|  |  |
| --- | --- |
| D:\Mohr\Arbeitsverzeichnis\Europäische Kommission\Mobile Communications & Technology Platform\EU Technology Platform\Logos NetWorld2020\NetWorld2020 large logos\NW 2020 LBB\NW2020.jpg | logo-5G-positive_ia-text |
|  | April 2020 |

**Strategic Research and Innovation Agenda 2021-27**

**European Technology Platform NetWorld2020**

**CONFIDENTIAL WITHIN SRIA EDITORS AND CONTIBUTORS**

**DO NOT DISTRIBUTE**

**“Smart Networks in the context of NGI”**

**2020**

**Table of Contents**

# Introduction

# Some Key Performance and Value Indicators towards 2030

# User Centric and Vertical Services

# System and Network Architecture and Control

# Radio Technology and Signal Processing

## Spectrum Re-farming and Reutilisation

Allocated frequency spectrum is one of the main factors that determines the system capacity. However, radio spectrum is a very scarce resource. Especially the lower frequency bands are precious and tightly regulated. In order to satisfy the high bandwidth demands of upcoming generations of mobile systems, it is crucial to reutilise the existing spectrum resources. While the traditional approach allocates a dedicated spectrum to each radio access technology (RAT), spectrum reutilisation between RATs offers a more efficient utilisation of resources and greater flexibility, e.g., for load-balancing. Spectrum reutilisation, also known as spectrum sharing*,* can be applied to licensed but also unlicensed bands.

A straightforward approach to inter-RAT spectrum reutilisation is *spectrum re-farming*. Re-farming performs static allocation of spectrum resources to different RATs. This method was already used to clear GSM spectrum to make it available for 3G. Because of its static nature, it has poor spectrum utilisation.

A more efficient utilisation is achieved by dynamic inter-RAT resource scheduling with optimised multi-RAT handover and interference coordination. Preferably, this is based on a centralised multi-RAT radio resource management. The signalling overhead can be reduced by decentralised strategies.

For the joint utilisation of licensed and unlicensed spectrum, adaptive strategies are required such as cognitive radio concepts, in which complete spectrum awareness, based on a combination of advanced SIGINT (Signal Intelligence) techniques and AI (Artificial Intelligence) mechanisms, is needed while co-existence with existing (e.g. analogue) services is studied. Such considerations and new concepts for spectrum licensing and reutilisation are particularly important in the context of new radio technologies such as the millimetre wave, optical wireless, and terahertz communications discussed below, which have a radically different interference footprint compared to conventional sub-6GHz communications. Their highly directional links and susceptibility to blockage reduce interference, which significantly increases the potential gains of spectrum sharing and simplifies its use [9].

Spectrum reutilisation is supported by multi-RAT connectivity, which allows the UE (User Equipment) to choose the best RAT depending on the link qualities. This added diversity not only increases the performance due to better spectrum utilisation, it also makes the network more robust and resilient towards shadowing effects, hence improving the reliability and availability. On the other hand, spectrum awareness and reutilisation can help to increase security at radio-level, through detection and countermeasure of threats like RF jamming or spoofing.

Future networks will support different services, enabled by network slicing based on a multi-RAT radio access. Multi-RAT connectivity can also make flexible use of licensed and unlicensed bands. E.g., data and voice traffic can be offloaded to WiFi or LTE small cells operating in unlicensed bands as an enhanced mobility concept. Hence, utilising unlicensed bands is important and technologies to bring the quality to the licensed spectrum level are open to study. This not only increases the overall throughput but also enables low latency.

Existing *short-range wireless communication* technologies, including WiFi, Bluetooth and Zigbee, share the same spectrum, e.g. in 2.4GHz. Co-existence of diﬀerent wireless network technologies in/near such a carrier frequency may cause radio interference, which can lead to relatively high error rate in data transmission. This problem happens especially in unlicensed bands. How to efficiently share the spectrum and improve co-existence needs careful considerations. Scalability and power efficiency are critical for success of a macro, micro, or pico network. Current short-range communication technology provides either high throughput with high power, or low throughput and low power consumption. Whereas IoT devices operate in a very low power mode in most of the time, they need to support a short-time high bandwidth transmission. Scalability is needed to support both short-time high throughput transmission and low power transmission. A unified and scalable architecture will be beneficial to support both low data rate (e.g. with Bluetooth, ZigBee, RFID, NFC, etc) and ultrahigh data rate (e.g. up to 100Gbps within 10m coverage). Further requirements to be considered include, e.g. scalable network topology supporting P2P (point-to-point), MP2MP (multipoint-to-multipoint), as well as the smart home and smart building coverage; more power/cost efficient designs, e.g. for zero-power consumption in some dedicated scenarios; and the capability of information and energy simultaneously transporting (IEST).

## Millimetre Waves and Cellular Networks

Millimetre wave (mmWave) have attracted large research interest in recent years due to the huge available bandwidth required to fulfil the today’s traffic demand. This is reflected in WLAN and WPAN standards: in the license free 60 GHz band, the IEEE802.11ad WLAN standard provides rates up to 8 Gbps and the upcoming IEEE802.11ay WLAN standard will provide rates up to 30 Gbps. The fifth generation (5G) wireless networks aim to use mmWave in mobile networks, where the transmitter/receiver nodes may be moving, channels may have a complicated structure, and the coordination among multiple nodes is difficult [10]. This year’s Winter Olympics in Korea already provided first glimpse at the 5G services powered by Korea Telecom with support from global equipment makers. This show case included a 28 GHz mmWave backhaul network for moving hotspots, such as buses. Additionally, the mmWave band in combination with mobile edge computing (MEC) is highly suitable for on-demand content (multi-media) delivery services, hence enabling the enhanced mobile broadband (eMBB). This combination of mmWave and MEC is the only way to satisfy both extreme communications requirements of ultra-high speed and low latency at the same time. Beyond 2020, MEC is expected to enable automated driving using mmWave based V2X/V2V links. This requires, however, cooperative perception and the exchange of HD dynamic map information between vehicles and radio units, to enhance the visibility area. The automated driving use case can be considered as the most important application of mmWave and MEC, which requires both ultra-high speed and low latency [10]. The high channel resolution due to the large bandwidth also makes mmWave technology a prime candidate for joint communication and radar, which can complement or replace conventional radar and lidar systems in autonomous vehicles [11]. Mobile virtual and augmented reality application represent a further future use case with similarly high requirements that may only be met by mmWave technology [12].

