_q frequency component of $q(t)$ was $A_q f_q) sin(\omega_q t)$. Thus it is necessary to change the phase of $q_{\mathrm{BB}}(t)$ to obtain the original $q(t)$ signal:

$$\begin{aligned} q_{\mathrm{BB}}(t) \text{ (phase shifted by } 90^\circ &= G_L A_q(f_q)\cos(\omega_q t - \pi/2) q(t) \\ &= G_L A_q(f_q)\sin(\omega_q t) \end{aligned} \tag{3.29}$$

and this can be implemented using a Hilbert transform just as was done in the SDR transmitter.

3.13 SDR Summary

The software defined radio described in this chapter used two stages of quadrature modulation in the SDR transmitter and two stages of quadrature demodulation in the SDR receiver. Each of the quadrature modulators and demodulators used ideal multipliers to multiply signals by an LO signal. There are other ways of implementing quadrature de/modulators but the principles of one stage of the quadrature modulator being implemented in a baseband DSP unit and the final analog RF modulator being analog; and one of the quadrature demodulator being analog and operating on the RF signal and the second stage quadrature demodulator being implemented in DSP operating on an intermediate frequency signal is essential to implementations of SDR.

3.14 Summary

This chapter considered the architectures of transmitters and receivers and in particular the implementation of modulators and demodulators. Since the beginning of radio there has been a tremendous increase in the sophistication of modulators and demodulators. The earliest AM demodulators required just a single diode and now the demodulators in software defined radio use hundreds of millions of transistors. Following early radios the discussion tracked the development of modulators and demodulators for cellular systems but the increasingly advanced techniques are used in other radio and radar systems. The 1G systems used FM modulation to transmit voice and simple FSK modulation to transmit limited digital data such as phone such as channel information. The 2G and 3G systems were confined to using just one or two types of modulation such four-state modulation. The level of complexity increased tremendously with 4G where now it it was essential to use software defined radio to support a very large number of different modulation formats including mostly various orders of QAM but also supporting legacy modulation formats. The 5G modulation and demodulation schemes are the same as those used in 4G with the 5G advances being in the move to millimeter-waves and beam steering. There is another layer of modulation used in 3G (i.e. wideband CDMA) which can layer multiple users in the same bandwidth. Strictly this is an access scheme and will be considered in Chapter 5. The 4G systems uses orthogonal frequency division modulation (OFDM) modulation. That is not considered here as it is a hybrid of an access scheme and a modulation scheme. It too will be considered in Chapter 5.

3.15 References

[1] S. Darlington, “On digital single-sideband modulators,” *IEEE Trans. on Circuit Theory*, vol. 17, no. 3, pp. 409–414, 1970.

[2] R. V. Hartley, “Modulating system.” 1918, uS Patent 1,287,982.

[3] F. Colebrook, “Homodyne,” *Wireless World and Radio. Rev*, 1924.

[4] D. Tucker, “The synchrodyne,” *Electronic Engineering*, vol. 19, pp. 75–76, Mar. 1947.

[5] E. Armstrong, “Some recent developments in the audion receiver,” *Proc. of the Institute of Radio Engineers*, vol. 3, no. 3, pp. 215–238, Sep. 1915.

[6] L. Lessing, *Man of High Fidelity: Edwin Howard Armstrong, a Biography*. J.B. Lippincott, 1956.

[7] B. A. Weaver, “A new, high efficiency, digital, modulation technique for am or ssb sound broadcasting applications,” *IEEE Trans. on Broadcasting*, vol. 38, no. 1, pp. 38–42, 1992.

3.16 Exercises

1. A superheterodyne receiver has, in order, an antenna, a low-noise amplifier, a bandpass filter, a mixer, a second bandpass filter, a second mixer, a lowpass filter, an ADC, and a DSP that will implement quadrature demodulation. Develop the frequency plan of the receiver if the RF input is at 2 GHz and has a 200 kHz single-channel bandwidth. The final signal applied to the ADC must be between DC and 400 kHz so that I/Q demodulation can be done in the DSP unit. Noise considerations mandate that the LO of the first mixer must be more than 10 MHz away from the input RF. Also, for a bandpass filter to have minimum size, the center frequency of the filter should be as high as possible. It has been determined that the appropriate trade-off of physical size and cost is to have a 100 MHz bandpass filter between the two mixers. (Note: 100 MHz is the center frequency of the bandpass filter.)

   (a) Draw a block diagram of the receiver and annotate it with symbols for the frequencies of the LOs and the RF and IF signals.

(b) What is the LO frequency f_{LO1} of the first mixer?

(c) What is the LO frequency f_{LO2} of the second mixer?

(d) Specify the cutoff frequency of the lowpass filter following the second mixer.

(e) Discuss in less than $\frac{1}{2}$ a page other design considerations relating to frequency plan, filter size, and filter specification.

2. Short answer questions. Each part requires a short paragraph of about five lines and a figure where appropriate to illustrate your understanding.

   (a) Explain the operation of a superheterodyne receiver.

   (b) Compare zero-IF and low-IF receivers.

CHAPTER 4

Antennas and the RF Link

4.1 Introduction

An antenna interfaces circuits with free-space with a transmit antenna converting a guided wave signal on a transmission line to an electromagnetic (EM) wave propagating in free space, while a receive antenna is a transducer that converts a free-space EM wave to a guided wave on a transmission line and eventually to a receiver circuit. Together the transmit and receive antennas are part of the RF link. The RF link is the path between the output of the transmitter circuit and the input of the receiver circuit (see Figure 4-1). Usually this path includes the cable from the transmitter to the transmit antenna, the transmit antenna itself, the propagation path, the receive antenna, and the transmission line connecting the receive antenna to the receiver circuit. The received signal is much smaller than the transmitted signal. The overwhelming majority of the loss is from the propagation path as the EM signal spreads out, and usually diffracts, reflects, and is partially blocked by objects such as hills and buildings.

The first half of this chapter is concerned with the properties of antennas. One of the characteristics of antennas is that the energy can be focused in a particular direction, a phenomenon captured by the concept of antenna gain, which can partially compensate for path loss. The second half of this chapter considers modeling the RF link and the geographical arrangement of antennas that manage interference from other radios while providing support for as many users as possible.

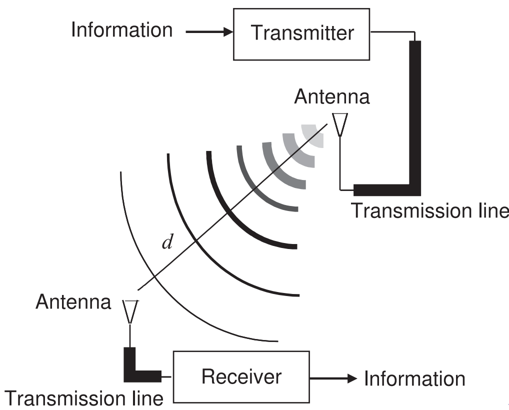
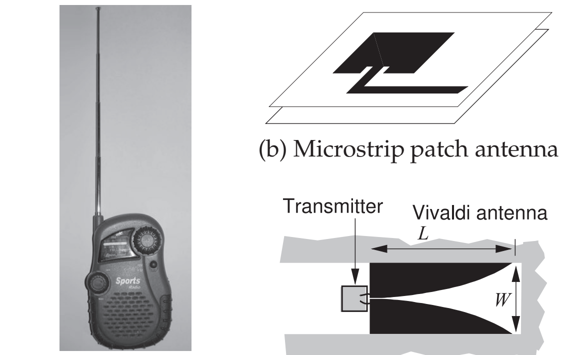


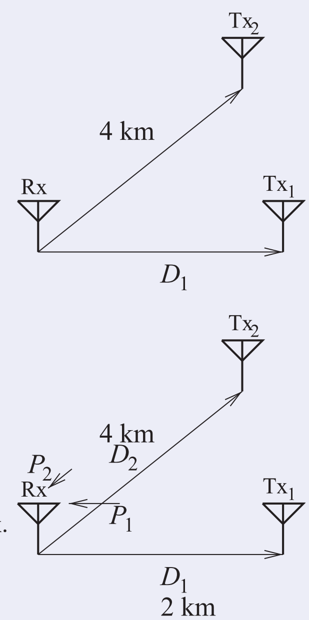
Figure 4-1: RF link.



(a) Monopole antenna (b) Microstrip patch antenna (c) Vivaldi antenna

Figure 4-2: Representative resonant, (a) and (b), and traveling-wave, (c), antennas.

EXAMPLE 4.1 Interference



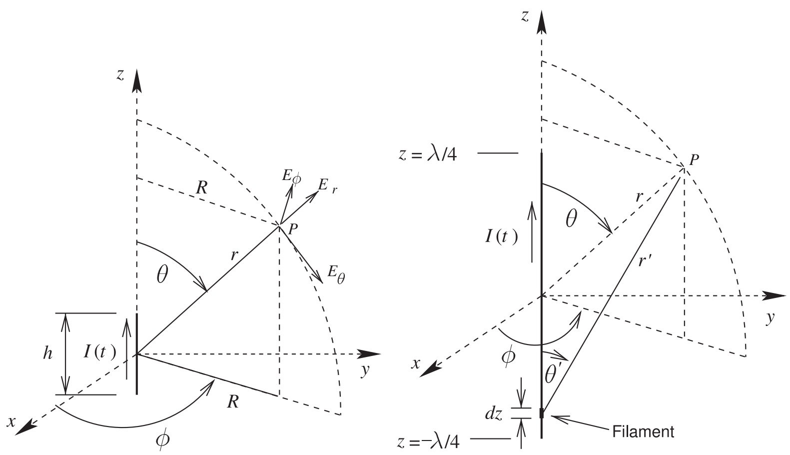
In the figure there are two transmitters, Tx_1 and Tx_2, operating at the same power level, and one receiver, Rx. Tx_1 is an intentional transmitter and its signal is intended to be received at Rx. Tx_1 is separated from Rx by $D_1 = 2$ km. Tx_2 uses the same frequency channel as Tx_1, and as far as RX is concerned it transmits an interfering signal. Assume that the antennas are omnidirectional (i.e., they transmit and receive signals equally in all directions) and that the transmitted power density drops off as $1/d^2$, where d is the distance from the transmitter. Calculate the signal-to-interference ratio (SIR) at Rx.

Solution:

$D_1 = 2\text{ km}$ and $D_2 = 4\text{ km}$.
P_1 is the signal power transmitted by Tx1 and received at Rx.
P_2 is the interference power transmitted by Tx2 and received at Rx.
So SIR $= \dfrac{P_1}{P_2} = \left(\dfrac{D_2}{D_1}\right)^2 = 4 = 6.02$ dB.

4.2 RF Antennas

Antennas are of two fundamental types: **resonant** and **traveling-wave antennas**, see Figure 4-2. Resonant antennas establish a standing wave of current with required resonance usually established when the antenna section is either a quarter- or half-wavelength long. These antennas are also known as standing-wave antennas. Resonant antennas are inherently narrowband because of the resonance required to establish a large standing wave of current. Figures 4-2(a and b) show two representative resonant antennas. The physics of the operation of a resonant antenna is understood by considering the time domain. First consider the physical operation of a transmit antenna. When a sinusoidal voltage is applied to the conductor



(a) Current filament ($h \ll \lambda$) (b) Half-wavelength long wire

Figure 4-3: Wire antennas. The distance from the center of an antenna to the field point P is r. $R = r \sin \theta$.

of an antenna, charges, i.e. free electrons, accelerate or decelerate under the influence of an applied voltage source which typically arrives at the antenna from a traveling wave voltage on a transmission line. When the charges accelerate (or decelerate) they produce an EM field which radiates away from the antenna. At a point on the antenna there is a current sinusoidally varying in time, and the net acceleration of the charges (in charge per second per second) is also sinusoidal with an amplitude that is directly proportional to the amplitude of the current sinewave. Hence having a large standing wave of current, when the antenna resonates sinusoidally, results in a large charge acceleration and hence large radiated fields.

When an EM field impinges upon a conductor the field causes charges to accelerate and hence induce a voltage that propagates on a transmission line connected to the receive antenna. Traveling-wave antennas, an example of one is shown in Figures 4-2(c), operate as extended lines that gradually flare out so that a traveling wave on the original transmission line transitions into free space. Traveling-wave antennas tend to be two or more wavelengths long at the lowest frequency of operation. While relatively long, they are broadband, many 3 or more octaves wide (e.g., extending from 500 MHz to 4 GHz or more). These antennas are sometimes referred to as **aperture antennas**.

4.3 Resonant Antennas

With a resonant antenna the current on the antenna is directly related to the amplitude of the radiated EM field. Resonance ensures that the standing wave current on the antenna is high.

4.3.1 Radiation from a Current Filament

The fields radiated by a resonant antenna are most conveniently calculated by considering the distribution of current on the antenna. The analysis begins by considering a short filament of current, see Figure 4-3(a). Considering the sinusoidal steady state at radian frequency ω, the current on the filament with phase χ is $I(t) = |I_0| \cos(\omega t + \chi)$, so that $I_0 = |I_0| e^{-\jmath\chi}$ is the phasor of the current on the filament. The length of the filament is h, but it has no other dimensions, that is, it is considered to be infinitely thin.

Resonant antennas are conveniently modeled as being made up of an array of current filaments with spacings and lengths being a tiny fraction of a wavelength. Wire antennas are even simpler and can be considered to be a line of current filaments. Ramo, Whinnery, and Van Duzer [1] calculated the spherical EM fields at the point P with the spherical coordinates (ϕ, θ, r) generated by the z-directed current filament centered at the origin in Figure 4-3. The total EM field components in phasor form are

$$H_\phi = \frac{I_0 h}{4\pi} \mathrm{e}^{-\jmath kr} \left(\frac{\jmath k}{r} + \frac{1}{r^2} \right) \sin\theta, \quad \bar{H}_\phi = H_\phi \hat{\phi} \tag{4.1}$$

$$E_r = \frac{I_0 h}{4\pi} \mathrm{e}^{-\jmath kr} \left(\frac{2\eta}{r^2} + \frac{2}{\jmath\omega\varepsilon_0 r^3} \right) \cos\theta, \quad \bar{E}_r = E_r \hat{\mathbf{r}} \tag{4.2}$$

$$E_\theta = \frac{I_0 h}{4\pi} \mathrm{e}^{-\jmath kr} \left(\frac{\jmath\omega\mu_0}{r} + \frac{1}{\jmath\omega\varepsilon_0 r^3} + \frac{\eta}{r^2} \right) \sin\theta, \quad \bar{E}_\theta = E_\theta \hat{\theta}, \tag{4.3}$$

where η is the free-space characteristic impedance, ε_0 is the permittivity of free space, and μ_0 is the permeability of free space. The variable k is called the **wavenumber** and $k = 2\pi/\lambda = \omega\sqrt{\mu_0\varepsilon_0}$. The $\mathrm{e}^{-\jmath kr}$ terms describe the variation of the phase of the field as the field propagates away from the filament. Equations (4.1)–(4.3) are the complete fields with the $1/r^2$ and $1/r^3$ dependence describing the near-field components. In the far field, i.e. large $r \gg \lambda$, the components with $1/r^2$ and $1/r^3$ dependence become negligible and the field components left are the propagating components H_ϕ and E_θ:

$$H_\phi = \frac{I_0 h}{4\pi} \mathrm{e}^{-\jmath kr} \left(\frac{\jmath k}{r} \right) \sin\theta, \quad E_r = 0, \quad \text{and} \quad E_\theta = \frac{I_0 h}{4\pi} \mathrm{e}^{-\jmath kr} \left(\frac{\jmath\omega\mu_0}{r} \right) \sin\theta. \tag{4.4}$$

Now consider the fields in the plane normal to the filament, that is, with $\theta = \pi/2$ radians so that $\sin\theta = 1$. The fields are now

$$H_\phi = \frac{I_0 h}{4\pi} \mathrm{e}^{-\jmath kr} \left(\frac{\jmath k}{r} \right) \quad \text{and} \quad E_\theta = \frac{I_0 h}{4\pi} \mathrm{e}^{-\jmath kr} \left(\frac{\jmath\omega\mu_0}{r} \right) \tag{4.5}$$

and the wave impedance is

$$\eta = \frac{E_\theta}{H_\phi} = \frac{I_0 h}{4\pi} \mathrm{e}^{-\jmath kr} \frac{\jmath\omega\mu_0}{r} \left(\frac{I_0 h}{4\pi} \mathrm{e}^{-\jmath kr} \frac{\jmath k}{r} \right)^{-1} = \frac{\omega\mu_0}{k}. \tag{4.6}$$

Note that the strength of the fields is directly proportional to the magnitude of the current. This proves to be very useful in understanding spurious radiation from microwave structures. Now $k = \omega\sqrt{\mu_0\epsilon_0}$, so

$$\eta = \frac{\omega\mu_0}{\omega\sqrt{\mu_0\epsilon_0}} = \sqrt{\frac{\mu_0}{\epsilon_0}} = 377\ \Omega, \tag{4.7}$$

as expected. Thus an antenna can be viewed as having the inherent function of an impedance transformer converting from the lower characteristic impedance of a transmission line (often $50\ \Omega$) to the $377\ \Omega$ characteristic impedance of free space. Further comments can be made about the propagating fields (Equation (4.4)). The EM field propagates in all directions except not directly in line with the filament. For fixed r, the amplitude of the propagating field increases sinusoidally with respect to θ until it is maximum in the direction normal to the filament.

The power radiated is obtained using the **Poynting vector**, which is the cross-product of the propagating electric and magnetic fields. From this the time-average propagating power density is (with the SI units of $\mathrm{W/m^2}$)

$$P_R = \frac{1}{2}\Re\left(E_\theta H_\phi^*\right) = \frac{\eta k^2 |I_0|^2 h^2}{32\pi^2 r^2}\sin^2\theta, \tag{4.8}$$

and the power density is proportional to $1/r^2$. In Equation (4.8) $\Re(\cdots)$ indicates that the real part is taken.

4.3.2 Finite-Length Wire Antennas

The EM wave launched by a wire antenna of finite length is obtained by considering the wire as being made up of many filaments and the field is then the superposition of the fields from each filament. As an example, consider the antenna in Figure 4-3(b) where the wire is half a wavelength long. As a good approximation the current on the wire is a standing wave and the current on the wire is in phase so that the current phasor is

$$I(z) = I_0\cos(kz). \tag{4.9}$$

From Equation (4.4) and referring to Figure 4-3 the fields in the far field are

$$H_\phi = \int_{-\lambda/4}^{\lambda/4} \frac{I_0\cos(kz)}{4\pi} \mathrm{e}^{-\jmath kr'} \left(\frac{\jmath k}{r'}\right)\sin\theta'\,dz \tag{4.10}$$

$$E_\theta = \int_{-\lambda/4}^{\lambda/4} \frac{I_0\cos(kz)}{4\pi} \mathrm{e}^{-\jmath kr'} \left(\frac{\jmath\omega\mu_0}{r'}\right)\sin\theta'\,dz, \tag{4.11}$$

where θ' is the angle from the filament to the point P. Now $k = 2\pi/\lambda$ and at the ends of the wire $z = \pm\lambda/4$ where $\cos(kz) = \cos(\pm\pi/2) = 0$. Evaluating the equations is analytically involved and will not be done here. The net result is that the fields are further concentrated in the plane normal to the wire. At large r, of at least several wavelengths distant from the antenna, only the field components decreasing as $1/r$ are significant. At large r the phase differences of the contributions from the filaments is significant and results in shaping of the fields. The geometry to be used in calculating the far field is shown in Figure 4-4(a). The phase contribution of each filament, relative to that at $z = 0$, is $(kz\sin\theta)/\lambda$ and Equations (4.10) and (4.11) become

$$H_\phi = I_0\left(\frac{\jmath k}{4\pi r}\right)\sin(\theta)\,\mathrm{e}^{-\jmath kr}\int_{-\lambda/4}^{\lambda/4}\frac{\sin(kz)}{4\pi}\sin(z\sin(\theta))dz \tag{4.12}$$

$$E_\theta = I_0\left(\frac{\jmath\omega\mu_0}{4\pi r}\right)\sin(\theta)\,\mathrm{e}^{-\jmath kr}\int_{-\lambda/4}^{\lambda/4}\sin(kz)\sin(z\sin(\theta))dz. \tag{4.13}$$

Figure 4-4(b) is a plot of the near-field electric field in the y-z plane calculating E_r and E_θ (recall that $E_\phi = 0$) every $90°$. Further from the antenna the E_r component rapidly reduces in size, and E_θ dominates.

A summary of the implications of the above equations are, first, that the strength of the radiated electric and magnetic fields are proportional to the

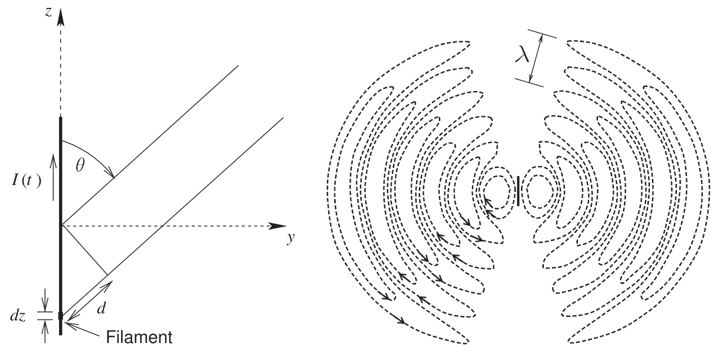


Figure 4-4: Wire antenna: (a) geometry for calculating contributions from current filaments of length dz with cordinate z ($d = -z\sin\theta$); and (b) instantaneous electric field in the y-z plane due to a $\lambda/2$ long current element. There is also a magnetic field.

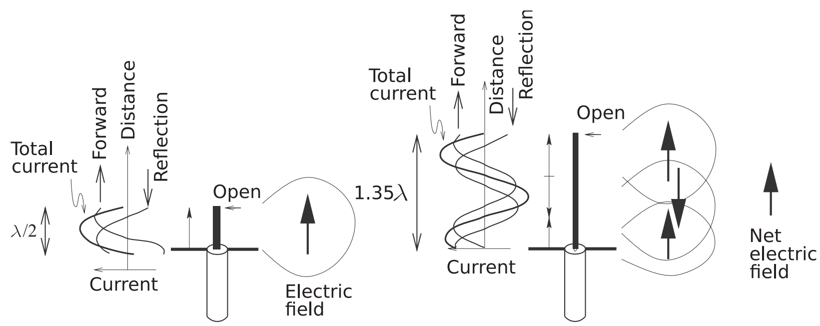


Figure 4-5: Monopole antenna showing total current and forward- and backward-traveling currents: (a) a $\frac{1}{2}\lambda$-long antenna; and (b) a relatively long antenna.

current on the wire antenna. So establishing a standing current wave and hence magnifying the current is important to the efficiency of a wire antenna. A second result is that the power density of freely propagating EM fields in the far field is proportional to $1/r^2$, where r is the distance from the antenna. A third interpretation is that the longer the antenna, the flatter the radiated transmission profile; that is, the radiated energy is more tightly confined to the x-y (i.e. $\Theta = 0$) plane. For the wire antenna the peak radiated field is in the plane normal to the antenna, and thus the wire antenna is generally oriented vertically so that transmission is in the plane of the earth and power is not radiated unnecessarily into the ground or into the sky.

To obtain an efficient resonant antenna, all of the current should be pointed in the same direction at a particular time. One way of achieving this is to establish a standing wave, as shown in Figure 4-5(a) At the open-circuited end, the current reflects so that the total current at the end of the wire is zero. The initial and reflected current waves combine to create a standing wave. Provided that the antenna is sufficiently short, all of the total current—the standing wave—is pointed in the same direction. The optimum length is about a half wavelength. If the wire is longer, the contributions to the field from the oppositely directed current segments cancel (see Figure 4-5(b)).

In Figure 4-5(a) a coaxial cable is attached to the monopole antenna below the ground plane and often a series capacitor between the cable and the antenna provides a low level of coupling leading to a larger standing wave. The capacitor also approximately matches the characteristic impedance of the cable to the input impedance Z_{in} of the antenna. If the length of the monopole is reduced to one-quarter wavelength long it is again resonant, and the input impedance, Z_{in}, is found to be $36\ \Omega$. Then a $50\ \Omega$ cable can

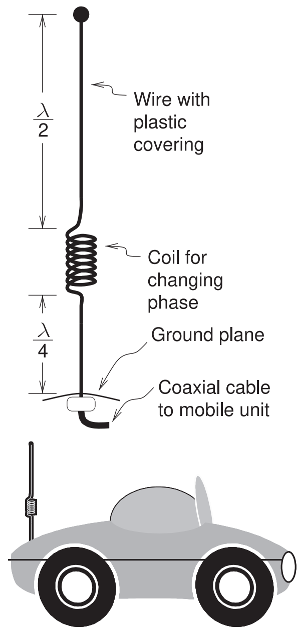


Figure 4-6: Mobile antenna with phasing coil extending the effective length of the antenna.

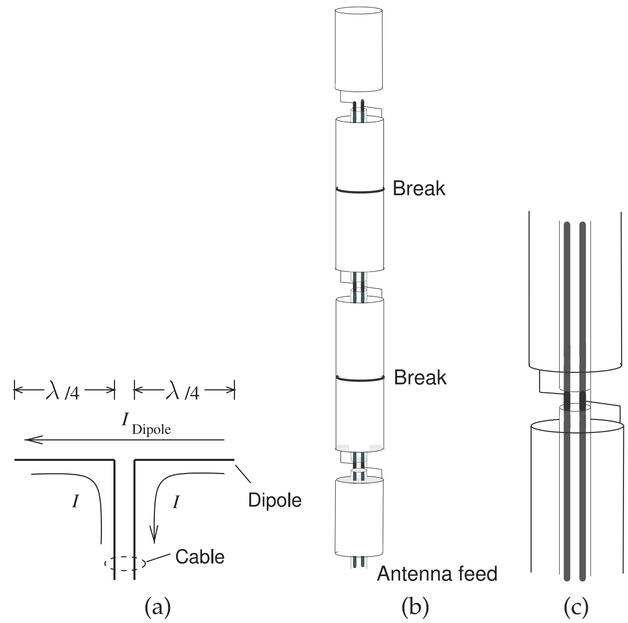
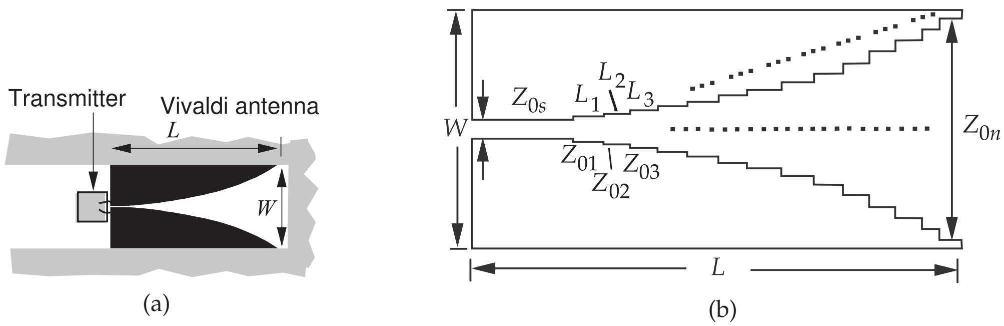


Figure 4-7: Dipole antenna: (a) current distribution; (b) stacked dipole antenna; and (c) detail of the connection in a stacked dipole antenna.

be directly connected to the antenna and there is only a small mismatch and nearly all the power is transferred to the antenna and then radiated.

Another variation on the monopole is shown in Figure 4-6, where the key component is the phasing coil. The phasing coil (with a wire length of $\lambda/2$) rotates the electrical angle of the current phasor on the line so that the current on the $\lambda/4$ segment is in the same direction as on the $\lambda/2$ segment. The result is that the two straight segments of the loaded monopole radiate a more tightly confined EM field. The phasing coil itself does not radiate (much).

Another ingenious solution to obtaining a longer effective wire antenna with same-directed current (and hence a more tightly confined RF beam) is the stacked dipole antenna (Figure 4-7). The basis of the antenna is a dipole as shown in Figure 4-7(a). The cable has two conductors that have equal amplitude currents, I, but flowing as shown. The wire section is coupled to the cable so that the currents on the two conductors realize a single effective current I_{dipole} on the dipole antenna. The stacked dipole shown in Figure 4-7(b) takes this geometric arrangement further. Now the radiating element is hollow and a coaxial cable is passed through the antenna elements and the half-wavelength sections are fed separately to effectively create a wire antenna that is several wavelengths long with the current pointing in one direction. Most cellular antennas using wire antennas are stacked dipole antennas.



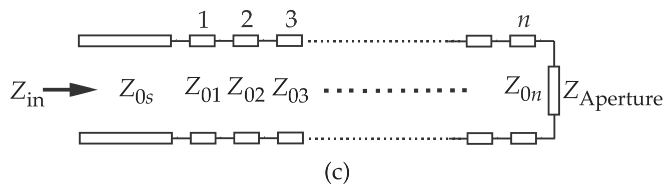


Figure 4-8: Vivaldi antenna showing design procedure: (a) the antenna; (b) a stepped approximation; and (c) transmission line approximation. In (a) the black regions are metal.

Summary

Standing waves of current can be realized by resonant structures other than wires. A microstrip patch antenna, see Figure 4-2(b), is an example, but the underlying principle is that an array of current filaments generates EM components that combine to create a propagating field. Resonant antennas are inherently narrowband because of the reliance on the establishment of a standing wave. A relative bandwidth of 5%–10% is typical.

4.4 Traveling-Wave Antennas

Traveling-wave antennas have the characteristics of broad bandwidth and large size. These antennas begin as a transmission line structure that flares out slowly, providing a low reflection transition from a transmission line to free space. The bandwidth can be very large and is primarily dependent on how gradual the transition is.

One of the more interesting traveling-wave antennas is the **Vivaldi antenna** of Figure 4-8(a). The Vivaldi antenna is an extension of a slotline in which the fields are confined in the space between two metal sheets in the same plane. The slotline spacing increases gradually in an exponential manner, much like that of a Vivaldi violin (from which it gets its name), over a distance of a wavelength or more. A circuit model is shown in Figures 4-8(b and c) where the antenna is modeled as a cascade of many transmission lines of slowly increasing characteristic impedance, Z_0. Since the Z_0 progression is gradual there are low-level reflections at the transmission line interfaces. The forward-traveling wave on the antenna continues to propagate with a negligible reflected field. Eventually the slot opens sufficiently that the effective impedance of the slot is that of free space and the traveling wave continues to propagate in air.

The other traveling-wave antennas work similarly and all are at least a wavelength long, with the central concept being a gradual taper from the

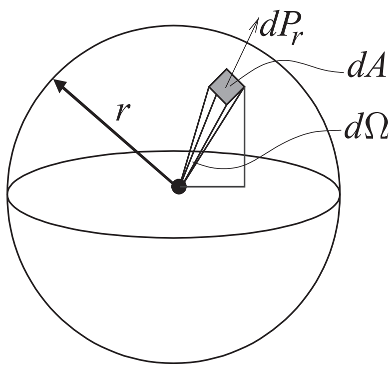


Figure 4-9: Free-space spreading loss. The incremental power, dP_r, intercepted by the shaded region of incremental area dA is proportional to $1/r^2$. The solid angle subtended by the shaded area is the incremental solid angle $d\Omega$. The integral of dA over the surface of the sphere, i.e. the area of the sphere is $4\pi r^2$. The total solid angle subtended by the sphere is the integral of $d\Omega$ over the sphere and is 4π steradians (or 4π sr).

characteristic impedance of the originating transmission line to free space. The final aperture is at least one-half wavelength across so that the fields can curl on themselves (i.e., loop back on themselves) and are self-supporting as they leave the antenna.

4.5 Antenna Parameters

This section introduces a number of antenna metrics that are used to characterize antenna performance.

4.5.1 Radiation Density and Radiation Intensity

Antennas do not radiate equally in all directions concentrating radiated power in (usually) one direction called the **main** (or **major**) **lobe** of the antenna. This focusing effect is called **directivity**. The power in a particular direction is characterized by the radiation density and the radiation intensity metrics. The radiation density, S_r, is the power per unit area with the SI units of $\mathrm{W/m^2}$, and will be maximum in the main lobe. Referring to Figure 4-9 with an antenna located at the center of the sphere of radius r and radiating a total power P_r, S_r is the incremental radiated power dP_r passing through the incremental shaded region of area, dA:

$$S_r = \frac{dP_r}{dA}. \tag{4.14}$$

S_r reduces with distance falling of as $1/r^2$ in free space. For a practical antenna S_r will vary across the surface of the sphere. The total power radiated is the closed integral over the surface S of the sphere:

$$P_r = \oint_S dP_r = \oint_S S_r dA. \tag{4.15}$$

An alternative measure of power concentration is the radiation intensity U which is in terms of the incremental solid angle $d\Omega$ subtended by dA so that $d\Omega = dA/r^2$ and (with the SI units of W/steradian or W/sr)

$$U = \frac{dP_r}{d\Omega} = \frac{dP_r}{dA} r^2 = r^2 S_r. \tag{4.16}$$

Isotropic Antenna

It is useful to reference the directivity of an antenna with respect to a fictitious isotropic antenna that has no loss and radiates equally in all directions so that S_r is only a function of r. Then integrating over the surface of the sphere yields the total radiated power

$$P_r|_{\text{Isotropic}} = \oint_S dP_r = \oint_S S_r dA = S_r \oint_S dA = S_r 4\pi r^2 = 4\pi U. \tag{4.17}$$

Since the isotropic antenna has no loss the power input to the antenna P_{IN} is equal to the power radiated $P_r = P_{\text{IN}}$. Thus for an isotropic antenna

$$S_r = \frac{P_r}{4\pi r^2} = \frac{P_{\text{IN}}}{4\pi r^2} \quad (4.18) \quad \text{and} \quad U|_{\text{Isotropic}} = r^2 S_r = \frac{P_{\text{IN}}}{4\pi r^2}. \quad (4.19)$$

Antenna Efficiency

Antenna efficiency, sometimes called the **radiation efficiency**, describes losses in an antenna principally due to resistive (I^2R) losses. Resonant antennas work by creating a large current that is maximized through the generation of a standing wave at resonance. There is a lot of current, and even just a little resistance results in substantial resistive loss. The power that is reflected from the input of the antenna is usually small. The total radiated power (in all directions), P_r, is the power input to the antenna less losses. The antenna efficiency, η_A is therefore defined as

$$\eta_A = P_r/P_{\text{IN}}, \tag{4.20}$$

where P_{IN} is the power input to the antenna and $\eta_A < 1$ and usually expressed as a percentage. Antenna efficiency is very close to one for many antennas, but can be 50% for microstrip patch antennas.

Antenna loss refers to the same mechanism that gives rise to antenna efficiency. Thus an antenna with an antenna efficiency of 50% has an antenna loss of 3 dB. Generally losses are resistive due to I^2R loss and mismatch loss of the antenna that occurs when the input impedance is not matched to the impedance of the cable connected to the antenna. Because of confusion with antenna gain (they are not the opposite of each other) the use of the term 'antenna loss' is discouraged and instead 'antenna efficiency' preferred.

4.5.2 Directivity and Antenna Gain

The directivity of an antenna, D, is the ratio of the radiated power density to that of an isotropic antenna with the same total radiated power, P_r:

$$D = \frac{S_r}{S_r|_{\text{Isotropic}}} = \frac{U}{U|_{\text{Isotropic}}} \tag{4.21}$$

where S_r and U refer to the actual antenna and the power densities and intensities are measured at the same distance from the antennas. For an actual antenna D is dependent on the direction from the antenna, see Figure 4-10. The maximum value of D will be in the direction of the main lobe of the antenna and this is called **directivity gain**.

The focusing property of an antenna is characterized by comparing the radiated power density to that of an isotropic antenna with the same input power. The antenna gain, G_A, is the maximum value of D when the power input $P_{\text{IN}} = P_r/\eta_A$ to the antenna and the isotropic antenna are the same:

$$G_A = \eta_A \max(D). \tag{4.22}$$

Antenna	Type	Figure	Gain (dBi)	Notes
Lossless isotropic antenna			0	
$\lambda/2$ dipole	Resonant	4-7(a)	2	$R_{\text{in}} = 73\ \Omega$
3λ diameter parabolic dish	Traveling	—	38	$R_{\text{in}} = \text{match}$
Patch	Resonant	4-2(b)	9	$R_{\text{in}} = \text{match}$
Vivaldi	Traveling	4-2(c)	10	$R_{\text{in}} = \text{match}$
$\lambda/4$ monopole on ground	Resonant	4-5(a)	2	$R_{\text{in}} = 36\ \Omega$
$5/8\lambda$ monopole on ground	Resonant	4-5(a)	3	Matching required

Table 4-1: Several antenna systems. $R_{\text{in}} = \text{match}$ for resonant antennas indicates that the antenna can be designed to have an input impedance matching that of a feed cable. Traveling-wave antennas are intrinsically matched.

Losses in the antenna are accounted for by the efficiency term η_A.

In Equation (4.22) G_A is a gain factor and is often expressed in terms of decibels (taking 10 times the log of G_A) but dBi (with 'i' standing for with-respect-to isotropic) is used to indicate that it is not a power gain in the same sense as amplifier gain. G_A instead is the ratio of power densities for two different antennas. For example, an antenna that focuses power in one direction increasing the peak radiated power density by a factor of 20 relative to that of an isotropic antenna has an antenna gain of 13 dBi. With care G_A can be often used in calculations of power as as with amplifier gain.

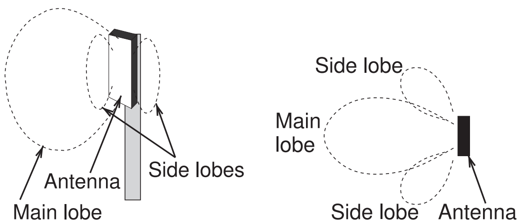
Since it is almost impossible to calculate internal antenna losses, antenna gain is invariably only measured. The input power to an antenna can be measured and the peak radiated power density, $P_D|_{\text{Maximum}}$, measured in the far field at several wavelengths distant (at $r \gg \lambda$). This is compared to the power density from an ideal isotropic antenna at the same distance with the same input power. Antenna gain is determined from

$$
\begin{aligned}
G_A &= \frac{\text{Maximum radiated power per unit area}}{\text{Maximum radiated power per unit area for an isotropic antenna}} \\
&= \frac{S_r|_{\text{Maximum}}}{S_r|_{\text{Isotropic}}} = 4\pi r^2 \frac{P_D|_{\text{Maximum}}}{P_{\text{IN}}} \qquad (4.23)\\
&= 4\pi \frac{\text{Maximum radiated power per unit solid angle}}{\text{Total input power to the antenna}} \\
&= 4\pi \frac{(dP_r/d\Omega)|_{\text{Maximum}}}{P_{\text{IN}}} = 4\pi r^2 \frac{(dP_r/dA)|_{\text{Maximum}}}{P_{\text{IN}}}. \qquad (4.24)
\end{aligned}
$$

The antenna gains of common resonant and traveling-wave antennas are given in Table 4-1. In free space the antenna gain determined using Equation (4.22) is independent of distance. Antenna gain is measured on an antenna range using a calibrated receive antenna and care taken to avoid reflections from objects, especially from the ground.

The losses of an antenna are incorporated in the antenna gain which is defined in terms of the power input to the antenna, see Equation (4.24). Thus in calculations of radiated power using antenna gain, there is no need to separately account for resistive losses in the antenna.

In summary, antennas concentrate the radiated power in one direction so that the density of the power radiated in the direction of the peak field is higher than the power density from an isotropic antenna. Power radiated from a base station antenna, such as that shown in Figure 4-11, is concentrated in a region that looks like a toroid or, more closely, a balloon squashed at its north and south poles. Then the antenna does not radiate much power into space and will concentrate power in a region skimming the surface of



(a) 3D view (b) Top view

Figure 4-10: Field pattern produced by a microstrip antenna.

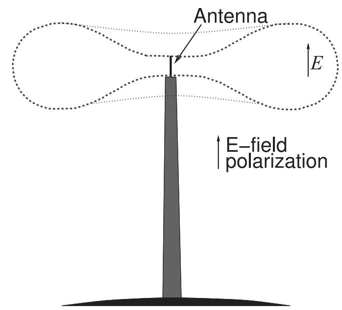


Figure 4-11: A base station transmitter pattern.

the earth. Antenna gain is a measure of the effectiveness of an antenna to concentrate power in one direction. Thus, in free space due to power spreading out the maximum power density (in SI units of $\mathrm{W/m^2}$) at a distance r is

$$P_D = \frac{G_A P_{\mathrm{IN}}}{4\pi r^2}, \tag{4.25}$$

where $4\pi r^2$ is the area of a sphere of radius r and P_{IN} is the input power.

Measurements of antenna gain are used to derive antenna efficiency. It is impossible to measure or simulate the resistive and dielectric losses of an antenna directly. Antenna efficiency is obtained using theoretical calculations of antenna gain assuming no losses in the antenna itself. This is compared to the measured antenna gain yielding the antenna efficiency.

EXAMPLE 4.2 Antenna Gain

A base station antenna has an antenna gain, G_A, of 11 dBi and a 40 W input. The transmitted power density falls off with distance d as $1/d^2$. What is the peak power density at 5 km?

Solution:

A sphere of radius 5 km has an area $A = 4\pi r^2 = 3.142 \cdot 10^8\ \mathrm{m}^2$; $G_A = 11$ dBi = 12.6.
In the direction of peak radiated power, the power density at 5 km is

$$P_D = \frac{P_{\mathrm{in}} G_A}{A} = \frac{40 \cdot 12.6\ \mathrm{W}}{3.142 \cdot 10^8 \mathrm{m}^2} = 1.603\ \mu\mathrm{W/m}^2.$$

EXAMPLE 4.3 Antenna Efficiency

A antenna has an antenna gain of 13 dBi and an antenna efficiency of 50% and all of the loss is due to resistive losses and resistance of metals is proportional to temperature. The RF signal input to the antenna has a power of 40 W.

(a) What is the input power in dBm?

P_{in} = 40 W = 46.02 dBm.

(b) What is the total power transmitted in dBm?

$P_{\mathrm{Radiated}} = 50\%$ of $P_{\mathrm{IN}} = 20$ W or 43.01 dBm.
Alternatively, $P_{\mathrm{Radiated}} = 46.02$ dBm $- 3$ dB $= 43.02$ dBm.

(c) If the antenna is cooled to near absolute zero so that it is lossless, what would the antenna gain be?

The antenna gain would increase by 3 dB and antenna gain incorporates both directivity and antenna losses. So the gain of the cooled antenna is 16 dBi.

4.5.3 Effective Isotropic Radiated Power

A transmit antenna does not radiate power equally in all directions and for a receiver in the main lobe of the transmit antenna it is as though there is an isotropic transmit antenna with a much higher input power. This concept is incorporated in the effective isotropic radiated power (**EIRP**):

$$\text{EIRP} = P_{\text{IN}} G_A. \tag{4.26}$$

This uses antenna gain to derive the total power that would be radiated by an isotropic antenna producing the same (peak) power density as the actual antenna. Regulatory limits on the power levels of transmitters are in terms of EIRP rather than the total radiated power. Sometimes **equivalent radiated power** (**ERP**) is used instead of EIRP with the same meaning.

