Chapter

Radio Systems and Computing at the Edge for IoT Sensor Nodes

Malcolm H. Smith

Abstract

Many Internet of Things (IoT) applications use wireless links to communicate data back. Wireless system performance limits data rates. This data rate limit is what ultimately drives the location of computing resources—on the edge or in the cloud. To understand the limits of performance, it is instructive to look at the evolution of cellular and other radio systems. The emphasis will be on the RF front-end architectures and requirements as well as the modulation schemes used. Wireless sensor nodes will often need to run off batteries and be low-cost, and this will constrain the choice of wireless communications system. Generally cheap and power efficient radio front ends will not support high data rates which will mean that more computing will need to move to the edge. We will look at some examples to understand the choice of radio system for communication. We will also consider the use of radio in the sensor itself with a radar sensor system.

Keywords: IoT sensor nodes, wireless systems, 5G, cellular, WLAN, Bluetooth, radar, RF front end, radio circuitry, bandwidth, power requirements

1. Introduction

Many IoT systems consist of networks of sensors, the data from which is brought together and processed to give some desired results. The method used to connect the sensor to the Internet to enable the transfer of data is a design factor that needs to be considered early on. It is not the case that throwing a lot of bandwidth at the problem and processing the data together in a central location are the solution in most cases. The answer is do not "use 5G" for all systems!

There are many trade-offs involved in the specification of a radio system to be used for Internet of Things (IoT). The main trade-off is the bitrate and power trade-off. Sending messages at a high bitrate requires more power than that at a low bitrate. There is also a cost and bandwidth trade-off: high-bitrate solutions require a wide bandwidth which, generally, costs more than the low bandwidth. Finally, there is the often-overlooked fact that high bitrate solutions require high fidelity of the radio which means more receive power. This limits their use to applications where the separation of transmitters and receivers is small.

If cheap high-bitrate wireless communications are not available, what are the alternatives if the sensor needs high bandwidth (video cameras being the most obvious example)? The most obvious solution is to use a wired communication link. This may not be a viable option for many IoT solutions where the access to wired links is not available. The alternative is to use computing at the edge.

By processing the signal at source—computing at the edge—we can decrease the bandwidth requirement, and, hence, we can use an alternative wireless communication solution.

First, we need to review digital communications as they pertain to wireless systems [1].

2. Review of digital communications systems

The wireless systems used in IoT applications will be digital systems. Although the world they are interfacing to is analogue, IoT systems will have a means of converting those analogue signals that come from the analogue world through sensors to digital signals. These digital signals will be processed ultimately in digital computers whether it be on the edge or in the cloud. It is, therefore, instructive to review digital communications systems.

2.1 Introduction

Modern digital communications systems are complex with many layers that interact with each other. In this chapter we are only going to consider the lowest layer of the Open Systems Interconnection (OSI) model, the physical layer. Furthermore we are going to consider this layer from the perspective of the analogue and RF components of the system as they are the parts that drive a lot of the trade-offs for power consumption that will be a major driving factor in the selection of a radio system for IoT.

2.2 Complex signals

In order to understand the concepts of digital communications systems and even the hardware implementation of them, it is necessary to understand complex signaling. We will therefore review the concept of complex signaling and in-phase and quadrature (IQ) modulation.

A complex signal can be represented as a function of time using a complex exponential:

$$s(t) = Ae^{j2\pi ft} \tag{1}$$

Expanding this exponential using Euler's formula gives us this complex exponential function represented using trigonometric functions:

$$s(t) = A\cos(2\pi ft) + j \times A\sin(2\pi ft) \tag{2}$$

where j is $\sqrt{(-1)}$, A is a scaler quantity representing the magnitude of the signal, and f is a scalar representing the frequency in Hz (or cycles per second).

Graphically this can be represented by a vector from the origin with a magnitude of A and an angle relative to the positive real axis of $2\pi ft$. This vector will spin around the origin in an anti-clockwise direction tracing out a circle. The number of times it spins in 1 second is given by the value of f, so for a value of 1 Hz (1 cycle per second), the vector will complete one rotation.

If we introduce another vector with a negative value for the frequency:

$$s(t) = Ae^{-j2\pi ft} = A\cos(-2\pi ft) + j \times A\sin(-2\pi ft) \tag{3}$$

we get a second vector that rotates in a clockwise direction at the same rate and tracing out the same circle. From the trigonometric identities:

$$\cos -\theta = \cos\theta \tag{4}$$

$$\sin -\theta = -\sin\theta \tag{5}$$

we can see that if we sum the two vectors, the imaginary components are canceled and we are left with a real cosine with twice the amplitude of the vector components. This is illustrated in **Figure 1**.

By shifting the vectors by 90°, we get a sine with twice the amplitude of the vector components. This leads us to the basic formulae relating real sinusoidal functions to complex exponential functions:

$$A\ \cos(2\pi ft) = A \times \frac{e^{j2\pi ft} + e^{-j2\pi ft}}{2} \tag{6}$$

$$A\ \sin(2\pi ft) = A \times \frac{e^{j2\pi ft} - e^{-j2\pi ft}}{2 \times j} \tag{7}$$

From these equations real signals separated by 90° can be used to generate complex signals. Signals at 90° such as this are referred to as quadrature signals and the individual component signals as in-phase (I) and quadrature (Q). These signals will be at low frequency and sit either side of 0 Hz (viz. DC)—they are what we call baseband signals. We need a method of transferring these baseband signals to radio frequency (RF) or microwave frequencies.

Using trigonometric identities, we can prove that:

$$A(t)\ \sin(2\pi ft + \theta(t)) = A(t) \times (\sin 2\pi ft \times \cos\theta(t) + \cos 2\pi ft \times \sin\theta(t)) \tag{8}$$

What this tells us is that we can upconvert a baseband signal represented by:

$$I_{BB} = A(t) \times \cos\theta(t) \tag{9}$$

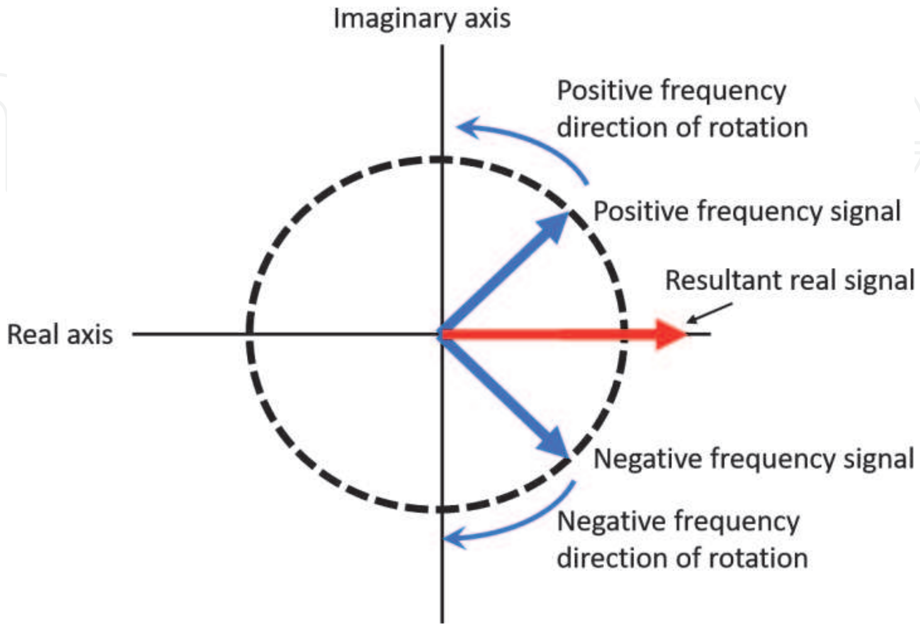
Imaginary axis
Positive frequency direction of rotation
Positive frequency signal
Resultant real signal
Real axis
Negative frequency signal
Negative frequency direction of rotation

Figure 1.
Real cosine signal as a vector sum of two complex signals: A positive frequency signal rotating anti-clockwise and a negative frequency signal rotating clockwise.

$$Q_{BB} = A(t) \times \sin\theta(t) \tag{10}$$

with another signal, known as a local oscillator (LO) signal, represented by:

$$I_{\text{LO}} = \sin 2\pi ft \tag{11}$$

$$Q_{\text{LO}} = \sin\left(2\pi ft + \frac{\pi}{2}\right) = \cos\left(2\pi ft\right) \tag{12}$$

We can implement Eqs. (8)–(12) in circuitry using circuits that can multiply two real signals (circuits of this sort are known as mixers) by signals generated from an oscillator with phase shifters (LO generator) and some sort of circuit to sum the resultant outputs (usually done by summing currents). Such a circuit is called a quadrature modulator or quadrature mixer, and the figurative block diagram is shown in **Figure 2** below with the local oscillator generation.

The receive operation is just the reverse of the above—the RF signal is split into *I* and *Q* components using a quadrature mixer driven from a quadrature LO signal. This is illustrated in **Figure 3**.

2.3 Encryption

On top of all the other parts of the data transmission system is the need to encrypt the data. This needs to happen for the data itself and for any control

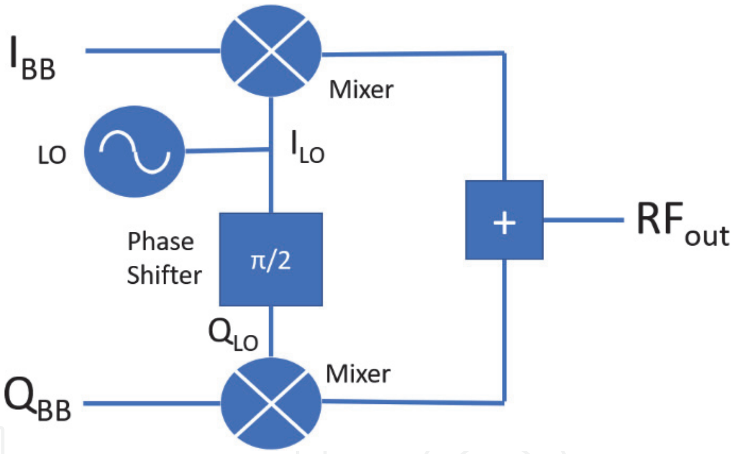

Figure 2.
Quadrature upconverter with local oscillator (LO) and quadrature local oscillator generation.

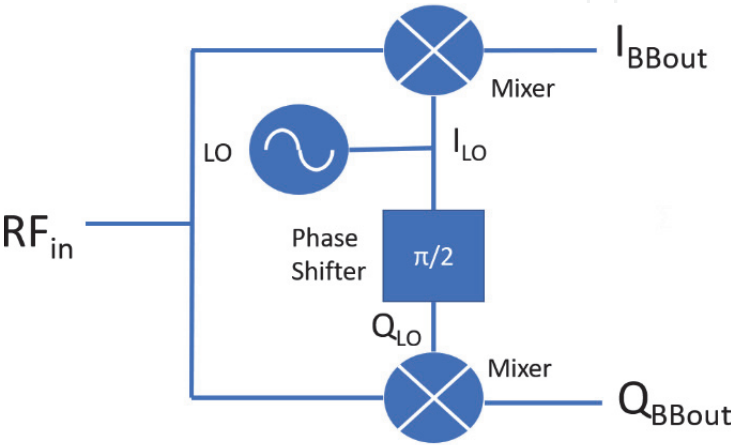

Figure 3.
Quadrature downconverter with local oscillator (LO) and quadrature local oscillator generation.

channels. It is obviously important to encrypt the data itself as this is what has value for the user. However, if control channels aren't encrypted, then an outside agent can read the control information and even take over the channel. This means it is necessary to encrypt control channels as well. Encryption will generally not add much to the overall size of data being sent and so does not affect the efficiency by sending more data. However, the computations involved in encryption do require extra power and so do lower efficiency.

2.4 Interleaving

When a radio receiver is moving, it goes through fades in signal strength due to two or more reflections of the signal interfering. This potentially leaves holes in the data stream, and these holes can lead to loss of information if related bits are sent together. In order to avoid this, the data bits can be separated so that consecutive bits are not sent together. An analogy would be a deck of cards arranged by suit. If four cards are taken together from a single place at random, it could mean, for instance, that the jack, queen, king, and ace of spades are all removed. If the pack is shuffled first, then the four cards taken from adjacent positions will not be related, and arranging the cards in suits, we should be able to spot the pattern even if individual cards are missing. Interleaving is like the shuffling of the pack in the example; however it is done in a controlled manner so that the data can be reconstructed.

2.5 Error coding and correction

Transmission of data will result in errors due to various random processes in the transmission. It is important to be able to a first level detect that an error has occurred and that the data is corrupted. It is much better to be able to correct errors that occur. The simplest method for correcting errors is to have the transmitter retransmit the data if an error is detected or no message acknowledging receipt of the data is received by the transmitter. This method is known as automatic repeat request (ARQ). An improved form of error correction, known as forward error correction, is achieved by adding extra bits to the data that is transmitted to enable both detection and correction of this data. Examples of error correction codes include convolutional codes that are added and processed on a bit-by-bit basis and Reed-Solomon codes or turbo codes that are added and processed on a block-by-block basis. If an error is detected in forward error correction that cannot be corrected, then ARQ can still be used and the data retransmitted. Adding error correction codes adds extra data to the transmission and so decreases the overall power efficiency of the system.

2.6 Modulation

Transmitting data as a series of "0's and 1's" over the transmission medium would require a very wide bandwidth. The required bandwidth is a function of the rise and fall time of the data rather than the clock rate and is significantly higher than the clock rate of the data. To transmit over air for a wireless system, the data needs to be bandlimited first. A filter must be used to bandlimit the data, but the performance of the filter is important. The filter needs to be sharp enough to filter the data to fit in an assigned bandwidth (radio channel) but not destroy the higher frequencies of the data, and so it must have a sharp cut-off. The time domain performance (impulse response) of the filter is also important. If a signal is bandlimited in the frequency domain, then it tends to spread out in the time domain. This is obviously undesirable with a digital signal as one data symbol would interfere with subsequent symbols. There are two approaches to the impulse response requirement:

1. Allow inter-symbol interference to happen, and correct or allow for it in the receiver. In this case the receiver will contain an equalizer to provide an inverse filter to the original filter (and any filtering from the channel as well). Alternatively, the receiver may implement the Viterbi algorithm to decode the data signal.

2. Use a filtering scheme; a root-raised cosine (RRC) filter is common, that has nulls at the subsequent symbols. In this case the filter on the transmit is matched with an identical filter in the receiver. In the case of an RRC filter, the impulse is a sinc function which naturally has nulls at integer time intervals. A data signal passed through such a filter will look like a sequence of overlapping sinc functions with the nulls of the previous sinc functions occurring at the peaks of the subsequent ones. This is illustrated in **Figure 4**.