Beyond 5G, it is expected that the data traffic due to mobile nodes (smart phone and tablets) will be more than 100 petabytes per month by 2023 [13], which is 10 times of the traffic in 2017. In Western Europe alone, the data traffic is expected to be as high as 12 petabytes per month, which amounts to 56 terabytes per person per month [14], hence, offering a huge potential to exploit mmWave bands and even the Terahertz frequency range (not considered so far by 5G). It is projected in [15] that the volume of traffic generated from smart phones will be 86% of the global data traffic by 2021 and among this more than 50% data will be offloaded to the fixed networks by means of Wi-Fi devices and small cells each month, while the remaining traffic will be covered by the cellular networks. Anticipating the requirements for short range communication beyond the year 2021, a very high data rate new Wi-Fi is inevitable. In order to achieve high data rates, one would require a large amount of contiguous bandwidth suitable for communications over short ranges, that is to be found beyond 100 GHz, for example around 140 GHz. The use of these frequency bands provides an excellent opportunity, since many antennas can be packed in a small area to direct a beam to the intended user.

An important business case for mmWave is in so-called ‘smart factories – Industry 4.0’. Due to its ability for spectrum re-use that enables multi-connectivity for high reliability, mmWave provides a complementary solution to low frequencies. Additionally, due to the high penetration losses, mmWave is inherently more secure against eaves-dropper and is a suitable candidate for industry environments. An additional feature of mmWave is sensing/positioning with high accuracy. This allows detection with higher spatial and velocity resolution that is suitable for both V2X and industry automation scenarios.

Small cells are to play a key role to cope with the increasing traffic demands in mobile network. These small cells connect to the core network via wired or wireless backhaul links. The dense deployment of small cells and a variety of services offered by the RAN having diverse requirements on throughput, latency and reliability, poses new challenges on backhaul links. One way to address these challenges is self-backhauling using mmWave, i.e., the access and backhaul share the same wireless channel. 3GPP stage 1 in its Release 15 [16] outlines the requirements for the self-backhauling in 5G networks. Among these requirements are the flexible partitioning of resources, autonomous configuration, multi-hop wireless connectivity, topology adaptation, and redundant connectivity.

One of the main challenges will be to manage the different network features introduced in 5G and developed beyond the first release and optimise them collectively. Diverse network components need to be integrated, such as D2D, self-backhauling, and multi-casting/broadcasting. While these technologies will be already available in 5G, the new challenge consists of extending them by advanced mmWave massive MIMO techniques, which are dynamically coordinated, considering interference and mobility. As mmWave network device density as well as the number of antenna elements and RF chains of devices increase, channel estimation overhead becomes a major challenge and intelligent solutions for initial access and beam training based on context information will be needed [17].

Due to high propagation loss and severe atmospheric effects, mmWave communication are more suitable for short range communication and small cells. However, the increase in the system capacity through densification is not linear and therefore potential benefits of using high frequency bands for longer ranges and macro cells need to be considered. Designing macro cells without appropriate planning and optimization may limit the potential gains of mmWave communication [18].

## Optical Wireless Communication

Despite the tremendous improvements due to the small cell concept and the allocation of new radio frequency (RF) spectrum, the continued exponential growth in mobile traffic [19] means that inevitably the RF part of the electromagnetic spectrum will not be sufficient to be able to drive the 4th industry revolution which is centred around data-driven economies and data-driven societies [20].

It is, therefore, natural to consider the infrared and visible light spectrum, both of which are part of the electromagnetic spectrum for future terrestrial wireless systems. In fact, wireless systems using these parts of the electromagnetic spectrum could be classified as nmWave wireless communications systems in relation to Section 3.2. Light based wireless communication systems will not be in competition with RF communications, but instead these systems follow a trend that has been witnessed in cellular communications by inspecting all the generations developed during the last 30 years. Light based wireless communications simply adds new capacity – the available spectrum is 2600 times larger than the entire RF spectrum.

An important advantage is that off-the-shelf optical devices can be used to harness these unregulated and free transmission resources. By using advanced devices, lab demonstrations showed 8 Gbps from single light emitting diodes (LEDs) and 17.6 Gbps using laser diodes (LEDs) [21]. Recently, a record of received data rates of 500 Mbps by using a single solar cell has been demonstrated. The use of these types of ‘data’ detectors has the appealing advantage of achieving simultaneous energy harvesting and high-speed data communication – a feature that will become ever more important in mobile machine-type communication (MTC) [22]. By 2026, it is expected that micro-LED technologies and spatial multiplexing techniques will be mature and cost effective such that white light based on different wavelengths will unlock throughput, leading to potentially 100 Gb/s plus for ultrahigh-data-rate VLC access points [23].

Networked and cellular wireless networks which are based on visible light communication (VLC) are referred to as LiFi (light fidelity) [24]. LiFi enables bi-directional networked communication including multiuser access and handover (please refer to Figure 2 for a taxonomy of different light communication approaches). The blue arrow in Figure 2 indicates that the major research efforts in the last 15 years have been focused on enhancing link data rates of intensity modulated (IM) / direction detection (DD) optical wireless communication systems. With the advent of LiFi the research focus has begun to shift to challenges related to networking issues using light.

As in RF networks, there are issues surrounding interference management and interference mitigation in LiFi networks. However, since, for example, there is no multipath fading because the detector sizes are much larger than the wavelength, techniques developed for RF systems may only be sub-optimum. There are also fundamental differences as a result of IM/DD, in that signals can only be positive and real-valued. Consequently, new LiFi-bespoke wireless networking methods must be developed. Moreover, because light can be confined spatially by using very simple and inexpensive optical components, interference can be controlled much easier. This feature also allows step-change improvements of the small cell concept as single cells might cover sub-m2 areas.