4.5.4 Effective Aperture Size

Effective aperture size is defined so that the power density at a receive antenna when multiplied by its effective aperture size, A_R, yields the power output from the antenna at its connector. Since antennas are linear and reciprocal it is is to be expected that there is a relationship between effective aperture size and antenna gain.

An antenna has an effective size that is more than its actual physical size. This is because of its influence on the EM fields around it. The power captured by an antenna is the effective aperture size (or area) multiplied by the transmitted power density. That is, the effective aperture size of an antenna is the area of the surface that captures all of the power passing through it and delivers this power to the output terminals of the antenna.

The effective aperture area of a receive antenna, A_R, is related to the receive antenna gain, G_R, as follows [2, 3] ((note that A_e is often used if it is not necessary to distinguish antennas):

$$A_R = \frac{G_R \lambda^2}{4\pi}, \tag{4.27}$$

where λ is the wavelength of the radio signal[1]. The effective aperture area of an antenna can have little to do with its physical size; e.g., a wire antenna has almost no physical size but has a significant effective aperture size.

If S_r is the transmitted power density at the receive antenna, the power received is

$$P_R = P_D A_R = P_D \frac{G_R \lambda^2}{4\pi}. \tag{4.28}$$

[1] The derivation of Equation (4.27) invokes a thought experiment with an antenna connected to a resistor in thermal equilibrium and each in separate thermally isolated chambers with black body walls [4, 5]. The available thermal (Johnson-Nyquist) noise power from the resistor is radiated by the antenna (incorporating antenna gain) and absorbed by the walls of the antenna's chamber. The same amount of power, as black body radiation from the walls of the antenna's chamber, must be received by the antenna (incorporating effective aperture size) and delivered to the resistor where it is dissipated as heat. At frequencies up to infrared and at room temperature (and hence at radio frequencies) the power density of black body radiation increases linearly with temperature and as the square of frequency (i.e. as $1/\lambda^2$) whereas Johnson-Nyquist noise is independent of frequency but increases linearly with temperature. The derivation is exact for an isotropic antenna and assumed to apply to all antennas.

Here P_R is the power delivered at the output connector of the receive antenna, as loss in the receive antenna is incorporated in G_R and A_R. The total power radiated by the transmit antenna in the direction of maximum power density is given by multiplying the power input to the transmit antenna, P_T,[2] by the antenna gain of the transmit antenna, G_T. This can be converted to the power density at a distance d (ignoring multipath effects),

$$S_r = \frac{P_T G_T}{4\pi d^2}, \tag{4.29}$$

and the power delivered by the receive antenna is

$$P_R = S_r A_R = \frac{P_T G_T}{4\pi d^2} \frac{G_R \lambda^2}{4\pi} = P_T G_T G_R \left(\frac{\lambda}{4\pi d} \right)^2. \tag{4.30}$$

That is,

$$P_R = P_T G_T G_R \left(\frac{\lambda}{4\pi d} \right)^2. \tag{4.31}$$

This equation is known as the **Friis transmission equation** or the **Friis transmission formula**.

Equation (4.31) provides a powerful tool for evaluating the power received in a communication system and modifications have been developed to account for impairments that are encountered. One of the assumptions in the development of Equation (4.31) is that the polarization of the receive antenna matches that of the transmitted signal. The transmit antenna will radiate a signal with an electric field with a particular polarization, i.e. E field orientation. Propagation through air has little effect on the polarization of a signal, although atmospheric conditions can rotate the polarization slightly. A polarization mismatch factor could be included in Equation (4.31). For ground-based communications, such as cellular communications, reflections and diffraction have a major impact on the power of the received signal and techniques for handling these effects are considered in the next section.

4.5.5 Summary

This section introduced several metrics for characterizing antennas:

Metric	Equation	Description
S_r	(4.14)	Radiated power density, W/m^2
U	(4.16)	Radiation intensity W/sr
η_A	(4.27)	Antenna efficiency
D	(4.21)	Antenna directivity
G_A	(4.23)	Antenna gain,used with a transmit antenna
A_e	(4.27)	Effective aperture area, used with a receive antenna
EIRP	(4.26)	Equivalent isotropic radiated power

[2] Note that this was P_{IN} previously and here the subscript T is used to indicate the power input to the transmit antenna.

EXAMPLE 4.4 Point-to-Point Communication

In a point-to-point communication system, a parabolic receive antenna has an antenna gain of 60 dBi. If the signal is 60 GHz and the power density at the receive antenna is 1 pW/cm^2, what is the power at the output of the receive antenna connected to the RF electronics?

Solution:

The first step is determining the effective aperture area, A_R, of the parabolic antenna. The frequency is 60 GHz, and so the wavelength (λ) is 5 mm. Also note that G_R = 60 dBi = 10^6. From Equation (4.27),

$$A_R = \frac{G_R \lambda^2}{4\pi} = \frac{10^6 \cdot 0.005^2}{4\pi} = 1.989\ \mathrm{m}^2. \tag{4.32}$$

Using Equation (4.28) P_D = 1 pW/cm^2 = 10 nW/m^2, the total power delivered to the RF receiver electronics (at the output of the receive antenna) is

$$P_R = P_D A_R = 10\ \mathrm{nW} \cdot \mathrm{m}^{-2} \cdot 1.989\ \mathrm{m}^2 = 19.89\ \mathrm{nW}. \tag{4.33}$$

4.6 The RF Link

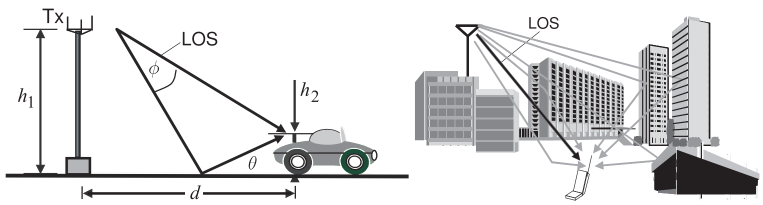
The RF link is between a transmit antenna and a receive antenna. Sometimes the RF link includes the antenna and sometimes it does not, this will be clear from the context, but usually it includes the antennas. The principle source of link loss is the spreading out of the EM field as it propagates. In the absence of any other effects (such as atmospheric loss and reflections), the power density reduces as $1/d^2$, where d is distance, and this is called the **line-of-sight** (**LOS**) situation. In this section the propagation path is first described along with its impairments including propagation on multiple paths between a transmit antenna and a receive antenna. The resonant scattering is described as well as multiple fading effects due to reflections, diffraction, rain, and other atmospheric effects.

4.6.1 Propagation Path

When the radiated signal reflects and diffracts there are multiple propagation paths that result in what is called fading, as the paths constructively and destructively combine at the receiver. The direct path and other paths, called multipaths, which are reflected or diffracted by ground, buildings, and other objects do not arrive at the receiver in phase leading to time-varying constructive and destructive combining, called multipath interference. Of these destructive combining is much worse as it can reduce a signal level below what it would be if propagation was in free space. In urban areas, 10 or 20 paths can have significant power in them and these combine at the receive antenna [6].

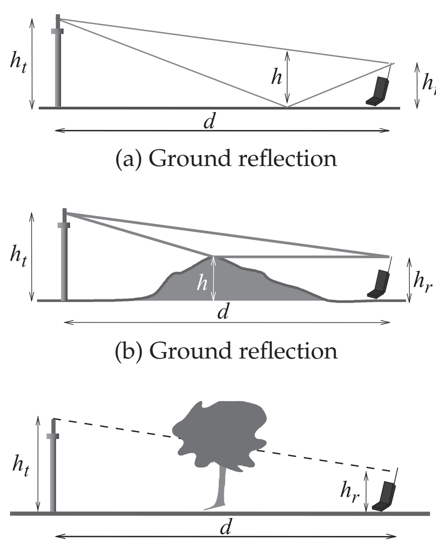
When the signal on one of the paths dominates, this is usually the LOS path, fading is called **Rician fading**. With LOS and a single ground reflection, the situation is the classic Rician fading as shown in Figure 4-12(a). This is a special situation, as the ground changes the phase of the signal upon reflection, generally by 180°. When the receiver is a long way from the base station, the lengths of the two paths are almost identical and the level of the signals in the two paths are almost the same. The net result is that these two signals almost cancel, and so instead of the power falling off by $1/d^2$, it falls off by $1/d^3$. When there are many paths and all have similar amplitude signals, fading is called **Rayleigh fading**. In an urban area such as that shown in Figure 4-12(b), there are many significant multipaths and the power falls off by $1/d^4$ and sometimes faster.

Common paths encountered in cellular radio are shown in Figure 4-13. As a rough guide, in the single-digit gigahertz range each diffraction and scattering event reduces the signal received by 20 dB. The knife-edge diffraction scenario is shown in more detail in Figure 4-14. This case is fairly easy to analyze and can be used to estimate the effects of individual obstructions. The diffraction model is derived from the theory of half-infinite screen diffraction [7]. First, calculate the parameter ν from the geometry of



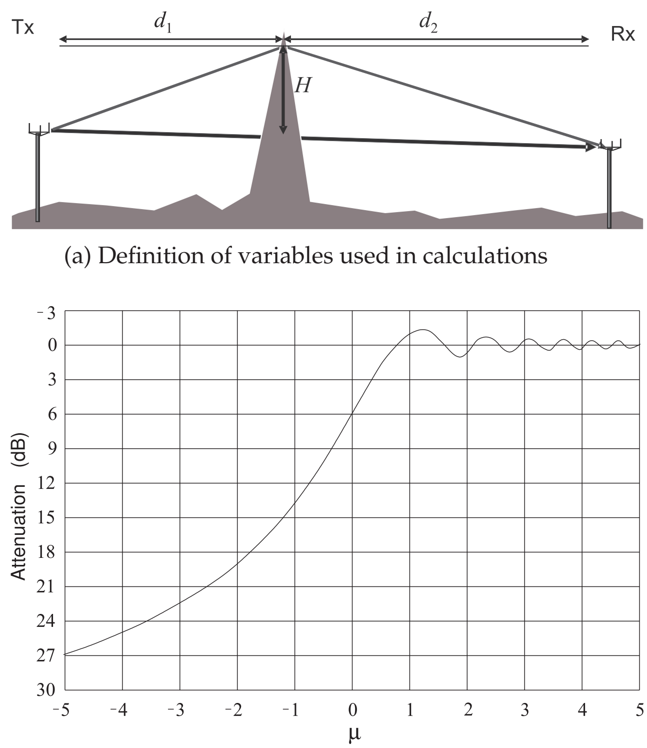
(a) Rician fading (b) Rayleigh fading

Figure 4-12: Multipath propagation: (a) line of sight (LOS) and ground reflection paths only; and (b) in an urban environment.



(c) Scattering through tree

Figure 4-13: Common paths contributing to multipath propagation.



(b) Chart for determining attenuation using Equation (4.34)

Figure 4-14: Knife-edge diffraction.

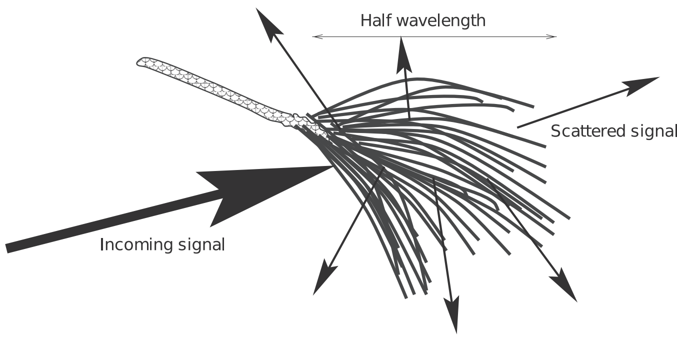


Figure 4-15: Pine needles scattering an incoming EM signal.

the path using

$$\nu = -H\sqrt{\frac{2}{\lambda}\left(\frac{1}{d_1}+\frac{1}{d_2}\right)}. \tag{4.34}$$

Next, consult the plot in Figure 4-14(b) to obtain the diffraction loss (or attenuation). This loss should be added (using decibels) to the otherwise determined path loss to obtain the total path loss. Other losses such as reflection cancellation still apply, but are computed independently for the path sections before and after the obstruction.

4.6.2 Resonant Scattering

Propagation is rarely from point to point as the path is often obstructed and so non-LOS (**NLOS**). One type of event that reduces transmission is scattering. The level of the effect depends on the size of the objects causing scattering. Here the effect of the pine needles of Figure 4-15 will be considered. The pine needles (as most objects in the environment) conduct electricity, especially when wet. When an EM field is incident an individual needle acts as a wire antenna, with the current maximum when the pine needle is one-half wavelength long. At this length, the "needle" antenna supports a standing wave and will re-radiate the signal in all directions. This is scattering, and there is a considerable loss in the direction of propagation of the original fields. As well, there is loss due to a needle not being a very good conductor, so EM energy is lost as heat. The effect of scattering is frequency and size dependent. A typical pine needle is 15 cm long, which is exactly $\lambda/2$ at 1 GHz, and so a stand of pine trees have an extraordinary impact on cellular communications at 1 GHz. As a rough guide, 20 dB of a signal is lost when passing through a small stand of pine trees. At 2 GHz, a similar impact comes from other leaves, and they do not need to look like wire antennas. Consider an oak leaf having a dimension of 7.5 cm. This is $\lambda/2$ at 2 GHz, another dominant cellular frequency. So scattering results, but now the loss is season dependent, with the loss due to scattering being considerably smaller in winter (when trees such as oaks do not have leaves) than in summer.

4.6.3 Fading

Fading refers to the variation of the received signal with time or when the position of transmit or receive antennas is changed. The variation in signal strength is much greater than would be expected from changes in path

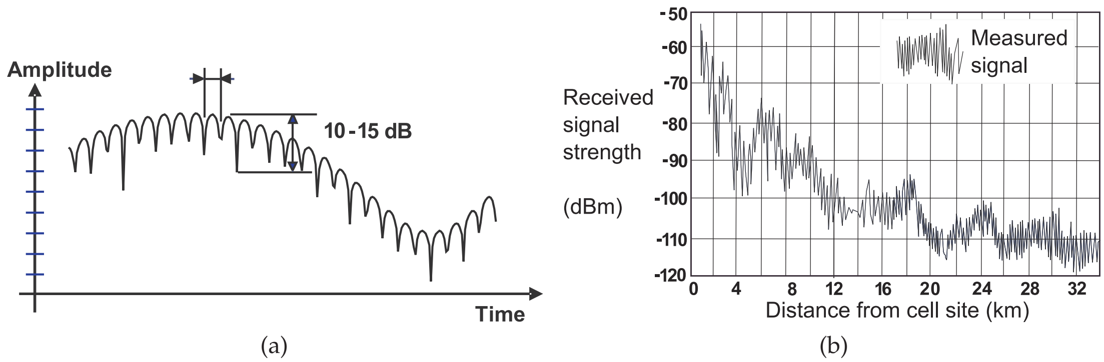


Figure 4-16: Fast and slow fading: (a) in time as a radio and obstructions move; and (b) in distance.

length. Several forms of fading are identified and associated with different physical effects. The most important fading types are flat fading, multipath fading, and rain fading.

Flat Fading

Temperature variations of the atmosphere between the transmit antenna and receive antenna give rise to what is called flat fading and sometimes called **thermal fading**. This fading is called flat because it is independent of frequency. One form of flat fading is due to refraction, which occurs when different layers of the atmosphere have different densities and thus dielectric permittivities increasing or decreasing away from the surface of the earth. The temperature profile can increase away from the earth surface or reduce depending on whether the temperature of the earth is higher than that of the air and is commonly associated with the beginning and end of the day. **Temperature inversions** can also occur where the temperature profile in air does not uniformly increase or decrease, thus causing a layer with a higher permittivity than that of the air above or below. RF energy gets trapped in this layer, reflecting from the top and bottom of the inversion layer. This is called **ducting**. In point-to-point communication systems the transmit and receive antennas are mounted high on towers and then reflection from ground objects is often small. In such cases flat fading is the most commonly observed phenomena and minor fluctuations of several decibels in receive signal level are common throughout the day. However, when temperature variations are extreme, ducting can severely impact communications reducing signal levels by up to 20 dB.

Shadow Fading

Shadow fading occurs when the LOS path is blocked by an obstruction such as a building or hill. The full signal strength returns when the obstruction or receiver has moved to restore the direct path. Shadow fading is also called slow fading and the characteristic of the channel is regarded as being relatively constant over a short time. The amplitude response varies in time and distance and is shown in Figure 4-16. This figure shows both fast fades that are 10 to 15 dB deep and slow or shadow fades that are 20 to 30 dB deep.

Multipath Fading

Multipath fading is the most common fading when either the transmit antenna or the receive antenna are close to the ground, near obstructing buildings or terrains, or inside a building. In such situations there are many reflections that combine destructively and constructively. Multipath fading is also called **fast fading**, as the characteristics of the channel can change significantly in a few milliseconds. Two types of multipath fading—Rician fading and Rayleigh fading—will be considered.

Multipath is principally a problem when the line of sight between transmit and receive antennas is obscured. However, where there is line of sight the signal reflecting from the ground immediately in front of a receive antenna can sometimes largely cancel the line-of-sight signal. In time, intervening objects can move and the propagation characteristics of the various paths can change due to thermal variations. All this adds to the randomness of fast fades. Multipath fading of 20 dB can occur for a small percentage of the time on time scales of many seconds when there are few propagation paths (e.g. in a rural area) to a large percentage of the time many times per second in a dense urban environment when there are many paths. Constructive combining does increase the signal level momentarily, but there is no advantage to this. Destructive combining can result in deep fades of 20 dB impacting communications and forcing the communication system to accommodate either by using higher average powers or using strategies such as multiple antennas or spreading the communication signal over a wide bandwidth since fades tend to be 500 kHz to 1 MHz wide at all frequencies.

Rician Fading

Rician fading occurs when there is one dominant RF path, usually the LOS path, and one or more other paths. The main situation is when there is an LOS path and a ground reflection as shown in Figure 4-12(a). The signal received is the sum of two signals:

$$v_r(t) = C\left[\cos(\omega_c t) + r\cos(\omega_c t + \phi)\right], \tag{4.35}$$

where $C\cos(\omega_c t)$ is the LOS signal, r is the ratio of the amplitude of the signal reflected from the ground and the LOS signal, and ϕ is the relative phase. When the distance between the transmitted and received signal is large, the ground reflection will be a glancing reflection and the LOS and reflected path will be almost exactly the same length. Ground reflection will usually introduce a 180° phase rotation in the reflected signal and r could be 0.8–0.9. Thus there will be nearly complete cancellation of the signal [8].

Rayleigh Fading

Rayleigh fading, or **fast fading**, in a multipath environment results from destructive cancellation as individual paths of similar amplitude drift in and out of phase as the receiver and sources of multiple reflections, diffractions, and refractions move. With multiple paths between a transmitter and a receiver, different components of the received signal arrive at different times. When the line of sight is blocked, the received signal on as many as 20 of these paths can have appreciable signal power. The time between when the first significant received signal is received and the last is received is called the

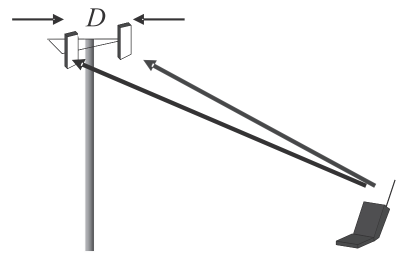


Figure 4-17: Two receive antennas used to achieve **space diversity** and overcome the effects of Rayleigh fading.

delay spread. The delay spread is largest inside buildings. A typical office building has a delay spread of 30 ns, and an office building with highly reflective walls and large open spaces has a **delay spread** of up to 250 ns [9–11], similar delay spreads are obtained at all RF frequencies [12]. Rayleigh fading is the type of fading that occurs when the inverse of the delay spread is small compared to the bandwidth of the received signal. Then when the multipaths combine destructively the entire signal within a bandwidth of 500 kHz to 1 MHz is suppressed. This bandwidth is experimentally observed and is related to the delay spread of individual paths, each with random amplitude and phase. If a communication signal has a bandwidth that is entirely within the suppression bandwidth of a Rayleigh fade, the situation is called frequency flat fading. Some communication signals have bandwidths greater than the Rayleigh fade bandwidth, so it is possible to recover from Raleigh fades if error correction is used.

The result of Rayleigh fading is rapid amplitude variation with respect to frequency. In Figure 4-16(a) the amplitude varies in time depending on the speed of moving objects. This graph could just as well be a plot of amplitude versus distance or amplitude versus frequency. A measured amplitude response is shown in Figure 4-16(b). Although somewhat random, the fades occur over distances roughly $\frac{1}{2}\lambda$ apart and 500 kHz wide. The 500 kHz width is almost independent of frequency from hundreds of megahertz to 100 GHz. The probability of receiving a signal x dB below the time-averaged received signal power is approximately 10^{-x} [13]. Rayleigh fading is named after the statistical model that describes this.

Fortunately deep Rayleigh fades are very short, last a small percentage of the time, and slight changes to the propagation environment can circumvent their effects. One strategy for overcoming fades is to use two receive antennas as shown in Figure 4-17. Here the signals received by two antennas separated by several wavelengths will rarely fade concurrently. In practice, the required separation for good decorrelation is found to be 10 to 20λ.

In an LOS wireless system (e.g. a point-to-point link without multipath), Rayleigh fading is due to rapidly changing atmospheric conditions, with the refractive index of small regions varying. These fades occur over a few seconds.

Rain Fading

Rain fading is due to both the amount of rain and the size of individual rain drops and fading occurs over periods of minutes to hours. Propagation through the atmosphere is affected by absorption by molecules in the air, fog, and rain, and by scattering by rain drops. Figure 4-18 shows the attenuation in decibels per kilometer from 3 GHz to 300 GHz. Curve (a) shows the

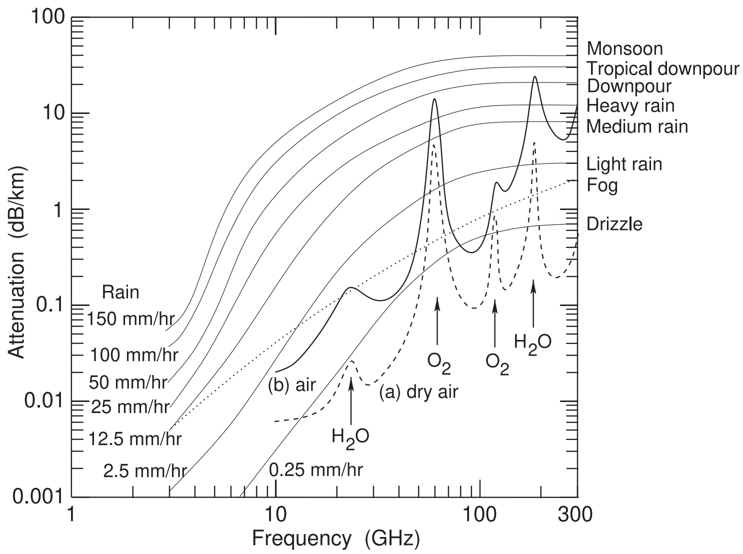


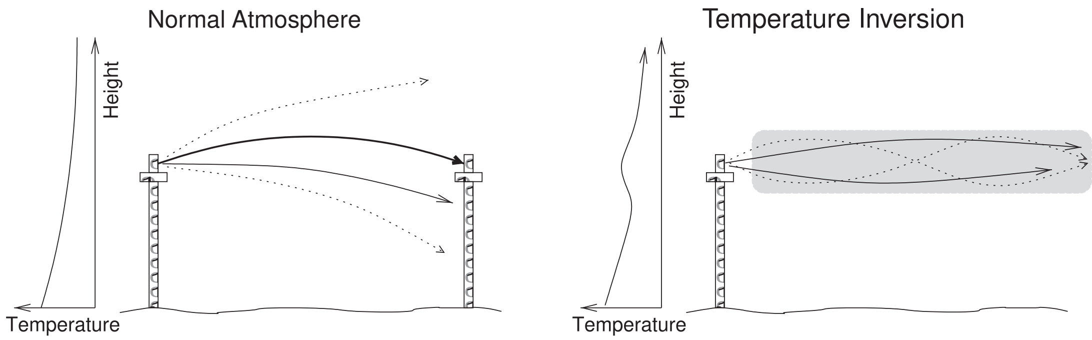
Figure 4-18: Excess attenuation due to atmospheric conditions showing the effect of rain on RF transmission at sea level. Curve (a) is atmospheric attenuation, due to excitation of molecular resonances, of very dry air at 0 °C, curve (b) is for typical air (i.e. less dry) at 20 °C. The attenuation shown for fog and rain is additional (in dB) to the atmospheric absorbtion shown as curve (b).

attenuation in dry air at 0°C with very low attenuation at a few gigahertz with several attenuation peaks as frequency increases. The first peak in attenuation is due to the resonance of water molecules in water vapor. The resonance peaks at 23 GHz but loss starts increasing before that. The next absorption peak is at 60 GHz due to the resonance of oxygen molecules. Two other absorption peaks are observed going up to 300 GHz. Curve (b) is for a less dry atmosphere at 20°C. The dotted curve shows the additional effect of fog on attenuation and then a family of curves shows the effect of rain on propagation. Below 10 GHz the effect of rain is very little and is safely ignored below 5 GHz. The attenuation due to rain increases with frequency and this derives largely from scattering and is related to the RF signal's wavelength relative to the circumference of rain drops.

Fading Due to Ducting

Normally the temperature and density of air, and thus its refractive index, drops with increasing height above earth. As a result, radio waves will be refracted toward the earth as shown in Figure 4-19(a). However, during periods of stable weather characterized by a high-pressure system, the temperature can rise with increasing height before eventually falling, creating what is called temperature inversion. This can occur at heights of tens to hundreds of meters. Temperature inversion is most common in summer, but can also occur when the temperature is dropping quickly, such as at sunset. At sunset the inversion layer may occur at a meter or so above the ground as the earth cools rapidly. The denser colder air above the ground has a higher permittivity than the air above and this results in an inversion layer that has a higher refractive index than the air above and below. RF waves can become trapped in the inversion layer, as any RF energy leaving the temperature inversion layer is refracted back into the layer. This effect is called ducting. Ducting can also occur when a cold air mass is overrun by warm air.

Fading due to ducting occurs when the receiver wanders in and out of the ducting layer, as the ducting layer is not stable, increasing and decreasing in



(a) Atmospheric refraction (b) Atmospheric ducting

Figure 4-19: Fading resulting from ducting: (a) normal atmospheric refraction (normally the temperature of air drops with increasing height and the lower refractive index at high heights results in a concave refraction); and (b) atmospheric ducting (resulting from temperature inversion inducing an air layer with higher dielectric constant than the surrounding air).

strength and in geometry.

Summary of Fading

The fades of most concern in a mobile wireless system are the deep fades resulting from destructive interference of multiple reflections. These fades vary rapidly (over a few milliseconds) if a handset is moving at vehicular speeds but occur slowly when the transmitter and receiver are fixed. Fades can be viewed as deep amplitude modulation, and so so modulation is restricted to phase shift keying schemes when a transmitter and receive moving at vehicular speeds relative to each other.

4.6.4 Link Loss and Path Loss

With transmit and receive antennas included in the RF link, the usual case, link loss is defined as the ratio of the power input to the transmit antenna (P_T) to the power delivered by the receive antenna (P_R). Rearranging Equation (4.31) and taking logarithms yields the total line-of-sight (LOS) link loss, $L_{\text{LINK,LOS}}$, between the input of the transmit antenna and the output of the receive antenna separated by distance d is (in decibels):

$$L_{\text{LINK,LOS}}|_{\text{dB}} = 10\log\left(\frac{P_T}{P_R}\right) = 10\log\left(\frac{P_T}{P_D A_R}\right) \tag{4.36}$$

$$= 10\log\left[P_T\left(\frac{4\pi d^2}{P_T G_T}\right)\left(\frac{4\pi}{\lambda^2 G_R}\right)\right] \tag{4.37}$$

$$= 10\log\left[\left(\frac{1}{G_T G_R}\right)\left(\frac{4\pi d}{\lambda}\right)^2\right] \tag{4.38}$$

$$= -10\log G_T - 10\log G_R + 20\log\left(\frac{4\pi d}{\lambda}\right). \tag{4.39}$$

The last term includes d and is called the LOS path loss (in decibels):

$$L_{\text{PATH,LOS}}|_{\text{dB}} = 20 \log \left(\frac{4\pi d}{\lambda} \right). \tag{4.40}$$

This is the preferred form of the expression for path loss, as it can be used directly in calculating link loss using the antenna gains of the transmit and receive antennas without the exercise of calculating the **effective aperture size** of the receive antenna.

An alternative definition of path loss comes directly from Equation (4.25) and is called the LOS path loss of the first kind:

$${}^{1}L_{\text{PATH,LOS}} = 4\pi d^2, \tag{4.41}$$

but this is not commonly used.

Multipath effects result in losses that are proportional to d^n [14, 15] so that the general path loss, including multipath effects, is (in decibels)

$$\begin{aligned} L_{\text{PATH}}|_{\text{dB}} &= L_{\text{PATH,LOS}}|_{\text{dB}} + \text{excess loss}|_{\text{dB}} \\ &= 20 \log \left(\frac{4\pi d}{\lambda} \right) + 10(n-2) \log \left(\frac{d}{1\ \text{m}} \right) \\ &= 20 \log \left[\frac{4\pi (1\ \text{m})}{\lambda} \right] + 10(2) \log \left(\frac{d}{1\ \text{m}} \right) + 10(n-2) \log \left(\frac{d}{1\ \text{m}} \right) \\ &= 10n \log[d/(1\ \text{m})] + C, \end{aligned} \tag{4.42}$$

where the distance d and wavelength λ are in meters, and C is a constant that captures the effect of wavelength. Here,

$$C = 20 \log \left[4\pi (1\ \text{m})/\lambda \right]. \tag{4.43}$$

Combining this with Equation (4.39) yields the link loss between the input of the transmit antenna and the output of the receive antenna:

$$L_{\text{LINK}}|_{\text{dB}} = -G_T|_{\text{dB}} - G_R|_{\text{dB}} + 10n \log[d/(1\ \text{m})] + C. \tag{4.44}$$

As you can imagine, a few constants, here n and C, cannot capture the full complexity of the propagation environment. Many models have been developed to capture particular environments better and incorporate mast height, experimental correction factors, and statistical parameters.

The path loss between two antennas is exactly the same in both directions when the frequency of the signal in each direction is the same. This is radio link reciprocity. Many communication systems use different frequencies in the two directions, and then the links are not reciprocal.

EXAMPLE 4.5 Power Density

A communication system operating in a dense urban environment has a power density roll-off of $1/d^{3.5}$ between the base station transmit antenna and the mobile receive antenna. At 10 m from the transmit antenna, the power density is $0.3167\ \text{W/m}^2$. What is the power density at the receive antenna located at 1 km from the base station?

Solution:

$P_D(10\ \text{m}) = 0.3167\ \text{W/m}^2$ and let d_c = 10 m, so at d = 1 km, the power density $P_D(1\ \text{km})$ is obtained from

$$\frac{P_D(1\ \text{km})}{P_D(10\ \text{m})} = \frac{d_c^{3.5}}{d^{3.5}} = \frac{10^{3.5}}{1000^{3.5}} = 10^{-7}, \tag{4.45}$$

so

$$P_D(1\ \text{km}) = P_D(10\ \text{m}) \cdot 10^{-7} = 31.7\ \text{nW/m}^2. \tag{4.46}$$

EXAMPLE 4.6 Link Loss

A 5.6 GHz communication system uses a transmit antenna with an antenna gain G_T of 35 dB and a receive antenna with an antenna gain G_R of 6 dB. If the distance between the antennas is 200 m, what is the link loss if the power density reduces as $1/d^3$? The link loss here is between the input to the transmit antenna and the output from the receive antenna.

Solution:

The link loss is provided by Equation (4.44),

$$L_{\text{LINK}}|_{\text{dB}} = -G_T - G_R + 10n\log[d/(1\ \text{m})] + C,$$

and C comes from Equation (4.43), where λ = 5.36 cm. So

$$C = 20\log\left(\frac{4\pi}{\lambda}\right) = 20\log\left(\frac{4\pi}{0.0536}\right) = 47.4\ \text{dB}.$$

With n = 3 and d = 200 m,

$$L_{\text{LINK}}|_{\text{dB}} = -35 - 6 + 10\cdot 3\cdot\log(200) + 47.4\ \text{dB} = 75.4\ \text{dB}.$$

EXAMPLE 4.7 Radiated Power Density

In free space, radiated power density drops off with distance d as $1/d^2$. However, in a terrestrial environment there are multiple paths between a transmitter and a receiver, with the dominant paths being the direct LOS path and the path involving reflection off the ground. Reflection from the ground partially cancels the signal in the direct path, and in a semi-urban environment results in an attenuation loss of 40 dB per decade of distance (instead of the 20 dB per decade of distance roll-off in free space). Consider a transmitter that has a power density of $1\ \text{W/m}^2$ at a distance of 1 m from the transmitter.

(a) The power density falls off as $1/d^n$, where d is distance and n is an index. What is n?
(b) At what distance from the transmit antenna will the power density reach 1 μW·m^{-2}?

Solution:

(a) Power drops off by 40 dB per decade of distance. 40 dB corresponds to a factor of 10,000 (= 10^4). So, at distance d, the power density $P_D(d) = k/d^n$ (k is a constant).
At a decade of distance, $10d$, $P_D(10d) = k/(10d)^n = P_D(d)/10000$, thus

$$\frac{k}{10^n d^n} = \frac{1}{10,000}\frac{k}{d^n}; \qquad 10^n = 10,000 \Rightarrow n = 4.$$

(b) At $d = 1$ m, $P_D(1\text{ m}) = 1\text{ W/m}^2$. At a distance x,

$$P_D(x) = 1\ \mu\text{W/m}^2 = \frac{k}{x^4}\text{m}^2 = \frac{k}{x^4} \rightarrow x^4 = \frac{1}{10^{-6}} \quad \text{and so} \quad x = 31.6\text{ m}.$$

4.6.5 Fresnel Zones

In long-distance wireless communications from one fixed site to another, the intent is to use the LOS path and avoid reflections. Such systems are called **point-to-point links**. Thus avoiding reflections is an important consideration in design. The dominant propagation paths in a point-to-point system are shown in Figure 4-20. The refracted wave path arises because of density variations in the air producing a permittivity profile that varies with height, as shown in Figure 4-21(a). This effect is called **beam bending**. The other important propagation path to consider is the reflected wave from the ground, which can be important if the ground is too close to the propagation path. Both of these effects will be considered in this section.

As radio waves propagate, they spread out in a plane perpendicular to the direction of propagation, the power density of the radio waves then reduces with distance from the centerline. One of the consequences of this is that an obstruction that is not in the LOS path can still interfere with signal propagation. The appropriate clearance is determined from the Fresnel zones, which are shown in Figure 4-22. The direct LOS path between the antennas has a length d. If there is a reflecting object near the LOS path, then there can be a second path between the transmit and receive antennas. The path from the first antenna to the circle defined by the first Fresnel zone and then to the second antenna has a path length $d+\lambda/2$, and so at the receive antenna this signal is $180°$ out of phase with the LOS signal and there will be

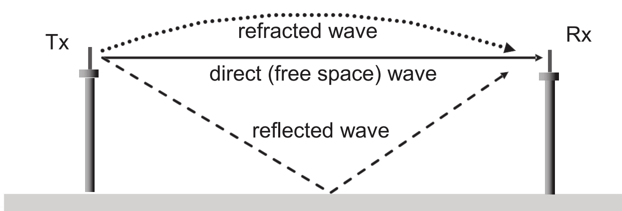


Figure 4-20: Three point-to-point characteristic propagation paths: line of sight, reflection, and refraction.

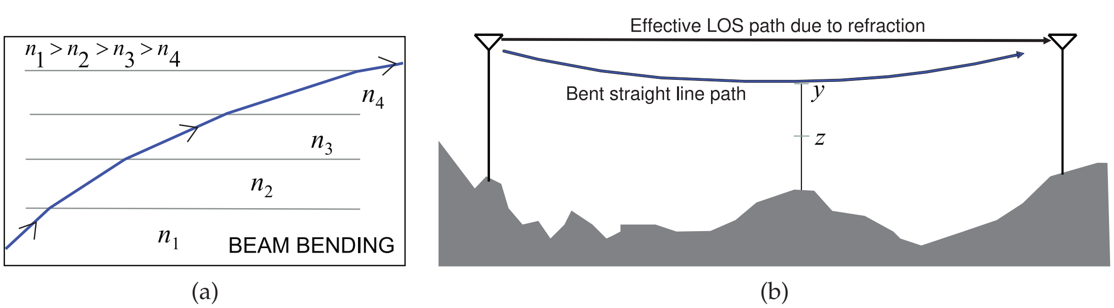


Figure 4-21: Beam bending by density variation in air: (a) refraction index profile with the air density reducing with height; and (b) incorporating beam bending in a curved-earth model.

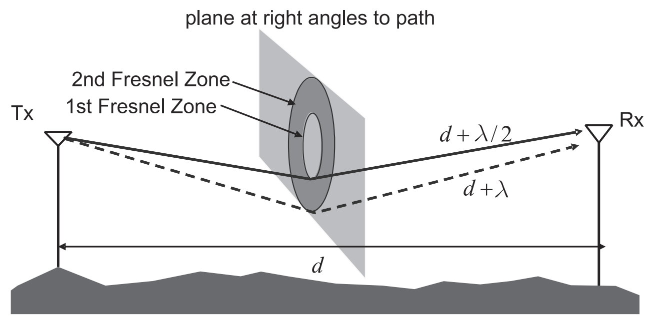


Figure 4-22: Fresnel zones in a plane perpendicular to the LOS path, which has length d.

cancellation. The radius of the nth Fresnel zone at a point P is

$$F_n = \sqrt{\frac{n\lambda d_1 d_2}{d_1 + d_2}}, \tag{4.47}$$

where d_1 is the distance from the first antenna to P, d_2 is the distance from the second antenna to P (so $d = d_1 + d_2$), and λ is the wavelength of the propagating signal. Ninety percent of the energy in the wave is in the first Fresnel zone. A guideline is that an obstacle should be separated from the direct path by a distance more than the radius of the first Fresnel zone. Antenna heights are increased so that the beam achieves one or more Fresnel zone clearances.

The above analysis can be used even with beam bending. A convenient way of accommodating beam bending is to use a curved-earth model, as shown in Figure 4-21(b), so that subsequent calculations can use LOS considerations [16]. The amount of beam bending to account for comes from experimental surveys. In Figure 4-21(b) the original clearance from a hill to the first Fresnel zone is z. With beam bending the clearance increases to y.

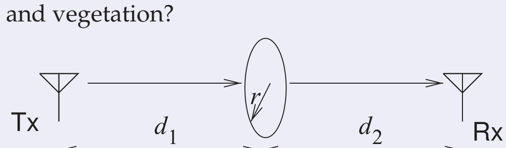
Since a receive antenna is unlikely to be large enough to capture all of the energy contained in the first Fresnel zone, the effect of signal blocking by an obstruction that encroaches the first Fresnel zone is of little concern. What is of concern is the destructive combining of a reflected signal with the signal in the main LOS beam.

EXAMPLE 4.8 Fresnel Zone Clearance

A transmit antenna and a receive antenna are separated by 10 km and operate at 2 GHz.

(a) What is the radius of the first Fresnel zone?
(b) What is the radius of the second Fresnel zone?
(c) To ensure LOS propagation, what should the clearance be from the direct line between the antennas and obstructions such as hills and vegetation?

Solution:



The radius of the Fresnel zone is calculated at the midpoint so that $d_1 = d_2 = d/2 = 5$ km. Also $f = 2$ GHz, $\lambda = 15$ cm.

(a) The radius of the first Fresnel zone is, from Equation (4.47),

$$r_1 = F_1 = \sqrt{\frac{\lambda d_1 d_2}{d_1 + d_2}} = \sqrt{\frac{\lambda d}{4}} = \sqrt{\frac{(0.15)(10^4)}{4}} = 19.36 \text{ m}.$$

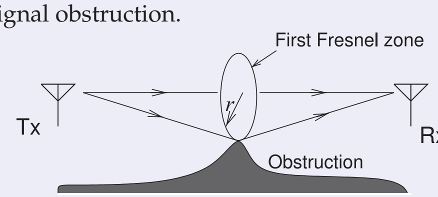
(b) At the midpoint the radius of the second Fresnel zone is

$$r_2 = F_2 = \sqrt{\frac{2\lambda d_1 d_2}{d_1 + d_2}} = \sqrt{\frac{\lambda d}{2}} = \sqrt{\frac{(0.15)(10^4)}{2}} = 27.39 \text{ m.}$$

(c) Ninety percent of the energy in the beam is contained within the first Fresnel zone of radius r_1. Obstructions within the first Fresnel zone will result in a significant fraction of the beam being blocked. In addition, reflections from the obstruction could destructively combine with the main beam and reduce the received signal level even beyond the reduction caused by part of the beam being blocked. Two criteria are commonly used to determine the clearance required to avoid signal obstruction.

One criterion used is that the minimum clearance between the direct beam and an obstruction is $0.6r_1$. So the minimum clearance is $0.6r_1 = 11.6$ m.

A more conservative criterion is that the minimum clearance should be $r_1 = 19.36$ m.



4.6.6 Propagation Model in the Mobile Environment

RF propagation in the mobile environment cannot be accurately derived. Instead, a fit to measurements is often used. One of the models is the Okumura–Hata model [17], which calculates the path loss as

$$L_{\text{PATH}}|_{(\text{dB})} = 69.55 + 26.16 \log f - 13.82 \log H + (44.9 - 6.55 \log H) \cdot \log d + c, \tag{4.48}$$

where f is the frequency (in MHz), d is the distance between the base station and terminal (in km), H is the effective height of the base station antenna (in m), and c is an environment correction factor (c = 0 dB in a dense urban area, c = -5 dB in an urban area, c = -10 dB in a suburban area, and c = -17 dB in a rural area, for f = 1 GHz and H=1.5 m).

More sophisticated characterization of the propagation environment uses ray-tracing models to follow individual propagation paths. The ray-tracing models are based on deterministic methods using terrain data and calculate paths accounting for obstruction and reflection analyses. Each refraction and reflection event is characterized either experimentally or through detailed EM simulations. Appropriate algorithms are applied for best compliance with radio physics. Commonly required inputs to these models include frequency, distance from the transmitter to the receiver, effective base station height, obstacle height and geometry, radius of the first Fresnel zone, forest height/roof height, distance between buildings, arbitrary loss allowances based on land use (forest, water, etc.), and loss allowances for penetration of buildings and vehicles. Such a technique was used to calculate the radio coverage diagrams on the inside front cover of this book.

There are many propagation models for different frequency ranges and different environments. The considerable effort put into developing reliable models is because being able to predict signal coverage is essential to efficient design of basestation layout.

4.7 Multipath and Delay Spread

Multipath was discussed previously in Sections 4.6.1 and 4.6.3 in the context of reduction in signal amplitude due to the destructive interference of a signal traveling on several paths. Another effect of multipath is that instances of a signal traveling from a transmitter to a receiver on different paths will arrive at a receiver with various delays. These delays can differ by tens or thousands of nanoseconds which is much longer than the period of a microwave signal, e.g. a 1 GHz signal has a period of 1 ns.