To transfer a digital bit stream over an analogue channel, it is necessary to use some form of modulation. In modulation, a periodic signal, carrier signal, has its frequency, its amplitude, or phase or both changed—modulated—by a second signal. For digital modulation, this second signal is a filtered version of the digital bit stream. Modulation schemes can be divided into many different groupings—phase modulation, frequency modulation, and amplitude modulation—but for the front end, the grouping that matters is non-linear also known as constant amplitude modulation (AM) and linear modulation. The bit stream is broken up into symbols. Each symbol can be one, two, three, or more bits long. These symbols are used to modulate the carrier. To send a data stream over a link with a given bandwidth, in general a symbol with more bits is needed. This will mean that both amplitude and phase will need to be modulated and the system will be more susceptible to errors.

The simplest modulation to visualize is amplitude modulation (AM). The amplitude of the carrier wave is modulated by the modulating signal. We are familiar with AM radio where the signal doing the modulation is an analogue signal. This modulation can be demodulated by using a simple diode, and in the simplest receivers, this diode was made from a crystal (lead sulphide) with a "cat's whisker' touching it leading to the term "cat's whisker radio" or "crystal radio" [2]. For digital bit-streams, "On–off keying" is the simplest form of amplitude modulation

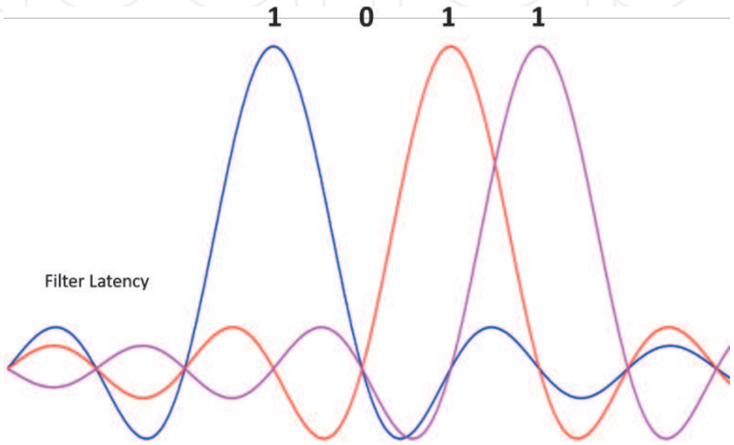

Figure 4.
A filter with a sinc impulse response being fed the bit sequence (1011). The filter latency means that the signals will be delayed with respect to the input.

and is known as amplitude shift keying (ASK). In practice the carrier is usually not switched on and off as this causes spurious signals around the transition but is switched between two amplitudes. A signal that is switched sharply between two different amplitudes is wideband, so a filtered signal is used leading to a smooth transition between the two signal levels (**Figure 5**).

In frequency modulation (FM), the frequency of the carrier is modulated. Frequency modulation is the FM in FM radio, and in this case the modulating signal is analogue [3]. The simplest example of FM use as a digital modulation would be using on–off keying where the carrier is switched between two frequencies. This is known as frequency shift keying (FSK). If more than two frequencies are used, it is possible to send more than one bit at a time—for instance, three bits would require eight different frequencies (**Figure 6**).

Again, to limit the bandwidth, the input signal is filtered, and the transition passes through the intermediate frequencies. A typical version of filtered FSK is GFSK which uses a Gaussian filter before to limit the bandwidth—GFSK is used in many radio systems, Bluetooth being a good example. An efficient form of

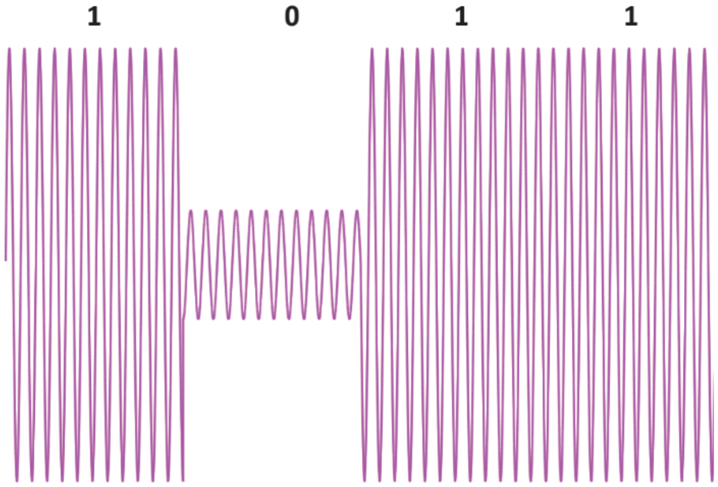

Figure 5.
Amplitude shift keying (ASK) with two levels (no filtering) modulated by the sequence [1011].

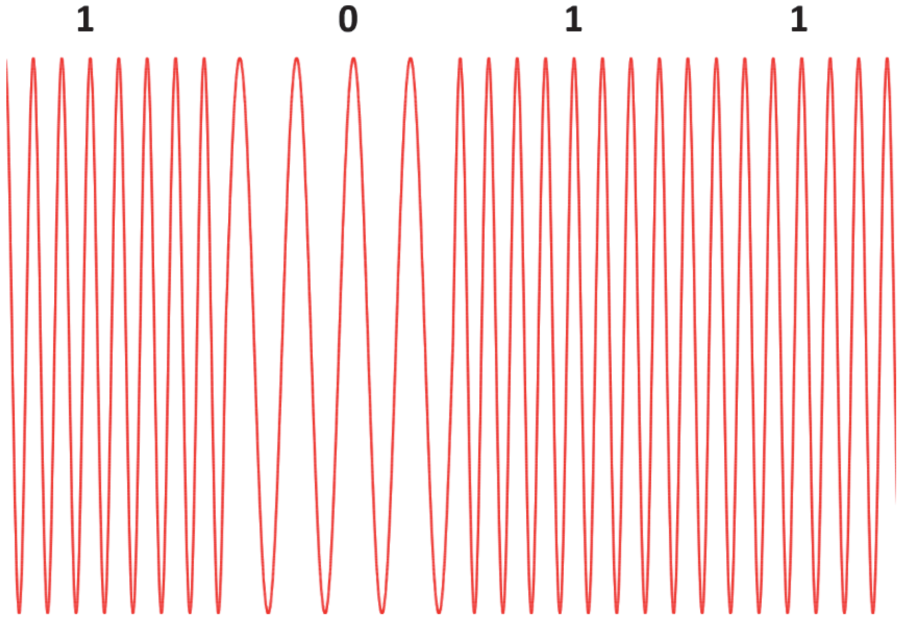

Figure 6.
Frequency shift keying (FSK) with two levels (no filtering) modulated by the sequence (1011).

frequency modulation is minimum shift keying (MSK). In MSK, the two frequencies used differ by half the bitrate, and this gives a very efficient modulation with a modulation index of 0.5. If Gaussian filtering is used to limit the bandwidth with MSK, the resulting modulation is known as GMSK, and this is the modulation scheme used in GSM. In FM radios (digital and analogue), the amplitude of the signal is not important, and so the radio is much more robust and tolerant of signal fading. Therefore, FM radio was the preferred medium for broadcast radio before the arrival of digital radio. However, FM requires a more complicated receiver than AM. As the amplitude is not important—it is a non-linear modulation—the transmitter can be simplified vs. a transmitter used for modulations where the amplitude varies (linear modulations).

In phase modulation (PM) schemes, the phase of the carrier is varied. Again, on–off keying can be used to vary the phase, such a modulation being known as phase shift keying (PSK). The simplest form is binary phase shift keying (BPSK) where only one bit is used, and the phase is varied by 180° to represent a binary "0" or "1". In theory, BPSK, like all pure phase modulations, is a constant envelope modulation (there is no change in the amplitude), but that would require an infinite bandwidth to accommodate an instantaneous 180° phase shift. In practice a BPSK where the bandwidth is limited will have amplitude modulation associated with it. Both unfiltered and filtered BPSK are shown in **Figure 7**. BPSK is the modulation used in Zigbee for the low band.

If a two-bit symbol is to be transmitted, then four phases must be used each 90° apart; such a modulation is known as quadrature phase shift keying (QPSK), and variants are used in many commercial radio systems. For a three-bit symbol, eight phases are needed each 45° apart, and this system is known as 8PSK. The constellation points for QPSK and 8PSK are shown in **Figure 8**. More bits can be added to the symbol, and the number of phases increased for each extra bit by a factor of two, but this rapidly becomes impractical as the tolerance of error decreases as each extra bit decreases the distance between phases by a factor of two. QPSK, 8PSK, and higher orders of PSK have an amplitude component and require linear transmitters. This also means that the power amplifier needs to be linear and so is less efficient than the switching power amplifiers used in non-linear modulations like GMSK or GFSK.

For amplitude modulation and frequency modulation systems, it is possible to consider only one scalar quantity (amplitude or frequency). For phase-modulated systems, we must look at amplitude and phase, and that requires us to look at a complex representation of the signal at baseband (real, imaginary). We can map the signal out on the complex plane and see what it looks like. For a QPSK signal, the four decision points are 90° apart. We can easily visualize the points (1,0), (0,1), (−1,0), and (0, −1). However, as long as the four points are 90° apart, the location

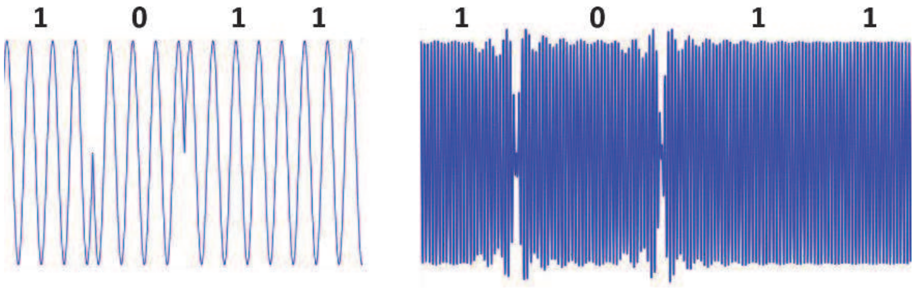

Figure 7.
Binary phase shift keying (BPSK). The left-hand side is without filtering on the bit stream which has shown modulating a low-frequency signal to make the transitions clearer; the right-hand side is with filtering of the data stream but with a higher frequency to make the resulting amplitude modulation clearer.

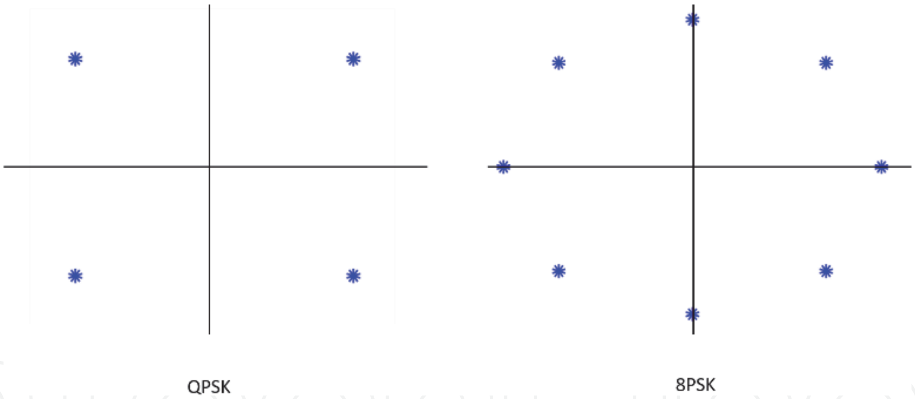

Figure 8.
The QPSK and 8PSK constellation points plotted in the complex domain.

of the initial point can be at any random phase around the circle, and so a 45° phase shift also gives valid points at $(\sqrt{2}, \sqrt{2})$, $(-\sqrt{2}, \sqrt{2})$, $(-\sqrt{2}, -\sqrt{2})$, and $(\sqrt{2}, -\sqrt{2})$. This arrangement always includes a path that passes through the origin—in other words the amplitude goes to zero—and this is bad for the power amplifier as it must be sufficiently linear to be able to handle this change through zero amplitude.

One solution to the need to pass through zero is to offset one of the in-phase and quadrature bit streams by half a clock cycle ensuring that the code points never change by more than 90°, so the modulation will never pass through zero. This scheme is known as offset QPSK (OQPSK). Another solution is to rotate the constellation of points by 45° at each symbol so that there is no path through the origin and the amplitude never goes to zero. This is known as $\pi/4$ QPSK. Both these schemes are used in radio standards to simplify the requirements for the power amplifier (PA). The constellation diagrams of QPSK and $\pi/4$ QPSK are shown in **Figure 9**.

The final sort of modulation we will consider is quadrature amplitude modulation (QAM). QAM modulates both the phase and the amplitude. Generally, it uses points that are in a square centred on the origin, so QPSK is actually a form of QAM. This means that powers of two that are also square numbers are preferred numbers of constellation points: 4 QAM (QPSK), 16 QAM, 64 QAM, 256 QAM, 1024 QAM, etc. A constellation diagram for a 16 QAM implementation is shown in **Figure 10**.

QAM can support high data rates with a relatively small bandwidth, and this makes it desirable in applications where a high bitrate is desired. However, because the spacing between constellation points is less than in other modulations and

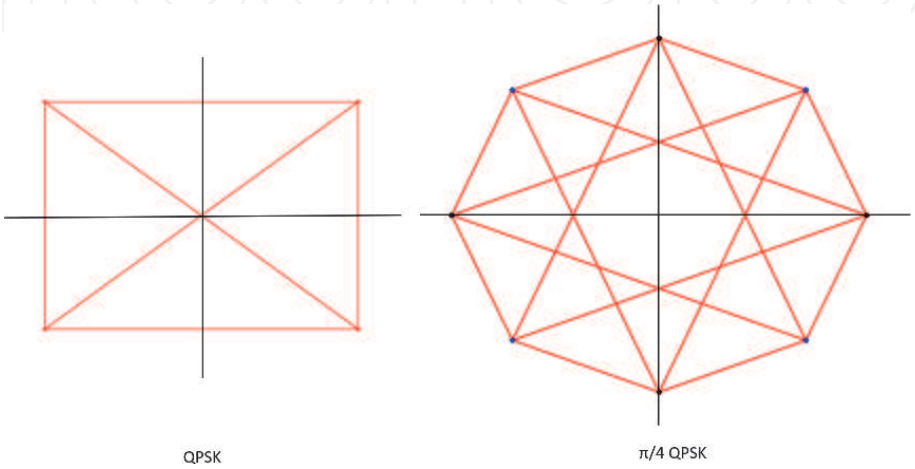

Figure 9.
QPSK and $\pi/4$ QPSK constellation diagrams. Note that the $\pi/4$ QPSK constellation does not pass through the origin simplifying the design of the PA.