Furthermore, due to the extremely small wavelength, the active detector sizes are very small, and massive MIMO structures can be implemented at chip-level. Edinburgh University, for example, has developed a massive MIMO LiFi chip composed of 49 avalanche photodiodes (APDs) on CMOS which, as a major breakthrough, requires only very low-voltage negative biases while achieving at least 10 dB APD gains. The size of the 49-dector die was merely 2.8 mm x 2.8 mm. This property can be used to develop unique and LiFi-bespoke MIMO systems, networked MIMO approaches, and new angular diversity techniques in conjunction with low computational complexity cooperative multipoint (CoMP) systems. Diversity techniques in LiFi systems are especially powerful to combat random blockages that naturally occur in a mobile scenario.

Moreover, the spatial confinement of signals in LiFi enables the development of radically new physical layer security concepts.

|  |
| --- |
|  |

1. A taxonomy of emerging light communication technologies

Free-space optical (FSO) is point-to-point long range optical wireless communications with target data rates of tens of Gbps primarily using laser diodes and coherent transmission. VLC has been used in the context of line-of-sight high-speed point-to-point communication, primarily using LEDs in conjunction with IM/DD. VLC systems are usually designed for ranges less than 100 m, as well as for bi-directional communication. Optical camera communication (OCC) in contrast is simplex communication using embedded CMOS camera sensors as data detectors. Due to the use of CMOS sensors, the achievable data rates are well below 1 Mbps. OCC is primarily used for indoor navigation, asset tracking and positioning. These applications assume some user mobility.

LiFi is currently being standardised in a Task Group within IEEE 802.11. The new LiFi standard has received the following reference: IEEE 802.11bb. Similarly, VLC is being standardised in IEEE 802.15.13, while OCC has been standardised in IEEE 802.15.7r1.

**Convergence with 3GPP access:** LiFi communication is bi-directional. Due to the abundance of optical spectrum, typically the visible spectrum is used for the downlink by piggy-backing on lighting systems, while the infrared spectrum is used for uplink transmission. The simplicity of IM/DD in conjunction with advanced layered modulation techniques [25] enable highly energy-efficient transmission systems suitable for the uplink. These highly energy-efficient and spectrum efficient modulation techniques are based on multicarrier approaches. Therefore, it could be argued that a *tight interaction between radio and optical components should be considered at the level of baseband processing*. Since OFDM transmission (e.g. 5G waveforms) is feasible on a free-space IM/DD optical link, it is definitely worth investigating the use of the same basic waveform and protocol stack for radio and LiFi systems. This would allow for a *common baseband processing platform* in both the small-cell transmitters and terminal receivers. Moreover, the 3GPP access-layer protocols are perfectly adapted to the use of downlink-only component carriers.

## Terahertz Communications

Wireless data rates have doubled every eighteen months for the last three decades. Following this trend, Terabit-per-second (Tbps) links are expected to become a reality within the next five years. While mmWave communications are a step in the right direction, the total consecutive available bandwidth in such systems is less than 10 GHz. Consequently, supporting Tbps would require a physical layer efficiency of 100 bit/s/Hz, which is several times higher than the state of the art.

In this context, **Terahertz-band (0.1–10 THz)** communication is envisioned as a key technology to satisfy the need for much higher wireless data rates [26]. This frequency band, which lies in between mmWave and the far infrared, supports huge transmission bandwidths: from almost 10 THz for distances below one meter, to multiple transmission windows, each tens to hundreds of GHz wide, for distances beyond several tens of meters. However, this very large bandwidth comes at the cost of a very high propagation loss. Moreover, for many years, the lack of efficient ways to generate and detect THz signals has hampered the use of the THz-band in practical communication systems.

To date, different technologies are being considered to close the so-called THz gap. In an *electronic* approach, the limits of silicon CMOS technology [27], silicon-germanium BiCMOS technology [28], and III-V semiconductor HEMT, mHEMT, HBT and Schottky diode technologies [29] are being pushed to reach the 1 THz mark. In a *photonics* approach, uni-travelling carrier photodiodes [30], photoconductive antennas [31], optical downconversion systems [32] or, more recently, quantum cascade lasers [33] are being investigated for high-power THz systems. In both approaches, fundamental device limits are being reached, as the frequency is ”too high” for electronic devices and the photon energy is ”too low” for photonic devices to efficiently operate at *true* THz frequencies.

More recently, the use of **graphene to develop novel plasmonic devices** for THz communications has been proposed. Graphene is a two-dimensional (2D) carbon-based material that has excellent electrical conductivity, which makes it very well suited for propagating extremely-high-frequency electrical signals [34]. Moreover, graphene supports the propagation of THz surface plasmon polariton (SPP) waves at room temperature. SPP waves are surfaced-confined electromagnetic waves generated by the global oscillation of electrons. By leveraging the properties of graphene, nano-transceivers [35] and [36] and nano-antennas [37] and [38] have been proposed and are being developed. These devices are intrinsically small, efficiently operate at THz frequencies, and can support very large modulation bandwidths. Moreover, graphene is ”just the first” of a new generation of 2D materials (such as MoS2 or Hb-N), which can be stacked to create new types of devices and leverage new physics.

In parallel to the development of new device technologies, there is a need to understand and model the THz-band channel. In the case of line-of-sight (LoS) propagation [39], the main phenomena affecting the propagation of THz waves are the spreading loss and the molecular absorption loss. The *spreading loss* accounts for the attenuation due to expansion of the wave as it propagates through the medium and is common to any wireless communication system. The *molecular absorption loss* accounts for the attenuation that a propagating wave suffers because a fraction of its energy is converted in vibrational kinetic energy in molecules (especially water vapor). In the case of non-line-of-sight (NLoS) propagation [40], in addition to the two aforementioned phenomena, high reflection loss, diffused scattering and diffraction by obstacles need to be captured. Ultimately, stochastic multi-path channel models are needed to statistically characterise the channel. In addition, there is a need to understand the channel characteristics in mobile environments, see e.g. [41].

In light of the capabilities of THz devices and the peculiarities of the THz-band channel, there is a need to develop new communication algorithms and networking protocols, tailored to THz communication systems. At the physical layer, new types of modulations are needed. For short-range communications (below one meter), the use of impulse-radio-like communication based on the transmission of one-hundred-femtosecond-long pulses following an on-off keying modulation spread in time has been proposed [42]. Such very short pulses are already at the basis of many THz sensing systems and can be generated and detected with current technologies. For longer communication distances, new **dynamic bandwidth modulations** [44] are needed to not only overcome but even leverage the unique distance-dependent bandwidth created by molecular absorption. Specialized (single-carrier) waveforms that are robust to phase noise, Doppler, and carrier frequency offset have been investigated as well [43].