4.7.1 Delay Spread

Each of the paths supports what is called a signal instance. For a single-tone signal the effect of various delays is for the signal instances to arrive at the receiver with very different phases. As well, the signal instances invariably have different amplitudes resulting in a composite received signal that is an unpredictable combination of constructive and destructive interference. Sometimes the constructive interference will result in a stronger composite received signal but this is not that important. What is much more significant is the destructive interference which will result in signal fades with the signal received being very small at times. Consider two signal instances $y_1(t)$ and $y_2(t)$, traveling from a transmitter to a receiver on paths 1 and 2 respectively. If $y_1(t)$ and $y_2(t)$ have the same amplitude and phase then the combined signal will only be increased by 6 dB over a signal that travels on just one path. However if $y_1(t)$ and $y_2(t)$ have the same amplitude but phases that differ by 180° then there will be total cancellation and no composite signal will be received. So destructive interference is much more significant than constructive interference.

For cellular radio, in a rural area there can be two or three significant paths, usually taken as paths having signal instances that are within 20 dB of the largest signal instance. In an urban environment there can be tens of significant paths because there are many reflections from buildings and often there is not an LOS path which would otherwise have the largest signal instance. With multiple paths it is unlikely that there will be total cancellation of the received signal but it is very likely that the received signal will be smaller than if there was one path.

Early cellular radio had narrow bandwidths, e.g. 200 kHz for 2G's GSM cellular system, and low-order modulation, e.g. 2 bits per symbol and it was sufficient to use the phase-based concepts of constructive and destructive interference to understand the effects of multipath. As channel bandwidths increased and the order of modulation increased it was necessary to consider the actual physical effect of the paths having different delays.

If there is an LOS path, say with no building in the way, then the LOS signal instance will be stronger than a signal instance that travels on any other path. It will also have the smallest delay. This is partly because each non-LOS (NLOS) path will travel further and hence spread out more, and also because there will be signal loss at reflections. If there is scattering or diffraction on a path the amplitude of this signal instance on that path will be even smaller. Even if there is an LOS path, if there are multiple NLOS paths then the combined NLOS signal instances could indeed be larger than the LOS signal instance. If there is not an LOS path then one of the NLOS paths will become

Freq-uency (GHz)	Tx-Rx distance (km)	Environ-ment	max. excess delay $\tau_{d,\max}$ (ns)	median excess delay $\tau_{d,\text{med}}$ (ns)	
0.43	3.8†	urban	1300	900	[18]
0.90	2.2	rural	800	–	[19]
0.90	5	rural	1900	–	[20]
0.90	2	suburban	3700	300	[21]
0.90	2.2	suburban	900	300	[19]
0.90	0.6	urban	700	200	[22]
0.90	3.5	urban	5000	1100	[23]
0.90	7.0	urban	15000	1900	[24]
0.90	1.0	urban	2000	700	[21]
0.90	13	mountain$	3800	–	[25]
0.90	6.0	mountain$	1800	–	[26]
1.35	3.8†	urban	1400	850	[18]
2.26	3.8†	urban	1400	800	[18]
5.75	3.8†	urban	1000	300	[18]
28	0.052	urban LOS	754	<200	[27]
28	0.097	urban NLOS	1388	200	[27]
38	0.2‡	urban LOS	12	1.5	[28]
38	0.2‡	urban NLOS	133	14	[28]

Table 4-2: Measured excess delays for various carrier frequencies, Tx-Rx distances and multipath environments. Data has been normalized where needed so that $\tau_{d,\max}$, is the maximum excess delay that is exceeded only 1% of the time.
† d ranges from 0.04 km to 3.8 km and here $\tau_{d,\max}$ is the maximum delay exceeded only 10% of the time.
‡ high gain steerable antennas.
$ urban area surrounded by mountains

the largest received signal instance and it may not have the shortest delay. Generally only signal instances that are within 20 dB of the largest signal instance must be considered and then the delay, also referred to as 'delay (20 dB)', that matters is the excess delay, τ_d, which is the difference between the earliest and latest arriving signal instances within that 20 dB window. It is essential that a communications link be maintained a certain percentage of the time, e.g. 99%, for a specified level of service. This minimizes the number of times packets must be re-transmitted and assures acceptable operation of a system. Thus the maximum excess delay, $\tau_{d,\max}$ becomes an important metric and determines the interval for which the received signal is not valid, the required guard band to prevent signals from an earlier symbol interval overlapping with the current symbol interval, and also the maximum useful bandwidth of a signal.

It is found that the delay spread increases with distance between the transmitter and receiver, the Tx-RX distance, increases with frequency, and increases with a richer multipath environment. For example, a rural area has few paths but large cell sizes are used and the Tx-Rx distance is relatively large resulting in a large excess delay but a dense urban area has small cells but a rich multipath environment with many reflections from buildings. Measured excess delays for various situations are given in Table 4-2. There is a lot of variation in excess delays but it is clearly reduced with smaller cells and a less rich multipath environment.

It is also found that as a mobile unit moves the excess delay and the number of significant paths varies tens or hundreds of times per second depending on the environment but also the mobility, e.g. pedestrian versus highway speeds. However over a millisecond the propagation is found to be effectively fixed and consequently systems are designed with data sent in packets that are usually no longer than a millisecond.

Examining Table 4-2 further it is observed that the 2G and 3G cellular bands are below 2 GHz and the delay spreads can be several milliseconds and even larger with larger Tx-Rx distances in an urban area. At the higher microwave frequencies up to 6 GHz covering the operating frequency range of 4G, the delay spread is smaller.

The 28 GHz and 38 GHz bands are where 5G operates with short Tx-Rx distances. Here the number of paths above 20 dB is around 4 to 8 for Tx-Rx separations of 35 to 193 m [27, 28].

The spreads are relatively low and this is largely a result of using high gain steerable antennas which are essential to 5G operation. The high gain antennas reduce the spread of multipath path lengths, and hence reduce the excess delay spread, concentrating the reflected paths close to the transmit and receive antennas.

The prior discussions on multipath effects and the RF link consider a single frequency tone propagating from a transmit antenna and a receive antenna. When a single-tone signal travels on multiple paths there will be destructive and/or constructive interference.

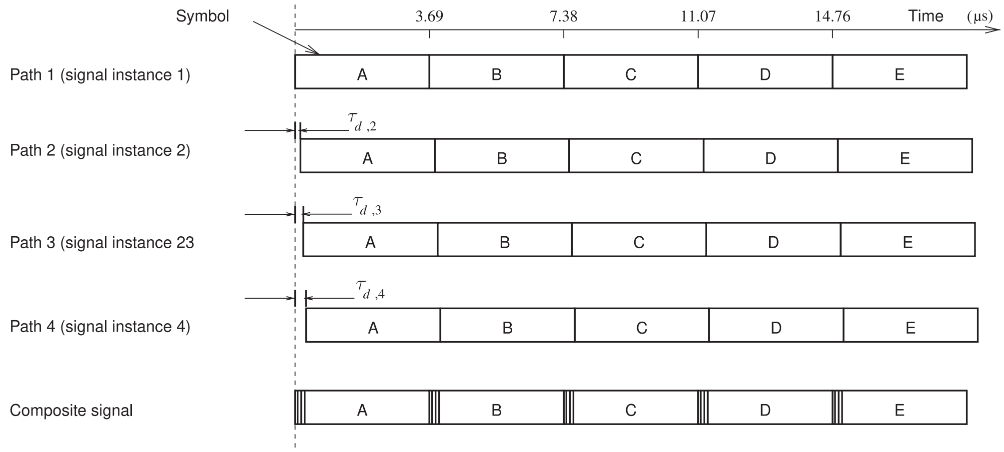
4.7.2 Intersymbol Interference

Intersymbol interference (**ISI**) occurs when a symbol traveling on one path interferes with the symbol traveling on another path. This is the result of the paths having different delays. The effect of excess delay on the received composite signal is indicated in Figure 4-23. Consider the situation where there are four paths with the excess delay of path 1 being 0 by definition, i.e. $\tau_{d,1} = 0$, the excess delay of path 2, $\tau_{d,2} = 100$ ns, and so on, $\tau_{d,3} = 200$ ns., and $\tau_{d,4} = 300$ ns. The 2G–5G cellular systems transmit a series of symbols, here A, B, C, D, and E, which have a symbol duration which differs by standard. In Figure 4-23 four paths are considered and the four signal instances each travel on a different path and are combined in the receiver's antenna to yield a composite received signal. The symbols of each of the signal instances overlap causing ISI.

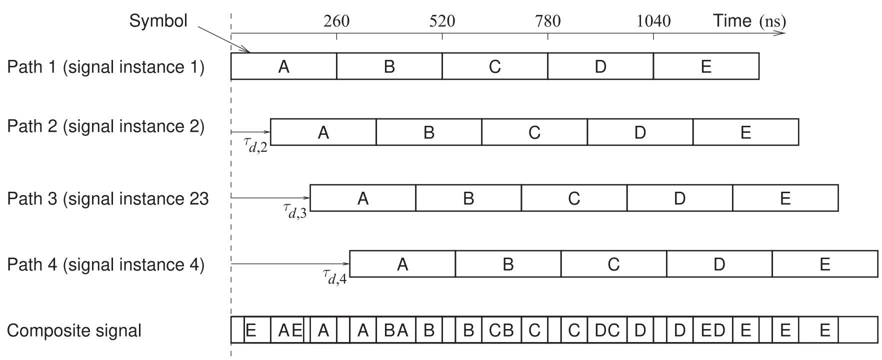
With the 2G GSM system symbol lengths are 3690 ns and the first 300 ns of a symbol instance traveling on the first path to arrive has interference from the previous symbol traveling on the other three paths, see Figure 4-23(a). In this case the first 300 ns of the 3690 ns of the symbol is garbled. In this example the delay spread is just 300 ns but this is conservative. Table 4-2 indicates that in some environments that the excess delay spread 1% of the time can be as long as the symbol interval for 2G/GSM (operating below 2 GHz). So even if the signal strength is high reception may not be possible because of ISI.

With 3G the packet length is 260 ns so with an excess delay spread of 300 ns for the four path example the symbol interference is severe, see Figure 4-23(b), and as noted 300 ns is a very conservative estimate of the excess delay spread. One of the features of 3G is a method for separating out the individual signal instances. Only when there are a very large number of significant paths does 3G have problems. In an urban area sometimes a communication link cannot be established with 2G and 3G even though the signal strength is high.

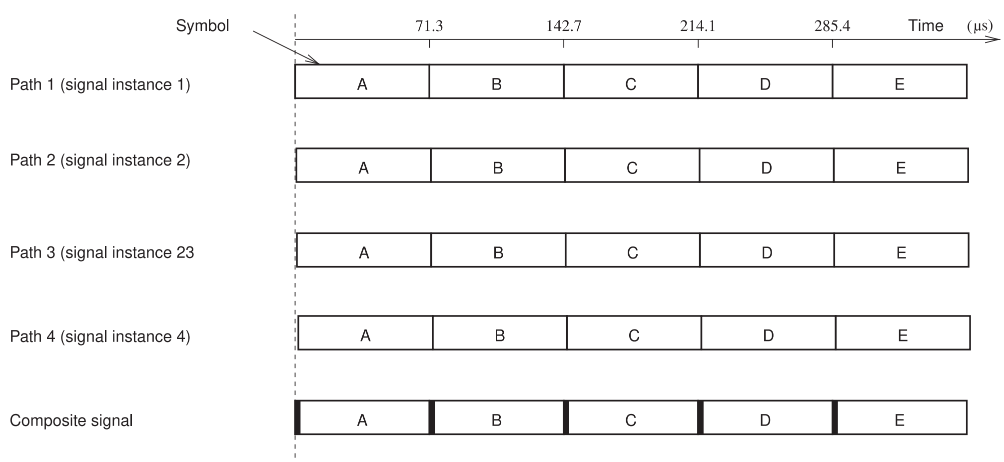
The 4G system minimizes the problem of ISI by having a very long symbol interval of 71.35 μs and a 300 ns excess delay spread is a small fraction of the symbol length. The symbol interval is considerably longer than any of the maximum delay spreads given in Table 4-2. Also the 4G standard



(a) Symbols in 2G GSM



(b) Symbols in 3G WCDMA



(c) Symbols in 4G

Figure 4-23: The effect of excess delay on symbol interference when there are four paths. The excess delay of each path, $\tau_{d,x}$, increases by 100 ns. Five symbols are shown: A, B, C, D, and E. The 2G GSM standard has a symbol duration of 3.6928 μs, the 3G WCDMA standard has a symbol duration of 260 ns, and the 4G OFDM standard has a symbol duration of 71.35 μs. The situation with 5G is similar to that for 4G.

has a guard time band known as a cyclic prefix. (The cyclic prefix is a bit more than a guard band as it involves repeating the end of symbol but this extra feature will be discussed latter.) The cyclic prefix can be either 4.7 μs, 5.2 μs, 16.7 μs, or 33.3 μs accommodating various excess delay spreads. This is more than enough to avoid the problem of ISI as any of the maximum excess delays listed in Table 4-2 are less than the maximum cyclic prefix length. The long symbol interval implies a very narrow bandwidth or sub-channel bandwidth. A 4G system communicates with a user using a very large number of sub-channels thus supporting high data rates.

4.7.3 Summary

Various excess delays on different paths causes intersymbol interference. High levels of intersymbol interference results in the failure to establish a communication link. Excess delay is a fundamental limitation with the 2G/GSM system and the only solution is to use very small cells in urban environments which have many significant signal paths. The 3G system employs a coding technique that enables the first few different paths to be resolved thus limiting the problem of ISI but not eliminating it. The 4G system, and 5G operates similarly, employs a long guard time, the cyclic prefix, to avoid ISI completely.

4.8 Radio Link Interference

The two major source of interference in a cellular system are self-interference resulting from excess delay spreads as discussed in the previous section, and interference, called radio link interference, from radios in other cells operating on the same frequency channel. This section discusses radio link interference and how it is controlled by using a frequency reuse plan.

4.8.1 Frequency Reuse Plan

Radio systems using the same channel are geographically spaced to control interference. In a cellular system, the coverage areas are arranged in cells represented by the hexagons in Figure 4-24. The actual shape of the cells is influenced by obstructions such as hills and buildings. In 1G and 2G cellular systems the cells are arranged in clusters such as the 3-, 7-, and 12-cell clusters shown, and the total number of channels available is divided among the cells in a cluster with the full set of frequency channels repeated in each cluster. (More about 3G–5G arrangements later.) The size of the clusters is the major component of what is called the **frequency reuse** plan. So a three-cell cluster has a spacing of approximately one cell diameter to the next cell using the same frequency channels. The signal level transmitted from the original cell will interfere with the signal in its corresponding cell in adjacent clusters. The level of interference is reduced with 7-cell and then 12-cell clusters. There is also background noise coming from cosmic sources as well as artificial sources, but in a cellular system, interference from other radios operating in the same system nearly always dominates.

The use of directional antennas at the base station increases the SIR. The interference pattern obtained using a **trisector antenna** (each segment of the antenna providing 120° of coverage) is shown in Figure 4-25. The trisector antenna can be arranged so that the transmit and receive antennas are

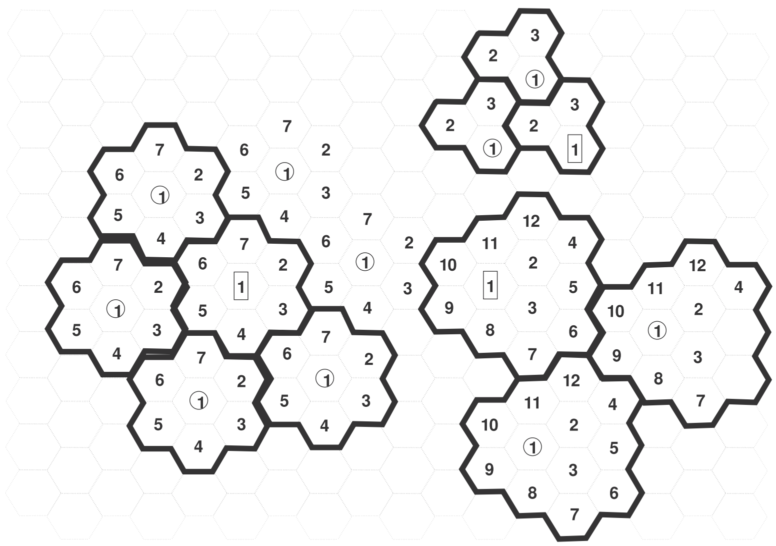


Figure 4-24: Cells arranged in clusters. Shown are 3-cell clusters (top right), 7-cell clusters (bottom left), and 12-cell clusters (bottom right).

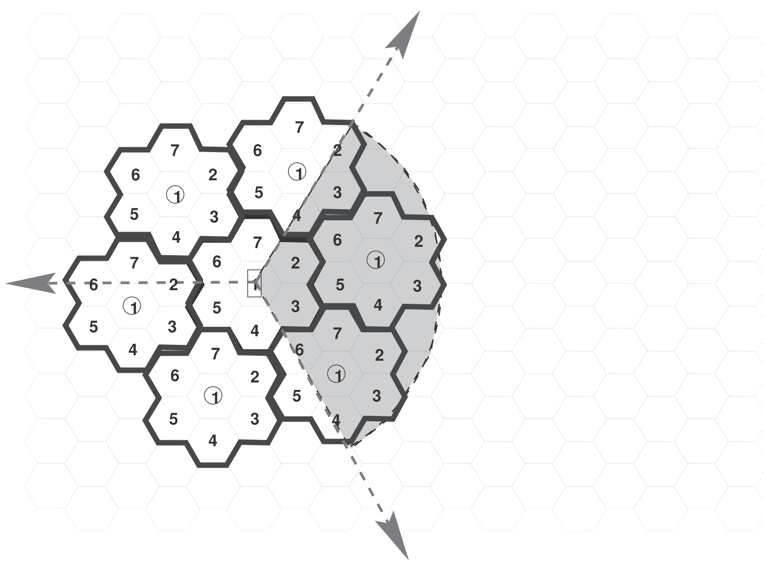


Figure 4-25: Cells in a cellular radio system with a trisector antenna, such as in Figure 4-26, showing the area of possible interference as a shaded region.

separated (see Figure 4-26), and can also be arranged to tilt the coverage toward the ground (e.g., see the antenna in Figure 4-2(l)). It is also possible to use higher-order antenna sectoring and to have smaller cells (achieved possibly by relatively low-power transmissions) to provide higher levels of coverage at critical regions such as intersections of roads. These smaller cells are referred to as microcells or picocells.

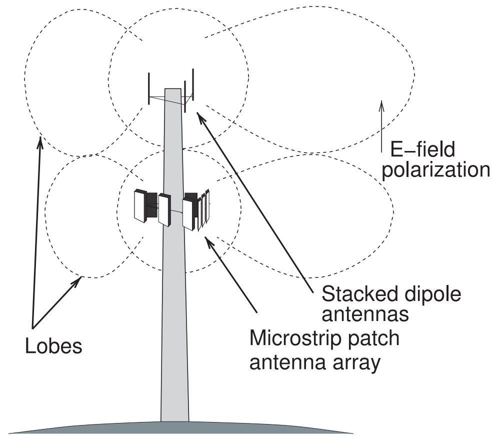


Figure 4-26: Base station with trisector stacked-dipole transmit and microstrip patch receive antennas.

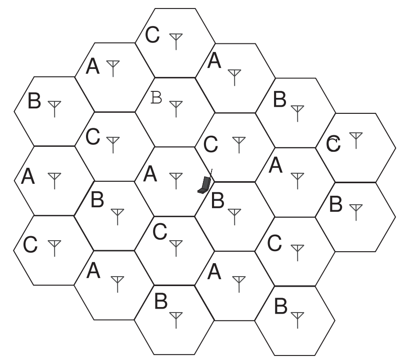
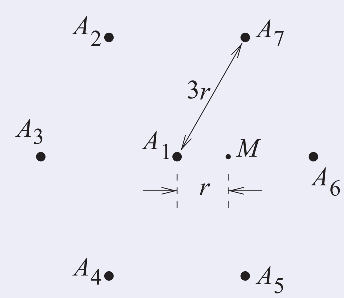


Figure 4-27: Three-cell cluster with mobile unit at the edge of cells A, B, and C.

EXAMPLE 4.9 Cellular Interference

In a cellular system, a signal is intentionally transmitted from a base station nominally located in the center of a cell to a mobile unit in the same cell. However, nearby transmitters using the same channel cause interference. In Figure 4-27, a mobile unit is located at the edge of a cell and uses frequency channel A. Many nearby transmitters also operate using channel A and the six nearest transmitters can be considered as causing significant interference. Consider that the mobile unit is a distance r from its cell's transmitter along the line connecting two channel A base stations, that the transmitters all operate at the same power level, and that the distance between base stations operating using channel A is $3r$. This three-cell cluster operates in a suburban area and the power density drops off with distance d as $1/d^3$ due to multipath effects. What is the SIR at the mobile unit?

Solution:

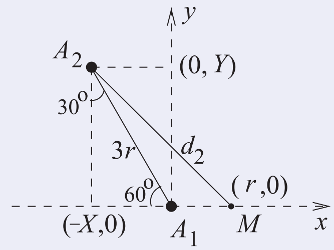


There are seven close towers transmitting signals on the same channel. Call these A_1–A_7 (M = mobile unit, A = transmitters). A_1 is the desired signal and A_2–A_7 are interferers. The distances from the transmitters to the mobile unit are d_1–d_7, and the powers from the transmitters received at the mobile unit are P_1–P_7.

Now SIR $= \dfrac{P_1}{P_2 + P_3 + P_4 + P_5 + P_6 + P_7}$.

This problem requires d_2–d_7 to be determined.

Consider the diagram to the right with $d_1 = r$. Also, the distance from A_2 to A_7 is $3r$. Now $d_3 = 3r + r = 4r$, $d_6 = 3r - r = 2r$. So $d_2 = \left[(X+r)^2 + Y^2\right]^{\frac{1}{2}} = \left(2.5^2 + 2.6^2\right) r = 3.607r = d_4$. ($X = 3r \sin 30^\circ = 1.5r$ and $Y = 3r \cos 30^\circ = 2.6r$.) Similarly,



$d_7 = d_5 = \left[(1.5-1)^2 + 2.6^2\right]^{\frac{1}{2}} = 2.648r.$

$$\frac{P_2}{P_1} = \frac{d_1}{d_2} = \frac{1}{3.607^3} = 0.0213 = \frac{P_4}{P_1} \qquad \frac{P_3}{P_1} = \frac{d_1}{d_3} = \frac{1}{4^3} = 0.0156.$$

$$\frac{P_7}{P_1} = \frac{d_1}{d_7} = \frac{1}{2.648^3} = 0.0539 = \frac{P_5}{P_1} \qquad \frac{P_6}{P_1} = \frac{d_1}{d_6} = \frac{1}{2^3} = 0.1250.$$

$$\text{SIR} = (0.0213 + 0.0156 + 0.0213 + 0.0539 + 0.1250 + 0.0539)^{-1} = 3.44 = 5.36 \text{ dB}.$$

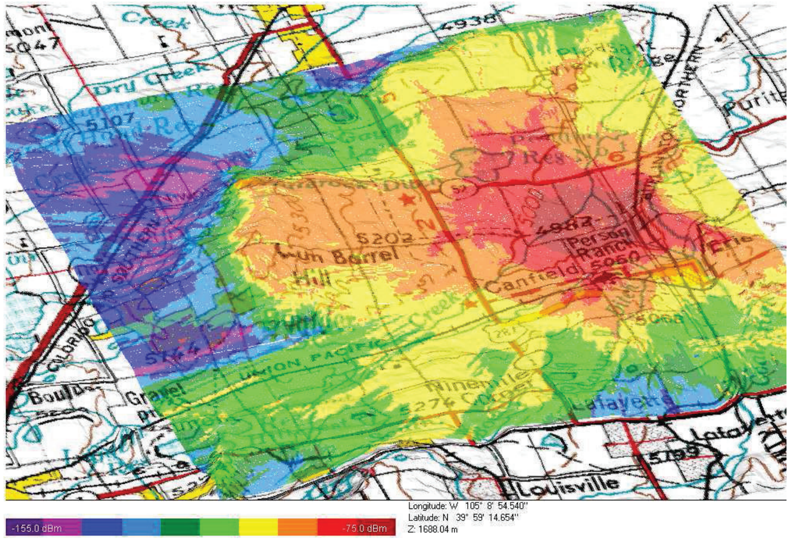
4.8.2 Summary

The clustering of cells as described above applies most directly to 1G and 2G cellular systems. The 3G system does not use clusters of cells but instead uses high frequency codes to separate users. There is a system awareness of managing interference but more complicated than the clustering arrangements. The 4G and 5G systems take frequency reuse algorithms to a whole new level performing global minimization of interference of wide areas. Still there is the basic concept that radios in adjacent cells operating on the same frequency channel can interfere with each other and careful attention must be given to frequency reuse.

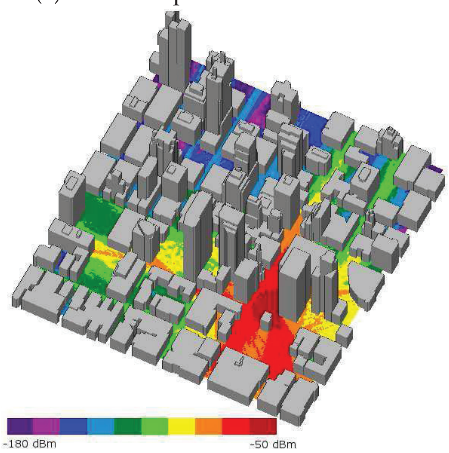
The region of signal coverage by a basestation is greatly influenced by terrain and is not a neat hexagon. See for example the coverage map for a rural area shown in Figure 4-28(a). To ensure adequate coverage basestations must be placed on a terrain-aware grid and not a regular grid. A wasted expense in operating a cellular system would be installing more basestations than are needed for optimum coverage. Thus cellular service providers simulate coverage before installing basestations. A significant characteristic of urban coverage as shown in Figures 4-28(b and c) is the "**urban canyon effect**" whereby coverage is concentrated along roads with signals bouncing off buildings directing signal propagation down the roadway. Even areas close to a transmitter but behind buildings are shadowed while distant areas near roads can have good coverage. This effect is also known as the urban **wave-guiding effect** or **wave-channeling effect**.. The signal blocking by buildings is shown in Figure 4-28(e) where it is seen that there is some diffraction over buildings but shadowing is severe after two or more diffraction events. The indoor environment also affects coverage as shown in Figure 4-28(d) for WiFi.

4.9 Antenna array

An antenna array comprises multiple radiating elements, i.e. individual antennas, and focuses a transmit beam in a desired direction. In Figure 4-29 the field pattern in the plane of the earth produced by an array of 30 antenna elements arranged horizontally is shown. The fields from each

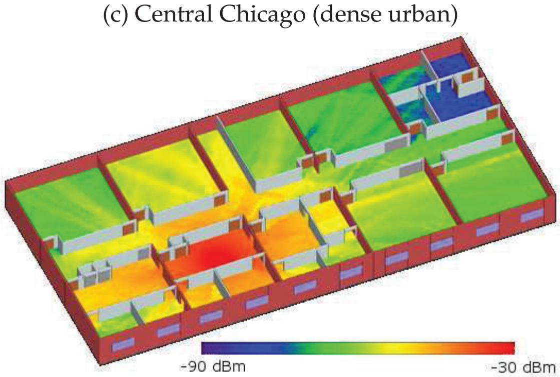


(a) Colorado plain east of Boulder and northwest of Denver (rural, gently rolling terrain).

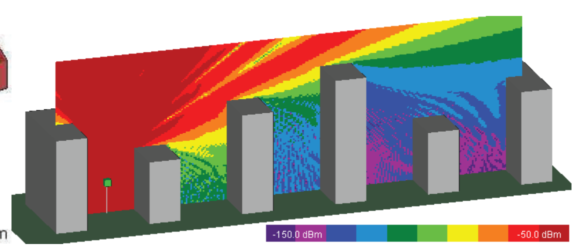


(c) Central Chicago (dense urban)

(c) Rosslyn Virgina (urban)



(d) Indoor coverage



(e) Verticle profile with low transmit antenna

Figure 4-28: Radio coverage profiles with 900 MHz transmitter except for (d) which is for 2.45 GHz. Calculated using Wireless Insite®. Copyright Remcom, Inc. Used with permission.

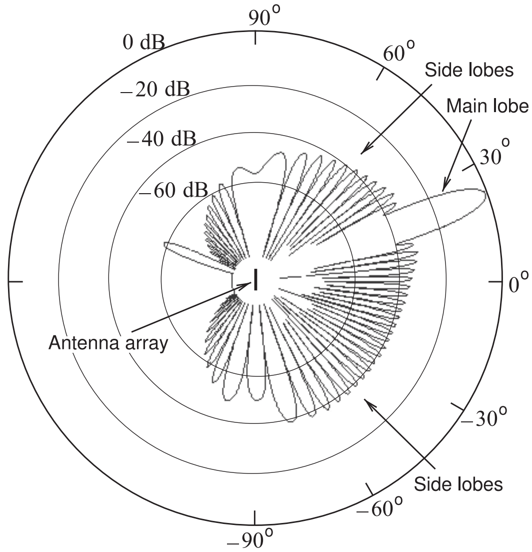


Figure 4-29: Electric field pattern from a 30 element array of antennas spaced 0.65λ appart. The sidelobe levels are about 40 dB below the power level of the main lobe. The same signal is presented to the antenna elements except that phases of the signal at each antennas is adjusted to produce a main beam directed at 20 degrees. The signals to each antenna are thus correlated. After [29].

antenna element combined to narrow and strengthen the main beam. Side (or grating) lobes are produced and these will result in some interference but their level here is 40 dB below that of the main beam. This is another way of managing interference in a cellular system but the direction to the mobile unit must be known. Antenna arrays are used in 5G. The affect of the array is to increase the power density of the main beam and the density relative to that of an isotropic antenna is called the directional gain of the array, D_{Array}, and is the product of the antenna gain, G_A, of an individual antenna element and the array gain, G_{Array}:

$$D_{\text{Array}} = G_{\text{Array}} G_A. \tag{4.49}$$

The maximum value of G_{Array} is N for an N element array. So the maximum directional gain of the array in dBi is

$$D_{\text{Array}}|_{\text{dBi}} = G_A|_{\text{dBi}} + 10 \log N. \tag{4.50}$$

4.10 Summary

This chapter discussed the impacts on communication integrity of imperfections in the RF link from the output of the transmitter to the input of the receiver. Prior to cellular communications becoming so important, only the LOS communication path was considered and the impact of fading, multiple reflections, and delay spread was regarded as bothersome. Most commonly the solution to the problems introduced by these effects was either to shift to another frequency, to increase the operating power, or install more basestations. With the advent of cellular communications, and digital communications in general, techniques have been developed to overcome the negative impacts of fading without boosting signal levels to the level where they severely impact the operation of other radios. So while many aspects of RF propagation are random, concepts and statistical models have been developed that enable design choices to be made that permit digital communication systems to operate in what would have once been regarded as a hostile environment.

4.11 References

[1] S. Ramo, J. Whinnery, and T. Van Duzer, *Fields and Waves in Communication Electronics*. John Wiley & Sons, 1965.

[2] C. A. Balanis, "Antenna theory: A review," *Proc. IEEE*, vol. 80, no. 1, pp. 7–23, 1992.

[3] C. Balanis, *Antenna Theory: Analysis and Design*. John Wiley & Sons, 2005.

[4] T. L. Wilson, K. Rohlfs, and S. Hüttemeister, *Tools of radio astronomy*, 6th ed. Springer, 2013.

[5] J. L. Pawsey and R. N. Bracewell, "Radio astronomy," *Oxford, Clarendon Press*, 1955.

[6] J. Doble, *Introduction to Radio Propagation for Fixed and Mobile Communications*. Norwood, MA, USA: Artech House, Inc., 1996.

[7] M. Born and E. Wolf, "Two-dimensional diffraction of a plane wave by a half-plane,," in *Principles of Optics: Electromagnetic Theory of Propagation, Interference, and Diffraction of Light*, 7th ed. Cambridge University Press, 1999, section 11.5.

[8] A. Abdi, C. Tepedelenlioglu, M. Kaveh, and G. Giannakis, "On the estimation of the k parameter for the rice fading distribution," *IEEE Communications Letters*, vol. 5, no. 3, pp. 92–94, Mar. 2001.

[9] A. Saleh and R. Valenzuela, "A statistical model for indoor multipath propagation," *IEEE J. on Selected Areas in Communications*, vol. 5, no. 2, pp. 128–137, Feb. 1987.

[10] D. Devasirvatham, "Time delay spread measurements of wideband radio signals within a building," *Electronics Letters*, vol. 20, no. 23, pp. 950–951, 8 1984.

[11] A. Acampora and J. Winters, "System applications for wireless indoor communications," *IEEE Communications Magazine*, vol. 25, no. 8, pp. 11–20, Aug. 1987.

[12] A. Siamarou and M. Al-Nuaimi, "Multipath delay spread and signal level measurements for indoor wireless radio channels at 62.4 ghz," in *2001 IEEE VTS 53rd Vehicular Technology Conf.*, 2001, pp. 454–458.

[13] R. Gibson and R. Murray, "Systems organization for multichannel cordless telephones," in *Int. Zurich Seminar on Digital Communicatio*, 1982, pp. 59–61.

[14] J. Andersen, T. Rappaport, and S. Yoshida, "Propagation measurements and models for wireless communications channels," *IEEE Communications Magazine*, vol. 33, no. 1, pp. 42–49, Jan. 1995.

[15] T. Sarkar, Z. Ji, K. Kim, A. Medouri, and M. Salazar-Palma, "A survey of various propagation models for mobile communication," *IEEE Antennas and Propagation Magazine*, vol. 45, no. 3, pp. 51–82, Jun. 2003.

[16] M. Murotani and K. Kohriyama, "Representation by approximate numerical formulas of radio-relay antenna beam bending due to atmospheric refraction," *IEICE Trans.*, vol. E72-E, no. 2, pp. 101–103, Feb. 1989.

[17] J. Seybold, *Introduction to RF Propagation*. John Wiley & Sons, 2005.

[18] P. Papazian, "Basic transmission loss and delay spread measurements for frequencies between 430 and 5750 mhz," *IEEE transactions on antennas and propagation*, vol. 53, no. 2, pp. 694–701, 2005.

[19] J. Van Rees, "Measurements of the wideband radio channel characteristics for rural, residential, and suburban areas," *IEEE Transactions on Vehicular Technology*, vol. 36, no. 1, pp. 2–6, 1987.

[20] J. A. Wepman, J. R. Hoffman, and L. H. Loew, "Characterization of macrocellular pcs propagation channels in the 1850-1990 MHz band," in *Proceedings of 1994 3rd IEEE International Conference on Universal Personal Communications*, 1994, pp. 165–170.

[21] E. S. Sousa, V. M. Jovanovic, and C. Daigneault, "Delay spread measurements for the digital cellular channel in toronto," *IEEE Transactions on Vehicular Technology*, vol. 43, no. 4, pp. 837–847, 1994.

[22] D. Devasirvatham, "Radio propagation studies in a small city for universal portable communications," in *38th IEEE Vehicular Technology Conference*, 1988, pp. 100–104.

[23] D. Cox and R. Leck, "Distributions of multipath delay spread and average excess delay for 910-MHz urban mobile radio paths," *IEEE Transactions on Antennas and Propagation*, vol. 23, no. 2, pp. 206–213, 1975.

[24] T. S. Rappaport, S. Y. Seidel, and R. Singh, "900-MHz multipath propagation measurements for us digital cellular radiotelephone," *IEEE Transactions on Vehicular Technology*, vol. 39, no. 2, pp. 132–139, 1990.

[25] J.-P. De Weck, P. Merki, and R. W. Lorenz, "Power delay profiles measured in mountainous terrain (radiowave propagation)," in *38th IEEE Vehicular Technology Conference*, 1988, pp. 105–112.

[26] T. Tanaka, S. Kozono, and A. Akeyama, "Urban multipath propagation delay characteristics in mobile communications," *Electronics and Communications in Japan (Part I: Communications)*, vol. 74, no. 8, pp. 80–88, 1991.

[27] Y. Azar, G. N. Wong, K. Wang, R. Mayzus, J. K. Schulz, H. Zhao, F. Gutierrez Jr, D. Hwang, and T. S. Rappaport, "28 GHz

propagation measurements for outdoor cellular communications using steerable beam antennas in new york city." in *ICC*, 2013, pp. 5143–5147.

[28] T. S. Rappaport, Y. Qiao, J. I. Tamir, J. N. Murdock, and E. Ben-Dor, "Cellular broadband millimeter wave propagation and angle of arrival for adaptive beam steering systems," in *2012 IEEE Radio and Wireless Symposium*. IEEE, 2012, pp. 151–154.

[29] Phased array radiation pattern, Phased_array_radiation_pattern.gif, By Maxter315 [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)], from Wikimedia Commons.

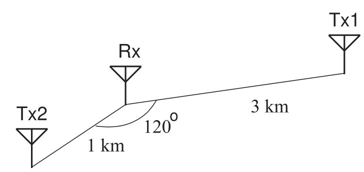
4.12 Exercises

1. An antenna only radiates 45% of the power input to it. The rest is lost as heat. What input power (in dBm) is required to radiate 30 dBm?

2. The output stage of an RF front end consists of an amplifier followed by a filter and then an antenna. The amplifier has a gain of 27 dB, the filter has a loss of 1.9 dB, and of the power input to the antenna, 35% is lost as heat due to resistive losses. If the power input to the amplifier is 30 dBm, calculate the following:
   (a) What is the power input to the amplifier in watts?
   (b) Express the loss of the antenna in dB.
   (c) What is the total gain of the RF front end (amplifier + filter)?
   (d) What is the total power radiated by the antenna in dBm?
   (e) What is the total power radiated by the antenna in mW?

3. The output stage of an RF front end consists of an amplifier followed by a filter and then an antenna. The amplifier has a gain of 27 dB, the filter has a loss of 1.9 dB, and of the power input to the antenna, 45% is lost as heat due to resistive losses. If the power input to the amplifier is 30 dBm, calculate the following:
   (a) What is the power input to the amplifier?
   (b) Express the loss of the antenna in decibels.
   (c) What is the total gain of the RF front end (amplifier + filter)?
   (d) What is the total power radiated by the antenna in dBm?
   (e) What is the total power radiated by the antenna in milliwatts?

4. In the figure below there are two transmitters, Tx_1 and Tx_2, operating at the same power level, and one receiver, Rx. Tx_2 is an intentional transmitter and its signal is intended to be received at Rx. Tx_1 uses the same frequency channel as Tx_2, but it transmits an interfering signal. [Parallels Example 4.1]



   (a) Assume that the antennas are omnidirectional and, being in a semiurban area, that the transmitted power density drops off as $1/d^3$, where d is the distance from the transmitter. Calculate the SIR at Rx. Express your answer in decibels.
   (b) Now consider a directional antenna at Rx while the transmit antennas remain omnidirectional. The antenna at Rx is directed toward the transmitter Tx_2 and the antenna gain is 6 dB. In the direction of Tx_1 the effective antenna gain is -3 dB. Now recalculate the SIR. Express your answer in dB.

5. Thirty five percent of the power input to an antenna is lost as heat, what is the loss of the antenna in dB.
6. Only 65% of the power input to an antenna is radiated with the rest lost to dissipation in the antenna, what is the gain of the antenna in dB? (This is not the antenna gain.)
7. The efficiency of an antenna is 66%. If the power input to the antenna is 10 W what is the power radiated by the antenna in dBm?
8. An antenna with an input of 1 W operates in free space and has an antenna gain of 12 dBi. What is the maximum power density at 100 m from the antenna?
9. A transmitter has an antenna with an antenna gain of 10 dBi, the resistive losses of the antenna are 50%, and the power input to the antenna is 1 W. What is the EIRP in watts?
10. A transmitter has an antenna with an antenna gain of 20 dBi, the resistive losses of the antenna are 50%, and the power input to the antenna is 100 mW. What is the EIRP in watts?
11. An antenna with an antenna gain of 8 dBi radiates 6.67 W. What is the EIRP in watts? Assume

that the antenna is 100% efficient.

12. An antenna has an antenna gain of 10 dBi and a 40 W input signal. What is the EIRP in watts?

13. An antenna with 5 W of input power has an antenna gain of 20 dBi and an antenna efficiency of 25% and all of the loss is due to resistive losses in the antenna. [Parallels Example 4.3]
    (a) How much power in dBm is lost as heat in the antenna?
    (b) How much power in dBm is radiated by the antenna?
    (c) What is the EIRP in dBW?

14. An antenna with an efficiency of 50% has an antenna gain of 12 dBi and radiates 100 W. What is the EIRP in watts?

15. An antenna with an efficiency of 75% and an antenna gain of 10 dBi. If the power input to the antenna is 100 W,
    (a) what is the total power in dBm radiated by the antenna?
    (b) what is the EIRP in dBm?

16. On a resonant antenna a large current is established by creating a standing wave. The current peaking that thus results establishes a strong electric field (and hence magnetic field) that radiates away from the antenna. A typical dipole loses 15% of the power input to it as resistive (I^2R) losses and has an antenna gain of 10 dBi measured at 50 m. Consider a base station dipole antenna that has 100 W input to it. Also consider that the transmitted power density falls off with distance d as $1/d^3$. Hint, calculate the power density at 50 m. [Parallels Example 4.2]
    (a) What is the input power in dBm?
    (b) What is the power transmitted in dBm?
    (c) What is the power density at 1 km? Express your answer as $\mathrm{W/m^2}$.
    (d) What is the power captured by a receive antenna (at 1 km) that has an effective antenna aperture of 6 cm^2? Express your answer in first dBm and then watts.
    (e) If the background noise level captured by the antenna is 1 pW, what is the SNR in decibels? Ignore interference that comes from other transmitters.

17. A communication system operating at 2.5 GHz includes a transmit antenna with an antenna gain of 12 dBi and a receive antenna with an effective aperture area of 20 cm^2. The distance between the two antennas is 100 m.
    (a) What is the antenna gain of the receive antenna?
    (b) If the input to the transmit antenna is 1 W, what is the power density at the receive antenna if the power falls off as $1/d^2$, where d is the distance from the transmit antenna?
    (c) Thus what is the power delivered at the output of the receive antenna?

18. Consider a point-to-point communication system. Parabolic antennas are mounted high on a mast so that ground effects do not exist, thus power falls off as $1/d^2$. The gain of the transmit antenna is 20 dBi and the gain of the receive antenna is 15 dBi. The distance between the antennas is 10 km. The effective area of the receive antenna is 3 cm^2. If the power input to the transmit antenna is 600 mW, what is the power delivered at the output of the receive antenna?

19. Consider a 28 GHz point-to-point communication system. Parabolic antennas are mounted high on a mast so that ground effects do not exist, thus power falls off as $1/d^2$. The gain of the transmit antenna is 20 dBi and the gain of the receive antenna is 15 dBi. The distance between the antennas is 10 km. If the power output from the receive antenna is 10 pW, what is the power input to the transmit antenna?

