(one6G)

Taking communications
to the next level

# 6G TECHNOLOGY OVERVIEW

WHITE PAPER

Fourth Edition - September 2024

one6g.org

# Executive Summary

6G is supposed to address the demands for consumption of mobile networking services in 2030 and beyond. These are characterized by a variety of diverse, often conflicting requirements, from technical ones such as extremely high data rates, unprecedented scale of communicating devices, high coverage, low communicating latency, flexibility of extension, etc., to non-technical ones such as enabling sustainable growth of the society as a whole, e.g., through energy efficiency of deployed networks. On the one hand, 6G is expected to fulfil all these individual requirements, extending thus the limits set by the previous generations of mobile networks (e.g., ten times lower latencies, or a hundred times higher data rates than in 5G). On the other hand, 6G should also enable use cases characterized by combinations of these requirements never seen before (e.g., both extremely high data rates and extremely low communication latency).

In this white paper, we give an overview of the key enabling technologies that constitute the pillars for the evolution towards 6G. They include: Terahertz frequencies (Section 1), 6G radio access (Section 2), next-generation MIMO (Section 3), integrated sensing and communication (Section 4), non-terrestrial networks (Section 5), multimodal sensing, computing, communication and control for 6G remote operation (Section 6), distributed and federated artificial intelligence (Section 7), intelligent user plane (Section 8), flexible programmable infrastructures (Section 9), and sustainability (Section 10). For each enabling technology, we first give the background on how and why the technology is relevant to 6G, backed up by a number of relevant use cases. After that, we describe the technology in detail, outline the key problems and difficulties, and give a comprehensive overview of the state of the art in that technology.

6G is, however, not limited to these ten technologies. They merely present our current understanding of the technological environment in which 6G is being born. Future versions of this white paper may include other relevant technologies too, as well as discuss how these technologies can be glued together in a coherent system.

# Table of contents

# 1. THz Frequencies

The digital evolution of our society requires communication services to be constantly improved. By means of enhanced multimedia services, 6G is expected to merge digital and physical worlds across all dimensions, providing users with a holographic, haptic, and multi-sense experience. According to the International Telecommunication Union (ITU), these applications will emerge during the next decade and will be characterized by tight requirements in terms of communication [1]. Additionally, some applications will require functionalities that are not currently provided by cellular systems, such as accurate sensing, mapping, and localization. Holographic telepresence represents an exemplary use case in this regard. Indeed, the transmission of raw 3D holograms requires more than 4 Tbps [2], while the capability to sense the environment will allow the network to predict users' movement without any explicit feedback, and enable an immersive remote experience. Similarly, high data rate and low latency communications required for factory automation will benefit from Tbps transmissions (Figure 1).

*Figure 1: Tbps transmissions for factory automation.*

To fill this gap, THz communications have been identified as one promising candidate for the physical layer in 6G, having the potential to enable data rates of Tbps, as well as providing sensing, mapping, and localization services [3]. At the World Radio Communication Conference 2019 (WRC-2019), ITU has identified 137 GHz of spectrum between 275 and 450 GHz which can be used for THz communications [4]. Together with the already allocated spectrum between 252 and 275 GHz, a total of 160 GHz is now available in the sub-THz range. In 2017, the IEEE 802 working group has completed the first wireless standard for carrier frequencies around 300 GHz (IEEE Std 802.15.3d-2017 [5][39]).

## 1.1 Use cases for THz communications

Two categories of use cases for THz communications are mentioned in [40], which are covered by [5], and applications where **at least one end of the link is mobile**. In addition, a third category covering use cases for **joint consideration of communication, localization and sensing** is mentioned in [18]. In the following, these three use cases are briefly described.

### 1.1.1 Fixed Point-to-Point Applications

THz communications have the potential to enable extremely high data rates, thanks to the large amount of radio resources available in these bands. However, signals at these frequencies are subject to severe propagation conditions, including high spreading loss and molecular absorption effect, which limit the communication range. To mitigate these phenomena, THz systems make use of multiple antennas and beamforming techniques that focus the transmit power into narrow beams. While extending the communication range, beamforming requires transmitters and receivers to align their beams before the actual communication can take place. This operation is challenging, especially when nodes are moving. For this reason, so far THz communications have been considered only for point-to-point applications. In particular, IEEE 802.15 Std 15.3d-2017 [5][39] defines four application scenarios, which have a strong demand for ultra-high data-rate transmissions. They are characterized by fixed point-to-point links where the location of the antennas is known, making device discovery and beam alignment obsolete. The four application scenarios are intra-device communication, close proximity communication, wireless links in data centres, and wireless backhaul/fronthaul links [6]:

- Intra-device communication [7] is targeting a wireless communication link within a device like a computer, camera, or video projector, making the use of cables inside devices obsolete. In this context, THz communications have the potential to make devices cheaper, lighter, more compact and efficient. Also, the absence of wired connections improves their reconfigurability and repair ability, since components can be easily replaced.

- Close proximity point-to-point communication [8][9], like kiosk downloading, is targeting wireless exchanges of large amount of data between two electronic devices such as smartphones, tablets, or hard disks. This use case enables the realization of Wireless Personal Area Networks (WPANs), where personal devices, such as smartphones, PCs, and monitors, can automatically interconnect without user intervention.

- Complementary wireless links in data centres [10][11][12] enable a faster reconfiguration of data centres, avoiding the deployment of large amount of fibre links and reducing the installation costs.

- Wireless backhaul and front haul links [13][14] in cellular networks enable the wireless connection of backhaul or front haul links to base stations, where fibre links are not available or too expensive.

### 1.1.2 Use cases where at least one end of the link is mobile

Use cases where at least one end of the link is mobile require algorithms for device discovery during link establishment [15], beamforming [16] and beam tracking [17]. Indeed, through efficient beamforming schemes, possibly assisted by lower frequency bands, 6G will enable seamless THz communications also in presence of user mobility. Such applications include extensions of WLAN-type (Wireless Local Area Network) applications [18], for example providing users in conference rooms or hotspots in public areas with ultra-high data rates. This functionality enables the intense use of virtual or augmented reality applications, e.g., in indoor environments and production lines. Also, future applications for in-flight [19] or in-train entertainment [20]. for a large number of users require ultra-high aggregated data rates. The latter requires backhauling for the aggregated data rate covering users in the moving vehicles [20] The use of ultra-high data rate applications in the context of vehicle-to-X communications requires capabilities enabling the exchange of these data between cars, or between cars and the infrastructure [21][22]. Space communications such as space/ground, inter-satellite and deep-space high speed data link can also benefit from THz bands. In comparison with free-space optical (FSO) communication, the THz link is less affected by atmospheric attenuation/scintillation and less limited by power and size [23][24]

### 1.1.3 Integrated Sensing and Communication (ISAC) at THz frequencies

6G is envisioned to exploit the specific propagation characteristics of signals in the THz spectrum to realize cellular networks with Integrated Sensing and Communication (ISAC) [25]. This new feature can improve the support to many envisioned use cases, providing situational awareness and context information. In particular, THz bands enable **sensing and imaging services**, such as radio astronomy and earth remote sensing [26], vehicle radars [27][28], chemical analysis [29], explosives detection [30], and moisture content analysis [37]. Moreover, THz signals can be used for wireless gas sensing, electronic smelling and pollution monitoring [18][31]. For example, in [31] the authors demonstrate the feasibility of using a THz communication link to infer the concentration of certain greenhouse gases. The same approach can be applied to monitor the presence of chemical hazards in critical places, such as factories or laboratories.

THz signals can also be used for medical imaging and material sensing, i.e., identify the shape and material composition of a certain object based on its spectral fingerprint. Therefore, the THz technology offers support to **e-health applications**, such as non-invasive tissue analysis [26][32] and measurement of glucose concentration [33], or **surveillance applications**, such as crowd monitoring and street surveillance [34].

Moreover, THz signals are strongly reflected by metal surfaces. This property can be exploited for **security** purposes, for example to detect the presence of weapons and in critical environments [18][26], such as in airports.

Finally, the directional nature of THz links makes them suitable to support accurate **localization and mapping services** [18], thus enabling new applications and use cases, such as high-resolution 3D mapping, massive twinning, immersive holographic telepresence, and interactive and cooperative robotics [25][35]. For example, the transmission of real-time, high-fidelity holograms may require accurate localization services, in such a way to closely model users' movements.

Thanks to the unique features of THz signals, it will be possible to realize radio access networks with ubiquitous radio sensing and unprecedented communication capabilities. This approach will promote a more efficient usage of the spectrum, enable new use cases, and avoid the need of using two separate systems, thus reducing costs [34].

The ISAC use cases discussed above are focused on those that are suitable for THz frequency bands. For a broader set of use cases also applicable to other bands, we refer the reader to Section 4.

*Table 1: THz use cases*

| Use case group | Use cases | Reference |
|---|---|---|
| Fixed Point-to-Point Applications | Intra-device communication | [7] |
| | Close proximity point-to-point communication | [8][9] |
| | Wireless links in data centres | [10][11][12] |
| | Wireless backhaul and fronthaul links | [13][14] |
| Use cases where at least one end of the link is mobile | Ultra-high data rate local area networks | [18] |
| | Virtual and augmented reality | [18] |
| | In-flight / in-train entertainment | [19][20] |
| | Backhauling for moving vehicles | [20] |

| Use case group | Use cases | Reference |
|---|---|---|
| | Vehicle-to-X communications | [21][22] |
| | Space communication | [23][24] |
| | Factory automation | [36] |
| Integrated sensing and communications | **Sensing and imaging services** | |
| | Radio astronomy and earth remote sensing | [26] |
| | Vehicle radars | [27][28] |
| | Chemical analysis | [29] |
| | Moisture content analysis | [37] |
| | Wireless gas sensing | [18] |
| | Electronic smelling | [18] |
| | Pollution and greenhouse gases monitoring | [31] |
| | **e-health applications** | |
| | Non-invasive tissue analysis | [26] [32] |
| | Medical imaging | [38] |
| | Measurement of glucose concentration | [33] |
| | **Surveillance and security applications** | |
| | Crowd monitoring | [34] |
| | Street surveillance | [34] |
| | Detect the presence of hidden objects and weapons | [18][26] |
| | Explosive detection | [30] |
| | **Accurate localization and mapping services** | |
| | Massive twinning | [25] |
| | Immersive telepresence | [25] |
| | Interactive and cooperative robotics | [25] |
| | High-resolution 3D mapping | [35] |

## 1.2 Characterization of THz wireless channels

The characterization of wireless channels is of paramount importance for the design of every wireless system. This is particularly true for THz bands, since the harsh propagation conditions experienced at these frequencies may have a strong impact on the system performance if not properly considered. In the following, we provide an overview of measurement and modelling approaches for the characterization of THz wireless channels.

### 1.2.1 THz channel measurements techniques

The measurement of wireless channels at THz frequencies poses significant challenges which makes the design of measurement systems more complex compared to lower frequency bands. For example, given the target bandwidth that is envisioned for THz systems, there is the need to perform measurements over large frequency bands, possibly exceeding 20 GHz. Also, proper characterization of the Doppler effect at these frequencies requires very high measurement rates [54] So far, three main techniques have been used for the characterization of THz channels: time-domain spectroscopy, vector network analysis, and broadband channel sounding [55].

Time-domain spectroscopy is a popular method that has been widely used for the determination of material properties in THz bands. This technique makes use of laser pulses with short time duration that are transmitted towards a sample of the material under investigation. A detector analyses amplitude and phase difference resulting after the interaction of the laser pulses with the sample, which can be exploited to derive the channel characteristics. This method was used to investigate the effect of atmospheric gases on THz signals, and to characterize the reflective properties of different materials.

Vector network analysis is another popular approach, which exploits analysers able to determine the frequency-dependent and complex-valued scattering parameters in order to derive the channel impulse response. This technique is suitable to measure static or slowly-varying channels over relatively short distances, given that the measurement endpoints need to be connected by wire.

Finally, broadband channel sounding architectures have also been developed. These systems are composed of a transmitter, which transmits a periodic pseudo-random binary sequence, and a receiver, which receives the sequence and performs a correlation to extract the channel impulse response. With this method, it is possible to perform instantaneous measurements of the full system bandwidth, and to measure the propagation effects even when the wireless channel varies over time.

The following table contains a list of THz channel measurements available in the literature.

*Table 2: THz channel measurements*

| Reference | Frequency | Method | Use Case |
|---|---|---|---|
| [41] | 260-400 GHz | VNA | Intra-device communication |
| [42] | 60 and 300 GHz | VNA | Intra-device communication |
| [43] | 220 – 340 GHz | VNA | Close proximity point-to-point communication |
| [44] | 300 GHz | VNA | Wireless links in data centres |
| [45] | 300 GHz | Broadband channel sounding | Wireless links in data centres |
| [46] | 300 GHz | Broadband channel sounding | In-flight / in-train entertainment |
| [56] | 300 GHz | Broadband channel sounding | Vehicle-to-X communications |
| [57] | 142 GHz | Broadband channel sounding | Factory automation |

### 1.2.2 Modelling approaches for THz channels

THz channel models can be divided into three main categories: deterministic, stochastic, and hybrid.

Deterministic models make use of a 3D representation of the surrounding environment and apply ray tracing or ray launching techniques to model multipath propagation. This approach enables the accurate modelling of wireless channels, but requires in-depth knowledge of the propagation environment and produces site-specific results. Deterministic THz channel models have been developed for various scenarios, including indoor and urban environments [47].

On the other hand, stochastic models are obtained by performing channel measurements and deriving a mathematical representation which ensembles the statistics of real channels. As such, stochastic models do not require a detailed knowledge of the propagation environment and therefore represent a valuable tool for system design. Several stochastic channel models targeting different use cases have already been proposed. For example, [48] proposed a cluster-based stochastic model derived from measurements in data centres, [49][50][51][52] introduced models for indoor environments, and [46] dealt with the train-to-infrastructure scenario.

Finally, hybrid models are considered as a middle ground, as they include both deterministic and stochastic components. An example of this approach is described in [53], where a hybrid ray-tracing-statistical model for THz communications is proposed.

The following table provides an overview of THz channel models available in the literature.

*Table 3: THz channel models*

| Reference | Scenario | Notes |
|---|---|---|
| [48] | Data centre | Cluster-based model derived from measurements |
| [49] | Laboratory, Conference room, Office | Cluster-based model derived from measurements |
| [46] | T2I inside station | Quadriga-based channel model with custom parameters derived from measurements |
| [9] | Kiosk downloading | Ray-based channel model derived from measurements |
| [51] | Indoor office | Ray-based channel model |
| [50] | THz indoor communications in a rectangular room | Geometric-statistical channel model for system-level simulation |
| [52] | Small indoor scenario | Abstract scattering model in the AoA/AoD/ToA domain for THz propagation simulations |
| [58] | Indoor | Extension of 3GPP 38901 for 100-300 GHz derived from ray-tracing simulations |
| [59] | Nanonetworks | Wideband multiple scattering channel model for THz frequencies |
| [60] | Indoor office | Spatial statistical channel model for an indoor office building up to 150 GHz |
| [61] | Different scenarios whose transmission distances range from tens of meters | Three-dimensional space-time-frequency non-stationary geometry-based stochastic model |

| Reference | Scenario | Notes |
|---|---|---|
| | to a few centimetres | |
| [62] | Intrabody Nanoscale | Novel channel model for intrabody communication in iWNSNs in the THz band |
| [63] | Scenarios with mobility | Three-dimensional space-time-frequency non-stationary massive multiple-input multiple-output channel model |
| [64] | Urban micro | Spatial statistical MIMO channel model for urban microcells at 142 GHz |

### 1.2.3 Modelling challenges and new approaches

Although many channel models targeting THz frequency bands have already been proposed, research on this topic is still in its infancy. Indeed, several open challenges prevent the characterization of THz propagation in many different scenarios and use cases, therefore new work has to be carried out.

The first challenge is related to the development of adequate measurement systems able to operate at high carrier frequencies and over large bandwidths. In this regard, the short coherence time of THz channels requires fast measurement speeds, while the high path loss experienced at these frequencies requires high dynamic range and sensitivity.

Moreover, next-generation wireless systems are expected to make use of ultra-massive MIMO antennas and reflective intelligent surfaces. The introduction of these new elements triggers new propagation phenomena that have to be taken into consideration, such as mutual coupling and near-field effects.

Finally, the ISAC paradigm requires new channel modelling methodologies, as communication and sensing operations need to be jointly considered.

### 1.2.4 Machine learning tools for channel modelling

Machine learning (ML) techniques could assist in more efficient channel modelling for THz bands. Three directions are highlighted below: scenario identification, multi-path components clustering and channel predictions as indicated in [69].

#### 1.2.4.1 Scenario Identification

As higher operating frequencies of wireless communication provide the possibility of large bandwidth and hence high data rates for communication channels, path loss becomes a limiting factor. To combat path loss, directional communication links are adopted together with the use of a large number of antennas. Moreover, the electromagnetic propagation on THz bands is more susceptible to dispersion in time and frequency domains, which leads to time-varying channels and Doppler spreads. The aforementioned channel peculiarities, which generates unusually large amount of measurement data and are susceptible to fast time variations, promote the use of ML inspired algorithms to detect the communication scenario, i.e. whether it is a LoS or a NLoS case.

Scenario identification in the context of LoS/NLoS channels is essential to applications involving localization tasks and to channel modelling in general. In particular, LoS channels are fundamental to the realization of THz communication, and knowing the presence of a LoS link might impact the feasibility of communication over THz frequencies. For instance, in the absence of a LoS link, a

system designer might resort to multi-hop communications through RISs or simply through relays with different mediating protocols.

#### 1.2.4.2 Channel Prediction

Machine learning, especially deep learning, has proved to be a powerful tool for modelling non-linearities of all kinds. In the context of channel modelling, the relation between desired channel model and the known/collected information also exhibits such non-linearity. Therefore, there has been plenty of work on using ML model, specifically neural networks, to construct mapping between collected measurements and the target channel model.

Given a certain area of communication setup, it is beneficial to have knowledge of the coverage of communication resources so as to adjust the design or improve the communication system. However, to get a precise description of the channel characteristics, such as path loss in a large area, the common approach of ray tracing entails high computational costs. To this end, combined with the recent advancement in computer vision, researchers have been working on using mature computer vision techniques to directly generate maps of radio resources given the 2D image description of the scenes as input.

In [70], satellite photos of a suburban area are pre-processed to explicitly represent the locations of buildings and other objects, then fed as input to generate the receiver signal strength at a single location by the convolutional neural network (CNN). Specifically, the receiving antennas' information is also encoded in the structure of the CNN. Levie et al. [71] adopted a similar approach, but with a complete radio map as output and simulated data for training and evaluation. Ates et al. [72] firstly categorize the path loss parameters and then applied CNN to classify the input scenarios represented by satellite images. In [73], CNN is applied to generate high-resolution radio maps on multiple metrics, based on the low-resolution images generated by fast but low-granularity simulation.

In THz communication, where the channel exhibits high directivity and suffers from blockage, the radio map generation techniques mentioned above are expected to identify direction and blockage more quickly than iterating on all possible rays as in ray tracing methods. Furthermore, recent advancements in computer vision of 3D point clouds also provide the possibility of generating "radio space" given the point clouds of a scene as input. For high frequency communication such as THz where the range of communication could be relatively short and heavily relying on LoS, using ML techniques in 3D vision might facilitate the generation of channel models in 3D scenarios.

#### 1.2.4.3 Clustering of multipath components

Many channel models assume that wireless propagation happens through a finite number of multipath components, each representing a plane wave travelling along a different path. Each multipath component is characterized by its complex amplitude, delay, direction of arrival and departure. Typically, multipath components that exhibit similar characteristics are grouped into a cluster. In this context, different machine learning techniques have been exploited to identify clusters in an automatic manner. In [74], authors presented the KPowerMeans, a variant of the popular K-means algorithm that accounts for the power of multipath components to compute clusters centroids. The same algorithm was used in [75] and [76] for the identification of multipath clusters at mmWave and THz frequencies. Li et al. [77] collected channel measurements at THz frequencies and applied the DBSCAN algorithm [78] to identify the multipath clusters. Chen et al. [79] exploited THz channel measurements and ray tracing simulations for clustering and matching of multipath components. First, the multipath components observed in real measurements are clustered using the DBSCAN algorithm. Then, the identified clusters are matched with those observed in a ray tracing simulator based on the multipath component distance (MCD) metric. In [80], authors proposed a clustering algorithm that identifies independent clusters based on a kernel density measure. In [81], a novel clustering approach based on Fuzzy-c-means is presented. In [82], authors proposed a new clustering algorithm based on the region competition algorithm [83], an optimization technique originally developed for image segmentation, and the kurtosis measure. In [84], clustering is treated as a sparsity-based optimization problem that exploits the physical

property that the power of multipath components decreases exponentially with respect to the delay.

### 1.2.5 Characterization of dielectric properties of materials

Given the short propagation range, the majority of use cases that have been identified for THz communications are indoor. In these environments, wireless signals interact with multiple objects causing reflection, diffraction, and transmission phenomena. The behaviour of these phenomena is influenced by the dielectric properties of the object, which depend on physical features such as shape, material, roughness, etc. For the deterministic modelling of wireless channels (e.g., by means of ray tracing) characterizing the dielectric properties of the interacting objects is important to achieve accurate results. The ITU-R recommendation P.2040-3 [85] provides guidelines and parameters for modelling the dielectric behaviour of common building materials. However, most of the parameters available in the document are valid up to 100 GHz, thus requiring further work for extending the model up to THz frequencies.

The methodologies that are commonly adopted for measuring the dielectric properties of materials can be divided into three main groups: (i) waveguides, (ii) resonators, and (iii) free space methods.

#### 1.2.5.1 Waveguides

The sample under test is inserted in a section of waveguide and characterized by measuring the transmission and reflection coefficients of the EM wave propagating along the line. This method provides high accuracy. However, it requires small tolerances for the preparation of the sample and its insertion into the waveguide. For frequencies beyond 100 GHz, this method becomes impractical because the size of the waveguide becomes extremely small (in the order of micrometres).

#### 1.2.5.2 Resonators

This method exploits the perturbation theory of resonant cavities, i.e., structures exhibiting a resonant behaviour at specific frequencies. When introducing an object inside the cavity, its resonant modes are perturbed in a way which depends on the dielectric properties of the object. After measuring the perturbation caused by the sample under test, it is possible to determine its permeability and permittivity by applying the cavity perturbation theory. Examples of structures that are commonly used for this purpose include split-cavity resonators, Fabri-Perot open resonators [86], and resonators based on Whispering Gallery Modes [87]. Despite providing accurate results, these methods are typically narrowband and not applicable to material with high losses.

#### 1.2.5.3 Free-space methods

The sample under test is placed between transmit and receive antennas. The transmit antenna generates a wireless signal which is reflected or refracted through the sample and measured by the receive antenna. By comparing the signal response with and without the sample, it is possible to determine the transmission and/or reflection coefficients. Then, the dielectric properties of the sample can be derived by applying the Nicolson-Ross-Weir algorithm [88]. This method does not require any contact with the sample, thus representing a suitable solution for performing non-destructive tests. Different techniques are available for conducting free-space measurements, the most popular being vector network analysis and time domain spectroscopy. With the latter recently numerous typical building materials have been measured for the purpose of using it in simulations for THz communications [89].

## 1.3 THz device technology

While front ends at E-band (71 - 76 GHz, 81 - 86 GHz) are already available in the market, the technology for links above 100 GHz is still at the prototype level. Two major obstacles have to be

resolved: (i) harsh propagation conditions, and (ii) equipment cost. Tens of dB of attenuation higher than microwave frequencies (e.g., due to rain, humidity and gases) for the same range pose serious technology constraints to satisfy the link budget. The low transmission power of solid-state power amplifiers can be partially compensated by a high antenna gain. However, at the increase of the gain (above 40 - 45 dBi) there are problems of sway in case of wind, large footprint and cost due to the required high fabrication accuracy.

On the other hand, THz high attenuation is also advantageous because it permits an effective spatial division, frequency reuse and low interference expanding the potential of THz technology.

In recent years, numerous THz wireless front end were presented up to 400 GHz with different technologies and performance [65][35]. Substantial advancements are reported in the development of chipset based on different processes such as CMOS, InP, GaAs, Si Ge and GaN, mostly for low power electronics to support multi-Gb/s transmission. The data rate exceeds 10s Gb/s for most of the prototypes. However, the range of all these systems is limited to tens of meters, with the need for very high gain antennas, due to lack of the required high transmission power for satisfying long link range. Gallium Nitride (GaN) is the most promising semiconductor process for high power, but presently is limited to about 100 GHz.

The short wavelength (e.g. 3 mm at 100 GHz, 1 mm at 300 GHz) makes fabrication and assembly difficult due to the small dimensions of parts that also require tight tolerances. With the increase of frequency, fabrication technologies need to be improved and be more affordable. The high manufacturing cost of THz equipment is a critical factor for its wide deployment.

At the same time, the short wavelength permits low size antennas and components for high integration and low footprint, enabling an easier installation and deployment with high density in urban environment, reducing the cost of site renting.

Point-to-point and point-to-multipoint distribution at THz frequency would permit 100s Gb/s/km$^2$ area capacity, needed for supporting 6G concepts. Two European projects, TWEETHER and ULTRAWAVE [66] explored the use of W-band (92 – 95 GHz), and D-band (141 – 148.5 GHz) for Point to multi Point distribution [67] and G-band (275 – 305 GHz) for point-to-point transport. The key target is a low cost per bit, competitiveness with the fibre and long range. The novelty of these projects is the introduction of a new generation of travelling wave tubes being able to produce more than one order of magnitude transmission power than a solid-state amplifier [68]. The high transmission power available (e.g., 10 W at D-band) permits long range links close to 1 km both in point to multipoint and point to point.

## 1.4 References

[1] ITU-T Focus Group Technologies for Network 2030 (FG NET-2030), “Network 2030 - A Blueprint of Technology, Applications and Market Drivers Towards the Year 2030 and Beyond,” White Paper, May 2019.

[2] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan and M. Zorzi, “Toward 6G Networks: Use Cases and Technologies,” in IEEE Communications Magazine, vol. 58, no. 3, pp. 55-61, March 2020.

[3] M. Lattva-aho, K. Lappänen, “Key drivers and research challenges for 6G ubiquitous wireless intelligence,” University of Oulu, 2019.

[4] T. Kürner and A. Hirata, “On the Impact of the Results of WRC 2019 on THz Communications,” 2020 Third International Workshop on Mobile Terahertz Systems (IWMTS), Essen, Germany, 2020.

[5] IEEE Standard for High Data Rate Wireless Multi-Media Networks--Amendment 2: 100 Gb/s Wireless Switched Point-to-Point Physical Layer, in IEEE Std 802.15.3d-2017 (Amendment to IEEE Std 802.15.3-2016 as amended by IEEE Std 802.15.3e-2017) , pp.1-55, Oct. 2018.

[6] B. Peng, K. Guan, A. Kuter, S. Rey, M. Patzold and T. Kuerner, "Channel Modeling and System Concepts for Future Terahertz Communications: Getting Ready for Advances Beyond 5G," in IEEE Vehicular Technology Magazine, vol. 15, no. 2, pp. 136-143, June 2020.

[7] Kürner T, Fricke A, Rey S, et al, "Measurements and modeling of basic propagation characteristics for intra-device communications at 60 GHz and 300 GHz'" Journal of Infrared, Millimeter, and Terahertz Waves, 2015.

[8] Kim S, Zajic, "Statistical modeling and simulation of short-range device-to-device communication channels at sub-THz frequencies," IEEE Transactions on Wireless Communications, 2016.

[9] He D., Guan K., Ai B. et al, "Stochastic channel modeling for kiosk applications in the terahertz Band," IEEE Transactions on Terahertz Science and Technology, 2017.

[10] Hamza A. S., Deogun J. S., Alexander D. R., "Wireless Communication in Data Centers: A Survey," IEEE Communications Surveys & Tutorials, vol. 18, no. 3, 2016.

[11] Davy A. S., Pessoa L., Renaud C., et al, "Building an end user focused THz based ultra high bandwidth wireless access network: The TERAPOD approach," Proceedings of the 9th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), Munich, 2017.

[12] Eckhardt J. M., Doeker T., Rey S., Kürner T., "Measurements in a Real Data Center at 300 GHz and Recent Results," Proceedings of 13th European Conference on Antennas and Propagation (EuCAP 2019), Krakow, Poland, 2019.

[13] C. Castro, R. Elschner, T. Merkle, C. Schubert, and R. Freund, "Experimental demonstrations of high-capacity THz-wireless transmission systems for Beyond 5G," IEEE Communications Magazine, vol. 58, no. 11, pp. 41-47, Nov. 2020.

[14] I. Dan, G. Ducournau, S. Hisatake, P. Szriftgiser, R. Braun and I. Kallfass, "A Terahertz Wireless Communication Link Using a Superheterodyne Approach," in IEEE Transactions on Terahertz Science and Technology, vol. 10, no. 1, pp. 32-43, Jan. 2020.

[15] Q. Xia and J. M. Jornet, "Expedited Neighbor Discovery in Directional Terahertz Communication Networks Enhanced by Antenna Side-Lobe Information," in IEEE Transactions on Vehicular Technology, vol. 68, no. 8, Aug. 2019.

[16] T. Merkle, A. Tessmann, M. Kuri, S. Wagner, A. Leuther, S. Rey, M. Zink, H.-P. Stulz, M. Riessle, I. Kallfass, T. Kürner, "Testbed for phased array communications from 275 to 325 GHz," 2017 IEEE Compound Semiconductor Integrated Circuit Symposium (CSICS), Miami, FL, 2017.

[17] B. Peng, Q., Jiao, and T. Kürner, "Angle of Arrival Estimation in Dynamic Indoor THz Channels with Bayesian Filter and Reinforcement Learning," Proc. 24th European Signal Processing Conference (EUSIPCO 2016), Budapest, Ungarn, September 2016.

[18] H. Sarieddeen, N. Saeed, T. Y. Al-Naffouri and M. -S. Alouini, "Next Generation Terahertz Communications: A Rendezvous of Sensing, Imaging, and Localization," in IEEE Communications Magazine, vol. 58, no. 5, pp. 69-75, May 2020.

[19] J. M. Eckhardt, T. Doeker and T. Kürner, "Indoor-to-Outdoor Path Loss Measurements in an Aircraft for Terahertz Communications," 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), Antwerp, Belgium, 2020.

[20] K. Guan, G. Li, T. Kürner, A. F. Molisch, B. Peng, R. He, H. Bing, J. Kim, Z. Zhong, "On Millimeter Wave and THz Mobile Radio Channel for Smart Rail Mobility," IEEE Transactions on Vehicular Technology, Vol. 66, No. 7, 2017

[21] J. M. Eckhardt, V. Petrov, D. Moltchanov, Y. Koucheryavy and T. Kürner, "Channel Measurements and Modeling for Low-Terahertz Band Vehicular Communications," in IEEE Journal on Selected Areas in Communications, vol. 39, no. 6, Jun. 2021.

[22] V. Petrov et al., “On Unified Vehicular Communications and Radar Sensing in Millimeter Wave and Low Terahertz Bands,” IEEE Wireless Communications, vol. 26, no. 3, Jun. 2019.

[23] Meltem Civa, Ozgur B. Akan, “Terahertz Wireless Communication in Space,“ ITU Journal on Future and Evolving Technologies, Volume 2, Issue 7, Oct. 2021

[24] Z. Chen et al., “A survey on terahertz communications,” in China Communications, vol. 16, no. 2, pp. 1-35, Feb. 2019.

[25] Wymeersch H, Shrestha D, De Lima CM, Yajnanarayana V, Richerzhagen B, Keskin MF et al. “Integration of Communication and Sensing in 6G: A Joint Industrial and Academic Perspective,” in 2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC 2021, 2021.

[26] P. de Maagt, “Terahertz technology for space and earth applications,” 2006 First European Conference on Antennas and Propagation, Nice, France, 2006.

[27] Y. Xiao, F. Norouzian, E. G. Hoare, E. Marchetti, M. Gashinova and M. Cherniakov, “Modeling and Experiment Verification of Transmissivity of Low-THz Radar Signal Through Vehicle Infrastructure,” in IEEE Sensors Journal, vol. 20, no. 15, Aug. 2020.

[28] D. Jasteh, M. Gashinova, E. G. Hoare, T. Y. Tran, N. Clarke and M. Cherniakov, “Low-THz imaging radar for outdoor applications,” 2015 16th International Radar Symposium (IRS), 2015.

[29] Fischer, B., Hoffmann, M., Helm, H., Modjesch, G., and Jepsen, P. U., “Chemical recognition in terahertz time-domain spectroscopy and imaging”, Semiconductor Science Technology, vol. 20, no. 7, 2005.

[30] Wang Gao, Xu Degang and Yao Jianquan, “Review of explosive detection using terahertz spectroscopy technique,” Proceedings of 2011 International Conference on Electronics and Optoelectronics, 2011.

[31] L. T. Wedage, B. Butler, S. Balasubramaniam, Y. Koucheryavy, and J. M. Jornet, “Climate Change Sensing through Terahertz Communications: A Disruptive Application of 6G Networks,” arXiv preprint, 2021, url: https://arxiv.org/abs/2110.03074

[32] N. Chopra, K. Yang, Q. H. Abbasi, K. A. Qaraqe, M. Philpott and A. Alomainy, “THz Time-Domain Spectroscopy of Human Skin Tissue for In-Body Nanonetworks,” in IEEE Transactions on Terahertz Science and Technology, vol. 6, no. 6, Nov. 2016.

[33] Torii T, Chiba H, Tanabe T, Oyama Y., “Measurements of glucose concentration in aqueous solutions using reflected THz radiation for applications to a novel sub-THz radiation non-invasive blood sugar measurement method,” DIGITAL HEALTH, Jan. 2017.

[34] A. Zhang, M. L. Rahman, X. Huang, Y. J. Guo, S. Chen and R. W. Heath, “Perceptive Mobile Networks: Cellular Networks with Radio Vision via Joint Communication and Radar Sensing,” in IEEE Vehicular Technology Magazine, vol. 16, no. 2, pp. 20-30, June 2021.

[35] Chaccour, C., Soorki, M.N., Saad, W., Bennis, M., Popovski, P., and Debbah, M., “Seven Defining Features of Terahertz (THz) Wireless Systems: A Fellowship of Communication and Sensing,” arXiv preprint, 2021, url: https://arxiv.org/abs/2102.07668.

[36] Gangakhedkar, Sandip, et al. “Use cases, requirements and challenges of 5G communication for industrial automation,” 2018 IEEE International Conference on Communications Workshops (ICC Workshops), 2018.

[37] Kashima, M., Tsuchikawa, S., and Inagaki, T, “Simultaneous detection of density, moisture content and fiber direction of wood by THz time-domain spectroscopy,” J Wood Sci 66, 2020.

[38] K. Humphreys et al., “Medical applications of terahertz imaging: a review of current technology and potential applications in biomedical engineering,” The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2004.

[39] V. Petrov, T. Kürner and I. Hosako, “IEEE 802.15.3d: First Standardization Efforts for Sub-Terahertz Band Communications toward 6G,” in IEEE Communications Magazine, vol. 58, no. 11, pp. 28-33, Nov. 2020.

[40] T. Kürner, D. Mittleman, T. Nagatsuma, “THz Communications - Paving the Way Towards Wireless Tbps,” Springer, 2022.

[41] N. Khalid and O. B. Akan, "Wideband THz communication channel measurements for 5G indoor wireless networks," 2016 IEEE International Conference on Communications (ICC), 2016, pp. 1-6, doi: 10.1109/ICC.2016.7511280.

[42] Kürner, T., Fricke, A., Rey, S. et al. Measurements and Modeling of Basic Propagation Characteristics for Intra-Device Communications at 60 GHz and 300 GHz. J Infrared Milli Terahz Waves 36, 144–158 (2015). https://doi.org/10.1007/s10762-014-0117-5

[43] Fricke, A., et al. (2016). Channel modelling document (CMD). IEEE 802.15 Plenary Meeting, Macau, 2016, DCN: 15-14-0310-19-003d.

[44] Cheng, C., & Zajic, A. (2020). Characterization of propagation phenomena relevant for ´ 300 GHz wireless data center links. IEEE Transactions on Antennas and Propagation, 68(2), 1074–1087.

[45] Eckhardt, J. M., Doeker, T., Rey, S., & Kürner, T. (2019). Measurements in a real data center at 300 GHz and recent results. In Proceedings of 13th European Conference on Antennas and Propagation (EuCAP 2019), Krakow.

[46] Guan, K., Peng, B., He, D., et al. (2019). Measurement, simulation, and characterization of train-to-infrastructure inside-station channel at the terahertz band. IEEE Transactions on Terahertz Science and Technology, 9(3), 291–306. https://doi.org/10.1109/TTHZ.2019. 2909975

[47] D. He, B. Ai, K. Guan, L. Wang, Z. Zhong and T. Kürner, "The Design and Applications of High-Performance Ray-Tracing Simulation Platform for 5G and Beyond Wireless Communications: A Tutorial," in IEEE Communications Surveys & Tutorials, vol. 21, no. 1, pp. 10-27, Firstquarter 2019, doi: 10.1109/COMST.2018.2865724.

[48] C. -L. Cheng, S. Sangodoyin and A. Zajić, "THz Cluster-Based Modeling and Propagation Characterization in a Data Center Environment," in IEEE Access, vol. 8, pp. 56544-56558, 2020, doi: 10.1109/ACCESS.2020.2981293.

[49] L. Pometcu and R. D’Errico, "An Indoor Channel Model for High Data-Rate Communications in D-Band," in IEEE Access, vol. 8, pp. 9420-9433, 2020, doi: 10.1109/ACCESS.2019.2960614.

[50] Choi, Y., Choi, JW. & Cioffi, J.M. A Geometric-Statistic Channel Model for THz Indoor Communications. J Infrared Milli Terahz Waves 34, 456–467 (2013). https://doi.org/10.1007/s10762-013-9975-5

[51] S. Priebe and T. Kurner, "Stochastic Modeling of THz Indoor Radio Channels," in IEEE Transactions on Wireless Communications, vol. 12, no. 9, pp. 4445-4455, September 2013, doi: 10.1109/TWC.2013.072313.121581.

[52] S. Priebe, M. Jacob and T. Kuerner, "AoA, AoD and ToA Characteristics of Scattered Multipath Clusters for THz Indoor Channel Modeling," 17th European Wireless 2011 - Sustainable Wireless Technologies, 2011, pp. 1-9.

[53] Y. Chen, Y. Li, C. Han, Z. Yu and G. Wang, "Channel Measurement and Ray-Tracing-Statistical Hybrid Modeling for Low-Terahertz Indoor Communications," in IEEE Transactions on Wireless Communications, vol. 20, no. 12, pp. 8163-8176, Dec. 2021, doi: 10.1109/TWC.2021.3090781.

[54] C. Han et al., "Terahertz Wireless Channels: A Holistic Survey on Measurement, Modeling, and Analysis," in IEEE Communications Surveys & Tutorials, vol. 24, no. 3, pp. 1670-1707, thirdquarter 2022, doi: 10.1109/COMST.2022.3182539.

[55] Kürner, Thomas, Daniel Mittleman, and Tadao Nagatsuma, eds. THz Communications: Paving the Way Towards Wireless Tbps. Springer, 2022.

[56] V. Petrov, J. M. Eckhardt, D. Moltchanov, Y. Koucheryavy and T. Kurner, "Measurements of Reflection and Penetration Losses in Low Terahertz Band Vehicular Communications," 2020 14th European Conference on Antennas and Propagation (EuCAP), 2020, pp. 1-5, doi: 10.23919/EuCAP48036.2020.9135389.

[57] S. Ju, Y. Xing, O. Kanhere and T. S. Rappaport, "Sub-Terahertz Channel Measurements and Characterization in a Factory Building," ICC 2022 - IEEE International Conference on Communications, 2022, pp. 2882-2887, doi: 10.1109/ICC45855.2022.9838910.

[58] Z. Hossain, Q. C. Li, D. Ying, G. Wu and C. Xiong, "THz Channel Model for 6G Communications," 2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), 2021, pp. 1-7, doi: 10.1109/PIMRC50174.2021.9569257.

[59] J. Kokkoniemi, J. Lehtomäki, K. Umebayashi and M. Juntti, "Frequency and Time Domain Channel Models for Nanonetworks in Terahertz Band," in IEEE Transactions on Antennas and Propagation, vol. 63, no. 2, pp. 678-691, Feb. 2015, doi: 10.1109/TAP.2014.2373371.

[60] S. Ju, Y. Xing, O. Kanhere and T. S. Rappaport, "Millimeter Wave and Sub-Terahertz Spatial Statistical Channel Model for an Indoor Office Building," in IEEE Journal on Selected Areas in Communications, vol. 39, no. 6, pp. 1561-1575, June 2021, doi: 10.1109/JSAC.2021.3071844.

[61] J. Wang, C. -X. Wang, J. Huang, H. Wang and X. Gao, "A General 3D Space-Time-Frequency Non-Stationary THz Channel Model for 6G Ultra-Massive MIMO Wireless Communication Systems," in IEEE Journal on Selected Areas in Communications, vol. 39, no. 6, pp. 1576-1589, June 2021, doi: 10.1109/JSAC.2021.3071850.

[62] H. Elayan, R. M. Shubair, J. M. Jornet and P. Johari, "Terahertz Channel Model and Link Budget Analysis for Intrabody Nanoscale Communication," in IEEE Transactions on NanoBioscience, vol. 16, no. 6, pp. 491-503, Sept. 2017, doi: 10.1109/TNB.2017.2718967.

[63] J. Wang, C. -X. Wang, J. Huang and H. Wang, "A Novel 3D Space-Time-Frequency Non-Stationary Channel Model for 6G THz Indoor Communication Systems," 2020 IEEE Wireless Communications and Networking Conference (WCNC), 2020, pp. 1-7, doi: 10.1109/WCNC45663.2020.9120570.

[64] S. Ju and T. S. Rappaport, "Sub-Terahertz Spatial Statistical MIMO Channel Model for Urban Microcells at 142 GHz," 2021 IEEE Global Communications Conference (GLOBECOM), 2021, pp. 1-6, doi: 10.1109/GLOBECOM46510.2021.9685929.

[65] H. Tataria, M. Shafi, A. F. Molisch, M. Dohler, H. Sjöland and F. Tufvesson, "6G Wireless Systems: Vision, Requirements, Challenges, Insights, and Opportunities," in Proceedings of the IEEE, doi: 10.1109/JPROC.2021.3061701.

[66] C. Paoloni, V. Krozer, F. Magne, T. Le, R. Basu, J. M. Rao, R. Letizia, E. Limiti, M. Marilier, G. Ulisse, A. Ramirez, B. Vidal, and H. Yacob, “D-band point to multi-point deployment with g- band transport,” in 2020 European Conference on Networks and Communications (EuCNC), pp. 84–88, 2020.

[67] J. Shi, L. Lv, Q. Ni, H. Pervaiz, and C. Paoloni, “Modeling and analysis of point-to-multipoint millimeter wave backhaul networks,” IEEE Transactions on Wireless Communications, vol. 18, no. 1, pp. 268–285, 2019.

