The physical layer,
a key layer for wireless networks

Philippe Ciblat & **Alain Sibille**,
Télécom Paris, Institut Polytechnique de Paris

Abstract:
The success of wireless networks since the 1990s stems from digital technology and the conception of a cellular network. These have made it possible to efficiently use the radio-frequency spectrum while observing the fundamental laws of electromagnetism. Since forty years ago, enormous progress has been made in using the spectrum, as the latter's high end has been increasingly put to more diversified uses. The physical layer of the Open Systems Interconnection (OSI) model [ROLI2016], which underlays the architecture of networks since the 1970s, handles both the transmission and emission of signals from the physical world via antennas and the reverse operations for receiving signals. This layer's characteristics are decisive for the performance of wireless networks, their coverage of geographical zones and their capacity to provide reliable services.

The success of wireless networks since the 1990s heavily depends on digital technology and the concept of a cellular network for efficiently using the radio-frequency spectrum by taking into account the fundamental laws of electromagnetism. The use of radio frequencies has increased enormously over the past forty years, as more and more of the spectrum has been allocated for a wide variety of services.[1]

This article focuses on the "physical layer" of the open systems interconnection model (OSI) underlying network architecture since the 1970s (ROLIN *et al*. 2016). The physical layer refers to the transmission and emission of electromagnetic signals from the physical world in a medium of propagation via an antenna, and to the reverse operation for receiving signals. Its properties determine the performance of wireless networks, the territorial coverage of radio signals and the capacity for providing specific services.

The fundamentals of radio communications

C. Shannon and W. Weaver laid the foundations for communication theory in 1949. Their *Mathematical Theory of Communication* contained, in particular, the formula whereby the channel capacity (C in bits/s) is, as shown in Figure 1, related to the power ratio of the signal received (P_r) to the noise (P_{bruit},) for a given frequency bandwidth, B (where P_{bruit} equals B/N_0, N_0 being the thermal noise power spectral density). Except in an ultra wideband, this formula indicates a quasi proportionality between capacity and bandwidth. For this reason, the general trend for several decades has been to open more of the spectrum to wireless networks.

Figure 1:

$$C = B \cdot \log_2\left(1 + \frac{P_r}{P_{bruit}}\right)$$

[1] This article has been translated from French by Noal Mellott (Omaha Beach, France). All websites were consulted in November 2020.

The second basic remark to draw from this formula is that the capacity varies logarithmically with the signal received (P_r). It is, therefore, of no avail to enormously increase the power of the transmitted signal (P_t, which is, of course, related to the received signal). Besides, a significant decrease in power does not enormously decrease the channel capacity. It is "on the border of the cell" [area of land served by a base station] that the capacity plunges along with the connection's quality.

This brings us to the equation of telecommunications (Figure 2) that relates P_r to P_t for a frequency (F), distance (D) and antennas with "gains" in reception and transmission (G_r and G_t); c is the speed of light. Notice that the received signal (P_r) varies with $1/D^2$; this does not amount to a very fast decrease given the slow variation of the capacity with P_r.

Figure 2:

$$\frac{P_r}{P_t} = G_r \cdot G_t \cdot \left(\frac{c}{4\pi DF} \right)^2$$

These two formulas point toward the core problems for wireless communications.

First of all, the channel capacity (C) is not just a bit (data) rate but the maximum bit rate possible for errorless communications. So, it is necessary to make the "real spectral efficiency" (in bits/s/Hz) approach as closely as possible to its theoretical maximum (C/B). In this way, this scarce natural resource can be put to an optimal use. This means manipulating binary signals and their transmission through the air, and then the reverse operation. Given noise and interference from various sources, all of this added together means that the binary signals normally received are occasionally subject to errors. Rather than increasing the power with no limits, it is preferable to use "error-correcting codes". Error detection and correction is a major field of research in communications since Shannon and Weaver's work. Furthermore, the bits to be transmitted also have to be distributed among various frequencies within the band (B). This involves modulation or the choice of a waveform.

A second very important remark is that the received signal (P_r) decreases with the frequency in a proportion: $1/F^2$ — a significant variation. When passing from 900 MHz (GSM) to 28 GHz (the millimeter wave bands foreseen for 5G), the decrease is almost one thousandfold! The explanation in physics is very simple: the higher the frequency, the smaller the antenna (indirectly proportional to the frequency), its "effective area" varying as a function of $1/F^2$.