20. An antenna has an effective aperture area of 20 cm^2. What is the antenna gain of the antenna at 2.5 GHz?

21. An antenna operating at 28 GHz has an antenna gain of 50 dBi. What is the effective aperture area of the antenna?

22. A 15 GHz receive antenna has an antenna gain of 20 dBi. If the power density at the receive antenna is 1 $\mathrm{nW/cm^2}$, what is the power at the output of the antenna? [Parallels Example 4.7]

23. A microstrip patch antenna operating at 2 GHz has an efficiency of 66% and an antenna gain of 8 dBi. The power input to the antenna is 10 W.
    (a) What is the power, in dBm, radiated by the antenna?
    (b) What is the equivalent isotropic radiated power (EIRP) in watts?
    (c) What is the power density, in $\mu\mathrm{W/m^2}$, at 1 km if ground effects are ignored?
    (d) Because of multipath effects, the power density drops off as $1/d^4$, where d is distance. What is the power density, in $\mathrm{nW/m^2}$, at 1 km if the power density is 100 $\mathrm{mW/m^2}$ at 10 m from the transmit antenna?

24. A communication system operating at 10 GHz uses a microstrip patch antenna as a transmit antenna and a dipole antenna as a receive antenna. The transmit antenna is directly connected to the

transmitter and the output power of the transmitter is 30 W. The transmit antenna has an antenna gain of 9 dBi and an antenna efficiency of 60%. The receive antenna has an antenna gain of 2 dBi and a radiation efficiency of 80%. The receive antenna is connected to a receiver by a 10 m long cable with a loss of 0.1 dB/m. The link between the transmit and receive antenna is sufficiently elevated that ground effects and multipath effects are insignificant.

(a) What is the output power of the transmitter in dBm?
(b) What is the EIRP in dBm?
(c) The transmitted power will drop off as $1/d^n$ (d is distance). What is n?
(d) What is the peak power density in $\mu\text{W/m}^2$ at 1 km?
(e) What is the effective aperture size of the receive antenna in m^2?
(f) If the radiated power density at the receive antenna is 1 $\mu\text{W/m}^2$, what is the signal power at the output of the receive antenna in dBm?
(g) What is the total cable loss in dB?
(h) What is the power presented to the receiver in dBm?

25. A communication system operating at 10 GHz uses a microstrip patch antenna as a transmit antenna and a dipole antenna as a receive antenna. The transmit antenna is connected to the transmitter by a 20 m long cable with a loss of 0.2 dB/m and the output power of the transmitter is 30 W. The transmit antenna has an antenna gain of 9 dBi and an antenna efficiency of 60%. The link between the transmit and receive antenna is sufficiently elevated that ground effects and multipath effects are insignificant.
(a) What is the output power of the transmitter in dBm?
(b) What is the cable loss between the transmitter and the antenna?
(c) What is the total power radiated by the transmit antenna in dBm?
(d) What is the power lost in the antenna as resistive losses and spurious radiation? Express your answer in dBm.
(e) What is the EIRP of the transmitter in dBm?
(f) The transmitted power will drop off as $1/d^n$ (d is distance). What is n?
(g) What is the peak power density in $\mu\text{W/m}^2$ at 1 km?

26. Stacked dipole antennas are often found at the top of cellphone masts, particularly for large cells and operating frequencies below 1 GHz. These antennas have an efficiency that is close to 90%. Consider an antenna that has 40 W of input power, an antenna gain of 10 dBi, and transmits a signal at 900 MHz.
(a) What is the EIRP in watts?
(b) If the power density drops as $1/d^3$, where d is the distance from the transmit tower, what is the power density at 1 km if the power density is 100 mW/m^2 at 10 m?

27. Consider an 18 GHz point-to-point communication system. Parabolic antennas are mounted on masts and the LOS between the antennas is just above the tree line. As a result, power falls off as $1/d^3$, where d is the distance between the antennas. The gain of the transmit antenna is 20 dBi and the gain of the receive antenna is 15 dBi. The antennas are aligned so that they are in each other's main beam. The distance between the antennas is 1 km. The transmit antenna is driven by a power amplifier with an output power of 100 W. The amplifier drives a coaxial cable that is connected between the amplifier and the transmit antenna. The cable loses 75% of its power due to resistive losses. On the receive side, the receive antenna is directly connected to a masthead amplifier with a gain of 10 dB and then a short cable with a loss of 3 dB before entering the receive base station.
(a) Draw the signal path.
(b) What is the loss and gain of the transmitter coaxial cable in decibels?
(c) What percentage of the power input to the receive coaxial cable is lost in the receive cable?
(d) Express the power of the transmit amplifier in dBW and dBm.
(e) What is the propagation loss in decibels?
(f) Determine the total power in watts delivered to the receive base station.

28. Consider a point-to-point communication system. Parabolic antennas are mounted high on a mast so that ground effects are minimal. Thus power density falls off as $1/d^{2.3}$, where d is the distance from the transmitter. The gain of the transmit antenna is 15 dBi and the gain of the receive antenna is 12 dBi. These antenna gains are normalized to a distance of 1 m. The distance between the antennas is 15 km. The output power of the receive antenna must be 1 pW. The RF frequency is 2 GHz; treat the antennas as lossless.
(a) What is the received power in dBm?
(b) What is the path loss in decibels?
(c) What is the link loss in decibels?
(d) Using the link loss, calculate the input power, P_T, of the transmitter. Express the answer in dBm.

(e) What is the aperture area of the receiver in square meters?
(f) Determine the radiated power density at the receiver in terms of the transmitter input power. That is, if P_T is the power input to the transmit antenna, determine the power density, P_D, at the receive antenna where $P_D = xP_T$. What is x in units of m^{-2}?
(g) Using the power density calculation and the aperture area, calculate P_T in watts.
(h) What is P_T in dBm? This should be the same as the answer you calculated in (d).
(i) What is the total power radiated by the transmit antenna in dBm?

29. Two identical antennas are used in a point-to-point communication system, each having a gain of 50 dBi. The system has an operating frequency of 28 GHz and the antennas are at the top of masts 100 m tall. The RF link between the antennas consists only of the direct line-of-sight path.
(a) What is the effective aperture area of each antenna?
(b) How does the power density of the propagating signal rolloff with distance.
(c) If the separation of the transmit and receive antennas is 10 km, what is the path loss in decibels?
(d) If the separation of the transmit and receive antennas is 10 km, what is the link loss in decibels?

30. A transmitter and receiver operating at 2 GHz are at the same level, but the direct path between them is blocked by a building and the signal must diffract over the building for a communication link to be established. This is a classic knife-edge diffraction situation. The transmit and receive antennas are each separated from the building by 4 km and the building is 20 m higher than the antennas (which are at the same height). Consider that the building is very thin. It has been found that the path loss can be determined by considering loss due to free-space propagation and loss due to diffraction over the knife edge.
(a) What is the additional attenuation (in decibels) due to diffraction?
(b) If the operating frequency is 100 MHz, what is the attenuation (in decibels) due to diffraction?
(c) If the operating frequency is 10 GHz, what is the attenuation (in decibels) due to diffraction?

31. A hill is 1 km from a transmit antenna and 2 km from a receive antenna. The receive and transmit antennas are at the same height and the hill is 20 m above the height of the antennas. What is the additional loss caused by diffraction over the top of the hill? Treat the hill as a knife-edge and the operating frequency is 1 GHz.

32. A 1 GHz point-to-point link has two major transmission paths. One is a LOS path and the other includes reflection from the ground so that the power density of the transmitted signal rollsoff as $1/d^{2.5}$ where d is the distance from the transmit antenna. At 10 m the power density from the transmit antenna is 100 $\mathrm{mW/m^2}$.
(a) What is the power density at 1 km.
(b) If the receive antenna has an antenna gain of 30 dBi, what is the effective aperture area of the receive antenna?
(c) What is the power of the signal at the output of the receive antenna?

33. Two identical antennas are used in a point-to-point communication system, each having a gain of 30 dBi. The system has an operating frequency of 14 GHz and the antennas are at the top of masts 100 m tall. The RF link between the antennas consists only of the direct LOS path.
(a) What is the effective aperture area of each antenna?
(b) How does the power density of the propagating signal rolloff with distance?
(c) If the separation of the transmit and receive antennas is 10 km, what is the path loss? Ignore atmospheric loss.

34. The three main cellular communication bands are centered around 450 MHz, 900 MHz, and 2 GHz. Compare these three bands in terms of multipath effects, diffraction around buildings, object (such as a wall) penetration, scattering from trees and parts of trees, and ability to follow the curvature of hills. Complete the table below with the relative attributes: high, medium, and low.

Characteristic	450 MHz	900 MHz	2 GHz
Multipath			
Scattering			
Penetration			
Following curvature			
Range			
Antenna size			
Atmospheric loss			

35. Describe the difference in multipath effects in a central city area compared to multipath effects

in a desert. Your description should be approximately 4 lines long and not use a diagram

36. Wireless LAN systems can operate at 2.4 GHz, 5.6 GHz, 40 GHz and 60 GHz. Contrast with explanation the performance of these schemes inside a building in terms of range.

37. At 60 GHz the atmosphere strongly attenuates a signal. Discuss the origin of this and indicate an advantage and a disadvantage.

38. Short answer questions. Each part requires a short paragraph of about five lines and a figure, where appropriate, to illustrate your understanding.
    (a) Cellular communications systems use two frequency bands to communicate between the basestation and the mobile unit. The bands are generally separated by 50 MHz or so. Which band (higher or lower) is used for the downlink from the basestation to the mobile unit and what are the reasons behind this choice?
    (b) Describe at least two types of interference in a cellular system from the perspective of a mobile handset.

39. The three main cellular communication bands are centered around 450 MHz, 900 MHz, and 2 GHz. Compare these three bands in terms of multipath effects, diffraction around buildings, object (such as a wall) penetration, scattering from trees and parts of trees, and the ability to follow the curvature of hills. Use a table and indicate the relative attributes: high, medium, and low. WAS 10(C)

40. Describe Rayleigh fading in approximately 4 lines and without using a diagram.

41. In several sentences and using a diagram describe Rayleigh fading and the impact it has on radio communications.

42. A transmit antenna and a receive antenna are separated by 1 km and operate at 1 GHz. What is the radius of the first Fresnel zone at 0.5 km from each antenna? [Parallels Example 4.8]

43. A transmit antenna and a receive antenna are separated by 40 km and operate at 10 GHz. [Parallels Example 4.8]
    (a) What is the radius of the first Fresnel zone at the midpoint between the antennas?
    (b) What is the radius of the second Fresnel zone?
    (c) To ensure LOS propagation, what should the clearance be from the direct line between the antennas and obstructions such as hills?

44. A transmitter and receiver operate at 100 MHz, are at the same level, and are separated by 4 km. The signal must diffract over a building half way between the antennas that is 20 m higher than the direct path between the antennas. What is the attenuation (in decibels) due to diffraction?

45. A transmitter and receiver operate at 10 GHz, are at the same level, and are 4 km apart. The signal must diffract over a building that is half way between the antennas and is 20 m higher than the line between the antennas. What is the attenuation (in dB) due to diffraction?

46. The path from a transmit antenna to a receive antenna is elevated so that ground and multipath effects are insignificant. The power radiated by the transmit antenna drop off as $1/d^n$ where d is distance, what is n?

47. A communication system has a power density roll-off of $1/d^{2.5}$ between a transmit antenna and a mobile receive antenna which are separated by 10 km. At 10 m from the transmit antenna, the power density is $10\ \mathrm{W/m^2}$. What is the power density at the receive antenna? [Parallels Example 4.5]

48. A 900 MHz communication system uses a transmit antenna with an antenna gain G_T of 3 dB and a receive antenna with an antenna gain G_R of 0 dB. If the distance between the antennas is 200 m, what is the link loss from the input to the transmit antenna and the output of the receive antenna if the power density reduces as $1/d^{2.5}$? [Parallels Example 4.6]

49. A transmitter has a power density of $100\ \mathrm{mW/m^2}$ at a distance of 1 m from the transmitter. The power density falls off as 33 dB per decade of distance. At what distance from the transmit antenna will the power density reach $1\ \mu\mathrm{W{\cdot}m^{-2}}$? [Parallels Example 4.7]

50. Describe at least two types of interference in a cellular system from the perspective of a mobile handset.

51. In a cellular system, a signal is intentionally transmitted from a base station nominally located in the center of a cell to a mobile unit in the same cell. However, nearby transmitters using the same channel cause interference. In Figure 4-27, a mobile unit is located at the edge of a cell and uses frequency channel A. Many nearby transmitters also operate using channel A and the six nearest transmitters can be considered as causing significant interference. Consider that the mobile unit is a distance r from its cell's transmitter along the line connecting two

channel A base stations, that the transmitters all operate at the same power level, and that the distance between base stations operating using channel A is $3r$. This three-cell cluster operates in a suburban area and the power density drops off with distance d as $1/d^{3.5}$ due to multipath effects. What is the SIR at the mobile unit? Express your answer in decibels. [Parallels Example 4.9]

52. A cellular system uses a three-cell cluster.
    (a) Treating cells as equal sized hexagons with towers in the center of each cell, draw the cell map including all cells within 3 cell diameters of the main channel. Label the main cell and other cells using the same frequencies as A.
    (b) Consider a mobile unit at the edge of a cell transmitting the intended signal from the tower at the center of that cell. Identify the interfering towers.
    (c) If ground effects and multipath effects are negligible, what is the power roll-off factor if the distance between a tower and the mobile unit is d? That is, what is n if power falls off as $1/d^n$?
    (d) If trisector antennas are used, identify the interfering cells and approximately determine the improvement in SIR compared to using a nonsectorized antenna? You do not need to do detailed calculations.

53. Describe trisector antennas in 4 lines and without using a diagram.

4.12.1 Exercises By Section

†challenging, ‡very challenging

§14.3	$1^\dagger$
§14.5	$2^\dagger, 3^\dagger, 4^\ddagger, 5, 6, 7^\dagger, 8^\dagger, 9^\dagger, 10, 11, 12, 13, 14, 15, 16^\ddagger, 17, 18, 19, 22$
§14.6	$23^\ddagger, 24^\ddagger, 25^\ddagger, 26^\ddagger, 27^\ddagger, 28^\ddagger, 29^\dagger, 30^\dagger, 31^\dagger, 32^\ddagger, 33^\dagger, 34^\ddagger, 35^\dagger, 36, 37, 38^\dagger, 39, 40, 41, 42^\dagger, 43, 44, 45, 46, 47, 48, 49$
§14.8	$50^\dagger, 51^\ddagger, 52^\ddagger, 53^\dagger$

4.12.2 Answers to Selected Exercises

2(e)	53.23 dBm
23(b)	63.1 W
17	0.251 μW
18	14.3 pW
25(e)	49.77 dBm
27(f)	708 pW
28	107.65 dB
32	7.16 m^2
42	8.66 m
51	7.33 dB

CHAPTER 5

RF Systems

5.1 Introduction

In the history of wireless communication, radar, and sensor systems there have been many standards and different types of systems. Following the development of the main cellular radio systems is a good proxy for the evolution of nearly all RF systems. Mobile radio up to and including 1G cellular radio was essentially analog with only simple signaling using tones or FSK modulation. There were many incompatible mobile radio and 1G systems as little thought was given to worldwide interoperability and radio companies were content with proprietary standards.

The situation continued with 2G cellular radio as there were many incompatible 2G systems with each only able to support one modulation method. Most of the radio functionality was in the analog hardware but VLSI was starting to become mature and it was possible to do limited error correction coding. At the basestation there was enough computing power to dynamically manage interference and implement simple system-level optimization. In the evolution to 3G there were many proposals all exploiting the increasing capability of VLSI. With the introduction of 3G, communication providers enforced the adoption of a single word-wide standard. Perhaps 'single' is a stretch as there were still several variations of the 3G implementation but a core set of standards supporting worldwide connectivity was adopted.

The path forward was becoming unmanageable so following the introduction of 3G an international consortium focused on developing a strategy to support cellular communications that was more capable and upwards compatible.

For the first time, with 4G a large number of modulation methods were supported with most of the modulation and demodulation functionality implemented in a DSP unit called the baseband processor. Changing modulation and demodulation formats was a simple procedure of running different code. This is software-defined radio with analog circuitry implementing just the translation from an analog baseband signal, really a modulated signal on a low-frequency intermediate frequency carrier, to the high-frequency RF signal. Modulation by the baseband data produced the baseband analog signal with the baseband processor implementing a digital version of an analog modulation format. With 5G additional capability is provided with the 4G standard becoming a subset of the 5G standard.

An understanding of systems is required to understand the specifications for RF hardware. This is particularly important as the actual performance required of hardware is often not directly related to the specifications developed by system designers. For example, one of the most important characteristics of digital radio systems is the bit error rate (BER). The BER is a quantity that cannot be determined until the components of a system are integrated. Thus in the design of subsystems, indirect measures such as intermodulation distortion (IMD) (referring to the generation of spurious signals when discrete tones are applied to a subsystem) are specified. The relationship between IMD and BER is weak. Clearly higher IMD tends to indicate a higher BER for the same technology, but the relative performance of different technologies cannot be evaluated this way. Thus an essential system design problem is developing sufficient and tractable criteria that enable subsystems to be locally designed and optimized, leading to an overall optimized system.

5.2 Broadcast, Simplex, Duplex, Diplex, and Multiplex Operations

One would think that defining these terms would be a simple matter. However, there have been different conventions in various segments of the telecommunications industry. Now that telecommunications are converging, universal definitions that accommodate all earlier uses are not possible. There is now standardization by the International Telecommunications Union (ITU) of the terms broadcast, simplex, duplex, diplex, and multiplex, as they relate to wireless communications [1, 2] (see Table 5-1). National Standards, however, do not need to conform to the ITU definitions. Examples of slightly different definitions come from the Telecom Glossary established as part of an American National Standard (ANS) [3]. There is greater specificity to radio in the ITU usage of the terms. In general the ITU definition should be used if the context is radio communications, but be careful if the interpretation of the terms is critical.

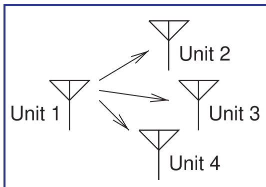
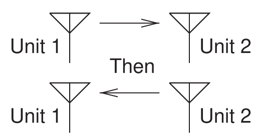
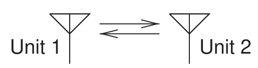
Schematic	ANS definition (2005) [3]	ITU definition (2004) [1]	ITU: obsolete usage [1]
	Broadcast	Broadcast	Simplex
	Simplex	Simplex	Halfduplex
	Duplex	Duplex	Fullduplex

Table 5-1: Definitions of operating modes of wireless communication systems. American National Standard (ANS) definition, International Telecommunication Union (ITU) definition, and deprecated (i.e., obsolete) usage

5.2.1 International Telecommunications Union Definitions

The **ITU** is the agency of the United Nations that coordinates the shared global use of the radio spectrum and establishes worldwide standards.

Broadcast operation refers to one-way communication in which there is only one transmitter and at least one, and perhaps more, receivers. The ITU defines broadcasting as [1, 2]:

> a form of unidirectional telecommunication intended for a large number of users having appropriate receiving facilities, and carried out by means of radio or cable networks.

Simplex operation is defined by the ITU as [1]

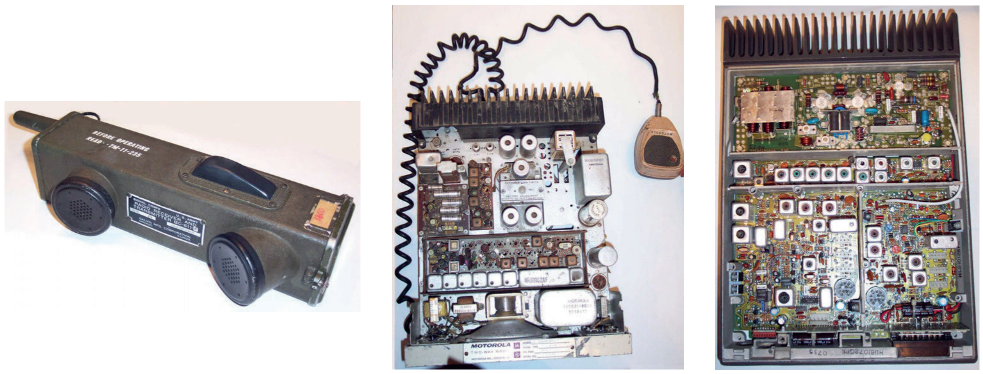
> designating or pertaining to a method of operation in which information can be transmitted in either direction, but not simultaneously, between two points.

So simplex operation is when there is communication from a first unit to a second unit. This is followed by a hand-off procedure and then communication from the second unit to the first. So the push-to-talk Family Radio Service (FRS) is an example of simplex communication as two people on opposite ends of the link cannot talk simultaneously. Simplex operation may use either one or two frequencies. That is, the forward link (from user 1 to user 2) may use the same frequency channel as the reverse link (from user 2 to user 1). However, simplex operation may use different frequency channels for the forward and the reverse link. The ITU defines "half duplex" as being synonymous with simplex, but the use of the term "half duplex" is deprecated (not recommended) [1, 2]. An early radio using simplex operation is shown in Figure 5-1(a).

Duplex Operation is defined by the ITU as [2]

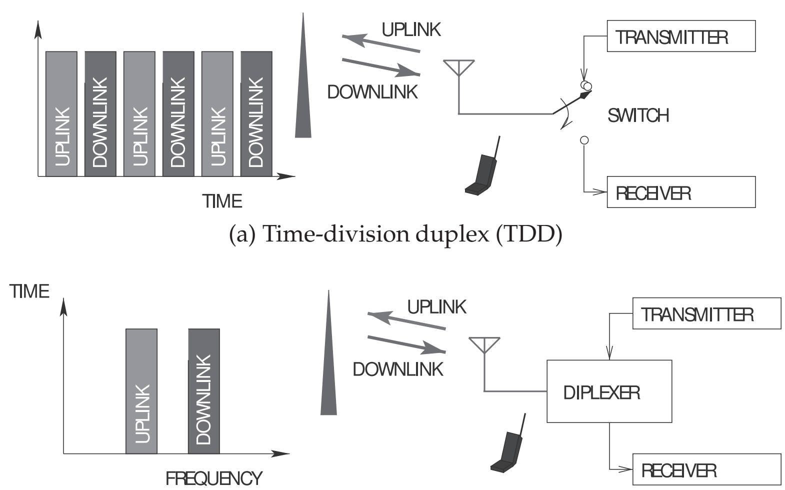
> the operating method in which transmission is possible simultaneously in both directions of a telecommunication channel.

Simultaneous two-way communication uses duplexing. In analog radio the links must be simultaneous and so operation requires two frequencies. But in digital radio with compression and expansion of digitized speech and



(a) (b) (c)

Figure 5-1: Early radios: (a) walkie-talkie with a push-to-talk (PTT) switch on top and using simplex communication; (b) Motorola business dispatcher two-way radio operating at 33.220 MHz designed in the 1960s as a dash mount unit; and (c) the Mitrek two-way radio designed by Motorola in 1977. The radio has two PC boards and was crystal controlled with a channel scanning control head. The radio was trunk mounted, with the control head, microphone, and speaker mounted under the dash board. The units in (b) and (c) are approximately 20 cm long.



(b) Frequency-division duplex (FDD)

Figure 5-2: Duplex schemes.

transmission of data, duplex operation can involve transmission of data in time slots that may not be simultaneous in each direction. If the time slots are short enough, say less than ten milliseconds, the effect will be that two-way voice communication is simultaneous. This is still called duplex operation. Early examples of radios using duplex operation are shown in Figure 5-1(b and c). A modern example is the cell phone.

The two predominant duplex schemes are **frequency-division duplex** (**FDD**) and **time-division duplex** (**TDD**), as illustrated in Figure 5-2. In these diagrams the mobile unit is shown communicating with a basestation, in which case transmission from the mobile unit to the basestation is called the **uplink** (**UL**), or less commonly the **reverse link** (**RL**), or **reverse path**.

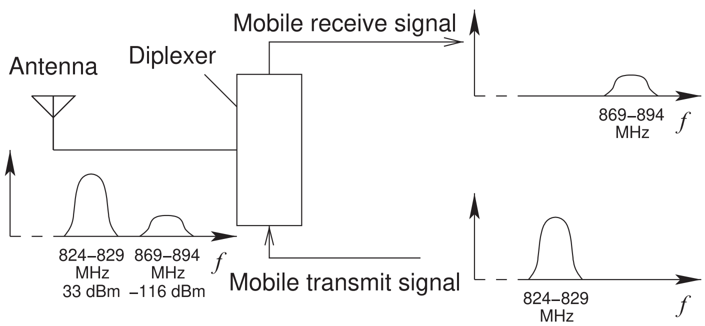


Figure 5-3: A diplexer which separates receive and transmit signals. A diplexer is a type of duplexer. However not all duplexers are diplexers.

Communication from the basestation and received by the mobile unit is called the **downlink**, (**DL**), **forward link** (**FL**), or **forward path**. In FDD it is necessary to use a filter to separate the uplink and downlink signals, as shown in Figure 5-3, as the two links are in use simultaneously.

In TDD, see Figure 5-2(a), the uplink and downlink are separated in time and a switch connects the antenna to the transmitter to send the uplink and then to the receiver to receive the downlink. The sequence is then repeated.

In FDD the uplink and downlink are at different frequencies and the transmitter and receiver are connected permanently to the antenna, see Figures Figure 5-2(b) and 5-3. FDD requires a diplex filter, which is a special filter with three ports that looks like a lowpass filter (usually for a handset) for the uplink when the transmitter is connected to the antenna (the uplink is generally at a lower frequency than the downlink) and a highpass filter for the downlink when the receiver is connected to the antenna.

5.2.2 Duplex Versus Diplex

The term duplex refers to a way of handling two communication channels. It derives from the term **multiplexing** (or **MUXing**) defined as [3]

> the combining of two or more information channels onto a common transmission medium. Note: In electrical communications, the two basic forms of multiplexing are time-division multiplexing (TDM) and frequency-division multiplexing (FDM). Synonym: multiplex.

Duplexing, or **duplex operation**, is used when there are two channels and in radio communications it nearly always refers to one transmit channel and on receive channel.

Diplex operation is defined as [3]:

> the sharing of one common element, such as a single antenna or channel, for transmission or reception of two simultaneous, independent signals on two different frequencies. Note: An example of diplex operation is the use of one antenna for two radio transmitters on different frequencies.

A **diplexer** is defined as [3]

> a three-port frequency-dependent device that may be used as a separator or a combiner of signals.

So in the context of cellular communications, a diplexer is a filter (see

Figure 5-3). It is a three-port filter that separates the transmitted and received signals, which are at different frequencies.

5.3 Cellular Communications

Cellular communications, as the name implies, are based on the concept of cells in which a terminal unit communicates with a basestation at the center of a cell. Each cell can be relatively small and a terminal unit travels smoothly from cell to cell with a connection transferring using what is called a hand-off process. For communication in closely spaced cells to work, interference from other radios must be managed. This is facilitated using the ability to recover from errors available with error correction schemes and the use of antenna beam-forming technology.

5.3.1 Cellular Concept

The cellular concept was outlined in a 1947 Bell Laboratories technical memorandum [4]. It described a system of frequency reuse with small geographical cells, and this remains the key concept of cellular radio. This was elaborated on in two articles published in 1957 and 1960 [5, 6]. The first widespread cellular radio system was the **Advanced Mobile Phone System** (**AMPS**), one of many 1G cellular radio systems, and was fully described by Bell Laboratories in a submission to the U.S. **Federal Communications Commission** (**FCC**) and in a patent filed on December 21, 1970 [7]. Bell Labs petitioned the FCC in 1958 for a frequency band around 800 MHz for a cellular system. The FCC, believing that it was better to allocate spectrum for the public good (such as radio, television, and emergency services) was reluctant to act on the petition. In 1968, pressure on the FCC became too great and an agreement was reached in principle to make frequencies available. Thus began the research and development of cellular systems in the United States. In 1961, Ericsson reorganized to address mobile radio, including cellular radio systems. Nokia did not begin developing 1G cellular systems until the 1970s. NTT was working away as well and began developing cellular radio systems in 1967 [8]. Meanwhile, in January 1969, the Bell System launched an experimental cellular radio system employing frequency reuse for the first time to achieve optimum use of a limited number of RF channels. The first commercial cellular system was launched by the Bahrain Telephone Company in May 1978 using Matsushita equipment. This was followed by the launch of AMPS by Illinois Bell and AT&T in the United States in July 1978.

In 1979 the **World Administrative Radio Conference** (**WARC**) allocated the 862–960 MHz band for mobile radio, leading to the FCC releasing, in 1981, 40 MHz in the 800–900 MHz band for "cellular land-mobile phone service." The service, as defined in the original documents, is (and this is still the best definition of cellular radio)

> a high capacity land mobile system in which assigned spectrum is divided into discrete channels which are assigned in groups to geographic cells covering a cellular geographic area. The discrete channels are capable of being reused within the service area.

The key attributes here are

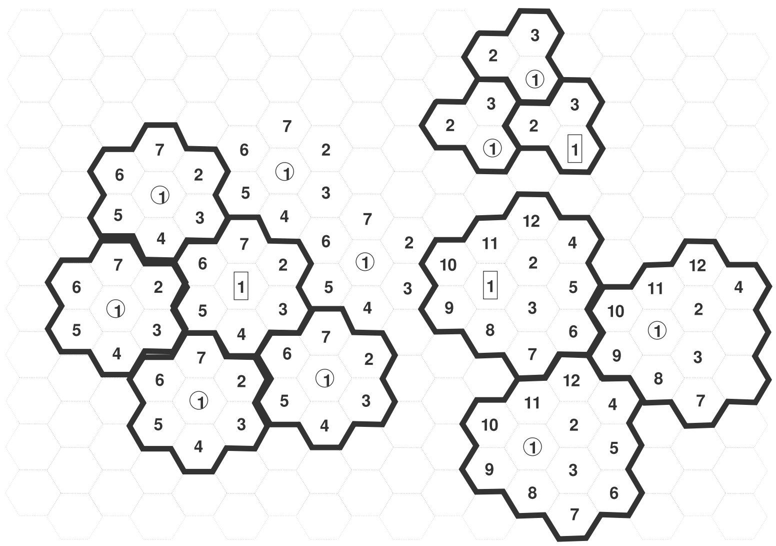


Figure 5-4: Cells arranged in clusters.

- High capacity. Prior to he cellular system, mobile radio users were not always able to gain access to the radio network and frequently access required multiple attempts.

- The concept of cells. The idea is to divide a large geographic area into cells, shown as the hexagons in Figure 5-4. The actual shape of the cells is influenced by obstructions such as hills and buildings, but the hexagonal shape is used to convey the concept of cells. The cells are arranged in clusters and the total number of channels available is divided among the cells in a cluster and the full set is repeated in each cluster. In Figure 5-4, 3-, 7-, and 12-cell clusters are shown. As will be explained later, the number of cells in a cluster affects both capacity (the fewer cells the better) and interference (the more cells per cluster, the further apart cells operating at the same frequency are, and so interference is less).

- Frequency reuse. Frequencies used in one cell are reused in the corresponding cell in another cluster. As the cells are relatively close, it is important to dynamically control the power radiated by each radio, as radios in one cell will produce interference in other clusters.

The shape of a cell depends on many factors. In a flat desert the coverage area of each basestation would be circular, so that with a cluster of cells there would be overlapping circles of coverage. (Power levels are adjusted to minimize the overlap of these circles.) Buildings, hills, lakes, etc., affect cell size. In a city, what is called the **urban canyon effect**, or **urban waveguide effect** greatly distorts cells and creates havoc in managing cellular systems [9–11]. In the urban canyon effect, good coverage extends for large distances down a street (see the inside front cover).

Achieving maximum frequency reuse is essential in achieving high capacity. In a conventional wireless system, be it broadcast or the mobile telephone service, basestations are separated by sufficient distance such that the signal levels fall below a noise threshold before the same frequencies are reused, as shown in Figure 5-5. There is clearly poor geographical use of the

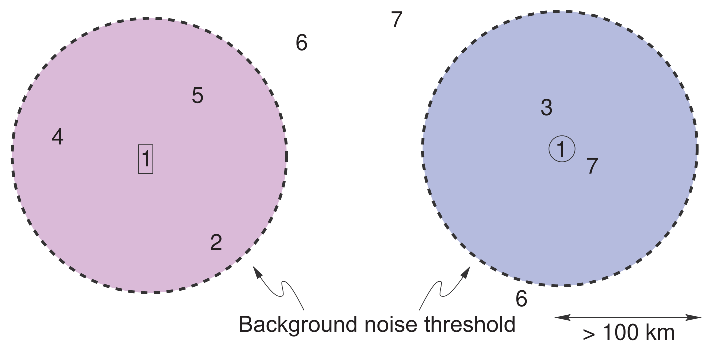


Figure 5-5: Interference in a conventional radio system. The two transmitters, 1, are at the centers of the coverage circles defined by the background noise threshold.

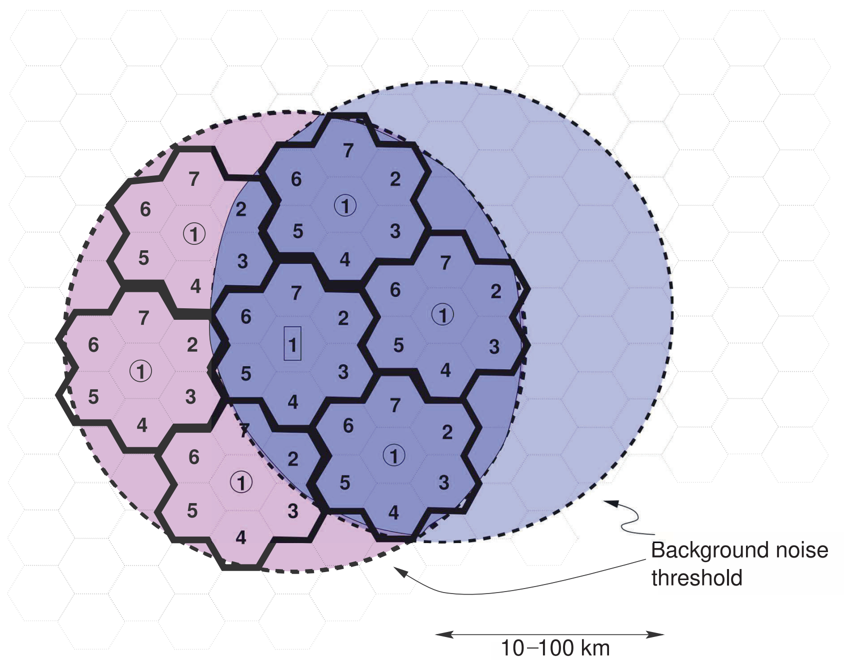


Figure 5-6: Interference in a cellular radio system.

spectrum here. The geographic areas could be pushed closer to each other at the expense of introducing what is called cochannel interference—a receiver could pick up transmissions from two or more basestations operating at the same frequency. This is strictly avoided so that interference in conventional radio systems results solely from background noise. In a cellular system, there is a radical departure in concept from this. Consider the interference in a cellular system as shown in Figure 5-6. The signals in corresponding cells in different clusters interfere with each other and the interference is much larger than that of the background noise. Generally only in rural areas and when mobile units are near the boundaries of cells will the background noise level be significant. Interference can also be controlled by dynamically adjusting the basestation and mobile transmit powers to the minimum acceptable level. Tolerating interference from neighboring clusters is a key concept in cellular radio.

The analog 1G AMPS system, which uses frequency modulation, has a qualitative minimum SIR of 17 dB (about a factor of 50) that was determined via subjective tests with a criterion that 75% of listeners ranked the voice quality as good or excellent. The seven-cell clustering shown in Figure 5-

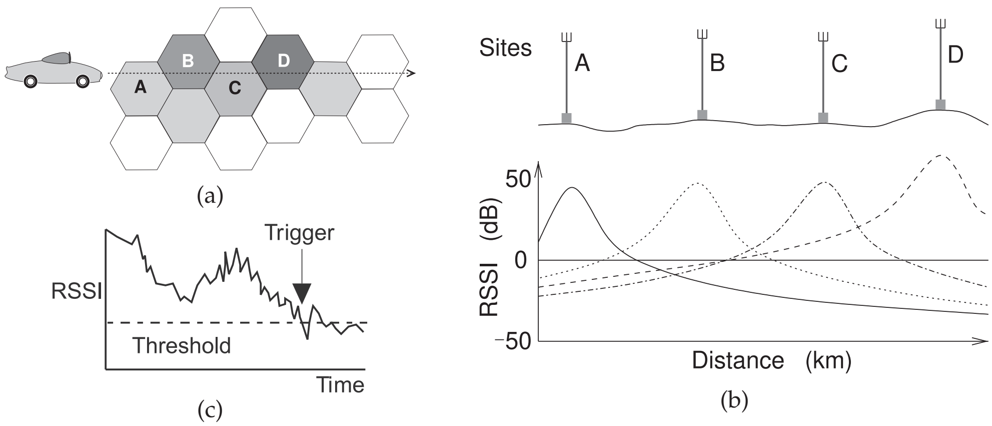


Figure 5-7: Process of handoff: (a) movement of a mobile unit through cells; (b) received signal strength indicator (RSSI) during movement of a mobile unit through cells; and (c) RSSI of a mobile unit showing the handoff triggering event.

6 does not yield this required minimum SIR. So either a 12-cell cluster is required or directed antennas are used, as these provide enough SIR.

The 2G and 3G digital cellular systems use error correction coding and can tolerate high levels of interference and can reuse frequency channels more efficiently. Indeed, in the 3G CDMA system the tolerance to interference is so high that the concept of clustering is not required and every frequency channel is available in each cell.

In 4G and 5G cellular radio interference must be low to enable high-order modulation to be used. The high modulation efficiencies more than makes up for the reduced frequency reuse.

5.3.2 Personal Communication Services

The personal communication services (**PCS**) concept was implemented in the early 1990s and was a development in the thinking of cellular communications. In PCS, the concept is that communication is from person to person, whereas in cellular radio communication as originally conceived, it was from terminal-to-terminal. The idea is that when a call is placed, an individual is being contacted rather than a piece of hardware. One way this is achieved is by using a card, a **subscriber identification module** (**SIM** card), to identify the user. A user can insert his or her SIM card into any (appropriate) handset and the handset becomes personalized. The term PCS is not commonly used now, as the concept has been incorporated in all evolved cellular systems.

5.3.3 Call Flow and Handoff

Mobile users can be expected to move frequently between cells and thus handoff procedures for transferring connections from one cell to the next are necessary. The triggering events that initiate handoff are shown in Figure 5-7. The main aspect is monitoring of the signal strength, the **received signal strength indicator** (**RSSI**), both in the handset for the signal received from a basestation and in the basestation for the signal received from a handset. If either of these falls below a threshold, computers in the basestation initiate a handoff procedure by polling nearby basestations for the RSSI they have for the user. If a suitable RSSI is found handoff proceeds and the other basestation takes control of the RF link to the mobile terminal.

5.3.4 Cochannel Interference

The minimum signal detectable in conventional wireless systems is determined by the received SIR. In cellular wireless systems the dominant interference is due to other transmitters in the cell and adjacent cells. The noise that is produced in the signal band from other transmitters operating at the same frequency is called cochannel interference. The level of cochannel interference is dependent on cell placement and the frequency reuse pattern. The degree to which cochannel interference can be controlled has a large effect on system capacity. Control of cochannel interference is largely achieved by controlling the power levels at the basestation and at the mobile units.

5.4 Multiple Access Schemes

Many schemes have been developed to enable multiple users to share a frequency band. The simplest scheme requiring the least sophistication in channel management is the **frequency division multiple access** (**FDMA**) scheme, shown in Figure 5-8(a), where the numerals indicate a particular user. In FDMA the available spectrum is divided by frequency, with each user assigned a narrow frequency band that is kept for the duration of a call. This can be conveniently implemented in analog radio. In duplex operation a frequency channel is assigned for the uplink as well as the downlink. An example of where FDMA is used is AMPS, where 30 kHz is assigned to each channel. Clearly this is wasteful of spectrum, as not all of the band is used continuously.

In the first digital access technique, **time division multiple access** (**TDMA**), shown in Figure 5-8(b), a bitstream is divided among a few users using the same physical channel. In the **GSM** mobile phone system,[1] a physical channel is divided into eight time slots and a user is allocated to each so that eight users can be supported in each physical channel. Thus the logical channels are divided in both frequency and time. In a TDMA system, the basestation transmits a continuous stream of data containing frames of time slots for multiple users. The mobile unit listens to this continuous stream, extracting and processing only the time slots assigned to it. On the reverse transmission, the mobile unit transmits to the basestation in bursts only in its assigned time slots. This is yet another complication to the RF design of TDMA systems. The RF circuitry operates in a burst mode with constraints on settling time. One of the advantages of GSM is that multiple slots can be assigned to the same user to support high data rates and time

[1] GSM stands for the 2G Global System for Mobile Communications and was formerly known as Groupe Spécial Mobile. The system was deployed worldwide in 1991. An interesting footnote is that the GSM group began discussions leading to the system in 1982. By the mid-1980s there were many different versions of the GSM system in Europe. The European Union (EU) intervened and all member countries adopted a single standard.

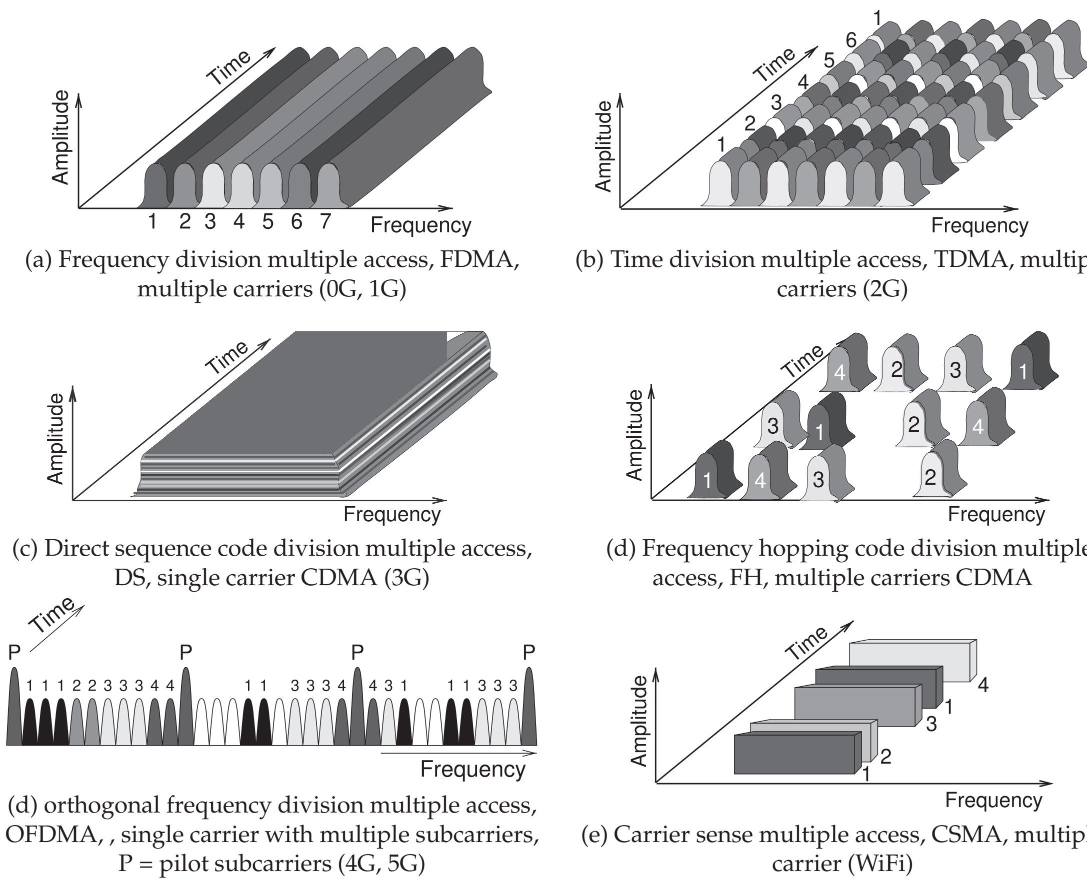
slots can be skipped.