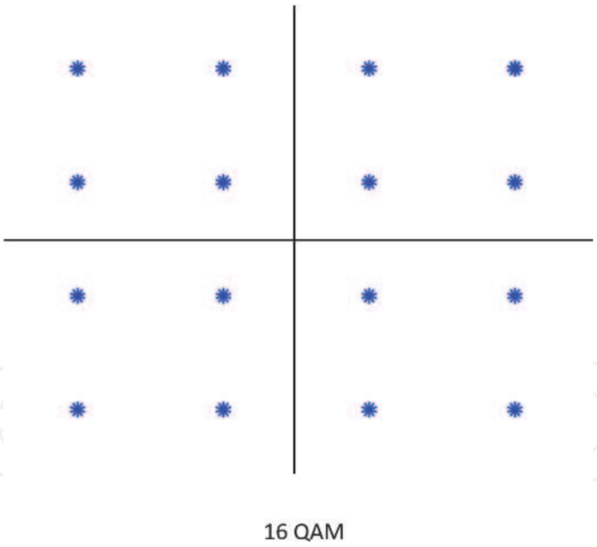

Figure 10.
Constellation points for 16 QAM modulation.

decreases as more points are added, QAM requires more precision in the transmit architecture and is less robust on the receive side to noise and fading. On the transmit side, the transmitter and the power amplifier need to be very linear. On the receive side, the whole receiver needs a much wider dynamic range than other modulations.

2.7 Channel access method

When sharing a communications medium, the users need to be allowed access to the medium in a controlled manner. There are several techniques that can be used to control access to the medium. Usually a combination of several of these methods is needed to get a system to work well.

The first, and perhaps most obvious, is spatial separation known as spatial division multiple access (SDMA). This is the principle used in the cellular network where an area is separated into cells with channel allocation such that neighboring cells do not share the same channels but channels are reused in cells that are sufficient far apart that they will not interfere with another.

The next technique is to separate the channels by frequency using a technique called frequency division multiple access (FDMA). We are all familiar with this from AM and FM radio stations as well as terrestrial broadcast television. In fact, all radios use this technique in some form or another; as radio signals of a certain type are restricted to certain bands, the access to most is controlled by a government agency. Within a radio band, users can be assigned channels that are also of different frequencies and are a sub-band of the overall radio band assigned to that system.

The next technique is time division multiple access (TDMA). In this case the resource is split into units of time generally called slots. A user is allocated a slot to transmit on and is quiet in other slots. There will generally be a matching slot (or slots) on the receive side to receive transmissions. TDMA is good for transmissions that are bursty in nature which many IoT applications are. TDMA is used in the GSM system.

The final technique is code division multiple access (CDMA). Each user is allocated a code that is unique to them and mathematically orthogonal to other user codes. The data stream is multiplied with a much faster version of the code and transmitted. On the receive side, the receiver uses the same code and again multiplies the incoming data by that code to decode the message. The principle is simple, but the actual implementation is complex because of the need to time align the code with the received signal and the presence of reflected copies of the signal.

2.8 Orthogonal frequency division multiplexing

Orthogonal frequency division multiplexing (OFDM) is a technique mainly used to transmit wideband data. OFDM has a number of advantages for the transmission of wideband data which is why it is used in all new wideband systems:

1. It is robust against narrow band interference from other radios.

2. It is robust against fading.

3. It deals with multi-path easily.

However, is also has disadvantages:

1. It is sensitive to Doppler shift.

2. It has a high peak-to-average ratio (PAR) which requires a linear power amplifier and a lot of current.

For IoT systems that aren't designed to be particularly mobile, the sensitivity to Doppler shift is probably not an issue. However the inefficiency of a good linear power amplifier will be a significant issue in systems where battery current is premium.

In an OFDM system, the radio band is split into a number of sub-bands with sub-carriers which are modulated separately. Each sub-carrier needs to be orthogonal to the others which gives the relationship:

$$\Delta f = \frac{1}{\mathrm{T_s}} \tag{13}$$

where Δf is the difference in the sub-carrier frequencies and T_s is the receive symbol duration. This gives an overall bandwidth of:

$$B \cong \Delta f \times N \tag{14}$$

where B is the bandwidth and N is the number of sub-carriers. Bandwidth efficiency is high as the orthogonal nature of the sub-carriers ensures that adjacent sub-bands do not interfere with the demodulation of a sub-band, so no guard bands are needed, and spectral efficiency is high (**Figure 11**).

Within each sub-band, the sub-carrier is modulated just as it would be in a single carrier system. QAM-based modulations are common starting with QPSK, although 802.11a allowed BPSK, and increasing the symbol size and modulation as the communications channel quality improves.

Usually the centre sub-carrier of the overall bandwidth is not modulated. This assists on in the receiver where it converts down to DC. Modern receivers are generally what is known as direct downconversion receivers, and the control of DC offsets is an issue. With no useful signal around DC, the DC offset can be eliminated using simple filtering schemes.

A variant of OFDM, orthogonal frequency division multiple access (OFDMA), shares the band between users rather than dedicating the whole band to a single user.

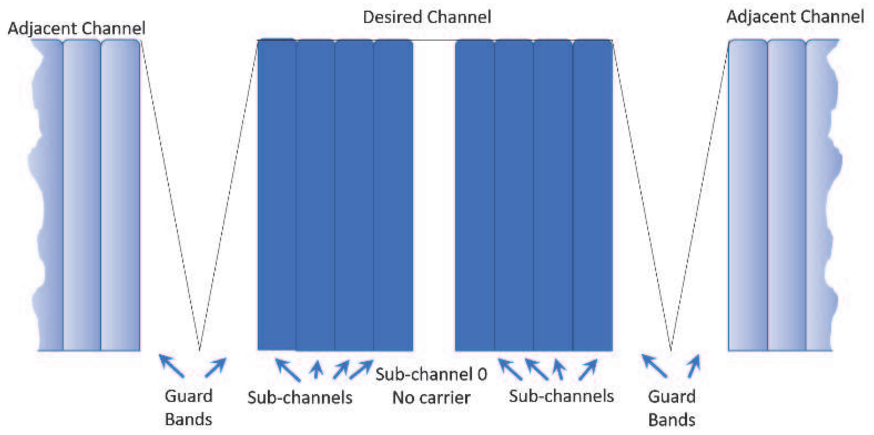

Figure 11.
OFDM channel structure.

This is done by designating each user a group of sub-carriers. It is used on the downlink of the cellular standards of the Long Term Evolution (LTE) family and 5G.

A variant of OFDM, single carrier frequency division multiple access (SC-FDMA), is used in the cellular standards on the uplink. In SC-FDMA, the bit stream is processed in a manner similar to OFDMA, but the parallel streams are then serialized to give a single carrier modulation. This gives a signal with much less PAR which requires a less linear power amplifier and transmit section and so saves power. As user equipment is battery driven and therefore sensitive to power consumption, this is necessary for the quality of service it can deliver customers.

2.9 Half- and full-duplex

How a radio system handles the relationship between transmit and receive has a large impact on the design of the radio front end and the power consumption of the overall radio. It is possible to transmit while receiving, and a lot of modern radio systems have this capability. These systems are known as full-duplex systems. Obviously, if they are to transmit while receiving, they cannot use the same frequency; otherwise they block themselves. (The exception here is radar which we will cover later.) The systems that do not permit transmission at the same time as reception are called half-duplex systems and are significantly simpler and more power efficient.

There are two approaches to duplexing: frequency division duplex (FDD) and time division duplex (TDD). FDD is the only approach that allows for full-duplex operation as the transmitter and receiver are operating at different frequencies. Usually this requires a front-end filter called a duplexing filter or duplexer that filters the transmit out on the receive path and filters transmit noise in the receive band out on the transmit path. Cellular radios usually separate the transmit and receive bands and so are capable of full-duplex operation even though some of the earlier standards did not call for it.

TDD is a half-duplex technique that is used a lot in connectivity radios (e.g. WiFi, Bluetooth) and is specified for some cellular radio standards but is not as common there. A TDD radio uses the same frequency for transmit and receive and will transmit and then switch to receive to listen for any reply.

3. Overview of radio transmit and receive architectures

Figure 12 is a block diagram showing the structure of a constant envelope transmitter. As the data is only coded in the phase, the amplitude contains no

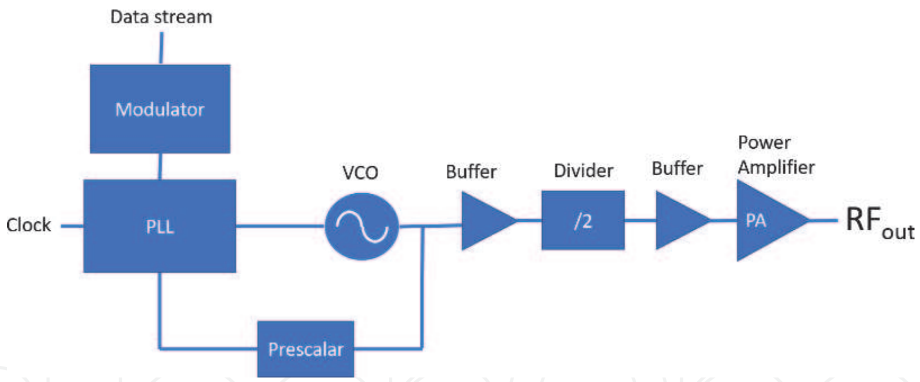

Figure 12.
Constant envelope transmit employing a PLL.

information. It is still necessary to control the amplitude in many radio systems to ensure that the signal received by the base station is not too large or too small. However, this amplitude control only needs to be accurate enough to ensure the signal is within a few dB of target and can be relaxed compared to a system that uses linear modulation.

With no requirement to provide modulation on the amplitude, only a circuit that modulates the phase is required, and a PLL is the best circuit for the job. The first transmitters of this type were introduced for GSM and used offset-loop PLLs to be compatible with the Cartesian baseband circuit outputs available at the time. These were quickly replaced by what are known as fractional N (frac-N) frequency synthesizers. In a frac-N frequency synthesizer, the divider is controlled by a sigma-delta modulator, and the divide ratio is constantly updated between integer values allowing the effective average ratio to be a fraction of the nominal integer divide ratio. This operation is carried out in the digital domain making circuit design easier.

The noise from the PLL and associated circuits is filtered by the PLL filter. This means that the noise far away from the carrier is dominated by the voltage-controlled oscillator (VCO), divider, and power amplifier (PA) noise. As noise away for the carrier is important in radio systems, this is a big advantage for these constant envelope transmitters. It generally means that the filtering at the output of the PA can be a simple harmonic trap filtering—filters out harmonics of the transmit signal—and does not need to filter close to the transmit signal. Harmonic trap filters can be implemented using standard passive components (inductors and capacitors) not specialized components (SAW, BAW, or FBAR filters), and these filters will generally have lower loss (meaning the overall transmitter is more efficient).

The design of the VCO is critical. It needs to have a very low noise profile as this noise will dominate the noise of the transmitter, and it needs to be able to drive the output well. Following the VCO is a divider. This is usually a divide by two dividers, so the VCO frequency is set at twice the desired output frequency. The use of a frequency of twice the transmit frequency is to stop feedback from the PA causing the VCO to shift off frequency. This effect is known as pulling and happens if the output signal is at the same frequency as the VCO frequency as the VCO will injection lock to the PA output signal. In some cases, even with a VCO being run at twice the transmit frequency, the VCO can still be pulled if the PA produces lots of second harmonic distortion. In this case a VCO frequency of four times the transmit frequency can be used with cascaded dividers that divide by two each. Running the VCO at twice or four times the transmit frequency obviously has power implications because the transistors will have less gain at higher frequencies. This is an unavoidable trade-off in these designs.

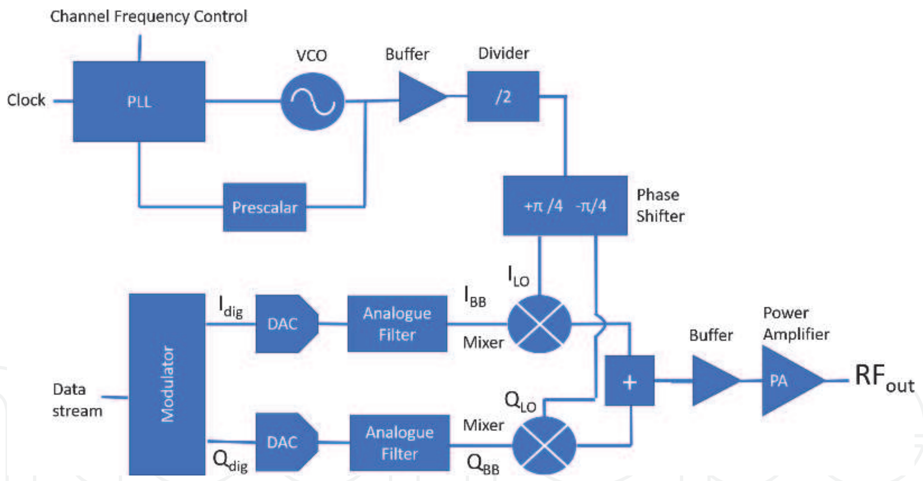

Figure 13.
Cartesian upconverter-based transmit for linear modulations with a PLL to control the local oscillator. Modulation is done on the digital domain giving digital I *and* Q *signals which are converted to analogue for upconversion.*

Figure 13 is a simplified block diagram of a typical transmitter used for cellular and connectivity solutions in smartphones. This sort of transmitter is known by many names, but *I* will refer to it as a Cartesian transmitter here. The baseband signal is coded as a complex signal with an in-phase (I) and quadrature (Q) component. This allows us to control both the amplitude and the phase and so is used for linear modulations where information is coded in both the amplitude and the phase.

It is obvious that this transmitter is a lot more complex than the constant envelope transmitter. In fact, the Cartesian transmitter uses most of the components of the constant envelope transmitter just to generate a constant local oscillator (LO) signal. The complexity not only means an increase in cost but also an increase in power consumption.

The linear modulation signal has both amplitude and phase components, and so the signal chain needs to be linear. This means it needs to have a constant DC bias current flowing and so is less efficient than a non-linear equivalent that can just operate as a digital logic gate switching between states. This inefficiency shows up particularly in the PA which can have half the efficiency of its non-linear counterpart.

One extra effect to consider for a transmitter suitable for linear modulations is the noise. It is a general rule that any component added to a system will add noise. Whether this added noise is large enough to affect the overall signal to noise of the system is somewhat a matter of design. In the case of the Cartesian transmitter, the signal path adds a substantial amount of noise to the signal being transmitted. In past the signal was usually cleaned up by adding a filter between the transmitter and the PA. This was easily achieved as the transmitter and the PA are usually on separate integrated circuits even now, and so the signal had to pass through the main circuit board to which a filter could be easily added. However, filters are expensive, and as the number of bands increased, the transmitter had to serve more bands which meant more filters, and some form of isolation was needed for unused bands. The filter was eliminated by cutting the noise of the transmit chain itself.