Independently of the modulation, and similar to any wired or wireless Tbps communication system, physical-layer synchronisation (both in time, frequency and phase) becomes a major challenge. The front-end non-idealities, e.g. non-linearity and phase noise, can severely impact the achievable throughput. Going to Tbps throughputs implies increasing the bandwidth to tens of GHz. This is another challenge for implementations for two reasons: first, ADCs and DACs in the tens of Gsamples/s are needed; second very wideband analogue baseband circuits are needed (half the RF bandwidth). The lack of digital-to-analogue and analogue-to-digital converters (DACs and ADS, respectively) able to handle multi-GHz bandwidth signals, limits the application of traditional digital signal processing and motivates the research and development of new mixed (digital and analogue) techniques where some traditionally digital functions such as synchronisation or equalisation can be moved to the analogue domain. Additional challenges include new channel coding strategies, which leverage the uniqueness of the THz-channel, or physical layer security schemes for THz-signals. Very generally, efficiencies become dominant bottlenecks: at 1 Tbps, an efficiency of 1 pJ/bit (impossible today if we consider the whole PHY) translates into 1 Watt of power consumption; similar considerations about implementation efficiencies in silicon technology (area efficiency (bit/s/mm²) and power density (W/mm²)) show huge challenges at Tbps rates.

Similarly, many challenges arise in the higher layers of the protocol stack. At the link layer, novel **MAC protocols** are required for THz-band communication networks, since classical solutions do not capture the peculiarities of this band. The very large available bandwidth almost eliminates the need for nodes to contend for the channel. The transmission of very short signals also minimises the chances for collisions. All these come at the cost of more complex synchronisation schemes between devices. Ideas to be explored for new MAC protocols include, among others, the development of receiver-initiated transmission schemes to ensure that the transmitter does not waste resources when the receiver is not available, especially when highly directional systems are used. Additional challenges also include packet size optimisation and adaptive error control strategies.

At the network layer, new **routing mechanisms** could be developed that take into account the availability of both classical active relaying nodes as well as novel passive dielectric mirrors, which can direct the signal towards its final destination. In addition, new routing metrics that consider the channel molecular composition and its impact on the available distance-dependent bandwidth need to be explored. At the transport layer, as wireless multi-Gbps and Tbps links become a reality, the aggregated traffic flowing through the network will dramatically increase. These will introduce many challenges at the transport layer regarding **congestion control** as well as end-to-end reliable transport. For example, we expect that a revision of the TCP congestion control window mechanism will be necessary to cope with the traffic dynamics of THz-band communication networks.

For the validation and refinement of the developed solutions, new **experimental platforms** and integrated testbeds will be needed. For the time being, these are mainly focused in the sub-THz windows (300 GHz, 650 GHz), but systems at *true* THz frequencies will be required. Finally, in parallel to all the scientific developments, work needs to be done towards regulation and standardisation of the THz-band [45].

Smart devices will be able to sense and monitor the environment based on the radiated communication signals. This will enable smart context-aware networking and new applications, which are based on the knowledge of certain features of objects in the surrounding, e.g. position, velocity, structure, or used frequency bands. Sensing not only includes the recognition of devices that radiate THz signals, but also passive imaging techniques can be integrated. Environment-aware communication will be a key component of next generation smart networking, which will push the integration of data analytics to a new level.

## Ultra-Massive MIMO

The grand challenge for mmWave, THz-band and optical communications is posed by the very high and frequency-selective path loss, which easily exceeds 100 dB for distances over just a few meters in the presence of LoS (line-of-sight) and becomes even worse in NLoS (non-line-of-sight) conditions. As a result, high-gain directional antennas are needed to communicate over distances beyond a few meters.

Similarly, as in lower frequency communication systems, antenna arrays can be utilised to implement MIMO communication systems, which are able to increase either the communication distance by means of beamforming, or the achievable data rates by means of spatial multiplexing. In the last few years, the concept of Massive MIMO has been introduced and heavily studied in the context of 5G systems [46], [47] and [48]. In such schemes, very large antenna arrays with tens to hundreds of elements are utilised to increase the spectral efficiency to communicate over a large distance. In these arrays, it is important to take mutual coupling between the antenna elements into account in a physically consistent way [49][50]. Very large antenna arrays have been proved to be very useful for mmWave communication systems [51] and [52]. When moving to the THz-band, antennas become even smaller and more elements can be embedded in the same footprint. However, linearly increasing the number of antennas is not enough to overcome the much higher path loss in THz-band.

In this context, the concept of **ultra-massive MIMO (UM-MIMO)** communications,enabled by very dense plasmonic nano-antenna arrays, has been recently introduced in [53] and [54]. Instead of relying on conventional metals, nanomaterials and metamaterials can be utilised to build plasmonic nano-antennas (see Section 3.4), which are much smaller than the wavelength corresponding to the frequency at which they are designed to operate. This property allows them to be integrated in very dense arrays with innovative architectures. For example, even when limiting the array footprint to 1 mm × 1 mm, a total of 1024 plasmonic nano-antennas designed to operate at 1 THz can be packed together, with an inter-element spacing of half plasmonic wavelength. Such plasmonic nano-antenna arrays can be utilised both at the transmitter and the receiver (1024×1024) to simultaneously overcome the spreading loss problem (by focusing the transmitted signal in space) and the molecular absorption loss problem (by focusing the spectrum of the transmitted signal in the absorption-free windows).

By properly feeding the antenna array elements, different operation modes can be adaptively generated. In **ultra-massive beamforming**, all the nano-antennas are fed with the same plasmonic signal, as in conventional beamforming. This mode can effectively overcome the very high attenuation at mmWave, the THz-band and optical frequencies and thus enhance the communication distance. Moreover, beamforming has the benefit of avoiding co-channel interference while exploiting the angle diversity by steering the narrow beam dynamically to the targeted angle directions. In **ultra-massive spatial multiplexing (UM-SM)**, physically or virtually grouped array elements can be assigned to communicate with an individual user. This mode uses multiple streams on a single carrier to increase the capacity per user and can be most effective when radio links operate in a high SNR regime and are bandwidth-limited. This mode improves the network throughput by means of spatial multiplexing, given that the UM-MIMO channel matrix is well-conditioned, or equivalently, provides sufficient diversity and rank. Obviously, any combination in between UM Beamforming and UM Spatial Multiplexing is possible.

