

Digital modulation methods

In digital modulation, an analog carrier signal is modulated by a digital bit stream. Digital modulation methods can be considered as digital-to-analog conversion, and the corresponding demodulation or detection as analog-to-digital conversion. The changes in the carrier signal are chosen from a finite number of M alternative symbols (the modulation alphabet).

Schematic of 4 baud (8 bps) data link.

A simple example: A telephone line is designed for transferring audible sounds, for example tones, and not digital bits (zeros and ones). Computers may however communicate over a telephone line by means of modems, which are representing the digital bits by tones, called symbols. If there are four alternative symbols (corresponding to a musical instrument that can generate four different tones, one at a time), the first symbol may represent the bit sequence 00, the second 01, the third 10 and the fourth 11. If the modem plays a melody consisting of 1000 tones per second, the symbol rate is 1000 symbols/second, or baud. Since each tone (i.e., symbol) represents a message consisting of two digital bits in this example, the bit rate is twice the symbol rate, i.e. 2000 bits per second. This is similar to the technique used by dialup modems as opposed to DSL modems.

According to one definition of digital signal, the modulated signal is a digital signal, and according to another definition, the modulation is a form of digital-to-analog conversion. Most textbooks would consider digital modulation schemes as a form of digital transmission, synonymous to data transmission; very few would consider it as analog transmission.

Fundamental digital modulation methods

The most fundamental digital modulation techniques are based on keying:

- In the case of PSK (phase-shift keying), a finite number of phases are used.
- In the case of FSK (frequency-shift keying), a finite number of frequencies are used.
- In the case of ASK (amplitude-shift keying), a finite number of amplitudes are used.
- In the case of QAM (quadrature amplitude modulation), a finite number of at least two phases, and at least two amplitudes are used.

In QAM, an inphase signal (the I signal, for example a cosine waveform) and a quadrature phase signal (the Q signal, for example a sine wave) are amplitude modulated with a finite number of amplitudes, and summed. It can be seen as a two-channel system, each channel using ASK. The resulting signal is equivalent to a combination of PSK and ASK.

In all of the above methods, each of these phases, frequencies or amplitudes are assigned a unique pattern of binary bits. Usually, each phase, frequency or amplitude encodes an equal number of bits. This number of bits comprises the symbol that is represented by the particular phase, frequency or amplitude.

If the alphabet consists of $M = 2^N$ alternative symbols, each symbol represents a message consisting of N bits. If the symbol rate (also known as the baud rate) is f_S symbols/second (or baud), the data rate is Nf_S bit/second.

For example, with an alphabet consisting of 16 alternative symbols, each symbol represents 4 bits. Thus, the data rate is four times the baud rate.

In the case of PSK, ASK or QAM, where the carrier frequency of the modulated signal is constant, the modulation alphabet is often conveniently represented on a constellation diagram, showing the amplitude of the I signal at the x-axis, and the amplitude of the Q signal at the y-axis, for each symbol.

Modulator and detector principles of operation

PSK and ASK, and sometimes also FSK, are often generated and detected using the principle of QAM. The I and Q signals can be combined into a complex-valued signal

$I+jQ$ (where j is the imaginary unit). The resulting so called equivalent lowpass signal or equivalent baseband signal is a complex-valued representation of the real-valued modulated physical signal (the so called passband signal or RF signal).

These are the general steps used by the modulator to transmit data:

1. Group the incoming data bits into codewords, one for each symbol that will be transmitted.
2. Map the codewords to attributes, for example amplitudes of the I and Q signals (the equivalent low pass signal), or frequency or phase values.
3. Adapt pulse shaping or some other filtering to limit the bandwidth and form the spectrum of the equivalent low pass signal, typically using digital signal processing.
4. Perform digital-to-analog conversion (DAC) of the I and Q signals (since today all of the above is normally achieved using digital signal processing, DSP).
5. Generate a high-frequency sine wave carrier waveform, and perhaps also a cosine quadrature component. Carry out the modulation, for example by multiplying the sine and cosine wave form with the I and Q signals, resulting in that the equivalent low pass signal is frequency shifted into a modulated passband signal or RF signal. Sometimes this is achieved using DSP technology, for example direct digital synthesis using a waveform table, instead of analog signal processing. In that case the above DAC step should be done after this step.

6. Amplification and analog bandpass filtering to avoid harmonic distortion and periodic spectrum

At the receiver side, the demodulator typically performs:

1. Bandpass filtering.
2. Automatic gain control, AGC (to compensate for attenuation, for example fading).
3. Frequency shifting of the RF signal to the equivalent baseband I and Q signals, or to an intermediate frequency (IF) signal, by multiplying the RF signal with a local oscillator sine wave and cosine wave frequency (see the superheterodyne receiver principle).
4. Sampling and analog-to-digital conversion (ADC) (Sometimes before or instead of the above point, for example by means of undersampling).
5. Equalization filtering, for example a matched filter, compensation for multipath propagation, time spreading, phase distortion and frequency selective fading, to avoid intersymbol interference and symbol distortion.
6. Detection of the amplitudes of the I and Q signals, or the frequency or phase of the IF signal.
7. Quantization of the amplitudes, frequencies or phases to the nearest allowed symbol values.
8. Mapping of the quantized amplitudes, frequencies or phases to codewords (bit groups).
9. Parallel-to-serial conversion of the codewords into a bit stream.
10. Pass the resultant bit stream on for further processing such as removal of any error-correcting codes.

As is common to all digital communication systems, the design of both the modulator and demodulator must be done simultaneously. Digital modulation schemes are possible because the transmitter-receiver pair have prior knowledge of how data is encoded and represented in the communications system. In all digital communication systems, both the modulator at the transmitter and the demodulator at the receiver are structured so that they perform inverse operations.

Non-coherent modulation methods do not require a receiver reference clock signal that is phase synchronized with the sender carrier wave. In this case, modulation symbols (rather than bits, characters, or data packets) are asynchronously transferred. The opposite is coherent modulation.