The VCO for the LO needs to be at a different frequencies from the transmit for the same reason as the constant envelope transmitter, namely, VCO pulling. If the VCO is not being operated at some integers multiple of the desired frequency, then extra circuits are needed to generate the desired frequency. The VCO can operate at twice the desired frequency, but if possible, four times makes more sense as it is easier to generate LO signals that are exactly 90° apart. Having LO in-phase and quadrature signals that are exactly 90° apart helps keep the number of errors down in the overall transmitter.

Figure 14 shows a simplified receiver. It looks somewhat like a Cartesian transmitter in reverse. The PA is replaced by a low noise amplifier (LNA), and again two LO signals that are 90° apart are used to downconvert to baseband. This scheme is called direct downconversion. In the past, heterodyne receivers were used, and the signal was converted down through one or more intermediate frequencies (IFs) before being brought down to the baseband. Consumer radio receivers, including those for the cellphone network where more performance is needed, are all direct downconversion receivers.

There are no pulling issues as such in the receiver, so the VCO for the LO can be at the same frequency as the receive signal, but running at twice or four times the frequency makes the generation of the quadrature components easier using a divided down VCO. For a full-duplex radio, it is possible to share the VCO between the receive and the transmit, but this requires a method to derive the transmit frequency and receive frequency from the same source. Two synthesizers can be used, but care must be taken to ensure they do not talk to each other.

On the receive side, as on the transmit side, most of the power consumption occurs in the RF blocks. These are the LNA, mixer, and LO buffer. The baseband filtering will not consume as much power, but the power consumption in the baseband filters is a function of the bandwidth of the signal: wider bandwidth leads to more power consumption. Having a narrower bandwidth not only lowers the power requirements for the filtering but also allows for more resolution in the ADC as it can run at a lower frequency than a wider bandwidth radio, and resolution and bandwidth tend to have a reciprocal relationship. With higher resolution on the ADC, it is possible to move more of the filtering and gain functions into the digital which makes the design easier, smaller, and often lower power.

Figure 15 shows the radio front end for a half-duplex radio (the transmit and receive never operate at the same time) and a full-duplex radio. Both use frequency division duplexing (as would be seen in a cellular radio system). Many elements overlap with the previous figures showing transmit and receive structures. This is so that you can see how everything fits together. In the cellular world, the front end started out as discrete components. Later, the LNAs were integrated into the transceiver chips, but the other RF components were kept separate and integrated into

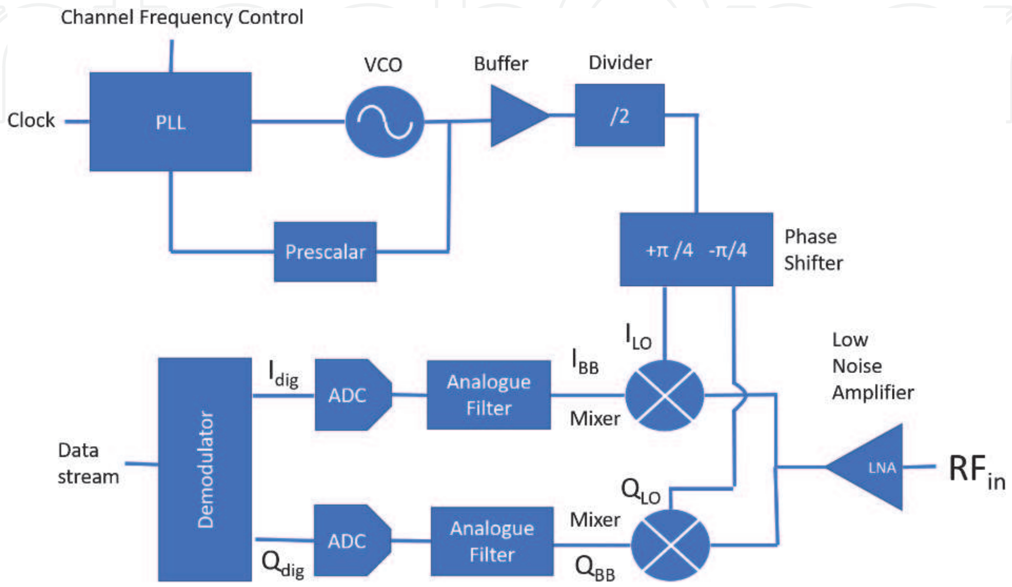

Figure 14.
Typical architecture for a direct downconversion receiver.

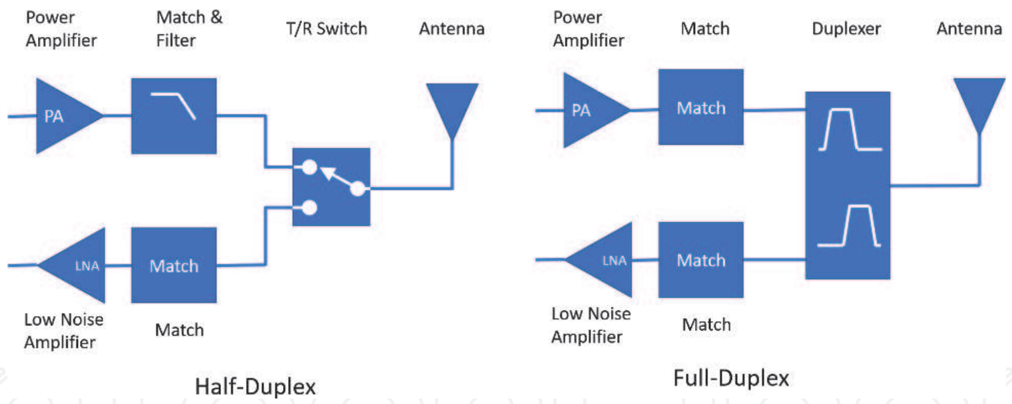

Figure 15.
Half-duplex and full-duplex front ends.

their own modules. The recent trend has been to take the LNA out of the transceiver chip and integrate it back with the RF front end in a module.

In the half-duplex radio, the filters can be relatively loose and made using standard passive components, so they tend to have less loss and can even be integrated with the power amplifier easily. Both the PA and LNA require matching circuits. For the PA the match is a power match and can often be combined with the trap filtering required. For the LNA a noise match is needed. The switch loss needs to be added to the filter loss on the transmit, but generally, the overall loss can be less than 2 dB. A receive channel filter is optional, and, in fact, modern, well-designed receivers do not need it, and so on the receive side, the front end can have less loss, be cheaper, and be relatively easy to integrate.

For a full-duplex radio, the duplexer needs to reject the transmit as much as possible on the receive side but also reject the transmit noise which lies in the receive band. The transmit noise in the receive band will be the limiting factor on the receiver performance. Duplexers are made using special processes—SAW, BAW, or FBAR—and add a significant cost to the front end. They tend to have a loss of 3 dB or more in the transmit path which means half the power put out by the PA is lost in the output network which directly translates to a need for a bigger battery.

The block diagram for a TDD radio front end is shown in **Figure 16**. Duplexing is done using time separation of transmit and receive, and the same frequency band is used for both. This means no duplexer is needed and the filtering can be handled by the same band filter for both transmit and receive. In some applications this filter can be omitted, but in applications where performance out of band is important (some of the TDD versions of the cellular standards and WiFi, for instance), it is often needed.

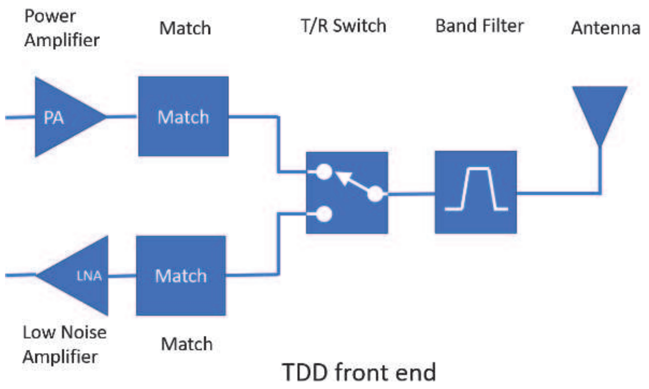

Figure 16.
TDD front end with band filtering.

4. Trade-offs in radio systems for communications

There are five choices that directly affect the amount of current needed in the RF front end for the transmit operation:

1. Frequency of the transmitter

2. Bandwidth of the signal

3. Number of radio bands supported

4. Type of modulation—linear or non-linear

5. Full-duplex or half-duplex

6. Output power

For the receiver, the list looks similar:

1. Frequency of the receiver

2. Bandwidth of the signal

3. Type of modulation—dynamic range required

4. Full-duplex or half-duplex

5. Presence of interfering signals

We will look at each individually.

4.1 Frequency of the transmitter and receiver

The frequency that the transmitter and receiver work at determines the current required. The higher the frequency, the more current are required as the transistors have less gain at higher frequencies. This immediately constrains the bandwidth of communication that can be handled as the bandwidth available increases with increase in frequency.

Another advantage of going low in frequency is that the signal will suffer less loss and can travel through obstacles better. If an IoT system is needed to operate throughout a building including traveling through walls, then it is better to operate at a lower frequency than at a higher one. This is readily seen in WiFi where the 2.4 GHz signal will be found to have more reach than the 5 GHz signal from a dual band router.

One disadvantage of using a low frequency is that the antenna size increases. If the system requires many antennas (as a MIMO system does), then this can be problematic.

4.2 Bandwidth of the signal

The bandwidth of the signal determines the noise level of the system as white noise is being integrated over the bandwidth. As the bandwidth increases, the noise level increases with it, so to maintain the same signal-to-noise ratio, the signal must be increased which increases the current. Also, at a given bias current, the gain-bandwidth product of transistors is fixed. As bandwidth increases, the bias current of transistors needs to increase to increase the gain-bandwidth product and maintain the gain at the same level.

4.3 Number of radio bands supported

Increasing the number of bands requires that some mechanism be found to stop all the circuitry for the radio bands that are not being used interfering with the transmission in the band that is being used. The usual method used to isolate bands is to incorporate a “band switch” in between the antenna and the output of each band. This switch is turned on when the band is to be used. However, when the power amplifiers for the radios are all on the same die, coupling increases, and the harmonics reaching the switch may be high enough that the switch does not have enough isolation. In the case where the switches are all on one die (the cheapest option as they can share circuitry), then the isolation can be further reduced by coupling.

The biggest issue with supporting a large number of bands is the necessity of providing filtering on these bands, especially in the case of full-duplex radios. As the number of bands and, therefore, filters increases, the cost increases. This is likely to lead to IoT solutions needing a separate RF front end for each geography covered so as to keep the number of bands manageable.

4.4 Type of modulation: linear or non-linear

The transmitter for non-linear modulation offers significant savings in power and complexity over a transmitter for linear modulation. A frequency synthesizer is all that is needed for most non-linear modulations, whereas a linear modulation would need the frequency synthesizer and a linear upconversion transmit. Finally, the power amplifier for a non-linear modulation can be a switching power amplifier; a linear modulation needs a linear power amplifier which is not as efficient.

Linear modulations offer the prospect of much higher bitrates, but this comes at the expense of power consumption. If you are trying to run at high data rates, then you will have to accept that your system will be power hungry.

4.5 Type of modulation: dynamic range required

On the receive side, the same type of receiver RF and analogue circuits are used for both linear and non-linear modulations, and so there is no difference seen between these types of modulation. However, the dynamic range required for the modulation has a large effect on the current requirements of the receiver. In general, a switch to higher orders of QAM means that a higher signal-to-noise ratio is required of the received signal than modulations like QPSK or GMSK, and so this means a wider dynamic range is required in the overall receiver.

4.6 Full-duplex or half-duplex

A full-duplex radio offers the prospect of higher bitrates on both the transmit and receive because both are working at the same time. Unfortunately, this means that more spectrum must be used to cover the transmit and the receive functions. A duplexer is necessary in this case, and this not only adds cost and complicates the front end design but also increases the loss in the transmit path. This increased loss leads to a lower efficiency for the transmitter and higher current draw for the same output power. On the receive side, the filtering must remove the large transmit signal, and this tends to lead to a larger loss for the desired signal as well. This loss must be compensated for with a lower noise figure in the receiver circuits, and this requires more current.

Half-duplex radio front ends are far simpler and have less loss. Most IoT data traffic is low data rate or bursty in nature, and this sort of traffic is well served by half-duplex radios. It should also be noted that the radio systems that use the ISM frequencies are half-duplex by necessity. As many IoT systems will make use of these radio systems, they are inherently half-duplex.

4.7 Transmit output power

Power is the product of current and voltage. Increasing the power requirements while keeping the voltage the same will require more current. For battery-powered devices, the voltage is fixed by the battery chemistry and by the reliability constraints of the integrated circuits being used, and so if more output power is required, then more current will need to be taken from the battery.

4.8 Presence of interfering signals

Large interfering signals on the receive side will affect the radio receiver by limiting the dynamic range of the receiver. If the signal is also large, then this may not be a problem. However if the desired signal is a lot smaller than the interfering signals, then the dynamic range of the receiver needs to be increased to receive both the interfering signal and the desired signal. In order to increase the dynamic range, it is necessary to increase the current in the circuits.

5. Overview of cellular radio systems

The most widely used and pervasive radio systems are cellular radio systems. Originally just an extension of the telephone service, cellular radio systems have grown into much more enabling data download and upload and giving access to the Internet virtually anywhere. For IoT applications that need mobility or access to the Internet from remote locations, a cellular radio will be the first consideration, and so we need to consider the alternatives. It should be noted that the need to pay royalties for IP included in the cellular specifications may make cellular a more expensive option than some of the other radios.

Cellular radios work on the principle of multiple transmitters and receivers separated into cells (hence the name), each controlling the communication with end users in the cell. The system works based on spatial separation: adjacent cells use a different set of frequencies, but cells sufficiently far away can re-use the same set of frequencies. Hand-off of mobile users is an important function of a cellular system, and a lot of resources are assigned to this function. A cellular network of cells each with its own base station is shown in **Figure 17**.