In addition, to maximise the utilisation of the mmWave- and THz-channel and enable the targeted Tbps-links, more than one spectral window could be utilised at the same time. In this direction, **multi-band UM-MIMO** enables the simultaneous utilisation of different frequency bands by leveraging the electrically tunable frequency response of graphene-based plasmonic nano-antennas. By tuning (virtually) grouped sub-arrays to different frequencies, a single UM-MIMO system can simultaneously cover multiple transmission windows. One of the key advantages is that the multi-band approach allows the information to be processed over a much smaller bandwidth, thereby reducing overall design complexity as well as improving spectral flexibility. In this direction, advanced **space-time-frequency coding and modulation techniques** need to be developed for the UM-MIMO systems to exploit all of the spatial, temporal and frequency diversities, and hence, promise to yield remarkable performance improvements.

Besides the challenges related to the plasmonic nano-antenna array technology, the realisation of UM-MIMO communication requires the development of novel **accurate channel models** able to capture the impact of both plasmonic nano-antenna arrays in transmission and reception, as well as the behaviour of a very large number of parallel THz-waves propagating in space. Existing MIMO or Massive MIMO channel models for lower frequency bands [46], [47], [48], [55], [56] and [57] cannot be utilised because they do not capture the peculiarities of the THz-band channel, including the frequency-selective absorption loss or the very high reflection loss. Similarly, the few THz Massive MIMO channel models developed to date [58][59] do not take into account the capabilities of plasmonic nano-antenna arrays, such as the sub-wavelength size and separation, and the opportunities this brings. Therefore, a 3D UM-MIMO channel model for ultra-broadband communications is needed.

A new and revolutionizing technique able to improve substantially the performance of wireless communication networks is smartly changing the propagation characteristics of the wireless channel through the use of **intelligent reflecting surfaces** (IRS), which are made of a large number of low cost passive reflecting elements able to independently change the amplitude and/or phase of the incident signal so as to achieve specific propagation effects [60][61]. The IRS can be used to improve the coverage, reduce interference levels, increase system capacity. Additionally, they can be employed to increase physical layer security and even support wireless power transference.

**Large intelligent surfaces** (LIS) [62] can be an even powerful technique. They are also surfaces covered by a large number of antenna elements but, contrarily to IRS, where the elements are passive, the LIS elements are low complexity antennas connected to a processing unit. These elements are activated according to the user location and its transmission requirements. This allows unprecedented capacity gains [62], as well as accurate positioning [63]. LIS can also be used for wireless transference of power.

Although the LIS is made of antenna elements with very low complexity, its implementation can still involve considerable challenges due to the large number of antennas and the associated circuitry. A natural option to overcome these difficulties is to split the LIS surface in small panels (say with 10x10 antennas), each one able to separate a small number of users (say, 2 or 4 users), and only a fraction of the panels is activated. Radio stripes [64] are another interesting variant of LIS, with the antennas placed over a stripe instead of a surface. As with LIS, radio stripes can enable huge capacity gains, enabling cell-free systems [65], together with accurate positioning, having the additional advantage of a very simple and low cost and implementation.

The communication using LIS and/or radio stripes schemes will require advanced, low complexity techniques for the signal separation, as well as new resource allocation spacial aspects (i.e., which antennas/panels are activated for a given user). Highly efficient, low complexity amplifiers, namely switched amplifiers, will be required, especially for systems operating at higher frequency bands (e.g., mm-waves and THz). To further improve the performance, disruptive techniques that take advantage of hardware imperfections such as nonlinear and/or memory effects can be employed [66][67].

The large dimensions of LIS and radio stripes (several tens of meters), together with the relatively short communication ranges (tens of meters or even less), leads to a near field communication effects, with its inherent potential and challenges. The channel estimation can be a considerable challenge due to the large number of parameters to estimate (this also applies for IRS). To overcome these difficulties, parametric channel estimation and tracking techniques [68] can be employed, eventually supported by positioning information.

In the different variations of ultra-massive MIMO schemes, **machine learning** (ML) can be used in certain scenarios. One scenario is when the available mathematical model is flawed or is only a coarse approximation of the real underlying model. An example is the linear channel model which neglects non-linearities caused by hardware and other practical effects. ML can be used to improve the solutions that are based on the linear model approximation. ML can also be effective to define detection schemes that take advantage of nonlinear effects [69][70]. A further example is molecular communications discussed in Chapter 9, where an accurate model describing the molecular channel is yet to be developed. Another scenario is where the optimal solutions are computationally demanding and/or not possible for practical hardware architectures. In this context, ML can be used to approximate those optimal solutions with lower complexity, albeit clearly at a performance loss. Examples include maximum likelihood detection, channel estimation, etc.

## Waveform, Multiple Access and Full Duplex

Cyclic prefix orthogonal frequency division multiplexing (CP-OFDM) has been adopted in several wireline and wireless standards such as ADSL, Wi-Fi, LTE, and has recently been proposed for 5G [71]. CP-OFDM divides the bandwidth into several orthogonal subcarriers. The orthogonality is preserved as long as the transmitters are synchronised to each other. Fine time and frequency synchronisation is then required to maintain the subcarrier orthogonality. However, strict synchronisation is limiting in certain scenarios. For example, sporadic access in internet of things (IoT) and machine-type communications (MTC) requires relaxed synchronisation schemes, in order to limit the length of the signalling overhead [72]. Ideally, the massive number of devices could just transmit their messages asynchronously; being only coarsely synchronised [72]. This could also be advantageous for low-latency communications. However, in multi-user asynchronous access, the CP-OFDM subcarriers are no longer orthogonal, which introduces high inter-carrier interference [73]. Therefore, CP-OFDM is no longer viable in such scenarios.

Several waveforms, e.g. filter bank multi-carrier (FBMC), generalised frequency division multiplexing (GFDM) and universal filtered multi-carrier (UFMC) may be more suitable since their subcarriers are better localised in the frequency domain, and therefore limit the inter-carrier interference. A good frequency localisation may also be beneficial due to other reasons, e.g. sensitivity to phase noise in mmWave, required accuracy of frequency-synchronisation, etc.