Another technique that can be used in digital radio is **spread spectrum** (**SS** [12]), the invention of which is attributed to Hedwig Kiesler Markey[2]. Most of the development of spread spectrum was secret up until the mid-1970s. There are two main types of spread-spectrum techniques used in multiuser communication: direct sequence, and frequency hopping. It is both an access technique and a way to secure communications. **Direct sequence code division multiple access** (**DS-CDMA**)[3] is shown in Figure 5-8(c). DS-CDMA mixes the baseband signal with a broadband spreading code signal to produce a broadband signal that is then used to modulate an RF signal [13–15]. This process is illustrated in Figure 5-9. The rate of the spreading code is referred to as the **chip rate** rather than the bit rate, which is reserved for the rate of the information-bearing signal. A unique code is used to

---

[2] The 1940 invention is described in [12]. The patent described a frequency-hopping scheme to render radio-guided torpedoes immune from jamming using a piano roll to hop among 88 carrier frequencies.

[3] DS-CDMA is usually referred to as just CDMA.



(a) Frequency division multiple access, FDMA, multiple carriers (0G, 1G)

(b) Time division multiple access, TDMA, multiple carriers (2G)

(c) Direct sequence code division multiple access, DS, single carrier CDMA (3G)

(d) Frequency hopping code division multiple access, FH, multiple carriers CDMA

(d) orthogonal frequency division multiple access, OFDMA, , single carrier with multiple subcarriers, P = pilot subcarriers (4G, 5G)

(e) Carrier sense multiple access, CSMA, multiple carrier (WiFi)

Figure 5-8: Multiple access schemes for supporting multiple users with each user indicated by a numeral except for (b), TDMA, where the numerals indicate time slots.

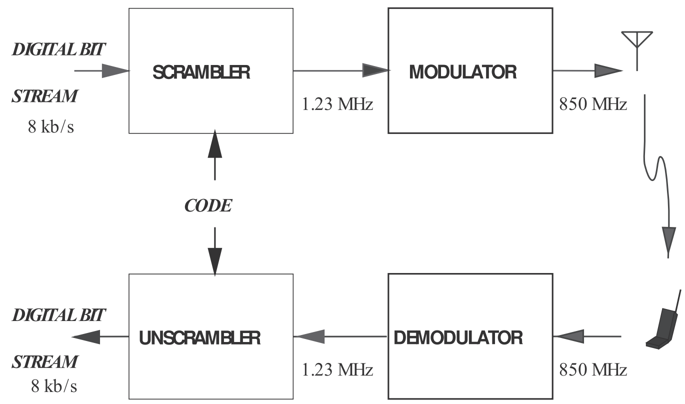


Figure 5-9: Block schematic of the operation of a CDMAOne spread-spectrum radio.

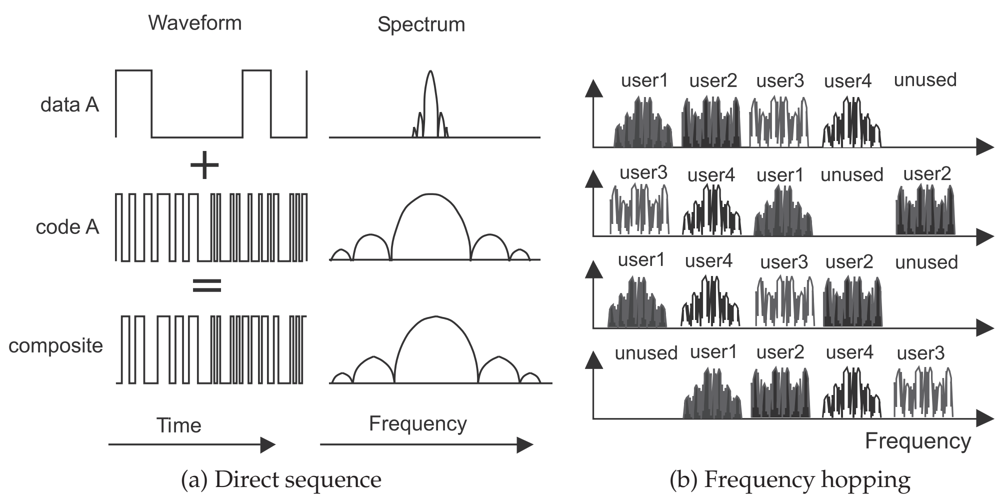


Figure 5-10: Basis of multiuser spread-spectrum communication.

"scramble" the original baseband signal, and the original signal can only be recovered using that particular code. The system can support many users, with each user assigned a unique code, hence the term code division multiple access (CDMA). Another interpretation of the DS-CDMA process is shown in Figure 5-10(a). One of the unique characteristics is that different users are using the same frequency band at the same time. CDMA is used in the 3G **wideband CDMA** (**WCDMA**) system.

Another form of CDMA access is the **frequency-hopping CDMA** scheme (**FH-CDMA**) shown in Figures 5-8(d) and 5-10(b). This scheme uses **frequency-hopping spread spectrum** (**FHSS**) to separate users. For one particular user, a code is used to determine the hopping sequence and hops can occur faster than the data rate. The signal can only be reassembled at the receiver if the hopping code is known. FH-CDMA is useful in unregulated environments, such as an ad hoc radio network, where it is not possible to coordinate multiple users. In this case, two users can transmit information in the same frequency band at the same time with error correction for occasional errors. FH-CDMA is commonly used in radios for emergency workers and by the military.

Another multiple access scheme is **carrier-sense multiple access** (**CSMA**), which is used in the WiFi system (IEEE 802.11). The access scheme is illustrated in Figure 5-8(e). This scheme is also used in hostile environments where various radios cannot be well coordinated. In CSMA, users transmit at different times, but without coordination. A terminal unit listens to the channel and transmits a data packet when it is free. Collisions are unavoidable and data are lost, requiring re-transmission of data, but the terminals use a random delay before re-transmitting data. In the case of WiFi there are several channels and if collisions are excessive another channel can be selected either as the preferred start-up channel or in evolved systems under central unit control.

Orthogonal frequency division multiple access (**OFDMA**) is used in **worldwide interoperability for microwave access** (**WiMAX**) and in 4G and 5G cellular radio. It overcomes many of the performance limitations encountered with CSMA as assignment of effective channels is used. OFDMA builds on orthogonal frequency division multiplexing (**OFDM**) introduced in Section 5.10.2.

5.5 Spectrum Efficiency

The concept of spectral efficiency is important in contrasting digital radio systems. Spectral efficiency has its origins in **Shannon's theorem**, which expresses the information-carrying capacity of a channel as [16–18]

$$\hat{C} = B_c \log_2(1 + \mathrm{SNR}), \tag{5.1}$$

where $\hat{C}$ is the capacity in units of bits per second (bit/s), B_c is the channel bandwidth in hertz, and SNR is the signal-to-noise ratio. N is assumed to be Gaussian noise, so interference that can be approximated as Gaussian can be incorporated by adding the noise and interference powers, and then it is more appropriate to use the SIR. Thus Equation (5.1) becomes

$$\hat{C} = B_c \log_2[1 + S/(N + I)] = B_c \log_2(1 + \mathrm{SIR}). \tag{5.2}$$

Shannon's theorem is widely accepted as the upper limit on the information-carrying capacity of a channel. So the stronger the signal, or the lower the interfering signal, the greater a channel's information-carrying capacity. Indeed, if there is no noise and no interference, the information-carrying capacity is infinite. Shannon's capacity formula indicates that increasing the interference level (lower SIR) has a more weakened effect on the decrease in capacity than may initially be expected; that is, doubling the interference level does not halve $\hat{C}$. This is the conceptual insight that supports the use of closely packed cells and frequency reuse, as the resulting increase in interference, and its moderated effect on capacity, is offset by having more cells and supporting more users.

Shannon's carrying capacity limit has not been reached, but today's radio systems are very close. Current systems operate with SNRs only a few decibels away from the limit [16]. Different modulation and radio schemes come closer to the limit, and two quantities will be introduced here to describe the performance of different schemes. From the capacity formula, a useful metric for the performance of modulation schemes can be defined. This is the modulation efficiency (also referred to as the **channel efficiency**,

channel spectrum efficiency, and **channel spectral efficiency**),

$$\eta_c = R_c/B_c, \tag{5.3}$$

where R_c (in bit/s) is the bit rate transmitted on the channel, so η_c has the units of bit/s/Hz. The unit is dimensionless, as hertz has the units of s^{-1}.

In a cellular system, the number of cells in a cluster must be incorporated to obtain a system metric [19]. The available channels are divided among the cells in a cluster, and a channel in one cell appears as interference to a corresponding cell in another cluster. Thus the SIR is increased and the capacity of the channel drops. System throughput increases, however, because of closely packed cells. So the system throughput is a function of the frequency reuse pattern. The appropriate system-level metric is the radio spectrum efficiency, η_r, which incorporates the number of cells, K, in a cluster:

$$\eta_r = \frac{R_b}{B_c K} = \frac{\eta_c}{K}\frac{R_b}{R_c}, \tag{5.4}$$

where R_b is the bit rate of useful information R_c, (R_c is higher than R_b because of coding). Coding is used to enable error correction, assist in identifying the start and end of a packet, and also to provide orthogonality of users in some systems that overlap users as with CDMA. The units of η_r are bit/s/Hz/cell. The decrease in channel capacity resulting from the increased SIR associated with fewer cells in a cluster (i.e., lower K) is more than offset by the increased system throughput.

So there are two definitions of spectral efficiency: the channel spectrum efficiency (also known as the modulation efficiency), η_c, which characterizes the efficiency of a modulation scheme, and the **radio spectrum efficiency**, η_r, which incorporates the added interference that comes from frequency reuse. Commonly both measures of efficiency are referred to as spectral efficiency, and then only the units identify which is being referred to.

One may well ask why efficiency is not expressed as a ratio of actual bit rate to Shannon's limit for a given set of conditions. There are several reasons:

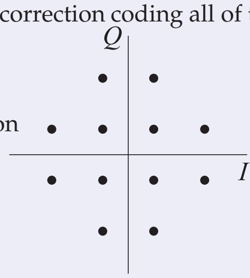
1. The historical use of bits per hertz to characterize a modulation scheme was used long before cellular systems came about.
2. Shannon's capacity limit is so high that back in the 1950s people would have been talking about extremely low efficiencies if performance was referred to the capacity limit.
3. Only additive white noise is considered, but this does not capture all types of interference, which can be multiplicative or partially correlated with the signal. Shannon's limit is not really a theoretical limit, there is no proof. In the digital communications world, much has been published about how close Shannon's capacity limit can be approached. In a direct LOS system, such as a point-to-point microwave link, the limit is now approached within a few percent. In MIMO systems, (see Section 5.10.5) the limit has been exceeded, prompting a redefinition of the limit when multiple transmit and receive antennas are used.

EXAMPLE 5.1 Modulation Efficiency

A radio uses a modulation scheme based on 16-QAM but the four constellation points corresponding to the largest signal are not used. Consequently the distortion that would occur in RF amplifiers is reduced. Even though there are 4 bits of information per symbol for the symbols that are actually used, not every possible combination of the bits is used. Ignoring error correction coding all of the bits modulated are information bits.

Solution:

(a) Draw the constellation diagram.



(b) How many symbols are there?
12.

(c) On average, how many bits of information are transmitted per symbol?
It takes 8 symbols to transmit 3 bits of information and 16 symbols to transmit 4 bits. With 12QAM 8 symbols are sent in the first clock tick interval and 4 symbols are borrowed from the second clock tick interval to provide 16 symbols combined and hence 4 bits of information. There are 8 symbols left over in the second clock tick interval and these can be used to send 3 bits. Thus over two clock tick intervals 7 bits of information are sent. Thus in one clock tick interval 3.5 bits are transmitted. So the number of bits of information sent by each symbol is 3.5 bits.

(d) What is the maximum possible modulation efficiency, η_c, in bit/s/Hz?
Ideally η_c is equal to the number of bits per symbol. However not all of the symbol transitions are of equal length and the bandwidth must be high enough to allow the longer transitions to take place in the same amount of time as the short transitions. However the maximum possible modulation efficiency is 3.5 bit/s/Hz.

5.6 Processing Gain

In digital radio the demodulated signal is a binary signal that includes information bits, coding bits, as well as bits that are in error which contributes to a raw bit error rate (BER). In the receiver decoding creates a smaller bitstream with just the information bits and which has a relatively low BER. Another way of looking at what happens is that the signal-to-interference ratio (SIR) of the decoded signal divided by the SIR of the raw signal represents a gain and this gain is called the processing gain. In 2G cellular radio this processing gain is also called the coding gain. With 3G spreading codes are used in addition to error correction codes with both reducing the bit error rate but now the gain from despreading is called spreading gain and the gain from error correction coding is called coding gain. The processing gain is now the product of spreading gain and coding gain. With 4G and 5G the functions of the spreading codes and error correction codes merge so just the term processing gain is used.

5.6.1 Energy of a Bit

Processing gain (G_P) occurs when demodulating and recovering an information bitstream and is the ratio of the SIR of the processed signal to the SIR of the unprocessed signal so G_P captures the amount that the SIR increases. G_P is a measure of the additional noise immunity obtained

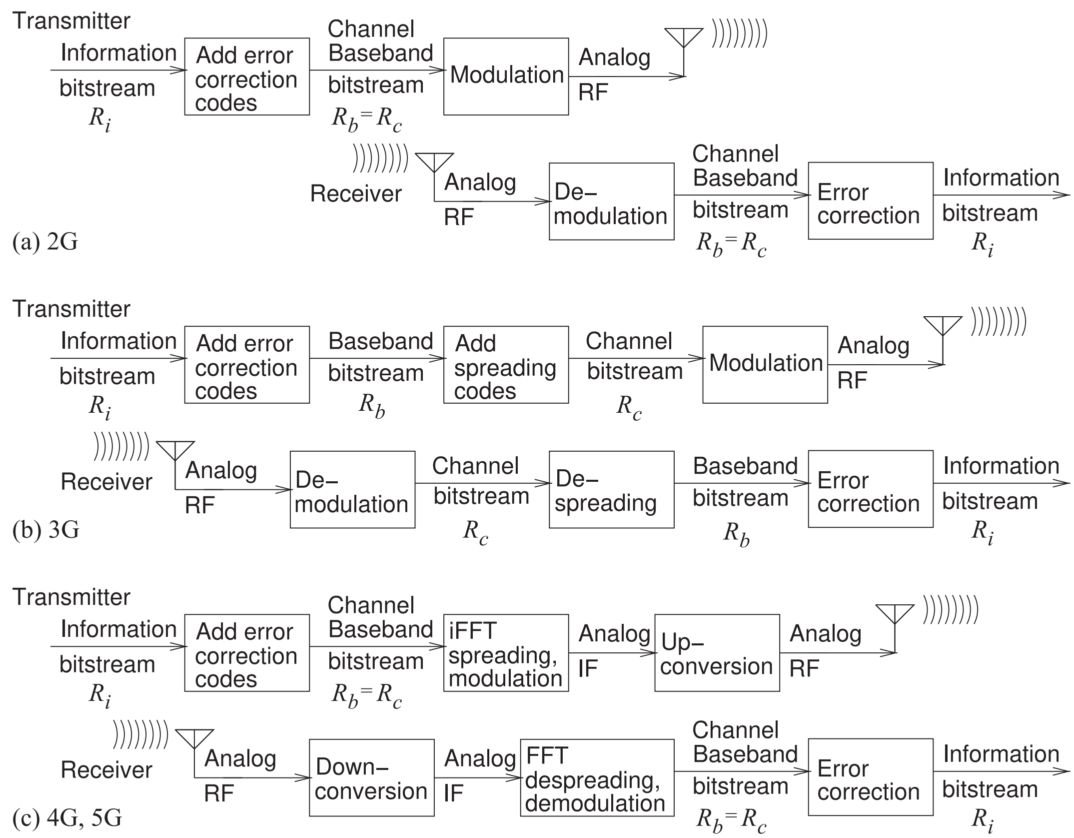


Figure 5-11: Flow of information in digital cellular radio with R_i, R_b, and R_c being the information, baseband, and channel bit rates, respectively.

through decoding and demodulation. There are various formulas for G_P depending on whether the processed signal is the baseband bitstream with or without error correction coding, and whether the unprocessed signal is the analog RF signal, the channel bitstream, or the channel symbol (i.e. chip) stream. Processing gain is an essential concept in cellular radio and really is what makes it work. In the following, reference will be made to three bitstreams in the transmitter which will be recovered in the receiver. The first is the **information bitstream** with a bit rate R_i (in units of bit/s). To this error correction coding can be used to produce a **baseband bitstream** with baseband bit rate $R_b > R_i$. In some generations of radio the baseband bitstream is modulated on a carrier and the baseband bitstream is the same as the channel bitstream. The bitstreams are shown in the depiction in Figure 5-11(a) of the information flow in 2G cellular radio.

In 3G cellular radio relatively fast spreading codes merged with the baseband bitstream to produce a much faster **channel bitstream** with bit rate $R_c \gg R_b$. The introduction of this separate third bitstream is the depiction in Figure 5-11(b) of the information flow in 3G cellular radio. The third bitstream is the **channel bitstream** with bit rate R_c and this is the bitstream that is modulated to produce a modulated signal with a constellation diagram where each symbol representing b bits.

In 4G and 5G cellular radio error correction codes are merged with the information bitstream to produce a baseband bitstream and this is spread during a first stage of the modulation process which produces an analog intermediate frequency modulated signal. Details of this process will be given later. There is not a separate spreading code so that the baseband and channel bitstreams are the same and $R_c = R_b$.

Derivation of the coding, spreading, and processing gains is based on the energy of a bit and the energy of the noise in the interval corresponding to a bit. The energy of a bit in the xth bitstream ($x = i, b, c$ for information, baseband, and channel respectively) is denoted $E_{b,x}$ and the energy of the noise corresponding to the duration of the bit is denoted $N_{o,x}$. The digital equivalent of the analog SIR is E_b/N_o (pronounced E-B-N-O for **EBNO**) and the effective SIR of the xth bitstream is

$$\mathrm{SIR}_{\mathrm{eff},x} = \frac{E_{b,x}}{N_{o,x}}. \tag{5.5}$$

A bitstream of course is a binary signal and the noise in the bitstream is manifested as binary errors in the bitstream. Consider a sequence of 7 bits, ideally 1001110, with one of the bits being in error so that the bitstream recovered is 1001010 where 1 bit in 7 is in error so $\mathrm{SIR}_{\mathrm{eff}}$ = 7.

5.6.2 Coding Gain

There are many types of error correction coding schemes with some schemes better for certain types of errors[4]. Generalizing, with each extra error correcting bit added to an information bitstream to produce the baseband bitstream, one error in the recovered bitstream can be corrected. Thus the processing gain due to coding (and often called just coding gain) in going from the baseband to the information bitstream is

$$G_{PC} = \frac{\mathrm{SIR}_{\mathrm{eff},i}}{\mathrm{SIR}_{\mathrm{eff},b}} = \frac{E_{b,i}/N_{o,i}}{E_{b,b}/N_{o,b}} = \frac{R_b}{R_i}. \tag{5.6}$$

This is a bit-wise processing gain as it applies to bitstreams. The coding gain here is not restricted to error correction coding as there are other types of codes called spreading codes that have a similar property of providing redundancy that can be used to remove errors. Not codes are 100% effective. However Equation (5.6) is the best simple measure of processing gain when only bitstreams are considered. It provides the coding gain of one bitstream derived from a second bitstream to which coding has been used to randomize and increase the rate of a bitstream and thus introduce redundancy.

[4] The error codes used in cellular radio are **forward error correction** (**FEC**) **codes** also called **channel codes**. There are many FEC codes broadly categorized as either block codes (because they work on blocks of data) or convolution codes (because they work on arbitrary-length bitstreams). The various FEC codes use different models (i.e. assumptions) about the types of errors encountered. The selection of the FEC code to use is a design choice based on the available computing power and the nature of the errors, e.g. random or long strings of errors.

5.6.3 Spreading Gain

Equation (5.6) can be rearranged so that the information EBNO is

$$\frac{E_{b,i}}{N_{o,i}} = G_{PC}\frac{E_{b,b}}{N_{o,b}}. \tag{5.7}$$

The processing gain determined in Equation (5.6) applies to bitstreams and does not include the effect of modulation. A second form of the processing gain relates the SIR of the analog RF signal, i.e. $\mathrm{SIR}_{\mathrm{RF}}$, to the EBNO of the baseband bitstream. To include the effect of modulation it is recognized that what is modulated and transmitted are symbols and the transitions from one symbol to another. The energy of a symbol is denoted E_s and this energy is shared by the b bits associated with the symbol. Thus a channel bit will have the energy $E_{b,c} = E_s/b$. In digital radio power levels of transmitted signals are adjusted so that at most one received channel bit can be in error per received symbol. Thus if received noise results in a symbol error it will affect just one bit. That is, the symbol noise will be the same as the bit noise, $N_{o,c} = N_o$ where N_o is the noise energy in the duration of one symbol. Thus the effective SIR of the channel bitstream is

$$\mathrm{SIR}_{\mathrm{eff},c} = \frac{E_{b,c}}{N_{o,c}} = \frac{E_{b,c}}{N_o} = \frac{E_s/b}{N_o}. \tag{5.8}$$

Development is completed by relating the energy of a symbol to the RF signal energy. For PSK modulation all of the symbols have the same energy. (The energy of each symbol is not the same for modulation formats such as QAM where symbols have different energy levels and a more sophisticated derivation is required than that provided here.) Also the noise energy in the duration of a symbol is the same for all symbols for all modulation formats. Thus the effective SIR of a symbol is

$$\mathrm{SIR}_{\mathrm{RF}} = \frac{E_s}{N_o}. \tag{5.9}$$

Thus the effective SIR of the channel bitstream is (combining Equations (5.8) and (5.9),

$$\mathrm{SIR}_{\mathrm{eff},c} = \frac{E_{b,c}}{N_{o,c}} = \frac{1}{b}\mathrm{SIR}_{\mathrm{RF}}. \tag{5.10}$$

The processing gain relating the EBNO of the recovered baseband bitstream to the SIR of the RF signal can be defined as the processing gain due to spreading and modulation and this is sometimes called just the **spreading gain** G_{PS}. It is defined here as the ratio of the effective SIR of the baseband bitstream and the RF SIR (and using Equation (5.10)):

$$G_{PS} = \frac{E_{b,b}/N_{o,b}}{\mathrm{SIR}_{\mathrm{RF}}} = \frac{1}{b}\frac{E_{b,b}/N_{o,b}}{\mathrm{SIR}_{\mathrm{eff},c}} = \frac{1}{b}\frac{\mathrm{SIR}_{\mathrm{eff},b}}{\mathrm{SIR}_{\mathrm{eff},c}}. \tag{5.11}$$

This can be rearranged so that the EBNO of the baseband bitstream is

$$\frac{E_{b,b}}{N_{o,b}} = G_{PS}\mathrm{SIR}_{\mathrm{RF}}. \tag{5.12}$$

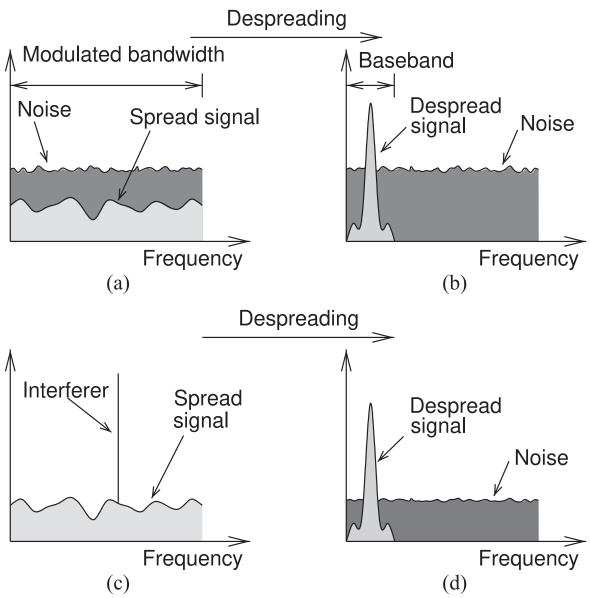


Figure 5-12: Increase in SNR obtained by despreading a spread-spectrum signal: (a) the spectrum of noise and a spread-spectrum signal that has a power below the level of the noise; (b) the despread signal, where the power distribution of the noise is not changed but the despread signal is confined to a narrower bandwidth called the baseband bandwith; (c) a signal with a single-tone interferer; and (d) after despreading where the power of the interferer spread across the the original bandwidth.

Equation (5.11) includes the ratio of the effective SIRs of two bitstreams and this was related to the bit rates of the bitstreams in Equation (5.6) where the faster bitstream is coded (or spread) to enable recovery from errors. Using the result in Equation (5.6), Equation (5.11) can be written as

$$G_{PS} = \frac{1}{b}\frac{\mathrm{SIR}_{\mathrm{eff},b}}{\mathrm{SIR}_{\mathrm{eff},c}} = \frac{1}{b}\frac{R_c}{R_b}. \tag{5.13}$$

Using this the EBNO of the baseband bitstream is

$$\frac{E_{b,b}}{N_{o,b}} = \mathrm{SIR}_{\mathrm{eff},b} = \frac{1}{b}\frac{R_c}{R_b}\mathrm{SIR}_{\mathrm{RF}} = \frac{1}{b}G_{PC}\mathrm{SIR}_{\mathrm{RF}}. \tag{5.14}$$

Then the EBNO of the information bitstream is

$$\frac{E_{b,i}}{N_{o,i}} = G_{PC}\frac{E_{b,b}}{N_{o,b}} = G_{PC}G_{PS}\mathrm{SIR}_{\mathrm{RF}} = G_P\mathrm{SIR}_{\mathrm{RF}} = \frac{1}{b}\frac{R_c}{R_i}\mathrm{SIR}_{\mathrm{RF}} \tag{5.15}$$

where the processing gain $\quad G_P = G_{PC}G_{PS}. \qquad (5.16)$

5.6.4 Spreading Gain in Terms of Bandwidth

The spreading gain that results from spreading a baseband bitstream in a transmitter and then despreading in a receiver is illustrated graphically in Figure 5-12. The modulated signal shown in Figure 5-12(a) is the "spread signal" which here has a power density which is below that of the noise. Following despreading all of the energy in the modulated RF signal is correlated with the spreading code (it was spread using the spreading code) and is collapsed to the despread signal shown in Figure 5-12(b). Only the signal correlated to the despreading code is collapsed to the smaller baseband bandwidth. Noise is rearranged with the spectral density of the noise unchanged so that the total noise energy in the baseband for the duration of a bit of data, i.e. the noise in the baseband bandwidth B_b, is greatly reduced by the ratio of the modulated bandwidth to the baseband bandwidth.

In Equation (5.13) R_c is related to the bandwidth B_m of the modulated carrier by the modulation efficiency as $\eta_c = R_c/B_m$(see Equation (5.3)). Also the minimum bandwidth of a baseband bitstream with a bit rate R_b is the baseband bandwidth $B_b = R_c$. So Equation (5.13) becomes

$$G_{PS} = \frac{\eta_c}{b}\frac{B_m}{B_b}. \tag{5.17}$$

Ideally for a modulation scheme $\eta_c = b$ so that

$$G_{PS,\text{ideal}} = \frac{B_m}{B_b}. \tag{5.18}$$

For all modulation schemes other than BPSK where η_c can be very close to b, the spreading gain will be less than $G_{PS,\text{ideal}}$ as in reality $\eta_c < b$, see Table 2-2. If both spreading and error correction coding are used the two processing gains in Equations (5.6) and (5.17) can be multiplied.

5.6.5 Symbol Error Rate and Bit Error Rate

With digital modulation, symbols (i.e., groups of bits) are transmitted rather than individual bits. As the modulation order increases, the number of symbols increases and there is a smaller margin between the symbols. That is, on a constellation diagram there are more symbols and the symbols are closer together. For the same signal and noise levels, as the modulation order increases the probability of a symbol error will increase and thus the symbol error rate (SER) increases. However the bit error rate (BER) increases more slowly than the SER. This is because if the SIR is high enough, the only possible errors will be nearest-neighbor errors (on a constellation diagram). If there are at least b bits per symbol the BER will be less than the **symbol error rate** by a factor of b as errors are at most one bit per symbol since the modulation order or the signal power level are adjusted to ensure this. If it is not possible to achieve a maximum of one bit error per symbol error then communication is lost. The rest of this section treats this analysis more mathematically.

The discussion begins with the SIR of the incoming RF signal. Through sampling of the RF signal at appropriate times (determined by the recovered carrier) discrete symbols are obtained. The digital form of SIR relates the energy of a symbol, E_s, to the noise and interference energy corresponding to the symbol (the noise and interference in the duration of the symbol). The height of the double-sided noise spectral density is conventionally taken as $N_o/2$, so the noise power corresponding to a symbol is N_o. Thus [16–18]

$$\frac{E_s}{N_o} = \text{SIR}. \tag{5.19}$$

Error in a digitally modulated communications system is first manifested as a symbol error that occurs when a symbol selected in a receiver is not the symbol transmitted. The probability of a symbol error is a function of E_s/N_o. However it is not always possible to develop a closed-form expression relating the two. For a BPSK system it can be derived. The probability of

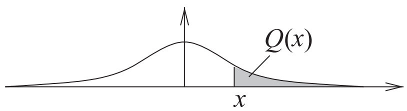


Figure 5-13: Gaussian distribution function showing the Q function as the area of the shaded region.

a symbol error, the SER, is [16, p. 187]

$$\mathrm{SER}_{\mathrm{BPSK}} = \Pr[\text{symbol error}] = \Pr_s^{\mathrm{BPSK}} = Q\left(\sqrt{\frac{2E_s}{N_o}}\right) = Q\left(\sqrt{2\cdot\mathrm{SIR}}\right), \tag{5.20}$$

where $Q(x)$ is known as the Q function [16] and is the integral of the tail of the Gaussian density function (see Figure 5-13). It can be expressed in terms of the error function $\mathrm{erf}(x)$ and complementary function $\mathrm{erfc}(x)$ [16, page 63]:

$$Q(x) = \tfrac{1}{2}\mathrm{erfc}\left(\frac{x}{\sqrt{2}}\right) = \tfrac{1}{2}\left[1 - \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right]. \tag{5.21}$$

For M-PSK [16, page 191] the SER is

$$\mathrm{SER}_{M-\mathrm{PSK}} = \Pr_s^{M-\mathrm{PSK}} \approx 2Q\left[\sqrt{\frac{2E_s}{N_o}}\sin\left(\frac{\pi}{M}\right)\right], \tag{5.22}$$

where M is the number of symbols (e.g. for 8-QPSK $M = 8$). To see the effect of higher-order modulation (i.e., higher M) consider Equation (5.22). As the order of modulation increases, M increases, and the argument of Q reduces, thus increasing SER.

The per-bit SIR is obtained after noting that each symbol can represent several bits. With a uniform constellation and the same number of bits per symbol, b, the signal energy received per bit is

$$E_b = E_s/b, \qquad b = \log_2 M, \tag{5.23}$$

and

$$M = 2^b. \tag{5.24}$$

For high SIR, a symbol error is the erroneous selection of a nearest neighbor symbol by the receiver. So with **Gray code mapping** (also called **Gray mapping**), such a symbol error results in only a single bit being in error. Thus the probability of a bit error, the BER, is

$$\Pr[\text{bit error}] = \Pr_b = \frac{1}{b}\Pr_s. \tag{5.25}$$

The final results are the bit error probabilities for BPSK and M-PSK [16, page 193]:

$$\mathrm{BER}|_{\mathrm{BPSK}} = \Pr_{b,\mathrm{BPSK}} = Q\left(\sqrt{2E_s/N_o}\right) = Q\left(\sqrt{2\cdot\mathrm{SIR}}\right) \tag{5.26}$$

$$\begin{aligned}\mathrm{BER}|_{M-\mathrm{PSK}} = \Pr_{b,M-\mathrm{PSK}} &= \frac{2}{b}Q\left[\sqrt{2E_s/N_o}\sin\left(\frac{\pi}{M}\right)\right]\\ &= \frac{2}{b}Q\left[\sqrt{2\cdot\mathrm{SIR}}\sin\left(\frac{\pi}{M}\right)\right].\end{aligned} \tag{5.27}$$

Thus for the same SIR the probability of bit errors, the BER, grows with higher-order modulation (i.e., larger M), but not as fast as SER. At the same time the number of bits transmitted increases. These extra bits are use to embed error correction codes. The net gain in throughput can be tremendous as long as the SIR is high enough.

EXAMPLE 5.2 Symbol Error Rate

Calculate the SER and BER for QPSK and 8-PSK if the SIR is 10 dB.

Solution:

SER is found by evaluating Equation (5.22). For SIR = 10 dB, E_s/N_o= $10^{(\mathrm{SIR_{dB}}/10)}$ = 10.
For QPSK, $M = 4$ and $b = 2$, and the SER is

$$\begin{aligned}\mathrm{SER_{QPSK}} &= \mathrm{Pr}_s^{M-\mathrm{PSK}} \approx 2Q\left(\sqrt{\frac{2E_s}{N_o}}\sin\left(\frac{\pi}{M}\right)\right) = 2Q\left[\sqrt{20}\sin\left(\frac{\pi}{4}\right)\right] \\ &= 2Q(3.162) = \left[1 - \mathrm{erf}(3.162/\sqrt{2})\right] = 0.001565. \end{aligned} \tag{5.28}$$

For 8-PSK, $M = 8$ and $b = 3$, and the SER is

$$\begin{aligned}\mathrm{SER_{8-PSK}} &= \mathrm{Pr}_s^{M-\mathrm{PSK}} \approx 2Q\left[\sqrt{\frac{2E_s}{N_o}}\sin\left(\frac{\pi}{M}\right)\right] = 2Q\left(\sqrt{20}\sin\left(\frac{\pi}{8}\right)\right) \\ &= 2Q(1.711) = 0.08701. \end{aligned} \tag{5.29}$$

The corresponding BERs are

$$\mathrm{BER_{QPSK}} = \tfrac{1}{2}\mathrm{SER_{QPSK}} = 0.000783 \quad \text{and} \quad \mathrm{BER_{8-PSK}} = \tfrac{1}{3}\mathrm{SER_{8-PSK}} = 0.0290. \tag{5.30}$$

5.6.6 Summary

The formulas for processing gain developed above are only approximate as many simplifications were made. One assumption is that the error correction coding is fully randomized bits but in practice there are many error correction codes that address particular types of errors. The aim of error correction coding is to recover from errors not necessarily to achieve processing gain.

Another caveat to the processing gain formulas developed here is that it considered the energy of a symbol as being the same for all symbols. While this is true for PSK modulation it is not true for all modulation formats. Considerable effort is put into developing better estimates that capture the essence of processing gain but in a more rigorous manner. Still the simple formulas provide the required insight into the effect of using error correction and spreading codes.

In 2G cellular radio processing gain is achieved using error correction codes alone and a separate spreading code is not used. Thus in 2G the baseband bits are also the channel bits used directly in modulating the RF carrier. The extra bits provide redundancy and the ability to recover from some errors resulting from noise. In 3G spreading codes as well as error correcting codes are used. A spreading code greatly increases the bit rate of the channel bitstream above that of the baseband bitstream and the bandwidth of the modulated carrier is much greater than that needed to transmit the baseband bitstream. Generally error correcting codes do not increase the bit rate by more than a factor of 2. Thus in 3G the processing gain achieved with the error correction codes is almost insignificant compared to

Table 5-2: Processing gain definitions. R_i is the information bit rate, $R_b > R_i$ is the baseband bit rate after error correction coding of the information bitstream, and $R_c > R_b$ is the channel bit rate after spreading the baseband bitstream. Reference is made to the xth bitstream with $x = i, b,$ or c indicating the information, baseband, or channel bitstreams respectively. $B_{b,x}$ is the energy of a bit in the xth bitstream, and $N_{o,x}$ is equivalent noise energy corresponding to a bit in the xth bitstream. E_s is the energy of a symbol with b bits per symbol. B_m is the bandwidth of the modulated carrier and B_b is the bandwidth of the baseband signal.

Description	Definition	Formula	Equation
Coding gain, processing gain from error correction coding. Calculated for bitstreams.	$G_{PC} = \dfrac{E_{b,i}/N_{o,i}}{E_{b,b}/N_{o,b}}$	$G_{PC} = \dfrac{R_b}{R_i}$	(5.6)
Processing gain from spreading calculated on a bitstream basis	$G_{PS} = \dfrac{E_{b,c}/N_{o,c}}{E_{b,b}/N_{o,b}}$	$G_{PS} = \dfrac{1}{b}\dfrac{R_c}{R_b}$	(5.13)
Processing gain from spreading in terms of bandwidth.	$G_{PS} = \dfrac{E_{b,b}/N_{o,b}}{\mathrm{SIR_{RF}}}$	$G_{PS} = \dfrac{\eta_c}{b}\dfrac{B_m}{B_b}.$	(5.17)
Processing gain	$G_P = \dfrac{E_{b,c}/N_{o,c}}{E_{b,i}/N_{o,i}}$	$G_P = G_{PC}G_{PS}$	(5.16)
EBNO	$E_{b,i}/N_{b,i}$	$E_{b,i}/N_{b,i} = G_P\mathrm{SIR_{RF}}$	(5.12)

that obtained from spreading the signal.

In 4G and 5G error correction coding also provides spreading and there is not a separate spreading code. The 4G and 5G systems have great complexity.

A summary of the key results for processing gain is given in in Table 5-2.

EXAMPLE 5.3 Processing Gain

A new communication system is being investigated for sending data to a printer. The system will use GMSK modulation and a channel with 10 MHz bandwidth and the baseband bit rate will be 1 Mbit/s. The modulation format will result in a spectrum that distributes (i.e. spreads) power almost uniformly over the 10 MHz bandwidth.

(a) What is the processing gain?

The most efficient way to spread the signal across the 10 MHz of available bandwidth is to use a combination of error correction and spreading. Referring to Table 2-2, GMSK has a modulation efficiency η_c of 1.354 bit/s/Hz, so that the channel bit rate R_c = 13.54 Mbit/s. Each symbol in GMSK represents two bits so $b = 2$. The information bit rate R_i = 1 Mbit/s and so coding at the rate of 12.54 Mbit/s will be used and $R_b = (R_i + 12.54\ \mathrm{Mbit/s}) = 13.54$ Mbit/s.

There are two processing gains. The coding gain is determined using Equation (5.6):

$$G_{PC} = \frac{R_b}{R_i} = \frac{13.54\ \text{Mbit/s}}{1\ \text{Mbit/s}} = 13.54 = 11.31\ \text{dB}.$$

The spreading gain, using Equation (5.13), is

$$G_{PS} = \frac{1}{b}\frac{R_c}{R_b} = \frac{1}{2}\frac{1\ \text{Mbit/s}}{1\ \text{Mbit/s}} = 0.5 = -3\ \text{dB}.$$

Thus the processing gain

$$G_P = G_{PC}G_{PS} = 6.77 = 8.31\ \text{dB}.$$

Another way of calculating the spreading gain is to consider the modulated bandwidth, $B_m = 10$ MHz and the baseband bandwidth $B_b = 1$ MHz (taken to be numerically equal to R_b). The processing gain due to spreading, from Equation (5.17):

$$G_{PS} = \frac{\eta_c}{b}\frac{B_m}{B_b} = \frac{1.354}{2}\frac{10\ \text{MHz}}{1\ \text{MHz}} = 6.770 = 8.31\ \text{dB}.$$

(b) If the received RF SIR is $\text{SIR}_{\text{RF}} = 6$ dB, what is the effective system SIR (or E_b/N_o) after the digital signal processor?

$$\text{Effective SIR} = \frac{E_{b,i}}{N_{o,i}} = G_P\text{SIR}_{\text{RF}} = 8.31\ \text{dB} + 6\ \text{dB} = 14.32\ \text{dB}.$$

EXAMPLE 5.4 Signal-to-Interference Ratio

At the output of a receiver antenna, the level of interfering signals is 1 pW, the level of background noise is 500 fW, and the level of the desired signal is 4 pW.

(a) What is the SIR? Note that SIR includes the effect of the signal, interference, and noise.
(b) If the processing gain is 20 dB and 16-QAM is the modulation scheme used, what is the effective system SIR, that is, what is the signal energy in a bit versus the noise energy in the duration of the bit (i.e., E_b/N_o)?

Solution:

(a) Interfering signal $P_I = 1$ pW, noise signal $P_N = 500$ fW, and signal $P_S = 4$ pW.

$$\begin{aligned}\text{SIR}_{\text{RF}} &= P_S/(P_I + P_N)\\ &= 4\ \text{pW}/(1\ \text{pW} + 0.5\ \text{pW}) = 4/1.5 = 2.667 = 4.26\ \text{dB}.\end{aligned}$$

(b) Processing gain $G_P = 20$ dB so, from Equation (5.15),

$$\text{Effective SIR} = \frac{E_{b,i}}{N_o} = G_P \cdot \text{SIR}_{\text{RF}} = 20\ \text{dB} + 4.26\ \text{dB} = 24.26\ \text{dB}.$$

5.7 Early Generations of Cellular Phone Systems

The evolution of cellular communications is described by generations of radio. The major mobile communication systems are outlined in Table 5-3. No 0G and 1G systems remain and 2G systems are being phased out with only the 2G GSM system still in use. Third generation (3G) offers a significant increase in capacity and is optimized for broadband data access. The 0G–2G systems were plagued by many incompatible systems. Even with 3G there were several incompatible systems but the system known as WCDMA became overwhelmingly dominant. With 4G the evolution coalesced and while there were many modulation methods a standardized interface was developed with the physical interface defined as common. The 4G radio, and also 5G, support many modulation methods and frequency bands and cellular radios support a large number of these combinations. In part this is because the bulk of the modulation is now performed in a DSP unit making the switching of modulation methods a simple process of using a different algorithm. The development of 4G, 5G (and beyond) has been guided by the Third Generation Partnership, 3GPP, which developed a plan called long term evolution (LTE) that ensured upwards compatibility as the capability of the systems were enhanced every year or two eventually becoming 5G.