5.1 Introduction

Cellular radio systems are so-called because the coverage area is broken up into adjacent cells, each covered by a single wireless communication receiving and transmitting station called a base station. The original cellular radio systems were analogue and did not use digital modulation or any of the other techniques used in

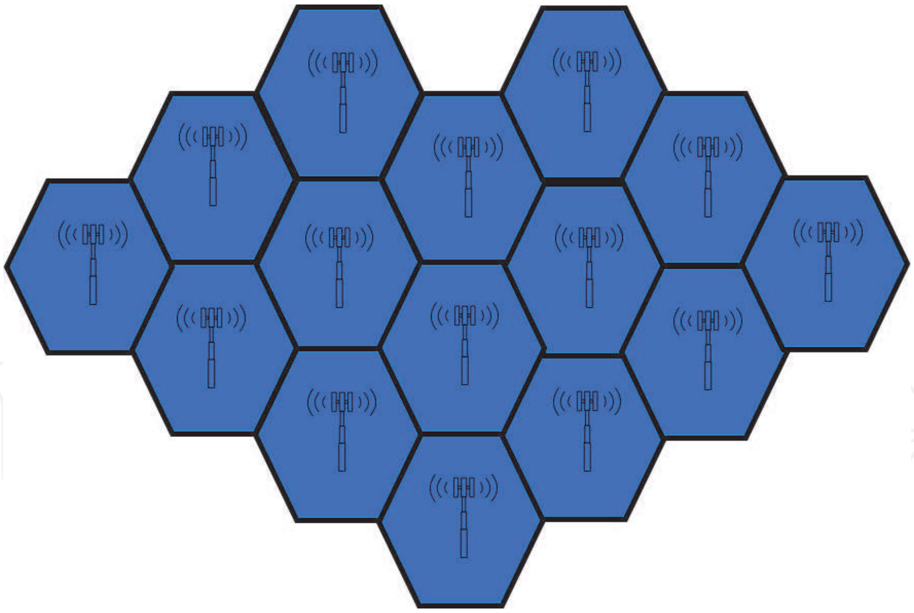

Figure 17.
A cellular radio layout with a base station at the centre of every cell serving that cell.

digital communications systems. We will not discuss those as they have been superseded and are of no relevance to IoT.

The first digital systems were the so-called 2G systems—analogue-based systems being 1G. There were multiple 2G systems deployed around the world—IS-136 (also known as TDMA) and IS-95 (generally called CDMA) in the USA and some other countries, PDC in Japan, and GSM throughout Europe and later most of the rest of the world.

The original 2G systems were circuit switched and served the voice market. Later, GPRS, a packet switched network, was added to the GSM system and allowed data communications to happen over the original GRPS network. Enhanced Data Rates for GSM Evolution (EDGE) was an upgrade of GPRS with a higher data rate. Some people used the names 2.5G to describe GPRS and 2.75G for EDGE.

High-speed data communications came with the 3G networks. All systems for 3G communications were based on CDMA technology. The GSM world migrated to Universal Mobile Telephone System (UMTS), while IS-95 CDMA networks migrated to CDMA2000. Further enhancements to the UMTS system— high-speed downlink packet access (HSDPA), high-speed uplink packet access (HSUPA), and high-speed packet access (HSPA) with further evolved versions adding a "+" (HSDPA+, HSUPA+, HSPA+)—allowed even higher data rates. The IS-95-based networks were able to upgrade to a new standard called evolution data optimized (EVDO). Following the naming convention adopted for higher-speed 2G networks, these enhancements were called 3.5G and 3.75G by some.

The GSM-based and IS-95 CDMA-based worlds came together with Long Term Evolution (LTE). For the first time, the world had a single cellular standard. This was named 4G by the marketing folks even though it did not meet the requirements for a next generation system as defined by the International Telecommunication Union (ITU) in its International Mobile Telecommunications Advanced (IMT Advanced) specification. In particular, while LTE is able to meet speeds of 100 MBps in the downlink and 50 MBps in the uplink, the IMT Advanced specification calls for 100 MBps to a fast moving vehicle and 1 GBps to a stationary device. The modulation in LTE changed to orthogonal frequency-division multiple access (OFDMA) on the downlink and single carrier frequency domain multiple access (SC-FDMA) on the uplink.

An enhanced version of LTE—LTE Advanced—does meet the requirements for a next generation system as defined by the International Telecommunication Union (ITU) in its International Mobile Telecommunications Advanced (IMT Advanced) specification.

The latest cellular standard, 5G, is being deployed now. With 5G, data rates have been increased up to a promise of 2 Gbps for some networks in ideal circumstances and latencies of 4 ms or less, with 1 ms being an oft-quoted target.

5.2 GSM/GPRS

The work to define a European cellular system started in 1983 under the Groupe Spécial Mobile (GSM) committee. This acronym later came to stand for Global System for Mobile Communications. The first systems were deployed in 1991. Over the years GSM and its packet radio version general packet radio service (GPRS) became the standard for cellular coverage. For IoT systems that need to work worldwide or in rural areas, GSM, or its IoT version extended coverage GSM IoT (EC-GSM-IoT), is still the system to use as it is the only one that is likely to have coverage.

GSM uses a combination of frequency division duplex (FDD), frequency division multiple access (FDMA), and time division multiple access (TDMA) to allow multiple users to use the system [4]. The transmit and receive frequency bands are separated (on the user side, the transmit is the lower-frequency band, and the receive is the higher-frequency band), with equal bandwidth dedicated to both. Each band is spit into channels of 200 kHz. Within a channel, TDMA is used to control access. Each channel is divided into frames of 4.615 ms each, and each frame is divided into eight time slots (or bursts) that are allocated between users. Users can use up to four of the eight slots in most systems, although there are higher-order specifications where more time slots are available, they tend not to be used. The maximum theoretically achievable bitrate is 114 kbps, but deployed systems do not get close to this limit.

The modulation used in GSM and GPRS is GMSK and is, therefore, a constant envelope modulation which lightens the requirements on the transmitter and power amplifier. Although there are classes of GSM that can be full-duplex, most implementations are half-duplex, and this removes the need for a duplexer. The combination of a constant envelope modulation and no requirement for a duplexer makes GSM transmitters very efficient.

The low bands for GSM allow for a transmission of up to 2 W (33 dBm) in the low band and 1 Watt (30 dBm) in the high band. The high output power with a timing advance specification in GSM that allows operation out to 35 km means that GSM is the ideal choice for operation in rural environments where, quite apart from the fact that it may be the only system available, the base stations will tend to be spread further apart.

It should also be noted that GSM has the capability to handle text messages in the form of short message service (SMS) which can be used to send small amounts of data. This capability could be useful for some IoT systems where only a small amount of data is typically sent. Each text message is 140 bytes although extensions allow more data to be sent by breaking that data up across multiple messages.

5.3 Edge

Enhanced Data Rates for GSM Evolution (EDGE) as the name implies is an extension of GSM and GPRS to further increase the maximum possible data rate [4].

The highest theoretical data rate is over 384 kbps, the threshold for a 3G system, but EDGE networks generally do not get near this number.

EDGE generally fits on top of the GSM system: it uses the same channel bandwidth as GSM (200 kHz) and similar filtering (Gaussian) and when the channel conditions are not good will use GMSK modulation. However, with the best conditions, it can use a new modulation: 8PSK with a 3π/4 shift between symbols to avoid the zero crossing. The symbol for this new modulation is three bits long, and so, theoretically, three times the data can be sent over the channel. The 8PSK modulation has an amplitude component, and the strict adjacent channel requirements inherited from GSM mean that, unfortunately, the power amplifiers and transmit chain used in EDGE have relatively low efficiencies (higher teens percent than over 50% for GSM). This meant that there was a relatively large market for a hybrid solution with full EDGE capability on the downlink but only GPRS modulation using GSM components on the uplink to save on battery life. EDGE networks have mostly been superseded by 3G and LTE networks, and so it has little relevancy for IoT.

5.3.1 EC-GSM-IoT

Extended coverage GSM IoT (EC-GSM-IoT) is an extension to the GSM specification to allow it to work for IoT devices. The system was specified in such a way that the base station could be upgraded in software. As no hardware changes are possible, there are no changes to the modulation. Devices that support EC-GSM-IoT can be GMSK only or GMSK and 8PSK capable.

5.4 The UMTS: 3G system

The Universal Mobile Telecommunications System (UMTS) often known as Wideband-CDMA is the first true 3G system. The initial specification is known as R99 (Release 99) and is controlled by an organization known as 3rd Generation Partnership Project (3GPP) [5]. The GSM specification [4] was also rolled into 3GPP and is controlled by the 3GPP.

Several changes were introduced including the replacement of the TDMA used in GSM and EDGE with a code division multiple access (CDMA) to enable sharing of the medium. The bandwidth of the channel moved from 200 kHz to 5 MHz. The filtering was changed from a Gaussian filter to an approximation to a root raised cosine (RRC) improving the inter-symbol interference characteristics. The modulation used on the downlink is QPSK and OQPSK is used on the uplink. This means that the modulation has an amplitude content, but the adjacent channel requirements in UMTS are not as stringent as in EDGE, and therefore the PA needs less linearity and can be more efficient, with efficiencies over 40% on the uplink being common.

The system is full-duplex, meaning a duplexing filter (duplexer) is needed. This filter has more loss than the simple harmonic trap filtering that can be used in GSM because it is a half-duplex system and adds to the transmit losses of UMTS. Baseband processing in the system changed completely with the move to CDMA, and a lot more processing was added. Overall power consumption went up considerably with UMTS. Smaller geometries for the chips and new techniques have brought that power consumption down, but it is still higher than what can be achieved for a GSM system if that GSM system is optimized.

Later the air interface was changed and new modes added—high-speed downlink packet access (HSDPA), high-speed downlink packet access (HSUPA), and high-speed packet access (HSPA). There were many changes including changes to modulation available where the channel could support a higher data rate. In HSDPA 16 QAM was added to the downlink. Later a further enhancement was added in Release 7 of the 3GPP specification and was known as evolved high-speed packet access (HSPA+) which used up to 64 QAM modulation and promised theoretical speeds of 337 Mbps in the downlink and 34 Mbps in the uplink.

As a legacy system, there may be some IoT applications that make use of the 3G standards, but in many parts of the world, 3G has been superseded by Long Term Evolution (LTE) systems. For rural applications, GSM remains the only system with significant coverage throughout the developed and developing world.

5.5 LTE: popularly known as 4G

Long Term Evolution (LTE) is the standard that most networks, certainly in cities, are operating on. It is usually called 4G to differentiate it from 3G even though it does not meet the criteria for 4G as given by the ITU. LTE has theoretical data rates up to 300 Mbps on the downlink and 75 Mbps on the uplink [6]. The system is designed to operate out to 100 km with what is defined as acceptable performance and so could be used in a rural setting. The bandwidths used for the signal are flexible and vary between 1.4 and 20 MHz. LTE is available in frequency division duplex (FDD) and time division duplex flavors. With the addition of carrier aggregation, even wider bandwidths are effectively available at the expense of complexity in the specification, design, and testing.

The modulation was changed to orthogonal frequency division multiple access (OFDMA) for the downlink and single carrier frequency division multiple access for the uplink to limit the peak-to-average ratio (PAR) so that the power amplifier does not have to operate backed off by a lot and so conserving power. With special filters and power supply control, the power consumption of an LTE smartphone is reasonable in an urban context with base stations spaced relatively closely.

The change to OFDMA increased the spectral efficiency of LTE vs. UMTS by a factor of up to five times. As spectrum must be bought from the government for substantial fees, this is obviously attractive for cellular carriers. We should bear in mind, however, that spectral efficiency may not be factor in an IoT application.

The LTE channel is divided by time and frequency into units called resource blocks (RB). Each RB is 0.5 ms long and 180 kHz wide made up of 12 15 kHz subcarriers. A user can be assigned a minimum of 2 RB in a 1 ms sub-slot. The more resource blocks a user is assigned, the higher the data rate is available to that user. The number of resource blocks available is dependent on the channel bandwidth and varies from 6 for a 1.4 MHz bandwidth up to 100 for a 20 MHz bandwidth. As can be seen, as each RB is 180 kHz, the 1.4 MHz bandwidth has 1.08 MHz of used bandwidth and 160 kHz of guard band on each side. For the 20 MHz case, the guard band would be 1 MHz on each side.

While the whole world is covered by four frequency bands for GSM (two in most of the world and two in the Americas), LTE requires a much larger number of bands to support worldwide operation. This requires a large number of filters and switches as well as power amplifiers and puts the cost up substantially.

5.5.1 LTE-M

Long Term Evolution Machine (LTE-M)-type communication and Narrow Band Internet of Things (NB-IoT) are low-power wide area network (LPWAN) protocols from the 3GPP as an extension of LTE that allows the cellular network to be used as an LPWAN. LTE-M was developed as a result of a realization that the LTE as it stood was not suitable for IoT and machine communications. The LTE-M targets are low device cost, long battery life, ability to support many devices in each cell, and deep coverage. LTE-M aims to achieve these objectives by allowing half-duplex operation, going to a single antenna only, adding a lower power class, supporting a lower data rate, and operating at a lower bandwidth.

There are two sorts of LTE-M defined in the 3GPP specifications—LTE-CAT-M1 and LTE-CAT-M2—but LTE-CAT-M1 is the most commonly deployed. LTE CAT-M1 uses a 1.4 MHz channel—the smallest available in the LTE specification. The system can support half-duplex and full-duplex operation. It is a single-antenna system (does not support receive diversity) which brings down the cost and power consumption.

5.5.2 NB-IoT

NB-IoT is a narrow band specification that fits into the LTE standard. NB-IoT can also be deployed in GSM networks and even in the guard bands at the edge of each frequency band in LTE. NB-IoT is aimed at high-density, low-bitrate support of IoT devices. Release 14 added support for mobility to NB-IoT (asset tracking is one of its largest use cases), but the degree of mobility is not a great as for the other cellular standards. Maximum supported data rates in the latest version of the specification are 127 kbps on the downlink and 159 kbps on the uplink; as always, real-world data rates are substantially less than this.

NB-IoT only supports half-duplex operation which makes the front end design simpler. It is a single-antenna system which also reduces complexity, cost, and power requirements. It uses a 200 kHz bandwidth which is how it is compatible with GSM. Within the 200 kHz bandwidth, the channel bandwidth is 180 kHz which fits with LTE and is the width of 1 RB. The 180 kHz of usable bandwidth in the downlink needs to be compatible with LTE and so uses OFDM with up to a maximum of 12 of the15 kHz sub-carriers. The uplink is also compatible with LTE and uses SC-FDMA with 15 and 3.75 kHz sub-carrier options.

The latency is specified at less than 10 seconds which is at least two orders of magnitude slower than other standards. For systems that need a low latency, this is going to be an issue.

5.6 5G

The latest addition to the cellular radio family is known, simply, as 5G. The 5G specification [7, 8] continues a trend from the earlier specifications of increasing the data rates over the link and also improving latency. Although improvements in latency (the delay over the network) were part of the aims of earlier standards, they became one of the main aims for 5G. The stated objectives for 5G are:

1. Data rate—5G is designed to deliver data rates ranging from 50 Mbps to 2 Gbps.

2. Latency—5G stated design target is 4 ms with 1 ms often being quoted as a target.

The data rate target is achievable in the lab but is more difficult to achieve in the field. To get close to data rates in the Gbps range, it is necessary for the transmitter and receiver to be relatively close.