[68] R. Basu, J. M. Rao, T. Le, R. Letizia, and C. Paoloni, “Develop- ment of a d-band traveling wave tube for high data-rate wireless links,” IEEE Transactions on Electron Devices, vol. 68, no. 9, pp. 4675–4680, 2021.

[69] C.-X. Wang, J. Huang, H. Wang, X. Gao, X. You, and Y. Hao, “6G Wireless Channel Measurements and Models: Trends and Challenges,” IEEE Vehicular Technology Magazine (VTMag), vol. 15, no. 1, pp. 22–32, 3 2020.

[70] H. Cheng, S. Ma, and H. Lee, “CNN-Based mmWave Path Loss Modeling for Fixed Wireless Access in Suburban Scenarios,” IEEE Antennas and Wireless Propagation Letters, vol. 19, no. 10, pp. 1694–1698, 2020.

[71] R. Levie, C¸. Yapar, G. Kutyniok, and G. Caire, “RadioUNet: Fast Radio Map Estimation With Convolutional Neural Networks,” IEEE Transactions on Wireless Communications (TWC), vol. 20, no. 6, pp. 4001–4015, 2021.

[72] H. Ates, S. Hashir, T. Baykas, and B. Gunturk, “Path Loss Exponent and Shadowing Factor Prediction From Satellite Images Using Deep Learning,” IEEE Access, vol. 7, pp. 101366–101375, 2019.

[73] X. Wang, Z. Zhang, D. He, K. Guan, D. Liu, J. Dou, S. Mumtaz, and S. AlRubaye, “A Multi - Task Learning Model for Super Resolution of Wireless Channel Characteristics,” in IEEE Global Telecommunications Conference (GLOBECOM 2022). Rio de Janeiro, Brazil: IEEE, 12 2022, pp. 952–957.

[74] N. Czink, P. Cera, J. Salo, E. Bonek, J.-P. Nuutinen, and J. Ylitalo, “A Framework for Automatic Clustering of Parametric MIMO Channel Data Including Path Powers,” in 64th IEEE Vehicular Technology Conference (VTC 2006). Montreal, ´ Canada: IEEE, 9 2006, pp. 1–5.

[75] C. Gustafson, K. Haneda, S. Wyne, and F. Tufvesson, “On mm-Wave Multipath Clustering and Channel Modeling,” IEEE Transactions on Antennas and Propagation, vol. 62, no. 3, pp. 1445–1455, 3 2014.

[76] C.-L. Cheng, S. Sangodoyin, and A. Zajic, “THz Cluster-Based Modeling and Propagation Characterization in a Data Center Environment,” IEEE Access, vol. 8, pp. 56544–56558, 3 2020.

[77] Y. Li, Y. Wang, Y. Chen, Z. Yu, and C. Han, “Channel Measurement and Analysis in an Indoor Corridor Scenario at 300 GHz,” in IEEE International Conference on Communications (ICC 2022). Seoul, South Korea: IEEE, 5 2022.

[78] E. Martin, K. Hans-Peter, and X. Xiaowei, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” in Proceedings of the Second International Conference on Knowledge Discovery in Databases and Data Mining (KDD), Portland, OR, 8 1996, pp. 226–231.

[79] Y. Chen, Y. Li, C. Han, Z. Yu, and G. Wang, “Channel Measurement and Ray-Tracing-Statistical Hybrid Modeling for Low-Terahertz Indoor Communications,” IEEE Transactions on Wireless Communications (TWC), vol. 20, no. 12, pp. 8163–8176, 12 2021.

[80] R. He, Q. Li, B. Ai, Y. L.-A. Geng, A. F. Molisch, V. Kristem, Z. Zhong, and J. Yu, “A Kernel-Power-Density-Based Algorithm for Channel Multipath Components Clustering,” IEEE Transactions on Wireless Communications (TWC), vol. 16, no. 11, pp. 7138–7151, 11 2017.

[81] C. Schneider, M. Bauer, M. Narandzic, W. A. T. Kotterman, and R. S. Thomae, “Clustering of MIMO Channel Parameters - Performance Comparison,” in 69th IEEE Vehicular Technology Conference (VTC 2009-Spring). Barcelona, Spain: IEEE, 4 2009.

[82] C. Gentile, “Using the Kurtosis Measure to Identify Clusters in Wireless Channel Impulse Responses,” IEEE Transactions on Antennas and Propagation, vol. 61, no. 6, pp. 3392–3395, 6 2013.

[83] S. C. Zhu and A. Yuille, “Region competition: unifying snakes, region growing, and Bayes/MDL for multiband image segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18, no. 9, pp. 884–900, 1996.

[84] R. He, W. Chen, B. Ai, A. F. Molisch, W. Wang, Z. Zhong, J. Yu, and S. Sangodoyin, “On the Clustering of Radio Channel Impulse Responses Using Sparsity-Based Methods,” IEEE Transactions on Antennas and Propagation, vol. 64, no. 6, pp. 2465–2474, 6 2016.

[85] ITU-R, Recommendation P.2040-3, “Effects of building materials and structures on radiowave propagation above about 100 MHz,” International Telecommunication Union, 2023.

[86] B. Salski, A. Pacewicz, and P. Kopyt, “Measurement errors and uncertainties in the complex permittivity extraction with a fabry–perot open resonator,” IEEE Trans. on Microwave Theory and Techniques, 2023.

[87] Barannik, A., N. Cherpak, A. Kirichenko, Y. Prokopenko, S. Vitusevich, V. Yakovenko, Whispering gallery mode resonators in microwave physics and technologies, Int. J. Microw. Wirel. Technol.9 (2017) 781–796.

[88] Nicolson, A.M., G.F. Ross, Measurement of the Intrinsic Properties of Materials by Time-Domain Techniques, IEEE Trans. Instrum. Meas.19 (1970) 377–382. https://doi.org/10.1109/TIM.1970.4313932.

[89] F. Taleb, G. G. Hernandez-Cardoso, E. Castro-Camus and M. Koch, "Transmission, Reflection, and Scattering Characterization of Building Materials for Indoor THz Communications," in IEEE Transactions on Terahertz Science and Technology, vol. 13, no. 5, pp. 421-430, Sept. 2023, doi: 10.1109/TTHZ.2023.3281773.

# 2. 6G Radio Access (6GRA)

## 2.1 Introduction and Motivation

6G radio access is expected to provide extended support for the traditional 5G use cases: enhanced mobile broadband (eMBB), ultra-reliable low latency communications (URLLC), and massive machine-type communications (mMTC) – that is, increased throughput, lower latency and increased reliability, and greater scalability. However, the requirements 'space' of 6G applications will be much wider, including flow-level timing metrics, service availability, service continuity, and energy efficiency. Therefore, the scope of 6G radio access expands well beyond just extending the capabilities of 5G. Specifically, the 6G radio access design must be flexible and resource-efficient, being capable of adapting in real-time, both at the infrastructure and the user terminal side, while also fulfilling the application requirements. Furthermore, it is essential to consider advanced flow-level timing metrics such as Age of Information (AoI), Value of Information (VoI) and semantics of information, as well as goal-oriented metrics that go well beyond the traditional packet-level timing metric: latency and reliability. Doing so would allow the network to capture the real requirements of the applications and, hence, conduct optimal resource slicing, allocation, and user scheduling with maximum energy efficiency.

## 2.2 Reference use cases

### 2.2.1 Tactile Internet

Focused on having mixed traffic (broadband and URLLC-like) in the same service: High data rates are needed for video/audio feedback whereas the control commands and sensory (i.e., touch) feedback must be transmitted with ultra-low latency and high reliability (URLLC-like). The problem becomes complex as 1) this combination of traffic takes place in both uplink and downlink communication; 2) the source of broadband and URLLC traffic may be from different devices; and 3) the URLLC-like traffic may not be fully periodic, and its pattern may change depending on the status of the control loop. Therefore, allocating pre-defined resources to URLLC traffic may lead to over-provisioning and a more efficient and flexible allocation is needed. Augmented, virtual, and extreme reality (AR/VR/XR) applications are examples where URLLC and broadband traffic coexist. Specifically, large volumes of data (video) are transmitted in the downlink while short feedback and control messages must be transmitted in the uplink with high reliability, even though a few consecutive packet losses might be tolerated.

### 2.2.2 Remote control of cyber-physical systems

This use case focuses on flow-level KPIs rather than per-packet reliability. 3GPP recently defined the survival time as the maximum time a cyber-physical control system may continue its operation without receiving an anticipated message [1]. Several other metrics such as Age of Information (AoI) have also been defined for systems transmitting updates, where the newest update immediately replaces previous ones. Nevertheless, AoI still maintains the traditional view of data transmission being the end goal of the communication process and makes the strong assumption that all the generated data belongs to updates of a specific process. Instead, value of information (VoI), semantics of information, and goal-oriented communications (GOCom) rely on harnessing the fact that data is communicated with the objective of achieving a particular end goal set by the application [31][43]—in this case, the control of a cyber-physical system. Only by considering this latter aspect it would be possible to escape traditional resource efficiency goals including, for example, energy per bit and spectral efficiency, and achieve unprecedented levels of resource efficiency by transmitting less data but equal amounts of information.

### 2.2.3 Industry 4.0

This use case focuses on scenarios that include IoT devices that transmit large amounts of data such as video surveillance cameras but also other devices that transmit small amounts of data in a periodic, quasi-periodic, or sporadic fashion. Furthermore, the environment must respond reliably and with ultra-low latency to emergency situations, such as power outages, increase in temperature, or malfunction of individual elements in, for example, an assembly line. Hence, the network must support different mechanisms to increase the efficiency of communication in these use cases. An essential question to answer to increase the efficiency of communication is how frequently the devices need to sample and transmit data. Therefore, considering the semantics of information is of utmost importance to reduce the amount of generated and transmitted data without affecting the amount of conveyed information [31], but also to schedule the users and allocate resources optimally.

### 2.2.4 Smart metering

This use case focuses on exploring and exploiting the influence of space and time in the behaviour of the sensor nodes. These influence the spatial-temporal correlation of the data, which must be considered to maximize energy efficiency. That is, in smart metering applications, the spatial distribution determines the potential impact of the interference from a user to its neighbours and, hence, sets the basis for the competition for resources. In addition, it also determines the nature of the collected data, as space induces correlation among the sensor readings, which can be learned and exploited to design deployment- and application-specific access mechanisms.

### 2.2.5 Smart city

This use case focuses on the coexistence of diverse services with heterogeneous requirements and traffic characteristics, including a combination of previously listed use cases, at a large scale. Coexisting services include, for example, VR/AR/XR, V2X communications, robots and cobots, smart metering, and e-health. This leads to potentially massive access scenarios with time-varying traffic demands that would benefit from dynamic scaling of the network resources and from mobile network elements such as drones, high-altitude platforms, and/or satellites to offload the terrestrial infrastructure.

## 2.3 SotA on the main topics towards 6G radio access

### 2.3.1 5G standardization and limitations in the radio access network (RAN)

Since its introduction in the 3rd Generation Partnership Project (3GPP) standardization in Release 15, the 5G standards have evolved by including mechanisms to reduce the user- and control-plane latency. A similar path has been followed by Internet of Things (IoT) technologies such as narrowband IoT (NB-IoT) and LTE-M, introduced in 3GPP Release 13 [2]. These enhancements include new numerologies in 5G, grant-free random access and advanced grant-based mechanisms, along with priority handling mechanisms. Regarding the frame structure, the new numerologies in 5G allow to reduce the length of a slot (the minimum resource unit for allocation) in the time domain by increasing the subcarrier spacing [5].

In the downlink, besides the traditional scheduling mechanism based on the size of the transmission buffers, 5G possesses pre-emptive scheduling and semi-persistent scheduling (SPS) mechanisms, which allow to achieve low-latency communication. Pre-emptive scheduling [7], usually referred to as puncturing [10][1], allows to cancel the transmission of low-priority (i.e.,

broadband) data in specific time-frequency downlink resources so these can be used for the transmission of high-priority (i.e., latency-sensitive) data instead. Then, an interrupted transmission must be sent to the affected UE. On the other hand, SPS assigns resources to a UE periodically, so these are available if needed. Data transmission in the downlink only occurs in connected mode. However, uplink data transmission can occur in both connected and disconnected modes.

Uplink access in connected mode is purely grant-based. Therefore, the users must first transmit a scheduling request (SR) [8] and then, they must receive a specific grant from the gNB that includes the resources allocated for their transmissions. Until Release 16, uplink scheduling supports SPS, along with cancellation indication (CI) and configured grant mechanisms. Like pre-emptive scheduling in the downlink, CI allows for out-of-order scheduling in the uplink. Specifically, the gNB can send a CI message to a user with uplink resources allocated previously after receiving an SR from a user with latency-sensitive data to transmit. In such a case, the user receiving the CI will cancel the transmission in the indicated resources.

The baseline access mechanism in disconnected mode (i.e., random access) in 5G, NB-IoT, and LTE-M is grant-based. This mechanism is known as the random access (RA) procedure and consists of a four-message handshake [8]. Initially, only after completing the procedure would the UEs be allowed to transmit data by first transmitting an SR. However, several enhancements have been made to the procedure to reduce the control overhead of data transmission. The most notable enhancement is the Type-2 RA procedure defined in Release 15 of 5G NR [6]. The type-2 RA procedure begins with the traditional preamble transmission, but it is accompanied by a short uplink data transmission. The selected preamble determines the specific time-frequency resources for the uplink transmission. This is, effectively, a grant-free access protocol whose capacity is limited by the number of preambles.

Despite the advances in scheduling mechanisms for orthogonal multiple access and contention-based random access, there are still several aspects that require a redesign of the radio access network towards 6G. For instance, resource efficiency and, consequently, the number of supported users is limited by strict rules for orthogonal resource allocation. As an example, services with URLLC-like requirements but with uncertainty in the activation in 5G are limited to using either time-consuming grant-based access (e.g., via SR and CI for previously allocated users [6]) or some form of SPS, which leads to resource wastage. Furthermore, even though multi-connectivity has been previously introduced, the cell-based infrastructure does not possess standardized mechanisms to fully exploit the multiple links to the infrastructure. Moreover, the implemented scheduling mechanisms and configuration parameters are usually static and must be designed and selected by mobile network operators (MNOs), which creates a mismatch between the pre-planned service provisioning capabilities of the infrastructure and the actual user demands. To fulfil the specific requirements of the advanced 6G applications while maximizing resource efficiency, self-optimization mechanisms for parameter selection and slicing of the radio access resources must be put into place, in combination with a wide variety of access mechanisms. This combination of capabilities is called Intelligent Edge [63].

### 2.3.2 Orthogonal and non-orthogonal RAN slicing and resource allocation

In the downlink, RAN slicing in 5G between eMBB and URLLC users corresponds to the resource allocation problem of incoming data with two different requirements. In [9], the authors proposed a deep reinforcement learning approach to allocate resources to both service types. The risk of a specific allocation (i.e., time and frequency resources) for URLLC users in the finite block length regime is used as the input to a learning agent. The URLLC users were uniformly distributed within the cell area, each URLLC packet is transmitted to a different URLLC user (i.e., no memory per user), and the error probability at each URLLC packet transmission was used to calculate the overall reliability. The approach is interesting but cannot be directly applied in practice due to the simplification of the calculation of the reliability of URLLC users. Nevertheless, it influenced other works that target the latency aspect of URLLC. For example, in [10], the slicing of the resources is

performed assuming that, if URLLC is scheduled with sufficiently low latency, the reliability aspects are covered. Hence, [10] focuses on the importance of the loss model for the punctured eMBB traffic, considering a linear, threshold-based, and quadratic loss model as a function of the punctured resources. Furthermore, [68] explores the potential of optimizing resource allocation with flexible numerology in frequency domain and variable frame structure in time domain, with services of with different types of requirements including URLLC.

Despite not being included in the 5G standardization, non-orthogonal multiple access (NOMA) presents interesting advantages when compared to orthogonal multiple access (OMA) schemes. Specifically, NOMA may result in an increased resource efficiency and scalability, along with lower latency than traditional OMA schemes. On the downside, NOMA schemes usually require a high level of contextual awareness (e.g., the ratio of channel qualities among users) and more complex encoding and decoding mechanisms.

An example of the latter is provided in [49] comparing the energy consumption of using OMA and NOMA as scheduling solutions for eMBB and URLLC traffic. In particular, the schemes proposed in [49] exploit the available channel state information (CSI) of the eMBB users, while relying on statistical CSI only for the URLLC users, to allocate the resources for both service types. Results in [49] show that NOMA attains lower power consumption than OMA in most cases, except when the average channel gain of the URLLC user is exceedingly high. Even in these cases, the gap between NOMA and OMA is negligible, showing the capability of NOMA to reduce the power consumption and guarantee close-to-the-optimal optimal performance in practically every condition.

In the uplink, eMBB may coexist with URLLC services, but also with other IoT-like services with diverse timing requirements. For example, services that require either a 1) high reliability with slightly relaxed latency or 2) a flow-level timing metric such as AoI [53].

The performance of NOMA and OMA mechanisms in the uplink has been studied in AoI-focused scenarios with homogeneous users [38]. Furthermore, OMA and NOMA mechanisms have been investigated with coexisting eMBB and IoT users, where the latter require either low latency and high reliability or AoI in a collision channel model [19][35]. Even though the collision channel model is not favourable for NOMA, results showed that NOMA may outperform OMA for latency-oriented services, whereas AoI-oriented users are less sensitive to the selection of access/slicing mechanism. The reason for this is that the inter-arrival times tend to dominate the AoI and, once a sufficiently high reliability is achieved, the differences in per-packet latency between NOMA and OMA have a small impact on AoI. The analyses conducted in [19][35] were recently extended in [20] to a scenario with an enriched channel model where capture might occur when the SINR of one of the overlapping signals is sufficiently high. Therefore, the diverse outcomes for a channel realization and their probabilities were considered to derive the performance of OMA and NOMA in the uplink. The results showed that the probability of capture greatly enhances the advantages of NOMA in the latency-oriented scenario, allowing the IoT users to achieve considerably low delay and a near-optimal performance for the eMBB users, which are unattainable with OMA.

In [42] the interplay between delay violation probability and the average AoI in a wireless multiple access channel with multipacket reception capability and heterogeneous traffic characteristics is studied. Further, the coexistence of a primary user that aims to minimize the AoI with secondary users that communicate among them in a cognitive network was studied in [32]. The latter illustrates how the AoI requirements of the primary user limit the aggregate throughput for the secondary users. Furthermore, it is shown that the aggregate throughput remains relatively stable as the number of secondary users becomes large.

### 2.3.3 Random access (RA) mechanisms: grant-based and grant-free

To achieve the performance requirements of the users, different RA mechanisms must be implemented depending on the traffic characteristics and the number of contending users. Differently from multiple access, where the users are known and they are assumed to be active, one of the main challenges in RA is activity detection, which must be performed before decoding the

data. Grant-based RA has been widely used in 4G and many other systems focused on broadband traffic, where the overhead of the initial handshake becomes negligible. Effectively, grant-based RA solves the activity detection problem by using orthogonal pilot signals (e.g., preambles in 4G and 5G) during an initial reservation phase. However, numerous works proved its inefficacy in massive RA scenarios (i.e., mMTC) [4][57]. Therefore, a significant amount of research has been performed on grant-free RA mechanisms [45] and on understanding the nature of packet transmissions in the finite-block length regime [45][67].

Even though slotted ALOHA mechanisms have been around for many years now, most advanced RA mechanisms are variations of these. For instance, the access mechanisms in NB-IoT are based on transmitting consecutive repetitions of the packets in multichannel slotted ALOHA channel. The repetitions in NB-IoT might effectively counteract the effects of fast fading but come at the cost of increased resource utilization and, hence, reduced access capacity (i.e., packets per time slot) [64]. To increase the capacity of the access channel, irregular repetition slotted ALOHA (IRSA), implements a degree distribution, which is a function to place the repetitions randomly across the frames [21]. Coded Pilot Access (CPA) is a closely related mechanism to IRSA, where data packets are accompanied by pilot signals that aid the decoding of the data in massive MIMO systems [52].

An area of research that has attracted significant attention recently is sparse estimation. Advances in computing platforms and algorithms can now be used to solve underdetermined systems of equations by exploiting the sparsity of the solution. Among these, compressed sensing (CS) has been widely explored for activity detection and estimation [9][27]. In massive access scenarios, CS can be applied to determine the activation and to estimate the channel coefficients of a relatively small subset of users from their overlapped non-orthogonal signals.

### 2.3.4 Multi-connectivity

Maintaining multiple links to the infrastructure allows users to increase throughput and reliability while reducing latency. This is achieved by exploiting the macro-diversity of the environment in terms of space (e.g., by connecting to different BSs) and/or frequency (e.g., by connecting to the same BS but on frequency different bands) which increases the resources allocated per user [65]. Therefore, while multi-connectivity may benefit specific users, careful allocation of resources is needed to avoid resource wastage and, hence, the number of users and their requirements must be considered. Furthermore, the dependencies between the different channels (i.e., links) can be exploited to minimize outage probability and, possibly, optimize resource allocation and link selection [11]. An algorithm for the link scheduling optimization that maximizes the network throughput for multi-connectivity in millimetre-wave cellular networks is proposed in [56]. The considered approach exploits a centralized architecture, fast link switching, proactive context preparation and data forwarding between millimetre-wave access points and the users.

Furthermore, a specific type of multi-connectivity where different interfaces is used is termed interface diversity. In [25], transmission policies and the benefits of interface diversity with an LTE and a WIFi interface were investigated for cyber-physical systems where periodic uplink transmissions take place. In [18], the effect of bursty traffic in an LTE and Wi-Fi Aggregation (LWA)-enabled network is investigated. The LTE base station routes packets of the same IP flow through the LTE and Wi-Fi links independently. Superposition coding is used at the LWA-mode Wi-Fi access point (AP) so that it can serve LWA users and Wi-Fi users simultaneously. Then, a congestion-aware random-access protocol is applied to avoid impeding the performance of the LWA-enabled network.

A detailed discussion of 3GPP support for interface diversity can be found in Section 5.4.3.

### 2.3.5 Self-optimization mechanisms

The joint configuration of the access parameters and resource slicing and allocation, both in a real-time and off-line manner, is a complicated task. To achieve an optimal use of network resources, the

performance requirements, traffic characteristics, channel conditions, and capabilities of the served users must be known. Naturally, the use of traditional optimization techniques, including convex optimization and dynamic programming, provides numerous benefits such as optimality guarantees and the conditions under which optimality can be achieved. Furthermore, the vast literature on these techniques allows us (under some conditions) to understand the process and the rationale for the obtained outcomes. Consequently, traditional optimization techniques should remain the preferred option for self-optimization mechanisms. However, the increased complexity caused by the use of detailed models for the channel, traffic, and capabilities of the users complicates the use of traditional optimization techniques, along with the design and analysis of the considered access schemes.

Machine learning techniques, which are envisioned to play a major role in 6G [24], can be used to complement traditional optimization techniques in overly complicated scenarios. For instance, realistic performance evaluation and optimization in random access scenarios present a major challenge. Thus, most of the developed access mechanisms oversimplify the different channel characteristics and performance requirements of the users, which might be greatly different. If these differences are neglected, access mechanisms may either suffer from low resource efficiency or fail to meet the performance guarantees of the users. ML techniques including, among others, deep learning, (deep) reinforcement learning, multi-armed or contextual bandits, can be used to achieve on-the-fly self-adaptation of the RAN, for example, for RAN slicing [9] and resource allocation. Furthermore, lightweight ML techniques could be deployed at the user side for the individualized adaptation of the users' behaviour (i.e., policy). However, some of the major challenges for applying ML and other data-driven optimization techniques are the amount of training data that is needed by the algorithms [13] and the impact of the knowledge and fundamental assumptions about the environment, for example, the channel [10].

### 2.3.6 Advanced Channel Coding and Modulation

Modern channel coding schemes like LDPC and Polar codes, introduced in the 5G NR specification, provide near-capacity error correction performance. However, when combining these with higher order modulation schemes, a gap in capacity exists [32]. One of the reasons for this loss is due to the non-optimal probability distribution of transmitted symbols. In [14] and [37] probabilistic shaping and geometric shaping schemes, respectively, are introduced that avoid this so-called shaping loss. It was shown in [59] that probabilistic shaping can be implemented jointly with polar coding, by employing the successive cancellation decoder both at encoded and decoder sides. Furthermore, an explicit code construction algorithm was presented. Moving forward to 6G, the requirements on energy efficiency, latency and throughput are only expected to increase. This demands further investigations of a hardware friendly implementation of efficient shaping schemes. Recently joint probabilistic and geometric shaping is used to design constellations in [52] with an autoencoder to learn the constellation over a wide range of SNR performing close to capacity.

Furthermore, 6G systems will have to support several types of traffic with extremely diverse requirements, including frame error rate, end-to-end data latency, data block size and transceiver power consumption. These requirements must be supported by the appropriate forward error correction techniques. For 5G, the application of polar codes was limited to control channel only, where small blocks of data are transmitted, due to the following reasons:

1. Long polar codes require large list size in the successive cancellation list decoder to achieve reliable performance. This results in a high decoding complexity.

2. The successive cancellation decoding algorithm and its derivatives suffers from high decoding latency.

However, small block length polar codes can also provide excellent performance with extremely low decoding complexity.

Supporting multiple types of codes results in excessive complexity and power consumption of the communication hardware. It is therefore highly desired to have a unified coding solution, which could be adapted to the requirements of various traffic types and block lengths and would facilitate different decoding approaches depending on the amount of computing power available at the receiver, affordable latency and target performance.

The crucial parameter which relates the performance and block length for a family of codes is the scaling exponent. Optimal scaling exponent of 2 is achieved by random codes [68], whereas Arikan polar codes and LDPC codes achieve an scaling exponent of 3.627 (for the case of the binary erasure channel) and 3, respectively [39][40]. Polar codes with large kernels were shown to asymptotically achieve m=2 [26]. Several constructions of polarization kernels are available [36][48]. However, their practical merits depend on the complexity of the kernel processing (also known as kernel marginalization) operation, i.e. computing the log-likelihood ratios arising in the successive cancellation decoding algorithm. In general, the kernel must be carefully optimized offline to ensure that its processing is simple enough, while its polarization properties are sufficiently good [60]. It is possible to show that for well-optimized kernels under successive cancellation list decoding it is possible to obtain both better performance and lower decoding complexity compared to the codes based on the Arikan kernel [58][61]. Furthermore, it is possible to implement code length adaptation by employing a family of shortened polarization kernels [62], which admit unified hardware implementation of the processing algorithm for kernels of different size.

The problem of high decoding latency, which is inherent for the decoding algorithms based on the successive cancellation approach, can be avoided by employing belief propagation decoding techniques. This, however, requires one either to transform the factorgraph of the original polar code [16], or to explicitly optimized the code structure for belief propagation decoding [30].

Further developments in this area may enable the design of a FEC scheme which can be scaled for different QoS requirements in terms of performance, complexity, and latency.

### 2.3.7 Access mechanisms in distributed MIMO architectures

One of the promises of 6G is to integrate a wide range of device types into a distributed architecture to achieve greater network deployment flexibility and coverage.

Cell-free architectures are a candidate for 6G where groups of radio heads or distributed antenna elements are controlled by a central entity. If the number of antennas in these elements is much greater than the number of users, then it is referred to as a cell-free massive MIMO network [69]. While cell-free massive MIMO networks provide numerous benefits when compared to traditional cell-based architectures such as greatly reducing the outage probability, having a shared RAN between several radio heads and controllers, introduces new challenges in the design of radio access mechanisms. For example, RAN slicing and resource allocation in cell-based architectures are centralized optimization problems, whereas they become distributed optimization problems in cell-free MIMO architectures, which increases their complexity [28].

Furthermore, Reflective Intelligent Surfaces (RIS) are an interesting alternative for network densification. By being passive elements rather than active, RIS lack several functionalities when compared to traditional base stations but, in exchange, provide a more flexible, cost-efficient alternative to eliminate coverage holes. While most of the research on RIS has focused on physical layer aspects, recent works are now focusing on medium access control (MAC) protocol-layer aspects to make RIS a reality. For example, [23] is one of the first to tackle the design of access policies based on the beam sweeping pattern of the RIS, which is needed to cover the area where line-of-sight to the base station is obstructed. It is observed that appropriate access policies can be defined based on the knowledge obtained during a training phase, which allows the users to learn the sweeping pattern and to identify the best time to transmit. However, the training phase must be carefully designed to avoid excessive overhead.

### 2.3.8 Radio access for semantic and goal-oriented communications

The seminal work by Shannon in 1948 set the foundation of modern communication systems by quantifying the maximum data rate that a noisy communication channel can support [50]. In Shannon's work, all messages were treated as having equal importance regardless of their semantic meaning or the underlying application for which communication is used, which sets the final goal of the communication. This assumption, together with the guarantee that separation of source and channel coding performs equally as a Joint Source and Channel Coding (JSCC) strategy in the infinite block-length regime, motivated the split of communication systems into two independent subsystems: the communication modules (i.e., lower layers of the protocol stack) and the application itself. This principle has defined the way most communications systems are designed today, whose main goal has been to reconstruct the messages at the receiver side with the highest fidelity possible.

Nevertheless, this classic view of communications covers the technical problem of communications, which is considered to be one out of the three levels of communication problems already outlined by Weiver in the introduction of Shannon's seminal work in 1959 [51]. Therefore, the classic view neglects the other two levels of problems. The semantic problem focuses on conveying the semantic content of the information to the receiver, while the individual symbols do not need to be perfectly reconstructed. The effectiveness problem focuses on achieving the ultimate goal of the system for which the messages are transmitted. Consequently, the effectiveness problem can be a new and appealing framework to address use cases of remote control of cyber-physical systems described above.

As mentioned earlier, the technical problem has been the focus in the design of communication systems until a few years ago. This is mainly because humans have always been assumed to be the ultimate consumers of information, and humans inherently address the semantics and effectiveness problems of communication. It is only with the advent of machines with advanced processing and communication capabilities, along with intelligence (i.e., the mechanisms and algorithms to make intelligent decisions), that the limitations of focusing only on the technical problem of communications are being revealed. With machines in fact, fidelity of reconstructed messages is not necessarily the most relevant criterion to guarantee optimal operation. Consider for example a robotic application where a robot collects sensing information and sends it to a central server for processing, after which the output is sent back to the robot. Depending on the task, e.g., object detection, there are clearly some segments of the sensor information more relevant for the optimal performance of the task than others, such as areas on the video feed of the robot where relevant objects lie. An optimal communication system for such a task should be designed with the awareness of this contextual relevance, as well as the timing requirements of the application, for example, to prevent the robot to constantly crash against obstacles.

In [29], the design of a Goal-Oriented Communication (GOCom) system is proposed, where a Goal-Oriented Encoder (GOE) and a Decoder (GOD), separated by a wireless channel, are jointly trained to generate a task output based on an input signal. Figure 2 shows the general system model, where the GOE is represented by the function $f_\theta$, with $f$ being the non-linear mapping parametrized by learnable parameters $\theta$ and the GOD is represented by a function $g_\phi$, with $g$ being a non-linear mapping parametrized by learnable parameters $\phi$.


*Figure 2: System model of GOCom system where x represents the input features, z represents the symbols encoded by the Goal-oriented encoder (GOE) and H is the channel transfer function. Then, $\hat{z}$ is the corrupted symbols by the channel with transfer function H, and $\hat{y}$ is the decoded output of the Goal-Oriented Decoder (GOD).*

The framework is general enough to handle any kind of learning task and communication channel, as long as the task and the channel are (almost everywhere) differentiable functions. Furthermore, [29] introduced algorithms for both supervised learning and Reinforcement Learning (RL) and presented the case study of GOCom for image transmission. Within the case study, two communications systems were designed, the first one focused on an image classification task, while the second one focused on an RL task. The simulation results showed that the intuition behind designing specialized communications systems for a particular task indeed holds and GOCom increases the performance when compared with JSCC, especially in bad channel conditions. Moreover, it was shown, that for the application of playing the Atari game "BreakOut" with RL, the agent is extremely sensitive to the distortion of reconstructed signals and fails drastically with a JSCC strategy, while by using our GOCom approach, the communication system effectively focuses on the relevant parts of the transmitted information.

Many other examples exist for the joint design of application and communication system in the context of JSCC [15], Sematic Communications [66], or Goal-Oriented Communications [41]. While all of them show great potential, the question of how to integrate such systems in current communication systems, which are based on the separation of source and channel coding principle, remains an open challenge with very little contributions up to this point. In [47], authors propose the inclusion of a "transversal" semantic layer across the protocol stack to exchange semantic information and enable such new ways of communication. Another example is given in [34], where instead of a transversal layer, a semantic layer is introduced at the application level. This layer is in charge of communicating with the physical layer of the communication system to exchange relevant information related to the transmitted information, as well as algorithmic information to semantically process the application data before transmission. Other solutions are being currently proposed [17][55], but they all require i) extensive modification of the protocol stack, and ii) the exchange of proprietary information, in particular raw data and algorithmic information, between application and communication system. These limitations make it impractical to implement such systems within current commercial communication and networking systems such as 3GPP.

To overcome these limitations, it is possible to include a pair of functions in the data plane of the MAC layer, called Extract (at the transmitter) and Combine (at the receiver). The job of the Extract function is to extract the payload of ICC applications from the upper-layers headers, so the payload can be transported directly to the resource mapping function of the physical layer, while the headers are sent through the standard path, i.e., the standard PHY layer functions are applied to the headers, since they represent unstructured data and need to be encoded and modulated. At the receiver side, the Combine function collects the received payload information, along with the headers, and combines them into packets to be delivered at higher layers. Figure 3 shows the data plane diagram including the pair of new functions.

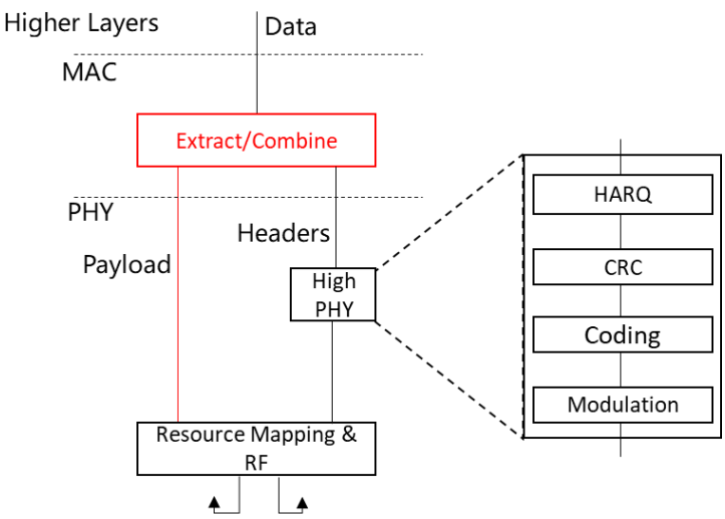



*Figure 3: Data plane including Extract/Combine functions.*

Note that the implication of treating the information coming from semantic applications in this way is that, effectively, the application is in charge of generating the baseband I/Q symbols to be transmitted over the radio access. A new type of scheduler is necessary to accommodate this new type of information and guarantee a certain Quality of Service (QoS) for such applications. Such a scheduler needs to be designed with the objective to allocate a certain number of resource elements (in time, frequency and/or space) per second for each user, given a certain expected SNR. Such information is user-dependent and would need to be provided by the user to the communication system at, for example, session establishment.

## 2.4 References


[1] 3GPP, "Service requirements for cyber-physical control applications in vertical domains," TS 22.104 V16.5.0, 2020.

[2] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2," TS 36.300, V13.14. Apr. 2020.

[3] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2," TS 36.300, V16.6. Sept. 2021.

[4] 3GPP, "Study on RAN Improvements for Machine-Type Communications," TR 37.868, Jul. 2011.

[5] 3GPP, "NR; Physical channels and modulation," TS 38.211 V17.0.0, Dec. 2021.

[6] 3GPP, "Physical layer procedures for control," TS 38.213 V16.3.0, 2020.

[7] 3GPP, "NR and NG-RAN Overall Description; Stage 2. TS 38.300 V15.3.1, Oct. 2018.

[8] 3GPP, "5G; NR; Medium Access Control (MAC) protocol specification," TS 38.321 V16.3.0, Jan. 2021.

[9] M. Alsenwi, N. H. Tran, M. Bennis, S. R. Pandey, A. K. Bairagi, and C. S. Song, "Intelligent Resource Slicing for eMBB and URLLC Coexistence in 5G and Beyond: A Deep Reinforcement Learning Based Approach," IEEE Transactions on Wireless Communications, vol. 12, no. 7, 2021.

[10] A. Anand, G. de Veciana and S. Shakkottai, "Joint Scheduling of URLLC and eMBB Traffic in 5G Wireless Networks," IEEE/ACM Transactions on Networking, vol. 28, no. 2, pp. 477-490, April 2020.

[11] M. Angjelichinoski, K. F. Trillingsgaard, and P. Popovski, "A Statistical Learning Approach to Ultra-Reliable Low Latency Communication," IEEE Transactions on Communications, vol. 67, no. 7, pp. 5153–5166, Jul. 2019.

[12] K. -L. Besser, P. -H. Lin and E. A. Jorswieck, "On Fading Channel Dependency Structures with a Positive Zero-Outage Capacity," IEEE Transactions on Communications, vol. 69, no. 10, pp. 6561-6574, 2021,

[13] K.-L. Besser, B. Matthiesen, A. Zappone, and E. A. Jorswieck, "Deep Learning Based Resource Allocation: How Much Training Data is Needed?", in Proc. IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC), 2020.

[14] G. Boecherer, F. Steiner and P. Schulte: "Bandwidth efficient and rate-matched low density parity-check coded modulation," IEEE Transactions on Communications, 63(12), pp. 4651-4665, 2015.

[15] E. Bourtsoulatze, D. B. Kurka, and D. Gunduz, "Deep joint sourcechannel coding for wireless image transmission," IEEE Transactions on Cognitive Communications and Networking, vol. 5, no. 3, pp. 567–579, 2019.

[16] S. Cammerer, M. Ebada, A. Elkelesh, and S. ten Brink, "Sparse Graphs for Belief Propagation Decoding of Polar Codes," in IEEE Inter. Symp. Inf. Theory (ISIT), June 2018.

[17] C. Chaccour, W. Saad, M. Debbah, Z. Han, and H. V. Poor, "Less data, more knowledge: Building next generation semantic communication networks." arXiv preprint arXiv:2211.14343, 2022.

[18] B. Chen, N. Pappas, Z. Chen, D. Yuan, J. Zhang, "Throughput and Delay Analysis of LWA with Bursty Traffic and Randomized Flow Splitting", IEEE Access, vol. 7, pp. 24667-24678, 2019.

[19] F. Chiariotti, I. Leyva-Mayorga, Č. Stefanović, A. E. Kalør, and P. Popovski, "Spectrum Slicing for Multiple Access Channels with Heterogeneous Services," Entropy, vol. 23, no. 6, p. 686, May 2021.

[20] F. Chiariotti, I. Leyva-Mayorga, Č. Stefanović, A. E. Kalør and P. Popovski, "RAN Slicing Performance Tradeoffs: Timing Versus Throughput Requirements," IEEE Open Journal of the Communications Society, vol. 3, pp. 622-640, 2022.

[21] J. W. Choi, B. Shim, Y. Ding, B. Rao, and D. I. Kim, "Compressed Sensing for Wireless Communications: Useful Tips and Tricks," IEEE Communications Surveys and Tutorials, vol. 19, no. 3, pp. 1527–1550, 2017.

[22] F. Clazzer, E. Paolini, I. Mambelli, and C. Stefanovic, "Irregular repetition slotted ALOHA over the Rayleigh block fading channel with capture," IEEE International Conference on Communications, pp. 4–9, 2017.

[23] V. Croisfelt, F. Saggese, I. Leyva-Mayorga, R. Kotaba, G. Gradoni and P. Popovski, "A Random Access Protocol for RIS-Aided Wireless Communications," in Proc. IEEE International Workshop on Signal Processing Advances in Wireless Communication (SPAWC), 2022.

[24] S. Dang, O. Amin, B. Shihada, and M.-S. Alouini, "What should 6G be?", Nature Electronics, vol. 3, pp. 20–29, 2020.

[25] I. Donevski, I. Leyva-Mayorga, J. J. Nielsen, and P. Popovski, "Performance trade-offs in cyberphysical control applications with multi-connectivity," Frontiers in Communications and Networks, vol. 2, 2021.

[26] A. Fazeli, H. Hassani, M. Mondelli and A. Vardy, "Binary Linear Codes With Optimal Scaling: Polar Codes With Large Kernels," in IEEE Transactions on Information Theory, vol. 67, no. 9, pp. 5693-5710, Sept. 2021

[27] Z. Gao, L. Dai, S. Han, C.-L. I, Z. Wang, and L. Hanzo, "Compressive Sensing Techniques for Next-Generation Wireless Communications," IEEE Wireless Communications, vol. 25, no. 3, pp. 144–153, Jun. 2018.

[28] D. Gunduz, P. De Kerret, N. D. Sidiropoulos, D. Gesbert, C. R. Murthy, and M. Van Der Schaar, "Machine Learning in the Air," IEEE Journal on Selected Areas in Communications, vol. 37, no. 10, pp. 2184–2199, 2019.

[29] M. A. Gutierrez-Estevez, W. Yiqun, and Z. Chan, "Learning to communicate with intent: An introduction," in Proc. IEEE PIMRC, 2023. arXiv preprint: arXiv:2211.09613 (2022).

[30] T. Koike-Akino and Y. Wang, "Protograph-Based Design for QC Polar Codes," *in Proc. IEEE International Symposium on Information Theory (ISIT),* 2021.

[31] M. Kountouris and N. Pappas, "Semantics-Empowered Communication for Networked Intelligent Systems," IEEE Communications Magazine, vol. 59, no. 6, pp. 96–102, Jun. 2021.

[32] A. Kosta, N. Pappas, A. Ephremides, and V. Angelakis, "Age of Information and Throughput in a Shared Access Network with Heterogeneous Traffic," in Proc. IEEE Global Communications Conference (GLOBECOM), 2018.

[33] F. R. Kschischang and S. Pasupathy: "Optimal nonuniform signaling for Gaussian channels," IEEE Transactions on Information Theory, vol. 39, no. 3, pp. 913-929, 1993.

[34] Q. Lan, D. Wen, Z. Zhang, Q. Zeng, X. Chen, P. Popovski, and K. Huang, "What is semantic communication? A view on conveying meaning in the era of machine intelligence," Journal of Communications and Information Networks, vol. 6, no. 4, pp. 336-371, 2021.

[35] I. Leyva-Mayorga, F. Chiariotti, C. Stefanovic, A.E. Kalør, and P. Popovski, "Slicing a single wireless collision channel among throughput- and timeliness-sensitive services," in Proc. IEEE Int. Conf. Commun. (ICC), 2021.

[36] H.-P. Lin, S. Lin and K. A. Abdel-Ghaffar, "Linear and nonlinear binary kernels of polar codes of small dimensions with maximum exponents", IEEE Transactions on Information Theory, vol. 61, no. 10, 2015.