The waveforms differ in whether they are orthogonal, whether and how they employ a cyclic prefix, and how the subcarriers are filtered to make them well localised in the frequency domain [74]. FBMC is quasi-orthogonal, performs a per sub-carrier filtering and eliminates the cyclic prefix, but care must be taken in the implementation since contrary to OFDM, GFDM and UFMC, it uses offset quadrature amplitude modulation (OQAM). GFDM also performs per-subcarrier filtering and reduces the overhead of the cyclic prefix by employing it for several symbols, instead of per symbol as in OFDM. However, its non-orthogonality introduces self-interference even if the transmitters are perfectly synchronised. This requires a more complex receiver using e.g. successive interference cancellation. UFMC eliminates the cyclic prefix and applies a filtering for a sub-band consisting of several subcarriers, where the subcarriers within a sub-band are orthogonal to each other but the sub-bands are non-orthogonal, introducing less inter-carrier interference compared to GFDM. Numerous comparisons between those waveforms have been made regarding implementation complexity, spectral efficiency, robustness towards multi-user interference (MUI) and resilience to power amplifier non-linearity etc., see e.g. [75] and [76].

Even if they have not yet been adopted in 3GPP, these post-OFDM waveforms are promising schemes, especially in asynchronous multiple access for massive IoT scenarios. Therefore, application-oriented research on algorithms and proof-of-concept implementations are needed to make them more mature.

Relaxing the orthogonality constraint generally leads to a more efficient and flexible use of the wireless channel. Non-orthogonal multiple access (NOMA) has attracted significant attention in recent years, as it does not only result in larger achievable rates for scheduled uplink and downlink transmissions, but also provide means to cope with packet collisions for MTC scenarios with grant-free access [77]. Challenges for NOMA research include

* *User pairing:* Current research in NOMA mostly considers pairing of two users with the same resource due to high interference accumulation at receiver. However, some publications suggest that with a careful design, more than two users can be paired to use same resource [78]. The application of dense users using same resource can be further extended by considering other candidates such as massive MIMO and mmWave.
* *Interference cancellation:* Several potential candidates for interference cancellation in NOMA have been discussed intensively. Successive interference cancellation (SIC), parallel interference cancellation (PIC) and hybrid interference cancellation are some possibilities [79]. Yet the challenge to find the optimal one is still broadly open. The main focus is to find balance between error rate performance, number of paired users, each user’s throughput and overall throughput.
* *Power control:* It is important to design intelligent power control scheme since power is a crucial resource in NOMA. The design of power control in NOMA can affect other performances such as receiver interference level and throughput. E.g. the work in [80], where the power constraint is jointly allocated in full-duplex NOMA, can be further extended to multi-cell scenario.
* *Physical layer security:* One of the earliest questions when NOMA was introduced is: Is NOMA safe? This question arose since in most NOMA cancellation techniques, one user can decode another user’s signal in its own device. Such an issue needs further investigations (see e.g. [81]).

Furthermore, advanced self-interference cancellation techniques enable in-band full duplex transceivers that offer a wide range of benefits, e.g., for bidirectional communication, cooperative transmission in heterogeneous networks, and cognitive radio applications [82].

## Coding and Modulation

Channel coding aims to correct errors to establish reliable communication and can be regarded as one of the most complex parts of the baseband transmission chain [83]. For decades, researchers sought for channel codes with good error correction performance approaching Shannon’s capacity limits with manageable complexity. Modern channel coding schemes such as Turbo, LDPC and Polar codes with excellent performance made their way into several communication standards after advancements in semiconductor technology. However, as the decoders for those codes are very complex, there will be implementation bottlenecks (w.r.t. computational complexity, algorithm parallelisation, chip area, energy efficiency, etc.) to be addressed for high throughput (e.g. when throughput is over multiple Gigabits per second) and/or low latency applications are targeted by future communication standards.

For Polar codes, the state of the art CRC aided successive cancellation list (CA-SCL) decoding doesn’t scale up well with throughput due to its serial nature of the algorithm. Hence, iterative algorithms like multi-trellis BP (belief propagation) decoding [84][85] may be considered. Furthermore, modified polar code constructions can be adopted to improve the performance of iterative BP algorithms. Approaches like unfolding the iterative decoders using deep neural networks can be used to improve the latency and throughput of the decoders [86][87].

Even though these modern coding schemes show near-capacity error correction performance for many channels (e.g. binary input additive white Gaussian channels, BI-AWGN), their combination with higher order modulation schemes (such as QAM) can lead to a sub-optimal performance. One reason for this degradation is the so-called ‘shaping loss’ caused by the probability distribution of the transmitted symbols [88]. In order to approach capacity, the transmitted symbols need to have a certain probability distribution (e.g. discrete Gaussian distribution is needed for the transmission over AWGN channels) and using uniformly distributed symbols results in a performance loss, which can be up to 1.53 dB on AWGN channels.

Several solutions for constellation shaping are proposed to compensate this loss. One option is to optimise the locations of the modulated symbols in the constellation diagram to obtain non-uniform constellations (NUC), as adopted in the ATSC3.0 standard [89]. This scheme is also called geometric shaping and shows improvements compared to uniform signalling. Another approach is the so-called probabilistic shaping [90][91][92], where a shaping encoder is employed to encode messages in a way that the transmitted codewords have a non-uniform probability distribution, resulting in a capacity achieving distribution when combined with simple QAM symbols. This approach is shown to perform close to channel capacity. Another feature of probabilistic shaping is that the probabilities of transmitted symbols can be changed to adapt the transmission rate without changing the FEC code. This is of particular importance since a single FEC code design is sufficient for rate-adaption. Considering the diverse requirements of future communications systems, several shaping encoders suitable for both high throughput and ultra-low latency (short blocks) have been proposed in the literature [91] and [93]. However, hardware implementation of efficient shaping encoders and decoders needs further investigations.

Constellation shaping provides significant improvements in terms of error correction performance. In general, signal shaping is a fundamental and important technology to further improve the spectral efficiency of wireless and wireline communication systems, as the shaping loss may be considered as one of the last gaps between Shannon’s information theory and the practical communication systems to be bridged.