With 5G a lot of new spectrum is being opened up. The radio interface specification, 5G NR (New Radio), defines two bands in the spectrum for 5G—FR1 (sometimes called “sub-6 GHz” even though it is specified up to 7.125 GHz) and

FR2 or millimeter wave (usually known as mmWave). The FR1 spectrum is now also broken into low-band and mid-band in most discussions. Low-band spectrum starts at 410 MHz and includes the current cellular spectrum up to the 2.4 GHz ISM band. Mid-band is the spectrum above the 2.4 GHz ISM band up to 6 GHz (which is the cut-off for the 5 GHz ISM spectral band) although the 3GPP is specified out to 7.125 GHz. Although networks are being rolled out in the low band, the mid-band is where the highest number of new networks is being introduced as these frequencies offer the possibility of getting towards the desired bit speeds. Networks at the very lowest frequencies often do not offer much performance advantage over advanced 4G networks at the same frequency.

The mmWave band is new frontier spectrum for cellular radios. Whereas the 3GPP specifications are specified for unwanted transmissions up to 12.75 GHz, the new mmWave spectrum has bands from 24.25 GHz up to 52.6 GHz. Operating at these high bands opens up the possibility to have 400 MHz bandwidths for signals as 400 MHz at a frequency of 25 GHz is only 1.6% of the band frequency whereas it is 100% of the band frequency at 400 MHz. Unfortunately, these high frequencies are much more difficult to work with. The signals at these frequencies are easily blocked, and the available transistor gain is much lower requiring either extra current to drive circuits or special transistors to get more gain at cost. Any system that needs to use the mmWave frequencies will need to have more base stations to cover the desired area.

The modulation schemes used in 5G are essentially the same as those used in LTE. OFDMA is used on the downlink, and SC-FDMA is used on the uplink. Sub-carriers of 15 kHz are used but now 30, 60, 120, and 240 kHz sub-carriers have also been added [9]. The sub-carriers are modulated using QPSK as the base modulation and 16 QAM, 64 QAM, and 256 QAM can be introduced as the link quality improves. As in LTE, the waveforms have a high PAR even on the uplink, and this leads to lower efficiency from the power amplifier.

Even in some of the low bands, channel bandwidths of 100 MHz are available, and this will add extra complications to the design of the transmit circuitry and power amplifiers as the circuits can experience “memory” effects. In a circuit experiencing memory effects, the circuit performance changes due to the amplitude of the signal but does not change back in time to process a change in amplitude. Wideband modulations put a larger stress on the circuit because it must react faster. Memory effects require that circuits not be driven at their limits which means that they have to be backed off and are less efficient.

Another objective defined in the yet to be released ITU-2020 Standard for 5G to be able to service 1,000,000 devices in a square kilometer. This is a 100× improvement on 4G systems. This has been referred to as “Massive IoT”. The 5G specifications as they stand are too power-hungry for most IoT applications which do not need the bitrate and latency advantages of 5G. The 3GPP is working on IoT specifications to be included in the 5G specification in an equivalent way to LTE-M and NB-IoT for LTE and EC-GSM-IoT for GSM. At the time of writing, these specifications are not available.

6. Overview of other radio systems

Outside the cellular world, there are a number of radio systems that are designed with IoT applications in mind. These radios work in the unlicensed industrial, scientific, and medical (ISM) bands of the radio spectrum. The use of these bands does not require a license, but they are still regulated bands, and any radio working in them must meet the regulatory requirements. There are a number of ISM frequency bands available, but the most commonly used are the 2.4 GHz band which is heavily used by radio communications devices and the 5.8 GHz band that is used for wireless local area networks (LAN).

Frequency spectrum is available at 902 to 928 MHz in Region 2 (the Americas)—known as 915 ISM—which is popular for IoT applications because it is less than 1 GHz, and so signals that travel better and transmit powers up to 1 W (30 dBm) are allowed. In Europe (part of Region 1), the main low band of cellular communications overlaps with this band, and so it cannot be used. In Europe there is a band at 868 MHz as part of the short-range device (SRD) spectrum, and that is used, but the output power of the transmitter is restricted to 25 mW (14 dBm), so it is less useful than the 915 ISM band.

Many of the alternatives to cellular systems use base stations to connect to nodes although they usually refer to these “base stations” as gateways or access points (in the case of wireless LAN). Unlike the cellular system, there is usually no handover protocol defined, and leaving one gateway requires a hard break and reconnection with the new gateway.

A network with a base station sitting at its centre and multiple sensors communicating through it is known as a star network. In some of the technologies outside the cellular world, it is possible to have another sort of network where nodes connect to each other and communicate with one another. Such a grid of nodes is known as a mesh. Star and mesh networks are shown in **Figure 18**.

It is, of course, possible to combine star and mesh networks. Such a network is shown in **Figure 19**.

The radio systems we will look at are Bluetooth Low Energy (BLE), Zigbee, 802.11ah, LoRaWAN, and Sigfox.

6.1 Bluetooth low energy

Bluetooth Low Energy (BLE) is an extension to the Bluetooth radio family controlled by the Bluetooth Radio Special Interest Group (SIG) specifically for IoT applications where long operation on battery is desired. Bluetooth Low Energy is not compatible with the previous versions of Bluetooth basic rate/enhanced data rate (BR/EDR). It is part of the Bluetooth 4.0 specification [10] and can be used alongside BR/EDR. The aim of BLE is to provide communications at the same range as BR/EDR at lower power and lower cost. It gives up voice capability and is purely data based in order to meet these aims.

BLE uses the same frequencies in the 2.4 GHz ISM as Bluetooth BR/EDR, but instead of 79 channels of 1 MHz each, it uses 40 channels of 2 MHz each. The modulation is Gaussian frequency shift keying (GFSK) which is a non-linear modulation which simplifies the design of the transmitter and uses less power than

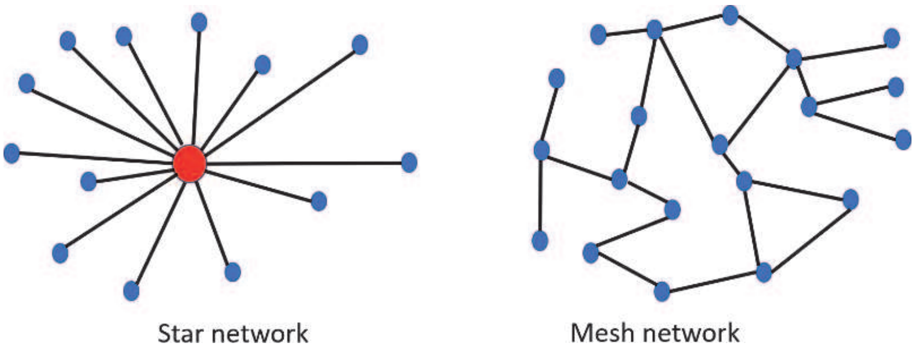

Figure 18.
Star network and mesh network.

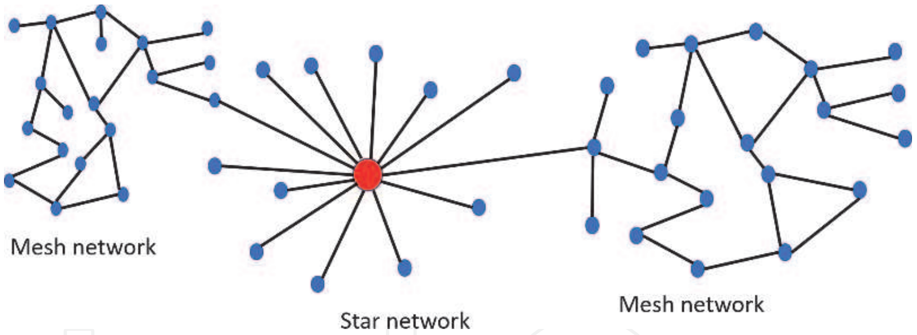

Figure 19.
Star network with mesh networks.

alternative linear modulations. In the Bluetooth 4.0 specification [10], the maximum output power was 10 mW (10 dBm). The maximum data rate is 1 Mbps.

One useful feature added with BLE is the concept of a mesh. BLE nodes can connect with their neighbors and pass a data transmission on to them. From one node it is possible to push a message across the whole mesh. This is another way of accessing an array of sensors rather than the central base station seen in cellular or gateways seen in other systems.

With the Bluetooth 5.0 specification [11], came some extensions to the BLE specification: higher power and higher data rate modes. The maximum power was increased tenfold to 100 mW (20 dBm) at a lower data rate of 500 or 125 kbps. This is expected to give a fourfold increase in range (power decreases with square of range from the transmitter, so a tenfold increase in power only sees a root-10 increase in distance). The maximum data rate was doubled to 2 Mbps but at a lower power.

BLE chipsets are designed to run off a standard coin cell battery and last up to 10 years. Although most BLE solutions will work for less than this you can still expect many years of operation.

There have been a number of cases of security vulnerabilities in Bluetooth being exposed. As ad hoc connections can be created, it is vulnerable to people pairing with devices which can, in the worst case, mean a loss of control. Also, one of the advantages of Bluetooth—the fact its communications can penetrate walls—is also a vulnerability as it can mean that a Bluetooth network can be accessed from outside a building.

Overall, Bluetooth, and in particular BLE, is an excellent radio protocol for IoT sensor nodes within a small geographic area providing the security issues are addressed. BLE offers no way to connect back to the Internet and hence cloud unlike cellular and some other systems; however it makes a good choice if the system needs to connect to multiple sensors through a local facility. In this case the system should also have computing at the edge capabilities to process data before sending the processed data into the cloud.

6.2 Zigbee

Zigbee is a networking protocol designed specifically for low-power and low-data rate devices. The Zigbee standard is developed by the Zigbee Alliance, but the physical (PHY—the lowest layer) and medium access control (MAC) layers are adopted from the Institute of Electrical and Electronic Engineers (IEEE) 802.15.4 specification [12]. This leaves only the network (NWK) and application (APL) layers in the Zigbee specification [13].

Zigbee is specified to work in the 868 MHz SDR, 915 MHz ISM, and 2.4 GHz ISM bands. In the 868 MHz band, one channel only is available, and BPSK modulation is used on that channel. Similarly, for the 915 MHz band, BPSK modulation is used, but there are 10 channels available. In the 2.4 GHz band, the modulation changes to OQPSK, and with the wider available bandwidth, 16 channels are available. ASK and OQPSK modulations are optionally available in the low bands for use in the case the 2.4 GHz channel is unavailable. Zigbee supports data rates in the 20–250 kbps range, so it is not suitable for many of the higher data rate applications but is well suited for many IoT sensor applications.

Output power in the 868 MHz band is limited to a maximum of 14 dBm which, with the BPSK PAR of about 2 dB, means the radio is limited to around a maximum of 12 dBm of output power. In the other bands, the regulatory limits are much high, and the highest power Zigbee radios can put out up to 20 dBm, but most transmit less. The line of sight range for Zigbee is quoted at 100 m maximum.

Two types of network can be set up with Zigbee devices: star and mesh. Combinations of the two types are possible. From this requirement, three different types of Zigbee device are needed: Zigbee coordinator, Zigbee router, and Zigbee end device. The Zigbee coordinator is the core of any Zigbee network. This is the node that will communicate with the outside world and control the whole network. Zigbee routers pass communications between each other but are not as capable as the Zigbee coordinator. Zigbee end devices will communicate only with either a Zigbee router device or the Zigbee coordinator and have minimal functionality. These will usually be the sensor nodes of the application. A Zigbee network is shown in **Figure 20**.

Zigbee is well designed for the purposes of connecting low data rate sensors in IoT applications at a local level. It is often criticized for not having enough security, so applications that use it may have to add their own data encryption on top which adds overhead and processing. It is not designed to connect back to the Internet and so to make the data available in the cloud.

6.3 802.11ah

The version of the wireless local area network (WLAN) specification designed to cover IoT usage cases is the IEEE 802.11ah specifications [14]. The other specifications in the 802.11 series (known as WiFi) are also useable in IoT applications but are primarily aimed at connectivity and are not optimized for IoT. The 802.11ah specification is also known as WiFi HaLow. While most of the wireless LAN specifications operate in the 2.4 and 5 GHz ISM bands (802.11ad, known as WiGig, operates at 60 GHz), 802.11ah will operate in the 868 and 915 MHz bands.

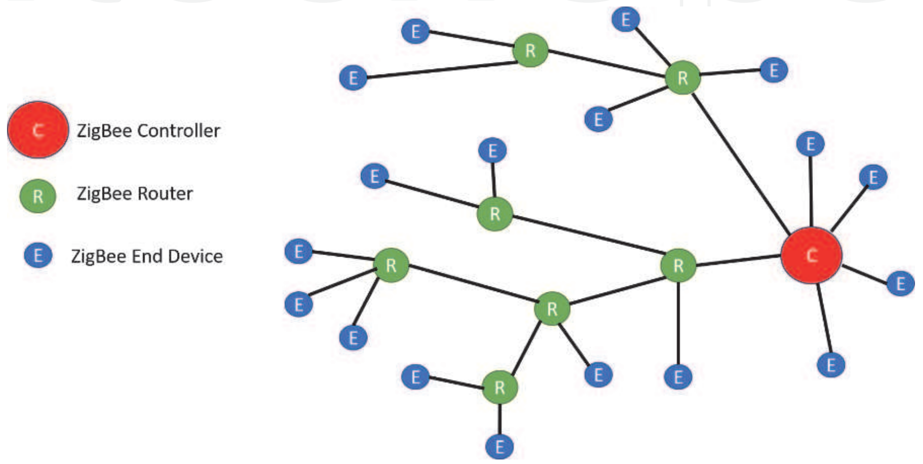

Figure 20.
Zigbee network illustrating the role of three device types.

WiFi HaLow is designed for longer haul communications than Bluetooth or Zigbee being capable of coverage up to 1 km. The system uses OFDM in 2 MHz channels, and the modulations used on the sub-carriers are all phase modulations going from BPSK up to 256 QAM. Available bitrates go from 150 kbps up to 234 Mbps.

Unfortunately, although WiFi HaLow offers some advantages—the ability to ramp the data rate over a wide range is a useful property to have—the standard has not been taken up by many companies, and no large company is producing chipsets to support it.

6.4 LoRaWAN

LoRaWAN is an low-power wide area network (LPWAN) networking technology built on top of the Long Range (LoRa) PHY layer protocol [15]. LoRa technology is available in the 433 and 868 MHz bands in Europe, the 915 MHz band in the USA and some other nations.

LoRa uses a unique modulation scheme: chirp spread spectrum [16]. The carrier is modulated much like an FSK signal, but in this case the frequency is either increased linearly or decreased linearly with a continuum of frequencies. Such a continuum of linearly increasing frequencies is called a chirp. Bandwidth is fixed, so the chirp can only move between a minimum and maximum frequency; however the rate at which it does that movement can vary. A chirp is shown in **Figure 21**.