[37] N. S. Loghin, J. Zöllner, B. Mouhouche, D. Ansorregui, J. Kim and S.I. Park: "Nonuniform constellations for ATSC 3.0," IEEE Transactions on Broadcasting, 62(1), 197-203, 2016.

[38] A. Maatouk, M. Assaad, and A. Ephremides, "Minimizing the Age of Information: NOMA or OMA?" in Proc. IEEE INFOCOM Workshops, vol. 65, no. 8, 2019, pp. 102–108.

[39] M. Mondelli, S. H. Hassani, and R. L. Urbanke, "Unified scaling of polar codes: Error exponent, scaling exponent, moderate deviations, and error floors," IEEE Trans. Inf. Theory, vol. 62, no. 12, pp. 6698–6712, Dec. 2016.

[40] M. Mondelli, S. H. Hassani, and R. L. Urbanke, "How to achieve the capacity of asymmetric channels," IEEE Trans. Inf. Theory, vol. 64, no. 5, pp. 3371–3393, May 2018.

[41] A. Mostaani, T. X. Vu, S. K. Sharma, V. -D. Nguyen, Q. Liao and S. Chatzinotas, "Task-Oriented Communication Design in Cyber-Physical Systems: A Survey on Theory and Applications," IEEE Access, vol. 10, pp. 133842-133868, 2022.

[42] N. Pappas, M. Kountouris, "Delay Violation Probability and Age-of-information Interplay in the Two-user Multiple Access Channel", IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), July 2019.

[43] N. Pappas and M. Kountouris, "Goal-Oriented Communication For Real-Time Tracking In Autonomous Systems," in Proc. IEEE International Conference on Autonomous Systems (ICAS), 2021.

[44] M. Pikus and W. Xu: "Bit-level probabilistically shaped coded modulation," IEEE Communications Letters, vol. 21, no. 9, pp. 1929–1932, Sept. 2017.

[45] Y. Polyanskiy, "A perspective on massive random-access," in Proc. IEEE Int. Symp. Inf. Theory (ISIT), Jun. 2017, pp. 2523–2527.

[46] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," IEEE Trans. Inf. Theory, vol. 56, no. 5, pp. 2307–2359, 2010.

[47] P. Popovski, O. Simeone, F. Boccrdi, D. Gündüz, and O. Sahin, "Semantic-effectiveness filtering and control for post-5G wireless connectivity." Journal of the Indian Institute of Science, vol 100, no. 2, pp. 435-443, 2020.

[48] N. Presman, O. Shapira, S. Litsyn, T. Etzion and A. Vardy, "Binary polarization kernels from code decompositions", IEEE Transactions on Information Theory, vol. 61, no. 5, May 2015.

[49] F. Saggese, M. Moretti and P. Popovski, "Power Minimization of Downlink Spectrum Slicing for eMBB and URLLC Users," IEEE Transactions on Wireless Communications, 2022.

[50] C. E. Shannon, "A mathematical theory of communication," The Bell system technical journal, vol. 27, no. 3, pp. 379-423, 1948.

[51] C. E. Shannon and W. Weaver, "The mathematical theory of communication," University of Illinois Press, Champaign, 1964.

[52] J. H. Sorensen, E. de Carvalho, C. Stefanovic, and P. Popovski, "Coded Pilot Random Access for Massive MIMO Systems," IEEE Transactions on Wireless Communications, vol. 17, no. 12, pp. 8035–8046, Dec. 2018.

[53] G. Stamatakis, N. Pappas, A. Traganitis, "Optimal Policies for Status Update Generation in an IoT Device with Heterogeneous Traffic", IEEE Internet of Things Journal, vol. 7, no. 6, June 2020.

[54] M. Stark, F. Aoudia and J. Hoydis: "Joint Learning of Geometric and Probabilistic Constellation Shaping", arXiv:1906.07748v3

[55] E. C. Strinati and S. Barbarossa. "6G networks: Beyond Shannon towards semantic and goal-oriented communications." Computer Networks, vol. 190, 2021.

[56] C. Tatino, I. Malanchini, N. Pappas, D. Yuan, "Maximum Throughput Scheduling for Multi-connectivity in Millimeter-Wave Networks", 16th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), May 2018.

[57] L. Tello-Oquendo, V. Pla, I. Leyva-Mayorga, J. Martinez-Bauset, V. Casares-Giner, and L. Guijarro, "Efficient random access channel evaluation and load estimation in LTE-A with massive MTC," IEEE Trans. Veh. Technol., vol. 68, no. 2, pp. 1998–2002, 2019.

[58] P. Trifonov, "Recursive Trellis Processing of Large Polarization Kernels," in Proc. *IEEE International Symposium on Information Theory (ISIT),* 2021.

[59] P. Trifonov, "Design of Multilevel Polar Codes with Shaping," in Proc. Of Int. Symp. On Inf. Theory (ISIT), 2022.

[60] G. Trofimiuk, "A Search Method for Large Polarization Kernels," in Proc. *IEEE International Symposium on Information Theory (ISIT),* 2021.

[61] G. Trofimiuk and P. Trifonov, "Window Processing of Binary Polarization Kernels," *IEEE Transactions on Communications*, vol. 69, no. 7, pp. 4294-4305, July 2021.

[62] G. Trofimiuk, "Shortened Polarization Kernels," in Proc. *IEEE Globecom Workshops (GC Wkshps),* 2021.

[63] X. Wang, Y. Han, V. C. M. Leung, D. Niyato, X. Yan and X. Chen, "Convergence of Edge Computing and Deep Learning: A Comprehensive Survey," IEEE Communications Surveys & Tutorials, vol. 22, no. 2, pp. 869-904, second quarter 2020, doi: 10.1109/COMST.2020.2970550.

[64] Y.-P. E. Wang, X. Lin, A. Adhikary, A. Grövlen, Y. Sui, Y. Blankenship, J. Bergman, and H.-S. Razaghi, "A Primer on 3GPP Narrowband Internet of Things," IEEE Communications Magazine, vol. 55, no. 3, pp. 117–123, Mar. 2017.

[65] A. Wolf, P. Schulz, M. Dorpinghaus, J. C. S. Santos Filho, and G. Fettweis, "How Reliable and Capable is Multi-Connectivity?" IEEE Transactions on Communications, vol. 67 no. 2, pp. 1506–1520, 2019.

[66] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang "Deep learning enabled semantic communication systems," IEEE Transactions on Signal Processing vol. 69, pp. 2663-2675, 2021.

[67] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Block-Fading Channels at Finite Blocklength," in Proc. Int. Symp. Wireless Commun. Sys. (ISWCS), Aug. 2013, pp. 410–413.

[68] L. You, Q. Liao, N. Pappas, D. Yuan, "Resource Optimization with Flexible Numerology and Frame Structure for Heterogeneous Services", IEEE Communications Letters, vol. 22, no. 12, Dec. 2018.

[69] J. Zhang, E. Bjornson, M. Matthaiou, D. W. K. Ng, H. Yang, and D. J. Love, "Prospective Multiple Antenna Technologies for Beyond 5G," IEEE Journal on Selected Areas in Communications, vol. 8716, no. c, pp. 1–24, 2020.

# 3. Next Generation MIMO

The introduction and further development of multi-antenna communication techniques in the fourth and fifth generation mobile radios had a significant influence on the increase of spectral efficiency. Yet, information theory shows that the potential benefits of multi-antenna technologies are still far from being fully exploited. In theory, the capacity of wireless networks can be increased as desired by adding more and more antennas, both co-located on the same antenna panel and geographically distributed over the service area. From a practical perspective, however, this presents many new challenges. Some promising trends in next generation multi-antenna technologies will be discussed in this section.

## 3.1 Fundamental MIMO gains

Using multiple antennas has a number of fundamental benefits. Generally, these benefits depend strongly on how well the channel is known at the transmitter and/or the receiver, on the properties of the propagation channel (multipath characteristics, attenuation), on the link type, i.e., whether the channel is a point-to-point channel or a multi-user channel, and whether the antennas can "cooperate", i.e., whether the transmitted or received signals can be processed jointly. Also, the design of MIMO solutions crucially depends on the particular scenario under consideration. There is no one-size-fits-all MIMO solution. When designing MIMO systems, one always needs to consider practical constraints. It is also important to consider the trade-off between the basic MIMO gains, which are as follows.

*Array gain* – By coherently combining the signals of M antennas (no matter what antenna distance and whether the antenna is operated in transmit or receive mode), the SNR can be improved by at most 10 log10(M) dB. This upper bound is achieved by maximum ratio combining. It is reasonable to distinguish between array gain and beamforming gain, although they are often used synonymously. Array gain typically denotes the improvement of the average SNR when using M antennas, corresponding to a horizontal shift of the performance curves plotted versus the SNR in dB. For pure LoS channels, this is identical to the beamforming gain. However, for fading channels with rich scattering, beamforming provides an additional diversity gain through reduction of SNR variation.

*Multiplexing and multiple access gain* – A fundamental benefit of coherent antenna combining is the ability to eliminate interfering signals, by controlling the antenna weights such that superimposing wave fronts cancel out for certain channels. This is exploited in different ways, e.g. for space division multiple access (SDMA) in case of multi-user channels, or for spatial multiplexing in case of point-to-point channels. A typical assumption in this context is that the number of signals is not greater than the number of antennas at transmitter or receiver side. With an M-antenna array one can separate up to M signals without interference. It should be noted, however, that there is a fundamental trade-off between maximizing SNR and separating signals. Certain channel constellations can even lead to SNR losses up to the cancellation of the desired signal. This can be avoided, though, by combining multi-antenna processing with scheduling. The multiplexing gain is defined as the asymptotic slope of the (sum) spectral efficiency vs. SNR in dB (for a given error probability).

*Antenna diversity gain* – Multipath propagation spread in the angle domain is causing small scale fading over space. If the antenna distances are sufficiently large, then each antenna experiences a channel that is uncorrelated with those of other antennas. This is exploited by transmitting or receiving redundant information in parallel via multiple antennas and merging it at the receiver. Therefore, antenna diversity makes the channel robust against fading. The diversity gain is defined as the asymptotic slope of the (average/maximum) error probability vs. SNR in dB (for a given spectral efficiency).

A further fundamental benefit of using multiple antennas is the ability to estimate the spatial direction of incoming signals or multipath components. Estimating the so-called angle-of-arrival

(AoA) is of major importance for modern positioning and localization techniques. This approach complements Time-Difference-of-Arrival (TDoA) based techniques, which are constrained by tight requirements on time synchronization between base stations. The achievable spatial resolution of the signal components is improved by increasing the number of antenna elements and the aperture of the array.

While the SNR gain is independent of the inter-antenna distance, antenna diversity generally benefits from placing the antenna elements far away from each other, which decorrelates the fading observed at each antenna. Likewise, multiplexing and multiple access benefit from a spatially distributed arrangement of antennas. This improves the so-called "channel rank" and helps to separate signals in space. If cooperative antennas are distributed over several base stations, then this is known as Cooperative Multipoint (CoMP), which has the added benefit to be effective against shadow fading (macro diversity). Moreover, it avoids the cell-edge effect through base station cooperation. The CoMP advantage comes at the cost of increased signalling overhead for exchanging CSI and for jointly processing the distributed signals.

Since MIMO gains depend on the number of jointly processed antenna elements and the aperture of the array, the trend is towards ever larger arrays, either co-located (so called "massive MIMO" [1]) or distributed over the service area. The distributed MIMO (D-MIMO) technology has been discussed for 15 years (see e.g. [2]) and gained recently renewed attention within the context of cell-free massive MIMO [3]. For 6G, the concept of multiple D-MIMO arrays, termed modular massive MIMO (mmMIMO) [4], is introduced. This comprises structured D-MIMO and several variants of CoMP, including joint transmission (JT) as special cases. Prototype implementation as well as standardization steps are illustrated in [4].

## 3.2 MIMO channel modelling

Channel modelling is a fundamental step in the modelling of any wireless communications system since it determines to a certain extent how close to reality the results obtained are and therefore the conclusions derived from them. Specifically, for high frequencies where the wavelength is quite small, the modelling of the MIMO channel becomes more challenging since even imperfections on the surfaces of building structures or objects may affect signal propagation.

In this context, several MIMO-centric channel models have been designed. Some of them are listed below. These models cover the frequency range up to 100 GHz. Note that, the modelling of THz and sub-THz channels >100 GHz poses another challenge due to the special propagation properties in the THz spectrum, which behaves like optical channels. This is discussed separately in Section 1.

- 3GPP [5]: The model proposed by the 3GPP is valid for frequencies between 0.5 GHz and 100 GHz. The scenarios supported by the model are urban micro-cell street canyon, urban macro-cell, indoor office, rural macro-cell and indoor factory [6]. The model also supports mobility and spatial consistency.
- METIS [7]: Different channel model approaches are provided (map-based model, stochastic model and hybrid model). The frequency range considered goes form 2 GHz to 60 GHz.
- mmMAGIC [8]: The model is based on the 3GPP model. It is focused on the modelling of the frequency dependence of large-scale parameters, ground reflections, intra clusters parameters, small scale fading, blockage, and building penetration, among others.
- NYU Wireless [9]: It is a statistical spatial channel model. Multiple measurement campaigns were conducted on frequencies ranging between 28 GHz and 73 GHz for indoor and outdoor environments. Omni-directional and directional antennas are considered.
- QuaDRiGA [10]: The model considers quasi-deterministic extensions to support spatial consistency and tracking of users. Extensive radio channel measurements in the 6-100~GHz range were taken to parameterize the model.

- IEEE 802.11ad/ay [11]: These semi-deterministic models are based on ray-tracing techniques. They are focused on short range communications for scenarios such as conference rooms, cubicles and living rooms at 60 GHz. For the modelling of propagation losses, the specular components are calculated by ray-tracing algorithms, and the components due to diffractions, diffuse scattering and transmission are aggregated in a stochastic way. Human blockage is also considered in terms of blockage probability and blockage attenuation.
- MiWEBA [12]: It consists of a quasi-deterministic channel model for 60 GHz. The model focuses on university campus, street canyon, hotel lobby, backhaul, and D2D scenarios and addresses several challenges such as spatial consistency and shadowing.

## 3.3 MIMO transceiver design

As the number of antennas in an antenna array increases, the antenna distance decreases for given space constraints, and at the same time, mutual coupling among them tends to increase. Mutual coupling influences the transmit power as well as the spatial noise correlation. Improved hardware models – in particular circuit theoretical methods have been developed [13][14], leading to a systematic framework [15] to achieve the best performance taking mutual coupling into account (see Figure 4). This framework is advantageous, because it maps between the typical information theoretic model and the circuit model, which enables the reuse of existing algorithms. Various aspects of such a system were analysed, including reciprocity [16]. Furthermore, the framework can be used in the design of the antenna array and of other analog components, where it is important to have good simulation or measurement results of the array impedance matrix and its radiation patterns. The design of decoupling and matching networks (DMNs) for such systems is essential, because the number of elements to realize a DMN grows quadratically with the number of antennas in general [17]. The design of two-port matching networks at the receiver, where their number of elements only grows linearly with the number of antennas was considered in [18][19]. Two new classes of power amplifiers, class M and class N, were introduced in [20] to improve their power efficiency by energy recycling. Smaller than half a wavelength antenna spacing can be promising, because it enables super-gain, meaning an array gain larger than the number of antennas. Depending on the scenario, super-gain is also achievable with antenna spacing larger than half a wavelength. For example, array directivity approaches the number of antennas squared in a ULA consisting of isotropic radiators transmitting into endfire direction as the antenna spacing goes to zero [15]. For a given super-directivity, the gain also depends on the radiation efficiency, and good efficiency requires a well-designed DMN. Prototypes consisting of UCA with 3 to 4 antennas with antenna spacings smaller than half a wavelength and a wideband DMN were demonstrated in [21], where the measured results are close to the simulation results.

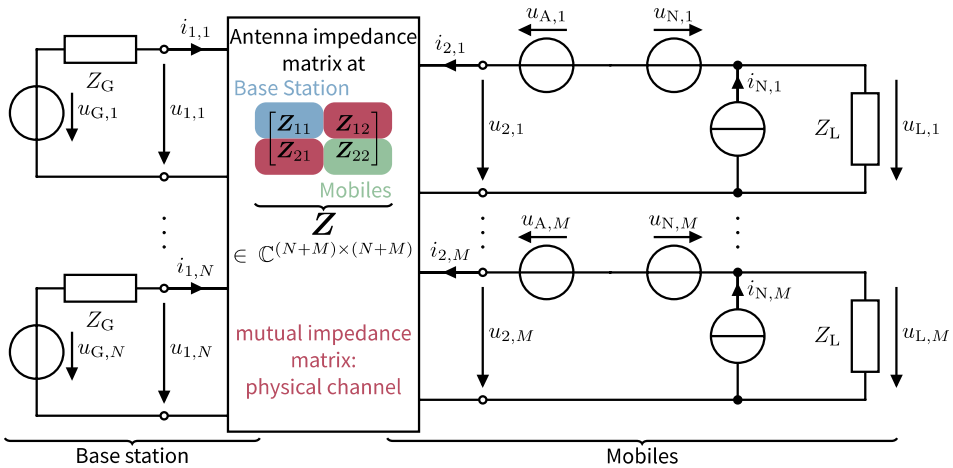



*Figure 4: Circuit model in the downlink*

Although massive MIMO is a promising technology, it comes with a cost of increased number of circuit components and hence increased energy consumption and signalling overhead. Note that, for massive MIMO antenna arrays, the design of fully digital transceivers implies as many RF chains as antenna elements. Therefore, the research community has shown some reluctance to embrace all-digital solutions, even though they are much more flexible and, in many cases, more powerful. This motivates the development of low-power fully digital approaches (e.g. at mmWave bands [22]). Such techniques, for example, are based on low-resolution data converters, use of constant envelope transmit signals and hybrid beamforming, as an alternative for a fully digital approach. As the focus was on decreasing the energy consumption at the receiver side, where the most power consuming elements are the analog-to-digital converters (ADCs), the achievable rate limits of communication systems employing low-resolution ADCs were studied first [23]. The analysis of a communication system with low-resolution ADCs is extended to the power efficiency afterwards [24]. At the transmitter side, the power amplifiers (PAs) are the most power consuming elements and constant envelope transmission is optimal for the PA's efficiency. One way to implement constant envelope transmit signals is to employ 1-bit digital-to-analog converters (DACs). The distortions due to the quantization at the DACs are considered in precoder design [25]. Numerous precoding techniques to improve the achievable rates (or to decrease the error rates) have followed after: some focused on developing a nonlinear precoder for the 1-bit quantization case [26][27], some on providing nonlinear and linear precoders for constant envelope signalling [28] and some have offered linear and nonlinear precoding techniques for the systems with uniform DAC quantization [29]. Also, works such as [30][31][32] provided an analysis of energy and spectral efficiency for the digital and hybrid beamforming systems employing low-resolution ADCs. Implementations based on low-resolution data converters (or constant envelope signalling) are promising for energy-efficient deployments of massive MIMO in 6G.

## 3.4 Line-of-sight (LoS) MIMO

In MIMO channels mainly dominated by a line-of-sight (LoS) path, spatial multiplexing gain can be extracted provided that the channel has a rank larger than one. However, the rank strongly depends on the regime where the antenna array is operated, namely the near-field or the far-field. The near-field applies if the distance between an antenna array and the corresponding communication device is smaller than the so-called Fraunhofer distance, which depends on the aperture of the antenna array and the used carrier frequency. For a fixed aperture of the antenna array, this distance scales proportionally to the carrier frequency; hence, at high frequency bands, the probability of operating in the near-field increases substantially with such an array. Since in the near-field the spherical model of the radio waves applies (as compared to the planar model in the far-field), this gives rise to linearly independent propagation paths, which set the basis for increasing the channel rank.

*(a) The LoS MIMO performance should be optimized by proper antenna placement at Tx and Rx arrays. As the optimum antenna placement is a function of the distance between each Tx-Rx antenna pair, the performance of LoS MIMO systems is susceptible to variations in the Tx-Rx distance and to misalignments of the arrays, e.g. due to rotation or tilting of the arrays. However, the capacity of a tilted array is determined by its projection onto the plane perpendicular to the line between the Tx and Rx arrays. Considering this, let us assume an Rx uniform circular array (UCA) facing that plane, but that the Tx array is tilted with respect to that plane. To compensate the effect of the tilt, the aim is to place the Tx antennas on the tilted Tx array plane, such that their projection on the plane perpendicular to the line between the Tx and Rx arrays results in an equivalent Tx UCA. For example, to compensate a Tx array tilt of -60 degrees, the antennas on the tilted Tx array plane should be located on an ellipse (see dashed green ellipse corresponding to subarray 1 in (b)*

*(b) Figure 5 below), such that its projection on the plane perpendicular to the line between the Tx and Rx arrays corresponds to a UCA with optimum diameter. Based on this, the Tx array can then be designed to consist of a set of concentric elliptical subarrays, so that the projection of each subarray on the plane which is*

*perpendicular to the line between the Tx and Rx arrays corresponds to a Tx UCA of optimum diameter (* **(b)**

**(c)** *Figure 5a). Hence, depending on the tilting angle of the Tx array, a different subarray can be selected to compensate the capacity loss arising at different array tilts [33], e.g., as shown in the example with 5 subarrays in* **(b)**

Figure 5b. In addition, although a tilted array can be detrimental at a fixed transmit distance, the tilting of an array can be leveraged towards improving the capacity of a system operating over varying Tx-Rx distances [34].

Furthermore, non-uniform array design of the Tx and Rx array, i.e., with unequal spacing between neighbouring antennas, can be considered as well to increase the robustness of LoS MIMO systems over varying Tx-Rx distances and tilts, by leveraging the flexibility in the placement of the antennas and configuration of the array geometry [70].

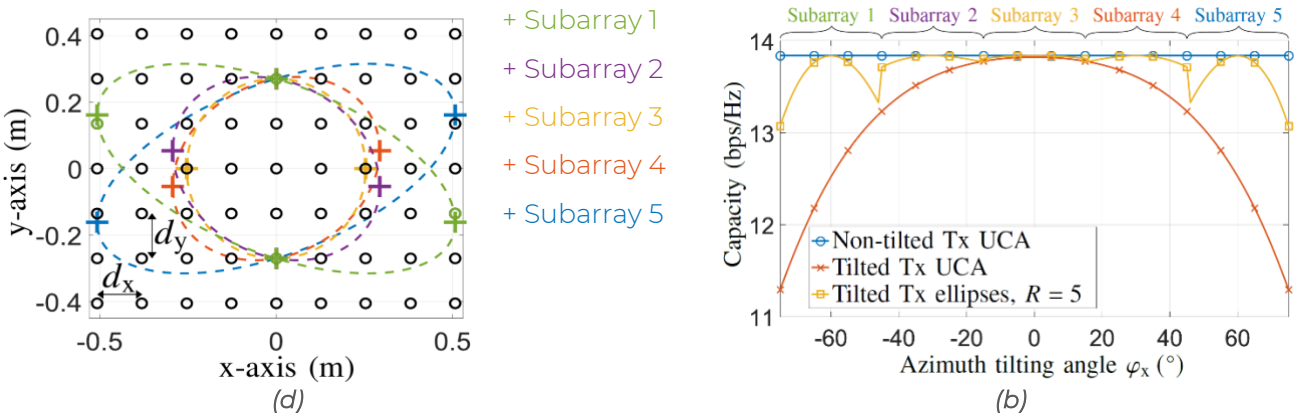



*Figure 5. (a) Example of R = 5 subarrays in different colours that can compensate different tilts at a fixed Tx-Rx distance of 53.8 m. A grid of antennas can also be used to approximate different subarrays. (b) Performance enhancement by selecting different subarrays at different tilting angles*

## 3.5 Multiuser MIMO

While the capacity and optimal transmit strategy of a point-to-point MIMO link with given array geometries is well known, the situation is more complicated for multiuser MIMO channels in a network context (known as "network MIMO"). Special cases, like the MIMO uplink and downlink channels (specifically, non-degraded Gaussian multiple access and broadcast channels) are well understood. For general interference channels, though, the capacity region (and thus the optimal transceiver) is unknown. In a (possibly meshed and cell free) network of access points, there are many degrees of freedom with regard to the choice of distributed access points that are used for transmitting and receiving information, e.g., the possible cooperation between the nodes, the availability and exchange of CSI, the casting type (unicast or multicast [35]), the link type (sidelink or infrastructure), as well as the allocation of resources (scheduling). MIMO is always at the core of such design choices and must be considered.

A fundamental problem in this context is the joint optimization of multiuser MIMO precoders and combiners, along with power allocation and scheduling. The scheduling ensures that the users are orthogonalized in frequency and time domains to avoid interference, while the power control ensures that the desired user rates are achieved in an energy-efficient manner. Such joint optimization is generally NP-hard, and the design of efficient near-optimal solutions is still an open challenge. This is particularly true for distributed cell-free solutions, as discussed in the following

section, where scalability is a major issue. Some solutions to perform the precoding locally are provided in [36].

## 3.6 Cell-free massive MIMO

Network densification is an important factor in realizing 6G services with extreme area spectral efficiency and low latency [37]. In the past, this has driven the development of CRAN and small cell architectures in combination with decentralized MIMO connectivity. However, such architectures still essentially follow the conventional cell-centric design philosophy. Above a certain deployment density, the cell-based approach becomes inefficient due to more and more frequent cell handovers and signalling bottlenecks, which means that the achievable area capacity reaches a plateau where further densification does not lead to corresponding gains [38]. This motivates a "cell-free" architecture, where the UE is moving through an array of radio units which are geographically distributed over the area. Cell-free massive MIMO combines the advantages of distributed systems and massive MIMO. The concept removes cells and cell boundaries with its fundamental idea to deploy a large number of distributed access points (APs) that are connected to a central processing unit (CPU) to serve all users in a wide area. Compared to conventional co-located massive MIMO, cell-free networks offer more uniform connectivity to all users - thanks to the macro diversity gain obtained from the distributed antennas.

In order to allow scalability, the user-centric dynamic cooperation clustering (DCC) scheme was introduced [39], where each UE can connect to nearby APs in a fully flexible manner. In [40], a less restrictive scalability was defined to include more systems, and several clustering approaches and resource allocation techniques were compared. The performance analysis of cell-free massive MIMO systems for different fading channel models [41][42] yields the general conclusion that it can achieve a huge performance gain in a variety of scenarios. Also, the energy efficiency of cell-free massive MIMO is shown to improve by approximately ten times compared to cellular massive MIMO [43][44]. Therefore, cell-free massive MIMO has become one of the most promising technologies in 6G wireless networks and has attracted extensive research interests from both academia and industry [45]. In the context of cell-free MIMO, well-established system procedures, like CSI acquisition, or pilot design, need to be revisited. A central challenge is to find a good compromise between centralized designs and scalable distributed designs. Good recent overviews are given, e.g., in [46][47].

## 3.7 Reconfigurable intelligent surfaces

Reconfigurable intelligent surfaces (RIS) are emerging as a potential key technology for beyond 5G systems. RIS represent low-cost wireless planar structures with *reconfigurable* properties for reflection, refraction, transmittance and absorption of impinging electromagnetic waves, which may be implemented by either purely passive or by (some) active components plus a corresponding controller for their configuration. Note that RIS is also known under the term "Intelligent Reflecting Surfaces" (IRS), which was introduced firstly, but has been replaced by the new term RIS, since their properties now go beyond the sole functionality of reflecting an impinging signal.

The works [48][49] tackled the channel estimation problem for RIS; nonetheless, the required training overhead scales proportionally to the number of RIS antenna elements, limiting the applicability of these schemes in practice. This is due to the fact that an RIS is very likely to have 100s or even 1000s of antennas. Later works, e.g., [50][51] have developed channel estimation schemes with limited overhead. In particular, [51] proposes to group RIS antennas together and to use fixed DFT coefficients for each RIS antenna group for channel probing, while a linear MMSE estimator exploiting the channel's spatial correlations is used at the UE to estimate the RIS channel. This has the effect of reducing the effective RIS channel dimension.

Going beyond the work in [51], which uses fixed RIS configuration parameters for the channel estimation, we propose in [52] to optimize those RIS configuration parameters with the aim to

minimize the channel estimation MSE or sum MSE in a system with multiple RIS. The optimization exploits spatial correlations of the channel between RIS and UE as side information and is based on an alternating optimization and projected gradient descent framework. The final result is the forming of RIS beams for channel estimation that direct the pilot signals into the eigenspace of the RIS channel's spatial correlation matrix containing highest power. Numerical results show superior channel estimation and data rate performance compared to the method in [51].

The circuit theoretic models discussed in Section 3.3 can also be applied to systems with RIS [53]. A circuit model is well suited to characterize the tuning circuits of the RIS elements (e.g. based on PIN diodes), as well as for the mutual coupling of the RIS unit cells. Such a model shows that when the phase of the unit cells is tuned, not only the phase of the reflection coefficient changes, but also its amplitude. Further, it enables to characterize the behaviour of the RIS more accurately.

The use of RIS in industrial indoor scenarios is highly promising, since favourable deployment of multiple RIS along the ceiling of a factory hall can provide LoS-like links to a UE positioned anywhere on the shop floor, thus yielding a homogeneous illumination inside the factory hall with a limited number of access points. The homogenous illumination is of supreme importance for facilitating industrial IoT communication, characterized by data transmission with high reliability under low latency constraints, i.e., URLLC type of service. In [54], we have investigated the achievable performance of industrial IoT services in RIS-enhanced factory deployments in the cm- and mm-wave frequency bands. It has been shown therein that the use of RIS can provide the homogeneous illumination for the whole range of frequency bands considered, given that the RIS are properly dimensioned, while capacity gains become more prominent with increasing the carrier frequency. Moreover, the use of (mainly) passive components in RIS and the comparatively inexpensive materials needed for RIS production allow for greener solutions of the wireless infrastructure in factories to yield homogeneous illumination, compared to a deployment based on (active) access points only. This problem has been subject of our investigation in [55], where a carbon footprint analysis has been carried out for an RIS-enhanced deployment compared to an infrastructure of access points only, yielding a reduction of the carbon footprint (considering production as well as use phase) by up to 66% (factor 3).

## 3.8 Security, resilience and reliability

Related to next generation MIMO, the physical layer security concepts, which exploit the spatial domain to generate novel security paradigms, are of major interest. Security and privacy issues raise increased interest, because other than inheriting vulnerabilities from the previous generations, 6G has new threat vectors from new radio technologies, such as the exposed location of radio stripes in ultra-massive MIMO systems and attacks against pervasive intelligence. Physical layer protection, deep network slicing, quantum-safe communications, artificial intelligence (AI) security, platform-agnostic security, real-time adaptive security, and novel data protection mechanisms such as distributed ledgers and differential privacy are the top promising techniques to mitigate the attack magnitude and personal data breaches substantially. The survey [56] identifies security preservation technologies in the physical layer, the connection layer, and the service layer as the pillars of 6G networks. In the overview paper [57], PHY security is explicitly mentioned as a sixth generation (6G) enabling technology (quote: "The strongest security protection may be achieved at the physical layer").

Also, Massive MIMO is beneficial for Secret Key Generation. Channel reciprocity-based key generation is an emerging physical layer-based technique to establish secret keys between devices [58]. A number of technical challenges in the channel reciprocity-based secret key generation driven by different duplex modes, massive MIMO and mmWave communications, and prototypes in IoT networks were approached in recent studies. An algorithm to generate pairwise secret keys in massive multi-user MIMO networks is derived in [59], which exploits the spatially correlated structure of the underlying massive MIMO channels. In [60], a two-band multiple-antenna loop-back key

generation algorithm is presented which solves the major issues of imperfect channel reciprocity, nearby attack, and high temporal auto-correlation.

## 3.9 Energy and cost efficiency

Energy and cost efficiency are key design principles for future MIMO implementations (see e.g. [61]). At the transceiver level, this can be achieved by hybrid analog-digital design and low-resolution data converters (or constant envelope signalling), as discussed above. In the future, energy efficiency is expected to become increasingly important in the context of network MIMO. For example, distributed MIMO designs based on local measurements and signal processing require less signalling overhead (and thus energy consumption) than centralized processing of all signals. However, distributed designs typically come at the cost of reduced data throughput and it is important to strike the right balance. Moreover, realistic, up-to-date energy models are required to properly evaluate all factors involved. While most studies focus on the energy radiated by the antennas, the bigger part of the total energy budget is actually consumed by the hardware (e.g., coolers and circuit energy consumption) [62]. Intelligent adaptation based on learning techniques can help the system self-optimize for energy efficiency. For example, parts of the network infrastructure can be dynamically switched off and on (sleep mode), depending on the traffic and service requirements. Furthermore, other design parameters can be considered, including renewable resources and energy harvesting as an integral part of the network.

## 3.10 Conclusion and connection to other topics

MIMO-based technologies have made a significant contribution to the success of 4G and 5G. But MIMO has not yet reached the limits of its full potential. There are still many new opportunities and challenges on the road to 6G and beyond. Especially the system-wide consideration of MIMO in a network context opens up a variety of new possibilities, depending on architectural assumptions (e.g. CRAN vs meshed), fronthaul-backhaul capabilities, frontend limitations, as well as the considered frequency band.

For THz and sub-THz, discussed in Section 1, realizing distributed coherent MIMO is a challenge. Specifically, THz MIMO would require extremely precise timings and synchronization for fully coherent base station coordination. This makes non-coherent combining techniques an attractive alternative.

The MIMO principle is also closely related to Non-Orthogonal Multiple Access (NOMA), which has been extensively studied for 5G [63], and discussed in Section 2. However, NOMA needs to be reinvestigated for a user-centric cell-free MIMO architecture. This holds in particular for novel grant-free or unsourced random-access approaches [64][65], that are required to support the huge number of devices in the future. Also, rate-splitting approaches, which are well-known from information theory [66], have recently regained significant interest for MIMO downlink communication [67] and other applications. In this case, the simultaneous transmission of common and private messages requires efficient algorithms for joint multicast and unicast beamforming. This may also be combined with advanced nonlinear precoding methods based on dirty paper coding [68][69].

The term "Next generation MIMO" stands for the development of spatial signal processing in a system context. In this section, some aspects have been discussed, but the selection is in no way

complete. More aspects will be discussed in other sections, e.g., the use of MIMO for Integrated Sensing and Communications in Section 4.

## 3.11 References


[1] E. Björnson, E. G. Larsson and T. L. Marzetta, "Massive MIMO: Ten myths and one critical question," in IEEE Communications Magazine, vol. 54, no. 2, pp. 114-123, February 2016.

[2] W. Choi et al., "Downlink performance and capacity of distributed antenna systems in a multicell environment", IEEE Trans. Wireless Commun., vol. 6, no. 1, Jan. 2007.

[3] You X., Wang D., Wang J. (2021) Massive Distributed MIMO and Cell-Free Systems Under Pilot Contamination. In: Distributed MIMO and Cell-Free Mobile Communication. Springer, Singapore.

[4] J. Jeon et al., "MIMO Evolution toward 6G: Modular massive MIMO in low-frequency bands," in IEEE Communications Magazine, vol. 59, no. 11, pp. 52-58, Nov. 2021.

[5] ETSI, "5G; study on channel model for frequencies from 0.5 to 100 Ghz (3GPP TR 38.901 version 16.1.0 Release 16)," 2020.

[6] T. Jiang et al., "3GPP standardized 5G channel model for IIoT scenarios: A survey," in IEEE Internet of Things Journal, vol. 8, no. 11, pp. 8799-8815, 1 June1, 2021.

[7] METIS 2020, "METIS channel model," METIS2020, Tech. Rep., Deliverable D1.4 v3, July 2015.

[8] K. Haneda, et al., "Measurement results and final mmMagic channel models," Deliverable D2.2, May 2017.

[9] S. Sun, G. R. MacCartney, and T. S. Rappaport, "A novel millimeter-wave channel simulator and applications for 5G wireless communications," in 2017 IEEE International Conference on Communications (ICC). IEEE, May 2017.

[10] S. Jaeckel, L. Raschkowski, K. Börner, and L. Thiele, "QuaDRiGa: A 3-D multi-cell channel model with time evolution for enabling virtual field trials,", IEEE Transactions on Antennas and Propagation, vol. 62, no. 6, pp. 3242–3256, Jun 2014.

[11] "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," in IEEE Std 802.11-2020 (Revision of IEEE Std 802.11-2016), vol., no., pp.1-4379, 26 Feb. 2021.

[12] R. J. Weiler, et al., "Quasi-deterministic millimeter-wave channel models in MiWEBA," EURASIP Journal on Wireless Communications and Networking, vol. 2016, no. 1, Mar 2016.

[13] J. W. Wallace and M. A. Jensen, "Mutual coupling in MIMO wireless systems: A rigorous network theory analysis," IEEE Trans. Wireless Commun., vol. 3, no. 4, pp. 1317–1325, Jul. 2004.

[14] C. Waldschmidt, S. Schulteis, and W. Wiesbeck, "Complete RF system model for analysis of compact MIMO arrays," IEEE Trans. Veh. Technol., vol. 53, no. 3, pp. 579–586, May 2004.

[15] M. T. Ivrlač and J. A. Nossek, "Toward a circuit theory of communication," IEEE Trans. Circuits Syst. I, vol. 57, no. 7, pp. 1663–1683, Jul. 2010.

[16] T. Laas, J. A. Nossek, S. Bazzi, and W. Xu, "On reciprocity in physically consistent TDD systems with coupled antennas," IEEE Trans. Wireless Commun., vol. 19, no. 10, pp. 6440–6453, Oct. 2020.

[17] D. Nie, B. M. Hochwald, and E. Stauffer, "Systematic design of large-scale multiport decoupling networks," IEEE Trans. Circuits Syst. I, vol. 61, no. 7, pp. 2172–2181, Jul. 2014.

[18] K. F. Warnick, B. Woestenburg, L. Belostotski, and P. Russer, "Minimizing the noise penalty due to mutual coupling for a receiving array," IEEE Trans. Antennas Propag., vol. 57, no. 6, pp. 1634–1644, Jun. 2009.

[19] Y. Hassan, “Compact multi-antenna systems: Bridging circuits to communications theory,” Dr. sc. dissertation, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland, Mar. 2018.

[20] B. Lehmeyer, “Receiver and transmitter topologies,” Dr.-Ing. dissertation, Technical University of Munich (TUM), Munich, Germany, Aug. 2018.

[21] J. Kornprobst, et al. “Compact uniform circular quarter-wavelength monopole antenna arrays with wideband decoupling and matching networks,” IEEE Trans. Antennas Propag., vol. 69, no. 2, pp. 769–783, Feb. 2021.

[22] E. Björnson and E. G. Larsson, “Digital millimetre wave beamforming for 5G terminals”, http://www.massive-mimo.net/, 2020.

[23] A. Mezghani and J. A. Nossek, "On ultra-wideband MIMO systems with 1-bit quantized outputs: Performance analysis and input optimization," 2007 IEEE International Symposium on Information Theory, 2007.

[24] A. Mezghani and J. A. Nossek, "Power efficiency in communication systems from a circuit perspective," 2011 IEEE International Symposium of Circuits and Systems (ISCAS), 2011.

[25] A. Mezghani, R. Ghiat and J. A. Nossek, "Transmit processing with low resolution D/A-converters," 2009 16th IEEE International Conference on Electronics, Circuits and Systems - (ICECS 2009), 2009.

[26] L. Chu, F. Wen, L. Li and R. Qiu, "Efficient nonlinear precoding for massive MIMO downlink systems with 1-bit DACs," in IEEE Transactions on Wireless Communications, vol. 18, no. 9, pp. 4213-4224, Sept. 2019.

[27] A. Li, F. Liu, C. Masouros, Y. Li and B. Vucetic, "Interference exploitation 1-bit massive MIMO precoding: A partial branch-and-bound solution with near-optimal performance," in IEEE Transactions on Wireless Communications, vol. 19, no. 5, pp. 3474-3489, May 2020.

[28] H. Jedda, A. Mezghani, A. L. Swindlehurst and J. A. Nossek, "Quantized constant envelope precoding with PSK and QAM signaling," in IEEE Transactions on Wireless Communications, vol. 17, no. 12, pp. 8022-8034, Dec. 2018.

[29] S. Jacobsson, G. Durisi, M. Coldrey, T. Goldstein and C. Studer, "Quantized precoding for massive MU-MIMO," in IEEE Transactions on Communications, vol. 65, no. 11, pp. 4670-4684, Nov. 2017.

[30] K. Roth and J. A. Nossek, "Achievable rate and energy efficiency of hybrid and digital beamforming receivers with low resolution ADC," in IEEE Journal on Selected Areas in Communications, vol. 35, no. 9, pp. 2056-2068, Sept. 2017.

[31] K. Roth, H. Pirzadeh, A. L. Swindlehurst and J. A. Nossek, "A comparison of hybrid beamforming and digital beamforming with low-resolution ADCs for multiple users and imperfect CSI," in IEEE Journal of Selected Topics in Signal Processing, vol. 12, no. 3, pp. 484-498, June 2018.

[32] F. Askerbeyli, W. Xu and J. A. Nossek, "Energy efficiency comparison of digital and hybrid precoding in 1-bit mmWave massive MIMO," in 2023 IEEE 97th Vehicular Technology Conference (VTC-Spring), 2023.

[33] M. Palaiologos, M. H. C. Garcia, R. A. Stirling-Gallacher and G. Caire, "Tilt compensation in UCA-based LoS MIMO systems with antenna selection," *2022 IEEE 33rd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Kyoto, Japan, 2022, pp. 1308-1313.

[34] M. Palaiologos, M. H. C. n. García, R. A. Stirling-Gallacher and G. Caire, "Tilting in UCA-based LoS MIMO systems," in *IEEE Wireless Communications Letters (Early Access)*, June 2023.

[35] M. Sadeghi, E. Björnson, E. G. Larsson, C. Yuen and T. L. Marzetta, "Max–Min fair transmit precoding for multi-group multicasting in massive MIMO," in IEEE Transactions on Wireless Communications, vol. 17, no. 2, pp. 1358-1373, Feb. 2018.

[36] G. Interdonato, M. Karlsson, E. Bjornson, and E. G. Larsson, “Local partial zero-forcing precoding for cell-free massive MIMO”, IEEE Trans. Wireless Commun., vol. 19, no. 7, pp. 4758–4774, Jul 2020.

[37] J. Zander and P. Mähönen, "Riding the data tsunami in the cloud: myths and challenges in future wireless access," in IEEE Communications Magazine, vol. 51, no. 3, pp. 145-151, March 2013.

[38] J. G. Andrews, X. Zhang, G. D. Durgin and A. K. Gupta, "Are we approaching the fundamental limits of wireless network densification?" in IEEE Commun. Mag., vol. 54, no. 10, pp. 184-190, Oct. 2016.

[39] E. Björnson and L. Sanguinetti, “Scalable cell-free massive MIMO systems,” IEEE Trans. Commun., vol. 68, no. 7, pp. 4247-4261, Jul. 2020.

[40] D. Prado-Alvarez, D. Calabuig, J. F. Monserrat, S. Bazzi and W. Xu, "Study of clustering solutions for scalable cell-free massive MIMO," in *IEEE Access*, vol. 11, pp. 26703-26711, 2023.