This reduction can be counterbalanced by making the antenna directional — by making it "listen" in a particular direction rather than in all directions. This leads to a gain in reception (G_r) that increases as a reciprocal of the fixed listening angle. For instance, if the specified angular sector is 10° azimuth and 6° altitude, a little more than this factor of 1000 can be recuperated. Furthermore, the same approach can be applied for a gain in transmissions (G_t). In both cases, this means that the antennas have to be pointed in the right direction. Among the advances made during the first decade of this millennium are multiple-input multiple-output antennas (MIMO) with electronic control over their direction. This is one reason for the assignment of millimeter wave bands to 5G.

Let us not forget that the transmitting and receiving antennas are separated by the medium of propagation: air or a vacuum (in "free" space). In the case of wireless networks however, there are several obstructions, first of all, the ground, not to mention buildings outside and their contents inside. The waves transmitted often cannot be sent in a straight line. Communications often have to take advantage of reflection (from walls, furniture, etc.) or diffraction (along the edges of obstructions). Diffraction happens more easily at low than high frequencies, where absorption is often stronger. As a result, obstacles in the millimeter wave bands are much more obstructive than below 6 GHz, for example. This tends to considerably reduce the range of radio connections.

A final, major problem is timing. When the transmitting or receiving station moves, or when the medium of propagation fluctuates, the system will vary over time. Such variations have to be taken into account when designing the physical layer. The problem is even more severe when these variations occur very fast (as in a high-speed train) or at high frequencies. In these cases, it is indispensable to counter the "Doppler effect", which causes signals to be mixed.

Advances during the past twenty years

Though developed in the 1980s and widely used in local WiFi networks, OFDM (orthogonal frequency division multiplexing) was not used in cellular networks till 4G. This now prevalent technique of modulation makes it easy to manage the interference between "digital symbols" created by a signal rebounding from obstacles. It is coupled with access for multiple users. Till now, multiple access has been based on the principle of orthogonality so as to avoid internal collisions within a cell. Given its efficiency, this technique figures, with a few minor modifications, once again in 5G.

As pointed out, MIMO represents one of the major innovations made during the past two decades. Its twofold objective is to reinforce the reliability of connections through multipath propagation (on condition that the antennas are spaced wide enough apart — typically a wavelength, which is easy to accomplish above 1 GHz), and to strongly increase the data rate since new paths of communication can then be used in parallel on the same channel to send different data signals. A compromise, sufficiently studied in the literature, is to devote the additional paths to improving reliability or else the bit rate. To apply this compromise, space-time codes have to be designed for beaming data efficiently between antennas. I might mention Alamouti coding (or, too, *code d'or*, a French invention). This approach naturally leads to "massive MIMO" with dozens of antennas in an array. As a result, spectral efficiency increases significantly, and many more users can be served. To its advantage, some antennas can be devoted to "beamforming" so as to focus the beam toward a particular user. The frequencies can then be reused.

Orthogonal frequency-division multiple access (OFDMA) helps avoid interference from neighboring cells. For the best use of resources since 3G, cells have been sharing the same frequencies and thus interfering with each other, especially along their borders. As a consequence, users on the border might experience serious transmission losses. To make up for this, base stations should collaborate. This is a major innovation because it reaches beyond the classical OSI model. A network's higher layers have to be used so that the two base stations' physical layers can reach an agreement about cooperation.

The techniques used are rather different depending on the level of cooperation authorized and the type of data exchanged. When base stations share streams of data, the paradigm is "virtual MIMO": the antennas are no longer colocated. This is a major challenge for the core network, since several base stations have to simultaneously receive the data from a single user. If these stations only share information about the channels of propagation (This is still complicated to do when the same frequencies are not used for both transmitting and receiving), the power or sub-bands to be used will be the only points agreed upon; and this technique will yield fewer advantages. These solutions, which had been taken under consideration for 4G but were seldom applied, are pertinent for 5G. They have been facilitated thanks to time division duplex (TDD), whereby the same frequency can be used in both directions of communication.