There are three bandwidths available—125, 250, and 500 kHz—as well as six different slopes (known as spreading factors (SF)). For high data rates, the chirp will have a high slope and a correspondingly lower SF, and for low data rates, the chirp will have a low slope and high SF [17]. Start frequency determines the coding. On the receive side, the received symbol is multiplied with the inverse chip to extract the data. The higher the spreading factor, the longer the symbol that will appear in the receiver and the higher the likelihood of correct demodulation.

The LoRa modulation is a constant envelope modulation and as such lends itself to a compact transmitter and low power consumption. The nature of the modulation means that LoRa transmission can travel further for the same output power than many competing technologies. This makes LoRaWAN deployment attractive for rural and outdoor applications.

LoRaWAN is a networking technology built on top of the LoRa physical layer. The LoRaWAN network is a star network with individual nodes connecting back to a central gateway. This gateway is set up by the private company or individuals building the network, and this is one of the advantages of LoRaWAN—users build their own network and control their own data. Connection from this gateway to the Internet can be over Ethernet or fiber or can also be over the cellular network.

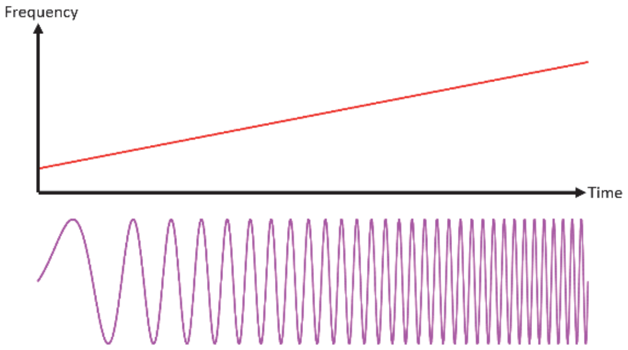

Figure 21.
Chirp signal frequency vs. time representation and time domain signal.

With LoRaWAN you are limited to 27 kbps data rate (there is an FSK option that can get up to 50 kbps). The other main limitation with LoRa technology is that it is only available on chips from Semtech Corporation. This limits the choice on performance and cost.

6.5 Sigfox

Sigfox is a French company that operates a network for IoT in the 868 MHz band in Europe and the 915 MHz band in the USA [18]. The Sigfox network is a star network with nodes communicating with base stations. Sigfox owns the base stations. The technology is designed for low data rate communications. The modulation is called Ultra Narrow Band (UNB) and makes use of a narrow bandwidth signal. The channels available are 200 kHz wide, but the bandwidth of the signal is only 100 Hz. With the narrow bandwidth signal, it is difficult to block the communication because any blocking signal needs to be right on top of the UNB signal. The carries are modulated using differential binary phase shift keying (DBPSK)—It is the value of the difference between a symbol and its previous symbol that is used for modulation not the symbol value itself and GFSK modulations. As the bandwidth of the communications is narrow, the data rate available is small—100–600 bits per second. Uplink messages are 12 bytes long, while downlink messages are 8 bytes long. A user is limited to 140 uplink messages a day and is only allowed to receive 4 downlink messages a day.

Sigfox is certainly something that should be considered if you have a low data rate application and do not want to communicate with the device that much—something that goes for a lot of sensor nodes. However, the message limits can be very restrictive.

6.6 Combining radio systems

As we have seen, there are many radio systems available and, those in this book are not an exhaustive list. However, every radio system has its advantages and disadvantages, and depending on the application, one radio system may be better than another.

Both BLE and Zigbee have the useful property of being able to support mesh networks. There are many instances where this may be useful: a network of sensors in a factory or a large office building or even a farm for instance. We could therefore envisage using Zigbee or BLE to network our sensors and bring the data back to a central place. At some point we will want to process that data. If the central location is in the offices of the factory, for instance, there may be no need to send the data any further. **Figure 22** shows a Zigbee network controlled from smartphone app using Bluetooth.

However, if the factory is part of a large network of factories, then the main office of the company may want to see the data. In this case we may want to make use of a cellular network to send the data back. In effect we have used the advantages of the two technologies—mesh technology for the local network and a secure, robust data link to send the results back. This would need some sort of local processing to put the data into a format that the main office can handle. **Figure 23** illustrates this with a Zigbee network connected through local processing to the cellular network.

We could send all the data from all the factories to the main office if we have a high-speed data link, but the head office may not want to process all that data. In this case, some form of computing at the edge is to process the data first and send a summary back (while storing all the data locally) to the main office. This also allows

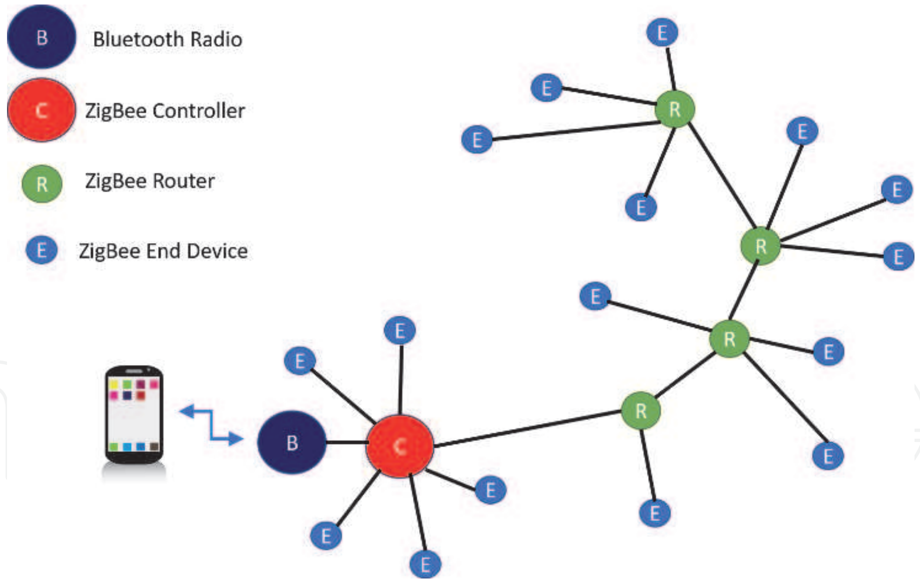

Figure 22.
A Zigbee network with Bluetooth link to smartphone control.

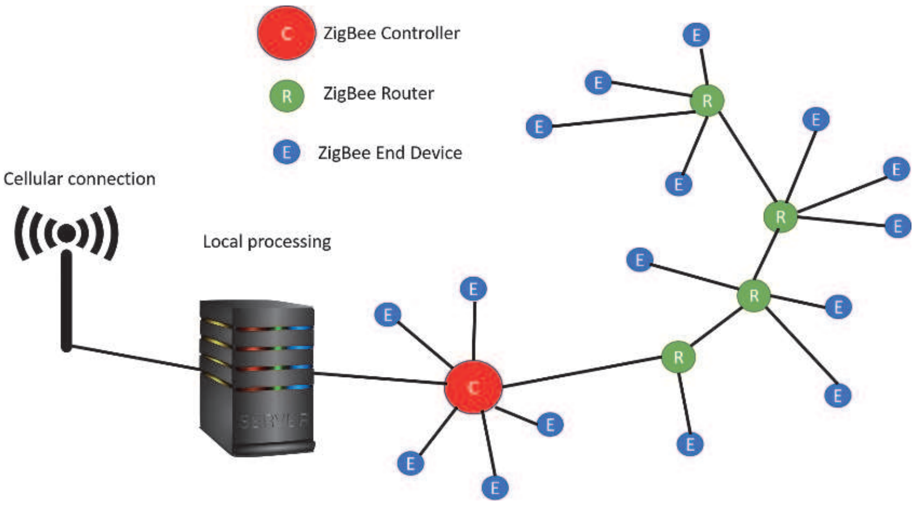

Figure 23.
Zigbee local network with local processing and cellular connection.

local management to catch problems quickly and start acting upon them before they get out of hand. If all decisions are taken at the main office based on data, they may miss problems because they have too many factories to check.

In the case of factories in remote areas with no access to high-speed data links, wired or not wired, it becomes necessary to compress the data sent back to the main office. In this case having local computing resources is a requirement. In fact, farms, oil facilities, mining facilities, highway rest stations, and many more places could all benefit by having networks of sensors linked up to a local processing facility with the data being sent back over another network.

7. Radio for sensing: introduction to radar sensors

We are familiar with radar in the context of ships and airplanes. These are large systems with large antennas. Recently advances in semiconductor technology have made it possible to integrate radars on chip. The initial application that started driving the development of radars was the reverse warning system for cars.

However, radars are now used for proximity detection in cars and are even being used in autonomous vehicle applications. With the availability of cheap integrated circuits from a number of manufacturers comes interest in using radars in a variety of applications.

Radar is divided into monostatic and bistatic types. Monostatic radars either use the same antenna for both receive and transmit or co-locate the antennas if they are separate. Bistatic radars separate the receive antenna from the transmit by a considerable distance. We will concentrate on monostatic radars here, not because bistatic radars are not interesting for IoT applications—they may well be—but because all the available integrated circuits are monostatic.

We can further divide monostatic radars into pulsed and continuous wave (CW) types. Most current radar chips are of a CW sub-type called frequency-modulated continuous wave (FMCW) which is the least complex yet powerful.

7.1 FMCW principles of operation

A CW radar transmits a continuous wave—a single frequency—which then bounces back off a target and is received back at the receiver. In the receiver it is mixed with the original transmitted tone, and this mixing brings it down to sit at 0 Hz (DC). Unfortunately this means we get no information about the distance of the object, but if the object moves, we see a shift in frequency received due to the Doppler effect, and this shift in frequency is seen as a shift away from DC in the downconverted signal. We can use this frequency shift to give us the velocity of the object. This is the principle of speed detection radar guns. CW radar is very cheap to build because it only requires a stable oscillator, downconversion circuitry, and some basic processing.

If we want to be able to detect distance—critical in applications like proximity warning systems—then we need to move to FMCW radar [19]. In FMCW radar, the transmit pulse is a frequency-modulated signal. Although it is possible to modulate the transmit signal with different signals, the most common and simplest is to use a sawtooth waveform or triangular pulse waveform. The signal in this case is known as a chirp. A chirp is a pure continuous wave which either linearly increases or decreases in frequency. Chirps, incidentally, are used in LoRa for communication. **Figure 24** shows the transmit and receive chirps represented as frequency vs. time and the relationship between them.

As can be seen, the transmit signal has changed frequency when the return pulse comes back. The difference in frequency may be small but it is measurable. Also as the transmit signal frequency slope is linear, the difference frequency, for a static

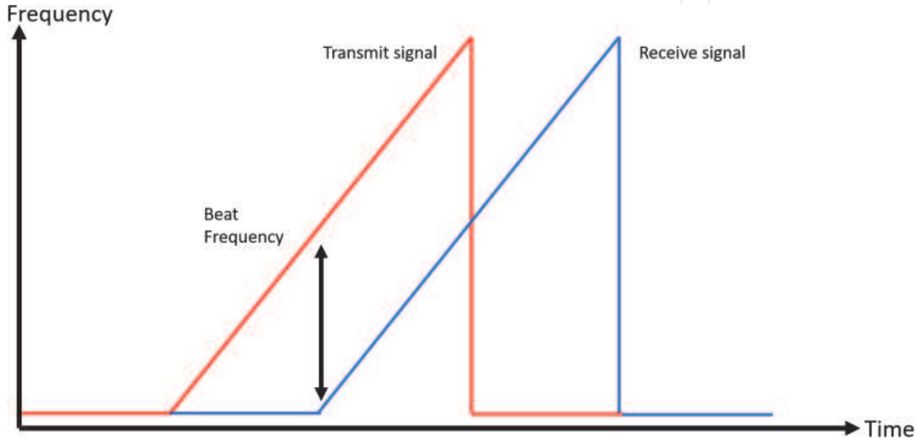

Figure 24.
Radar transmit and receive signals and their relationship.

object, will be a single frequency known as a beat frequency. From the beat frequency, we can calculate the distance to the object using the following equation:

$$f_{\text{beat}} = \frac{2 \times D}{c} \times \frac{df}{dt} \tag{15}$$

where $\mathrm{f_{beat}}$ is the beat frequency, D is the distance to target, c is the speed of light (3×10^8 m/s), and $\frac{df}{dt}$ is the slope of the chirp.

If the object is moving away or towards the antenna, then we will get Doppler shift in the frequency, and we will be able to detect that as well. We will not be able to detect it if an object is moving around a circle that is equidistant from the antenna. In effect we have a two-dimensional system—distance to the object and velocity. The velocity is given from the Doppler frequency using the following equation:

$$f_{\text{Doppler}} = \frac{2 \times v}{\lambda} \tag{16}$$

where f_{Doppler} is the Doppler frequency, v is the velocity, and λ is the wavelength at the radar frequency.

7.2 Moving to a 4D system

With a single antenna, we have distance to a target and velocity of the target relative to the antenna. If the antenna is idealized as an isotropic antenna (equal transmission in all directions), then we will have a sphere around the antenna at the distance to the target where the target could be. By adding a second antenna, we would have two spheres of potentially different size, and the target would lie somewhere on a circle at the intersection of the two antenna's spheres. Fortunately, real antennas are not isotropic and tend to be directional. In this case our spheres become distorted, and our circle becomes an arc. This is much more useful.

The antennas used with radar chips can be built into the package or are on the board. In either case they are patches of one sort or another which will transmit in much better away from the board or package than into it and behind it. By putting two antennas next to each other with at least half a wavelength distance at the radar frequency between them, we can determine where the target is on a plane stretching from left to right of the radar but not if it is above or below it. By adding a third antenna above the other two, we now know where, exactly, the target is in three dimensions relative to the antennas.

We already know that we can extract velocity towards or away from a single antenna but not the velocity of an object moving in a sphere around the antenna. However with more than one antenna, the object will always be moving relative to at least one of the antennas. So in the three-antenna configuration, the target will have velocity components that we can extract towards and at right angles to the path towards the antenna for each antenna. We can, therefore extract velocity information as well as direction of travel. This, in effect gives us four dimensions—the three special dimensions and velocity.

A common configuration of transmits in FMCW radar integrated circuits is two transmits and four receives. This configuration can be used to obtain a full 4D picture of the environment in front of the radar.

We can also extract other information about the target. In general, the return signal will also tell us how big the target is. This does, however, depend on the material the target is made of. Metal reflects radar radiation much better than, say, a human does or even a brick wall, and so a metal plate will appear relatively large. However, a human will reflect more than a dog, and so we are able to infer whether we have a human or an animal in front of the radar which may be useful in some remote monitoring applications.

7.3 FMCW signal processing

The processing required for FMCW radars is relatively simple. The digitized signal from the received signal is collected into a frame of a given number of samples, and this frame is fed into a discrete Fourier transform (DFT) usually implemented using an fast Fourier transform (FFT) algorithm. This gives us a spectrum of the signal where individual peaks will represent object and the frequency associated with that peak can be used to calculate the distance to the object. A second level of processing is then used to extract the velocity information. The spectra generated in the first stage are collected into another frame which now has two dimensions—time along one axis and frequency along the other. A DFT, again using an FFT algorithm, is run on this frame, and this gives us the velocity information at each frequency.