[41] Ö. Özdogan, E. Björnson, and J. Zhang, “Performance of cell-free massive MIMO with Rician fading and phase shifts,” IEEE Trans. Wireless Commun., vol. 18, no. 11, pp. 5299-5315, Nov. 2019.

[42] Z. Wang, J. Zhang, E. Björnson, and B. Ai, “Uplink performance of cell-free massive MIMO over spatially correlated Rician fading channels,” IEEE Commun. Lett., vol. 25, no. 4, pp. 1348-1352, Apr. 2020.

[43] H. Q. Ngo, L. Tran, T. Q. Duong, M. Matthaiou, and E. G. Larsson, “On the total energy efficiency of cell-free massive MIMO,” IEEE Trans. Green Commun. Netw, vol. 2, no. 1, pp. 25-39, 2018.

[44] H. V. Nguyen, V. D. Nguyen, O. A. Dobre, S. K. Sharma, S. Chatzinotas, B. Ottersten, and O. S. Shin, “On the spectral and energy efficiencies of full-duplex cell-free massive MIMO,” IEEE J. Sel. Areas Commun., vol. 38, no. 8, pp. 1698-1718, Aug. 2020.

[45] J. Zhang, S. Chen, Y. Lin, J. Zheng, B. Ai, and L. Hanzo, “Cell-free massive MIMO: A new next-generation paradigm,” IEEE Access, vol. 7, pp. 99878-99888, Sep. 2019.

[46] H. Ammar, R. Adve, S. Shahbazpanahi, G. Boudreau, K. Srinivas. “User-centric Cell-free Massive MIMO Networks: A Survey of Opportunities, Challenges and Solutions”, 2021.

[47] Ö. T. Demir, E. Björnson, L. Sanguinetti et al., “Foundations of user-centric cell-free massive MIMO,” Foundations and Trends® in Signal Processing vol. 14, no. 3-4, pp. 162–472, 2021.

[48] T. L. Jensen and E. De Carvalho, "An optimal channel estimation scheme for intelligent reflecting surfaces based on a minimum variance unbiased estimator," in 2020 IEEE Internat. Conf. on Acoustics, Speech and Signal Proc. (ICASSP), 2020.

[49] Q. U. A. Nadeem, H. Alwazani, A. Kammoun, A. Chaaban, M. Debbah and M. -S. Alouini, "Intelligent reflecting surface-assisted multi-user MISO communication: Channel estimation and beamforming design," in IEEE Open Journal of the Communications Society, vol. 1, pp. 661-680, 2020.

[50] Z. Wang, L. Liu and S. Cui, "Channel estimation for intelligent reflecting surface assisted multiuser communications: Framework, algorithms, and analysis," in IEEE Trans. Wireless Commun., vol. 19, no. 10, pp. 6607-6620, Oct. 2020.

[51] Ö. T. Demir and E. Björnson, “RIS-assisted massive MIMO with multi-specular spatially correlated fading”, in 2021 IEEE GLOBECOM, 2021, pp. 1-6.

[52] S. Bazzi and W. Xu, “IRS parameter optimization for channel estimation MSE minimization in double-IRS aided systems,” IEEE Wireless Commun. Lett., vol. 11, no. 10, pp. 2170-2174, Oct. 2022.

[53] M. Di Renzo, F. H. Danufane, and S. Tretyakov, “Communication models for reconfigurable intelligent surfaces: From surface electromagnetics to wireless network optimization,” Proc. IEEE, vol. 110, no. 9, pp. 1164–1209, Sep. 2022.

[54] M. Schellmann, “Capacity boosting by IRS deployment for industrial IoT communication in cm- and mm-wave bands,” EuCNC 2022.

[55] L. Stobbe, M. Kaiser, M. Schellmann, J. Eichinger, “IRS deployments in future factory: Carbon footprint analysis of different network configurations,” EuCNC and 6G Summit 2023.

[56] V. -L. Nguyen, P. -C. Lin, B. -C. Cheng, R. -H. Hwang and Y. -D. Lin, "Security and Privacy for 6G: A Survey on Prospective Technologies and Challenges," in IEEE Communications Surveys & Tutorials, vol. 23, no. 4, pp. 2384-2428, 2021.

[57] Shakiba-Herfeh M., Chorti A., Vincent Poor H. (2021) Physical Layer Security: Authentication, Integrity, and Confidentiality. In: Le K.N. (eds) Physical Layer Security. Springer, Cham.

[58] G. Li, C. Sun, J. Zhang, E. Jorswieck, B. Xiao, A. Hu, "Physical Layer Key Generation in 5G and Beyond Wireless Communications: Challenges and Opportunities", vol. 21, no, 5, pp. Entropy, 2019. https://www.mdpi.com/1099-4300/22/6/679.

[59] G. Li, C. Sun, E. Jorswieck, J. Zhang, A. Hu, Y. Chen, "Sum Secret Key Rate Maximization for TDD Multi-User Massive MIMO Wireless Networks", IEEE Trans. on Information Forensics and Security, vol. 16, pp.968-982, 2021.

[60] G. Li, Y. Xu, W. Xu, E. Jorswieck, and A. Hu, "Robust Key Generation With Hardware Mismatch for Secure MIMO Communications", IEEE Trans. on Information Forensics and Security, vol. 16, pp. 5264-5278, 2021.

[61] E. Björnson and E. G. Larsson, "How Energy-Efficient Can a Wireless Communication System Become?", 52nd Asilomar Conference on Signals, Systems, and Computers, 2018, pp. 1252-1256.

[62] R. L. G. Cavalcante, S. Stanczak, M. Schubert, A. Eisenblaetter and U. Tuerke, "Toward Energy-Efficient 5G Wireless Communications Technologies: Tools for decoupling the scaling of networks from the growth of operating power," in IEEE Signal Proc. Mag., vol. 31, no. 6, pp. 24-34, Nov. 2014.

[63] 3GPP, TR 38.812, "Study on Non-Orthogonal Multiple Access (NOMA) for NR", Dec. 2018.

[64] Y. Polyanskiy, "A perspective on massive random-access," 2017 IEEE International Symposium on Information Theory (ISIT), 2017, pp. 2523-2527.

[65] A. Fengler, G. Caire, P. Jung, S. Haghighatshoar, "Massive MIMO Unsourced Random Access", arXiv: 1901.00828.

[66] B. Rimoldi and R. Urbanke, "A Rate-Splitting Approach to the Gaussian Multiple-Access Channel," in IEEE Transactions on Information Theory, vol. 42, no. 2, pp. 364-375, March 1996.

[67] Y. Mao, B. Clerckx and V.O.K. Li, "Rate-Splitting Multiple Access for Downlink Communication Systems: Bridging, Generalizing and Outperforming SDMA and NOMA", EURASIP Journal on Wireless Communications and Networking, 2018.

[68] Y. Mao and B. Clerckx, "Beyond Dirty Paper Coding for Multi-Antenna Broadcast Channel with Partial CSIT: A Rate-Splitting Approach," in IEEE Transactions on Communications, vol. 68, no. 11, pp. 6775-6791, Nov. 2020.

[69] M. Y. Şener, R. Böhnke, W. Xu and G. Kramer, "Dirty Paper Coding Based on Polar Codes and Probabilistic Shaping," in IEEE Communications Letters, vol. 25, no. 12, pp. 3810-3813, Dec. 2021.

[70] C. Bencivenni, M. Coldrey, R. Maaskant and M. V. Ivashina, "Aperiodic Switched Array for Line-of-Sight MIMO Backhauling," in *IEEE Antennas and Wireless Propagation Letters*, vol. 17, no. 9, pp. 1712-1716, Sept. 2018.

# 4. Integrated Sensing and Communication (ISAC)

## 4.1 State-of-the-art on Integrated Sensing and Communication

It is widely believed that next-generation mobile radio systems will be designed for simultaneous communication and sensing, by exploiting the sensing capabilities of radio frequency (RF) signals in the mmWave and THz bands. In fact, the increased connectivity and bandwidth offered by the 6G technology will enable cooperative devices to exploit data fusion strategies to infer accurate positions of passive targets for applications including traffic monitoring (i.e., traffic, vehicles monitoring, pedestrian detection), collision avoidance between autonomous guided vehicles (AGVs) (see uses cases below), assisted living, as well as accurate localization and tracking of passive objects, to overcome the necessity to equip all the targets with active systems, and human-machine interface [1]. Sensing takes many forms, ranging from detecting the presence of an object to its position, speed, and specific micro-Doppler signature, up to imaging of an environment. Consequently, there is an increasing demand for systems exhibiting both sensing and communications capabilities. However, sensing and communications have traditionally been performed separately by different entities, functions, and/or frequency bands [2][3]. There are several possible options for integrated sensing and communication (ISAC), which include:

- Integration at high level: where the sensing and communication systems are separated and information is exchanged to help in some way the mutual functioning.
- Integration at scheduling: that is sensing and communications signals are multiplexed in time, frequency, and space, enabling the two functions to share the spectrum and partially share hardware resources.
- Full integration: in this case, sensing and communication systems are fully integrated and share the hardware and the frequency band. This approach is also known as joint sensing and communication (JSC); it exploits the waveforms transmitted by a communication network to perform sensing. To fully integrate sensing and communications, wireless systems will be designed to support both functions together, using the same spectral resources and hardware and thereby reducing cost, power consumption, latency, and size.

ISAC can encompass both sidelink communications such as V2V for vehicular networks (with great applications to autonomous driving) and also complicated mobile/cellular networks with multiple nodes, which can potentially revolutionize the current communication-only mobile networks [4].

ISAC at mmWaves may reach significant performance levels by exploiting multiple antennas in transmission and reception [5][6]. Indeed, the potential use of large-scale antenna arrays due to shorter wavelengths [7], and thanks to their larger bandwidths, compared to those available in the traditional cellular band (up to 400 MHz), mmWaves could be very advantageous not only for communication but also for sensing [8][9][10]. Furthermore, MIMO technology can provide high-capacity links to the users (e.g., through spatial multiplexing), while array processing at the sensing receiver can perform accurate direction of arrival (DoA) estimation [11]. Furthermore, by exploiting sub-THz and THz bands technologies, the potential gains to sensing are enormous in terms of resolution and accuracy, especially for short-range applications.

To maximize the advantages of such a fully integrated solution for 6G, it is essential to investigate different approaches and address several key technical challenges [12]: i) integrate sensing into existing communication waveforms (e.g., orthogonal frequency division multiplexing (OFDM) [13]); ii) optimize power and spectral allocation by suitable transmission parameters for the combined use of sensing and communications [14]; iii) optimize alternative waveforms for ISAC, e.g., orthogonal

chirp division multiplexing (OCDM) [15], affine frequency division multiplexing (AFDM) [16], or orthogonal time-frequency-space (OTFS) modulation [17], to name a few; and iv) understand the requirements and performance of different system setups. For monostatic arrangements, it is important to understand the requirements in terms of dynamic range and tolerance to self-interference caused by the direct coupling of the transmitter and the receiver. For this reason, on top of basic passive RF isolation and radar-domain digital suppression methods, efficient active RF and digital self-interference cancellation methods are necessary [18]. On the contrary, for bistatic and multistatic setups, which do not require full-duplex operations, it is crucial to understand the impact of signalling overhead, sensors' position, and synchronization.

To conclude, ISAC represents a key innovation for 6G that facilitates new applications and potentially revolutionizes the current mobile network concept. However, this poses challenges that will require considerable research effort.

## 4.2 Current 3GPP Standardization Status and 6G Pre-standardization Activities

6G standardization at 3GPP is expected to start in late 2025 and the goal will be to support a fully native, integrated and optimized ISAC system. This may potentially use enhanced waveforms, new frequency bands (including the 7-24 GHz, sub-THz and THz bands), while supporting new optimized transmission modes and procedures.

In the meantime however, the 3GPP Service and System Aspects working group 1 (SA1) has completed in 2023 a Study and Work Item on use cases and requirements for future enhancement of the 5G new radio (NR) system to provide ISAC services [19][20][21][22]. This resulted in identifying 32 use case and their requirements for various applications using 5G NR. The study considered 3GPP, non-3GPP and non-RF sensing techniques and paved the way for future 3GPP 5G NR releases. The identified 32 use cases for ISAC [19] cover a wide range of applications including:

- Object and intruder detection for smart homes, on a highway, for railways, for factory, for predefined secure areas around critical infrastructure.
- Collision avoidance and trajectory tracking of UAVs, vehicles, and AGVs.
- V2x Automotive maneuvering and navigation.
- Public safety search and rescue.
- Rainfall monitoring and flooding.
- Health and sports monitoring.

The SA1 work on use cases has been followed by a RAN study on ISAC channel models [23], which started in Jan 2024 and is scheduled to be completed at the end of 3GPP release 19.

In addition to the ongoing 3GPP standardization work to support ISAC in the 5G NR, key industrial organizations (5GAA and 5G ACIA) have started ISAC work items and an industrial study group (ISG) on ISAC [24] was established in 2023, which all focus on pre-standardization activities for 6G ISAC.

## 4.3 Key 6G Use cases

### 4.3.1 Vehicular scenarios

One application of high relevance for the next-generation mobile networks is intelligent transportation, which refers to services to improve traffic safety and increase efficiency, e.g., vehicle-

to-vehicle (V2V) and vehicle-to-everything (V2X) communication. Furthermore, high accuracy position and velocity estimations are critical parameters for safety applications; for instance, they can assist the driver when executing safety-related manoeuvres. Furthermore, using ISAC, environmental knowledge of both passive and active objects in the vicinity (which may block or reflect the desired communications signal) can be obtained together with the communications link. This information helps improve the reliability of the communication link to the target vehicle via beam management and resource allocation procedures. This scenario is split into two use cases: V2x Uu link based ISAC and V2V based ISAC.

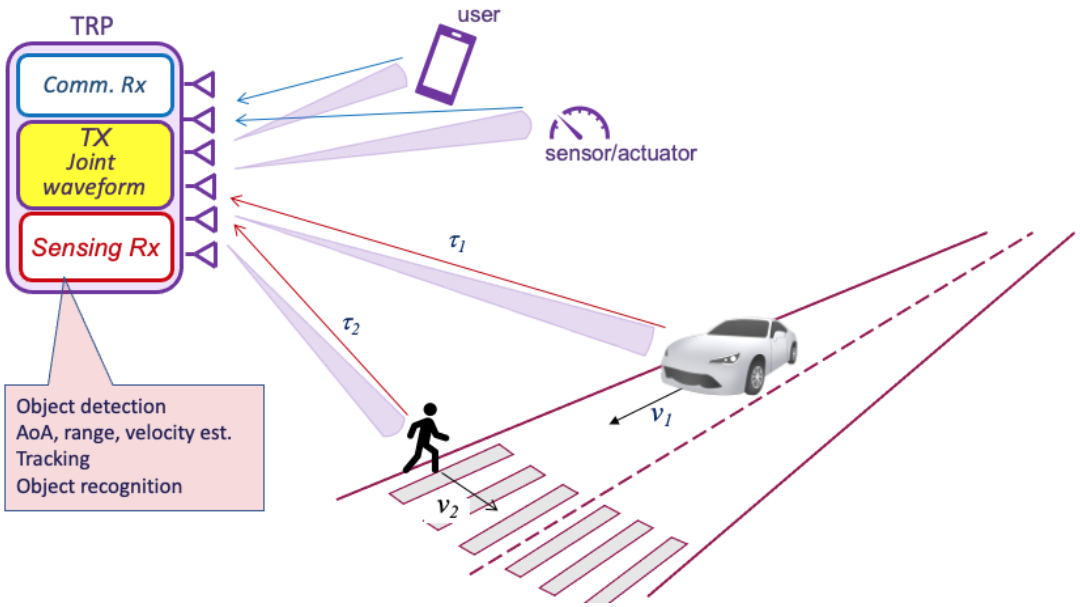



*Figure 6: Joint sensing and communication in a vehicular scenario (monostatic configuration)*

#### 4.3.1.1 V2x Uu link based ISAC

The V2x Uu link can provide a range of V2x services, including high-definition map updating, traffic and emergency notifications, and support for tele-operated driving. This link could be re-used to support sensing in a monostatic or multi-static arrangement. A monostatic arrangement is shown in Figure 6. The transmitted and received signals are jointly processed by the serving base station (access point or RSU) for such configuration. This requires full-duplex operation at the base station. A multi-static arrangement is shown in Figure 7. This configuration has some advantages, including the ability to use the traditional communication waveform (since full duplex is not required) and the capacity to illuminate objects (passive and active) from multiple angles. This latter feature guarantees spatial diversity that can facilitate the detection and classification of objects. Multi-static sensing requires that the transmitters and receivers are in different positions, synchronized, and can process and combine the received signals by exploiting the communication between them.

While general communications can deal with hundreds of ms delays, autonomous vehicle applications require delays in the order of tens of ms [25][26][27]. In such scenarios, sensing should provide robust, high-resolution obstacle detection in the order of a few decimetres.

The upcoming 6G technology, leveraging both massive MIMO antenna arrays and the mmWave and THz spectrum, is expected to address future autonomous vehicle network requirements. Additionally, large-scale antenna arrays can form pencil-shaped beams that accurately point to directions of interest by compensating for path loss (sensing loss decays with the fourth power of distance in free space) and improving DoA estimation accuracy. Therefore, it would make sense to equip vehicles or road infrastructure sensors with ISAC systems like those sketched in Figure 6 with a monostatic base station (BS) or in Figure 7 with a multi-static arrangement employing sensor

fusion. However, several issues need to be investigated in that context, such as specific mmWave channel models and constraints, given that fast-changing scenarios accompany high mobility.

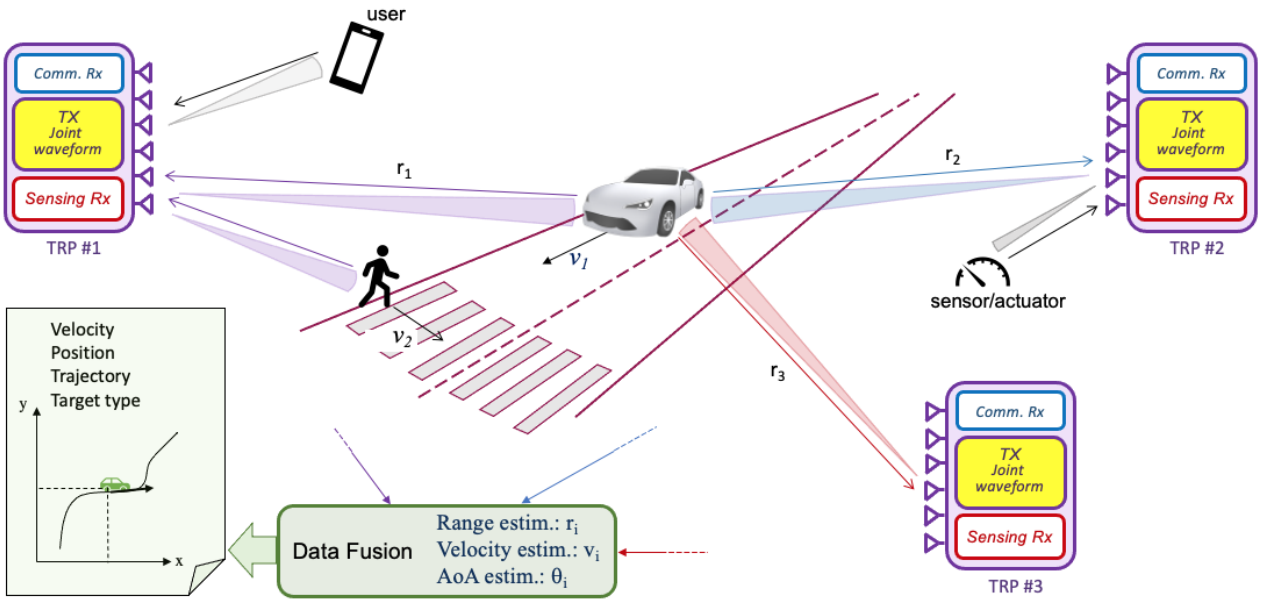



*Figure 7: Joint sensing and communication in a vehicular scenario (multi-static/distributed configuration)*

#### 4.3.1.2 V2V based ISAC

V2x sidelink communications represent another important use case. Here the sidelink can be re-used to provide sensing of nearby vehicles and pedestrians and can be used to complement the existing sensors (i.e., cameras, radars, Lidar, etc.) on the vehicle. The relative location of the closest objects to the vehicle in the line of sight is particularly important for autonomous driving. As before, different levels of integration can be used for this use case. However, as this is a monostatic scenario, full-duplex technology on the vehicle performing the sensing is necessary for the full integration of communication and sensing in the sidelink.

### 4.3.2 Industrial scenarios

ISAC is also anticipated to facilitate the real-time acquisition of object positions in industrial settings, resulting in heightened environmental awareness, while simultaneously reducing the expenses and maintenance associated with separate sensing and communication infrastructures. The benefits of ISAC in industrial scenarios are twofold: safety and network reliability. Regarding safety, sensing enables the identification of uncooperative objects (e.g., carts and forklifts) and obstacles in prohibited areas or workers in close proximity to machinery. This enables accurate detection and localization, allowing the implementation of necessary accident-prevention measures. In terms of network reliability, the industrial propagation environment is complex and susceptible to communication disruptions caused by obstructions such as large metallic objects (e.g., machinery or robots). By harnessing sensing-assisted communication systems enabled by ISAC, this challenging issue can be mitigated. Sensing information can be utilized to assist the communication link through techniques like beam management. Specifically, sensing capabilities can be employed to detect and pinpoint incoming obstacles, estimate their speed and direction, and predict their future trajectory several seconds in advance. This allows the network sufficient time to establish a backup link before the existing one experiences an outage.

Research work for ISAC in industrial settings has already started and is the subject of various EU funded projects including the TIMES project [28], funded by the Horizon JU SNS 2022 program.

## 4.4 Ongoing research and open problems

ISAC research is currently focusing on multiple directions.

Machine learning-based approaches based on soft information for the localization of things have been proposed in [3] for accurate positioning to overcome the limitations of classical techniques. In particular, soft information encapsulates all the information from measurements and contextual data at the UE at a given position, including sensing measurements (e.g., using radio signals), digital maps, and UE profiles.

The impact of bistatic configuration on sensing capabilities has been investigated in [29] and [31]. These studies analyse a multi-beam JSC system in terms of achievable sensing coverage. The optimization of beamforming for bistatic MIMO sensing is discussed in [30]. In [32], the principle of cooperative passive coherent location (CPCL) is introduced, where a passive bistatic configuration exploits the received communication signal as a reference for correlation. This way, CPCL effectively reuses the entire communication payload for target detection.

Recently, [33] proposed the design of new waveforms that combine OFDM and OTFS to enable JSC and [34] presented a low-complexity receiver for target parameter estimation. Additionally, in terms of waveforms, new results on the use of AFDM waveform for ISAC are also presented in [35], which highlights the unique feature of this waveform to reduce complexity for full duplex monostatic sensing. Regarding the impact of the array configuration, antenna selection at the transmitter site is leveraged in [36] towards improving the performance of an ISAC system. For IoT scenarios, communication primarily relies on packet-based transmission, and [37] investigates the impact of packet length on the sensing/communication trade-off. They examine a JSC system under the finite block-length regime. For V2x scenarios and specifically for V2V based ISAC, the combined use of sidelink resources for communications and sensing for different densities of vehicles and numerologies is shown in [40] and [41].

As for sensor fusion strategies (depicted in Fig. 7), tracking algorithms in [38], [39] and [40] are used to fuse local target position estimates from multiple monostatic sensors. Data fusion can enhance sensing performance or maintain the same performance as a single sensor while releasing radio resources for communication. The trade-off between sensing and communication in OFDM-based networks is studied in [41] for multistatic arrangements. [42] proposes an ISAC architecture based on distributed multisensor MIMO, analogous to multiuser MIMO in mobile communication. Furthermore, [43] discusses fundamental requirements for propagation measurement and modelling in the context of distributed MS-MIMO ISAC.

## 4.5 References

[1] T. Wild, V. Braun, and H. Viswanathan, “Joint design of communication and sensing for beyond 5G and 6G systems,” IEEE Access, vol. 9, pp. 30 845–30 857, 2021.

[2] M. Chiani, A. Giorgetti and E. Paolini, “Sensor radar for object tracking,” Proceedings of the IEEE, vol. 106, no. 6, pp. 1022-1041, Jun. 2018.

[3] A. Conti, S. Mazuelas, S. Bartoletti, W. C. Lindsey and M. Z. Win, "Soft information for localization-of-things," Proceedings of the IEEE, vol. 107, no. 11, pp. 2240-2264, Nov. 2019.

[4] F. Liu, Y. Cui, C. Masouros, J. Xu, T. X. Han, Y. C. Eldar, and S. Buzzi, “Integrated sensing and communications: Towards dual-functional wireless networks for 6G and beyond,” preprint, arXiv:2108.07165, 2021.

[5] J. A. Zhang, X. Huang, Y. J. Guo, J. Yuan, and R. W. Heath, “Multibeam for joint communication and radar sensing using steerable analog antenna arrays,” IEEE Trans. Veh. Technol., vol. 68, no. 1, pp. 671– 685, Jan. 2019.

[6] J. A. Zhang, M. L. Rahman, K. Wu, X. Huang, Y. J. Guo, S. Chen, and J. Yuan, "Enabling joint communication and radar sensing in mobile networks – a survey," IEEE Commun. Surveys Tuts., pp. 1–41, 2021.

[7] F. W. Vook, A. Ghosh, and T. A. Thomas, "MIMO and beamforming solutions for 5G technology," IEEE MTT-S International Microwave Symposium (IMS2014), 2014, pp. 1–4.

[8] L. Pucci, E. Matricardi, E. Paolini, W. Xu, and A. Giorgetti, "Performance analysis of joint sensing and communication based on 5G new radio," in Proc. IEEE Globecom Work. on Advances in Network Localization and Navigation (ANLN), Madrid, Spain, Dec. 2021.

[9] L. Pucci, E. Paolini, and A. Giorgetti, "System-Level Analysis of Joint Sensing and Communication based on 5G New Radio," IEEE J. on Selected Areas in Comm., vol. 40, no. 7, pp. 2043-2055, July 2022.

[10] R. Thomä, T. Dallmann, S. Jovanoska, P. Knott, A. Schmeink, "Joint communication and radar sensing: An overview," European Conference on Antennas and Propagation (EuCAP), pp. 1-5, 2021.

[11] Y. L. Sit, C. Sturm, J. Baier, and T. Zwick, "Direction of arrival estimation using the MUSIC algorithm for a MIMO OFDM radar," in Proc. IEEE Radar conference, pp. 0226–0229, 2012.

[12] Q Wang, A. Kakkavas, X. Gong, R. A. Stirling-Gallacher, "Towards Integrated Sensing and Communications for 6G," in 2nd IEEE Int. Symposium on Joint Communications and Sensing (JC&S), pp. 1-6, March 2022.

[13] M. Braun, "OFDM radar algorithms in mobile communication networks," Ph.D. dissertation, Karlsruhe Institute of Technology, 2014.

[14] G. Kwon, A. Conti, H. Park and M. Z. Win, "Joint Communication and Localization in Millimeter Wave Networks," in IEEE Journal of Sel. Topics in Signal Proc., vol. 15, no. 6, pp. 1439-1454, Nov. 2021.

[15] L. G. d. Oliveira, M. B. Alabd, B. Nuss, and T. Zwick, "An OCDM radar communication system," in 14th European Conference on Antennas and Propagation (EuCAP), Mar. 2020, pp. 1–5.

[16] A. Bemani, N. Ksairi, and M. Kountouris, "AFDM: A full diversity next generation waveform for high mobility communications," in IEEE Int. Conf. Commun. Work. (ICC Workshops), Jun. 2021, pp. 1–6.

[17] R. Hadani, S. Rakib, M. Tsatsanis, A. Monk, A. J. Goldsmith, A. F. Molisch, and R. Calderbank, "Orthogonal time frequency space modulation," in Proc. IEEE Wireless Communications and Networking Conference (WCNC), pp. 1–6, 2017.

[18] C. Baquero Barneto, T. Riihonen, M. Turunen, L. Anttila, M. Fleischer, K. Stadius, J. Ryynänen, and M. Valkama, "Full-duplex OFDM radar with LTE and 5G NR waveforms: Challenges, solutions, and measurements," IEEE Trans. Microw. Theory Techn., vol. 67, no. 10, pp. 4042–4054, Oct. 2019.

[19] "Feasibility Study on Integrated Sensing and Communication (Release 19)", 3GPP TR 22.837 V19.3.0 (2024-03)

[20] "Service requirements for Integrated Sensing and Communication (Release 19)," 3GPP TR 22.137 V19.1.0 (2024-03).

[21] "RAN Chair's Summary of Rel-19 Workshop", 3GPP RWS-230488, 3GPP RAN Plenary Workshop, June 2023.

[22] "Input from TSG SA Rel-19 Workshop: Consolidated SA WG2 Rel-19 Topics for moderated discussions", 3GPP SP-230759, 3GPP SA Plenary Workshop, June 2023.

[23] RP-234069, "New SID: Study on channel modelling for Integrated Sensing and Communication (ISAC) for NR," Nokia, Nokia Shanghai Bell, Dec. 2023.

[24] https://www.etsi.org/committee/isac

[25] K. V. Mishra, A. Zhitnikov, and Y. C. Eldar, “Spectrum sharing solution for automotive radar,” in Proc. IEEE Vehicular Technology Conference (VTC Spring), pp. 1–5, 2017.

[26] F. Liu, C. Masouros, A. P. Petropulu, H. Griffiths, and L. Hanzo, “Joint radar and communication design: Applications, state-of-the-art, and the road ahead,” IEEE Trans. on Comm., vol. 68, no. 6, pp. 3834–3862, 2020.

[27] S. H. Dokhanchi, B. S. Mysore, K. V. Mishra, and B. Ottersten, “A mmWave automotive joint radar-communications system,” IEEE Trans. Aerosp. Electron. Syst., vol. 55, no. 3, pp. 1241–1260, Jun. 2019.

[28] http://www.times6g.eu/

[29] L. Pucci, E. Matricardi, E. Paolini, W. Xu, and A. Giorgetti, “Performance of a 5G NR-based Bistatic Joint Sensing and Communication System," IEEE Int. Conf. on Comm. (ICC) – Workshop, pp. 73-78, May 2022.

[30] T. Laas, R. Boehnke, and W. Xu, “Optimal Beamforming for bistatic MIMO sensing,” arXiv:2405.01197, 2024.

[31] F. Zabini, E. Paolini, W. Xu, and A. Giorgetti, ‘’Joint Sensing and Communication with Multiple Antennas and Bistatic Configuration,” IEEE Int. Conf. on Comm. (ICC), Rome, Italy, May 2023.

[32] R.S. Thomä, C. Andrich, G. Del Galdo, M. Döbereiner, M. Hein, M. Käske, G. Schäfer, S. Schieler, C. Schneider, A. Schwind, P. Wendland, “Cooperative Passive Coherent Location – A Promising 5G Service to Support Road Safety,” IEEE Communications Magazine, Vol.: 57, Issue: 9, September 2019, pp. 86-92.

[33] D. Tagliaferri et al., "Integrated Sensing and Communication System via Dual-Domain Waveform Superposition," in IEEE Transactions on Wireless Communications, vol. 23, no. 5, pp. 4284-4299, May 2024.

[34] T. Bacchielli, L. Pucci, E. Paolini and A. Giorgetti, "Performance Analysis of a Low-Complexity OTFS Integrated Sensing and Communication System," *2023 IEEE 98th Vehicular Technology Conference (VTC2023-Fall)*, Hong Kong, Hong Kong, 2023, pp. 1-6.

[35] A. Bemani, N. Ksairi and M. Kountouris, "Integrated Sensing and Communications With Affine Frequency Division Multiplexing," in *IEEE Wireless Communications Letters*, vol. 13, no. 5, pp. 1255-1259, May 2024,

[36] M. Palaiologos, M. H. Castaneda, T. Laas, R. A. Stirling-Gallacher, G. Caire, “Joint Antenna Selection and Covariance Matrix Optimization for ISAC Systems”, accepted for publication at 2024 IEEE International Conference on Communications Workshops (ICC Workshops).

[37] F. Zabini, E. Paolini, W. Xu, and A. Giorgetti, "Joint Sensing and Communications in Finite Block-Length Regime,” IEEE Global Comm. Conf. (Globecom), Dec. 2022.

[38] E. Favarelli, E. Matricardi, L. Pucci, E. Paolini, W. Xu, and A. Giorgetti, ”Tracking and Data Fusion in Joint Sensing and Communication Networks,” IEEE Global Comm. Conf. (Globecom) - Workshop, Dec. 2022.

[39] E. Favarelli, E. Matricardi, L. Pucci, E. Paolini, W. Xu, and A. Giorgetti, ‘’Sensor Fusion and Extended Multi-Target Tracking in Joint Sensing and Communication Networks,” IEEE Int. Conf. on Comm. (ICC), Rome, Italy, May 2023.

[40] E. Favarelli, E. Matricardi, L. Pucci, E. Paolini, W. Xu, and A. Giorgetti, "Map Fusion and Heterogeneous Objects Tracking in Joint Sensing and Communication Networks," European Radar Conf. (EuRAD), Berlin, Germany, Sept. 2023.

[41] E. Matricardi, L. Pucci, E. Paolini, W. Xu, and A. Giorgetti, “Performance Analysis of a Multistatic Joint Sensing and Communication System,” IEEE Int. Symp. on Personal, Indoor and Mobile Radio Comm. (PIMRC), Toronto, Canada, Sepy. 2023.

[42] R.S. Thomä, T. Dallmann, “Distributed ISAC Systems – Multisensor Radio Access and Coordination,” European Microwave Week (EuMW 2023), Focused Session "Joint Communication and Radar Sensing - a step towards 6G", Berlin, Germany, Sept. 2023.

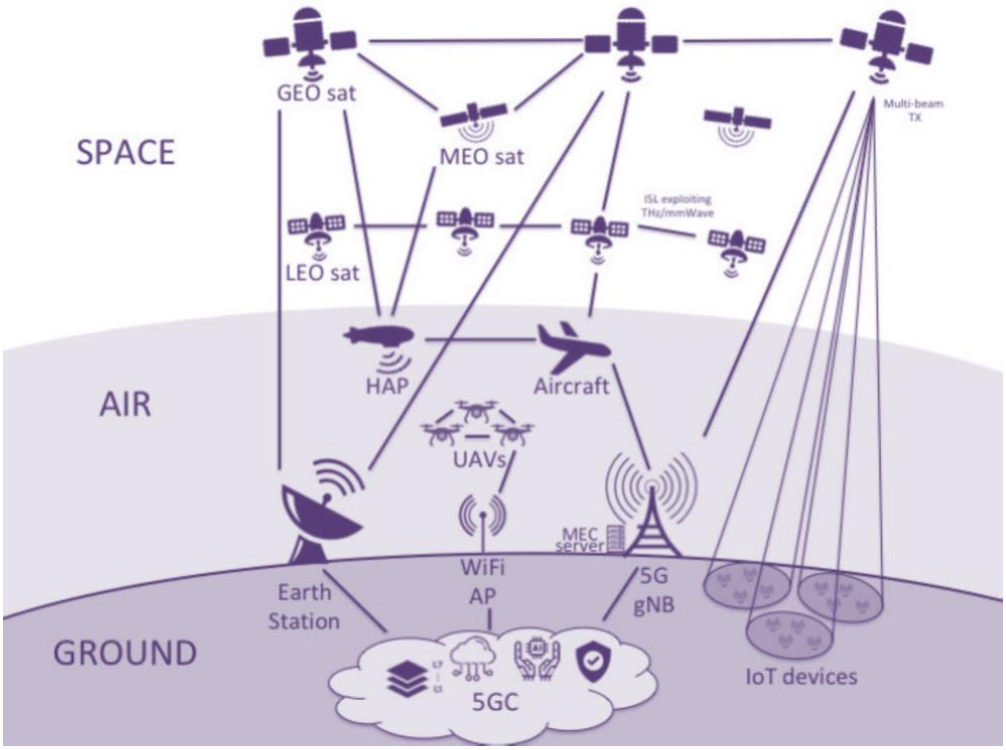
[43] R. S. Thomä, C. Andrich, J. Beuster, H. C. A. Costa, S. Giehl, S. J. Myint, C. Schneider, and G. Sommerkorn, "Characterization of multilink propagation and bistatic target reflectivity for distributed multi- sensor ISAC,” arXiv:2210.11840, 2023.

[44] N. Decarli, S. Bartoletti, A. Bazzi, R. A. Stirling-Gallacher and B. M. Masini, "Performance Characterization of Joint Communication and Sensing With Beyond 5 G NR-V2X Sidelink," in *IEEE Transactions on Vehicular Technology*, doi: 10.1109/TVT.2024.3365770

[45] C. Giovannetti, N. Decarli, S. Bartoletti, R. A. Stirling-Gallacher and B. M. Masini, "Target Positioning Accuracy of V2X Sidelink Joint Communication and Sensing," in IEEE Wireless Communications Letters, vol. 13, no. 3, pp. 849-853, March 2024, doi: 10.1109/LWC.2023.3346937.

# 5. Non-terrestrial networks in 6G

6G wireless mobile networks are expected to provide a plethora of services to billions of globally distributed users, including not only humans, but also machines, Internet of Things (IoT) sensors and the more sophisticated (future) devices that are required for applications with stringent performance requirements, such as Augmented Reality (AR) / Virtual Reality (VR), holographic communications, etc. Due to the drastic increase in the number of users, their need for ubiquitous connectivity and the economically driven interest to deploy infrastructure in geographical locations that are more densely populated, the necessity to co-operate Terrestrial Networks (TNs) with Non-Terrestrial Networks (NTNs) is arising. Such cooperation can provide cost-effective communications for users in sparsely populated areas, during emergency/disaster response scenarios, and further extend connectivity to oceans, the sky and space.




*© F. Rinaldi et al., Non-Terrestrial Networks in 5G & Beyond: A survey [1]*

NTNs refer to radio access networks where the access nodes are carried on platforms hovering high above the terrestrial surface in the air (high-altitude platform systems – HAPS, such as drones or balloons) or in space (satellites). It is noted that the NTN access nodes may be built on transparent or regenerative architectures, where the former realizes the access node as a Radio Frequency relay, while the latter allows to implement parts or all of the layers of a 5G base station according to the functional split options. For the next generation of mobile networks, such NTNs are highly attractive due to several reasons: First, they represent the means for ultimately fulfilling the ever-given promise of mobile networks to *provide ubiquitous connectivity anytime and everywhere*, removing availability gaps in coverage maps and providing the required resilience to withstand disasters. Second, they will *complement terrestrial radio networks* by providing additional communication links via independent access nodes, which may be used for offloading broadcast traffic in hotspot areas or for improving communication reliability and availability during daytime and energy efficiency at night-time. Third, network-operated mobile mission-critical applications in the context of IoT and vehicular-to-everything (V2X) communications, which require high availability with guaranteed latencies and accurate localization, are becoming more and more widespread, and hence there is an increasing demand for *guaranteed service provision, seamless service continuity*

*and precise positioning* anytime and anywhere. NTN have the potential to become the puzzle piece for completing a system concept fulfilling this demand.

Especially the first two benefits of NTNs, ubiquitous coverage and traffic offloading during peak hours, have already been recognized by 3GPP, and hence new study and work items targeting the integration of NTNs in 5G-NR were initiated in Rel.16, which were continuously developed further in Rel. 17 and 18 and are followed up in Rel.19 to date [2]-[5]. The research community has also started to unveil the potential energy savings that are achievable by offloading the low traffic at terrestrial cells at night-time to NTN hypercells and by allowing the deep sleep of terrestrial, power-hungry macro stations that would otherwise remain active to provide blanket coverage [6]. However, up to now, industry and 3GPP have mainly focused on using NTNs for the coverage extension of enhanced mobile broad band (eMBB) services and narrow-band IoT applications. Those services do not require much network coordination and can tolerate the large latencies that may occur due to the significant propagation distances between the earth surface and the satellite when using the conventional *transparent (or "bent-pipe") satellite architecture*, where the satellite simply acts as a Radio Frequency relay. In addition to delay tolerance, eMBB services are mostly best-effort, thus setting rather moderate demands on the Quality of Service (QoS). This eases the challenges of integrating NTN into the *existing* 5G-NR architecture, offering limited possibilities for substantial enhancements on the overall system design of the radio access network only. Considering, however, the energy efficiency challenge of current networks and the fact that 5G-NR has introduced two additional core services with properties differing substantially from eMBB services, namely massive machine type communication (mMTC) and ultra-reliable low-latency communication (URLLC), we envision that the steps taken by 3GPP towards NTNs can only be considered the first steps of a long pathway towards the full integration of TNs and NTNs.

## 5.1 Non-Terrestrial Networks for advanced services

With the introduction of novel mechanisms and features for the support of mMTC and URLLC services in 5G, mobile networks could be used for the first time to perform tasks for continuous monitoring and control services requiring highly reliable connections with a guaranteed QoS. This led to the emergence of several ground-breaking use cases in the IoT and V2X context, which were built on these 5G services and enabled autonomous operations at very high levels of comfort. In particular for vehicles, which can drive virtually anywhere and may thus easily get out of network coverage, guaranteed network connectivity frequently becomes an issue in daily operations. Hence, enhancing mobile networks with NTN components and functionalities with integrated support for the full set of 5G services and beyond (thus also considering new services to be supported by future network evolutions) is expected to become a key enabler for providing true ubiquitous connectivity and seamless service continuity for those use cases.

Given this objective, however, a question arises: *How can satellites be used to provide connectivity for all service types, in particular for those built on high reliabilities and low latencies*? To begin with, it should be noted that the latency constraints of typical services from the IoT and V2X realm lie in the order of several tens of milliseconds (typically 20-30 ms), and hence they are less strict than the minimum latency constraint supported by 5G URLLC services of a single millisecond. This means that those IoT and V2X services provided via satellite can bear propagation delays of a few milliseconds, but not much more than that. Satellites are operated in different orbits, reaching from the low earth orbit (LEO) at altitudes between 200 and 2000 km (measured from the earth surface), over the medium earth orbit (MEO) with altitudes between 2000 and 35,786 km, to the geostationary orbit (GEO) at 35,786 km altitude. In the very low earth orbit (vLEO), satellites circle the earth at altitudes between 200 and 400 km and thus can facilitate relatively short latencies of a few milliseconds due to the reduced propagation distances. Through the implementation of mass production techniques and the exploitation of reusable launchers, the latest generation of vLEO satellites is expected to come at modest capital and operational costs. Inherent to this satellite generation is a *regenerative satellite architecture* featuring powerful on-board processing, which allows these satellites to function as "network nodes in the sky" with compute and storage

capabilities, turning them into network edges. By using the on-board processing capabilities for (distributed) edge computing, 5G services and artificial intelligence (AI)-based applications can be directly executed at those satellite nodes. Such novel features of the vLEO satellites will allow lower latency applications, since they rely on a direct point-to-point communication between the satellite and the device and avoid the 3-point link via satellite and gateway ground station as in conventional satellite communications using the bent-pipe architecture, where the one-way signal always needs to travel twice the distance between earth surface and the satellite. By using satellites of the latest generation in the vLEO orbit, the propagation delay can indeed be reduced to a few milliseconds, hitting the target as formulated above, which will eventually enable the provision of URLLC services typical for the above use cases via NTN enhanced mobile networks. However, none of the above benefits will be realized without a true convergence of TNs and NTNs into a single, unified network capable of seamlessly managing and orchestrating eMMB, mMTC and URLLC services.