5.8 Early Generations of Radio

5.8.1 1G, First Generation: Analog Radio

The initial cellular radio system was analog, with the dominant system being AMPS, the attributes of which are given in Table 5-4. This is a relatively simple system, but appropriate for the low levels of integration of the 1980s, as most of the functionality could be realized using analog circuits. The first-

Table 5-3: Major mobile communication systems with the year of first widespread use.

System	Year	Description
0G		Broadcast, no cells, few users, analog modulation
MTS	1946	Mobile telephone service, halfduplex, operator assist to establish call, push to talk
AMTS	1965	Advanced mobile telephone system, Japan, duplex, 900 MHz
IMTS	1969	Improved mobile telephone service, duplex, up to 13 channels, 60–100 km (40–60 mile) radius, direct dial using dual-tone multifrequency (DTMF) keypad
0.5G		FDMA, analog modulation
PALM	1971	(also Autotel) Public automated land mobile radiotelephone service, used digital signaling for supervisory messages, technology link between IMTS and AMPS
ARP	1971	Autoradiopuhelin (car radio phone), obsolete in 2000, used cells (30 km radius) but not hand-off, 80 channels at 150 MHz, simplex and later duplex
1G		Analog modulation, FSK for signaling, cellular, FDMA
NMT	1981	Nordic mobile telephone, 12.5 kHz channel, 450 MHz, 900 MHz
AMPS	1983	Advanced mobile phone system, 30 kHz channel
TACS	1985	Total access communication systems, 25 kHz channel, widely used in Europe until 1990s, similar to AMPS
Hicap	1988	NTT's mobile radiotelephone service in Japan
Mobitex	1990	National public access wireless data network, first public access wireless data communication services including two-way paging network services, 12.5 kHz channel, GMSK
DataTac	1990	Point-to-point wireless data communications standard (like Mobitex), wireless wide area network, 25 kHz channels, maximum bandwidth 19.2 kbit/s (used by the original Blackberry device)
2G		Digital modulation
PHS	1990	Personal handyphone system, originally a cordless phone, now functions as both a cordless phone and as a mobile phone elsewhere
GSM	1991	Global system for mobile communications (formerly Groupe Spécial Mobile), TDMA, GMSK, constant envelope, 200 kHz channel, maximum 13.4 kbits per time slot (at 1900 MHz), 2 billion customers in 210 countries
DAMPS	1991	Digital AMPS (formerly NADC [North American digital cellular] and prior to that as U.S. Digital Cellular [USDC]), narrowband, $\pi/4$DQPSK, 30 kHz channel
PDC	1992	Personal Digital Cellular, Japan, 25 kHz channel
CDMAOne	1995	Brand name of first CDMA system known as IS-95, spread spectrum, CDMA, 1.25 MHz channel, QPSK
CSD	1997	Circuit switched data, original data transmission format developed for GSM, maximum bandwidth 9.6 kbit/s, used a single time slot

Table 5-3 continued.

System	Year	Description
2.5G		Higher data rates
WiDEN	1996	Wideband integrated dispatch enhanced network, combines four 25 kHz channels, maximum bandwidth 100 kbit/s
GPRS	2000	General packet radio system, compatible with GSM network, used GSM time slot and higher-order modulation to send 60 kbits per time slot, 200 kHz channel, maximum bandwidth 171.2 kbit/s
HSCSD	2000	High-speed circuit-switched data, compatible with GSM network, maximum bandwidth 57.6 kbit/s, higher quality of service than GPRS
2.75G		Medium bandwidth data—1 Mbit/s
CDMA2000	2000	CDMA, upgraded CDMAOne, double data rate, 1.25 MHz channel
EDGE	2003	Enhanced data rate for GSM Evolution, compatible with GSM network, 8-PSK, TDMA, maximum bandwidth 384 kbit/s, 200 kHz channel
3G		Spread spectrum
FOMA	2001	Freedom of mobile multimedia access, first 3G service, NTT's implementation of WCDMA
UMTS		Universal mobile telephone service, 5 MHz channel, data up to 2 Mbit/s
WCDMA	2004	Main 3G outside China
OFDMA	2007	Evolution to 4G (downlink high bandwidth data)
1xEV-DO		(IS-856) Evolution of CDMA2000, maximum downlink bandwidth 307 kbit/s, maximum uplink bandwidth 153 kbit/s
TD-SCDMA	2006	Time division synchronous CDMA, China, uses the same band for transmit and receive, basestations and mobiles use different time slots to communicate, 1.6 MHz channel
GAN/UMA	2006	Generic access network, formerly known as Unlicensed Mobile Access, provides GSM and GPRS mobile services over unlicensed spectrum technologies (e.g., Bluetooth and WiFi)
3.5G		
UMTS/HSDPA	2006	Upgraded WCDMA, High-speed downlink packet access, download of 7.2 Mbit/s
EV-DO Rev A	2006	CDMA2000 EV-DO revision A, downlink to 3.1 Mbit/s, uplink to 1.8 Mbit/s
3.75G		
UMTS/HSUPA	2007	High-speed uplink packet access, upload speeds to 5.76 Mbit/s
EV-DO Rev B	2008	CDMA2000 EV-DO revision B, downlink to 73 Mbit/s, uplink to 27 Mbit/s
UMTS/HSPA	2009	Upgraded WCDMA, High-speed packet access, downlink to 40 Mbit/s, upload to 10 Mbit/s. Eventually added CA and MIMO
3.9G	2009	WiMAX 1 (IEEE 802.16), 10 MHz bandwidth; IP-based; branded as 4G by service providers; MIMO + OFDMA, downlink of 37 Mbit/s, uplink 17 Mbit/s (for 2×2 MIMO); WiMAX 2, IEEE 802.16m, 20 MHz bandwidth, downlink of 110 Mbit/s, uplink 70 Mbit/s; not allowed in many countries
3.9G	2011	Long-term evolution (LTE); up to 20 MHz channel bandwidth, IP-based; branded as 4G by service providers Low latency (for VoIP) + MIMO + OFDMA, downlink of 100 Mbit/s
4G	2013	LTE-advanced, downlink of 1 Gbit/s fixed, 100 Mbit/s mobile, variable bandwidths of 5–40 MHz
5G	2019	Millimeter waves with beam steering and massive MIMO; Mesh networks and cognitive radio

Table 5-4: Attributes of AMPS.

Property	Attribute
Number of physical channels	832; 2 groups of 416 channels, each group has 21 signaling channels and 395 traffic or voice channels
Bandwidth per channel	30 kHz
Cell radius	2–20 km
Base-to-mobile frequency	869–894 MHz (downlink)
Mobile-to-base frequency	824–849 MHz (uplink) 45 MHz between transmit and receive channels
Channel spacing	30 kHz
Modulation	FM with peak frequency deviation of ±12 kHz Signaling channel uses FSK Can send data at 10 kbit/s
Access method	FDMA
basestation ERP	100 W per channel (maximum)
Channel coding	None
RF specifications of mobile unit	
Transmit RF power	3 W maximum (33 dBm) (600 mW for hand-held)
Transmit power control	10 steps of 4 dB attenuation each, minimum power is −4 dBm
Receive sensitivity	−116 dBm
Receive noise figure	6 dB measured at antenna terminals
Receive spurious response	−60 dB from center of the passband

generation systems handled analog 3 kHz voice transmissions with very limited ability to transmit digital information limited to signaling.

5.8.2 2G, Second Generation: Digital Radio

The second generation (2G) of cellular radio is characterized by digitization. Many different types of 2G digital systems were installed around the world. The 2G systems can transmit data and voice at rates of 8–14.4 kbit/s. This can be contrasted to the wireline phone system where, once signals reach the exchange, the 3 kHz analog signals are sampled at 64 kbit/s to achieve an undistorted signal representation. These cellular systems sacrifice some voice quality but use reasonably sophisticated algorithms that use the characteristics of speech to achieve greater than a factor of four compression.

In North America the first digital system introduced was the **digital advanced mobile phone system** (**DAMPS**) (originally known as **North American Digital Cellular** [**NADC**] and as the EIA/TIA interim standard **IS-54**). The system was designed to provide a transition from the then current 1G analog system to a fully digital system by reusing existing spectrum. The idea was that system providers could allocate a few of their channels for digital radio out of the total available. As analog radio was phased out, more of the channels could be committed to digital radio. The main motivation behind this system is that it provided three to five times the capacity of the analog system for the same bandwidth. The 2G GSM system provides a similar increase in capacity, and is compatible with the **Integrated Services Digital Network** (**ISDN**) which was the protocol used with the wired telephone system. The GSM system was initially (early 1990s) dominant in Europe and had the advantage that it did not need to coexist

Table 5-5: Attributes of the GSM system. Uplink and downlink frequencies are for the GSM-900 implementation, see Table 5-6 for other GSM implementations. Slow frequency hoping improves robustness.

Property	Attribute
Number of channels	125 (for GSM-900)
Bandwidth per channel	200 kHz
Channel spacing	200 kHz
Cell radius	2–20 km
Base-to-mobile frequency	935–960 MHz (GSM-900)
Mobile-to-base frequency	890–915 MHz (GSM-900)
Modulation	GMSK, Slow frequency hopping (217 hops/s)
Access method	TDMA, 8 slots per frame, user has one slot, each frame is 4.615 ms and each slot is 577 μs. There are 148 data bits and 8.25 guard bits in a slot.
Symbol duration	3.6828 μs
Transmit rate	270.833 kbit/s

Table 5-6: GSM frequency bands. GSM channels have a bandwidth of 200 kHz. The base-to-mobile transmission is the downlink and the mobile-to-basestation transmission is the uplink. GSM-900 and GSM-1800 are used in most of the world.

Band	Uplink (MHz)	Downlink (MHz)	Duplex spacing (MHz)
GSM-900	890–915	935–960	45
GSM-1800	1710–1785	1805–1880	95
GSM-900 extended	876–915	921–960	45
PCS-1900	1850–1910	1930–1990	80
GSM-850 (Americas)	824–849	869–894	45
GSM-450	450.4–457.6	460.4–467.6	10
GSM-480 (Nordic, Eastern Europe, Russia)	478.8–486	488.8–496	10

with uncoordinated 1G analog phone systems. The attributes of the GSM system are shown in Tables 5-5 and 5-6.

From an RF design perspective, the main differences between analog and digital standards are

1. The RF envelope. In AMPS, FM was used, which produces a constant envelope RF signal. Consequently, high-efficiency saturation mode amplifiers (such as Class C) can be used. In most digital modulation schemes the modulation results in a nonconstant envelope. This is true for PSK modulation, as the transition from one symbol to another does not follow a circle on the constellation diagram. The information contained in the amplitude of the RF signal is just as important as the information contained in the phase or frequency of the signal. Consequently, with digital radio saturation mode amplifiers that severely distort the amplitude characteristic must be avoided.
2. Bursty RF. In an analog system, RF power is continually being transmitted. In a digital system, transmission is intermittent and the RF signal is bursty. Therefore an RF designer must be concerned about turn-on transients and thermal stability of the power amplifier.

5.9 3G, Third Generation: Code Division Multiple Acces (CDMA)

Originally there were multiple 3G cellular radio systems but the one that became dominant uses code division multiple access. This section begins with a precursor to 3G and which is now called 2.5G cellular radio and this is relatively narrowband, 1.23 MHz-wide, CDMA. The 3G system uses 5 MHz-wide wideband CDMA (WCDMA).

5.9.1 Generation 2.5: Direct Sequence Code Division Multiple Access

CDMA, or more specifically CDMAOne, was initially promoted as being third generation, but the definition now is that data rates of at least 2 Mbit/s must be supported in 3G. Thus CDMAOne is now referred to as 2.5G. A depiction of spread spectrum is shown in Figure 5-14, in which a very fast code is superimposed on a slower data sequence and the combined code is used to modulate a carrier. The same fast code is used to extract the baseband signal from the received bitstream. The effect of the fast code is to greatly spread out the baseband signal, transforming perhaps a 12 kbit/s baseband bitstream into an RF signal with a bandwidth of 1.23 MHz.

The key feature of the DS-CDMA system is the use of lengthy codes to spread the spectrum of the signal that is to be transmitted. In the case of voice, an 8 kbit/s bitstream, for example, with error correction coding becomes a 12.5 kbit/s baseband bitstream that is mixed with a much faster code that is unique to a particular user. Thus the 8 kbit/s bitstream becomes a 1.23 MHz-wide analog baseband signal. This signal is then modulated up to RF and transmitted. On the receiver side, the demodulated RF signal can only be decoded using the original fast code. Use of the original code to decode the signal rejects virtually all interference and noise in the received signal. Despreading distributes the signal and noise components differently. Upon despreading, the noise is still distributed uniformly in frequency while the information-bearing signal is concentrated in a narrow bandwidth, the bandwidth of the baseband signal. Tremendous processing gain is available using this spreading and despreading approach.

The mechanism that increases SNR in DS-CDMA is shown in Figure 5-12. The SNR is enhanced by grouping the signal energy in a narrower bandwidth and the noise is reduced, as only the noise in a narrower

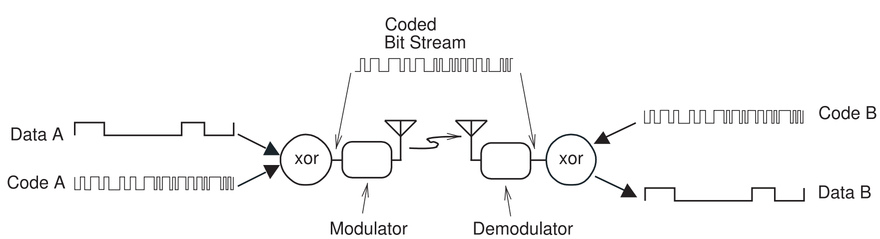


Figure 5-14: Depiction of direct sequence CDMA transmission. If code B is code A, then data B is the same as data A with some corruption by noise.

Table 5-7: Attributes of the CDMAOne system.

Property	Attribute
Bandwidth per channel	1.23 MHz
Channel spacing	1.25 MHz (20 kHz guard band)
Cell radius	2–20 km
Base-to-mobile frequency	869–894 MHz
Mobile-to-base frequency	824–849 MHz 45 MHz between transmit and receive channels.
Modulation	OPSK
Access method	CDMA 64 radio channels per physical channel Forward: 55 traffic channels, 7 paging channels 1 pilot channel 1 sync channel Reverse: 55 traffic channels, 9 access channels
RF specifications of mobile unit	
Transmit power control	1 dB power control

bandwidth is important. In despreading a signal with a single-tone interferer, as seen in Figure 5-12(c), the interferer is spread as noise while the signal energy is concentrated in the bandwidth of the baseband signal. Since orthogonal codes are used, many radio channels can be supported on the same radio link. CDMA can support approximately 120 radio channels on the same physical channel. Another important feature is that the same 120 channels can be reused in adjacent cells, as the information bitstream of each user can still be extracted. Thus there is no need for clustering as in the 2G systems. The attributes of the cellular 2.5G CDMAOne system, an immediate precursor to the 3G WCDMA system, are given in Table 5-7.

5.9.2 Multipath and Rake Receivers

In a line-of-sight CDMA system the code must be aligned with the data stream received so that the stream can be decoded correctly to reveal the original data. If there are multiple paths then the paths will, in general, have different delays and often the delay differences are more than the chip duration (the time allocated to transition from one symbol to another). The CDMA signal is then said to be transmitted over a **dispersive multipath channel**. This introduces complexity in aligning codes. The problem that arises is shown in Figure 5-15. The original data are shown in (a) and this

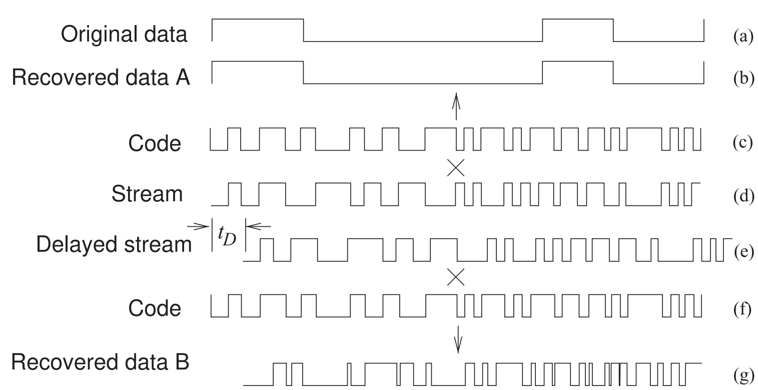


Figure 5-15: Decoding of a direct sequence coded stream showing a code aligned with the data stream and a data stream delayed by t_D.

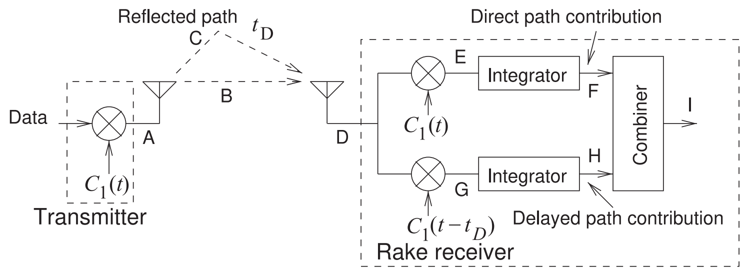


Figure 5-16: Rake receiver used to decode two paths, a direct path and a path with excess delay t_D. The integrators eliminate incorrectly recovered data.

is coded using the code in (c) yielding the stream in (d). The stream is just the exclusive or (XOR) of the data and code. On receiving, the coded data stream, (d), is despread using the original code (c). If the code and the data stream are aligned, then an XOR operation results in the data being recovered correctly by XORing (c) and (d) to obtain (b), the original data A. However, if the stream is delayed, as shown in (e), then XORing the delayed stream and the code (which is not delayed) will result in recovered data B, (g), which has no relationship to the original data in (a). The solution to this problem is to appropriately delay the despreading code for each propagation-delay path.

When there are reflections of the transmitted signal, versions of the signal followig different paths will arrive at the receiver with different delays. The **rake receiver**, introduced in the 1950s, recovers the data from each propagation delay path and combines the signal from each path to produce a combined signal with a higher SNR than that obtained using just the line-of-sight component of the signal [20–23].

A two-path version of the rake receiver is shown in Figure 5-16. The transmitter spreads the data using a code $C_1(t)$ to produce a stream at A that is transmitted over two paths C and B. Path C is delayed with respect to path B by a time t_D and the signal at the receive antenna, D, is a combination of the signal following the two paths. On the top path of the receiver, called a finger, the direct signal (that followed path B) is despread by the original code $C_1(t)$ and, if it is appropriately aligned with the signal that followed path B, it produces a signal at E that will contain the low-frequency user data plus a high-frequency component resulting from the wrong code being used to despread the delayed data stream that followed path C. The signal at E is lowpass filtered to produce the original data at F as the high frequency component is eliminated. Since the operation is implemented in DSP, an integrator is used over the duration of a data bit. The combined received signal, D, is despread using a delayed code, $C_1(t - t_D)$, to produce a signal at G that is integrated so that the data at H is the original D due to the component that followed path C. The signals at F and H, sampled after the duration of a bit, will both contain the original data plus some noise and the noise at the two points will be uncorrelated. When F and H are combined, the signal will combine coherently while the noise will combine incoherently, thus improving the SNR.

The rake receiver can be generalized to have many fingers. After despreading in each finger of the rake receiver, each delayed component is demodulated and the results are combined. The rake receiver is so named because each finger sweeps up information, resembling the tines on a garden rake collecting leaves. Since each component contains the original information, if the magnitude and time of arrival (phase) of each component is computed at the receiver (through a process called channel estimation),

Table 5-8: Attributes of the 3G WCDMA system. 3G operates in many frequency bands and just one representative band, Band 3, is considered here. The uplink modulation in WCDMA is unusual applying data to the I channel of an I/Q modulator and control data to the Q channel. Here the I channel (and Q channel) has two or four levels classified as **pulsed amplitude modulation** (**PAM**). † per user.
∗ shared, non-MIMO. ‡ shared, 2-cell carrier aggregation, MIMO. The 3G standards do not specify minimum information transmit rates per user.

Property	Attribute
Number of channels	
Bandwidth per channel	5 MHz
Channel spacing	5 MHz
Cell radius	2–20 km
Downlink frequency	1805–1880
Uplink frequency	1710–1785
Modulation (downlink)	QPSK, 16QAM, 64QAM
Modulation (uplink)	PAM on I and on Q
Access method	CDMA
Symbol rate	3.84 Msymbols/s
Modulation bandwidth	3.84 MHz
Symbol duration	260 ns
Information rate (original)†	200 kbit/s up- & down-link
Information rate (HSPA+)∗	40 Mbit/s downlink
Information rate (HSPA+)‡	168 Mbit/s downlink
Information rate (HSPA+)†	10 Mbit/s uplink

then all the components can be added coherently to improve the information reliability. This method of combining is called **maximal-ratio combining** (**MRC**), a method of **receiver diversity** combining in which

1. the signals from each channel are added together,
2. the gain of each channel is made proportional to the rms signal level and inversely proportional to the mean square noise level in that channel, and
3. different proportionality constants are used for each channel.

It is also known as **ratio-squared combining** and **predetection combining**. MRC is optimum combining for independent AWGN channels.

CDMA, WCDMA, and WLAN (WiFi) units use multipath signals and combine them to increase the SNR at the receivers. In contrast, the narrowband 2G systems cannot discriminate between the multipath arrivals, and multipath has negative impact. The rake receiver in Figure 5-16 has two fingers, but the practical limit on the number is based on the expected delay spread of the multipaths divided by the chip duration. Only delays that are an integer multiple of a chip duration are needed in a rake receiver [21, 22]. A trade-off must be made in terms of the amount of circuitry and the DC power available. In the original CDMA cell phone system (circa 2000), three fingers were used in handsets and four to five in basestations. Many more are used today.

5.9.3 3G, Wideband CDMA

Third-generation radio, (**3G**), is coordinated by the **Third Generation Partnership Project** (**3GPP**). This is a collaborative agreement of standards development organizations and other related bodies for the production of a complete set of globally applicable technical specifications for mobile communication systems (see http://www.3gpp.org). Initially the efforts of 3GPP were directed at establishing the 3G standards but the scope has expanded and it develops evolving standards for the transition to 4G and now 5G systems. Some of the attributes of 3G are given in Table 5-8.

The 3G mobile systems support variable data rates depending on demand and the level of mobility. Switched-packet radio techniques are required to support this bandwidth-on-demand environment. Here the physical channel is shared (i.e., **packet switched**) rather than the user being assigned a physical channel for exclusive use (referred to as **circuit switched**).

The drive for 3G systems was partly fueled by the saturation of 2G systems in many places and a desire to increase revenues by supporting high-speed data. Prior to the rollout of 3G systems, the increased demand primarily resulted from an increased consumer base rather than the emergence of significant data traffic. The increased subscriber base was addressed by 2.5G systems, which have some of the 3G concepts but only partially implemented. The driving concept of 3G was the development of a standard that supports high-speed data, global roaming, and supports advanced features including two-way motion video and internet browsing.

Third-generation cellular radio is defined by the **International Telecommunications Union** (ITU) [24] and is formally called International Mobile Telecommunications 2000 (**IMT-2000**). The basic requirements are for a system that supports data rates up to 2 Mbit/s in fixed environments ranging down to 144 kbit/s in wide area mobile environments. In 1999 the ITU adopted five radio interfaces for IMT-2000:

1. IMT-DS direct-sequence CDMA, more commonly known as WCDMA;
2. IMT-MC multi-carrier CDMA, more commonly known as CDMA2000, the successor to CDMAOne (specifically international standard IS-95);
3. IMT-SC time-division CDMA, which includes time division CDMA (TD-CDMA) and time division synchronous CDMA (TD-SCDMA);
4. IMT-SC single carrier, more commonly known as EDGE; and
5. IMT-FT frequency time, more commonly known as DECT.

The dominant choice for 3G is WCDMA. In October 2007 the ITU Radiocommunication Assembly included WiMAX-derived technology, specifically **orthogonal frequency division multiple access** (**OFDMA**, see Figure 5-8(d)) and MIMO, in the set of IMT-2000 standards as the sixth radio interface. 3GPP [25] provides a migration strategy for cellular communications through a process called **long-term evolution LTE** and through a number of releases each building on prior infrastructure and adding capabilities.

EDGE has intermediate data speeds between those of GSM and WCDMA. The following terms are also used to describe networks using the 3G WCDMA specification: Universal Mobile Telecommunication System (**UMTS**) (UMTS, in Europe), UMTS Terrestrial Radio Access Network (UTRAN), and **Freedom of Mobile Multimedia Access** (**FOMA**, in Japan). UMTS is the 3G successor of the GSM standard, with the air interface now using WCDMA. The terminology used in UMTS, listed in part in Table 5-9, is based on the terminology used in GSM, with subtle differences. UMTS was first deployed in Japan in 2001. The term WCDMA describes the physical interface and protocols that support it, while UMTS refers to the whole network. A large number of frequency bands are designated for 3G, see Tables 5-10 and 5-11.

The 3GPP timeline is summarized in Figure 5-17. The CDMA2000 and WCDMA paths become the single LTE path beyond 3G. The CDMA2000 (the IS-2000 standard) path builds on the original CDMA system defined by the IS-95 standard and commonly known as CDMAOne (the **IS-95** standard). CDMA2000 1xEV-DO is the first evolution of CDMA2000 that meets the ITU basic specification for 3G. **Evolution-data optimized** (**EV-DO**), combines CDMA and TDMA for higher data throughput.

Table 5-9: UMTS terminology.

Term	Description
AuC	Authentication center
GGSN	Gateway GPRS support node
GMSC	Gateway MSC
HLR	Home location register
ISDN	Integrated services digital network
MSC	Mobile switching center
Node B	Basestation
PSTN	Public switched telephone network
RNC	Radio network controller
SGSN	Serving GPRS support node
UE	User equipment
USIM	Universal subscriber identity module
VLR	Visitor location register

Table 5-10: Spectrum assignments for 3G, [25, Release 99].

Band	Uplink (MHz)	Downlink (MHz)	Available spectrum					
			NA	LA	EMEA	ASIA	Oceania	Japan
1	1920–1980	2110–2170		■	■	■	■	■
2	1850–1910	1930–1990	■	■		■	■	■
3	1710–1785	1805–1880		■	■			
4	1710–1755	2110–2155	■		■			
5	824–849	869–894	■	■	■	■	■	■
6	830–840	875–885						■
7	2500–2570	2620–2690						
8	880–915	925–960		■	■	■	■	■
9	1749.9–1784.9	1844.9–1879.9						■
10	1710–1770	2110–2170						
11	1427.9–1447.9	1475.9–1495.9						■
12	698–716	728–746	■					
13	777–787	746–756	■					
14	788–798	758–768	■					
15–18	reserved							
19	832.4–842.6	877.4–887.6						■
20	832–862	791–821			■			

The WCDMA and LTE evolution is defined by releases beginning with an initial release in 2000 known as Release 99 (Rel-99) [26]. This saw the beginning of the Third Generation Partnership Project (3GPP) specifying and controlling the evolution of cellular communications through 3G, 4G, and now 5G. The releases are designed to protect the installed investment in cellular systems while providing a migration path.

5.9.4 Summary

WCDMA 3G and 4G are widely deployed. While 3G/WCDMA it will be gradually replaced by 4G and 5G it has particular aspects that could see the system remain for many years.

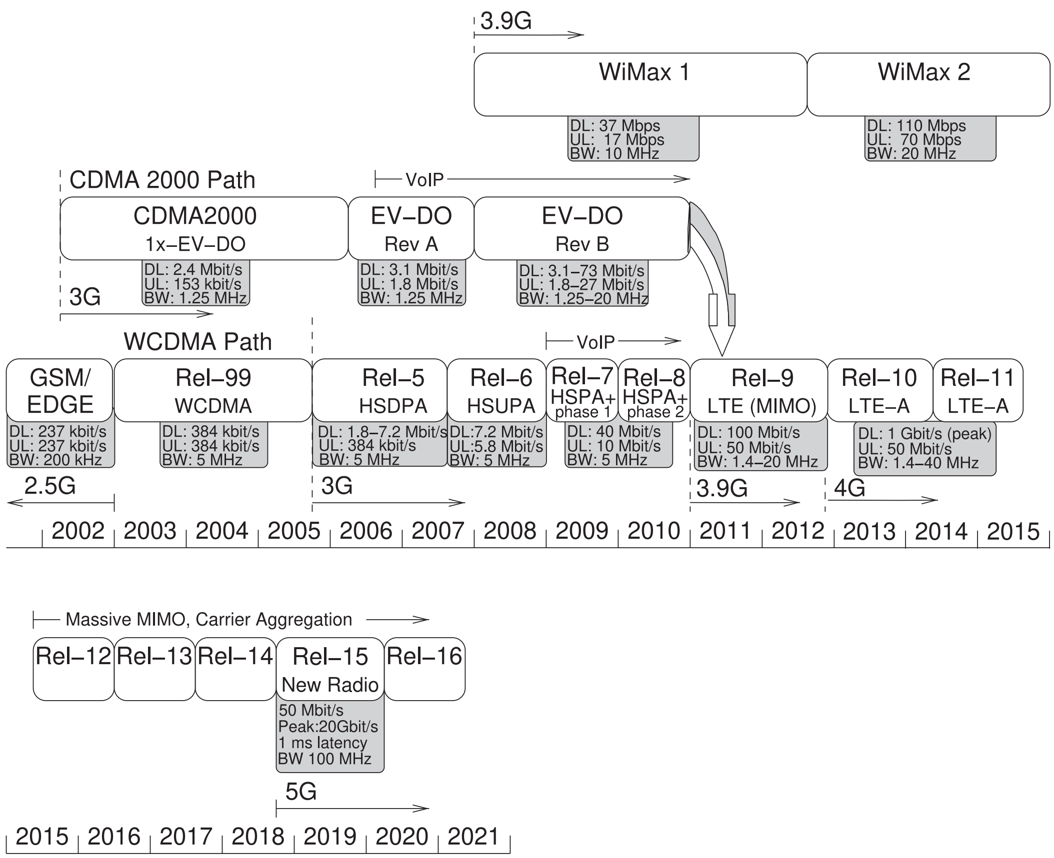


Figure 5-17: Timeline for implementation of 3G, 4G and 5G. DL indicates the downlink data rate; UL indicates the uplink data rate; BW indicates the channel bandwidth. Development supports Internet protocol (IP) and voice over IP (VoIP). The two 3G paths become a single long term evolution (LTE) path. LTE is the concept of smoothly evolving through 4G and into 5G utilizing existing infrastructure and adding capability [25]). Begining with Release 99 (Rel-99) the timeline is controlled by 3GPP . The dates refer to the first significnat commercial availability of the standard and the official release dates can be found at http://www.3gpp.org.

Band	Uplink (MHz)	Downlink (MHz)
33	1900–1920	1900-1920
34	2010–2025	2010–2025
35	1850–1910	1850–1920
36	1930–1990	1930–1990
37	1910–1930	1910–1930
38	2570–2620	2570–2620
39	1880–1920	1880–1920
40	2300–2400	2300–2400

Table 5-11: Spectrum assignments for TDD 3GPP [25, Release 8].

5.10 4G, Fourth Generation Radio

The 4G cellular system provides downlink data rates of 100 Mbit/s while mobile and 1 Gbit/s while stationery. The uplink data rates are much lower at 50 Mbit/s. The maximum rates are assigned to a cellular service area and are shared but rely on a user's average data rate being much lower. Providing these data rates relies on a number of significant advances over 3G:

- Orthogonal frequency division multiplexing (OFDM). Using a very large number of narrowband channels, e.g. 1200, with each narrowband channel having the highest-order modulation possible. Each narrowband channel has its own subcarrier which is individually modulated. Gets around the multipath problem where poor channel characteristics affecting a narrow range of frequencies does not limit overall bit rates.
- High-order modulation. A terminal unit supports a very large number of modulation formats, several PSK modulation methods and QAM methods ranging from 16-QAM to 1024-QAM. Modulation and demodulation are implemented in a DSP unit.
- Cyclic prefix (CP). Enables the use of efficient discrete Fourier transforms in the transmitter. Simplifies hardware by yielding lower PMEPR than would be obtained if a zero-level guard band was provided to overcome interference caused by multipath.
- Multiple input, multiple output (MIMO). MIMO uses multiple antennas to transmit and receive signals exploiting independent transmit-receive paths to increase the total capacity between a basestation and terminal unit. Multiplies data rates by the number of transmit or receive antennas (which ever is less).
- Pervasive access. Terminal units support many access technologies (e.g., 2G–4G, WiFi, Bluetooth etc.) and seamlessly switching among them. Since 4G is an IP-based system data can be sent on any network, e.g. WiFi utilization reduces the demands on the cellular network.

These advances require tremendous computing power made available by advanced low-power VLSI and developments in low-power RF modems. Another critical aspect of 4G (and now 5G) is deployment of a well thought out road map called Long Term Evolution (LTE) that enables incremental (although significant) advances building on previously implemented enhancements.

The LTE timeline is shown in Figure 5-17, 4G effectively began with Release-9 implemented first at the end of 2010 with this release introducing MIMO and OFDM (to be defined latter). There are fundamental changes to the physical interface, and 4G uses Internet Protocal (IP) exclusively including for voice (voice over IP—VoIP). Subsequent releases introduced more concepts and increased data coverage. 4G and now 5G can not coexist with 3G systems and separate bands must be allocated. Eventually 3G will go away. It is worth noting that 5G is upwards compatible with 4G.

The ITU World Radiocommunication Conference has allocated many bands to 4G ranging from 450 MHz to 5850 MHz. There is not a single band that can be used world-wide but cellular handsets are designed to support

Table 5-12: Selected LTE bands used in 4G.

Band	Mode	Uplink (MHz)	Downlink (MHz)	Bandwidths (MHz)	Regions
3	FDD	1710–1755	1805–1880	1.4, 3, 5, 10, 15, 20	Asia, Parts of Africa, Europe, Parts of Latin America, Oceania
12	FDD	699–716	729–746	1.4, 3, 5, 10	North America, Asia, Parts of Africa, Europe, Parts of Latin America
40	TDD	2300–2400		5, 10, 15, 20	Asia, Parts of Europe, Oceania
41	TDD	2496–2690		5, 10, 15, 20	Parts of Africa, Parts of Asia, USA

Bandwidth	Resource blocks	Subcarriers (downlink)	Subcarriers (uplink)
1.4 MHz	6	73	72
3 MHz	15	181	180
5 MHz	25	301	300
10 MHz	50	601	600
15 MHz	75	901	900
20 MHz	100	1201	1200

Table 5-13: Resources for different LTE bandwidths.

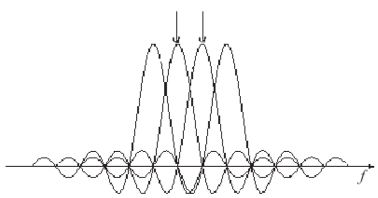


Figure 5-18: OFDM spectrum with four subcarriers showing orthogonality.

multiple bands. The features of representative bands are shown in Table 5-12. Channel bandwidth, with a channel having a single carrier, range from 1.4 MHz to 20 MHz, see Table 5-13. Each carrier has correlated subcarriers each modulated to produce a subchannel with a 15 kHz-wide bandwidth.

5.10.1 Orthogonal Frequency Division Multiplexing

In OFDM, data is divided into many bitstreams with each bitstream modulating a **subcarrier** producing a relatively narrowband subchannel and a user being assigned multiple subchannels. In 4G these subchannels are 15 kHz-wide in normal operation and 7.5 kHz wide in a rich multipath environment with a large excess delay spread as described in Section 4.7. The concept behind this is that having a narrow bandwidth, the duration of a symbol is long and will exceed the excess delay spread. This minimizes the impact of intersymbol interference (ISI) when a symbol traveling on a fast path overlaps with the component of a previous symbol traveling on a longer path. Also each subchannel can be modulated with the maximum order modulation enabled by that particular channel's characteristics. For this to work the subcarriers must be orthogonal and precisely spaced in frequency and time. The orthogonality is shown in the spectrum of Figure 5-18, where the arrows at the top indicate sampling points for two subcarriers. The orthogonality of the subcarriers is seen by noticing that the peak of one subcarrier is at the zeros of the other subcarriers. When one subcarrier is sampled, the contribution from all other carriers is zero; they are orthogonal.

The spectra of the subcarriers overlap, but this does not matter.

In 4G groups of 12 subcarriers lasting 0.5 ms are grouped in a resource block which is the minimal allocation to a user. In communications with a handset, a basestation (called a B node in 4G and 5G) allocates a specific number of subcarriers. These may not be contiguous on the downlink, as shown in Figure 5-8(d), but are contiguous on the uplink. The number of subcarriers allocated to a particular user varies according to the data rate requirements and the order of modulation used with each subcarrier being adjusted to accommodate the subchannel characteristics including fading and interference.

In communicating to a single user, the total bitstream is error encoded and then the bitstream is divided up so that different parts of the bitstream are sent over multiple subchannels thus spreading the coded data. As a result 4G implements spectrum spreading but at a much lower coding rate than the spreading in 3G. This further mitigates the impact of multipath which can affect the integrity of individual subchannels. Signal strength and interference, and hence SIR, can differ for each channel and this is compensated for by having different bit rates in each subchannel and spreading the data plus error correction coding.

A second effect of multipath is referred to as delay spread. Transmitted signal components following different paths arrive at a receiver at different times. If the time duration of a symbol is short, which is the case if wideband modulation is used, then the components of a signal, corresponding to one symbol, following longer paths could arrive at the same time as the components of the next symbol following shorter paths. This causes intersymbol interference. With the longer symbol duration with OFDM and a guard band interval (which is now relatively short) between symbols, intersymbol interference can be greatly reduced.

Implementation

OFDM could be implemented by using separate modulators and demodulators for each subcarrier. Instead the separate modulators and demodulators are replaced by a **fast Fourier transform** (**FFT**) and an **inverse FFT** (**iFFT**), respectively, implemented in a DSP unit called the baseband processor. In the transmitter the iFFT produces a modulated signal at a low intermediate carrier frequency. For example, a 1.4 MHz bandwidth OFDM signal could be a DSB-SC signal with a center frequency, i.e. IF carrier frequency, of 700 kHz but otherwise look identical to the final transmitted OFDM signal. Then a SSB-SC up-converter translates the OFDM signal to a radio frequency signal. Each of the time-varying frequency inputs of the iFFT is centered at a subcarrier frequency. Each of the frequency inputs is a sequence of time-samples of a slowly varying modulated signal implementing the selected modulation method for that subchannel. On receive the input of the FFT is the sequence of time samples of the modulated signal centered on the IF carrier which has been down-converted from RF. Each of the outputs of the FFT correspond to a time sample of an individual modulated subcarrier.

Since the OFDM signal combines many individually modulated subcarriers the PMEPR of OFDM is large but grows relatively slowly with the number of subcarriers. The higher PMEPR, compared to a single modulated carrier, introduces microwave design challenges especially for mixers and am-

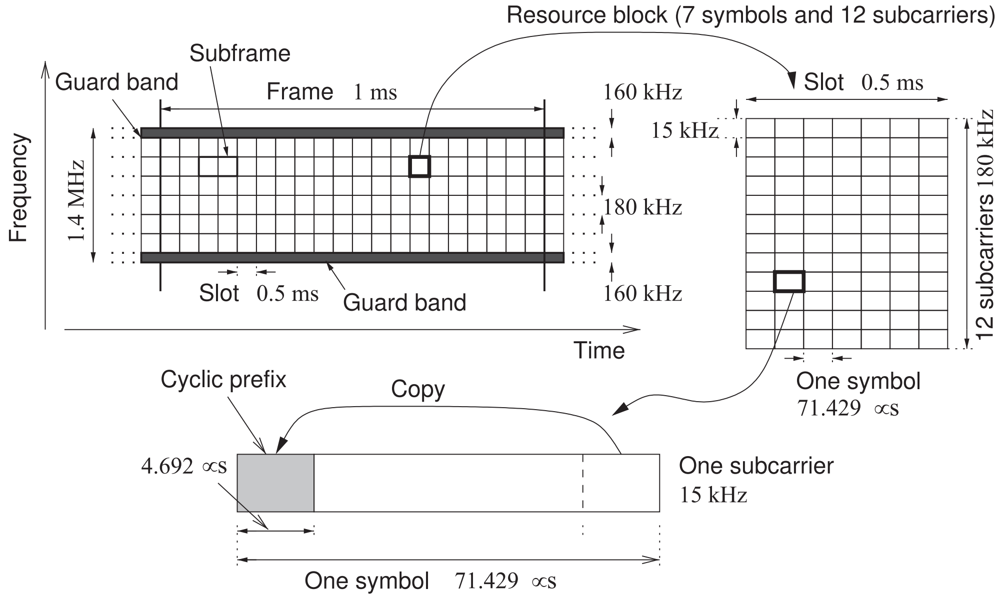


Figure 5-19: Components of a frame in 4G for a 1.4 MHz-bandwidth channel. The 10 ms-long frame, top left, has an array of resource blocks each with a slot duration of 0.5 ms and a bandwidth of 180 kHz. This frame has a 1.08 MHz data bandwidth with 160 kHz guard bands for a total bandwidth of 1.4 MHz. (The guard-band differs for channels having other bandwidths. Each resource block has on seven OFDM symbols with each comprised of seven symbols during a symbol interval. There are 84 symbols (Seven OFDM symbols) per resource block (with 1024-QAM having 10 bits per subcarrier 840 bits are transmitted per resource block).

plifiers. However, since most of the heavy lifting is done in DSP which being numerical is distortion-free, the up-converter is simplified .

Resource Blocks

The way that OFDM is implemented in 4G is illustrated in Figure 5-19 for a 1.4 MHz-bandwidth modulated signal. The subchannels in 4G are organized in a frame that is 10 ms long. Within the frame subchannels are grouped into **resource blocks** each of which has a duration of 0.5 ms and a bandwidth of 180 kHz. The resource block is the smallest unit allocated to a user and a user can have many simultaneous resource blocks and many sequential resource blocks. A resource black has many subcarriers and symbol intervals and the exact length and bandwidth of a single symbol-subcarrier segment has a normal mode used most of the time, and two extended modes that handle severe multipath environments. In the normal mode, each resource block has seven 71.4 µs-long symbols per subcarrier and there are 12 subcarriers. (The extended modes have longer symbol duration and 7.5 kHz bandwidth.)

In OFDM there are subchannels that are reserved for pilot tones that send known data streams and these can be used to characterize the channel. The pilot subchannels are denoted as P in Figure 5-8(d). The dedicated pilot subchannels enable carrier recovery. Therefore it is not necessary to restrict modulation by avoiding transitions through the origin of the constellation diagram (or more precisely the phasor of the modulated subcarrier). Thus near-ideal modulation efficiency is possible for the data subcarriers.

Summary

OFDM has been tremendously successfull and only had to wait for the capabilities of low-power VLSI to advance. It is now widely used in communications including WiFi, wired communications (e.g. digital subscriber line (DSL)) as well as in 4G and 5G.