This processing needs to be performed for each antenna, and then the results need to be combined to give a full 4D picture of the world in front of the antenna. This signal processing is specialized but involves well-known algorithms and so is not particularly arduous to implement. FFT algorithm implementations are available in dedicated hardware blocks, and it is possible to build a hardware engine to perform the processing which would be significantly power-saving over a software implementation on a general-purpose processor. With dedicated hardware blocks to perform most of the processing required for the radar and a general-purpose processor to perform user programmed functions, we could build a flexible and high-performance imaging sensor unit which runs on relatively low power.

7.4 FMCW radar and video

Radar is a different sort of imaging sensor. It does not offer the sort of high-resolution images that are available from cameras, but it does have advantages over cameras in many applications. In general an FMCW radar installation will be of lower power and require less bandwidth from the communications systems than a video installation would. Even if the video images are processed locally, FMCW processing is a lot simpler than image processing for many applications and so requires smaller memories and computers and less power. Radar will also work in fog, rain, sleet, and smoke making useful in a wide range of applications. For applications that require identifying individuals or certain states (wearing a helmet or not, for instance), radar will not do the job, and image processing will be necessary. However, for some applications there are power constraints, a variety of atmospheric conditions may be encountered, or there is a privacy concern—monitoring of changing rooms or public toilets for instance—so for other applications, radar is a better fit. The next few sections will go over some of these applications.

7.5 Remote monitoring

FMCW radar is a good choice for monitoring applications. Video is used extensively in the security industry, but humans are notoriously bad at monitoring video feeds, and a system based on image processing and some sort of machine learning would be power-hungry. Generally, we are only interested in things that move—it is impossible to steal or damage something without getting close to it and physically moving it. Radar can tell you not only that something is moving but also how fast, in what direction, and even roughly how big the moving object is. It can also tell you how far away something is, so monitoring fluid level in a tank is also another good application of FMCW radar. Other non-security applications like automatic door opening (where the system can tell people are approaching the door rather than just passing) and people counting are also an extension of this monitoring.

7.6 Heart beat and breathing

For people that are close enough, FMCW is able not only to detect breathing but also heartbeat [20]. This is potentially extremely useful in many applications. An obvious example is disaster relief where radar can be used to detect people trapped in the rubble of a collapsed building. In home monitoring of the elderly is another potential application and one where radar is a better fit than cameras as it does not violate the privacy of the old person being monitored. There are numerous potential uses in the medical field including monitoring of sleep quality and looking for signs of sleep apnoea.

7.7 Gesture

Google has a project, Soli, that is developing solutions in using radar for gesture recognition [21]. Doppler radar can detect movement, and this movement can be deliberately coordinated to convey information—gestures. Gestures can be inferred by using machine learning (ML) to recognize which particular radar signature belongs to which particular gesture. Using this system gestures can be used to control equipment around us much like a television remote control.

7.8 Gait

As we have information about an object's position in all three cardinal dimensions and its velocity (with direction on the velocity), radar can be used to extract gait information. From the gait information, we will be able to identify a person from the way they walk. However, gait is also related to mental faculty. In fact, there is an ongoing medical research on using gait to give an early indication of neurodegenerative diseases like Alzheimer's disease and Parkinson's disease. In the future it may be possible to combine gait analysis with other analyses in a home monitoring application for the elderly based on FMCW radar devices.

8. IoT, radios, and edge computing

Cellular radio systems are set up with mobility in mind. Also, as the original systems were often voice, the transmit and receive systems were set up with equal bandwidth and data rate considerations. As data communications came to dominate, the systems that adjusted for the asymmetry in the data needs made the adjustment assuming the data needs would be higher on the downlink. For instance, streaming video needs a high bandwidth and high data rate on the downlink but not much on the uplink. This, as if happens, fits well with the requirements of mobile devices: mobile devices run off battery systems and so have limited power available. Keeping the upload data rate and modulation simple allows for a more efficient PA and hence saves power.

Many IoT systems will consist of fixed nodes sending data back to a fixed node with very little need (or possibly even ability) to receive data. In effect they need support for a higher data rate on the uplink than on the downlink. This is the opposite of the mobile client case, and so radio systems designed with a mobile client in mind may not be optimal for many IoT applications. It should also be obvious that the performance of the transmitter for an IoT sensor node is of much more importance than the performance of the receiver. In fact, in the simplest sensor nodes that continuously blast data out, there is not even a need to have a receiver.

8.1 IoT sensor systems and radio requirements

In an IoT system with the position of sensor nodes fixed, there is no need to support overheads in the radio system for mobile clients. This means overhead to support handover and multiple base stations are not needed. Fading as a result of the client device's movement will not be an issue; however changes in the environment around the radio may still contribute to fading although it should not be as deep as in a mobile application. This means it would be prudent to include interleaving.

The requirements for a wireless system for IoT depend on the usage case. For a system that sends back temperature data every few hours for instance, the wireless system used to send the data back would not need to handle a high data rate, but it would probably be desirable that it is robust and of low power. Also, the distance that needed to be covered would need to be considered, and so transmit power and transmit efficiency would be a factor in the system selection. With a low data rate requirement, the link robustness could be somewhat traded off as ARQ error correction would be possible. If the system needed to operate off a battery, the power dissipation of the wireless link would be a big factor if not the biggest factor in overall power dissipation. This would make the choice of wireless link critical.

A system sending back a continuous video feed would need to have a much higher data rate than a simple system returning a single sensor reading. It would need to be much more robust as it would not be able to make much use of ARQ without a significant bandwidth overhead. Battery operation is not practical unless the battery is large or there is some mechanism for battery recharging (solar, wind power, diesel generator).

It is possible to envisage a system which needs different requirements at different times. For instance, in a system to automate agriculture, we may use a combination of sensors strategically placed to measure temperature, moisture, sunlight, and other conditions. These sensors could be linked via a low-power radio system like Bluetooth or Zigbee back to a local control centre in the middle of the fields being monitored. The local control area processes the data from the sensors (an example of computing at the edge) and is linked back to a central control centre over a cellular link. The central control centre controls agricultural operations over a wide geographic area.

Normally, the local control centre would not need to send much data back—temperature, hours of sunshine during the day, light intensity, soil moisture, and so on. However, there may be circumstances where we need to inspect the location quickly—a sudden rise in temperature in one area indicating a fire and sensors in one area suddenly going down. As it is a remote location, we do not want to send someone out to check because it may be too late by the time the inspection team gets there. This is a realistic prospect in countries with a rapidly aging population where there is no manpower available to work in agriculture. In this case we would want to send a robot or even a drone over to inspect from the local control centre. If we must drive the robot or fly the drone from the central control centre, we will need a high bandwidth link to the local control centre and from there to the robot or drone. The solution is to put resources close to the drone or robot—people or computing. Having computing resources available at the edge to control the drones makes a lot of sense.

Even if the drone or robot is fully autonomous and the fields and local control centre are equipped with a 5G network so any video sent back is received at full definition, we still need a high-speed link to get that video data back to central control centre, so we can see what is going on. In this case we need to be able to ramp the communications speed up by several orders of magnitude which is not possible with today's radio technology. It may even be necessary to install a point-to-point wireless link if the only way decisions can be made is to get the data back. To get away from the need to have a high-speed link, it is necessary to move the computing resources close where they can monitor the video data. A system that can recognize a fire and send a message "there is a fire" is more useful than "we have a problem".

8.2 Bandwidth, bitrate, latency, and distance

As bandwidth increases, the bitrate that can be supported with the simplest modulations increases with it. If the bandwidth is fixed (as it is by regulatory requirements for wireless systems), then the only way to increase bitrate is to use deeper phase and amplitude modulations (16–64 QAM). A wide bandwidth system using a higher-order modulation will quickly run into physical limits: the higher modulation is pushing the acceptable noise level down as a larger signal-to-noise ratio (SNR) is needed, yet the wider the bandwidth means more noise, as noise is integrated over bandwidth. If you are transmitting at full power, then the only way to get more SNR is to move the transmitter closer to the receiver (or vice-versa).

If your application needs both high bitrate and needs to work over long distances, you need to reconsider your system. Is the sensor you are using appropriate? A lot of solutions use vision (a video camera) with AI software to perform their task. Video images consume a lot of bandwidth. Is there not some other sort of sensor that can do the job? Could it be done using radar, for instance? If the answer truly is no, then you will need to consider moving the computing closer to the camera. If you can do all or the bulk of the processing at the edge, what is the size of message that needs to be sent?

If the system has a high-bandwidth and a low-latency need, then the only choice will be 5G. The low-latency requirement will force you to move your computing to the edge because of the delay between the base station and the server. Although 5G promises 4 ms latency, the latency over the air is already twice that in current generation products, and the delay between the base station and the server can be 20 ms or more.

9. Conclusions

There are many radio systems available for IoT applications. Some may be suitable, and many will not.

Most IoT applications will not have access to either a wired power supply or a wired communications system. Systems built to work in these applications are, by default, going to need a wireless system to connect to the outside world. They are, by their nature, also going to be running off batteries, and so power consumption is a critical design specification for them. High data rate systems are going to consume more power than lower data rate systems. It is probably the case that moving the computing to the edge will remove the need for a high-speed link and may make it possible for a sensor node off a battery.

For wireless communications systems, a general rule is higher data rate requires more power. Also, the higher the bandwidth, the more difficult it is to set the system up: the position and angle of the antennas, what other devices are working around the system, and what are the obstacles in the way between the transmitter and receiver, amongst many reasons, all play into performance. Wherever there are high-speed data needs, the transmitter and receiver need to be close to each other. Adding computing closer to the edge can decrease the data requirements and make the implementation more practical.

One of the most useful and overlooked wireless systems is the SMS system that comes with all cellular systems. You can use SMS on 5G, 4G, 3G, and 2G systems throughout the world. It works on 2G systems in some of the more inaccessible locations. In order to make use of the SMS system, it is necessary in many instances to process the data first to compress it before sending. With a computing at the edge system, it becomes possible to make use of a truly useful system for getting data back.

Finally, if low latency is needed in a 5G system, the computing will need to be as close as possible to the base station. 5G promises 4 ms latency—we aren't there yet with more like 10 ms latency over the air only. If the data has to go back to the central servers to be processed, the latency would be more like 30 ms or more than 4 ms—hardly what 5G promised.

Author details

Malcolm H. Smith
AnalogueSmith (S) Pte. Ltd, Singapore

*Address all correspondence to: analoguesmith@gmail.com

References

[1] Rappaport TS. Wireless Communication Principles and Practice. 1st ed. Upper Saddle River, New Jersey: Prentice Hall; 1996. ISBN: 0-13-375536-3

[2] Bose JC. Detector for Electrical Disturbances, U.S. Patent 755,840, 1904

[3] Armstrong EH. Radio Signaling, U.S. Patent 1,941,068, 1933

[4] 3GPP. GSM/EDGE Radio Transmission and Reception, TS 45.005 Version 16.1.0; 2 April 2020; Third Generation Partnership Project (3GPP), 2020

[5] 3GPP. User Equipment (UE) Radio Transmission and Reception (FDD), TS 25.101 Version 16.1.0; 2 April 2020; Third Generation Partnership Project (3GPP), 2019

[6] 3GPP. Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) Radio Transmission and Reception, TS 36.101 Version 16.1.0; 8 April 2020; Third Generation Partnership Project (3GPP), 2020

[7] 3GPP. NR; User Equipment (UE) Radio Transmission and Reception; Part 1: Range 1 Standalone, TS 38.101 Version 16.1.0; 8 April 2020; Third Generation Partnership Project (3GPP), 2020

[8] 3GPP. NR; User Equipment (UE) Radio Transmission and Reception; Part 2: Range 1 Standalone, TS 38.101 Version 16.1.0; 9 April 2020; Third Generation Partnership Project (3GPP), 2020

[9] 5G Physical Layer Specifications—5G NR, Medium [Internet]. Available from: https://medium.com/5g-nr/5g-physical-layer-specifications-e025f8654981 [Accessed: 14 April 2020]

[10] Bluetooth SIG. Bluetooth Specification, Version 4.2; 2 December 2014; Bluetooth SIG. 2014

[11] Bluetooth SIG. Bluetooth Core Specification, Version 5.2; 31 December 2019; Bluetooth SIG. 2019

[12] IEEE. IEEE Standard for Low-Rate Wireless Networks, 802.15.4; 5 December 2015; Institute of Electrical and Electronic Engineers (IEEE). 2015

[13] Farahani S. Zigbee Wireless Networks and Transceivers. 1st ed. Burlington, MA: Newnes; 2008. ISBN: 9780750683937

[14] IEEE. IEEE Standard for Information technology—Telecommunications and information exchange between systems—Local and metropolitan area networks—Specific requirements—Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 2: Sub 1 GHz License Exempt Operation, 802.11ah; 7 December 2016; Institute of Electrical and Electronic Engineers (IEEE). 2016

[15] Technical Marketing Workgroup 1.0, LoRaWAN™ What is it? A Technical Overview of LoRa® and LoRaWAN™, LoRa Alliance [Internet], 2015. Available from: https://lora-alliance.org/resource-hub/what-lorawanr [Accessed: 14 April 2020]

[16] Mobilefish.com. LoRa/LoRaWAN tutorial 12: Modulation Types and Chirp Spread Spectrum; 2 October 2018; YouTube Video [Internet]. Available from: https://www.youtube.com/watch?v=lg0eZWZFKiE [Accessed: 14 April 2020]

[17] Mobilefish.com. LoRa/LoRaWAN tutorial 13: Symbol, Spreading Factor and Chip; 4 October 2018; YouTube Video [Internet]. Available from:

https://www.youtube.com/watch?v=0FCrN-u-Vpw [Accessed: 14 April 2020]

[18] Sigfox Homepage [Internet]. Available from: https://www.sigfox.com/en [Accessed: 14 April 2020]

[19] Autonomous Stuff. Introduction to Mmwave Sensing: FMCW Radars; 14 April 2018; YouTube Video [Internet]. Available from: https://www.youtube.com/watch?v=8cHACNNDWD8 [Accessed: 14 April 2020]

[20] Ahmad A, Roh JC, Wang D, Dubey A. Vital signs monitoring of multiple people using a FMCW millimeter-wave sensor. In: 2018 IEEE Radar Conference (RadarConf18); Oklahoma City, OK. 2018. pp. 1450-1455

[21] Google ATAP. Welcome to Project Soli; 30 May 2015; YouTube Video [Internet]. Available from: https://www.youtube.com/watch?v=0QNiZfSsPc0&t=3s [Accessed: 14 April 2020]