## 5.2 Terrestrial and Non-Terrestrial Network Convergence

Many of the recent works [2]-[5], including those of 3GPP, have only just begun with the integration of NTN and are primarily focused on the RAN-end of the mobile network. There is still a long way to go before achieving a full synergy, though. If all the current as well as future mobile network services can be provided also through NTN nodes, then the NTN domain may readily become an integral part of the future mobile system, expanding the overall network structure to the 3rd dimension and forming an organic 3D network. This goes clearly beyond the "integration of NTNs" into the existing network design of 5G-NR, as currently pursued by 3GPP.

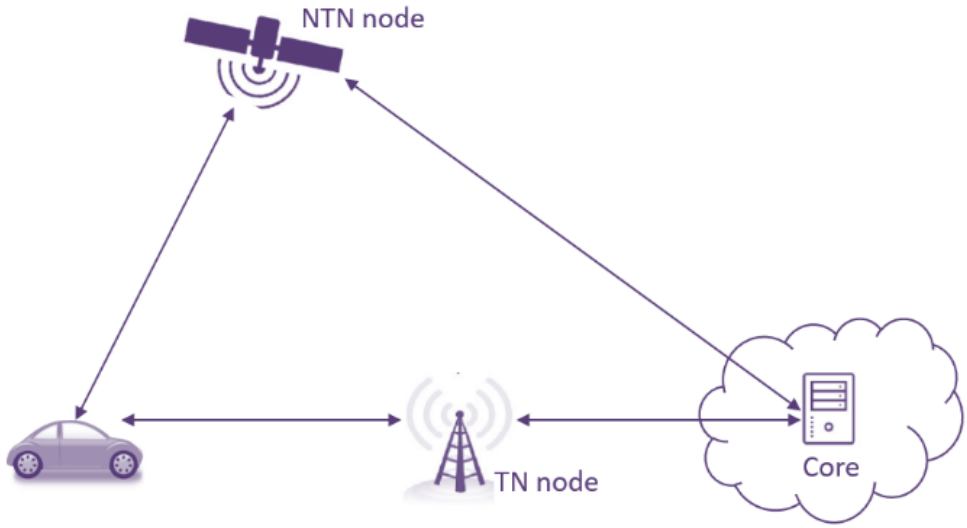



*Figure 8: Vision of TN and NTN convergence*

If we now consider the network being constituted of two different network domains, one being a TN and the other being an NTN, while both equally support the same set of services, then a completely new system design based on a novel network architecture should be derived to yield a convergence of the TN and NTN domains, uniting the best features from those two worlds in the most efficient way. With a fully converged network of this kind, the type of connection may become transparent to the user, meaning that users will not be aware of whether they are connected via a TN or an NTN node – or even via both at the same time (Figure 8). This type of network requires a novel network architecture supporting a large variety of network nodes with different properties and capabilities. For example, NTN nodes will be continuously moving and thus frequent handovers are required, even if the user is static. Satellites will also have their own on-board resources for compute, storage and networking. However, they will be strictly limited and may not be easily enhanced or upgraded, as compared to TN nodes like terrestrial base stations, where computing

and storage capabilities can be easily extended, e.g., by installing additional computing devices for multi-access edge computing (MEC). With the TN and NTN nodes in a mobile network, the number of network edges is significantly increased. Considering that each of the elements (TN node, NTN node and core) may host heterogeneous resources to be used for MEC, there will be a plethora of MEC-capable edge nodes, whose resources should be efficiently managed for the orchestration of the individual network services as well as for the implementation of AI functionalities in a distributed fashion. A truly converged mobile system with TN and NTN nodes for radio access supported by a mobile core is the future network environment that is envisioned for 6G, which calls for the development and investigation of the key enabling technologies for its realization.

## 5.3 Use cases for integration of NTN and TN

In this section, we give an overview of promising use cases building on integrated or converged NTN and TN in beyond 5G systems.

### 5.3.1 Use case 1: Sensor Networks in the Agricultural Sector

Smart agriculture is a scenario that involves several improvements to the classical agriculture scenario due to the introduction of ICT technologies. One of them is related to the employment of wireless sensor networks within crop fields. These sensors can manage different tasks related to the monitoring of plant growth and the overall agricultural environment. They can be deployed at different stages (e.g., different seasons), operate in different conditions (e.g., in open fields or inside the soil), and be of different kinds (e.g., cameras and temperature or humidity sensors). Properly adjusting the irrigation system operation depending on the environmental conditions and the plant growth in particular areas and the addition of the proper amount of fertilizer depending on the soil conditions before sowing are just a few examples of the traditional agricultural processes that could be improved.

However, especially in case of vast agricultural fields, the number of deployed sensors can be significant, as well as the amount of generated data to analyse and correlate to make decisions. All these data have to be forwarded to a central node (located within the agricultural environment or not) running proper AI applications. One of the issues to face is that agricultural areas are typically located in geographical regions not well covered by the terrestrial infrastructure. Besides, the terrestrial infrastructure may have been sized considering the typically low number of inhabitants in these regions and, hence, may not be well equipped for a significant increase in the number of connected nodes and consequent increase of the data traffic volume.

The employment of UAVs can help to achieve this goal. UAVs may be periodically deployed to operate as access nodes, collecting the data generated by sensors and forwarding them to the central node. Data forwarding can be delayed by implementing the store-and-forward paradigm onboard UAVs. In this way, UAVs collect data while they are flying above the target area and subsequently forward them only when they land back at their ground station. This can contribute to lowering UAV energy consumption, and thus extending their operational time, as well as lowering their hardware complexity, avoiding to equip them with long-range communication interfaces going beyond those used to communicate with the sensors.

### 5.3.2 Use case 2: Disaster relief situations

Disaster relief scenarios arise in crisis conditions that lead to: i) a sudden and unpredicted unavailability of either a part or the entire terrestrial infrastructure, for a limited period of time; ii) the need for Search & Rescue (S&R) operations in remote areas, i.e., areas without the support of a Terrestrial Network (TN). Such unavailability might be caused by a natural disaster (e.g., earthquake, flood, or fire) or be man-made (e.g., terrorist attack). In this context, there are two different services which shall be guaranteed: i) with higher priority, it is fundamental to promptly establish a

temporary network in the crisis area for the first responders, in order to coordinate and organize the rescue operations; and ii) with a slightly lower priority, the network operators providing service in the area shall be granted the possibility to initiate the restoring of the terrestrial network infrastructure.

In this framework, the availability of a non-terrestrial access, which can be either airborne (UAVs, HAPS) or spaceborne (LEO, MEO, GEO), is fundamental for the very first phase of the crisis. Moreover, while the terrestrial infrastructure is progressively being restored, dual connectivity solutions can also be implemented, in which the terrestrial and non-terrestrial segments are jointly exploited for the above purposes. It is worthwhile mentioning that a rescue operation might also be required in remote areas where no terrestrial infrastructure is present. In such cases, the exploitation of a non-terrestrial component is the only solution to allow the proper coordination of search and rescue operations. As such, this case can be considered as included in the former one with the same service flow and requirements, excluding the joint terrestrial/non-terrestrial connectivity option.

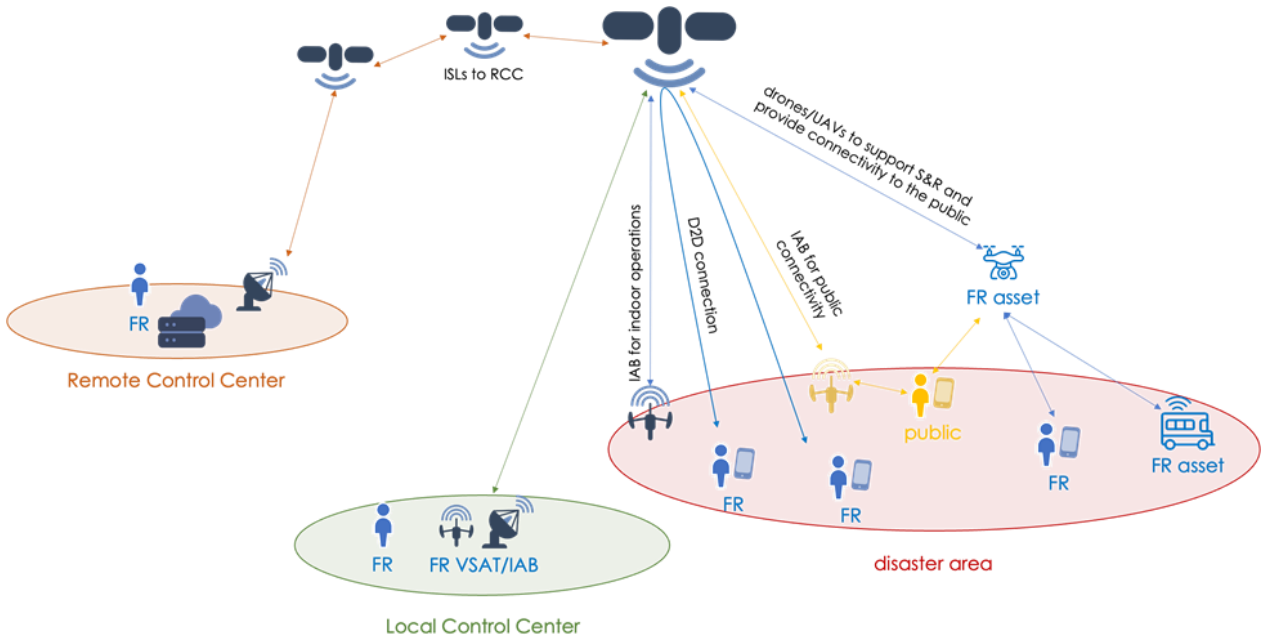



*Figure 9. Overview of the disaster relief scenario.*

As shown in Figure 9, in the aftermath of a disaster, the availability of a non-terrestrial component should allow information exchange: i) among the first responders in the disaster area; ii) between the first responders in the disaster area and a local control centre close to the disaster area; and iii) between the first responders and the remote command and control centre. The required capacity can be either limited, e.g., for sensor or location reporting, or large, e.g., for Augmented Reality (AR) and/or live stream helmets. It is also possible that indoor connectivity might be needed for the rescue operation; in such conditions, Integrated Access and Backhaul (IAB) nodes are required [7],[8].

In the context of disaster relief scenarios, it shall be mentioned that ITU-R Report M. 2291-2 provides indications on the use of International Mobile Telecommunications (IMT) technologies to support broadband Public Protection and Disaster Relief (PPDR) operations.

More detailed information on this scenario is partially available in 3GPP TR 22.822 [9], which addressed similar conditions in the “Temporary use of a satellite component” use case.

### 5.3.3 Use case 3: Energy saving enabled in integrated NTN/TN networks

In the realm of 6G telecommunications, NTNs have the potential to usher in a paradigm shift—a shift that goes beyond providing ubiquitous coverage and holds the potential to revolutionize our digital landscape while preserving precious energy resources. This section elaborates on two corresponding use cases, targeting the reduction of the energy consumption in the integrated network.

#### 5.3.3.1 Energy Consumption-based load balancing in NTN/TN networks

Climate change, carbon-footprint and environmental sustainability have become important issues that need immediate attention. Hence, it is imperative that 6G systems provide services that take reduction in energy consumption as a native factor in their service offerings. In an integrated TN/NTN environment with multiple heterogeneous access networks, different network entities have different characteristics, and various network entities operate under different conditions, especially under different loads. However, the higher load handled by a network entity, the greater its Energy Consumption. Moreover, different types of loads have different EC characteristics, i.e., the EC does not grow linearly with different types of loads. In this regard, this use case highlights the need for actively balancing and re-balancing the load across all available network entities in a heterogeneous, integrated TN/NTN environment to facilitate the goal of reducing Energy Consumption of mobile telecommunication networks. Such a form of active measurement, monitoring, balancing and re-balancing of the load across various network entities with respect to their energy consumption will ensure energy-efficient utilization of the resources and reduce the OPEX for MNOs.

#### 5.3.3.2 Advanced carrier shutdown in integrated NTN/TN networks

Traditionally, terrestrial cellular networks have operated around the clock, with macro base stations relentlessly churning out signals to provide blanket coverage. However, this round-the-clock activity exacts a significant toll on energy resources, contributing to environmental concerns and escalating operational costs. The emergence of NTN-based hypercells offers a game-changing solution to this dilemma.

The new concept/use case we have in mind is simple yet profound: During the night-time hours when network traffic ebbs to a minimum, low-traffic terrestrial macro base stations can offload their traffic to energy efficient NTN-powered hypercells. This offloading enables power-hungry macro base stations responsible for blanket coverage to be shut down. These dormant macro base stations, which would otherwise remain active, consume substantial energy even when scarcely being used. Despite advanced energy-saving solutions, today, about 40% of base stations in a network remain operational to provide coverage [6].

But how can this vision be realized efficiently? The answer lies in a truly integrated terrestrial and NTN and the development of cooperative algorithms designed to minimize the terrestrial network's energy consumption through satellite and high-altitude platform station (HAPS) and/or unmanned aerial vehicle (UAV) offloading, while maintaining the quality of experience of users. The integration must combine technological innovation with sustainability, setting the stage for a future where our digital connectivity coexists harmoniously with the conservation of our planet's resources. For the details on a technical solution and its performance, the reader is referred to Section 5.4.1.

In short, as we delve deeper into the night, this use case envisions a new energy efficiency solution, where terrestrial networks go to sleep while NTN hypercells take over, reducing our carbon footprint and illuminating a brighter, more sustainable future.

### 5.3.4 Use case 4: Multiple-Access Networks in integrated NTN/TN Network

In the current 3GPP mobile telecommunication standards, integration of Non-Terrestrial Networks (NTN) into Terrestrial Networks (TN) has garnered a lot of attention, where Releases 15-18 discuss solutions for NR & NG-RAN to support NTN (TR 38.811 Release 15 [2], TR 38.821 Release 16 [3], TR 23.700-27 Release 18 [4], TR 23.700-28 Release 18 [5]), while Release 19 discusses improvements for optimized performance for terminals, capacity performance on up link, notification of service area of a broadcast service, support for an NTN architecture with 5G system functions on board the NTN vehicle and use of RedCap devices within FR1 NTN. Therefore, an integrated TN/NTN telecommunication environment is anticipated in 6G. Moreover, we believe that Multiple Radio Access (MA) Technologies like LTE, NR, THz, NTN, Wi-Fi, to name a few, will be available simultaneously to Users in an integrated NTN/TN environment. Hence, User's Equipments (UE) will be equipped with radio capabilities to utilize services offered through these multitude of technologies in order to maximize their throughput and reliability. While, network operators would maximize the use of their heterogeneous technology infrastructure and resources. Therefore, inter-operation among such heterogeneous networks is deemed imperative (refer to TS 22.261 clause 6.3 [10]) and multiple access technologies must co-exist and operate efficiently in an integrated NTN/TN environment. This brings forward many new challenges that must be addressed.

Therefore, this use case aims to discuss the challenges and requirements for TN/NTN integration in the presence of multiple heterogeneous access networks along with challenges and requirements for sustainable operation of telecommunication networks in such an environment through elaborating on the following issues, which address specific aspects within the larger context of this use case.

#### 5.3.4.1 Seamless Service Continuity in a TN-NTN integrated Environment with Multiple Access Technologies

In today's mobile networks, handover is performed to ensure session continuity. However, current handover management solutions are complex. In a multi-access integrated TN/NTN environment, where even NTN nodes will be mobile, intricate co-ordinations will be needed between participating Access Networks and Core Networks, resulting in latency, scalability and QoE issues (refer to TS 37.340 [11]), which may worsen the effects of handover, resulting in increased Handover Interruption Time (HIT) and Handover Failures (HoF). This use case highlights the need to address the issue of simplifying handovers in an integrated multi-access TN/NTN environment e.g., through efficiently leveraging multiple simultaneous connections from the network and enabling UEs to distribute their traffic across all available connections based on the capability of the access networks. Essentially, we need solutions to ensure UEs can experience reliable connectivity and global coverage with seamless service continuity between various access networks with no HIT and HoF during mobility.

#### 5.3.4.2 Application Influenced Traffic Steering in a Multiple Access Environment

Utilizing heterogeneous resources efficiently in a multi-access integrated NTN/TN network is a challenge for the mobile network operator. On the other hand, applications and services want to ensure a certain QoE for their Users, but are unaware of the network characteristics that an end-user is experiencing. Hence, this use case highlights the need for network operators to partner with 3rd party applications and service providers and include their requirements in the service offerings to Users. This partnership when realized could lead to indications such as e.g., instructions to the UEs on how to distribute their traffic for a session on the available connections. This will ensure users receive a much better QoE while the resources in network and in end-point applications/services are utilized efficiently.

#### 5.3.4.3 Traffic Engineering with Multiple Access Technologies

When multiple services co-exist, one or more services could monopolize available resources based on their requirements e.g., high bandwidth, low latency, etc. Therefore, efficient traffic distribution over all available links is a crucial requirement for a network provider in order to efficiently utilize the resources in a heterogeneous integrated NTN/TN network with the goal of ensuring an overall good Quality of Experience (QoE) for users. From practice, it has been observed that every network has some congested links, while some links are underutilized. Therefore, it is crucial to engineer the traffic effectively in order to avoid congestion. Hence, this use case highlights the need for dynamically distributing the traffic in a multi-access integrated NTN/TN mobile network. With dynamic traffic distribution through leveraging multiple access networks we could ensure that no network is left to operate on its own and experience adverse effects resulting from multiple different types of application traffic and User loads.

#### 5.3.4.4 Networking, Compute and Energy aware scheduling in integrated NTN/TN networks

In 6G, billions of globally distributed users are expected, ranging from humans to vehicles to IoT devices. Such a heterogeneity in users will naturally introduce varying requirements from the mobile networks. Further, advanced applications that 6G is expected to support like Artificial Reality (AR)/Virtual Reality (VR), holographic communications, etc., will run multiple simultaneous compute, storage and transport sessions with stringent performance requirements. This use case highlights the need to schedule various different types of User sessions and their tasks onto network entities in an integrated TN/NTN environment based on the characteristics of the network entities, the requirements of the tasks and their expected EC with the overarching goal of reducing the EC of the network. With energy-efficient scheduling, end-to-end energy consumption of user sessions and networks can be controlled and services can be scheduled with the goal of not only satisfying the requests of Users but also to reduce the overall EC resulting from execution of User requests.

## 5.4 Enabling technologies for the use cases

This section will elaborate on selected enabling technologies for the use cases introduced in the previous section.

### 5.4.1 NTN for improving energy efficiency

As introduced earlier in Section 0, the exponential growth in wireless traffic, driven by the proliferation of mobile devices and the advent of 6G technology, poses significant energy efficiency challenges for radio access networks. A critical issue is the "zero bit non-zero watt" problem, which arises as many coverage macrocells are kept active even when there are no users to ensure that the network is always ready for potential users. This results in unnecessary energy consumption and operational costs. One of the aims of the work presented here is to tackle this problem by integrating Non-Terrestrial Network (NTN) components that can dynamically manage network traffic and reduce energy consumption.

The concept of integrating NTN into terrestrial networks presents a promising solution to the “zero bit non-zero watt” problem. Our solution involves offloading traffic from terrestrial macrocells to NTN hypercells, enabled by energy-efficient satellites, HAPS, or even tethered UAVs. By dynamically offloading traffic, terrestrial coverage macrocells can be switched off or transitioned to low-power states during periods of low traffic, thereby reducing overall energy consumption and helping us achieve the “zero bit zero watt” paradigm. Each type of NTN node offers unique performance trade-offs and complementary advantages towards this end:

- Satellites: Ideal for offloading outdoor users at night-time due to their extensive coverage and the limited capacity needs during this time of day, although they may face latency and signal degradation issues.
- HAPS: Suitable for offloading indoor users, HAPS provide better spatial reuse, signal penetration, and lower latency, making them more effective for urban environments.
- Tethered Drones: These provide additional offloading capabilities in high-demand areas, offering flexibility and targeted coverage efficiently as tethering allows to leave the most of the radio on the ground and bring electricity to the drone.

To optimize this offloading process and assess the potential benefits of the proposed solution, a heuristic algorithm has been designed to minimize the hourly energy consumption of the terrestrial network while ensuring the quality of service for users. The algorithm prioritizes offloading from terrestrial base stations with the least traffic volume first, thereby maximizing energy savings by allowing more base stations to enter sleep mode. Specifically, the algorithm sorts base stations by hourly traffic rate, offloading traffic from those with lower rates until the capacity constraints at the NTN hypercells are met.

The selection of the appropriate NTN node (satellite, HAPS, or UAV) to offload each specific user depends on the user's location, signal quality, and requirements, as well as the local network capabilities. To ensure a smooth transition of users from terrestrial to NTN cells, robust handover mechanisms are essential. To guarantee quality of service, sleeping macrocell stations will have to be activated on time. A true integration of terrestrial and NTN networks will enable the collection of relevant data and the execution of informed decisions to optimize this energy-saving solution and in turn network performance effectively. This method helps save the load-independent energy consumption of more terrestrial base stations, thus enhancing overall energy efficiency.

To assess performance gains, we focus our analysis on an integrated NTN network based on HAPS and use system-level simulations with realistic traffic and energy consumption models:

- Traffic Modelling: Our simulation uses mobile traffic data from Milan (2015) with 1419 traffic traces, each for a base station (BS), reported hourly over two months. To reflect current trends, we scaled these traces to match recent data from 960 4G/5G BSs in China (2020). We followed a four-step process: averaging the original traces to construct weekly profiles, scaling these profiles to recent data, selecting the most representative scaled trace, and constructing updated weekly profiles. This ensured our simulation data was realistic and up-to-date.
- Energy Consumption Modelling: We modelled RAN energy consumption using data from over 12,000 4G/5G macrocells in China, incorporating sleep modes. The model includes baseline energy consumption in sleep mode (E0), baseband processing energy (EBB), energy consumed by RF chains (ETran), power amplifier static energy (EPA), and energy for data transmission (Eout). This comprehensive model, fitted with real energy consumption values, allows accurate estimation of BS energy usage based on traffic load, ensuring precise simulations. Additionally, NTN nodes are considered self-sufficient, contributing to energy savings without consuming additional power.

Our study performs a parametric analysis to evaluate the influence of key parameters, including the elevation angle of HAPS, the percentage of traditional buildings, and the proportion of indoor user equipment (UEs). This analysis helps to understand the conditions under which maximum energy savings can be achieved. Our heuristic algorithm is employed to minimize the hourly energy consumption of the terrestrial network. As mentioned earlier, the algorithm prioritizes offloading from BSs with the least traffic volume first, allowing more BSs to enter sleep mode and thereby maximizing energy savings. A Monte Carlo simulation process is used to account for variability and uncertainty in the parameters, such as the probability of line-of-sight (LOS) conditions, indoor vs. outdoor UEs, and traditional vs. energy-efficient buildings. This approach ensures the robustness of the simulation results. Table 4 provides a summary of the parameters of our simulation.

*Table 4: Simulation parameters*

| Parameter | Value |
|---|---|
| Environment | Dense Urban |
| $l_B$, maximum fraction of offloaded BSs | 0.6 |
| $B$, channel bandwidth [MHz] | 20 |
| $f_c$, carrier frequency [GHz] | 2 |
| $d$, HAPS height [km] | 20 |
| $P_{tx}$, transmit power [dBm] | 43 |
| $G_{element}$, transmitter gain (single element) [dBi] | 8 |
| $G_{rx}$, receiver gain [dBi] | 0 |
| $n$ rows [-] | 1 |
| $m$ columns [-] | 4 |
| $T_N$, antenna Noise Temperature [K] | 290 |
| $N_p$, thermal noise power [dBm] | -100.96 |
| $\alpha$, elevation angle values | [60°, 70°, 80°, 90°] |
| Portion of traditional building range | (30%, 70%) |
| Percentage of indoor user range | (60%, 90%) |

The experimental results indicate substantial energy savings through the integration of HAPS into terrestrial RANs. By offloading traffic to HAPS, energy consumption can be reduced by up to 29% across the entire week, with even higher savings of up to 41% during night-time hours. The analysis shows that during the day, only a small fraction (3% to 15%) of the total traffic is handled by HAPS, while most of the traffic is still managed by terrestrial BSs. Despite this limited offloading, energy savings are achieved due to the reduced activity of power-hungry BSs. The parametric analysis, summarized in Figure 10, reveals the following insights:

- Elevation Angle: The elevation angle has the most significant impact on energy savings. Differences of up to 6 percentage points in energy savings are observed between the smallest and largest angles.
- Indoor UEs: The proportion of indoor UEs affects energy savings, with additional savings of approximately 3.6% when the percentage of indoor UEs decreases from 90% to 60%.
- Traditional Buildings: The percentage of traditional buildings shows a localized effect, with differences of around 2.5% in energy savings.

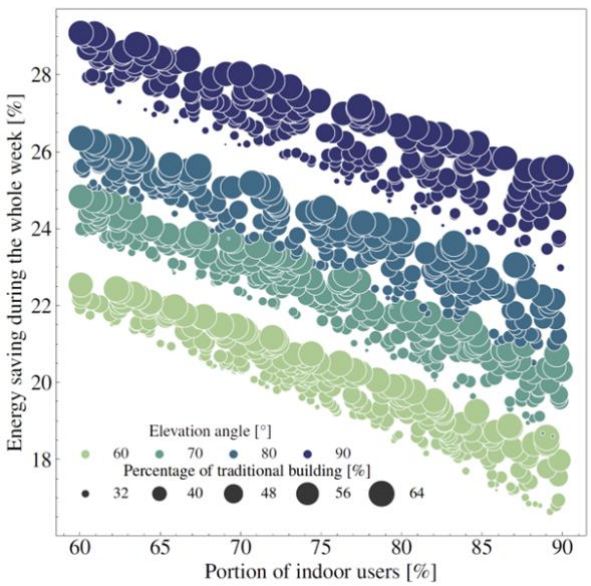



*Figure 10: Energy saving at each trial (y-axis) with respect to elevation angle (sample colour: the darker the colour, the larger the angle), portion of indoor users (x-axis), and percentage of traditional buildings (sample size: the larger the size, the more the traditional buildings).*

Overall, the study highlights that higher elevation angles, fewer indoor users, and a greater number of traditional buildings contribute to higher energy savings. The findings underscore the potential of integrating HAPS into RANs to achieve significant energy efficiency improvements, emphasizing the importance of optimizing these parameters for maximum benefit. For further details, the reader is referred to [6].

### 5.4.2 Dynamic spectrum sharing between NTN and TN

The increasing demand for ubiquitous connectivity and the extensive deployment of cellular networks has prompted significant concerns about energy consumption in mobile networks. A novel framework called BLASTER (Bandwidth spLit, user ASsociation, and PowEr contRol) aims to enhance energy efficiency and network performance by dynamically managing resource allocation in an integrated terrestrial and non-terrestrial network (TN-NTN) [12]. The framework focuses on optimizing the association of user equipment (UE), base station (BS) activation, and bandwidth allocation between terrestrial and non-terrestrial layers to adapt to varying traffic conditions effectively.

In the proposed integrated TN-NTN system, low-earth orbit (LEO) satellites complement the terrestrial networks to extend coverage and improve connectivity, especially in remote and rural areas. Both the terrestrial and non-terrestrial tiers operate in the S-band. The NTNs not only help in providing extensive coverage, but also significantly support reducing the energy consumption by enabling dynamic BS activation based on traffic load. This is particularly useful during low-traffic scenarios, such as during night-time, where NTN can be used to provide blanket coverage and most terrestrial macro BSs (MBSs) can be turned off accordingly to save energy.

The implementation of BLASTER involves a detailed simulation setup, where the network comprises both terrestrial MBSs and MBSs hosted on LEO satellites. The total system bandwidth is strategically allocated between these two tiers, and the association of UEs to BSs is governed by the strength and quality of the received signal and the current load on each cell (in TN as well as NTN). This approach ensures an optimal load distribution and minimizes performance degradation due to overloaded BSs.

BLASTER's strategy is to distribute the network load optimally across the terrestrial and non-terrestrial tiers. By adjusting the BS transmit power, activating or deactivating BSs, and managing bandwidth allocation, BLASTER can reduce the average daily total power consumption of the network by up to 45% compared to traditional networks following 3GPP recommendations, as shown in Figure 11. Moreover, the system achieves an average throughput increase of about 250%, indicating a substantial enhancement in network capacity during high-traffic periods.

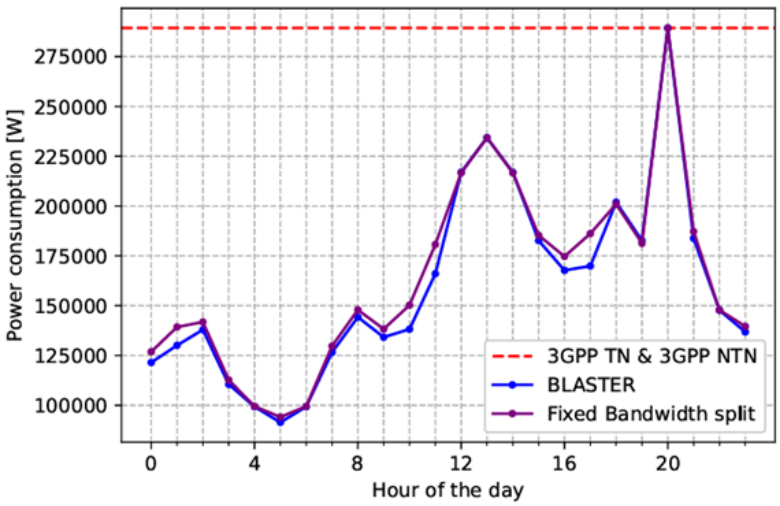



*Figure 11: Network Power consumption throughout the day.*

Simulation results underline the efficacy of BLASTER in dynamic traffic scenarios typical in rural settings. The network adapts to hourly variations in UE density, optimizing resource utilization and maintaining quality of service (QoS) across the network. During peak traffic hours, BLASTER's ability to manage resources stands out, as it significantly enhances throughput while keeping energy consumption in check, as shown in Figure 12. Conversely, in low-traffic periods, substantial energy savings can be realized by leveraging NTNs to handle a larger share of the network load, allowing for the deactivation of terrestrial BSs without sacrificing coverage or connectivity.

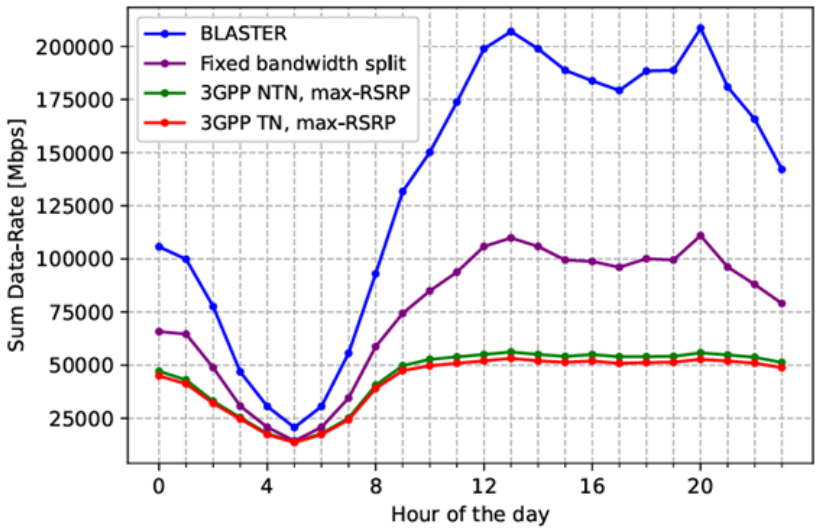



*Figure 12: Network Sum Data-Rate throughout the day.*

In conclusion, BLASTER exemplifies a significant step forward in the design of energy-efficient and high-performance mobile networks. By seamlessly integrating terrestrial and non-terrestrial technologies, it not only addresses the immediate challenges of energy consumption and network coverage, but also provides a scalable model for future network deployments. The flexibility and efficiency of BLASTER suggest that such integrated approaches will be crucial in evolving network infrastructures to meet the expanding demands of global connectivity, especially in underserved regions.

### 5.4.3 Multi-Access

Seamless communication and global coverage are considered important requirements/KPIs for supporting many 6G use cases like V2X, robotics, AR/VR and multi-modal use cases which need to transmit ultra-reliable time-sensitive multi-modal data like simultaneous transmission of audio, video, haptic, accurate-position, speed, velocity, and derived data like body posture, facial expressions, head movement, objects, and operating environment data. In parallel, as discussed in Use Case 0, multiple access technologies will be available in 6G in an integrated NTN/TN environment and UEs will be equipped with multiple radio capabilities for leveraging such heterogeneous radio technologies. Due to availability of Multiple Access Technologies, 3GPP has been exploring multi-connectivity since Release 15.

Even though 3GPP provides multi-connectivity solutions like Multi Radio Dual Connectivity (MR-DC) to support the non-standalone 5G architecture (refer to 3GPP TS 37.340 [11], Release 15) and Access Traffic Steering, Switching & Splitting (ATSSS) to balance the traffic load between 3GPP and Wi-Fi networks (refer to 3GPP TS 24.193 [13], Release 17-19), it is not truly multi-connectivity, as it limits the UE to a maximum of only two simultaneous connections. Moreover, MR-DC and ATSSS are difficult to scale and require complex, intricate configurations, with some variants also requiring MPTCP/MPQUIC protocols in the network and in the UEs. Unfortunately, this may lead to TCP port collision and exhaustion when UEs create a large number of flows. Hence, MR-DC and ATSSS may not be sufficient to fulfil the requirements of above mentioned 6G use cases, as they need multiple

simultaneous connections to deliver multi-dimensional data and services, where each stream of data and service will have its own set of requirements. Moreover, these solutions add additional overhead during mobility resulting from complex and intricately coordinated handovers, Handover Failures (HoF) and unavoidable Handover Interruption Times (HIT), inadvertently affecting seamless mobility especially in heterogeneous network environments.

Today, NTN networks cannot be directly incorporated into 3GPP to enable global coverage as NTN components cannot operate independently with 3GPP core network entities like, e.g., AMF. The necessary functionality, mechanisms and apparatus to extended 3GPP services to NTN are not yet in place. However, since NTN will become an integral part of 6G (in addition to legacy accesses), there is for the first time an opportunity for several different access technologies to be standardized within a single 6G communication system. In this regard, an obvious and indisputable problem is how will different access types be used in 6G? We believe that it could proceed in one of the two possible directions:

1. by continuing the current practice of patching the new system to support the legacy access types or

2. by building a system that is multi-access by design, wherein an access is treated as just an access and not as the entire system.

The latter of the two possibilities, i.e., multi-access by design, is what we advocate for. In principle, access network and core network are two separate pieces of a mobile network, and any changes to the access network should be supported in the core network, but not lead to the creation of a whole new core network. Hence for 6G, it is paramount to create a multi-access system by design, capable of supporting any and all types of access.

#### 5.4.3.1 Challenges for 6G Mobile Communication Systems

The following challenges and requirements from the corresponding use cases further demonstrate the need for multi-connectivity in future mobile networks.

- Seamless Global Coverage: in addition to increasing rate of end-user mobility, the network itself could be mobile in 6G e.g., NTN, fast trains providing sub networks to their passengers, etc. Hence, the challenge of seamless service continuity, especially in a heterogeneous integrated NTN/TN environment under various mobility scenarios requires further research.

- Multi-modal Communications: as mentioned above, many 6G applications require simultaneous compute, storage and communication sessions with very low latency and no-service interruptions. Today's standard practice of providing a single session to UE over 1-2 access will longer be sufficient. Since heterogeneous networks will be available in the future mobile networks, allocating multiple sessions over the resources in heterogeneous networks based on the requirements of the application will be necessary. However, managing multiple simultaneous sessions across heterogeneous networks needs further research.

- Diverse User Requirements: Since a multitude of advanced and complex applications and services are expected to be available in 6G like AR, V2X, haptic-based services, etc., to name a few examples, users of these applications will have varied types of requirement unlike the traditional mobile users e.g., high reliability, including during mobility, increased throughput, low latency and service continuity, multiple antennas, However, satisfying such diverse User requirements while simultaneously, ensuring optimally resource utilization in a heterogeneous network environment needs further research.

- Autonomic Core and RAT Network Evolution: traditionally, every new radio access technology has led to a new corresponding Core Network, resulting in a tight integration between the RAN and Core Networks. Moreover, in order to co-exist with legacy networks, additional enhancements are provided e.g., in 5G multiple inter-working architectures are

provided in order to co-exist with 4G and non-3GPP networks. If this trend continues, then 6G will have to provide inter-working architectures for 4G, 5G, NTN and non-3GPP networks, and succeeding generations will then have to support all the preceding (legacy) networks. This design choice will certainly not be scalable. The design trend of tightly integrating the RAN and Core networks with separate inter-working architectures to operate with legacy networks poses a major challenge for future mobile networks. Further research is needed to develop a 6G Core architecture with true separation of RAN and Core (remove tight integration between RAT and Core) such that, a 6G core can simultaneously operate with any and all types of RATs including legacy RATs and RATs that will be developed in the future.

### 5.4.4 UAV-aided Wireless Networks

In the realm of civil applications, Unmanned Aerial Vehicles (UAVs) have witnessed a notable surge in utility, finding roles in delivery services, video monitoring, photogrammetry, and beyond, even within urban landscapes. Advancements in technology are poised to render professional UAVs safer, capable of bearing heavier payloads, and extending their flight durations. For these reasons, we can predict an incredible increase in the use of UAVs in urban environments.

In this regard, in the dynamic arena of telecommunications, the integration of drones as flying Base Stations (BS) heralds a fundamental shift in network architecture. While UAVs fulfil their primary functions, such as aerial monitoring via onboard cameras, there exists the potential for them to also serve as Unmanned Aerial Base Stations (UABSs), thereby providing wireless communication services when required. This adaptation positions UABSs as adaptable tools within the arsenal of 6G networks, capable of strengthening the terrestrial infrastructure to meet the connectivity needs of ground users.

Moreover, UABSs offer unparalleled flexibility and scalability for future networks. Their on-demand deployment capabilities enable swift responses in ad hoc scenarios, ranging from delivering connectivity to remote disaster-stricken areas to enhancing agricultural operations through sensor networks and optimizing vehicular communications, as discussed in the previous sections. Unlike terrestrial infrastructure, UABSs are not bound by roads, unaffected by traffic congestion, and possess robust connectivity with both ground-based users and terrestrial base stations, thanks to their high likelihood of maintaining Line of Sight (LoS).

Nevertheless, the seamless integration of drones into terrestrial networks presents challenges necessitating careful consideration. Foremost among these challenges are regulatory barriers concerning airspace management and spectrum allocation, underscoring the imperative for close collaboration between telecom operators and regulatory authorities to delineate clear deployment guidelines. Furthermore, ensuring smooth interoperability and handoff between terrestrial base stations and airborne counterparts demands the implementation of robust network orchestration mechanisms. These challenges underscore key research trends driving the trajectory towards 6G standardization, including trajectory optimization [14], radio resource management [15], and the overarching investigation into coordination between airborne and terrestrial base stations, using for example mechanisms such as Integrated Access and Backhaul [16].

## 5.5 References

[1] F. Rinaldi, H.L. Maattanen, J. Torsner, S. Pizzi, S. Andreev, A. Iera, Y. Koucheryavy, and G. Araniti, “Non-terrestrial networks in 5G & beyond: A survey,” IEEE access, 8, pp.165178-165200, Sep. 2020. DOI: 10.1109/ACCESS.2020.3022981

[2] 3GPP TR 38.811, “Study on New Radio (NR) to support non-terrestrial networks (Release 15)”, 09-2020

[3] 3GPP TR 38.821, “Solutions for NR to support Non-Terrestrial Networks (NTN) (Release 16)”, 12-2019

[4] 3GPP TR 23.700-27, “Study on 5G System with Satellite Backhaul (Release 18)”, 12-2022

[5] 3GPP TR 23.700-28, “Study on Integration of satellite components in the 5G architecture; Phase 2 (Release 18)”, 03-2023

[6] T. Song, D. López Pérez, M. Meo, N. Piovesan, D. Renga, “High Altitude Platform Stations: The New Network Energy Efficiency Enabler in the 6G Era,” submitted to IEEE Globecom 2023, Dec. 2023. Preprint available at: https://arxiv.org/abs/2307.00969

[7] HORIZON-JU-SNS-2022 Project 6G-NTN (6G Non-Terrestrial Networks), D2.1 "Use cases definition," March 2023. Available: https://www.6g-ntn.eu/public-deliverables/

[8] HORIZON-JU-SNS-2022 Project 5G-STARDUST (Satellite and Terrestrial Access for Distributed, Ubiquitous, and Smart Telecommunications) D2.1 "Scenarios, use cases, and services," August 2023. Available: https://www.5g-stardust.eu/deliverables/

[9] 3GPP TR 22.822, “Study on using Satellite Access in 5G (Release 16)”, 06-2018

[10] 3GPP TS 22.261, “Service requirements for the 5G system; version 19.6.0”, 2024-03

[11] 3GPP TS 37.340, “Multi-connectivity; version 15.7.0 (Release 15)”, 10-2019

[12] H. Alam, A. de Domenico, F. Kaltenberger, D. López Pérez, “On the Role of Non-Terrestrial Networks for Boosting Terrestrial Network Performance in Dynamic Traffic Scenarios,” to be published in IEEE PIMRC 2024, Sep. 2024. Preprint available at: https://arxiv.org/abs/2405.14053

[13] 3GPP TS 24.193, “Access Traffic Steering, Switching and Splitting (ATSSS); version 16.0.0 (Release 16)”, 07-2020

[14] H. Bayerlein, P. De Kerret and D. Gesbert, "Trajectory Optimization for Autonomous Flying Base Station via Reinforcement Learning," in Proc. IEEE SPAWC 2018, doi: 10.1109/SPAWC.2018.8445768

[15] D. Ferretti, S. Mignardi, R. Marini, R. Verdone and C. Buratti, "QoE and Cost-Aware Resource and Interference Management in Aerial-Terrestrial Networks for Vehicular Applications," in IEEE Transactions on Vehicular Technology, doi: 10.1109/TVT.2024.3372310

[16] 3GPP TS 38.874 “Study on integrated access and backhaul (Release 15)” 2019-01

# 6. Multimodal sensing, computing, communication, and control for 6G remote operation

Maintaining a human model and digital representation of the physical environment for remote operation use cases over 6G involves the exchange of multimodal data including haptic, position, velocity, interaction forces, as well as audio, visual, gestures, head movements and posture, eye contact, facial expressions, user's emotion etc. Typically, haptic information is composed of two distinct types of feedback: kinaesthetic feedback (providing information on force, torque, position, velocity, etc.) and tactile feedback (providing information on surface texture, friction, etc.). The former is perceived by the muscles, joints, and tendons of the body, while the latter is consumed by the mechanoreceptors of the human skin. While the exchange of kinaesthetic information closes a global control loop with stringent latency constraints, this is typically not the case with the delivery of tactile impressions. In case of non-haptic control, the feedback is audio/visual and there is no notion of a closed control loop. With voice and data applications driving the designs of modern communication systems, the future Internet will enable exchange of haptic and other sensory information which in turn will be a paradigm shift toward Internet of skills and senses [1][2].