5.10.2 Orthogonal Frequency Division Multiple Access

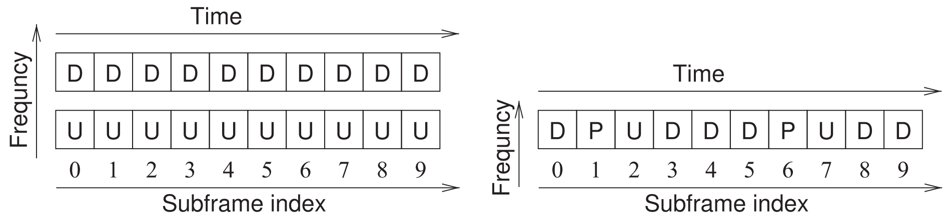
Orthogonal Frequency division multiple access (**OFDMA**), also known as **multiuser OFDM**, is the multiuser version of OFDM that supports simultaneous communication by multiple users and so is an access technology. The access scheme is illustrated in Figure 5-8(d) where one channel is illustrated and could have a bandwidth ranging from 1.4 MHz to 20 MHz. There could be 1201 subchannels so not all of these are shown. The subchannels are grouped in contiguous groups of 12 subchannels to form a 1 ms-long resource block (24 7.5 kHz-wide subchannels in extended mode used with a very rich multipath environment). So what are numbered in Figure 5-8(d) are resource blocks. In OFDMA, generally one or more resource blocks are assigned to a particular user and these do not need to be contiguous in the downlink but are contiguous in the uplink as then the PMEPR is lower

This lessens the demand on the power amplifier in the terminal unit. Effectively there is a single carrier on the uplink and hence the uplink access scheme is also called **single-carrier frequency division multiple access** (**SC-FDMA**) or **single-carrier orthogonal frequency division multiple access** (**SC-OFDMA**). On the downlink the basestation is using all of the subchannels simultaneously communicating with multiple terminal units but in the uplink each terminal unit only uses a few. Users share pilot subchannels, designated as P in Figure 5-8(d).

5.10.3 Cyclic Prefix

The cyclic prefix (**CP**) refers to the scheme used to accommodate excess delay spread resulting from multipath. Since symbols sent on different paths arrive at the receive antenna with various delays, there can be overlap of a symbol arriving on a fast path with a previous symbol arriving on a longer path. This could result in intersymbol interference (ISI). In 4G there could have been a zeroed (i.e. no signal) guard interval to eliminate ISI. Instead, in 4G, a cyclic prefix (**CP**) is used whereby a symbol is prefixed repeating the end part of a symbol. One symbol of a resource block is shown at the bottom of Figure 5-19. Each symbol of the resource block is 66.7 μs long with a cylic prefix which normally is 4.7 μs long but longer for the first symbol of the resource block and when in extended mode which is used in severe multipath environments.[5] The 4.7 μs long CP corresponds to a

[5] In the normal mode the first CP is 5.2 μs long and there are two extended modes. One is 16.7 μs long and another is 33.3 μs long and then the bandwidth of each subchannel is reduced to 7.5 kHz. In normal mode the overhead is 7% and in extended mode the CP overhead is 25%. If the first extended CP is used then there are 6 OFDM symbols per resource block.



(a) Frequency division duplex (FDD) (b) Time division duplex TDD

Figure 5-20: Duplex schemes used in 4G with ten subframes in a 10 ms frame.

difference in path lengths of 1.4 km. This cyclic prefix provides the required guard interval. There are several advantages to copying the end of a symbol and repeating it before the symbol. This reduces the PMEPR that would result if the guard interval had no signal, and it enables a discrete Fourier transform to be used in processing the signal. This is essential to efficient VLSI implementation of OFDM.

The repetition of the cyclic prefix for each symbol in a resource block is equivalent to repeating the end of each OFDM symbol at the start of an OFDM symbol (the aggregate of symbols across all 12 subcarriers constitutes an OFDM symbol). The repetition enables the start of each OFDM symbol to be determined as the CP is correlated to the end of the symbol. The cyclic prefix also helps in characterizing the channel and removing distortion within the OFDM symbol

5.10.4 FDD versus TDD

Frequency division duplex (FDD) and time division duplex (TDD) are two duplex schemes supported in 4G although only one mode or the other is supported in a particular band, e.g. see Table 5-12. FDD and TDD are arranged in frames of 10 ms comprising ten subframes of 1 ms duration, see Figure 5-20. In FDD, Figure 5-20(a), there are separate uplink, denoted U, and downlink, denoted D, frequency bands and uplink and downlink transmissions are simultaneous. The use of paired spectrum requires a good diplexer to isolate the receiver and transmitter. In TDD, Figure 5-20(b), uplink and downlink use the same band and there are separate uplink, denoted U, and downlink, denoted D, transmissions in different frequency bands. In TDD, uplink and downlink transmissions are simultaneous. Channel propagation is the same in both directions at least over 10 ms as long as mobility is not too high. TDD allows dynamic allocation of uplink and downlink capacity.

5.10.5 Multiple Input, Multiple Output

Multiple input, multiple output (**MIMO**, pronounced my-moe) technology uses multiple antennas to transmit and receive signals. The MIMO concept was developed in the 1990s [27, 28] and implemented in 4G and 5G, and several WLAN systems. There are several aspects to MIMO. First, each transmit antenna sends different data streams simultaneously on the same frequency channel as other transmit antennas. The most interesting feature

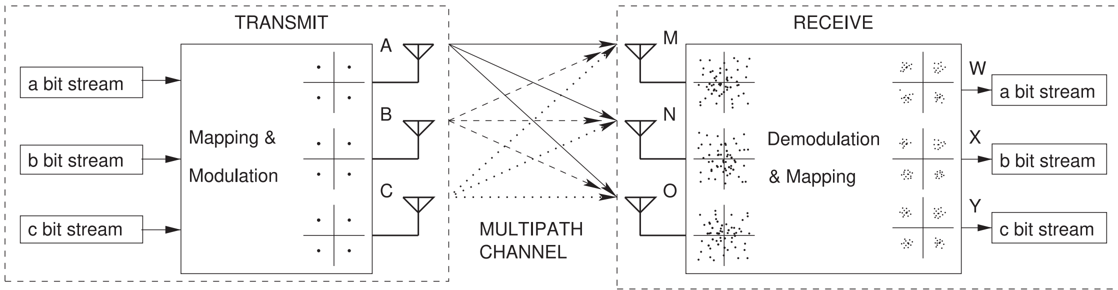


Figure 5-21: A MIMO system showing multiple paths between each transmit antenna and each receive antenna.

is that MIMO relies on signals traveling on multiple paths between an array of transmit antennas and an array of receive antennas. In a conventional communications system the various paths result in interference and fading, but in MIMO these paths are used to carry more information. In a MIMO system, each path propagates an image of one transmitted signal (from one antenna) that differs in both amplitude and phase from the images following other paths. Effectively there are multiple connections between each transmit antenna and each receive antenna, see Figure 5-21. For simplicity, three transmit antennas and three receive antennas are shown. However, MIMO can work with as few as two transmit antennas and two receive antennas. In MIMO a high-speed data-stream is split into several slower data streams, shown in Figure 5-21 as the a, b, and c bitstreams. The distinct bitstreams are separately modulated and sent from their own transmit antenna, with the constellation diagrams of the transmitted modulated signals labeled A, B, and C. The signals from each of the transmit antenna reaches all of the receive antennas by following different uncorrelated paths.

The output of each receive antenna is a linear combination of the multiple transmitted data streams, with the sampled RF phasor diagrams labeled M, N, and O. (It is not really appropriate to call these constellation diagrams.) That is, each receive antenna has a different linear combination of the multiple images. In effect, the output from each receive antenna can be thought of as the solution of linear equations, with each transmit antenna-receive antenna link corresponding to an equation. Continuing the analogy, the signal from each transmit antenna represents a variable. So a set of simultaneous equations can be solved to obtain the original bitstreams. This is accomplished by demodulation and mapping using knowledge of the channel characteristics to yield the original transmitted signals modified by interference. The result is that the constellation diagrams W, X, and Y are obtained. The composite channel can be characterized using known test signals. Special coding called **space-time** (or **spatio-temporal**) coding embedded in the transmitted data-stream also enables estimation of the communication matrix. Space-time coding encodes each transmitted data-stream with information that can be used to update the channel characterization.

The capacity of a MIMO system with high SIR scales approximately linearly with the minimum of M and N, $\min(M, N)$, where M is the number

Modulation scheme	Capacity bit/s/Hz (bits per second per hertz)				
	Non-MIMO	MIMO with $M = 2, N = 2$			
	$M = 1, N = 1$ maximum	SIR 0 dB	SIR 10 dB	SIR 20 dB	SIR 30 dB
BPSK	1	1.2	2	2	2
QPSK	2	1.6	3.7	4	4
8-PSK	3	1.6	4.8	6	6
16-PSK	4	1.6	4.9	7.5	8

Table 5-14: Capacity of MIMO schemes with PSK modulation for different received SIRs compared to the maximum capacity of a conventional non-MIMO scheme. M is the number of transmit antennas, N is the number of receive antennas. Data from [31].

of transmit antennas and N is the number of receive antennas (provided that there is a rich set of paths) [29, 30]. So a system with $M = N = 4$ will have four times the capacity of a system with just one transmit antenna and one receive antenna. Table 5-14 presents the capacity of a MIMO system with ideal PSK modulation and two transmit and two receive antennas. This is compared to the capacity of a conventional (non-MIMO) system. The capacity is presented in bits per second per hertz and it is seen that significant increases in throughput are obtained when SIR is high. MIMO is a successful way to increase capacity and is included in modern WiFi, radars, and other communication systems.

In summary, MIMO systems achieve throughput and range improvements through four gains achieved simultaneously:

1. Array gain resulting from increased average received SIR obtained by coherently combining signals. To exploit this the channel must be characterized. This increases coverage and quality of service (QoS).
2. Diversity gain obtained by presenting the receiver with multiple identical copies of a given signal. This combats fading. This also increases coverage and QoS.
3. Multiplexing gain by transmitting independent data signals from different antennas to increase throughput. This increases spectral efficiency.
4. Cochannel interference reduction. This increases cellular capacity.

5.10.6 Carrier Aggregation

Carrier aggregation (CA) is one of the main features that 4G introduced. In carrier aggregation bitstreams on different carriers are combined to yield a higher overall bit rate than one carrier can support. This enables short-term high bit rates to be transmitted to a terminal. The full concept supports combining of bit-streams with carriers that are in different frequency bands (called inter-band CA), the same frequency band (called intra-band CA), different cells, and even in a combination of licensed and unlicensed (think WiFi) bands. It is 4G LTE-A combining carrier aggregation and MIMO that achieves, and can even surpass, the peak 1Gbit/s goal of 4G. A typical goal of 4G LTE A is to combine five carriers. The 4G standards, 3GPP release 13, supports 32-carrier aggregation and up to 3 Gbit/s.

5.10.7 IEEE 802.11n

The IEEE 802.11n WiFi system is not 4G but this is discussed here to indicate that many of the concepts that are incorporated in wireless systems

Table 5-15: Data rates used in the 802.11n standard. MCS is the modulation and coding scheme index. GI is the guard interval. With a 20 MHz-wide channel there are 52 data subcarriers, and an additional 4 subcarriers, called pilots, enable carrier recovery and monitor the channel characteristics. With a 40 MHz-wide channel there are 108 data subcarriers and 6 pilot subcarriers. R_i = information bit rate, R_c = coded bit rate. R_c is the bit rate of information bits plus bits added for error correction. Thus a coding rate of 5/6 means that for every 5 information bits there is an additional (redundant) code bit.

MCS index	Spatial streams	Modulation type	Coding rate R_i/R_c	Data rate (Mbit/s)			
				20 MHz channel		40 MHz channel	
				800 ns GI	400 ns GI	800 ns GI	400 ns GI
0	1	BPSK	1/2	6.50	7.20	13.50	15.00
1	1	QPSK	1/2	13.00	14.40	27.00	30.00
2	1	QPSK	3/4	19.50	21.70	40.50	45.00
3	1	16-QAM	1/2	26.00	28.90	54.00	60.00
4	1	16-QAM	3/4	39.00	43.30	81.00	90.00
5	1	64-QAM	2/3	52.00	57.80	108.00	120.00
6	1	64-QAM	3/4	58.50	65.00	121.50	135.00
7	1	64-QAM	5/6	65.00	72.20	135.00	150.00
8	2	BPSK	1/2	13.00	14.40	27.00	30.00
9	2	QPSK	1/2	26.00	28.90	54.00	60.00
10	2	QPSK	3/4	39.00	43.30	81.00	90.00
11	2	16-QAM	1/2	52.00	57.80	108.00	120.00
12	2	16-QAM	3/4	78.00	86.70	162.00	180.00
13	2	64-QAM	2/3	104.00	115.60	216.00	240.00
14	2	64-QAM	3/4	117.00	130.00	243.00	270.00
15	2	64-QAM	5/6	130.00	144.40	270.00	300.00
16	3	BPSK	1/2	19.50	21.70	40.50	45.00
17	3	QPSK	1/2	39.00	43.30	81.00	90.00
18	3	QPSK	3/4	58.50	65.00	121.50	135.00
19	3	16-QAM	1/2	78.00	86.70	162.00	180.00
20	3	16-QAM	3/4	117.00	130.70	243.00	270.00
21	3	64-QAM	2/3	156.00	173.30	324.00	360.00
22	3	64-QAM	3/4	175.50	195.00	364.50	405.00
23	3	64-QAM	5/6	195.00	216.70	405.00	450.00
···	4	...	...	...	...	...	...
31	4	64-QAM	5/6	260.00	288.90	540.00	600.00

first appear in WiFi. IEEE 802.11n is an example of a data communication system that combines many advanced concepts. The 802.11n system uses MIMO, OFDM, and many dedicated subcarriers for continuous channel characterization and carrier recovery. The 802.11n standard uses several modulation types, as shown in Table 5-15, and it supports the use of either a 20 MHz- or 40 MHz-wide channel. There can be one, two, three or four spatial streams and these are equal to the minimum of the number of transmit and receive antennas. If the number of transmit or receive antennas is more than the number of spatial streams, the extra antennas are used for receiver or transmit diversity. Under the best conditions with high SNR, negligible correlation of the transmit-to-receive antenna paths, and four transmit and four receive antennas, the **modulation and coding scheme** (**MCS**) index 31 is used. With this index, modulation uses 64-QAM and 5/6

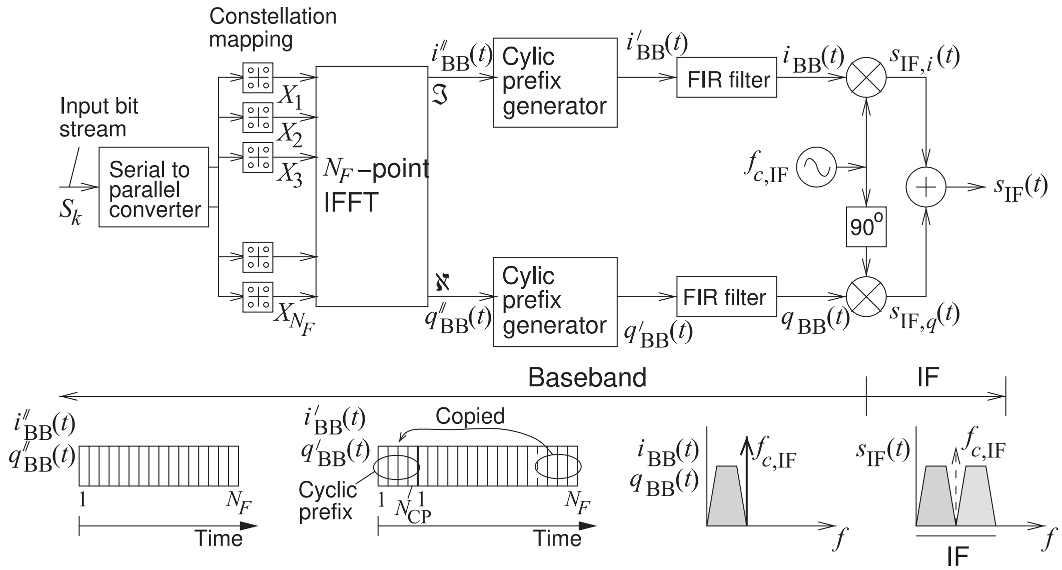


Figure 5-22: Block diagram of an OFDM modulator.

coding. This mode supports a data rate of 600 Mbit/s if a 400 ns guard interval is used. This guard interval, rather than the 800 ns guard interval, is used when there is low delay spread. If the SNR is lower, a lower-order modulation scheme is used.

5.10.8 OFDM Modulator

The basic structure of an OFDM modulator is shown in Figure 5-22. From the perspective of a single user, a bitstream S_k is divided up by a serial-to-parallel converter to produce multiple slower bitstreams each of which is mapped onto complex signals x_i which become the frequency inputs of an iFFT. The outputs of the iFFT are real and imaginary time-domain baseband signals $i'_{\mathrm{BB}}(t)$and $q'_{\mathrm{BB}}(t)$. At each symbol interval ($i'_{\mathrm{BB}}(t)$,$q'_{\mathrm{BB}}(t)$) indicates a symbol. The $i'_{\mathrm{BB}}(t)$and $q'_{\mathrm{BB}}(t)$ pass through a cyclic prefix generator that copies the end of the symbol and prefixes it to the beginning of the symbol. Then the output of the cyclic prefix generators are shaped by an FIR filter (generally a raised cosine filter) to produce the shaped I/Q signals $i_{\mathrm{BB}}(t)$and $q_{\mathrm{BB}}(t)$ which are input to a quadrature modulator with an intermediate carrier frequency to produce a DSB-SC intermediate frequency signal modulated signal $s_{\mathrm{IF}}(t)$. For example, if the frequency range of $i_{\mathrm{BB}}(t)$and $q_{\mathrm{BB}}(t)$ range from just above DC to just below 700 kHz the choice of $f_{c,\mathrm{IF}} = 700$ kHz results in a 1.4 MHz wide modulated signal $s_{\mathrm{IF}}(t)$ which is a DSB-SC OFDM modulated signal. The entirety of the OFDM modulator in Figure 5-22 is implemented digitally in a DSP unit.

5.10.9 Summary of 4G

At the physical layer the 4G standard introduces and combines OFDM, carrier aggregation (CA), and MIMO to achieve tremendous data rates that,

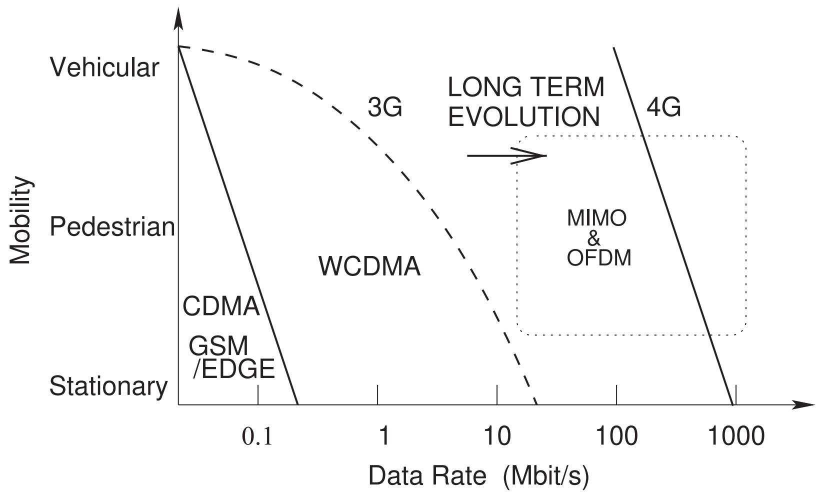


Figure 5-23: Data rate capacity of 4G as the long term evolution of 3G.

in a fully implemented 4G system achieves low-latency data download at mobile rates up to 100 Mbit/s and rates while stationary of 1 Gbit/s. OFDM is the technology that sends many relatively slow bitstreams over narrow bandwidth sub-channels to efficiently support multiple users making best use of the channels available and circumventing many multipath effects. It replaces the use of spreading codes in CDMA to support multiple access and circumvent some multipath problems. CA combines multiple bitsteams sent on different carriers and even from different basestations to increase the short term overall bit rate. MIMO uses multiple, i.e. M, transmit antennas, at the basestation, and multiple, i.e. N, receive antennas, at the handset, and exploits uncorrelated communication paths between pairs of transmit and receive antennas. The (ideally) zero correlation is made possible by multipath and the correlation is lowest if there is no line-of-sight path. If the de-correlation is complete the data capacity is increased by a function of the minimum of the number of transmit antennas, i.e. $\mathrm{MIN}(M, N)$. If there is some correlation, perhaps due to coupling between the receive antennas, then a **M**IMO capacity factor $H \leq \mathrm{MIN}(M, N)$ is defined. With MIMO it is necessary to modify Shannon's capacity limit as MIMO systems can exceed the limit defined for a single channel. Shannon's capacity limit for a MIMO system becomes [29]

$$\hat{C} = B_c \log_2(1 + \mathrm{SIR} \cdot H). \tag{5.31}$$

This indicates that the channel carrying capacity is greatly increased, especially when the SIR is high. The data rate capacity of 2G systems (GSM and CDMA) is contrasted with the capacity of 3G (WCDMA) and 4G systems in Figure 5-23.

5.11 5G, Fifth Generation Radio

As with 4G there will be long term evolution of the 5G standard and this will be upwards compatible with 4G. Even though 4G was touted as the long term evolution of 3G, at the physical layer it was fundamentally incompatible. 5G is fundamentally compatible with 4G at the physical layer and it is expected that cellular phone services will continue to function much the same as if 4G was being used. The first iteration (i.e. evolution) of 5G is called **N**ext Radio (**N**R) and begins with 3GPP Release 15, see Figure 5-17, and operation began in the second half of 2018.

The full 5G standard supports three use cases and a significantly improved level of service over 4G. One of the uses cases is **enhanced mobile broadband** (**eMBB**) with peak data rates of 20 Gbit/s and sustained data rates of 100 Mbit/s. High mobility of up to 500 km/s is supported. Another use case is **massive machine-to-machine communication** (**mMTC**) supporting the **internet of things** (**IoT**) with high density of devices (up to 10^6/km^2, long range, low data rate (1 kbit/s–100 kbit/s), and enabling 10 year battery life. The third use case is ultra reliability and low latency (URLLC) with less than 1 ms air interface latency and less than 5 ms overall end-to-end latency, 99.9999% reliability, data rates up to 10 Mbit/s, and supporting high mobility.

Many scenarios require ultra-reliable and very low end-to-end latency, perhaps as low as 1 ms. The early deployments of 4G had latencies of 80 ms or more and varied considerable across service providers. Over the years since 4G was launched latencies have reduced considerably but in 4G it is not possible to have air-interface latencies less than the 10 ms frame rate of 4G. A fundamental limit in 4G is due to the 0.5 ms time duration of a slot, see Figure 5-19. One concept in 5G is to have more slots per 1 ms subframe maintaining the subframe for upwards compatibility. In 4G there are two slots per subframe. Possible developments in 5G are 1) to have more slots per subframe and 2) (what seems to be similar) to have more symbols per slot. With 16 slots per subframe the slot duration is 0.0625 ms. The long-term evolution of 5G will continue and achieve overall latencies of 1 ms or less.

An important concept of 5G is to be fully compatible with the 4G infrastructure and the only new infrastructure required is that to support new services. There were two central ideas behind 4G: OFDM and MIMO and these are used in 5G OFDM. OFDM in 4G used two fixed sub-carrier spacing, but 5G has a number of different sub-carrier bandwidths including a subchannel bandwidth of 480 kHz, and a low bandwidth subchannel bandwidth of 7.5 kHz to accommodate high speed mobility As with 4G, carrier aggregation is used but now the bitstreams on up to 32 carriers are combined. Also channel bandwidths are up to 400 MHz compared to the 20 MHz RF bandwidth maximum of 4G.

5.11.1 Mesh Radio

In 5G a wireless mesh removes the need for a fixed basestation to communicate directly to an end unit [32]. If there is a node between the basestation and a mobile user, an intervening node, possibly another mobile user, can be used effectively as a relay. This reduces overall power requirements and levels of interference, which in turn leads to greater data carrying capacity. This concept can be extended to use multiple intervening nodes forming a dense mesh with considerable tolerance to multipath and interference effects. One significant benefit of this can be seen by noting that in an urban environment power can fall off by the fourth power of distance. Another benefit is that the impact of fading can be greatly reduced, as the paths in the mesh will fade independently. This will be augmented by many small cells called femtocells and picocells handling traffic in small geographic areas such as buildings, airports, and sporting facilities.

5.11.2 Cognitive Radio

Cognitive radio exploits the fact that much of the EM spectrum remains unused even while the bands reserved for cellular communications have reached capacity. This situation is a consequence of the allocation of bands for dedicated services that may not be active at a particular time and place. Examples are unused bands reserved for broadcast television channels and bands reserved for communication in times of emergencies. A cognitive radio senses its environment and adapts in real time to a user's communication needs by temporarily borrowing unused spectrum [33, 34]. As a result spectrum is used more efficiently. A cognitive radio avoids causing interference with the communications of other users by sensing spectrum use, changing frequency, adjusting power level, and altering transmission protocols. If a licensed band is borrowed, then a cognitive radio discontinues use of this part of the spectrum if the licensee of the band becomes active. This behavior is also called **dynamic spectrum management**.

5.11.3 Massive MIMO

MIMO in 4G uses multiple antennas at the basestation and at terminals to increase overall data rates provided that there are multiple paths between the transmitter and receiver. For example, with two basestation antennas transmitting separate bitstreams but in the same frequency band and two uncorrelated receive antennas, known as 2x2 MIMO, the theoretical maximum data rate is twice what could be supported by line-of-sight communication. In 5G there can be a very large number of transmit antennas even though there are few receiver antennas per terminal unit as multiple terminal units mean that there can effectively be a very large number of receive antennas.

5G operates at frequencies below 6 GHz and at millimeterwave frequencies. Below 6 GHz 8x8 MIMO is supported in the standard but there can be a hundred or more basestation antennas. For example. With 128 basestation antennas and four receiver antennas per terminal unit, 32 terminal units can be supported in the same frequency band. Overall the system has much higher throughput, theoretically a maximum of 128 times more than a single line-of-sight link could support. This concept, i.e. not all of the receive antennas need be on the same platform, is called massive MIMO.

Massive MIMO uses characterization of the channel as in conventional MIMO but uses this information to form multiple beams each directed at each terminal unit. This situation is shown in Figure 5-24 where there are multiple transmit/receive antennas at each basestation but relatively

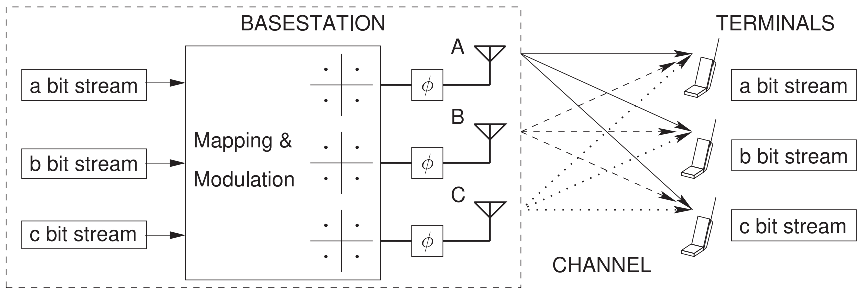


Figure 5-24: A massive MIMO system with beams from the basestation antennas to each terminal unit.

few transmit/receive antennas at each terminal unit. At the start of each correlation interval, i.e. the time over which a channel does not change significantly, each terminal unit transmits a code to the basestation. Each terminal unit uses an orthogonal code and so the basestation is able to characterize the channels from each of the terminal units to the basestation. The basestation is then able to use its antennas in a phased array to form a beam to transmit data just to the intended terminal. The real situation is a little different as a single beam is not formed but instead a weighted signal is broadcast from each of the basestation antennas with the effect being that all of the power following multiple paths comes together at the intended terminal unit. Still this process is called beamforming. Different beams are used for each basestation. In the second half of the correlation interval the terminal unit transmits back to the basestation and the basestation uses its beamforming in reverse to effectively create multiple receive beams each directed at a terminal units. Of course these are not strictly beams but use the channel model to combine the signals from each of the basestation receive antennas to effectively receive signals with the highest signal-to-interference ration from each terminal.

Massive MIMO has a performance advantage even with line-of-sight as long as the terminal units are not on top of each other as the paths from each of the basestation antennas to each of the terminals will be uncorrelated. Beam steering can be also used to concentrate transmitted energy in a particular direction.

5.11.4 Active Antenna Systems

Adaptive antennas, or **active antenna systems** (**AAS**), are also known as **smart antennas** and adapt an array of antennas to either direct an antenna beam between a transmitter and a receiver, eliminate interference, or use the **spatial signatures** provided by diverse propagation paths to transfer more information than can be sent over a single link [35]. All this is done using signal processing rather than hardware reconfiguration.

In one approach a smart antenna is used to switch between a number of fixed, usually narrow, antenna beams in what is analogous to a highly sectored antenna array. Switching requires knowledge of the angle of arrival and the beams track a mobile unit. A simple version of this concept is used in 3G and 4G radio. As well as focusing the available power on the intended communication nodes, beam forming through beam switching also minimizes interference.

Adaptive antennas increase data transfer rates, reduce interference, and reduce the amount of transmit power required.

5.11.5 Microwave Frequency Operation

Low frequency operation refers to 5G operating below 6 GHz. The bands being adopted for early stage deployment are the 700 MHz, 3300–4200 MHz and 4400–5000 MHz bands but the full coverage is not available in all countries. Many are predicting that the bulk of 5G communication will be at these frequencies, particularly at 3.5 GHz, up until the mid or late 2020s. It is important to 5G that there be globally available spectrum. The 700 MHz band is of particular interest as it will provide wide area coverage and deep penetration into buildings and so is a candidate for providing high reliability links important for the 5G mMTC applications. At this frequency 5G is compatible with 4G but with evolved characteristics such as supporting up to 32 carrier aggregation and 8×8 MIMO instead of the five carrier aggregation and (usually) maximum 3×3 MIMO of 4G. Also 5G in the low frequency range supports 256 QAM modulation but then the order of QAM modulation supported with 4G has evolved with latter 3GPP releases.

5.11.6 Millimeter-Wave Operation

The very high data rates of 5G are obtained by operating at millimeter-wave (mm-wave) frequencies where much more bandwidth is available. The bands being focused on are the 24.25–29.5 GHz and 37–40.5 GHz. (Yes, technically 24.25–29.5 GHz is not millimeter-waves which requires a wavelength to be 10 mm or less, but that is being called mm-waves in the 5G community.) At mm-wave frequencies very tight beamforming can be achieved if the same overall antenna size is used. However there is a penalty, signal attenuation is high and it is not possible to send signals through walls and into buildings. As such window mounted units are required to receive signals. Millimeter-wave operation is envisioned to use 2×2 MIMO only, but still use massive MIMO, be useful for low speed mobility, but provide very high data rates.

5.11.7 Non Orthogonal Multiple Access

One of the distinguishing features of 4G and 5G is OFDMA, an **orthogonal multiple access** (**OMA**) scheme. OFDMA in (both both 4G and) 5G allocates a dedicated resource block to each user and orthogonality ensures that there is almost no inter-user interference. The system utilization is optimum when each resource block is fully utilized. However this is not the case with IoT devices, the support for which is one of the main features of 5G. When IoT devices communicate very little information is usually exchanged and so most of the resource block used by an IoT device is unused. The expected growth of 5G will see a great proliferation of IoT devices with these devices having diverse data rate and latency requirements. Communicating with each IoT device by allocating a dedicated resource block does not use the capacity effectively. In addition, using the same basestation transmit power for each user is said to be 'unfair' in that each user typically has a different channel quality with a user having a poorer channel quality needing to operate with lower-order modulation. The 'unfairness' arrises because the channel capacity is not equally shared but it would be if the the basestation transmitted higher powers to the user with a poorer channel. Non orthogonal multiple access (**NOMA**) is designed to addresses this imbalance and to ensure optimum use of each resource block and hence of the communication system.

The key concept of NOMA is that low-rate devices will share a resource block, they will use differing symbol rates, and the transmit power level from the basestation to each user will be adjusted according to the quality of the communication channel [36, 37]. In NOMA, and in contrast to OFDMA, there will be inter-user interference and schemes have been developed that will enable individual users to be separated. One possible NOMA scheme to separate users by using code domain multiplexing with each user sharing a resource block being assigned a unique code. The concept is very similar to that used in WCDMA but applied at the level of a resource block.

The dominant NOMA scheme in 5G and the one to be first deployed (as of the time of writing this book) in a future 3GPP release of the 5G standard is **multi-user superposition transmission** (**MUST**). **Successive interference cancellation** (**SIC**) is the scheme used to separate users.

In MUST users are allocated different power levels and will use different symbol rates. For example, a distant user with a poor quality channel will receive more power in NOMA and a user with a good channel will receive less since the maximum power available from the basestation transmitter is fixed. The result is that the throughput of the close and distant users will be similar. This is called 'fairness' and is particularly important when the system is operating at or near capacity. A further advantage of NOMA is that the demands on the basestation transmitter hardware is reduced. One of the consequences of having a common symbol rate in OFDMA is that the PMEPR is higher than it would be if there was a range of symbol rates. Thus NOMA will result in the total RF signal being transmitted from the basestation having a lower PMEPR.

It is expected that there will be a significant increase in throughput as many users, think of IoT devices, only need to communicate at very low data rates. Current cellular systems reach an abrupt limit on their throughput. In NOMA overload is supported at the cost of inter-user interference. Studies have shown that a threefold increase incapacity is possible with NOMA schemes that tolerate overload and inter-user interference [37].

5.11.8 Summary

The 5G system introduces many system optimization strategies and a high-level of system optimization across perhaps hundreds of basestations is required to achieve full potential. The 5G systems will inevitably be followed by 6G and this is the subject of current research. The defining characteristic is that 6G is being targeted to operate above 100 GHz and provide enormous bandwidths since very wide bandwidths are available.

5.12 6G, Sixth Generation Radio

Sixth generation radio will operate in the 95 GHz to 275 GHz range, and will provide data rates of 100 Gbit/s [38]. While 5G hopes to provide these data rates at 60 GHz this may not come to pass as the required spectral efficiency of 14 bits/s/Hz and requires 60 GHz hardware with very high dynamic range and this is unlikely to be available for some time. With 6G these data rates will be possible with much lower-order modulation and hence reduced demands on hardware Regulatory bodies have allocated several bands in this range. In the USA these bands are 116–123 GHz, 174.8–182 GHz, 185–190 GHz, and 244–246 GHz for a total of 21 GHz of spectrum. Atmospheric losses can be large at these frequencies but, except for a peak of absorption around 140 GHz, this is more than compensated by the pencil-like beams generated by many-element phased array antennas possible at these high frequencies. The 100+ GHz frequencies have wavelengths of 3 mm or less enabling functionality not available at 60 GHz and below. These include precise positioning with millimeter resolution, near visual-quality imaging through fog and clouds, wireless cognition (off-loading large data sets from a mobile units for fixed computation), and unique sensing modalities by exploiting the many molecular resonances that occur above 100 GHz. The overwhelming challenge is the development of physical hardware and the development of array processing technologies [38].

5.13 Radar Systems

Radar uses EM signals to determine the range, altitude, direction, and speed of objects called targets by looking at the signals received from transmitted signals called radar waveforms. Common radar bands and applications are given in Table 5-16. The earliest use of EM signals to detect targets was demonstrated in 1904 by Christian Hülsmeyer using a spark gap generator [39]. This system was promoted as a system to avoid collisions of ships and detected the direction of targets only. Research contributed to further developments, with a significant acceleration during World War II. Radar is now a word in its own right, but in 1941 the term RADAR was created as an acronym for radio detection and ranging.

Table 5-16: IEEE radar bands and applications.

Band	Frequency	Wavelength	Application
HF	3–30 MHz	10–100 m	Over-the-horizon radar, oceanographic mapping
VHF	30–300 MHz	1–10 m	Oceanographic mapping, atmospheric monitoring, long-range search
UHF	0.3–1 GHz	1 m–30 cm	Long-range surveillance, foliage penetration, ground penetration, atmospheric monitoring
L	1–2 GHz	15–30 cm	Satellite imagery, mapping, long-range surveillance, environmental monitoring
S	2–4 GHz	7.5–15 cm	Weather radar, air traffic control, surveillance, search, IFF (identify, friend or foe)
C	4–8 GHz	3.75–7.5 cm	Hydrological radar, topography, fire control, weather
X	8–12 GHz	2.5–3.75 cm	Cloud radar, air-to-air missile seeker, maritime, air turbulence, police radar, high-resolution imaging, perimeter surveillance
Ku	12–18 GHz	1.7–2.5 cm	Remote sensing, short-range fire control, perimeter surveillance; pronounced “kay-you”
K	12=8–27 GHz	1.2–1.7 cm	Police radar, remote sensing, perimeter surveillance
Ka	27–40 GHz	7.5–12 mm	Police radar, weapon guidance, remote sensing, perimeter surveillance, weapon guidance; pronounced “kay-a”
V	40–75 GHz	4–7.5 mm	Perimeter surveillance, remote sensing, weapon guidance
W	75–110 GHz	2.7–4 mm	Perimeter surveillance, remote sensing, weapon guidance

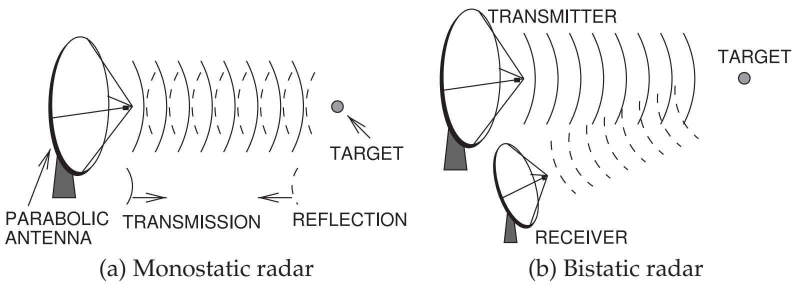


Figure 5-25: Radar system: (a) monostatic radar with the same site used for transmission of the radar signal and receipt of the reflection from the target; and (b) bistatic radar with different transmit and receive sites.

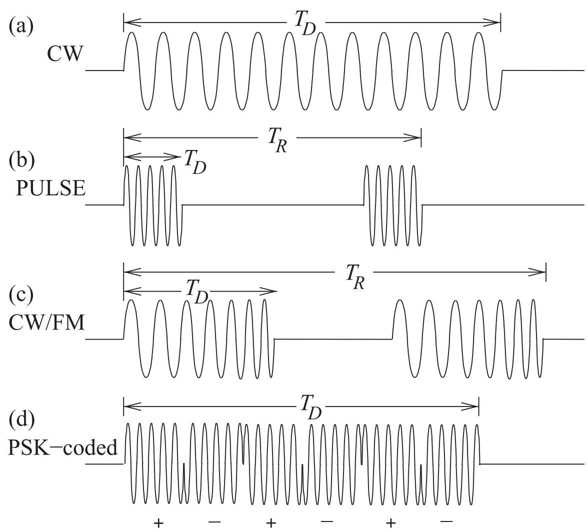


Figure 5-26: Radar waveforms: (a) continuous wave; (b) pulsed wave; (c) frequency modulated continuous wave; and (d) phase-encoded (PSK-coded) waveform.

In a radar system, typically a high-gain antenna such as a parabolic antenna is used to transmit a radar signal, but always a high-gain antenna is used to receive the signal. If the same antenna is used for transmit and receive (possibly two similar antennas at the same site) the system is called a **monostatic radar** (see Figure 5-25(a)). Radar with transmit and receive antennas at different sites is called **bistatic radar** (shown in Figure 5-25(b)).

In a monostatic radar using the same antenna for transmit and receive, the space is painted with a radar signal and the received signal is captured after a propagation delay from the antenna to the target and back again. A radar image can then be developed. In many radars the receive antenna is mechanically steered and often a regular rotation is used. With so-called synthetic aperture radars, a platform such as an aircraft moves the radar in one direction and a one-dimensional mechanical or electrical scan enables a two-dimensional image to be developed.

The categories of radar waveforms are shown in Figure 5-26. The continuous wave (CW) waveform shown in Figure 5-26(a) is on all or most of the time and is used to detect a reflection from a target. This reflected signal is much smaller than the transmitted signal and it can be difficult to separate the transmitted and received signals. A monostatic CW radar architecture is shown Figure 5-27(a), where a circulator[6] is used to separate the transmitted and received signals. The received signal is converted to digital form using

[6] For the circulator shown, power entering port 1 of a circulator leaves at port 2 and power entering port 2 is delivered to port 3. So ports 1 and 3 are isolated.

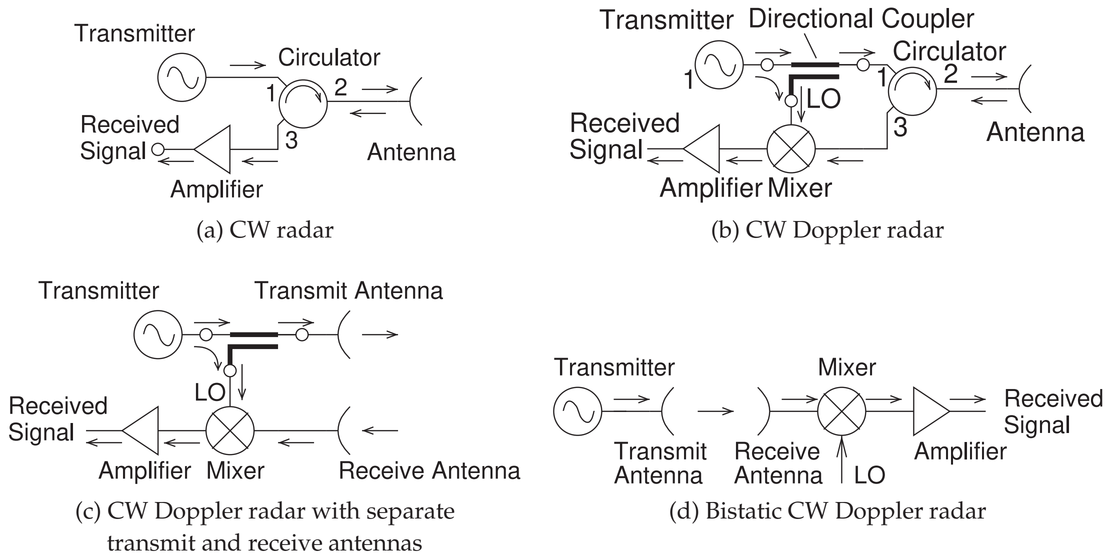


Figure 5-27: Radar architectures.

an ADC and the bandwidth of the ADC with the required dynamic range determines the limit on the bandwidth of the radar signal. Generally the broader the bandwidth, the better the radar system is at identifying objects. A CW signal can be used to develop an image, but it is not good at determining the range of a target. For this, a pulsed radar waveform, as shown in Figure 5-26(b), is better. The repetition period, T_R, is more than the round-trip time to the target, thus the time interval between the transmitted signal and the returned signal can be used to estimate range. Direction is determined by the orientation of the antenna.