## Positioning and Sensing

Especially in the massively connected world of the “Internet of Things” (IoT), it is getting more and more important to be aware of where all these “things” are located, e.g. via positioning and sensing. Mobile radio-based technologies envisaged for a future system will play an essential role in providing high accuracy positioning of the “things”. State-of-the-art communications systems like 4G LTE support positioning in a non-cooperative way, i.e. in the downlink, several base stations send reference signals and the UE measures them and, in the uplink, a UE sends reference signals and several base stations measure them. This is good enough to support the requirements imposed by the FCC for localisation of emergency calls (so-called E-911), where an accuracy in the order of 50 m can be required [94]. There are use cases for future mobile communications, e.g. V2X, smart factory and others, however, where a higher localisation accuracy is required. For instance, for V2X vulnerable road user discovery, an accuracy as high as 10 cm may be required (see e.g. [95]). Currently, 3GPP is considering a positioning accuracy of 50 cm for the future 5G NR standard, due to the **higher frequencies and large signal bandwidths**, **dense networks** as well as improved **device-to-device** communications available for 5G. In general, wireless sensing can be used to monitor and record the physical conditions of the environment, which can in turn be utilized for more power-efficient communications, such as in communication with integrated sensing. High accuracy sensing without weakening the wireless communication should be an indispensable feature for future networks, including short-range communication.

While it is by now known that MIMO systems improve spatial diversity and result in spatial multiplexing gains, their power in improving positioning accuracy has not yet been fully exploited. Large antenna arrays at the BS (base station) result in very fine angular sampling, which can be leveraged for positioning methods. Further, existing positioning methods only work well in strong LoS environments in general. Many environments, however, experience strong multipath which cause performance degradations and reduces position accuracy. For that reason, the existing methods need to be revised or new methods need to be developed to accommodate multipath propagation. Such methods can additionally leverage the presence of large antenna arrays at the BS [96]. Clearly, having multiple antennas at the UE can improve positioning. In particular, the ability for a receiver to measure the **time-of-arrival, angle-of-arrival, and angle-of-departure** of distinct multipath components improves not only to ability of the UE to exploit the LoS path (including the possibility to determine the UE’s orientation), but also to **map the environment**, in order to determine location and extent of dominant reflectors. Such **radar-like abilities** can occur in either bistatic operation (piggybacking on standard positioning reference signals) [98], or in monostatic operation (requiring full-duplex processing at the BS) [97]. The price to pay is the complexity cost of the associated simultaneous localization and mapping algorithms [98], which likely need to be solved through mobile edge computing. Moreover, fully harnessing these physical dimensions would require **novel signals in temporal spatial and frequency domain** [99].

Once such radar-like abilities of communications systems are available, a convergence of radar and communications technologies is very likely. However, passive radar technologies depending on RF (communications) transmitters that are not under control of the locating entity cannot be used for critical applications where **service availability and reliability** is crucial. Active radar is necessary. The more active radar systems will be employed, the more interference will be experienced. Radio resource management is one means to cope with the interference. Well-known technologies from cellular communications can be employed using communications links for exchanging such control information. Ultimately, new waveforms can be deployed for combining radar and communication [100][101].

Cooperation can boost the positioning accuracy [102][103], especially in massively connected scenarios. In cooperative positioning, the user equipment (UEs) can send and receive signals and exchange their position relevant information. If the density of UEs is large, it is likely that there are line of sight (LOS) propagation conditions to each UE from several UEs, which is significantly increasing achievable localisation **accuracy and coverage**. There are two different approaches to position calculation, a centralised approach where a central entity calculates the position and a decentralised approach where UEs calculate their position based on the position estimates of the UEs in their vicinity. With slidelink communication in 5G, new opportunitiesor localization and sensing arise, not only in signal design, but also protocols, and algorithms. Important use cases are in the vehicular and drones contexts, where relative location information from cooperative links can have direct implications for safety and global situational awareness.

Accurate positioning can be leveraged to enable **location-aware communications** [104], e.g. design of narrow beams targeted towards the intended user in traditional cellular systems, facilitate autonomous driving, etc. These effects will become increasingly pronounced as communications systems shift to ever higher carrier frequencies (0.1 THz and beyond). Furthermore, accurate positioning is a prerequisite for emerging industrial and factory applications. Therefore, in contrast to legacy systems, positioning has a big impact on the operation of future communication systems. For these reasons, investigating new positioning paradigms, e.g. for joint communication and positioning, is essential, as it can further improve spectral efficiency, energy efficiency, and reduce latency. Similar to other applications such as ultra-massive MIMO, **machine learning** can be used in the case of positioning where complex propagation environments cannot be accurately modelled.

## Massive Random Access

The future vision of IoT envisages a very large number of connected devices, generating and transmitting very sporadic data. The challenge here is how to coordinate such a network without spending the whole network resources and node energy in protocol overhead. Modern information theoretic research has formalised this problem as follows: consider a number of nodes, each of which makes use exactly of the same code, which is hardwired into the device for system simplicity and cost reasons. These nodes access a common transmission resource at random in a very sporadic manner. The receiver (e.g., a base station) must decode the superposition of codewords without knowing a priori who is transmitting [105]. After decoding the messages (payload), the ID of the transmitter can be found as part of the message, if necessary. For example, in some applications it is important to know the transmitter, but there are applications in which it is important to get the data and not the identity of the transmitter. The challenge now is to design such new random-access codes for which the superposition of up to K distinct codewords can still be uniquely decoded.

This new random-access paradigm is inherently related to **group testing**: A set of statistical procedures for which it is possible to identify the presence of certain individual agents by sampling combinations thereof [106] and [107]. A related setting consists of coded slotted Aloha, where sparse codes with iterative message passing decoding are developed along multiple random transmissions, to effectively eliminate interference by a sort of low-complexity successive interference cancellation [108]. The performance can be further improved using low-rate channel codes in combination with multi-user detection at the physical layer [109].

A related problem consists of activity detection using a receiver with a large antenna array: In this case, users are given unique signature sequences and transmit at random in a completely uncoordinated way. The base station has multiple antenna observations and must identify the “active set” of users that are transmitting. This problem is related to **compressed sensing** where the sparse vector to be estimated is the vector of 0s and 1s, denoting “absence” or “presence” of the transmitters. Modern techniques based on approximated message passing (AMP) can be used for this purpose [110] and preliminary research results show the exact trade-off between the length of the signature sequences (protocol overhead) and the number of active users, such that the probability of identification error can be made as small as desired [111] and [112]. Compressed sensing-based multi-user detection may also be combined with coded random access schemes [113].