In addition to enabling multimodal data transfer, 5G and beyond mobile networks aim to deliver real-time control, for which, low latency is of critical importance. Hence, the interoperability among different communication technologies, with different capabilities, e.g., 5G Ultra-Reliable Low Latency Communication (URLLC), and IEEE Time Sensitive Networking (TSN), is essential. 3GPP Release 16 featured various architectural enhancements for seamless integration of the two technologies; however, realizing TSN-like functionality over 5G (and beyond) networks still needs several challenges to be addressed. In such converged scenarios, a teleoperated robot can be used to interact and operate (from a distance) by any connected device. In a tele-robotic system, a human operator controls the movements of the robot from a distance by sending signals to the robot to control it, while receiving the feedback that tells the operator the robot has followed the instructions.

In this section, we refer to the term telerobot as a robot controlled at a distance by a human operator, regardless of the degree of robot autonomy. Sheridan [3] makes a finer distinction, which depends on whether all robot movements are continuously controlled by the operator (manually controlled teleoperator), or whether the robot has partial autonomy (tele-robot and supervisory control). By this definition, the human interface to a tele-robot is distinct and not part of the tele-robot. Haptic interfaces that mechanically link a human to a tele-robot nevertheless share similar issues in mechanical design and control and include haptic interface development. These controls and feedback signals are called telemetry. In a more sophisticated form of teleoperation known as Telepresence, the human operator can see what the robot "sees," which gives the operator a sense of being on location and provides a user with an augmented reality (AR) experience.

Telerobotics encompasses a highly diversified set of fundamental issues and supporting technologies. More generally, tele-robots are representative of human-machine systems that must have sufficient sensory and reactive capability to successfully translate and interact within their environment. The fundamental design issues encountered in the field of telerobotics, therefore, have significant overlap with those that are and will be encountered in the development of veridical virtual environments (VEs) [3] or distributed virtual environments (DVEs) in virtual reality environment such as metaverse, a shared immersive virtual environment.

The evolution of mobile radio networks beyond 5G is expected to enable new applications and enhance telerobotics applications, such as remote operation for industry sites, tele-surgery to extend the coverage of medical services, tele-edutainment and so on. 6G can unlock robotic applications potentials by facilitating existing requirements and addressing new requirements that have not been previously discussed or analysed in detail from robotics perspective. In the following,

we define a number of tele-operations use cases and define their requirements and key performance indicators (KPIs) to explore potential gaps in existing connectivity technologies.

## 6.1 Use cases for multimodal remote operation in 6G systems

Remote operation has attracted significant industrial interest particularly due to the pandemic to prevent physical contact between the expert and the workers in the factory, so that the expert has to work from remote locations (e.g., remote inspection, remote certification, remote maintenance, or remote repair). TACtile and MultiModal (TACMM) use cases for remote operation, including multi-user, multi-device applications, and their requirements, performance indicators, network functionalities and service flow are being defined and collected as inputs relevant for 6G network design.

### 6.1.1 Remote facilities: Remote operation and maintenance for uncrewed sites

Remote facilities are uncrewed industrial/manufacturing platforms where human operator(s) are remote and no human is onsite (e.g., for high-risk or not easily accessible sites), with robots cooperating and capable of further adopting the details of cooperation. Telerobotic operations may require Human Robot Interaction (HRI), meaning multimodal information and action exchanges (e.g., through vocal, visual and tactile means) between human and robot to perform a task in remote facilities. Wireless infrastructure is required to provide highly reliable coverage for remote operation (e.g., by using XR and haptic controllers) between the remote operator and onsite robots as well as communication links between the robots cooperating together and between robots and the edge/cloud computing entities.

An example for remote facility is offshore drilling which is a dangerous job for rig personnel. Moving this role onshore increases operation safety yet the remote operators need an immersive experience (including tactile feedback) from the cyber drill for remote operation and maintenance on uncrewed platforms. The platform will be controlled remotely by an onshore facility, with a range of innovations, from a digital twin for trials to uncrewed wellhead platforms, among the new approaches planned for the remote facility. In addition to the risk, accessibility issues such as Covid-related restrictions for workers, have demonstrated the clear benefits that robotics, AI, and machine learning technologies can bring to the industry. Therefore, tactile and multimodal robotic operation could help enhance safety in oil and gas operations, as well as help to improve efficiency, reduce emissions and cut costs.

#### Pre-condition

(1) mobile robots can be equipped with 6G ISAC capable UE. (2) Robot(s) interact with human with implicit or non-implicit task assignment. (3) Robot with local control and AI capability (4) Human operator(s) are remote and run the operation and maintenance (6) no human is onsite (unscrewed industrial platforms).

#### Description

Robot designed for direct interaction with a human. Tele-operation robots may require Human Robot Interaction (HRI), meaning information and action exchanges (e.g., through vocal, visual and tactile means) between human and robot to perform a task remotely and with no onsite human intervention by means of a user interface.

## Service Flows

- The remote operator(s) interacting with the off-shore robot(s) get authorized and have permission to perform remote procedures
- The remote operator(s) are trained over Human Robot Interaction (HRI) interface
- Robot actuators execute seamlessly to the given commands from remote operator throughout the operation over network.

## Key techniques

IEEE 1918.1 and 3GPP XRM & Metaverse features describe main techniques to enable collaborative operation:

- Low latency and reliable communication over regional distance
- Multimodality synchronization
- Service availability
- Predictive digital twin

## Post-conditions:

The rig maintenance is performed remotely and safely without the need to send engineers to the rig.

## Potential Requirements:

- Environment model accuracy (including sensing, intelligence, communication)
- Environment model convergence time
- Connectivity latency, reliability, deterministic, service coverage, etc.
- Multimodal synchronization for communication and control for distributed remote operators

Table 5 [SA1 22.104] highlights which of the 5G KPIs needs to be reviewed to extend the service area requirements for 6G uncrewed robotic tele-operation.

*Table 5: 5G KPIS [3GPP22.104] with highlighted potential requirements on the service area coverage*

| Use case [3GPP 22.104] | Characteristic parameter | | | | Influence quantity | | | |
|---|---|---|---|---|---|---|---|---|
| | Availability: target value [%] | Reliability: Mean Time btw Failure | **e2e latency** | Bit rate | Message Size [byte] | Transfer Interval | # of active UEs | **Service Area** |
| Local Robo surgery | > 99.999 99 9 | > 10 years | **< 2 ms** | 2 Mbit/s to 16 Mbit/s | 250 to 2,000 | 1 ms | 1 | **room** |
| Tele-surgery | > 99.999 9 | > 1 year | **< 20 ms** | 2 Mbit/s to 16 Mbit/s | 250 to 2,000 | 1 ms | **< 2 per 1,000 km²** | **national** |
| Robo Tele-diagnosis | > 99.999 | >> 1 month (< 1 year) | **< 20 ms** | 2 Mbit/s to 16 Mbit/s | ~80 | < 20 ms / 100 km² | **< 20 per 100 km²** | **regional** |
| Direction: UL/DL; Survival time: transfer interval; UE speed: static; *Clock sync<50Us*; Jitter 3-30 ms :3D video, <2ms : haptic | | | | | | | | |

### 6.1.2 Dual-User Shared Teleoperation (DUST): Tele-surgery

Dual-User Shared Teleoperation (DUST) configuration refers to structures that have two leader stations for the collaboration of two operators on one follower console. Tele-operation robots are tightly collaborating with humans and may require HRI, meaning information and action exchanges (e.g., through vocal, visual and tactile means) between human and robot to perform a task in the same space by means of a multimodal user interface. Furthermore, the 6G network enables remote monitoring and control of the collaborative robots, allowing for greater flexibility in managing production lines. For example, if a worker needs to adjust a robot's movement or programming, they can do so from a remote location using a mobile device connected to the 6G network.

#### Pre-conditions

Robots are equipped with 6G ISAC capable UE. (2) Robot(s) interact with human with implicit or non-implicit task assignment. (3) Robot with local control and AI capability (4) human (expert) in the loop (Human expert is remote and human operator is on-site) (5) The local and remote surgeons are trained and have permission to perform such remote procedures.

#### Description

Collaboration helps improve the task execution in comparison with the same task done individually. The concept of haptic-enabled negotiation between two operators can be made feasible using this framework to control the follower console collaboratively through haptic communication. Tele-operation scenario to allow skilled surgeons/operators to monitor, advise and take over as required from a distant location.

The depicted example scenario in Figure 13 is tele-surgery/tele-diagnosis- same challenges apply for industry remote operation where an engineer is remotely supporting/operating the onsite staff. Cage-free hybrid cell setup based on the industry set up will add the safety requirements for onsite workers.

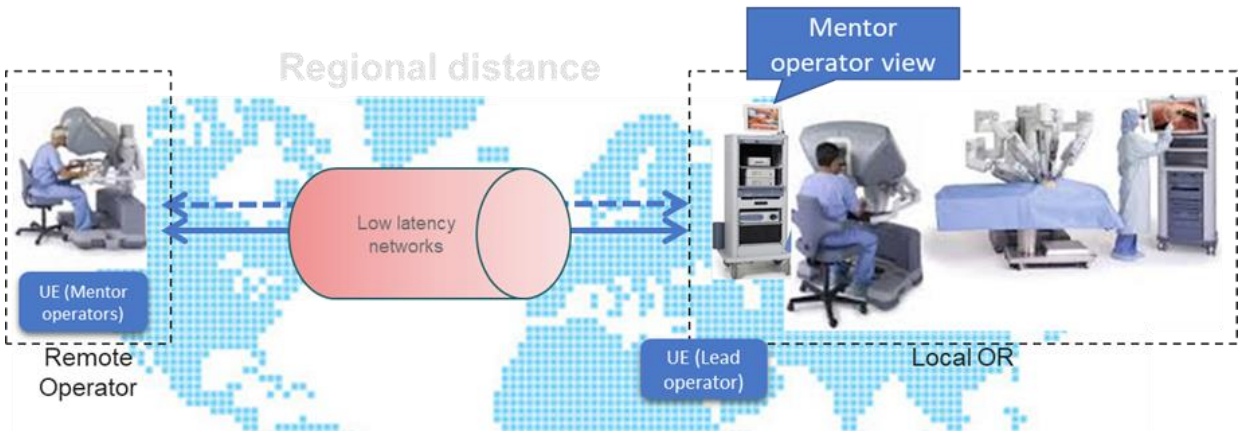



*Figure 13: 6G-Tactile-DUST Operation room (OR).*

#### Key techniques

IEEE 1918.1 [1] and 3GPP TACMM [15] and XRM [16] features describe main techniques to enable collaborative operation:

- Low latency and reliable communication over regional distance
- Multimodality synchronization

- Service availability
- Safety-rated monitored stop, collision avoidance
- Safety zone setup (cobotic cells/hybrid cells)

## Service Flows

- The patient/local surgeon authorizes the remote medical intervention.
- The surgeon was trained and has permission to perform such remote procedure.
- The surgeon controls the patient side card remotely through his surgeon console to perform the operation

## Post-conditions

Patient was operated in remote areas, where there is a lack of experienced medical personnel to perform the needed medical intervention. The patient has now his/her health conditions improved.

## Potential Requirements

- The system shall consider edge console computing in order to reduce the latency in the network: low latency (<20 ms), achievable data throughput, roundtrip interaction delay (<50 ms for ultra-low latency applications and <100 ms for low-latency applications);
- The surgeon console and the patient side console shall have connection to THz access point (Section 1);
- The system shall consider the use of AI to manage the patient's data in the 6G network and help in the medical interventions;
- The system can support haptic and tactile technology;
- The system can support multi-party (source) data stream synchronization in very low latency environments;
- The system shall consider AR in order to enable holographic image transfer with ultra-efficient data transmission techniques;
- Environment model accuracy (including sensing, intelligence, communication);
- Environment model convergence time;
- Safety zone update time (including dynamic human behaviour model);
- Connectivity latency, reliability, deterministic, service coverage, etc.;
- Sensing accuracy;

Multimodal synchronization for communication and control;

Table 6 shares a summary of 5G KPIs requirements and highlights where the number of active UE's need to be reviewed to extend in the service area requirements for 6G DUST.

*Table 6: [3GPP 22.104] blue arrow highlights where the 5G KPIs need to potentially be extended for Telesurgery use case.*

| Use case [3GPP 22.104] | Characteristic parameter | | | | Influence quantity | | | |
|---|---|---|---|---|---|---|---|---|
| | Availability: target value [%] | Reliability: Mean Time btw Failure | **e2e latency** | Bit rate | Message Size [byte] | Transfer Interval | # of active UEs | **Service Area** |
| Local Robo surgery | > 99.999 999 | > 10 years | **< 2 ms** | 2 Mbit/s to 16 Mbit/s | 250 to 2,000 | 1 ms | 1 | **room** |
| Tele-surgery | > 99.999 9 | > 1 year | **< 20 ms** | 2 Mbit/s to 16 Mbit/s | 250 to 2,000 | 1 ms | **< 2 per 1,000 km²** | **national** |
| Robo Tele-diagnosis | > 99.999 | >> 1 month (< 1 year) | **< 20 ms** | 2 Mbit/s to 16 Mbit/s | ~80 | < 20 ms / 100 km² | **< 20 per 100 km²** | **regional** |
| Direction: UL/DL; Survival time: transfer interval; UE speed: static; *Clock sync<50Us*; Jitter 3-30 ms :3D video, <2ms : haptic | | | | | | | | |

### 6.1.3 Tele-Edutainment

The development of extended reality (XR) technologies (e.g., VR, AR, MR) provides an opportunity to develop new education and training practices. By adding layers of contextually relevant information and delivering it visually or by activating other senses with the help of wearable and handheld devices, AR offers an enhanced picture of the physical environment. Immersive technologies can make learning a fun and enriching experience, especially for young students, through the use of immersive classrooms. Virtual environments have the capability to develop children's core skills and imagination by providing endless opportunities for exploration and interaction. When engaging in XR environments, the learning surrounding should be enriched with tactile feedback to enhance the level of immersion and create a more realistic environment for the user. By adding touch, learners will also experience social connection with the environment, but also with other human learners. The learning experience can be enhanced by combining XR techniques with AI-based object and language recognition, mnemonic-based training exercises and sensory stimulation that relate to the personal context of the learner.

#### Pre-conditions

- XR environment
- AI
- haptics
- human-in-the-loop

#### Description

Tele-edutainment can be seen as a real-time collaborative activity, which is characterised by simultaneous support of very high throughput and very low latency across multiple concurrent and synchronized communications channels. The effectiveness of this use case depends on the ability to predict the user's movement and emotional expressions, thus, rapidly adapting the data as needed. This is the "user interactivity challenge," and it enforces ultra-high bandwidth and ultra-low latency (to ensure interactivity with the content). In addition, perfect synchronization of concurrent flows will be needed. Contrary to other types of multimedia services (e.g., UHD streaming), ultra-low latency is required even if dealing with pre-recorded content that does not involve real-time interaction with a remote party, as the user still interacts with the content simply by changing the viewing angle and position. Further, highly precise 3D models of the human face, body, and clothing, as well as recognition and prediction of human movements and facial emotions, and streaming the data in real time requires compression and decompression, which come at the price

of additional computation, which will heavily influence the latency incurred. Therefore, a higher level of compression trades off computation bandwidth and latency vs required networking bandwidth and latency, and vice versa. Even with compression, streaming of data generated by the capture process requires massive bandwidth. For example, a capture of a 3D frame including mesh geometry is about 5 MB with compression so for real-time performance, an average per frame transmission of 1-2 Gbit/sec with overhead for compression (< 10 ms) is a must. While people can experience time delays of approximately 16 milliseconds or greater, the data requirement of 3D models is assumed to be at terabytes. The real-time holographic transmission requires 10 Gbps or higher bitrate to use current compression techniques.

### Key techniques

holographic communications.

### Potential Requirements

- low latency (<20 ms), achievable data throughput, roundtrip interaction delay (<50 ms for ultra-low latency applications and <100 ms for low-latency applications);
- Support for processing capability at the edge with < 5 ms round-trip time
- support of multi-party (source) data stream synchronization in very low latency environments;
- Network quality of service: prioritization of real-time traffic and mechanisms to ensure consistent performance across different network conditions; possibly support for dedicated network slices.
- The system shall consider AR in order to enable holographic image transfer with ultra-efficient data transmission techniques.
- The system may provide multimedia conversational communication services between two or more users to enable real-time interaction.
- Robust security and privacy
  - End-to-end encryption for all communications (e.g., AES-GCM with 256-bit key length for symmetric encryption or Curve25519 for asymmetric encryption) with encryption performance at latency overhead: < 2ms for encryption/decryption and at least 5 Gbps throughput to handle high-quality VR video streams.
  - Secure authentication and access control systems: Secure, encrypted channels for real-time updates of shared virtual objects and verification of data integrity for synchronized elements: < 50ms.

## 6.1.4 Meta-collaboration: XR enabled collaborative engineering

Since the industrial age, engineering design has become an extremely demanding activity. Collaborative and concurrent engineering occur as a concept and methodology at the end of the last century and was defined as a systematic approach to integrated co-design of products and their related processes. The diversity and complexity of actual products require collaboration of engineers from different geographic locations to share ideas and solutions with customers and to evaluate products development. VR and AR technologies have found their ways into critical applications in industrial sectors such as aerospace engineering, automotive engineering, and medical engineering. The range of technologies include Cave Automatic Virtual Environment (CAVE) environments, reality theatres, power walls, holographic workbenches, individual immersive

systems, head-mounted displays, tactile sensing interfaces, haptic feedback devices, multi-sensational devices, speech interfaces, and mixed reality systems. One of the key challenges is how to enable a distributed virtual environment (DVE) allowing multiple users from different geographical locations (some of them are present at the same location) to interact over a network.

All the performed tasks and knowledge acquired both in the physical world and metaverse will be shared in a DVE (with the corresponding 6G communication subscriptions) for collaborative and cooperative engineering in their product design with engineers participating locally and remotely, e.g., to work together with several partner companies to design and produce a new model of aero-plane engine. Some engineers use mobile phones or computers (as well as the necessary XR devices and tactile gloves) to attend such engineering meetings.


*Figure 14: a: XR enabled collaborative and concurrent engineering in product design; (Source: https://vrtech.wiki/); b: Illustration of Collaborative Workspace (Source: ESI-Icido GmbH) [3GPP metaverse feature]*

## Pre-conditions

(1) A distributed virtual environment set up for collaborative engineering in the product design with engineers participating locally and remotely, (2) engineers use mobile phones, computers, the necessary XR and multimodal devices, with the corresponding 6G communication subscriptions, to attend such engineering meetings, (3) mobile metaverse services have subscribed with different network slice for these different types of service, and different QoS for different flows accordingly for better user experience (4) strict security requirements for user identity management and data security to protect the sensitive business information.

## Key techniques

IEEE 1918.1 and 3GPP TACMM, XRM and Metaverse features describe the main techniques to enable collaborative operation:

- Low latency and reliable communication over local and regional distances;
- Multimodality synchronization;
- Service availability.

## Potential Requirements

- Latency, data rate, and reliability;

- Mobile metaverse media support among multiple users;
- Native support for mobile metaverse services;
- User Identity management;
- Data security.

Table 7 [18] highlights where the 6G KPIs need to be reviewed to extend the number of active UEs in the service area requirements for 6G metaverse tele-collaboration.

*Table 7: Typical QoS requirements for multimodal streams [3GPP TACMM feature]*

| | Haptics | Video | Audio |
|---|---|---|---|
| Jitter (ms) | ≤ 2 | ≤ 30 | ≤ 30 |
| Delay (ms) | ≤ 50 | ≤ 400 | ≤ 150 |
| Packet loss (%) | ≤ 10 | ≤ 1 | ≤ 1 |
| Update rate (Hz) | ≥ 1000 | ≥ 30 | ≥ 50 |
| Packet size (bytes) | 64-128 | ≤ MTU | 160-320 |
| Throughput (kbit/s) | 512-1024 | 2500 - 40000 | 64-128 |

### 6.1.5 AMR remote control in a future factory

Autonomous Mobile Robots (AMRs) play a critical role in the industrial sector by automating material handling and transportation tasks, replacing manual labour. These vehicles are equipped with advanced sensors and collision avoidance systems to ensure safe operations, especially when interacting with humans and other machines. AMRs offer the flexibility of being programmable and reprogrammable, allowing them to handle various tasks and adapt to changing production requirements. As a result, AMRs contribute to the development of efficient and secure industrial assembly and operation lines. AMR can be divided in operation in indoor, outdoor and both indoor and outdoor areas. These environmental conditions have an impact on the requirements of the communication system, e.g., the handover process, to guarantee the required cycle times [17].

As smart manufacturing and Industry 4.0 trends continue to mature, AMR technologies are receiving increasing attention due to their significant applications in various industry verticals. However, the current development and implementation of AMRs primarily follow a control-centric paradigm [28], focusing on the design of complex control algorithms and leveraging sensory capabilities. The shortcomings associated with this approach are becoming apparent. For instance, it is unable to handle versatile operational environments and has difficulties in perceiving the environment. Communication-based control offers an alternative to the conventional control-centric approach while overcoming many of its limitations [29][30]. This alternative approach places emphasis on the exchange of information and communication among the various components of a robot system, by establishing or leveraging effective communication channels. It fosters coordination and collaboration among different subsystems of a robot. The concept of communication-centric control has been extensively explored in the literature and has demonstrated significant benefits, particularly when combined with the capabilities of advanced communication networks. That paradigm will simplify the design of AMR systems, to create a more streamlined and interconnected network of subsystems, so as to offer new opportunities for efficiency and productivity. An illustration of the Toshiba's AMR testbed for handling pallets with varying sizes and shapes is shown in Figure 15.

*Figure 15: Illustration of the AMR testbed for handling pallets with varying sizes and shapes.*

## Pre-conditions

1) AMR node equipped with a rich set of sensors and able to share the locally sensed information observations through the 6G wireless interface/network. (2) AI/ML techniques network management/support of the networks. (3) Communication and control co-design techniques.

## Description

It is anticipated that more benefits can be realized by embracing future 6G for AMR control. While the specifications and standards of 6G are still being defined, its most prominent features are ultra-reliable and low-latency connections, native AI support, and cloud/edge intelligence enhancements. As a result, the following advancements can be envisioned in AMR solutions empowered by 6G. (1) Enhanced AMR robot coordination ability: With 6G connectivity, the coordination ability of AMR robots can be further strengthened. Through rapid data sharing, these robots can exchange operational states swiftly, enabling them to make informed decisions based on collected observations. (2) Empowered real-time data processing and remote control: The ultra-reliable and low-latency connections offered by 6G will empower AMR systems with real-time data processing and remote-control capabilities. This becomes particularly valuable in specialized application scenarios, such as those involving high temperatures or pressures that restrict human on-site operations. In such cases, remote control of AMRs proves extremely helpful. (3) Improved scalability: Leveraging the native AI support of the 6G network, AMR control can implement various types of distributed ML algorithms. This enables AMR control to be conducted in a partitioned and hierarchical manner, significantly enhancing the scalability of the AMR system [31][32]. (4) Enhanced co-design: The communication and control co-design approach involves joint design and optimization of communication and control systems, where the former plays the critical role of communicating control information and feedback signals among system components. This approach can be further empowered by incorporating 6G connectivity, as it is expected to offer enhanced robustness, agility, and deterministic behaviour. By leveraging the capabilities of 6G, the complexity of co-design can be reduced, leading to improved performance in AMR systems. (5) Automated planning of complex tasks: By harnessing the native AI capabilities of 6G and leveraging associated cloud/edge intelligence, the future AMR system is expected to achieve automated planning of complex tasks. This includes determining optimal paths and collaborative methods for the robots. The control policies of the AMR system can be managed in a unified manner, streamlining operations.

## Key techniques

- Cloud, distributed and hierarchical computing and machine learning supported by 6G networks;
- multi-service edge-intelligence control paradigm;

- Autonomous planning of AMR;
- Communication technologies that employ different frequency bands (IEEE 802.11 2.4-GHz and 5-GHz frequency bands).

### Potential Requirements

- Low response times (<20ms) and deterministic connectivity
- AMR self-localization (centimetre-level) leveraging the 6G ISAC feature
- Holistic information/task exchange and synchronization mechanism among AMR robots for communication and control co-design.

## 6.1.6 Remote driving

Remote driving, also known as teleoperation, is an emerging field in the domain of autonomous vehicles that aims to enable human operators to remotely control vehicles from a distant location [33]. With advancements in communication technology and the increasing deployment of autonomous vehicles, remote driving has gained significant attention for its potential applications in various industries such as transportation, logistics, and public safety. Remote driving systems typically involve a combination of high-speed data transmission, sensor feedback, and control interfaces that allow operators to perceive and interact with the vehicle's surroundings in real-time.

Multimodal learning and communication are effective enablers for interaction and understanding between human operators and autonomous vehicles. Through the integration of multiple sensory modalities, such as visual, auditory, and haptic feedback, operators can gather comprehensive information about the vehicle's surroundings and make informed decisions. Multimodal communication can facilitate bidirectional information exchange between the human operator and the vehicle. The communication requirements for remote driving include:

- Traffic Direction: Bidirectional traffic between the human operator and the vehicle.
- Traffic Type: Includes control signals, sensor data, video/audio streams, and system status updates.
- Burst Size: The burst size can vary depending on the specific application and the frequency of data updates. For instance, control signals might have small burst sizes as they convey concise instructions, while video/audio streams could have larger burst sizes due to the continuous transmission of sensory information.
- Reliability: Reliability is a critical requirement for remote driving communication.
- Latency: Latency requirements typically aim for values below 100 milliseconds, and certain applications might require even lower latencies in the range of 10 to 50 milliseconds.
- Average Data Rate: The average data rate is the amount of data transferred per unit of time. It depends on the specific communication requirements of the remote driving system, including the complexity of the control signals, resolution of video streams, and the frequency of sensor data updates.

### Precondition

Vehicles equipped with cameras, LiDAR, radar, GPS, accelerometers, steering and braking systems.

### Technology

Current standardization activities for the next generation wireless connectivity in vehicular networks are based on orthogonal frequency division multiplexing (OFDM) as by 3GPP NR V2X. In 3GPP TR 22.885, there exist the following three different types of V2X: Vehicle-to-Vehicle (V2V), Vehicle-to-Infrastructure (V2I), Vehicle-to-Pedestrian (V2P) Communications. The support of V2X by 3GPP is provided over the PC5 interface by sidelink (SL) transmissions. The main characteristic is the use of mini-slot scheduling, which allows to schedule transmission of a slot for latency-critical services. Improved localization of vehicles is made possible using large antenna arrays, and the exchange of safety-based (V2X, ETSI-compliant) messages. Another aspect is the definition of the resource allocation modes, where the base station schedules the resources, and, which lets the user equipment (UE) autonomously select the SL transmission resources. In the USA, The Federal Communications Commission (FCC) has endorsed the C-V2X as the standard for automotive uses.

### Potential requirements

- Remote driving encompasses various key features, including driver alerting and environmental awareness. Level 1 communication primarily relies on CAM (Cooperative Awareness Messages) and DENM (Decentralized Environmental Notification Messages). The transmission frequency can reach up to 10 Hz, such as in the case of emergency vehicle warnings, or lower frequencies for roadwork warnings;
- sensor orchestration that can be distributed with tight latencies;
- Distributed antenna deployment with 10-100 GHz bandwidth in support of fast association/disassociation with RSUs and strict air-interface jitter (<1 ns);
- Fully AI-based communication and sensing is required to support control and programmability;
- For this use case, the Connected and Autonomous Vehicles (CAVs) require (i) High-frequency Communications Communication systems operating in the mmWave (30 -100 GHz), sub-THz (100 - 300 GHz), THz (300 GHz - 10 THz) and optical (> 10 THz) bands are envisioned to be pillars of 6G V2X networks [34], (ii) resilient Network Technologies and smart V2X protocol architectures to handle the complex mission critical interactions that require flexible computing and communication frameworks;
- Other important requirement is related to sensing and localization which 6G V2X connectivity can handle through orchestration of sensing by sharing the measurements over a common virtual bus, by joint communication and sensing networks, hybrid positioning when GNSS signals are discontinued.

## 6.1.7 Emergency access to buildings in case of disasters

First responder services such as fire fighters and search and rescue units are frequently exposed to challenging access to indoor environments in a search for affected persons or animals. The main obstacles include sudden changes of connection quality and positioning accuracy while moving between outdoor and indoor environment, and lack of information about the indoor environment. Information about the location and current status of victims are also usually unavailable or possibly inaccurate. In such a situation, the team might easily take wrong decisions that will reflect on the time spent in search and rescue mission and increase the chance of mission failure.

This use case requires multimodal communication, sensing and positioning to demonstrate how the team of humans and uncrewed autonomous mobile vehicles can access the building. During the mission critical emergency operation, it is of utmost importance to maintain uninterrupted and Quality of Service (QoS)-aware access to communication and compute resources to receive relevant

multimodal data streams and execute needed AI inference models while receiving reliable information on position and status of affected entities within the building. The future solutions in 6G environment will use evolving wireless infrastructure (6G, but also augmenting next-generation Wi-Fi and Li-Fi technologies) supported by UAV nodes and next-generation 6G ambient IoT sensors [101] integrated with AI/Edge compute platform components both attached to humans/autonomous vehicles and deployed at network edge close to base stations to support the required resilience and efficiency. Exploiting AR/XR technologies, positioning in an unknown indoor environment and interaction with surrounding sensors, future first responders may rely heavily on multimodal learning, inference and multimodal positioning services to support optimal decision making [102].

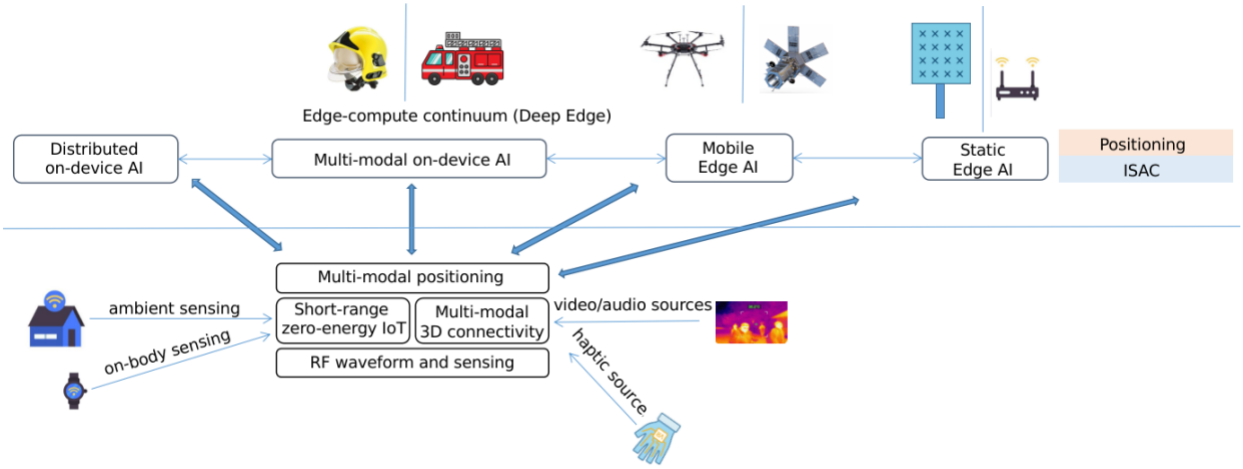



*Figure 16: Enabling technologies for 6G multimodal public safety services.*

## Precondition

Multimodal communication, sensing and positioning may significantly improve the prospects of search and rescue mission by augmenting the perception of human and robotic actors through exchange of multimodal content between the relevant devices and infrastructure. To be able to benefit from the full capacity of such services, the future buildings in urban environments should be covered by both outdoor and indoor cellular technologies augmented with next-generation Wi-Fi and Li-Fi deployments. Additional 3D networking support from on-site deployed UAVs is needed in case of major disruptions of cellular infrastructure. Access to beyond-5G multimodal services through AR/XR equipment (e.g., AR/XR glasses) and their seamless information exchange with surrounding ambient indoor IoT sensors underlies the possibility for augmented perception of emergency crews. This is complemented by further improvements of indoor positioning, tracking and mapping including the seamless transition from outdoor to indoor environments that should be provided by upcoming wireless cellular technologies at any location.

## Technology

The discussed scenario depends on the number of technological enablers:

- Integrated and flexible air interface for 3D wireless communication and sensing, that supports seamless user communication with terrestrial and aerial nodes and acquisition and exchange of sensing information
- Further integration of AI in 6G RAN that goes beyond current 3GPP efforts, enabling efficient and distributed execution of multimodal learning algorithms at end users and edge infrastructure

- Pervasive ambient short-range low-energy (including zero-energy) IoT connectivity through novel massive IoT-enabling sidelink interfaces that support sidelink relaying and multi-hop ambient IoT configurations
- Evolution of multimodal high-throughput communications to support future AR/XR services and their integration with surrounding ambient IoT and on-board or edge-AI compute infrastructure for provisioning future augmented perception services.
- Enhanced positioning, tracking and mapping indoor infrastructure integrated in 6G that exploits all available multimodal information sources to establish highly accurate positioning of objects in indoor environment and seamless positioning service transition between outdoor and indoor spaces.

### Potential requirements

Various service KPIs will have to be improved compared to today's technologies in order to enable efficient in-building disaster relief services:

- Latency is anticipated to be further reduced to sub-1 millisecond level, with objectives targeting as low as 0.1 ms in certain ultra-reliable and low-latency communication scenarios. This improvement is pivotal for applications necessitating instantaneous response times among indoor autonomous vehicular and aerial units and their mutual interaction and interaction with human units in real-time.
- Data rate expectations for 6G are set to reach unprecedented peaks, aiming for up to 1 Tbps to accommodate the exponential increase in data consumption driven by immersive augmented and virtual reality applications, enhanced holographic content, and advanced media streaming services. This requirement will be relevant for indoor disaster recovery missions by enabling richer and more immersive crew experiences.
- Localization accuracy will see significant improvements, with 6G networks expected to achieve centimetre-level precision, facilitated by integrated sensing and communication technologies. This high level of accuracy will be instrumental in safety scenarios requiring precise location information of many surrounding entities, supporting both outdoor and indoor uncrewed ground or aerial navigation, and automated exchange of critical positioning information for augmented reality services.
- Integrating AI with 6G will lead to more intelligent and autonomous network operations, including predictive reactions during search and rescue missions, dynamic resource allocation and decision making, and enhanced network security. AI integration will play a critical role in optimizing network performance and user experiences while minimizing operational costs and energy consumption.

## 6.2 State-of-the-art and discussion topics

Significant recent interest in multimodal remote operation use cases is evidenced by academic publications [1]-[12], industrial demonstrations [2] and recently started research projects for remote operation which demands a reliable and stable network operation. For high precision tele-operations Tactile and multimodal robots can be used in a variety of use cases such as Tactile Internet with Human-in-the-Loop (TaHiL) [23], smart factory [24], emergency [25], and medical use cases [26].

Tactile Internet is indeed the innovation leap enabling wireless remote control of collaborative robots over mobile network, which is currently revolutionizing several sectors such as healthcare, manufacturing, agriculture, construction, and logistics [6]. As 2G and 5G were the media for

communicating voice and mobile data, Tactile Internet is the medium where haptic communications take place [2]. However, to fully realize Tactile Internet, apart from extremely high reliability (seven nines) and low latency (less than 1 ms), 6G needs to provide also extremely high data rate to support the complementary high definition 4K/8K 180°/360° video for XR/VR applications to provide a fully immersive experience for the remote human operator [6].

Since the introduction of 1G to 4G mobile networks, every odd-numbered generation (1G and 3G) has introduced new mobile services (voice and data, respectively) to business customers, while the even-numbered generations (2G and 4G) have democratized this and extended those services to ordinary customers [6]. Extrapolating this trend, besides enhanced mobile broadband for consumers, 5G was the first generation introducing URLLC and promising the introduction of Tactile Internet. However, URLLC and haptic communication services are currently limited to vertical industries. Hence, 6G is expected to be the first generation extending URLLC and haptic communications from vertical industries to every individual citizen to provide an infrastructure to support remote control and mobile robotic solutions for everyone – realizing the vision of Personal Tactile Internet [6]. In terms of communication requirements for Tactile Internet applications, there are further constraints that must be faced. With regards to sensors and data acquisition, many of these devices must share their information at high data rates to ensure the Quality of Experience (QoE) of the user. This aspect is currently being researched and studied [7] to improve scheduling algorithms at MAC level. These enhancements include the introduction of Quality Indicators (QI) required for the application to modify the performance of the network, and thus the user QoE.

In parallel, the Internet has drastically evolved over the past three decades - going from the fixed Internet, connecting personal computers, which fuelled the economy in 1990s and 2000s, to the mobile Internet, connecting mobile phones and tablets in late 2000s and 2010s. The Internet then evolved to Internet of Things, connecting every single object (big or small) in the physical environment, which enabled a wide spectrum of smart applications based on IoT connectivity. The introduction of IoT was also a starting point for the coexistence of radio-based sensors (e.g., Radar and UWB indoor localization) with wireless communication interfaces. Now, with the deployment of 5G and the realization of Tactile Internet, the Internet is evolving into the Internet of Skills, enabling the delivery of specialized skills (surgery, maintenance, engineering, design, etc.) through the Internet anywhere in the world rather than content delivery over the current Internet [2]. This relies on real-time multimodal communications using voice, video, and haptic data across the globe. However, the required latency of the Internet of Skills is between 1 ms and 10 ms, which reduced the communication distance between 100 Km and 1000 Km, under typical network conditions [10]. Using AI for predicting the model and the network latency, this can be extended to tens of thousands of kilometres [8].

Digital twin is another cornerstone of remote operation and networked control use cases over 6G, which produces a digital replica of the physical and biological worlds at every spatial and time instant unifying the user experience across the physical, digital, and human worlds. To create an accurate and up-to-date twin of the physical world for seamless interactions with the remote operator and controller, an enormous amount of capacity with ultra-low latency and high reliability is required. Any excess delay, jitter, or packet loss in updating the digital twin and/or applying the control signal may significantly compromise the user experience and even destabilize the control system, which may result in severe consequences [9]. Apart from Terabit per second data rate, massive capacity, and supporting URLLC at scale, 6G is expected to 1) integrate radio sensing into RAN infrastructure for better use of spectrum and unified hardware design supporting both functionalities; 2) unify computing resources distributed across multiple edge nodes and the cloud that can be managed by different operators; 3) support multi-sensor fusion for better environmental awareness and mixed-reality experience; and 4) enable actuation to control the physical processes. On the user device side, new human-machine interfaces need to be developed for multimodal interaction (including haptic modality) between the human operator and the digital/physical world. These augmented reality user interfaces will enable efficient and intuitive human control of the physical, virtual or biological worlds. Last but not least, new security and privacy preservation and trust mechanisms need to be developed in order counteract new security attacks on multimodal 6G communication networks, e.g. jamming attacks on industrial networks

coming from outside the industrial facility, which makes physical protection insufficient [9]. A key to the 6G development is to provide a unified, distributed, and agile framework to enable data pipelines coming from the entire sensing stratum that includes distributed endpoints (e.g., embedded devices) with sensing and actuating capabilities, edge platforms (e.g., edge/cloud computing servers, fog nodes, micro-datacentres), and remote cloud datacentres [100].

In the following, we provide existing activities in the area including projects, platforms, and data sets available for multimodal research and analysis.

### 6.2.1 Multimodal teleoperation related projects

Multimodal sensing, communications and control enabled remote operation use cases, testbeds and pilots dominate demonstration aspects of majority of recent 6G-focused research initiatives. Extreme and conflicting requirements for data rates, latency, reliability, AI/ML integration present in multimodal remote teleoperation use cases described in the previous section need to be verified in pilot proof-of-concept demonstrators which is a commonplace ambition of 6G research initiatives. In Europe, as part of 6G Smart Networks and Services (SNS) framework within Horizon Europe research and innovation funding managed by European Commission, the first batch of 35 projects have been approved. In the table below, we present a representative sample of research projects, mainly funded by 6G SNS scheme, that develop use cases, testbeds or field trials in the domain of remote multimodal operation over 6G.

*Table 8: Multimodal teleoperation related projects*

| Related projects | Field trials and use cases | Description |
|---|---|---|
| Umbrella | Swarm robotics and other Industrial IoT use | This cobot testbed is part of a larger open Industrial IoT testbed which has been deployed in the UK. It relates to warehouse robotics, which aims to permit flexible experimentation using different end devices to evaluate algorithms or new practical application scenarios as each robot node supports multiple sensors and a number of wireless communication technologies [35]. |
| 5G-encode | AR/VR to support design, manufacturing and training<br>Monitoring and tracking of time sensitive assets<br>Wireless real-time in-process monitoring & analytic | 5G-encode is a collaborative project in the UK, aiming to develop clear business cases and value propositions for 5G applications in the manufacturing industry. |
| VERGE | XR-driven edge-enabled industrial B5G applications<br>Edge-assisted Autonomous Trams | The first use case focuses on the design of industrial products that require collaboration of engineers from different locations. Using 6G XR tools, designers can work on the same product design simultaneously thus avoiding costly and time-consuming travel. In the second use case, autonomous trams integrate multimodal sensors such as lidar, radars, GNSS, IMU, infrared and visible cameras to promptly detect and avert critical situations. Due to a high computational complexity, computation is partly offloaded to the edge, where it is fused with other smart city IoT data for improved situational awareness. |

| Related projects | Field trials and use cases | Description |
|---|---|---|
| ADROIT-6G | Extreme eMBB for immersive Extended Reality<br>Extreme URLLC for collaborative robots | The ADROIT-6G use cases will aim to: 1) demonstrate the proposed AI network architecture to support holographic telepresence services, 2) focus on distributed and collaborative robot and drone systems. |
| TARGET-X | Idiada automotive testbed<br>RWTH Aachen Industry Campus Europae | Idiada focuses on 5G and beyond technology use cases and demonstrations in the domain of cooperative perception of connected vehicles and digital road twins. RWTH Aarched Industry Campus Europae is a large-scale 5G industrial testbed for connected robotics (lineless mobile assembly lab), connected construction site and connected and smart energy grid. |
| TRIALS-NET | Large number of trials for beyond 5G | Support for a large number of beyond 5G use cases in the domain of infrastructure, transportation, eHealth, culture, tourism and entertainment |
| FIDAL | Field trials beyond 5G | Three large-scale test infrastructures in Greece, Norway and Spain targeting the augmentation of human capabilities, media and vertical industry players to perform advanced large-scale field trials. |
| IMAGINEB5G | Large-scale experimentation facilities for various use cases and field trials. | Four advanced 5G experimental facilities in Norway, Spain, Portugal and France to support firefighting and forest surveillance, port surveillance and inspection, crisis management, media production, localisation for transportation and logistics, telepresence-aided maintenance, Industry 4.0, education, smart agriculture and forestry. |
| 6G-SANDBOX | Automated experimentation capabilities through a rich toolbox. | Fully configurable, manageable and controlled end-to-end networks, composed of both digital and physical nodes on distributed infrastructure located in Malaga, Oulu, Athens and Berlin. |
| 6G-BRICKS | Building reusable testbed infrastructure | Deliver an evolvable 6G experimentation facility that will integrate 6G technologies and federate two testbeds to validate and showcase advanced use cases in holographic communication, metaverse and digital twinning. |
| 6G-XR | experimental research infrastructure to enable next-generation XR services | 6G-XR will develop an experimental infrastructure demonstrating the performance of key B5G/6G candidate technologies, components, and architectures and focusing on demanding immersive applications such as holographic, digital twins and XR/VR. |

### 6.2.2 Multimodal Datasets

Many of the possible tele-operation use cases rely to some extent on extended reality (XR), as a group of technologies that seek to enhance and expand human perception of the physical world. The interactions between the multimodal sensors and the wireless network are studied in [59], where the data exchange between the algorithms and their hardware requirements are analysed, leading to the identification of the required network key performance indicators (KPIs).