As for the previously mentioned error-correcting codes, the major advances made in theory during the 1990s ("turbo-codes", another French invention, and low density parity checking, LDPC) came into application during the 2000s in all systems in operation. This has made it possible to approach very closely the fundamental limit of point-to-point communications that results from Shannon and Weaver's formula.

Emerging developments

First of all, it is worth mentioning that communication systems have to respond to more and more applications with different qualities of service (ultra-high bit rates or very low latency or high energy efficiency or…). Since each system or standard has to be polyvalent, any single standard will have to be made for many lines of technical tools and widely varying frequencies.

The millimeter wave bands (6-60 GHz) will definitely soon be put to use. These very high frequencies require that we review several principles and earlier studies since the conditions of propagation are so different. In particular, receiving stations have to rely on a mixed analog-digital technology, but this is more realistic than an all-digital solution at such high frequencies.

As pointed out in the previous section, coordination between points of access and even between users has to be reinforced, above all because systems are increasingly heterogeneous with cells of varying scales (pico-, nano-, micro-, macro-, etc.) or even with device-to-device communications that do not pass through base stations (but do need their agreement).

I would now like to present a nonexhaustive list of problems with their solutions that networks are encountering given the new qualities of service required by certain applications:

- Multiple access for systems with ever more users in a single cell, whatever the cell's size, forces us to reconsider the nonorthogonal techniques that had been abandoned because of their relative inefficiency for 3G. At the time, their level of complexity could not be managed. This return of what is called NOMA (non-orthogonal multiple access) is promising owing to the recent advances in technology and algorithms that make it possible to manage more efficiently interference not only between cells but also within a cell.
- For many an industrial (robotics, driverless vehicles, etc.) or financial (high-speed trading) application, latency — the time needed to receive a signal — is a crucial factor. This means that short packets of bits have to be used. As we know, these packets do not fall in line with the predictions made by Shannon and Weaver's theory, since one of its fundamental premises is that packets are of a sufficient length. As a consequence, many results or optimizations ensuing from Shannon's formulas have to be reconsidered. Furthermore, with regard to low latency, headers for routing packets are still necessary even though they do not contain user information. However their consumption of resources is not to be overlooked. This raises questions about producing signals at a low cost. A final point: the need for synchronization has sharply increased.
- Traffic has grown exponentially because of video streaming (video-sharing websites, video products or time-delayed television); and the same information is conveyed very often, thus leading to network congestion and inefficiency. To remedy this, the storage of these data has to be decentralized nearer to users (stored, for example, in base stations or with others users). A method being studied is to identify the status of data throughout the transmission and storage processes. One problem is to make a decision about the files to be stored given that videos vary (normally declining) in popularity over time: Where to store files? How often to refresh their data? A second problem is to design a "smart" method for coding the data, in particular for the files to be saved on decentralized, distant servers. By intelligently decoupling the files and putting them in the right places, new coding techniques could be used to significantly decrease the requirements made on the network. Nonetheless, these concepts, still too new, must undergo many developments in various configurations and then be put into practice with realistic codes and modulations.

Conclusion

The development of wireless networks has not stopped. Talk has now just started about 6G networks. Given that a generation *N* lasts about ten years, it will be time to start thinking about generation *N+1* when *N* starts being rolled out. The physical layer is a key aspect of all wireless networks. While the advances gradually made since the 1990s have reached "Shannon's limit", the diversity of contemporary applications has definitely complicated the situation, as can be seen in the heterogeneity of: networks, the quality of service required, the modes of access and their optimization. "Edholm's law", which predicts an exponential growth of bit rates very close to Moore's well-known law, is, like the latter, based on innovations made during previous decades that take at least ten years to move from the laboratory to the marketplace (CHERRY *et al*. 2004).

Hopefully, this article has thrown light on the vibrancy of technology, in particular the technology related to the physical layer, which has served as the grounds for this growth.

References

CHERRY S. (2004) "Edholm's Law of Bandwidth", IEEE SPECTRUM, 1 July, available at https://spectrum.ieee.org/telecom/wireless/edholms-law-of-bandwidth.

ROLIN P., TOUTAIN L., TEXIER G., PAUL O. & CHAUDET C. (2016) *Les réseaux: Principes fondamentaux*, (Paris: Éditions Lavoisier).

SHANNON C. & WEAVER W. (1949) *The Mathematical Theory of Communication* (Champaign, IL: University of Illinois Press).