In both cases the massive random-access and the activity detection problems, a significant research effort must be made in order to bring the abovementioned theoretical ideas to practice and to solid and principled system design. Furthermore, even the basic theory needs to be extended, for example, to encompass asynchronism and presence of unknown parameters, such as phase and frequency offsets, and random fading coefficients, for which the current theory has only partial answers.

In a second step, this line of research should consider waveforms adapted for low-latency sporadic access for the cyber-physical systems characteristic of the tactile Internet [114]. Here, sub-ms latencies may be required in order to control moving or even flying objects (passenger drones) or other similar scenarios requiring the combination of ultra-reliable communication with centralised control systems. Similar mechanisms will also be required for evolved Industry 4.0 applications [115]. It is envisaged that the physical-layer transport mechanisms will be associated with real-time cloud computing (mobile edge computing) in proximity to the radio network to implement the necessary control loops. This concerns primarily sub-6GHz access for the uplink and massive connectivity of objects to wireless infrastructure. The objective is to provide solutions for the evolution of cellular IoT uplink waveforms and protocols that scale to huge number of connected devices with stringent energy and potentially latency constraints.

## Wireless Edge Caching

Wireless communication networks have become an essential utility for citizens and businesses. Wireless data traffic is predicted to increase by 2 to 3 orders of magnitude over the next five years [15] and [116]. The implications of these trends are very significant: while continued evolution is to be expected, the maturity of current technology (e.g., LTE-Advanced for cellular and IEEE 802.11ac for WLAN) indicates that the required orders of magnitude throughput increase cannot be achieved by an incremental “more-of-the-same” approach. As far as wireless capacity is concerned, the forthcoming 5th Generation (5G) of standards and systems is focused to a certain extend on the traditional view of “increasing peak rates” [117]. In contrast, it is widely recognised that a major driver of the wireless data traffic increase is on-demand access to multimedia content (Wireless Internet) [15] [116]. Peak rates do not necessarily yield an improved user Quality of Experience (QoE). For example, typical video streaming requires rates ranging from ~400 kbps (standard quality) to ~2 Mbps (high quality). What really matters for the end user QoE is the availability and stability of such rates, so that a video can be played anywhere, at any time, and without interruptions. Also, we observe that the users’ content consumption pattern and the operators' data plans are dramatically mismatched. For instance, a standard monthly data plan in the EU includes ~3 Gbytes of LTE traffic at a cost ranging between 15 and 50 EUR, while a single movie requires ~1.5 Gbytes of data, such that the whole plan would be depleted by streaming ~2 movies.

In light of the above considerations, a novel content-aware approach to wireless network design is needed. Such novel approach should support the paradigmatic shift **“from Gigabits per second to a few Terabytes per month for all”**. More precisely, the special features of on-demand multimedia content can be leveraged in order to deliver a target of ~1 $TB/month of content data to each user in a scalable and cost-effective manner. This target is far more challenging than achieving Gbps peak rates, which have been already demonstrated by various “5G-ready” experimental platforms [118] and [119].

Meeting this challenge requires a **profound and non-incremental advance** in the information theoretic foundations, in the coding and signal processing algorithms, and in the wireless network architecture design, in order to exploit the potential gain of content-awareness.

Recent research in information theory and wireless communication has shown that content distribution over a wireless network (e.g., on-demand video streaming) can be made much more efficient than current state-of-the-art technology by caching content at the wireless edge [120], [121], [122] and [123][123]. This means pre-storing segments of the content files at the base stations, at dedicated “helper” nodes, and also in the user terminals.

Traditional caching (e.g., prefix caching) decreases the transmission load by the fraction of data already present (pre-cached) at the destination. With these novel modern techniques, based on extensive use of network coding, it is possible to show that a constant (non-vanishing) per-user throughput can be achieved while the number of users grows to infinity. We refer to this behaviour as “full throughput scalability” [124]. For the sake of concreteness, consider the analogy with conventional TV broadcasting: in this case, leveraging the broadcast property of the wireless medium, an infinite number of users can be served with a finite transmission resource, i.e., a finite bandwidth and transmit power. For example, this approach is taken in the so-called enhanced multicast-broadcast multimedia service (eMBMS) in 4G networks. Now, the reason for which eMBMS turned out not to be a huge success is that users do not consume wireless multimedia as they used to consume traditional live TV: they wish “on-demand” services, to access what they want at the desired time and location, and not at the time decided by a TV broadcaster. With on-demand delivery, the broadcast nature of the wireless medium cannot be exploited in a direct and trivial manner. In fact, streaming services today treat the on-demand traffic as unicast individual traffic, as if the content was individual independent data. An important consideration here is security. The data can be stored on user’s local cache that depends on the demand of other users in the network. This leads to the possibility of spying and tampering. Authors in [125] formulate a shared-link caching model with ‘private demands’ with the goal to design a two-phase private caching scheme with minimum load while preserving the privacy of the demands of each user with respect to other users.

Treating on-demand content as unicast traffic is highly inefficient, since it does not exploit the huge redundancy inherently contained in the users’ requests, which concentrate on a relatively small set of very popular files, especially in video-server services where the library of popular movies can be controlled by the service provider, and can be updated at a relatively slow pace (e.g. the library is refreshed every day/week/month). Such redundant requests arrive to the server in an asynchronous way, such that the probability that many users wish to stream the same file at the same time is basically zero. Coded caching techniques have the ability of turning the unicast traffic (on-demand streaming) into a coded multicast traffic, such that again the scalability of broadcasting a common message is recovered and full throughput scalability is achieved.

Beyond these very compelling theoretical results, a significant knowledge gap must be filled to make these ideal of practical value. Therefore, a significant research effort needs be made e.g. in the following areas:

* Coding (e.g., combining edge caching with modern multiuser MIMO physical layer schemes);
* Protocol architectures (e.g., combining edge caching with schemes for video quality adaptation such as Dynamic Adaptive Streaming over HTTP (DASH) [126]);
* AI/ML based content popularity estimation and prediction, to efficiently update the cached content [127].

# Optical Networks

…