An XR offloading IP traffic dataset is available in [60], which can be used for simulation or prototyping. It contains traffic traces that have been captured using a novel implementation of a Transmission Control Protocol (TCP) for XR offloading. Besides, a full stack 5G Radio Access Network

(RAN) emulator fully scalable in terms of users and throughput handling capabilities is described in [61]. DeepSense 6G [62] is a multimodal dataset that comprises coexisting multimodal sensing and communication data, such as camera, GPS data, LiDAR, and radar, collected in 30+ scenarios and over several locations covering a diverse vehicular, drone, RIS, human, robots for indoor/outdoor scenarios are designed based on realistic wireless environments to enable different novel applications. A list of other available datasets mainly from IEEE DataPort platform are listed in Table 9.

*Table 9: Publicly available multimodal dataset*

| Description | Feature | Dataset/analysis | Papers/others |
|---|---|---|---|
| Dataset for Influence of Visual and Haptic Feedback on the Detection of Threshold Forces in a Surgical Grasping Task [63] | Tactile, multimodality, Haptics, surgical robotics | Anova analysis; Force threshold vs Iteration analysis; Threshold force analysis | Data in brief, Elsevier 2022 [64] |
| Opportunity++: A Multimodal Dataset for Video- and Wearable, Object and Ambient Sensors-based Human Activity Recognition [65] | AI and machine learning research focused on the multimodal perception and learning of human activities | 19.75 hours of sensor data annotated with multiple tracks; sensors placed on the objects produced a total of 3.92 hours of annotated data; Overall, the dataset comprised of more than 24000 unique annotations | Frontiers in Computer 2021 [66] |
| VibTac-12: Texture Dataset Collected by Tactile Sensors [67] | Signal feature extraction method and the Fourier transform as input to machine learning classifiers | 3D accelerometer recordings; sound recordings | IEEE Access 2020+ publication record from 2010 [67] |
| A Unified Perception Benchmark for Capacitive Proximity Sensing Towards Safe Human-Robot Collaboration (HRC) [68] | The perception gap between tactile detection and mid-range | Speed and Separation Monitoring (SSM)<br>Power and Force Limiting (PFL)<br>Capacitive Proximity Sensors (CPSs) | IEEE ICRA, 2021 Xi'an, China + Based on ISO/TS 15066 [68] |
| ExoNet Database: Wearable Camera Images of Human Locomotion Environments [69] | Large-scale hierarchical database of high-resolution wearable camera images (i.e., egocentric perception) of legged locomotion environments | Over 5.6 million RGB images of indoor & outdoor real-world walking environments<br>Approximately 923,000 images in ExoNet were human-annotated using a novel, 12-class hierarchical labelling architecture | Frontiers in Robotics and AI 2020 [70] |
| Multimodal grasp data set: A novel visual–tactile data set for robotic manipulation [71] | Integration of visual and tactile data for understanding the grasping process and deeper analysis of grasping issues | 2550 sets data, including tactile, joint, time label, image, and RGB and depth video | International Journal of Advanced Robotic Systems, Sage Journals 2019 [71] |

| Description | Feature | Dataset/analysis | Papers/others |
|---|---|---|---|
| Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection [72] | First experiment: Learning-based approach to hand-eye coordination for robotic grasping from monocular images.<br>Second experiment: test transfer between robots, how data from a different set of robots can be used to aid learning | First experiment, about 800,000 grasp attempts using between 6 and 14 robotic manipulators with differences in camera placement and gripper wear and tear.<br>Second experiment with different robotic platform and 8 robots to collect a dataset consisting of over 900,000 grasp attempts. | International Journal of Advanced Robotic Systems, Sage Journals; 2019 [72] |
| Complex urban dataset with multi-level sensors from highly diverse urban environments [73] | Light Detection and Ranging (LiDAR) and image data acquired in complex urban environments | LiDAR sensor data (2D and 3D), navigation sensor data with commercial-level and high-level accuracy, stereo camera images of the environment, position information of the vehicle estimated through simultaneous localization mapping (SLAM) | International Journal of Advanced Robotic Systems, Sage Journals; 2019 [73] |
| Create: multimodal dataset for unsupervised learning and generative modelling of sensory data from a mobile robot [74] | Mobile robot that can navigate and learn multimodal representations of its environment | Stereo RGB images, stereo audio, inertial measurement unit (accelerometer, gyroscope, magnetometer), odometry, battery current, motor velocities, infrared and contact sensors, atmospheric pressure, air temperature | IEEE Dataport 2018, updated 2022, Linked paper: Optimality of inference in hierarchical coding for distributed object-based representations [75] |
| Hard dataset and normal dataset for robotic tactile sensing [76] | Tactile sensors for detecting grasping stability and slip detection during lifting of objects | PVT points of robot arms and grasping parameters/timestamp of gripper, tactile files with timestamps, images of each grasping | IEEE Dataport 2022, Linked paper: A Robotic Grasping State Perception Framework with Multi-Phase Tactile Information and Ensemble Learning [77] |
| Tactile Information Dataset TacAct: Physical Human-Robot Interactions [78] | Tactile intelligence for human-contact safety to complement visual and auditory information for human-robot interaction | Real-time pressure distribution of 50 subjects who performed a total of 24,000 touch actions | Data available on Zenodo, doi: 10.5281/zenodo.5138841 [78]https://zenodo.org/records/5138841 |

### 6.2.3 Standardisation and related industry initiatives

Relevant Standardisation activities for multimodal remote operation are further discussed in the following:

#### IEEE

Tactile internet P19181.1 standard [1] was the first standardization activity for haptic communication started in 2017. IEEE P1918.1 has been further adopted in 3GPP 5GS Specs [27] in TACMM [15], XRM features [16], Cyber-CAV [17] and Metaverse [18] as well as in IEC TC100 PT63448 [19], standardizing ultra-low latency communication for control-centric applications, addressing the use case

requirements and enabling technologies. IEEE P1955 [20] is a recent standard for 6G empowering Robotics: Use Case Scenarios, Requirements, Architectural Impact, and Technical Assumptions. This standard defines a framework for 6G empowering robotics applications. The framework incorporates potential robotics use case requirements, 6G enabling technologies, and defines the architectural impact necessary to enhance robotics and automation processes.

## ETSI

The European Telecommunication Standard Institution (ETSI)'s Industry Specification Group (ISG) on ISAC [21] concentrates on pre-standards research efforts on ISAC with a focus on use cases, channel models, key performance indicators, and evaluation assumptions, for subsequent evaluation by standards organizations. This includes defining a prioritized set of 6G use cases and sensing types, developing advanced channel models for these use cases, specifying performance indicators and evaluation methods, studying system and radio access network architectures for 6G, addressing privacy and security aspects in the context of 6G sensing data, and studying the potential impact of widespread ISAC deployments.

## IETF

IETF Tactile Internet was also discussed in a number of groups within IETF, primarily as a use case that demands improved networking technologies that satisfy its stringent resource requirements namely, INTAREA working group [81] presents the service requirements for haptic and Tactile Internet use cases. Activities in Network Function Virtualization (NFV) Research Group (NFVRG) [22] suggests that a combination of radio access and core network components must be isolated into network slices for addressing specific requirements of emerging use cases, such as Tactile Internet services.

## 3GPP

From the outset, 3rd Generation Partnership Project (3GPP) 5G System (5GS) was designed to provide service-based, highly reliable (e.g., URLLC, Time Sensitive Communications (TSC), Edge Computing) low latency communications and enablers for Industrial Automation, e.g., Network Analytics and Network Slicing. Although great progress was achieved, many existing and new Use Cases still remain to be addressed, e.g., UCs with stringent requirements, as those needed to support tactile and multimodal communication services over the 5G system. To address these challenges, 3GPP TR 22.847 [15] (multimodality SA1 feature) and 3GPP TR 23.700 (XRM SA2 feature) [16] intend to create a gap analysis between new potential requirements and existing requirements and functionalities supported by 3GPP. Especially, for use cases that are immersive real-time experiences, including closed-loop feedback and control under varying DoFs. Requirements for use cases under consideration include (but are not limited to), parallel transmission of multiple modality representations associated with the same application. Also, their reliability, availability, security, privacy, charging, and the identification of KPIs for specific use cases are considered. 3GPP multimodality feature provides an example of new requirements motivating 3GPP 5GS enhancements to meet the needs of demanding applications as those seen in the healthcare industry. In residential/small environments, this gateway is referred to as evolved Residential Gateway (eRG) (3GPP TR22.858) and they are considered from a 3GPP System perspective as User Equipment (UE). Nevertheless, they also provide connectivity and Quality of Service (QoS) handling to other devices connected behind their realm, e.g., Personal IoT networks (PIN) (3GPP TR 22.859), which might correspond to sensors or actuators remotely controlled by e.g., a factory worker or a physician. It is thus quite possible that in an end-to-end scenario where a surgeon wearing a tactile glove, which is a wireless device connected through an eRG, may connect to actuators in a remote location, far away from the surgeon. In that case, the eRG at each end of the connection will need to satisfy reliability and latency constraints, using technologies that may further extend the existing 5GSs. To support the XR KPIs requirements, prediction of, or fast adaptation to, RF conditions changes is critical. Key vertical UCs (especially for factory and process automation) are studied in

3GPP System Aspects (SA) WG1 (SA1) (3GPP22.104) and SA6 (3GPP TR 23.745) [81]. XR-based services are an essential part of "Metaverse" services to provide immersive experience accessed either by users in the proximity or remotely. The 3GPP Localized Mobile Metaverse Services in Study Item (Rel19) [18] supports the XR KPI requirements, prediction and adaptation where coordinating input perception/sensing data from different user devices (such as sensors and cameras) and coordinating output data to different devices at different destinations to support the same application is required.

Besides multimodal communication based on device-based sensing, a key role in wireless communication standardization, is actively involved in advancing the standardization for ISAC by defining the service requirements and system architecture aspects [82]. The early study items for ISAC are focusing on both communication-assisted-sensing and sensing-assisted-communication, while currently Rel. 19 puts a great emphasis on communication-assisted-sensing. There are 32 industry use cases for various industries identified for functionality requirements. In parallel, discussions of ISAC for 6G have also been started in various industry forums, including the 3GPP SA1 study, which has identified use cases with requirements for Vehicular-to-X (V2X), unmanned aerial vehicles (UAVs), 3D map reconstruction, smart city, smart home, factories, healthcare, and applications for the maritime sector [83].

### ITU-T

The Focus Group on Technologies for Network 2030, which concluded its work in Summer 2020, studied the capabilities of networks for the year 2030 and beyond. The group selected TI as a representative use case for network 2030, among other use cases such as holographic type communications and Space terrestrial integrated network. Recent Metaverse focus group (MV-FG) was established under TSAG in December 2022. The group aims to analyse the technical requirements of the metaverse to identify fundamental enabling technologies in areas from multimedia and network optimization to digital currencies, Internet of Things, digital twins, and environmental sustainability [96].

## 6.3 Multimodal Sensing, computing, communications and control

Multimodal sensing, computing, communications for remote control and operation can be achieved based on a fully integrated and unified, next-generation architecture serving the edge-to-cloud compute continuum, integrated sensing and communication and communication and control co-design as three key architecture technology aiming at serving the most demanding applications and use cases, aligned with the 6G standardization discussed in subsection 6.2.3.

Firstly, the integration of high-performance computing devices into cellular networks with the performance expected for 6G, gives rise to the development of multimodal computing concept. This term that englobes sensing, control, communications, and computing, can be applied in most of the industry 4.0 and 5.0 scenarios. For example, in industry use cases, the requirements vary in terms of latency and data throughput, although high security and reliability are a must. The extraction of data if processed in distanced servers (i.e., the cloud) can significantly increase the latency but also pose risks to the security and data privacy. To this end, the notion of edge computing and edge devices are introduced to address such concerns, and to provide computing capabilities within users' locality. Secondly, to deliver an immersive remote-control experience, the remote operator requires the perception of the shared operation environment (i.e., a digital model) for decision-making process and control. This can be achieved by processing and interchanging the data generated from external (ambient) and internal (on-board) sensors or from RF sensing technique such as networked sensing. Finally, the communication–control co-design methods can also optimize remote control operation by including both communication and control constraints in the design.

The high-level illustration of the concept and its key enabler technologies is shown in Figure 17.

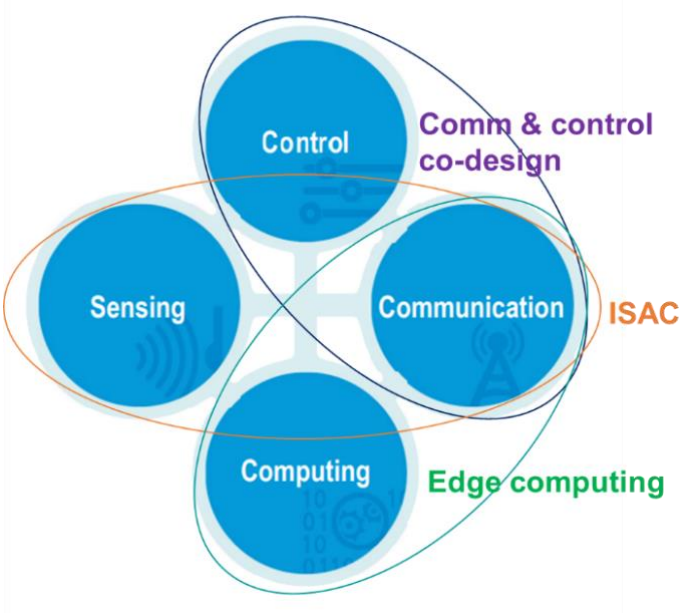



*Figure 17: A high-level representation of the 6G logical building blocks for multimodal remote operation*

### 6.3.1 Edge computing

With the fast growth of multimodal edge devices and network infrastructure, 6G mobile communications enable resource-intensive applications by delivering quick, efficient, and intelligent hyper-connectivity. Edge computing has become a promising solution towards efficient and low-latency multimodal communications [97][98]. Currently, there are two approaches about the location to integrate edge devices in the mobile networks. In case of use cases with a limited area and thus, a limited number of radio infrastructure, edge devices can be connected to the Core Network via User Plane to transfer data and Control Plane (Network Exposure Function) to have access to Network information, as specified by ETSI [98]. However, this approach could increase data latency in large deployed networks. In this case, edge devices could be integrated directly in the access network, reducing the distance to the multimodal user equipment, but then requiring mobility support since processes must move from one or a set of edge devices to another (set) to maintain the distance with the end devices. More specifically, edge computing reduces the requirement for data transfer to centrally located servers or the cloud, resulting in improved real-time processing capabilities. Especially in the case of multimodal communications, edge computing contributes towards:

- Reduced latency: Multimodal communications frequently require real-time processing of heterogeneous data across several modalities. Latency is considerably decreased by transferring computations near the edge, closer to the edge device level and data generation. This is especially critical in time-sensitive applications and services (e.g., autonomous driving, remote operation) that require real-time decision-making.

- Reliability: Edge computing further enhances resiliency of multimodal communications by avoiding a central point of failure, thereby improving the reliability of the system.

- Reduced communication overhead: Multimodal communications often involve frequent transmissions of large volume of data across different network domains. Transmitting vast amount of data to centrally located network entities or cloud infrastructures could potentially lead to significant communication and signalling overhead, causing network delays. However, edge computing enables processing and analysis of multimodal data

closer to data generation, contributing towards communication overhead reduction by decreasing data exchanges with the core network.

- Privacy preservation: Although centralized computing for multimodal communications could provide accurate analytics, it may also raise some privacy concerns. This relies on the fact that highly sensitive or private information (e.g., personal images, audio, or biometric information) may pose restrictions in data exchange. Edge computing performs processing and analysis of multimodal information locally, without requiring any raw data exchanges to third parties. This approach enhances data privacy, since sensitive information is kept in a trustworthy environment.

Consequently, edge computing architectures have been adopted and proposed for enabling low-latency multimodal communications, including a high level three-layer hierarchical scheme comprised of a sensing, an edge computing and a cloud computing layer.

- Sensing Layer: The sensing layer's primary function is to gather raw multimodal data from diverse sensory devices, conduct lightweight tasks such as filtering, and communicate important inputs to the hierarchy's subsequent tiers.

- Edge Computing Layer: The edge computing layer receives as input data the output of the underline layer and performs demanding tasks such as data pre-processing, feature extraction, data fusion, lightweight ML model training, and inference tasks. In critical applications, it is essential to analyse and act real-time on the provided multimodal data towards quick decision making.

- Cloud Computing Layer: The cloud computing layer could further improve edge multimodal communications. The cloud computing layer may be utilized for long-term storage, centralized analytics, ML model training, and edge device orchestration.

Section 6.5.2 further discusses distributed Federated AI at the edge for low-latency and reliable multimodal communications.

### 6.3.2 Multimodal communications and control co-design

Cloud, communication and control systems are generally designed independently, which leads to substantial wireless resource consumption. In communication–control co-design methods (aka Integrated communication and control) [86], the strong correlations between control and communication systems are in two folds: 1) control optimization problem with communication constraints and 2) communication optimization problem with control constraints. To apply the communication–control co-design, the essence of the co-design problem should be explored and then build an effective co-design model to help with communication and control optimization. For instance, for AMR UC (Section 6.1.5), a communication-control co-design can help lowering the requirements on the coding rate and lower the consumption of wireless resources as well as achieving better results in the probability of system instability and the number of admissible AMRs.

Additionally, communications and control co-design introduce various KPIs beyond the conventional ones used in communication networks. For example, age of information (AoI), could be used for measuring information freshness in remote monitoring and control scenarios [87]. The age of incorrect information (AoII) [88], urgency of information (UoI) [89] as well as value of information (VoI) [90] are other KPIs that have been considered to quantify the application layer performance, as an alternative to latency, throughput or jitter.

Each of these metrics captures a different aspect of control performance and offer great potential for future industrial networks, adding a new dimension to control and communications co-design. Despite being well-accepted and employed by the research community, the adoption of such novel KPIs in today's mobile networks has not been considered and is an open question for future standardization efforts.

### 6.3.3 Multimodal sensing and communication

Multimodal sensing can be implemented with multiple sensors in the same device or through the fusion of data from several distributed sensors. In both cases, communication is key to convey the information to where it will be processed, enabling the offload of this processing to powerful devices, as well as data fusion when needed. At the same time, sensors that estimate the position of objects often rely on wireless transmission, used for localization or radar techniques. Synergistically, sensing can help communications as well. For example, beam training can be aided by positioning information and camera images [85], and the same applies to multiple access [91]. The fact that sensing and communications must eventually make use of the same (scarce) radio-spectrum has motivated many recent approaches to combine sensing and communications, using either the same frequencies by multiplexing, the same radio interface (same signal format and radio resources) by a joint design, and/or the same hardware.

The so-called integrated sensing and communication and sensing (ISAC) are seen as key ingredients in the evolution towards 6G [60]. These are thoroughly described in Section 4. Simultaneously performing multimodal sensing and communications motivates new requirements in the waveform design, which needs to be more energy-efficient to be able to cover large sensing ranges [92]. Several works address the multiplexing of signals dedicated to sensing and communications [92][94][95], while it is not clear yet how an optimum waveform from the ISAC point of view would look like. Also, new ways of scheduling the radio resources dedicated to ISAC need to be devised. Multiplexing of multimodal information (audio, video, haptic, tactile) for efficient transmission over B5G/6G air-interface is an important challenge which has not been addressed in 5G.

The fusion of data gathered from multiple sensors that is shared and combined through wireless communications is a powerful tool to enhance the environment awareness, achieving cooperative perception [93]. Local information processed locally in any device can be combined with that obtained at the edge nodes, reducing the needs for local sensing and computing. To maximise the efficiency, the multi-sensor information fusion, the information sharing strategy and the communications need to be designed in a coherent way.

6G Networked sensing can also complement device-based sensing to increase the sensing coverage and reliability of the final perception. This is as part of the environmental sensor fusion or as data stream by taking advantage of the dense deployment of 6G nodes. For instance, in remote robotic operations, active multimodal sensing and data collection by robot internal sensors or ambient sensors can be used for 6G network services optimisation (such as coverage, rate). Networked sensing enabled by RAN and device sensing are new key technologies and enhance environmental sensor fusions gained from device-based sensing. It is obvious that 6G networked and device sensing are complementing state of the art sensors. It is clear that the dense deployment of 6G nodes increases the sensing coverage and reliability of the final perception.

Furthermore, the ISAC co-existence of a diverse range of sensing modalities, including vision, audio, motion, environmental, biomedical sensing, etc., will provide information for applications such as augmented reality, immersive experiences, smart surveillance, healthcare monitoring, and environmental monitoring [103]. ISAC based use cases and enhanced 6G capabilities will provide a paradigm shift to the 5G infrastructure and mobile services [104], such as deploying energy-efficient and spectrally-efficient solutions [105]- [109]. There are several user-assisted enhancements that ISAC signalling will revolutionize such as edge computing, where adding a nominal computational burden at the user's end, several benefits could be achieved, including a) bringing computation and data storage closer to the network edge, b) reducing end-user latency, c) possessing real-time processing capabilities, and d) enabling context-aware services. Accordingly, ISAC based networks can support a diverse range of applications. Another enhancement can be achieved in terms of environment monitoring, since by integrating sensors, devices, etc., into the edge-computing unit, wireless networks can provide real-time environmental insights to facilitate diverse applications, e.g., smart city initiatives, environmental sustainability, proactive decision-making, etc. Accordingly, this enables the networks to collect and analyse vast amounts of data for various modern applications.

Within the indoor factory setting and of particular relevance to ISAC for Industrial IoT (IIoT), the following use cases have been considered: detection of Automatic Guided Vehicles (AGV) and measurement of proximity to humans, collision avoidance for Autonomous Mobile Robots (AMR) and AGVs, the combined use of sensing and localization for accurate positioning of AGVs and AMRs, gesture detection and recognition, etc. Additionally, THz frequencies can be used for multimodal sensing and communication, for applications such as localization and tracking; simultaneous imaging, mapping, and localization, such as in virtual urban city and indoor factories, augmented human sense applications such as in remote surgery, detection of product defects, sink leakage detection, etc., where very high range and cross-range resolutions are required; and gesture recognition, emotion recognition, heartbeat detection, fall detection, respiration detection, sneeze sensing, intrusion detection, etc., can be implemented in a smart hospital in the foreseeable future. For instance, a robot arm equipped with ISAC-THz module is used to represent a human arm holding a THz imaging camera. The prototype is built to operate at 140 GHz carrier frequency with a bandwidth of 8 GHz [110]. In terms of multimodal sensing and communication, one should note several challenges that exist in ISAC based deployment, such as channel modelling for the co-design, maintaining network reliability and security, cost effectiveness, efficient dual-function waveform design, regulatory considerations, and large data volume handling and processing. In this regard, several solutions such as efficient dual-function ISAC waveform design [111], [112], and channel estimation approaches for sensing and communication co-design [113].

## 6.4 TACMM device capabilities and functionality impact

On top of audio-visual information, next generation immersive communications will involve TACMM information. Immersive realities will blur the boundary between the physical and the virtual worlds, inspiring new types of interactions between them. Such disruption will enrich communication experiences of users and bring forward new use cases such as 3D telepresence, online interactive sports with unprecedented realism, and game changing immersive learning, and many others. The same paradigm could be applicable to human-machine interaction and collaboration and robotic tele-operation in, e.g., industrial environments, and to set a cornerstone for the next industrial revolution. To this end, significant progress has been made in recent years in new terminals equipment and sensor systems and data capturing techniques, data processing and computing frameworks, and rendering and display devices. TACMM is expected to coexist with XR and holographic communications delivering technologies that will enable users to have lifelike experiences of the physical world, in the virtual world, or a combination of both, with interactions using 3D audio-visual and haptic information exchange. In concrete, multi-sensory XR will leverage human senses and perception (e.g., visual, auditory, olfactory, and tactile) into XR content, to provide a highly realistic immersive experience. This will be achieved by combining multiple disciplines, of ultra-high performance and low latency communications, computing and networking, AI-ML, computer vision, biology, robotics, etc., providing a fusion of real and virtual worlds.

### 6.4.1 Multimodal teleoperation terminal equipment and capabilities

Terminal Equipment (TE) provides the functions necessary for the operation of the access protocols by the user, a functional group on the user side of a user-network interface [58]. The TE's 'upper layers' may be in more than one device, e.g., VR glasses, gloves, etc. New TEs with sensory/actuator integration result in new user equipment (UEs) capabilities and functionalities required by future multimodal applications, further increases the demand on processing power required to execute the computation-intensive and low-latency tasks.

Simple, intuitive, multimodal, and easy to use interface to interact with devices, services, and surroundings, are needed. While current handheld smartphones are mainly used to provide audio/video services, immersive remote operation would require additional device capabilities such as smart glasses, haptic wearables, etc., which can be distributed to other connected devices for providing enhanced user experience. The use of such dedicated devices is driven by the

technologies like telepresence, haptic user interactions, holographic displays and content distribution, multi-dimensional imaging, and extended reality, and metaverse applications. New UE types could be classified and used by health, entertainment or industry verticals. Some examples are as following:

- Connected nano-things and bio-nano-things, which could interact with biological systems and act as sensors and actuators to interact with the environment. That could be part of development for multimodal user interaction and human-machine interactions;
- New sensory/actuator integration results in new user equipment capabilities and functionalities required by future holographic XR/multisensory XR applications, further increases the demand on processing power required to execute the computation-intensive and low-latency tasks;
- Implantable miniature sensors and 'nanosensors' devices, such as implants, stents, lenses and pumps are mainly used for medical purpose, e.g., a battery-operated pacemaker. Also, Biomolecular Sensing devices (DNA microarrays, Chemical sensors) that can potentially be used for people who may have devices on or beneath their skin that monitor heart rate, glucose, or oxygen saturation and help control chronic conditions like diabetes or respiratory disease. The monitoring could enable people to live at home instead of having to move into an assisted living facility;
- Wearable: Sensors that are embedded in some type of garments/textile to monitor. Example sensors are electrocardiogram (ECG), Electroencephalogram (EEG), Electromyography (EMG), blood variation pressure (BVP), galvanic skin resistance (GSR), Photoplethysmography (PPG), accelerometers, and glucose sensors. Use case examples can be paramedics and firefighters may eventually be required to use wearables that track their heart rates, emotion and stress levels.
- Tattooable: Ultra-thin electric mesh for human skin, or temporary skin that can store data and deliver drugs—and electronic second skins made of microscopic semiconductors;
- Ambient Sensors: Sensors monitoring user's activities (e.g., passive infrared sensor (PIR), RF imaging, surveillance cameras - fixed or on drones);
- Actuators: Haptic devices to provide feedback for navigation or alarm (haptic buzz), display interface (braille language), interaction (haptic gloves/belt/suits), and remote industrial control;
- Robot hand and arm: A robot hand, especially one with multiple fingers, is necessary for conducting various tasks in daily life. The robot hand should be able to make flexible human-like motions using a design inspired by the human and integrated with robot arms to construct systems with a high number of degrees of freedom (DoF). Skin or touch sensing is desirable, which sets requirements for haptic signal amplitude range, temporal and spatial resolution. To match the human hand for remote operation, a robot hand should be equipped with multimodal tactile sensors with different types of tactile sensing modalities (thermal, fast adapting and slow adapting afferents). The development of autonomous dexterous robotic manipulation systems is a complex process of an interdisciplinary nature involving such diverse research fields as computer vision, force control, motion planning, grasping, sensor fusion, digital signal processing, human-robot interaction, learning and tactile and multimodal sensing.

### 6.4.2 TACMM functionalities

TACMM communication combined with XR will pose new requirements in forthcoming 6G architectures. In general, XR has stringent latency requirements for accurate and smooth content

playback based on user motions. Furthermore, in order to achieve low content delivery latency, an ultra-high data transmission rate is required for delivering XR content. The computing capability of both network servers and user devices dominates the performance of interactive XR applications, and limited computing capability in the network can be another bottleneck for XR content delivery.

Multilateral TACMM communication procedures will take as a working assumption that even if there is no direct TACMM interaction between two users, they can still share haptic information. For example, the data format and content of the transmitted TACMM information may differ from those of the transmitted haptic information in direct haptic interactions. The end-to-end (e2e) delay tolerance of TACMM communications can be as low as 1 ms while the packet rate of TACMM data could reach more than 1,000 packets per second. In practice, the e2e delay requirement of TACMM communication is defined by perceptual sensitivity of receivers, the changing aspects of TACMM interaction, and the concrete operation or interaction among any others factors. In addition, the reliability of TACMM communication in immersive gaming is required to be above 99.9%, whereas higher reliability is required (i.e., above 99.999%) in telesurgery and remote machine manipulation.

6G architecture must embrace a number of technologies to tackle the challenges of TACMM communications. An indicative sample of them is mentioned in the following (i.e., a full and in-depth analysis does not fall within the scope of this White Paper):

- The nominal communication delay of 1 ms for TACMM, is translated to a physical-layer delay of less than 0.1 ms.
  - Queuing delay in the downlink transmissions could be reduced, if TACMM data preempt the data of other types. Non-orthogonal multiple access (NOMA) can improve spectrum efficiency and reduce channel access delay of TACMM in uplink transmissions.
  - Grant-free user scheduling has been exploited to remove the scheduling delay by saving transmission resources, which could be assigned to TACMM-based devices.
  - In packet retransmissions, interference in multiple access should be dealt with. In grant-free NOMA, such interference can be managed by device activity detection and successive interference cancellation (SIC).
- In addition to low latency, guaranteed latency is an important need of teleoperation and can allow for predictive algorithms to operate reliably in, e.g., teleoperation.
  - Guaranteed latency is often harder to achieve in mobile communications, and require deterministic approach to various processes that are stochastics in nature.
- The communication reliability of TACMM could be improved by adopting several solutions offered in the literature.
  - Low-density parity-check (LDPC) codes and short polar codes could fit with the requirements of TACMM communications.
  - Massive multiple-input and multiple-output (MIMO), IRS, and multi-connectivity techniques.
  - NOMA can improve retransmission efficiency where the transmit power of a device can be optimised to retransmit the required minimum redundant bits for satisfying the reliability requirement.
- Low delay and high reliability of TACMM communications, could be enforced by network slicing and in concrete via:
  - Resource reservation in the network slice for TACMM communications.
  - When various tele-operation slices are considered and when delay should be considered, customised transmission resource reservation should be applied.
  - AI-based learning methods should be leveraged for the traffic patterns of TACMM-based scenarios to efficiently assign resources.

In short, a battery of advanced signal processing techniques, distributed edge computing and AI-driven innovations should enhance the proposed 6G architectures in order to efficiently integrate TACMM communications, tackle their challenges and satisfy their requirements.

Furthermore, the distributed user equipment requires additional functionalities and interfaces to be exploited in 6G architecture, e.g., for multimodal synchronisation to provide highly reliable immersive and inter-related multimodal communications in 6G systems. This is due to a crucial requirement in scenarios where the application experience is consumed over multiple devices. Multimodal data belonging to the same application/session must be delivered, and ultimately presented to the user at the same time addressing different synchronisation threshold tolerance. A high-level illustration of the multimodal data flow over 6G communication system for an operator using multiple devices (i.e., VR and haptic gloves) to remotely operate a robotic hand in an industrial domain is shown in Figure 18.

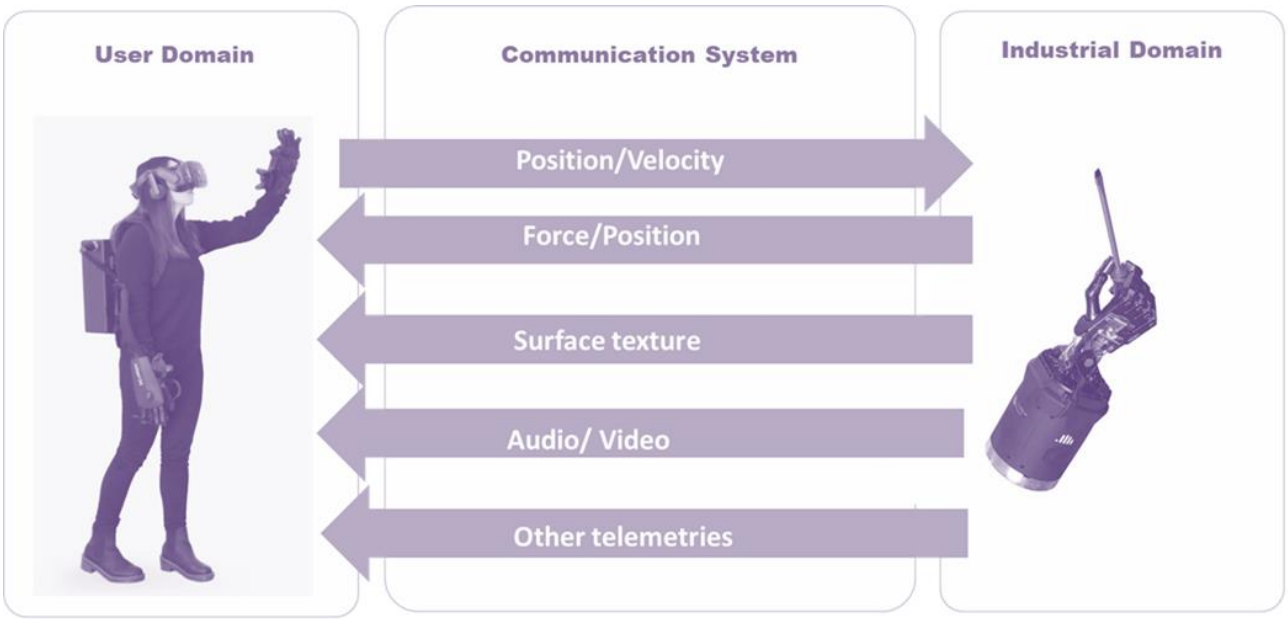



*Figure 18: High level illustration of an E2E Multimodal remote operation example using a robotic hand*

## 6.5 AI/ML techniques for multimodal operation

Audio and visual information have been traditionally captured, processed and communicated across networks for remote consumption. With tactile Internet, communicating tactile information is driving current trends in connected robotics [36]. Including smell and taste information in future multisensory communication systems will round up remote perception of all five human senses. Furthermore, although humans are restricted to visual sensing in visible light band, the machines are not. They can make sense and learn the environment by sensing across different bands [44]. Establishing environmental features (e.g., location and movement of objects) can be done across different bands which drives current interest in integrated communication and sensing.

In the above scenarios, machines will collect and learn from multimodal data in order to execute one or more tasks. In order to do so in an efficient manner, cross-modal perception can be applied for loss recovery in multimodal applications which demands design of combined classifiers [39] for the purpose of Data recovery while maintaining precision and resource efficiency trade-off. Furthermore, they will have to exploit fundamental principles of multimodal and multitask learning. In particular, learning and decision making from multimodal data usually involves learning suitable multimodal representations from which different decision-making tasks can be executed with high efficiency. In this section, we review learning methods that are likely to have significant impact on the design and implementation of multimodal sensing and communication systems.

### 6.5.1 Multimodal learning techniques

The gains of integrating AI in domains requiring analysis of high-dimensional unstructured data, as in the case of multimodal communications, are evident. In order for the AI to provide an interpretation and understanding of such complex systems, correlation and extraction of latent information from multimodal data is required. The underlying motivation to use correlated multimodal data is that complementary information could be extracted from each of the modalities considered for a given learning task, yielding a richer representation that could be used to produce much improved performance compared to using only a single modality [37]. Towards this direction, multimodal learning aims at creating models that can process and relate information from multiple modalities. Such information consists of data from various sensors monitoring common phenomena, towards being exploited in a complimentary manner in order to identify patterns and interpret a complex task. The following five challenges in multimodal learning have been identified [38]:

- Representation: The heterogeneity of multimodal sensory data poses significant challenges in terms of learning latent representations.
- Translation: The relationship (mapping) between different modalities towards interpret the multimodal information is another core challenge, e.g., different ways to translate an image or an audio.
- Alignment: To align different modalities, a model has to measure similarities between them and has to deal with long-range dependencies. Other challenges in multimodal alignment include the absence of annotated datasets, the development of acceptable similarity metrics between modalities, and the presence of numerous accurate alignments.
- Fusion: Another challenge is to provide predictions, combining data from two or more modalities. This relies on the fact that different modalities generalise at different rates and as a result a joint training strategy isn't always optimal.
- Co-learning: Finally, another challenge is to transfer knowledge between modalities, their representation, and their predictive models, especially in the case of modalities with limited resources (e.g., annotated data).

As previously mentioned, different modalities generalise at different rates and as such a multimodal fusion approach is required, involving better combination decisions of multimodal data, either by removing correlations between modalities or representing the fused data in a lower-dimensional common subspace. The fused data can then be exploited towards anticipatory decision making. Some well-known techniques for multimodal fusion are listed below:

- Early Fusion: A method of modality combination at the input level before fed to an ML model [40].
- Late Fusion: Late fusion involves an independent training of models for each modality and the integration of their outputs (predictions) throughout the decision-making process.
- Intermediate Fusion: Combining modalities at intermediate level of model training.

Currently, deep learning, a subfield of representational learning, focuses on the use of artificial neural networks (ANN) to automatically generate suitable representations or hidden features from raw data for specific tasks, resulting in a richer representation [41] that can contribute towards better decision making. As a result, it is critical to develop joint embeddings in order to better describe such notions by using the complementarity of multimodal data. A list of well-studied models for multimodal representation learning include:

- Multimodal Deep Autoencoders (MDAs): MDAs are neural network-based models that learn sequential and joint representations of different modalities and use a decoder to learn to

reconstruct the initial input [42][43]. Indicative examples of such encoders could be considered a Long-Short Term Memory encoder, a Transformer encoder, a Convolutional encoder etc.

- Graph-based Neural Networks (GNNs): GNNs are neural network-based models that can be directly applied and capture relationships and interactions between multimodal graph structures [44].
- Generative Adversarial Networks (GANs): GANs are known for their ability to generate realistic artificial data in the original space. As an extension of their main functionality, GANs can be used as means that allow cross-modal translation from one initial space (e.g., text) to another initial space (e.g., image), enabling the capture of joint distributions of different modalities [45].
- Multimodal Zero-Shot Learning (MZSL): MZSL comprehensively integrates the benefits of both multimodal learning and ZSL, resulting in models with greater generalization ability. The multimodal ZSL aims to distinguish unseen objects in the principal modality with the assistance of a secondary modality in which the entities are seen [46].

The outcome of the above described approaches could be further exploited by AI-driven approaches such as forecasting ML models towards early event detection for anticipatory decision making. The provided forecasts can provide insights into prospective outcomes and enable proactive adaptations.

### 6.5.2 Federated Learning Techniques

Multimodal learning can be efficiently applied in complex sensing tasks such as person tracking, activity detection, audio and video analysis. However, integrating multimodal learning in a wireless network environment, where data from heterogeneous data sources or modalities is distributed across devices and locations, poses new challenges. Data sizes for distinct modalities (e.g., video vs. audio) can be substantially asymmetric, resulting in considerable network delays or decreased performance accuracy [47]. Therefore, distributed federated learning (FL) paradigm has emerged as alternative solution, enabling collaborative training of models, while reducing network delays, preserving data privacy and decreasing data exchanges.

Federated learning has improved the security of multimodal learning by enabling a distributed and privacy-enhancing approach. In a multimodal system, devices can provide data for FL either through a single modality or multiple modalities, such as cameras and sensors in the same scene. The FL model is trained locally by fusing data collected by different sensors extract common or correlated representations between multiple modalities, without requiring any data exchange. A server with the role of coordinator combines encrypted FL model parameters to update a global FL model that will be re-distributed among the involved devices. The use of FL accelerates the learning process, while saving communication costs. FL in multimodal systems is classified as horizontal, vertical, and transfer learning based on how the data is spread in the feature and sample space [48], as described in the following:

- Horizontal FL (HFL): HFL considers participants that share the same feature space (same multimodal data types) but may differ in the sample space.
- Vertical FL (VFL): In contrast to HFL, VFL is applicable to the scenarios where different participants have the same number of multimodal data samples but differ in feature space.
- Transfer FL (TFL): TFL is a supplement to horizontal and vertical settings, and is suitable for scenarios where the sample space and feature space of each participant have less overlap.

### 6.5.3 Multitask Learning Techniques

In multimodal setting, we are usually interested in exploiting different modalities for different learning tasks. This is traditionally done separately and independent models are designed to perform different tasks across different input modalities. However, by focusing on a single task in a presence of parallel learning process on related tasks, we might ignore useful information and representations obtained from these related tasks. Multitask learning (MTL) exploits such relationships and advocates for sharing the representations between the related tasks, enabling better generalization capabilities of the resulting models for each of the tasks [47]. In general, any scenario in which one is required to optimise across multiple loss functions leads to multitask learning. Even in single-task scenarios, exploiting the domain-specific information of related tasks may improve performance and reduce training complexity.

Multitask learning is simply integrated into deep neural networks framework using two main approaches. In hard parameter sharing, most of the DNN architecture is shared across the tasks, except for several final task-specific output layers. In soft parameter sharing, different DNN architectures are used for different tasks under certain regularisation constraints that ensures parameter similarity. Many intuitive arguments on why multitask learning is useful and improves single-learning approach includes implicit increase of sample size, extracting more relevant or reusing (eavesdropping on) similar features, favourising joint representations, and by introducing regularisation effects.

Multitask learning has a recent history of successful applications across the mainstream machine learning fields such as Natural Language Processing [48] or Computer Vision [49]. A number of recent studies investigated multitask learning in the domain of wireless communications, e.g., in the context of connected autonomous vehicles [50], network management and orchestration [51], IoT network traffic prediction [52]. Multitask learning naturally fits multimodal sensing and communications, leading to exploration of multitask multimodal learning for training multimodal navigation agents [53] and remote robotic grasping [54]. Multitask learning is also suitable for an evolving concept of semantic or task-oriented communications. Multitask deep learning models may be developed as a core part of a semantic communication systems enabling serving different tasks based on multiple modalities [55].