The CW architecture can also be used with **pulsed radar**. In pulsed radar, the signal received contains the desired target signal, multipath and echoes, and clutter, as shown in Figure 5-28(a). These effects also appear in the signal received in CW radar, but it is much easier to see in pulsed radar. Identifying clutter and multipath effects is a major topic in radar processing. Alternative waveforms, especially digitally modulated waveforms, aid in extracting the desired information. The frequency modulated waveform, or chirp waveform, in Figure 5-26(c), will have a reflection that will also be chirped, and the difference between the frequency being transmitted and that received indicates the range of the target, provided that the target is not moving. The radar architecture that can be used to extract this information is shown in Figure 5-27(b). The directional couplers tap off a small part of the transmit signal and use it as the LO of a mixer with the received signal as input. A similar architecture is shown in Figure 5-27(c), but now separate transmit and receive antennas are used to separate the transmitted and received signals rather than using a circulator. Better separation of the transmitted and received signals can be obtained this way. The frequency of the IF that results is proportional to the range of the target.

An important concept in radar is that of **radar cross section** (**RCS**) denoted σ (with SI units of m^2). The RCS of a target is the equivalent area that

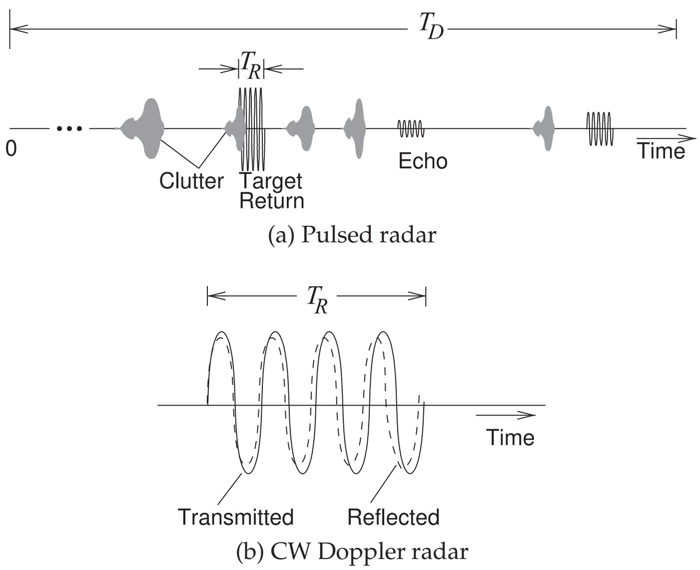


Figure 5-28:
Radar returns. The strength of the reflected signal depends on the radar cross section (RCS), σ, of the target. $\sigma = 0.01\ \mathrm{m}^2$ for a bird, $< 0.1\ \mathrm{m}^2$ for a stealth aircraft, $1\ \mathrm{m}^2$ for a human, 2–6 m^2 for a conventional fighter plane, $100\ \mathrm{m}^2$ for a large commercial aircraft, $200\ \mathrm{m}^2$ for a truck, 10,000–100,000 m^2 for a container ship.

intercepts the power of a transmitted signal and re-radiates all of that power isotropically to produce the observed power density at a receiver [40]. The RCS therefore depends on the frequency of the transmitted signal, the size of the target, the incident and reflected angle of the signal reflected by the target, and the reflectivity of the target.

The power of the signal reflected by the target and captured by the receive antenna is given by the **radar equation** [40]:

$$P_R = \frac{P_T G_T A_R \sigma F^4}{(4\pi)^2 R_T^2 R_R^2}, \tag{5.32}$$

where P_T is the transmit power delivered to the transmit antenna, G_T is the antenna gain of the transmitter, A_R is the effective aperture area of the receive antenna, F is the pattern propagation factor, R_T is the distance from the transmitter to the target, and R_R is the distance from the target to the receiver. F captures the effective loss due to multipath, which was captured in communications by introducing a signal dependence of $1/d^n$, where n ranges from 2 to 4 depending on the environment. In free space, $F = 1$. If the same antenna is used for transmit and receive, and multipath is not important (so $F = 1$), then Equation (5.32) becomes

$$P_R = \frac{P_T G_T A_R \sigma}{(4\pi)^2 R^4}, \tag{5.33}$$

where $R = R_T = R_R$. In Section 4.5.4 the antenna effective area was related to the antenna gain. From Equation (4.27),

$$A_R = \frac{G_R \lambda^2}{4\pi}, \tag{5.34}$$

and so, with $G_R = G_T$, Equation (5.33) becomes

$$P_R = \frac{P_T G_T^2 \lambda^2 \sigma}{(4\pi)^3 R^4}. \tag{5.35}$$

Table 5-17: Doppler frequency shifts for targets moving toward a radar at speed v_R.

Radar frequency, f_T	Relative speed, v_R 1 m/s	100 km/hr	1000 km/hr
500 MHz	3.3 Hz	92.6 Hz	925.9 kHz
2 GHz	13.3 Hz	370.4 Hz	3.704 kHz
10 GHz	66.7 Hz	1.852 kHz	18.519 kHz
40 GHz	266.7 Hz	7.407 kHz	74.074 kHz

If a signal of frequency f_T is transmitted and reflected from a moving target there is a **Doppler shift**, f_D, and the received signal will be at frequency

$$f_R = f_T + f_D \quad \text{(target moving toward the radar)} \tag{5.36}$$

$$f_R = f_T - f_D \quad \text{(target moving away from the radar)}, \tag{5.37}$$

where $f_D = 2v_R f_T/c$. (5.38)

In Equation (5.38) v_R is the radial component of the speed of the target relative to the radar, and c is the speed of light. Typical Doppler shifts are shown in Table 5-17.

If the target is moving, there will be a Doppler shift. If the target is moving toward a CW radar, then the frequency of the returned signal will be higher, as shown in Figure 5-28(b). A similar architecture to that used with chirp radar can be used (see Figures 5-27(bc)). The concept can be extended to bistatic radars, but now the local oscillator reference must be generated. As was seen previously, the frequency of the transmit carrier can be recovered for digitally modulated signals such as the PSK-encoded signals shown in Figure 5-26(d). Advanced high-end radars use digital modulation and CDMA-like waveforms and exploit space-time coding. Typically the radar waveforms to be transmitted are chirped, which is a technique that takes the desired transmit waveform and stretches it out in time so that it can be more efficiently amplified and more power can be transmitted. At the receiver, the radar signal is compressed in time so that it corresponds to the original transmit signal prior to chirping.

It should be apparent that radars and radar waveforms can be optimized for imaging or for exploiting Doppler shifts to track moving targets. Imaging is suitable when there is little clutter, such as looking into the air. However, it is difficult to detect targets such as cars that are moving on the ground. So-called **ground moving target indication** (**GMTI**) **radar** relies on Doppler shifts to discriminate moving targets, and then the ability to image accurately is compromised. Considerable effort is devoted to developing waveforms that are difficult to detect (stealthy) and are optimized for imaging or GMTI.

5.14 Summary

Radio frequency and microwave design continues to rapidly evolve, responding to new communication, radar, and sensor architectures. The long-term evolution also exploits opportunities made available with larger-scale monolithic integration and by advances in high-performance, low-power digital signal processing.

If filters and other hardware in a communication receiver are ideal, the final E_b/N_o and BER achieved are directly related to the RF SIR as described in this chapter. With practical filters and analog hardware there is a performance degradation and an SIR higher than the theoretical SIR is required to achieve a particular BER, or E_b/N_o. The difference is

Modulation	Implementation margin
BPSK	0.5 dB
QPSK	0.8 dB
8-PSK	1–1.6 dB
16-QAM	1.5–2.1 dB
CDMAOne	0.5 – 1 dB
WCDMA	2 dB

Table 5-18: Implementation margins for modern communication receivers [41, p. 328], [25, 42]. These implementation margins are what can be achieved by good system designs.

captured by the **implementation margin**, k, usually specified in decibels. To achieve a specific BER, SIR must greater than the theoretical SIR by k. The implementation margin is therefore a measure of the performance of RF hardware and is an important metric in design and in planning design. The implementation margin captures many imperfections. A design group and a company learns what it can achieve, e.g. see Table 5-18, and uses this in budgeting design costs and planning design effort. The design cost of RF systems are considerable and the ability to manage the design process and be able to estimate design effort is critical to timely success. Higher implementation margins result from the choice of lower-performing technologies, perhaps resulting from a compromise of performance, cost, and design effort.

5.15 References

[1] "Radio regulations," International Telecommunications Union, 2004.

[2] "Recommendation ITU-R V.662-3 (1986-1990-1993-2000), updated in 2005 for editorial reasons only, 2005." International Telecommunications Union.

[3] "American National Standard T1.523-2001, Telecom Glossary 2011," available on-line with revisions at http://glossary.atis.org, 2011, sponsored by Alliance for Telecommunications Industry Solutions.

[4] D. Ring, "Mobile telephony—wide area coverage," *Bell Laboratories Technical Memorandum*, Dec. 1947.

[5] J. Schulte, H.J. and W. Cornell, "Multi-area mobile telephone system," *IRE Trans. on Vehicular Communications*, vol. 9, no. 1, pp. 49–53, Jun. 1957.

[6] W. Lewis, "Coordinated broadband mobile telephone system," *IRE Trans. on Vehicular Communications*, vol. 9, no. 1, pp. 43–48, Jun. 1957.

[7] A. Joel, "Mobile communication system," US Patent 3 663 762, May 16, 1972.

[8] F. Ikegami, "Mobile radio communications in japan," *IEEE Trans. on Communications*, vol. 20, no. 4, pp. 738–746, Aug. 1972.

[9] "Cost final report, http://www.lx.it.pt/cost231."

[10] F. Ikegami, T. Takeuchi, and S. Yoshida, "Theoretical prediction of mean field strength for urban mobile radio," *IEEE Trans. onAntennas and Propagation,*, vol. 39, no. 3, pp. 299–302, 1991.

[11] J. Doble, *Introduction to Radio Propagation for Fixed and Mobile Communications*. Norwood, MA, USA: Artech House, Inc., 1996.

[12] H. Markey and G. Antheil, "Secret communication system," US Patent US Patent 397 412, 04 11, 1942.

[13] K. Gilhousen, I. Jacobs, L. Weaver Jr., and E. Armstrong, "Spread spectrum multiple access communication system using satellite or terrestrial repeaters," US Patent 4 901 307, Feb 13, 1990.

[14] A. Viterbi, *CDMA Principles of Spread Spectrum Communications*. Addison-Wesley, 1995.

[15] V. Garg, K. Smolik, and J. Wilkes, *Applications of CDMA in Wireless/Personal Communications*. Prentice Hall, 1996.

[16] J. Barry, E. Edwards, and D. Messerschmitt, *Digital Communication*, 3rd ed. Kluwer Academic Publishers, 2004.

[17] D. Smith, *Digital Transmission Systems*, 3rd ed. Kluwer Academic Publishers, 2004.

[18] G. Lewis, *Communications Systems: Engineers' Choices*, 3rd ed. Focal Press, 1999.

[19] W. Lee, "Spectrum efficiency in cellular [radio]," *IEEE Trans. on Vehicular Technology*, vol. 38, no. 2, pp. 69–75, May 1989.

[20] R. Price and P. E. Green Jr., "Anti-multipath receiving system," US Patent 2 982 853, May 02, 1961.

[21] R. Ziemer and W. Tranter, *Principles of Communications: Systems, Modulation, and Noise*, 5th ed. Wiley, 2001.

[22] T. Rappaport, *Wireless Communications Principles and Practice*. Prentice Hall, 1996.

[23] K. Cheun, "Performance of direct-sequence spread-spectrum rake receivers with random spreading sequences," *IEEE Trans. on Communications*, vol. 45, no. 9, pp. 1130–1143, Sep. 1997.

[24] International Telecommunications Union, at http://www.itu.int.

[25] "3rd generation partnership project (3gpp)," http://www.3gpp.org.

[26] Third Generation Partnership Project specifications, at http://www.3gpp.org/specifications.

[27] G. Foschini, "Layered space-time architecture for wireless communication in a environment when using multi-element antennas," *Bell Labs Technical J.*, vol. 1, no. 2, pp. 41–59, Autumn 1996.

[28] G. Raleigh and J. Cioffi, "Spatio-temporal coding for wireless communications," in *Global Telecommunications Conf., 1996)*, vol. 3, Nov. 1996, pp. 1809–1814.

[29] A. Goldsmith, S. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of mimo channels," *IEEE J. on Selected Areas in Communications*, vol. 21, no. 5, pp. 684–702, Jun. 2003.

[30] D. Gesbert, M. Shafi, D. shan Shiu, P. Smith, and A. Naguib, "From theory to practice: an overview of mimo space-time coded wireless systems," *IEEE J. on Selected Areas in Communications*, vol. 21, no. 3, pp. 281–302, Apr. 2003.

[31] W. He and C. Georghiades, "Computing the capacity of a MIMO fading channel under psk signaling," *IEEE Trans. on Information Theory*, vol. 51, no. 5, pp. 1794–1803, May 2005.

[32] D. Benyamina, A. Hafid, and M. Gendreau, "Wireless mesh networks design—a survey," *IEEE Communications Surveys Tutorials*, vol. 14, no. 2, pp. 299–310, quarter 2012.

[33] B. Wang and K. Liu, "Advances in cognitive radio networks: A survey," *IEEE J. of Selected Topics in Signal Processing*, vol. 5, no. 1, pp. 5–23, Feb. 2011.

[34] J. Wang, M. Ghosh, and K. Challapali, "Emerging cognitive radio applications: A survey," *IEEE Communications Magazine*, vol. 49, no. 3, pp. 74–81, Mar. 2011.

[35] D.-C. Chang and C.-N. Hu, "Smart antennas for advanced communication systems," *Proc. IEEE*, vol. 100, no. 7, pp. 2233–2249, Jul. 2012.

[36] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (noma) for cellular future radio access," in *2013 IEEE 77th vehicular technology conference (VTC Spring)*. IEEE, 2013, pp. 1–5.

[37] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-Lin, and Z. Wang, "Non-orthogonal multiple access for 5g: solutions, challenges, opportunities, and future research trends," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 74–81, 2015.

[38] T. Rappaport, Y. Xing, O. Kanhere, S. Ju, A. Alkateeb, and G. Trichopoulos, "Wireless communications and applications above 100 GHz: opportunities and challenges for 6g and beyond," *IEEE Access*, vol. 7, 2019.

[39] Hulsmeyer, "Wireless transmitting and receiving mechanism for electric waves," germany Patent US Patent 810 150, Jan. 16, 1906.

[40] G. W. Stimson, *Introduction to Airborne Radar*, 3rd ed. Scitech Publ., 2014.

[41] E. Tozer, *Broadcast Engineer's Reference Book*. Focal Press, 2004.

[42] A. Richardson, *WCDMA Design Handbook*. Cambridge University Press, 2005.

[43] K. Gard, H. Gutierrez, and M. Steer, "Characterization of spectral regrowth in microwave amplifiers based on the nonlinear transformation of a complex gaussian process," *IEEE Trans. on Microwave Theory and Techniques*, vol. 47, no. 7, pp. 1059–1069, Jul. 1999.

5.16 Exercises

1. Research at Bell Labs in the 1960s showed that the minimum acceptable SIR for voice communications is 17 dB. This applies to analog modulated signals, but not digitally modulated signals, where BER is important. Consider a seven-cell cluster. If the power falls off as $1/d^3$, where d is distance, determine the worst possible SIR considering only interference from other radios. The worst situation will be when a mobile handset is at the edge of its cell. To do this you need to estimate the distance from the handset to the other basestations (in neighboring clusters) that are operating at the same power levels. Consider the cells to be hexagons. Develop a symbolic expression for the total interference signal level at the handset, assuming that all basestations are radiating at the same power level, P. You can use approximate distances. For example, each distance can be expressed in terms of integer multiples of cell radii, R. Is the 17 dB SIR achieved using a 7-cell cluster?
2. Describe the following concepts.
   (a) Clusters in a cellular phone system.
   (b) Multipath effects in a central city area compared to multipath effects in a desert.
3. Describe the concept of clusters in a cellular phone system in four lines.
4. Short answer questions on modulation and spectral efficiency.
   (a) What is the PMEPR of a phase modulated signal?
   (b) In five lines explain your understanding of spectral efficiency as it relates to bits per hertz. That is, how can you have a spectral efficiency of n bit/s/Hz where n is more than 1? [Note that sometimes this is expressed as bit/s/Hz as well as bit/s/Hz.]
   (c) What is the modulation efficiency of a QPSK-modulated signal? Ignore the impact of the number of cells in a cluster.
5. A cellular communication system uses a frequency reuse plan with seven cells per cluster to obtain the required minimum SIR. If a QPSK system is used, what is the radio spectrum efficiency in terms of bit/s/Hz/cell if all transitions on the constellation diagram are allowable? Assume that there is no coding.
6. A cellular communication system uses GMSK modulation and a frequency reuse plan with three cells per cluster. What is the radio spectrum efficiency in terms of bit/s/Hz/cell? Assume that there is no coding.
7. A cellular communication system uses $\pi/4$-DQPSK modulation and a frequency reuse plan with five cells per cluster. What is the radio spectrum efficiency in terms of bit/s/Hz/cell? Assume that there is no coding.
8. A frequency reuse plan has three cells per cluster. If ideal 16-QAM modulation is used, what is the system spectral efficiency (in bits/s/Hz/cell)?
9. A 2G cellular communication system uses a frequency reuse plan with seven cells per cluster. If ideal QPSK modulation is used, what is the system spectral efficiency (in bits/s/Hz/cell)?
10. A 2G system has three cells per cluster. If $3\pi/8$-8PSK? modulation is used, what is the system spectral efficiency (in bits/s/Hz/cell)?
11. A cellular communication system uses $3\pi/8$-8PSK modulation and a frequency reuse plan with three cells per cluster. What is the radio spectrum efficiency in terms of bit/s/Hz/cell? Assume that there is no coding.
12. A communication system uses a modulation with a modulation efficiency of 5 bit/s/Hz. Ignore coding so that $R_b = R_c$. What is the radio spectral efficiency in terms of bit/s/Hz/cell if there are three cells per cluster?
13. A proposed modulation format has a modulation efficiency of 3.5 bit/s/Hz. Antenna sectoring and required SNR lead to a system with seven cells per cluster. You can ignore the impact of coding so you can assume that $R_b = R_c$. What is the radio spectral efficiency in terms of bit/s/Hz/cell modulated signal?
14. A cellular radio system uses a frequency reuse plan with 12 cells per cluster. If ideal 8-PSK modulation is used, what is the system spectral efficiency in terms of bit/s/Hz/cell?
15. Consider a cellular system having a frequency reuse plan with seven cells per cluster to obtain the minimum signal-to-interference ratio. If ideal QPSK modulation is used, what is the system spectral efficiency in terms of bits per second per hertz per cell?
16. A monostatic free-space 10 GHz pulsed radar system is used to detect a fighter plane having a radar cross section, σ, of $5\ \mathrm{m}^2$. The antenna gain is 30 dB and the transmitted power is 1 kW. If the minimum detectable received signal is -120 dBm, what is the detection range?

17. An antenna with a gain of 10 dB presents an RF signal with a power of 5 dBm to a low-noise amplifier along with noise of 1 mW and an interfering signal of 2 mW.
    (a) What is the RF SIR? Include both noise and the interfering signal in your calculation. Express your answer in decibels.
    (b) The modulation format and coding scheme used have a processing gain, G_P, of 7 dB. The modulation scheme has four states. What is the ratio of the energy per bit to the noise per bit, that is, what is the effective E_b/N_o after despreading?

18. The receiver in a digital radio system receives a 100 pW signal and the interference from other radios at the input of the receiver is 20 pW. The receiver has an overall gain of 40 dB and the noise added by the receiver, referred to the output of the receiver, is 100 nW.
    (a) What is the RF SIR at the output of the receiver?
    (b) If 16-QAM modulation with a modulation efficiency of 2.98 bit/s/Hz is used and the processing gain is 30 dB, what is the effective SIR after despreading, i.e. what is $E_{b,\mathrm{eff}}/N_{o,b}$?

19. A new communication system is being investigated for sending data to a printer. The system will use GMSK modulation and a channel with 25 MHz bandwidth and an information bit rate of 10 Mbit/s. The modulation format will result in a spectrum that distributes power almost uniformly over the 25 MHz bandwidth. [Parallels Example 5.3]
    (a) What is the processing gain?
    (b) If the received RF SIR is 6 dB, what is the effective system SIR (or $E_{b,i}/N_{o,i}$) after DSP? Express your answer in decibels.

20. A 4 kHz bandwidth voice signal is coded by a vocodor as an 8 kbit/s data stream. Coding increases the data stream to 64 kbit/s. What is the processing gain that can be achieved at the receiver if QPSK modulation is used with a modulation efficiency of 1.4 bit/s/Hz?

21. A receive antenna with a gain of 10 dB presents a signal with a power of 5 dBm to a low-noise amplifier along with noise of 1 mW and an interfering signal of 2 mW.
    (a) What is the SIR at the input to the amplifier? Express your answer in decibels.
    (b) BPSK modulation is used and coding results in a processing gain, G_P, of 7 dB. What is the ratio of the energy per bit to the noise power per bit (i.e., what is E_b/N_o) after despreading?

22. If a received RF signal has an SIR of -5 dB and the processing gain (calculated from bit rates) that can be achieved for the modulation and coding used is 15 dB, what is the E_b/N_0 after processing? There are 4 bits per symbol.

23. A signal with a power of 13 dBm is input to a low-noise amplifier along with noise of 1 mW and an interfering signal of 2 mW.
    (a) What is the SIR at the input to the amplifier? Express your answer in decibels.
    (b) QPSK modulation is used and coding results in a processing gain, G_P, of 13 dB. What is the ratio of the effective energy per bit to the noise power per bit (i.e., what is E_b/N_o) after despreading?

24. Short answer questions. Each part requires a short paragraph of five lines and a figure.
    (a) Explain how OFDM reduces the impact of multiple paths in a wireless communication system.
    (b) Explain how MIMO exploits multipath to enhance the capacity of a digital communication system.

25. A coding rate of 2/3 is required to manage transmission errors in a 54 Mbit/s data link. That is, the information bit rate is 54 Mbit/s. What is the total bit rate required (including data and coding bits)?

26. The channel bandwidth in the GSM cellular phone system is 200 kHz and the GMSK modulation scheme used has a spectral efficiency of 1.354 bit/s/Hz.
    (a) What is the data rate of one frequency channel?
    (b) A time slot is 577 μs long. How many bits are there in one (i.e. a duration of 8.25 bits). How many data bits are there in a GSM time slot?
    (c) A GSM frame duration is 4.615 ms long and has eight time slots and a voice user has one time slot every frame. How many data bits per second are available to a single user?

27. Consider an OFDM system with 48 subcarriers carrying data and which uses 16-QAM modulation of each subcarrier and a coding rate of 2/3. There are also four pilot subcarriers that are used for frequency and phase reference, to ensure that spectral lines are not created, and to facilitate carrier recovery. The pilot carriers can be ignored in this problem so consider 48 subcarriers. The modulation efficiency achieved

for this particular implementation of 16-QAM is 2.98 bit/s/Hz.

(a) How many symbols are there for each subcarrier? That is, how many points are there in the constellation diagram for one subcarrier?
(b) How many coded bits are there on each subcarrier? That is, how many bits per symbol are there for each subcarrier?
(c) Considering all of the data subcarriers, how many coded bits are there per OFDM symbol? [Hint, there are 16 subcarriers, so for each OFDM symbol there will be 16 subcarrier symbols.]
(d) Considering the coding rate, determine the number of data bits per OFDM symbol. That is, ignore coding bits.

28. Consider an OFDM system with 12 subcarriers carrying data and which uses 8-PSK modulation of each subcarrier and a coding rate of 3/4. Pilot subcarriers they can be ignored in this problem so consider all 12 subcarriers.
(a) How many symbols are there for each subcarrier? That is, how many points are there in the constellation diagram for one subcarrier?
(b) How many coded bits (code + data) are there on each subcarrier? That is, how many bits per symbol are there for each subcarrier?
(c) Considering all of the data subcarriers, how many coded bits are there per OFDM symbol? [Hint, there are 12 subcarriers, so for each OFDM symbol there will be 12 subcarrier symbols.]
(d) Considering the coding rate, determine the number of data bits per OFDM symbol. That is, ignore coding bits.

29. A digital radio system transmits a baseband digital signal of 100 Mbit/s over a channel that is 300 MHz wide. The digital modulation scheme effectively fills the 300 MHz channel with uniform power.
(a) What is the processing gain that can be achieved with this system?
(b) Consider that the signal received and delivered to the input of the receiver front end is 100 pW and the interference from other radios delivered to the receiver front-end is 20 pW. What is the SIR at the input to the receiver electronics?

30. The L1 band of the global positioning system (GPS), is centered at 1.57542 GHz and has two overlapping spread-spectrum encoded signals. The stronger of these is the coarse acquisition (C/A) signal with an information bit rate of 50 bits/s and a transmission rate of 1.023 million chips per second using BPSK modulation with an RF bandwidth of 2.046 MHz. In ideal conditions the C/A signal received has a power of -130 dBm. A GPS receiver has an antenna noise temperature is of 290 K.
(a) What is the processing gain?
(b) What is the noise in dBm received in the 2.046 MHz bandwidth?
(c) What is the SNR in decibels?
(d) If a C/A signal is received from each of 10 satellites (so there are 9 interfering signals), what is the total interference power for one satellite's C/A signal?
(e) With respect to just one of the C/A signals, what is the SINR (signal to interference plus noise ratio) at the receiver?
(f) If the receiver does not contribute noise, what is the effective SNR of the despread bitstream from each satellite?
(g) If the required minimum effective SNR is 6 dB, what is the minimum acceptable power, in dBm, of the GPS signal received from one satellite?

31. GLONASS is the Russian satellite navigation system with one of two open signals called the L1OF band at 1600.995 MHz. The system uses DSSS encoding and BPSK modulation and each GLONASS satellite transmits on a different frequency. The symbol rate is 511,000 chips/s, the bandwith of the transmitted signal is approximately 540 kHz, and there are 50 information bits per second.
(a) What is the system's processing gain?
(b) What is the noise in dBm received in the 540 kHz bandwidth?
(c) If the required system minimum effective SNR is 6 dB, what is the minimum acceptable power, in dBm, of the received signal? Assume that the receiver is noiseless.

32. A new communication system uses a channel that is 100 MHz wide and uses direct sequence CDMA to efficiently spread a 5 Mbit/s digital signal over the full channel.
(a) What is the processing gain that can be achieved with the received signal?
(b) The analog signal at the output of the receive antenna has an SIR of 1 dB, what is energy per bit divided by the noise per bit?

33. A deep-space communication system will use direct sequence spread spectrum to code a data stream of 10 kbit/s then modulate the signal to transmit over a 5 GHz link to a ground sta-

tion. Since the propagation loss is very high it has been determined that the processing gain must be 50 dB. If the link has a bandwidth of 1 MHz, what is the maximum baseband bit rate (in bit/s) that can be supported?

34. A direct sequence spread spectrum code of 10 Mbit/s is used to code a 4 kbit/s data steam that is modulated using $3\pi/8$-8PSK modulation to produce an RF signal at 1900 MHz. The modulation efficiency of $3\pi/8$-8PSK modulation is 2.7 bit/s/Hz.
    (a) What is the bandwidth of the RF signal?
    (b) What processing gain can be achieved in the receiver?

35. An OFDM system with 12 data subcarriers, uses a coding rate of 3/4, and each subcarrier uses 16-QAM modulation (with a modulation efficiency of 2.7 bit/s/Hz) with a bandwidth of 250 kHz. What is the maximum data rate supported?

36. An OFDM system with 48 subcarriers carrying data uses 16-QAM modulation of each subcarrier and a coding rate of 2/3. The actual modulation efficiency for the 16-QAM system here is 2.7 bit/s/Hz. What is the maximum data rate supported in Mbit/s when the bandwidth of each modulated subcarrier is 312 kHz?

37. Explain using sentences and a diagram how OFDM reduces the impact of multiple paths in a wireless communication system.

38. Explain using sentences and a diagram how MIMO exploits multipath to enhance the capacity of a digital communication system.

39. A free-space 2 GHz pulsed monostatic radar system transmits a 2 kW pulse and has a minimum detectable received signal power of -90 dBm. What is the antenna gain required to be able to detect a target with a radar cross section of $10\ \mathrm{m}^2$ at 10 km?

40. A 10 GHz bistatic radar has a minimum detectable received signal power of -150 dBm, an antenna gain of 26 dB, and a required range of 100 km. What is the transmitted pulse power in dBm needed to detect a
    (a) conventional fighter aircraft having an RCS of $5\ \mathrm{m}^2$?
    (b) a stealth aircraft with an RCS of $0.05\ \mathrm{m}^2$?

41. The L5 band is a new public band of the GPS system and is centered at 1.176.5 GHz. The coarse acquisition (C/A) signal has an information bit rate of 50 bits/s and a spread-spectrum transmission rate of $10.23 \cdot 10^6$ chips per second using BPSK modulation and occupying an RF bandwidth of 20.46 MHz. The noiseless GPS receiver has an omnidirectional antenna with a noise temperature is 290 K.
    (a) What is the system's processing gain?
    (b) What is the noise in dBm in the 20.46 MHz bandwidth?
    (c) If overlapping C/A signals are received from 10 satellites, what is the total interference power for the signal from one satellite? The power of a C/A signal is S.
    (d) If the required system minimum effective SNR is 6 dB, what is the minimum acceptable received power, in dBm, of the signal from one satellite?

5.16.1 Exercises By Section

†challenging, ‡very challenging

§5.3 $1^\ddagger, 2^\dagger, 3^\dagger, 4^\ddagger, 5^\dagger, 6^\dagger, 7^\ddagger$
§5.5 $8^\dagger, 9, 10, 11^\dagger, 12, 13, 14^\dagger, 15^\dagger, 16^\dagger, 17^\dagger, 18^\dagger, 19, 20, 21$
§5.6 $22, 23, 24, 25^\ddagger, 26^\dagger, 27^\dagger, 28^\dagger, 29^\dagger$
§5.8 $30^\dagger, 31^\dagger, 32^\dagger, 33^\ddagger, 34^\ddagger, 41^\dagger$
§5.10 $35, 36, 37, 38$
§5.13 $39, 40$

5.16.2 Answers to Selected Exercises

1	12.5 dB	6	0.45bit/s/Hz/cell	27(d)	128
8(d)	3 bit/s/Hz	32(b)	14 dB	28(d)	27
19	0.1	34(a)	3.704 MHz	29(b)	6.99 dB
21(b)	7.228 dB	25	81 Mbit/s		

Appendix

5.A Mathematics of Random Processes

This appendix presents the essential mathematics required to describe the statistical properties of a random process such as noise. Also, it is difficult to analyze circuits with digitally modulated signals unless the signals are treated as being random with high-order statistics. Several statistical terms are introduced to describe the properties of a random variable X and how its value at one time is related to its value at other times. So X will, in general, vary with time (i.e., it can be written as $X(t)$) and its value at a particular time will be random.

This section presents all of the probability metrics required to describe noise, interference, and digitally modulated signals. A random process in time t is a family of random variables $\{X(t), t \in T\}$, where t is somewhere in the time interval T. The probability that $X(t)$ has a value less than x_1 is denoted by $P\{X(t) \leq x_1\}$. This is also called the **cumulative distribution function** (**CDF**), and sometimes called just the **distribution function** (**DF**):

$$F_X(x_1) = P\{X \leq x_1\}. \tag{5.39}$$

In general, since $X(t)$ varies with time, the CDF required to handle noise, that is, random voltages and currents, will depend on time. So the CDF used with noise and interference in communications will have two arguments, and the multivariate CDF is

$$F_X(x_1; t_1) = P\{X(t_1) \leq x_1\}. \tag{5.40}$$

That is, $F_X(\infty, t) = 1$ and $F(-\infty, t) = 0$. F_X is being used with two arguments, as it is necessary to indicate the value of X at a particular time. $F_X(x_1; t_1)$ is said to be the first-order distribution of $X(t)$.

Another probability metric often used is the **probability density function** (**PDF**), f, which is related to the CDF as

$$F_X(x_1; t_1) = \int_{-\infty}^{x_1} f(x, t_1).dx\ . \tag{5.41}$$

Another property that needs to be captured is the relationship between the value of the random variable at one time, t_1, to its value at another time, t_2. Clearly, if the variables are completely random there would be no relationship. The statistical relationship of x at t_1 and at t_2 is described by the **joint CDF**, $F_X(x_1, x_2; t_1, t_2)$, which is the second-order distribution of the random process:

$$F_X(x_1, x_2; t_1, t_2) = P\{X(t_1) \leq x_1, X(t_2) \leq x_2\}. \tag{5.42}$$

This is the joint CDF, or probability, that $X(t_1)$ will be less than x_1 at time t_1 and also that $X(t_2)$ will be less than x_2 at time t_2. So in the case of noise, the joint CDF describes the correlation of noise at two different times. In general, the nth order distribution is

$$F_X(x_1, \ldots, x_n; t_1, \ldots, t_n) = P\{X(t_1) \leq x_1, \ldots, X(t_n) \leq x_n\}. \tag{5.43}$$

If the random process (i.e., X) is discrete, then the **probability mass function** (**PMF**) is used for the probability that x has a particular value. The general form of the PMF is

$$p_X(x_1, \ldots, x_n; t_1, \ldots, t_n) = P\{X(t_1) = x_1, \ldots, X(t_n) = x_n\}, \tag{5.44}$$

and the general form of the PDF, used with continuous random variables, is (from Equation (5.41))

$$f_X(x_1, \ldots, x_n; t_1, \ldots, t_n) = \frac{\partial^n F_X(x_1, \ldots, x_n; t_1, \ldots, t_n)}{\partial x_1 \ldots \partial x_n}. \tag{5.45}$$

The statistical measures above describe the properties of random variables and capture the way a variable is related to itself at different times, and how two different random processes are related to each other at the same time and at different times. Such characterizations are based on the **expected value** of a random variable. The expectation of a random variable $X(t)$ is the weighted average of all possible

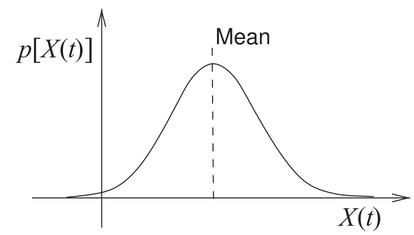


Figure 5-29: Gaussian distribution.

values of the random variable. The weighting is the probability for a discrete random variable, and is the probability density for a continuous random variable. So the expected value of the random variable $X(t)$ is

$$E[X(t)] = \begin{cases} \sum_{-\infty}^{\infty} x(t)p_X(x,t) & \text{for a discrete random variable} \\ \int_{-\infty}^{\infty} x(t)p_X(x,t)dx & \text{for a continuous random variable.} \end{cases} \tag{5.46}$$

This is just the mean of $X(t)$ defined as

$$\mu_X(t) = \bar{X}(t) = \langle X(t) \rangle = E[X(t)]. \tag{5.47}$$

The mean is also called the first-order moment of $X(t)$. $E[\,]$ is called the expected value of a random variable and the term is synonymous with the **expectation**, **mathematical expectation**, **mean**, and **first moment** of a random variable. The symbols $\langle\,\rangle$ are a clean way of specifying the expectation. In general a computer program would need to be used to calculate the expected value. However, for some assumed probability distributions there are analytic solutions for $E[\,]$.

The nth-order moment of $X(t)$ is just the expected value of the nth power of $X(t)$:

$$\mu_n' = E[X^n(t)] = \langle X^n(t) \rangle = \begin{cases} \sum_{-\infty}^{\infty} x^n(t)p_X(x,t) & \text{for a discrete random variable} \\ \int_{-\infty}^{\infty} x^n(t)p_X(x,t)dx & \text{for a continuous random variable.} \end{cases} \tag{5.48}$$

Thus the **second moment**, sometimes called the **second raw moment**, is $\mu_2' = \langle X^2(t) \rangle = E[X^2(t)]$. A more useful quantity for characterizing the statistics of a signal is the **second central moment**, which is the second moment about the mean. The second central moment of a random variable is also called its **variance**, written as σ^2 or as μ_2:

$$\begin{aligned} \sigma^2 = \mu_2 &= E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2] = E[X^2] - 2\mu E[X] + \mu^2 \\ &= E[X^2] - 2\mu^2 + \mu^2 = E[X^2] - \mu^2 = E[X^2] - (E[X])^2 \\ &= \langle X^2(t) \rangle - \mu^2 = \langle X^2(t) \rangle - \langle X^2(t) \rangle . \end{aligned} \tag{5.49}$$

The variance is a measure of the **dispersion** of a random variable (i.e., how much the random variable is spread out). Variance is one of several measures of dispersion, but it is the preferred measure when working with noise and digitally modulated signals. The **standard deviation**, σ, is the square root of the variance σ^2. It is also common to denote the variance of X as σ_X^2, and in general, the variance can be a function of time, $\sigma^2(t)$.

It is common to approximate the statistical distribution of a digitally modulated signal as a Gaussian or normal distribution. This distribution is shown in Figure 5-29 and is mathematically described by its probability distribution

$$p[X(t)] = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(X(t)-\mu)/(2\sigma^2)}, \tag{5.50}$$

where the mean of the distribution is μ and the variance is σ^2. The third and higher moments of the Gaussian distribution are zero. That is one of the reasons why this distribution is so commonly used as an approximation. Analysis using the Gaussian distribution is much simpler than it would be for other distributions. More realistically, a digitally modulated signal has I and Q components and the distribution of each of these should be approximated as a Gaussian distribution. Such a distribution is called a **complex Gaussian distribution** and the analysis of the distortion produced by an amplifier using the complex Gaussian distribution is more accurate. Using a more sophisticated distribution, e.g. using the moments calculated from the actual digitally modulated signal, provides even greater accuracy in analysis [43] but now the complexity is beyond manual calculation.

If a digitally modulated signal is completely random, then there would be no correlation between the value of the signal at one time and its value at another time. However, there is a relationship and the relationship is described by the signal's **autocorrelation function**. The autocorrelation function is used to describe the relationship between values of a function separated by different instants of time. For a random variable, it is given by

$$R_X(t_1, t_2) = E[X(t_1)X(t_2)]. \tag{5.51}$$

When the random variable is discrete, the one-dimensional autocorrelation function of a random sequence of length N is expressed as

$$R_X(i) = \sum_{j=0}^{N-1} x_j x_{j+i}, \tag{5.52}$$

where i is the so-called **lag parameter**. The autocovariance function of $X(t)$ is given by

$$\begin{aligned} K_X(t_1, t_2) &= E[\{X(t_1) - \mu_X(t_1)\}]E[\{X(t_1) - \mu_X(t_1)\}] \\ &= R_X(t_1, t_2) - \mu_X(t_1)\mu_X(t_2) \end{aligned} \tag{5.53}$$

while the variance is given by

$$\sigma_X(t) = E[\{X(t) - \mu_X(t)\}]^2 = K_X(t_1, t_1). \tag{5.54}$$

A random process $X(t)$ is stationary in the strict sense if

$$F_X(x_1, \ldots, x_n; t_1, \ldots, t_n) = F_X(x_1, \ldots, x_n; t_1 + \tau, \ldots, t_n + \tau) \tag{5.55}$$

$\forall t_i \in T$, $i \in$ **N**. If the random process is **wide-sense stationary** (**WSS**), then it is stationary with order 2. This means that in most cases only its first and second moments (i.e., the mean and autocorrelation) are independent of time t and only dependent on the time interval τ. More precisely, if a random process is WSS, then

$$\begin{aligned} &E[X(t)] = \mu \quad \text{(i.e., its mean is a constant)} \\ \text{and} \quad &R_X(t, s) = E[X(t)X(s)] = R_X(|s - t|) = R_X(\tau). \end{aligned} \tag{5.56}$$

Note that the autocorrelation of a WSS process is dependent only the time difference τ. A random process that is not stationary to any order is **nonstationary** (i.e., its moments are explicitly dependent on time).

The discussion now returns to the properties of a Gaussian random process. A Gaussian random process is a continuous random process with PDF of the form

$$f_X(x, t) = \frac{1}{\sqrt{2\pi\sigma(t)^2}} \exp\left(\frac{-\mu(x, t)^2}{2\sigma(t)^2}\right), \tag{5.57}$$

where $\mu(x, t)$ represents the mean of the random process and $\sigma(t)$ represents the variance. A normal random process is a special case of a Gaussian random process in that it has a mean of zero and a variance of unity. A **Poisson random process** is a discrete random process with parameter $\lambda(t) > 0$ and has a PDF given by

$$p_X(k) = P(X(t) = k) = \mathrm{e}^{\lambda t} \frac{(\lambda t)^k}{k!}, \tag{5.58}$$

where $\lambda(t)$ is generally time dependent. The mean and variance of a Poisson random process is $\lambda(t)$. So, for a Poisson random process,

$$\mu_X = \mathrm{E}[X(t)] = \lambda(t) \tag{5.59}$$

$$\sigma_X^2 = \mathrm{Var}(X(t)) = \lambda(t). \tag{5.60}$$

The statistical measures above are those necessary to statistically describe digitally modulated signals and also describe most noise processes.

Index

MICROWAVE AND RF DESIGN: RADIO SYSTEMS

Third Edition

Microwave and RF Design: Radio Systems is a circuits- and systems-oriented approach to modern microwave and RF systems. Sufficient details at the circuits and sub-system levels are provided to understand how modern radios are implemented. Design is emphasized throughout. The evolution of radio from what is now known as 0G, for early radio, through to 6G, for sixth generation cellular radio, is used to present modern microwave and RF engineering concepts. Two key themes unify the text: 1) how system-level decisions affect component, circuit and subsystem design; and 2) how the capabilities of technologies, components, and subsystems impact system design. This book is suitable as both an undergraduate and graduate textbook, as well as a career-long reference book.

KEY FEATURES

- The first volume of a comprehensive series on microwave and RF design
- Open access ebook editions are hosted by NC State University Libraries at: https://repository.lib.ncsu.edu/handle/1840.20/36776
- 31 worked examples
- An average of 38 exercises per chapter
- Answers to selected exercises
- Coverage of cellular radio from 1G through 6G
- Case study of a software defined radio illustrating how modern radios partition functionality between analog and digital domains
- A companion book, *Fundamentals of Microwave and RF Design*, is suitable as a comprehensive undergraduate textbook on microwave engineering

ABOUT THE AUTHOR

Michael Steer is the Lampe Distinguished Professor of Electrical and Computer Engineering at North Carolina State University. He received his B.E. and Ph.D. degrees in Electrical Engineering from the University of Queensland. He is a Fellow of the IEEE and is a former editor-in-chief of *IEEE Transactions on Microwave Theory and Techniques*. He has authored more than 500 publications including twelve books. In 2009 he received a US Army Medal, "The Commander's Award for Public Service." He received the 2010 Microwave Prize and the 2011 Distinguished Educator Award, both from the IEEE Microwave Theory and Techniques Society.

OTHER VOLUMES

Microwave and RF Design
Transmission Lines
Volume 2
ISBN 978-1-4696-5692-2

Microwave and RF Design
Networks
Volume 3
ISBN 978-1-4696-5694-6

Microwave and RF Design
Modules
Volume 4
ISBN 978-1-4696-5696-0

Microwave and RF Design
Amplfiers and Oscillators
Volume 5
ISBN 978-1-4696-5698-4

ALSO BY THE AUTHOR

Fundamentals of Microwave and RF Design
ISBN 978-1-4696-5688-5

Published by NC State University

Distributed by UNC Press

NC STATE UNIVERSITY