### 6.5.4 Summary of ML multimodal techniques analysis

Table 10 provides a comprehensive overview of the latest multimodal machine learning techniques applied to both private and publicly available datasets. It encapsulates the state-of-the-art models referenced above, detailing the analysis conducted, the solutions developed and the advancements achieved in multimodal and tactile technologies.

Multimodal Learning (Section 6.5.1) encompasses several advanced AI techniques, notably the use of Multimodal Deep Autoencoders (MDAs) and Generative Adversarial Networks (GANs). MDAs, demonstrated through tasks such as digit classification and network simulation, emphasise the denoising process to achieve high accuracy in data reconstruction and clustering. GANs, especially in video hyperlinking applications, leverage the power of Conditional GANs (CGANs) to create high-quality embeddings that improve cross-modal mappings. These models not only perform effectively in their respective domains but also provide a framework for enhancing data representation across multiple modalities. Future study is required to expand these techniques to more complex scenarios and improve their generalisation capabilities by exploring different architectures and processes.

Federated Learning (Section 6.5.2) is explored through Multimodal Federated Learning (MMFL), which distributes the learning process across multiple clients, each handling different data modalities. This approach is particularly beneficial in scenarios where data privacy and localisation are crucial, as it allows the development of deep learning models without centralised data sharing. By employing techniques such as hierarchical gradient blending and optimal aggregation of local updates, MMFL can achieve significant performance improvements. However, challenges remain

in handling computational costs for clients with incomplete data and ensuring the real-time applicability of these models. Future research is directed towards unsupervised learning and data augmentation to mitigate these issues.

Multitask Learning (Section 6.5.3) integrates tasks such as speech recognition, event detection, and network traffic prediction within a unified framework. This approach benefits from sharing deep network layers across related tasks, which improves the generalisation performance and efficiency of the models. Examples include the use of CNNs and LSTMs for audio-visual event detection and leveraging deep architectures like LSTMs for tracking IIoT backbone network traffic. MTL models show promising results in reducing latency and maintaining high accuracy. Further research is required for extending these models to handle more modalities and improving their scalability and robustness in diverse application scenarios.

*Table 10: Dataset, models, analysis, and solutions for AI/ML multimodal techniques*

| Model | Dataset | Analysis | Solution | Results |
|---|---|---|---|---|
| **AI/ML Technique: Multimodal Learning** | | | | |
| Multimodal Deep Autoencoders (MDAs) | Variations of MNIST digit classification as discussed on Larochelle et al., 2007 | Denoising autoencoders map corrupted examples to uncorrupted ones, and can be seen as a way to define and learn a manifold. | Stacked denoising autoencoders (SdA-3) | SdA-3 algorithm performs on par or better than the best other algorithms |
| Multimodal Deep Autoencoders (MDAs) | Data generated using the discrete-event network simulator (NS3) | Enhancing performance using clustering techniques like k-Means and HDBSCAN, and ensemble techniques like Weighted Voting model enriched by base models like Naïve Bayes, MLP and SGD. | Ensemble Weighted Voting Model, combines the predictions from 4 supervised ML models: DT, RF, k-NN, and SVM | High accuracy 96% for Device, 92.72% for Service and 93.68% for Network |
| Generative Adversarial Networks (GANs) | BBC video hyperlinking dataset (MediaEval 2014) | Evaluate classical and state-of-the-art multimodal AEs and BiDNNs. Obtain multimodal embeddings with CGANs. Generator network visualizes crossmodal mappings, and obtains high-quality embeddings. | Multimodal mbeddings with conditional enerative adversarial networks (CGANs) | Improved multimodal embeddings over initial representations. |
| Multimodal Zero-Shot Learning (MZSL) | Multiple datasets (e.g., Animals with attributes, Caltech-UCSD-Birds, Oxford flowers) | ZSL distinguishes unseen objects in one modality with the aid of another. Multimodal learning can endow ZSL with high-quality multimodal representation and the derived MZSL model has enhanced generalization. | Numerus ZSL and MZSL models e.g. E-PGN, MFF, TF-VAEGAN, FREE, f-VAEGAN-D2, and other | Improved recognition for both seen and unseen classes but still challenging for high accuracy on all classes. |

| Model | Dataset | Analysis | Solution | Results |
|---|---|---|---|---|
| **AI/ML Technique: Federated Learning** | | | | |
| MMFL based on modal conditions, the distribution of modalities and the availability of modal annotations | Kineics-400, RealWorld2, IEMOCAP, ModelNet40, Vehicle sensor, mHealth dataset, UR fall detection dataset | Classification based on modality distribution and annotations. | Homogeneous Multimodality, Heterogeneous hybrid modality. Supervised MMFL, Semi-supervised and unsupervised MMFL | Effective gradient blending, and optimal local update aggregation. Potential of FL in multidomain, multitask deep learning without sharing data. |
| **AI/ML Technique: Multitask Learning** | | | | |
| End-to-End multimodal models | Lip Reading in the Wild (LRW) Dataset | Speculative inference on multimodal data streams optimize multimodel model's accuracy while minimising end-to-end latency. | The speech recognition model predicts the inference class | Slight accuracy drop (1%) but significant latency speedup (2–128x) |
| | Audio-Visual Event (AVE) dataset | | Event detection model used to classify audio-visual events | |
| | STISEN dataset | | Activity recognition with dedicated CNN to extract features that combine through fully-connected layers to infer output. | Significant latency speedup (7–24×) with minimal accuracy drop (1.4%) |
| A general deep NN architecture for NLP | PropBank dataset, Wikipedia common words, Stanford Named Entity Recognizer | Training NNs on related tasks improves features produced and thus improves generalization performance. Great performance for NLP tasks when sharing the lookup-tables of each task. | All tasks were trained using the NN. POS, NER, and chunking tasks were trained with the window version. SRL with 1 convolution layer, hidden units, and lookup tables | Embeddings obtained in the word lookup table were excellent. SRL improves by learning auxiliary tasks. |
| UniT: transformer with separate encoders for each input modality followed by a decoder | COCO Dataset, Visual Genome (VG) dataset, GLUE benchmark, VQAv2 dataset, SNLI-VE dataset | Jointly learns tasks from different domains, including object detection as a vision-only and language-only tasks. Jointly train the model on object detection and VQA tasks. Then, language understanding tasks and SNLI-VE as an additional joint vision-and-language reasoning task. | Training on object detection as a vision-only task and VQA as a multimodal task that requires jointly modeling the image and the text modalities | Strong performance on multiple tasks with a compact set of shared parameters. |

| Model | Dataset | Analysis | Solution | Results |
|---|---|---|---|---|
| LSTM and MTL-based deep architecture | Real dataset sampled via the Abilene and GÉANT backbone networks | The method predicts IIoT backbone network traffic by extracting the spatial and temporal features of TM. LSTM tracks the spatial features of network traffic features. Multitask regression to train an additional task, which improves generalisation. | PCA method, the sparsity regularized matrix factorization (SRMF) method. | Accurate tracking of IIoT network traffic with enhanced generalisation |
| The Dual-Attention model, Gated-Attention and Spatial-Attention | Chaplot et al. (2017) instructions dataset | Jointly learning semantic goal navigation and embodied question answering through fusion of textual and visual modalities. Adaptation of baselines using multimodal fusion techniques. | Two naive baselines, Image and Text only. Two baselines on prior semantic goal navigation, Concat and Gated-Attention. Two baselines on Question Answering models, FiLM and PACMAN. | Achieves superior accuracy for SGN and EQA, compared to baseline models. Spatial auxiliary tasks lead to better performance for all models. |

## 6.6 Conclusions on multimodal operation

Advancements in multimodal and tactile technology, including high levels of co-presence in real time, which demand ultra-reliable low latency platforms over distance to offer high quality interpersonal communication experiences remotely. Compared to 5G that offers haptic services to business customers (Non-Public 5G networks), 6G is expected to democratise tactile and multimodal communication to mass consumers enabling remote skillset/labour delivery, in a similar way the Internet democratise the knowledge sharing. This is revolutionizing many sectors, including, healthcare, education, manufacturing, mining, warehousing, and so forth.

In this section, we have highlighted the 6G technology capabilities that can enhance remote operation and control in several vertical use cases including industry, education, emergency response and health. Multimodal related academic and industry research, and in multiple SDOs discussed in this section, need to advance in parallel, initially with focus on 6G vertical requirements and architectural design, and later through harmonization of computing, communication and control and their interfaces and APIs.

Considering the existing technology enablers, we highlight a number of open areas of research that a) are required for the interoperability between the discussed standards in order to realise a communication platform within millisecond and sub-milliseconds latency depending on the use case requirements, b) highlight advancements in other technologies (e.g., AI/ML) that can support improvement in haptic control and multimodality scenarios, c) present terminal innovations that can evolve into immersive experiences and communications including bi-directional tactile and multimodality remote operation.

## 6.7 References

[1] O. Holland, E. Steincach, R. Venkatecha Parasad, Q. Liu, Z. Dawy, A. Aijaz, N. Pappas, K. Chandra, V. S. Rao, S. Oteafy, M. Eid, M. Luden, A. Bhardwaj, X. Liu, J. Sachs, and J. Araujo, “The IEEE 1918.1 “Tactile Internet” Standards Working Group”, Proceedings of The IEEE | Vol. 107, No. 2, 2019

[2] A. Aijaz, M. Dohler, A. H. Aghvami, V. Friderikos, and M. Frodigh, “Realizing the tactile internet: Haptic communications over next generation 5G cellular networks,” IEEE Wireless Communications, vol. 24, no. 2, 2017

[3] National Academies of Sciences, Engineering, and Medicine. 1995. Virtual Reality: Scientific and Technological Challenges. Washington, DC: The National Academies Press. [Online] https://doi.org/10.17226/4761

[4] Zh. Hou, C. She, Y. Li, D. Niyato, M. Dohler, and B. Vucetic, “Communications for Tactile Internet in 6G: Requirements, Technologies, and Challenges”, IEEE Communications Magazine, 2021

[5] A. Alkhateeb, G. Charan, T. Osman, A. Hredzak, J. Morais, U. Demirhan, and N. Srinivas, “DeepSense 6G: A Large-Scale Real-World Multi-Modal Sensing and Communication Datasets,” available on arXiv, 2022. [Online] https://www.DeepSense6G.net

[6] G. P. Fettweis and H. Boche, “The Personal Tactile Internet—And Open Questions for Information Theory”, IEEE BITS the Information Theory Magazine, Volume: 1, Issue: 1, 2021

[7] S. Lagén, B. Bojović, K. Koutlia, X. Zhang, P. Wang, and Q. Qu, "QoS Management for XR Traffic in 5G NR: A Multi-Layer System View and End-to-End Evaluation," in IEEE Communications Magazine, vol. 61, no. 12, pp. 192-198, December 2023, doi: 10.1109/MCOM.015.2200745

[8] Z. Hou, C. She, Y. Li, D. Niyato, M. Dohler and B. Vucetic, "Intelligent Communications for Tactile Internet in 6G: Requirements, Technologies, and Challenges," in IEEE Communications Magazine, vol. 59, no. 12, pp. 82-88, December 2021, doi: 10.1109/MCOM.006.2100227

[9] H. Viswanathan and P. E. Mogensen, "Communications in the 6G Era," in IEEE Access, vol. 8, pp. 57063-57074, 2020, doi: 10.1109/ACCESS.2020.2981745

[10] G. Mountaser, T. Mahmoodi, and O. Simeone, “Reliable and Low-Latency Fronthaul for Tactile Internet Applications”, IEEE Journal on Selected Areas in Communications, 2018

[11] J. Bolarinwa, A. Smith, A. Aijaz, A. Stanoev, Ma. Sooriyabandara, and M. Giuliani, “Haptic Teleoperation goes Wireless: Evaluation and Benchmarking of a High-Performance Low-Power Wireless Control Technology”, IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR) 2022, 2210.07212.pdf (arxiv.org)

[12] M. Dohler, T. Mahmoodi, M. A. Lema, M. Condoluci, F. Sardis, K. Antonakoglou, and H. Aghvami “Internet of Skills, where Robotics meets AI, 5G and the Tactile Internet”, European Conference on Networks and Communications (EuCNC), 2017

[13] L. Liu, S. Zhang, R. Du, T. X. Han, S. Cui, “Networked Sensing in 6G Cellular Networks: Opportunities and Challenges,” 2206.00493.pdf (arxiv.org)

[14] Ch. Sarathchandra, S. Robitzsch, M. Ghassemian, and U. Olvera-Hernandez, “Enabling Bi-directional Haptic Control in Next Generation Communication Systems: Research, Standards, and Vision”, IEEE CSCN Conference, 2021

[15] 3GPP TR 22.847 - Supporting tactile & multi-modality comm services (TACMM); Stage 1 (Release 18)

[16] 3GPP TR 23.700 - Study on architecture enhancement for XR and media services (XRM); Stage 1 (Release 18)

[17] 3GPP TS 22.104 V18.3.0 (2021-12) - Service requirements for cyber-physical control applications in vertical domains; Stage 1 (Release 18)

[18] 3GPP TS 22.856 - Study on Localized Mobile Metaverse Services (Release 19)

[19] IEC TC100 PT 63448 - Low and Ultra-low Latency Communication and Control Systems

[20] IEEE P1955- Standard for 6G Empowering Robotics: Use Case Scenarios, Requirements, Architectural Impact, and Technical Assumptions, [online] https://standards.ieee.org/ieee/1955/11660/

[21] ETSI Integrated Sensing and Communication (ISAC) ISG, [online] https://www.etsi.org/committee/2295-isac

[22] C. J. Bernardos, A. Rahman, J. Zuniga, L. Contreras, P. Aranda, and P. Lynch, "Network virtualization research challenges," IRTF draft draftirtf-nfvrg-gaps-network-virtualization-10, 2019

[23] Tactile Internet: with Human-in-the-Loop, edited by Frank H.P. Fitzek, Shu-Chen Li, Stefanie Speidel, Thorsten Strufe, Meryem Simsek, Martin Reisslein [Chapter 3 - Human–robot cohabitation in industry]

[24] Verticals URLLC Use Cases and Requirements by NGMN Alliance- February, 2020

[25] P. Kostoulas, S. Oteafy, and P. Chatzimisios, "Fire-Fighting Drones: A Use Case for Tactile Internet," IEEE International Conference on Communications (ICC), pp. 4613-4618, 2022

[26] G. Kolovou, S. Oteafy, and P. Chatzimisios, "A Remote Surgery Use Case for the IEEE P1918. 1 Tactile Internet Standard," IEEE International Conference on Communications (ICC), pp. 1-6, 2021

[27] 3GPP TS 22.261 V18.6.0 (2022-03) - Service requirements for the 5G system; Stage 1 (Release 18)

[28] M. A. Dehghani and M. B. Menhaj, "Communication free leader–follower formation control of unmanned aircraft systems", Robotics and Autonomous Systems, pp. 69-75, 2016

[29] A. Aijaz, A. Stanoev, and M. Sooriyabandara, "Toward real-time wireless control of mobile platforms for future industrial systems", arXiv preprint arXiv:1907.01979, 2019

[30] X. X. Teh, A. Aijaz, A. Portelli, and S. Jones, "Communications-based formation control of mobile robots: modeling, analysis and performance evaluation", Proceedings of the 23rd International ACM Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems, 2020

[31] A. Aijaz, N. Jiang, and A. Khan, "Toward Multi-Service Edge-Intelligence Paradigm: Temporal-Adaptive Prediction for Time-Critical Control over Wireless", IEEE Internet of Things Magazine 6.1, pp. 96-101, 2023

[32] M. De Ryck, M. Versteyhe, and F. Debrouwere, "Automated guided vehicle systems, state-of-the-art control algorithms and techniques", Journal of Manufacturing Systems 54, pp. 152-173, 2020

[33] A. Sarker, H. Shen, M. Rahman, M. Chowdhury, K. Dey, F. Li, Y. Wang, and H. S. Narman "A review of sensing and communication, human factors, and controller aspects for information-aware connected and automated vehicles", IEEE transactions on intelligent transportation systems 21.1, pp. 7-29, 2019

[34] M. Mizmizi, M. Brambilla, D. Tagliaferri, Ch. Mazzucco, M. Debbah, T. Mach, R. Simeone, S. Mandelli, V. Frascolla, R. Lombardi, M. Magarini, M. Nicoli, and U. Spagnolini, "6G V2X technologies and orchestrated sensing for autonomous driving." arXiv preprint arXiv:2106.16146, 2021

[35] T. Farnham, S. Jones, A. Aijaz, Y. Jin, I. Mavromatis, U. Raza, A. Portelli, A. Stanoev, and M. Sooriyabandara, "Umbrella collaborative robotics testbed and IoT platform", IEEE 18th Annual Consumer Communications & Networking Conference (CCNC), 2021

[36] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, and S. Levine, "More than a feeling: Learning to grasp and regrasp using vision and touch", IEEE Robotics and Automation Letters, vol. 3, no. 4, pp. 3300–3307, 2018

[37] D. Ramachandram and G. W. Taylor, "Deep Multimodal Learning: A Survey on Recent Advances and Trends," in IEEE Signal Processing Magazine, vol. 34, no. 6, pp. 96-108, Nov. 2017, doi: 10.1109/MSP.2017.2738401

[38] T. Baltrušaitis, C. Ahuja and L. -P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423-443, Feb. 2019, doi: 10.1109/TPAMI.2018.2798607

[39] W. C. Skeeman, Ri. Kapoor, and P. Ghosh, "Multimodal Classification: Current Landscape, Taxonomy and Future Directions", ACM Computing Surveys (CSUR), 2021

[40] G. Barnum, S. Talukder, and Y. Yue, "On the Benefits of Early Fusion in Multimodal Representation Learning", 2020, [Online] https://doi.org/10.48550/arXiv.2011.07191

[41] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1798–1828, Aug. 2013

[42] P. Vincent, H. Larochelle, Y. Bengio, and P.A. Manzagol, "Extracting and composing robust features with denoising autoencoders", In Proceedings of the 25th international conference on Machine learning (ICML '08), Association for Computing Machinery, New York, NY, USA, pp. 1096–1103, 2008, [Online] https://doi.org/10.1145/1390156.1390294

[43] N. Koursioumpas, S. Barmpounakis, I. Stavrakakis, and N. Alonistioti, "AI-driven, Context-Aware Profiling for 5G and Beyond Networks," in IEEE Transactions on Network and Service Management, vol. 19, no. 2, pp. 1036-1048, June 2022, doi: 10.1109/TNSM.2021.3126948

[44] A. Holzinger, B. Malle, A. Saranti, and B. Pfeifer, "Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI", Information Fusion, Volume 71, 2021, pp. 28-37, ISSN 1566-2535, [Online] https://doi.org/10.1016/j.inffus.2021.01.008

[45] V. Vukotić, C. Raymond, and G. Gravier, "Generative Adversarial Networks for Multimodal Representation Learning in Video Hyperlinking", In Proceedings of the ACM on International Conference on Multimedia Retrieval (ICMR '17) 2017, Association for Computing Machinery, New York, NY, USA, 416–419. [Online] https://doi.org/10.1145/3078971.3079038

[46] W. Cao, Y. Wu, Y. Sun, H. Zhang, J. Ren, , D. Gu, , and X. Wang, , "A review on multimodal zero-shot learning", WIREs Data Mining and Knowledge Discovery, 13 (2), 2023, e1488, [Online] https://doi.org/10.1002/widm.1488

[47] T. Li, J. Huang, E. Risinger, and D. Ganesan, "Low-latency speculative inference on distributed multi-modal data streams", In Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '21). 2021, Association for Computing Machinery, New York, NY, USA, 67–80, **[Online]** https://doi.org/10.1145/3458864.3467884

[48] YM. Lin, Y. Gao, MG. Gong, SJ. Zhang, YQ. Zhang, and ZY. Li "Federated Learning on Multimodal Data: A Comprehensive Survey", Mach. Intell. Res. 2023, **[Online]** https://doi.org/10.1007/s11633-022-1398-0

[49] S. Ruder, "An overview of multi-task learning in deep neural networks", arXiv preprint arXiv:1706.05098. 2017

[50] R. Collobert, and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning", In Proc. ICML, pp. 160-167, 2007

[51] R. Hu, and A. Singh, "Unit: Multimodal multitask learning with a unified transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1439-1449, 2021

[52] B. Yang, X. Cao, K. Xiong, C. Yuen, Y.L. Guan, S. Leng, L. Qian, and Z. Han, "Edge intelligence for autonomous driving in 6G wireless system: Design challenges and solutions", IEEE Wireless Communications, 28(2), pp. 40-47, 2021

[53] R. Li, W. Liang, C. Peng, X. An, Z. Zhao, and H. Zhang, "Network AI management & orchestration: A federated multi-task learning case. In 2021 IEEE GlobeCom Workshops, pp. 1-6, 2021

[54] L. Nie, Wang, X., Wang, S., Ning, Z., Obaidat, M.S., Sadoun, B. and Li, S., "Network traffic prediction in industrial Internet of Things backbone networks: A multitask learning mechanism", IEEE Transactions on Industrial Informatics, 17(10), pp.7123-7132, 2021

[55] D.S. Chaplot, L. Lee, R. Salakhutdinov, D. Parikh, and D. Batra, "Embodied multimodal multitask learning", 2019, arXiv preprint arXiv:1902.01385

[56] A. Wong, Y. Wu, S. Abbasi, S. Nair, Y. Chen, and M.J. Shafiee, "Fast GraspNeXt: A Fast Self-Attention Neural Network Architecture for Multi-task Learning in Computer Vision Tasks for Robotic Grasping on the Edge." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2292-2296, 2023, [Online] https://arxiv.org/abs/2304.11196

[57] T. Nishio, Y. Koda, J. Park, M. Bennis, and K. Doppler, "When wireless communications meet computer vision in beyond 5G," IEEE Communications Standards Magazine, 5(2), pp.76-83, 2021

[58] Terminal Equipment definition - source: ITU-T I.112

[59] D. González Morín, P. Pérez, and A. García Armada, "Toward the distributed implementation of immersive augmented reality architectures on 5G networks", IEEE Communications Magazine, vol. 60, no. 2, pp. 46–52, 2022. doi: 10.1109/MCOM.001.2100225

[60] D. G. Morin, D. Medda, A. Iossifides, P. Chatzimisios, A. Garcia Armada, A. Villegas, P. Perez "An extended reality offloading IP traffic dataset and models", IEEE Transactions on Mobile Computing, 2023 [Online] https://arxiv.org/abs/2301.11217

[61] D. Gonzalez Morin, M. J. Lopez Morales, A. Perez Pablo Garcia Armada, and A. Villegas, "Fikore: 5G and beyond RAN emulator for application level experimentation and prototyping", 2022. [Online] http://arxiv.org/abs/2204.04290

[62] A. Alkhateeb, G. Charan, T. Osman, A. Hredzak, J. Morais, U. Demirhan, and N. Srinivas, "DeepSense 6G: A Large-Scale Real-World Multi-Modal Sensing and Communication Datasets," available on arXiv, 2022

[63] J. -J. Cabibihan, A. Y. Alhaddad, T. Gulrez and W. J. Yoon, "Influence of Visual and Haptic Feedback on the Detection of Threshold Forces in a Surgical Grasping Task," in IEEE Robotics and Automation Letters, vol. 6, no. 3, pp. 5525-5532, July 2021, doi: 10.1109/LRA.2021.3068934, [Online] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/QGKEUW

[64] J.-J. Cabibihan, A. Y. Alhaddad, T. Gulrez, and W. J. Yoon, "Dataset for influence of visual and haptic feedback on the detection of threshold forces in a surgical grasping task," Data Brief, vol. 42, no. 108045, p. 108045, 2022, [online] https://ieee-dataport.org/documents/dataset-influence-visual-and-haptic-feedback-detection-threshold-forces-surgical-grasping

[65] M. Ciliberto, V. Fortes Rey, A. Calatroni, P. Lukowicz, and D. Roggen, "Opportunity ++: A Multimodal Dataset for Video- and Wearable, Object and Ambient Sensors-based Human Activity Recognition", November, 2021, doi: 10.21227/vd6r-db31

[66] M. Ciliberto, V. Fortes Rey, A. Calatroni, P. Lukowicz, and D. Roggen, "Opportunity++: A multimodal dataset for video- and wearable, object and ambient sensors-based human activity recognition," Front. Comput. Sci., vol. 3, 2021, [Online] https://ieee-dataport.org/open-access/opportunity-multimodal-dataset-video-and-wearable-object-and-ambient-sensors-based-human

[67] O. Kursun and A. Patooghy, "An Embedded System for Collection and Real-Time Classification of a Tactile Dataset," in IEEE Access, vol. 8, pp. 97462-97473, 2020, doi: 10.1109/ACCESS.2020.2996576, [online] https://ieee-dataport.org/open-access/vibtac-12-texture-dataset-collected-tactile-sensors

[68] S. Ergun *et al.,* "A Unified Perception Benchmark for Capacitive Proximity Sensing Towards Safe Human-Robot Collaboration (HRC)," 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 2021, pp. 3634-3640, doi: 10.1109/ICRA48506.2021.9561224,

[online] https://ieee-dataport.org/open-access/supplementary-material-unified-perception-benchmark-capacitive-proximity-sensing-towards

[69] B. Laschowski, W. McNally, A. Wong, and J. McPhee, “ExoNet database: Wearable camera images of human locomotion environments,” Front. Robot. AI, vol. 7, 2020

[70] B. Laschowski, W. McNally, A. Wong, and J. McPhee, “Environment classification for robotic leg prostheses and exoskeletons using deep convolutional neural networks,” Front. Neurorobot., vol. 15, 2022, [online] https://ieee-dataport.org/open-access/exonet-database-wearable-camera-images-human-locomotion-environments

[71] T. Wang, C. Yang, F. Kirchner, P. Du, F. Sun, and B. Fang, “Multimodal grasp data set: A novel visual–tactile data set for robotic manipulation,” Int. J. Adv. Robot. Syst., vol. 16, no. 1, p. 172988141882157, 2019, [online] https://github.com/tsinghua-rll/Visual-Tactile_Dataset

[72] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, “Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection,” Int. J. Rob. Res., vol. 37, no. 4–5, pp. 421–436, 2018, [online] https://sites.google.com/site/brainrobotdata/home

[73] J.-Y. Jeong, Y. Cho, Y.-S. Shin, Hyun Seog Roh, and A. Kim, “Complex urban dataset with multi-level sensors from highly diverse urban environments,” The International Journal of Robotics Research, vol. 38, no. 6, pp. 642–657, Apr. 2019, doi: https://doi.org/10.1177/0278364919843996, [online] https://ieee-dataport.org/open-access/create-multimodal-dataset-unsupervised-learning-and-generative-modeling-sensory-data

[74] S. Brodeur and J. Rouat, "Optimality of inference in hierarchical coding for distributed object-based representations," 15th Canadian Workshop on Information Theory (CWIT), Quebec City, QC, Canada, 2017, pp. 1-5, doi: 10.1109/CWIT.2017.7994828

[75] S. Brodeur, S. Carrier, J. Rouat, "CREATE: Multimodal Dataset for Unsupervised Learning and Generative Modeling of Sensory Data from a Mobile Robot", January 30, 2018, IEEE Dataport, doi: 10.21227/H2M94J, [online] https://ieee-dataport.org/open-access/create-multimodal-dataset-unsupervised-learning-and-generative-modeling-sensory-data

[76] G. Yan, A. Schmitz, S. Funabashi, S. Somlor, T. P. Tomo and S. Sugano, "A Robotic Grasping State Perception Framework With Multi-Phase Tactile Information and Ensemble Learning," in IEEE Robotics and Automation Letters, vol. 7, no. 3, pp. 6822-6829, July 2022, doi: 10.1109/LRA.2022.3151260

[77] G. Yan, "Hard dataset and Normal dataset for robotic tactile sensing", February 2022, IEEE Dataport, doi: 10.21227/94km-0873, [online] https://ieee-dataport.org/documents/hard-dataset-and-normal-dataset-robotic-tactile-sensing

[78] P. Wang, J. Liu, F. Hou, D. Chen, Z. Xia and S. Guo, "Organization and Understanding of a Tactile Information Dataset TacAct For Physical Human-Robot Interaction," 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 2021, pp. 7328-7333, doi: 10.1109/IROS51168.2021.9636389, [online] https://zenodo.org/records/5138841

[79] 3GPP TR 22.804- Study on Communication for Automation in Vertical domains (CAV)

[80] 3GPP TR 22.916- Study on Network of Service Robots with Ambient Intelligence (SOBOT)

[81] 3GPP TR 23.745- Study on application layer support for Factories of the Future in 5G network

[82] 3GPP TR 22.837 V19.0.0, Feasibility study on Integrated Sensing and Communication (Release 19), June 2023

[83] A. Kaushik, R. Singh, M. Li, H. Luo, S. Dayarathna, R. Senanayake, X. An, R. S.-Gallacher, W. Shin, and M. Di Renzo, "Integrated Sensing and Communications for IoT: Synergies with Key 6G Technology Enablers," IEEE Internet of Things Magazine, In Press

[84] Ch. Sarathchandra, M. Ghassemian, and M. Kheirkhah, “Tactile internet service requirements,” IETF draft-sarathchandra-tactile-internet-01, 2021

[85] W. Xu, F. Gao, S. Jin and A. Alkhateeb, "3D Scene-Based Beam Selection for mmWave Communications," in IEEE Wireless Communications Letters, vol. 9, no. 11, pp. 1850-1854, Nov. 2020, [Online] doi: https://doi.org/10.1109/LWC.2020.3005983

[86] Y. Halevi and A. Ray, "Integrated Communication and Control Systems," Journal of Dynamic Systems, Measurement, and Control Dec 1988, Vol. 110/367

[87] R. Yates, Y. Sun, D. R. Brown, S. K. Kaul, E. Modiano, and S. Ulukus, "Age of Information: An Introduction and Survey", IEEE Journal on Selected Areas in Communications, 2021

[88] A. Maatouk, S. Kriouile, M. Assaad, and A. Ephremides, "The Age of Incorrect Information: A New Performance Metric for Status Updates", IEEE/ACM Transactions on Networking, 2020

[89] X. Zheng, S. Zhou, and Z. Niu, "Urgency of Information for Context-Aware Timely Status Updates in Remote Control Systems", IEEE Transactions on Wireless Communications, 2020

[90] O. Ayan, M. Vilgelm, M. Klugel, S. Hirche, and W. Kellerer, "Age-of-information vs. value-of-information scheduling for cellular networked control systems, IEEE/ACM International Conference on Cyber-Physical Systems, 2019

[91] S. Seth and B. Singh, "Sensing Based Contention Access for 6G Low Latency Networks," Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), Grenoble, France, 2022, pp. 309-313, [Online] doi: https://doi.org/10.1109/EuCNC/6GSummit54941.2022.9815795

[92] Z. Wei, H. Qu, Y. Wang, X. Yuan, H. Wu, Y. Du, K. Han, N. Zhang, and Z. Feng"Integrated Sensing and Communication Signals Toward 5G-A and 6G: A Survey," in IEEE Internet of Things Journal, vol. 10, no. 13, pp. 11068-11092, July, 2023, doi: https://doi.org/10.1109/JIOT.2023.3235618

[93] G. Cui, W. Zhang, Y. Xiao, L. Yao, and Z. Fang, "Cooperative Perception Technology of Autonomous Driving in the Internet of Vehicles Environment: A Review", Sensors 2022, 22, 5535. [Online] https://doi.org/10.3390/s22155535

[94] X. Chen, Z. Feng, Z. Wei, P. Zhang, and X. Yuan, "Code-Division OFDM Joint Communication and Sensing System for 6G Machine-Type Communication," in IEEE Internet of Things Journal, vol. 8, no. 15, pp. 12093-12105, Aug 2021, [Online] doi: https://doi.org/10.1109/JIOT.2021.3060858

[95] O. Lang, C. Hofbauer, R. Feger, and M. Huemer, "Range-Division Multiplexing for MIMO OFDM Joint Radar and Communications," in *IEEE Transactions on Vehicular Technology*, vol. 72, no. 1, pp. 52-65, Jan. 2023, [Online] doi: https://doi.org/10.1109/TVT.2022.3203205

[96] ITU-T Focus Group on metaverse (FG-MV), [Online] https://www.itu.int/en/ITU-T/focusgroups/mv/Pages/default.aspx

[97] B. Allen, How Multimodal AI Will Shape the Edge, [Online] https://embeddedcomputing.com/technology/ai-machine-learning/how-multimodal-ai-will-shape-the-edge

[98] A. M. Sampath, Edge Computing For Multimodal, Intermodal Mobility, 2022,2023, [Online]: https://www.mobilityoutlook.com/features/edge-computing-for-multimodal-intermodal-mobility

[99] S. Kekki, W. Featherstone, Y. Fang, P. Kuure, A. Li, A. Ranjan, D. Purkayastha, F. Jiangping, D. Frydman, G. Verin, K.W. Wen, K. Kim, R. Arora, A. Odgers, L. M. Contreras, and S. Scarpina, "MEC in 5G networks", ETSI whitepaper No. 28 June 2018 URL: MEC in 5G networks - June 2018 (etsi.org)

[100] D. Vukobratović, N. Bartzoudis, M. Ghassemian, F. B. Saghezchi, P. Li, A. Aijaz, R. Martinez, X. An, R. R. Venkatesha Prasad, H. Lüders, and S. Mumtaz "Distributed Sensing, Computing, Communication, and Control Fabric: A Unified Service-Level Architecture for 6G", IEEE GlobeCom 2023, [Online] https://arxiv.org/submit/5012485

[101] A. Alhammadi, A. Abraham, A. Fakhreddine, Y. Tian, J. Du and F. Bader, "Envisioning the Future Role of 3D Wireless Networks in Preventing and Managing Disasters and Emergency Situations.," 2024, arXiv preprint arXiv:2402.10600

[102] Y. Shi, K. Yang, T. Jiang, J. Zhang, and K. B. Letaief, "Communication-Efficient Edge AI: Algorithms and Systems," IEEE Commun. Surv. Tutor., vol. 22, no. 4, pp. 2167-2191, 2020

[103] A. Kaushik *et al.*, "Toward Integrated Sensing and Communications for 6G: Key Enabling Technologies, Standardization, and Challenges," in *IEEE Communications Standards Magazine*, vol. 8, no. 2, pp. 52-59, June 2024, [Online] doi: https://doi.org/10.1109/MCOMSTD.0007.2300043

[104] A. Kaushik, Y. C. Eldar and O. A. Dobre, "Integrated Sensing and Communications for Evolution of Next Generation Networks," IEEE Communications Technology News, Apr. 2023

[105] O. Dizdar, A. Kaushik, B. Clerckx and C. Masouros, "Energy Efficient Dual-Functional Radar-Communication: Rate Splitting Multiple Access, Low-Resolution DACs, and RF Chain Selection," IEEE Open Journal of the Communications Society, vol. 3, pp. 986-1006, June 2022

[106] A. Kaushik, E. Vlachos, J. Thompson, M. Nekovee, and F. Coutts, "Towards 6G: Spectrally Efficient Joint Radar and Communication with RF Selection, Interference and Hardware Impairments," IET Signal Processing, vol. 16, no. 7, pp. 851-863, Sept. 2022

[107] E. Vlachos and A. Kaushik, "Covariance-Based Hybrid Beamforming for Spectrally Efficient Joint Radar-Communications," IEEE International Conference on Communications (ICC), pp. 3553-3558, May-June 2023

[108] A. Kaushik, E. Vlachos, C. Masouros, C. Tsinos and J. Thompson, "Green Joint Radar-Communications: RF Selection with Low Resolution DACs and Hybrid Precoding," IEEE International Conference on Communications (ICC), pp. 3160-3165, May 2022

[109] A. Kaushik, C. Masouros and F. Liu, "Hardware Efficient Joint Radar-Communications with Hybrid Precoding and RF Chain Optimization," IEEE International Conference on Communications (ICC), pp. 1-6, June 2021

[110] Integrated Sensing and Communications (ISAC) – From Concept to Practice, HuaweiTech, Nov. 2022, [Online] https://www.huawei.com/en/huaweitech/future-technologies/integrated-sensing-communication-concept-practice

[111] A. Bazzi and M. Chafii, "On Integrated Sensing and Communication Waveforms With Tunable PAPR," *IEEE Transactions on Wireless Communications*, vol. 22, no. 11, pp. 7345-7360, Nov. 2023

[112] A. Kaushik, A. Arora, C. Tsinos, C. Masouros, F. Liu, and S. Chatzinotas, "Waveform Design for Joint Radar-Communications with Low Complexity Analog Components," *IEEE International Symposium on Joint Communications & Sensing*, pp. 1-5, 2022

[113] V. B. Shukla, A. Kaushik, and V. Bhatia, "Channel Estimation and Hybrid Precoding Design for Joint Radar Communication," IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS) Workshop, pp. 1-6, Dec. 2023.

# 7. Distributed Federated AI

Artificial intelligence (AI) and Machine Learning (ML) are among the key technologies shaping the future of the internet and the world. They are significantly changing the way data is collected and analysed to gain better and more important insights on key processes and support decision making in numerous application fields e.g., smart cities, industry 4.0, e-health, smart agriculture etc. The heterogeneity of today's large-scale ubiquitous networks and the need to fulfil the diverse requirements of their users in the best possible way, also mandate the usage of AI/ML approaches [46].

AI represents a tool to solve networking problems that were previously deemed intractable due to their tremendous complexity or the lack of the necessary models and algorithms. A common approach to build an AI-enabled system is to stream all data from source devices to the cloud and perform the model building/training as well as the inference there. However, the large amounts and complexity of data that need to be exchanged, often exceed the network infrastructure capabilities causing challenges with regard to data communication overhead, network delays, privacy and costs. To overcome these challenges, distributed learning and inference techniques have been proposed. In this approach AI components devoted to training/inference tasks are distributed at the edge devices thus alleviating the need to transfer huge amounts of data to the cloud aiming at exploiting the available computing resources in the best possible way.

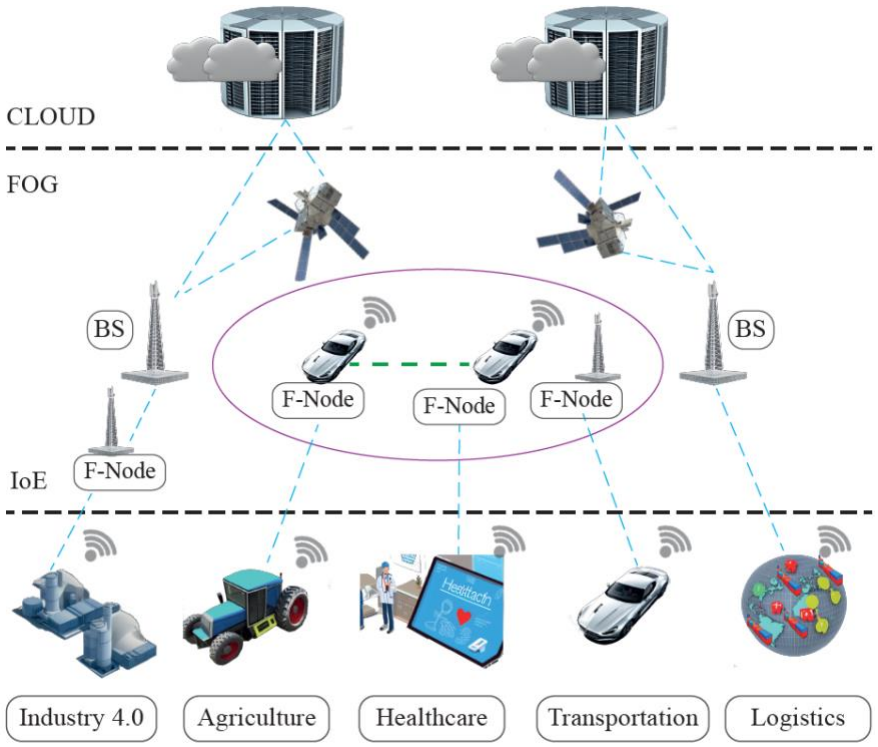



*Figure 19: A carrier network along with connected terminals forms a distributed platform.*

Specifically, a carrier network along with connected terminals forms a distributed platform, where computational resources are distributed from the core to the edge going down to deep-edge resources (such as smart homes, smart factories, smart cars, drones etc.), as depicted in Figure 19. Data is also gathered and collected all over that distributed system including on and from end devices (e.g. vehicles, mobile terminals), on BSs, on routers, on NF instances, on application servers etc. Therefore, there is a great potential in distributed learning and execution; to offload

computation tasks, and, thus, to increase the speed of learning; to bring computation closer to data, decreasing latency, transmission overhead, and cost; and to conform to privacy constraints.

In this section, we provide example use cases, as well as an overview of the state of art and the most promising developments in the exploitation of distributed AI techniques for network management, orchestration and overall optimization. The evolution of the network to support novel distributed AI deployments is also discussed.

## 7.1 Use cases

Distributed learning focuses on information management in systems composed of several components that work together to achieve a desired goal, and applies multi-agent systems to learn and manage the behaviour of independent agents and for the development of complex multi-agent systems [43]. The traditional areas of distributed AI solutions are health information systems, commerce, energy distribution and traffic control. However, many more cases are now being considered, since the vast availability of data in numerous everyday activities has generated new opportunities but centralized application of AI is unfeasible or unpractical.

An example use case includes service provisioning and coordination in carrier networks, with services being executed across multiple distributed network nodes on demand. To process incoming traffic, service components have to be instantiated and traffic assigned to these instances, taking capacities, changing demands, and Quality of Service (QoS) requirements into account. This challenge is usually solved with custom approaches designed by experts, relying on unrealistic assumptions or on knowledge which is not available in practice (e.g., a priori knowledge). Besides, many of these solutions are centralized, and hence suffer from scalability problems. Distributed reinforcement learning solutions [38][39] have shown to be able to address these issues, outperforming existing approaches while requiring much fewer resources to ensure high success rates on real-world network topologies.

Another use case includes computation task offloading for V2X applications. Such applications are usually computation intensive, e.g., inferring from large neural networks or solving non-convex optimization problems. These applications currently reside in the vehicle's onboard units (OBU), but the growing complexity of these tasks calls for alternative options such as offloading to distributed edge clouds [6gEco]. As in the previous case, computation offloading decisions are applied in centralized manner also, relying on unrealistic assumptions such as the availability of global state information at this centralized decision-making component, and hence are unpractical and not scalable. Distributed reinforcement learning approaches, where each vehicle can decide based on its local state information what task to offload and when, are shown to be of great interest, providing great improvements compared with SOTA [41][42], as discussed shortly in Section 7.3.

Finally, autonomous navigation is one of the fields that presents many challenges due to the influence of many factors in decision making, since it must consider nearby vehicles, pedestrians, cyclists, road lanes, among others (see Figure 20). Currently one of the most used AI models is deep learning because we can obtain much more accurate models allowing us to run a safe autonomous driving. Normally the approach used is the one in which automobiles store a huge amount of data centrally on servers in order to perform the offline learning phase. However, a better approach would be via a decentralized model, where the network is a continuous learning network, where cars can share their neural network models and improve their own model in real time [44].

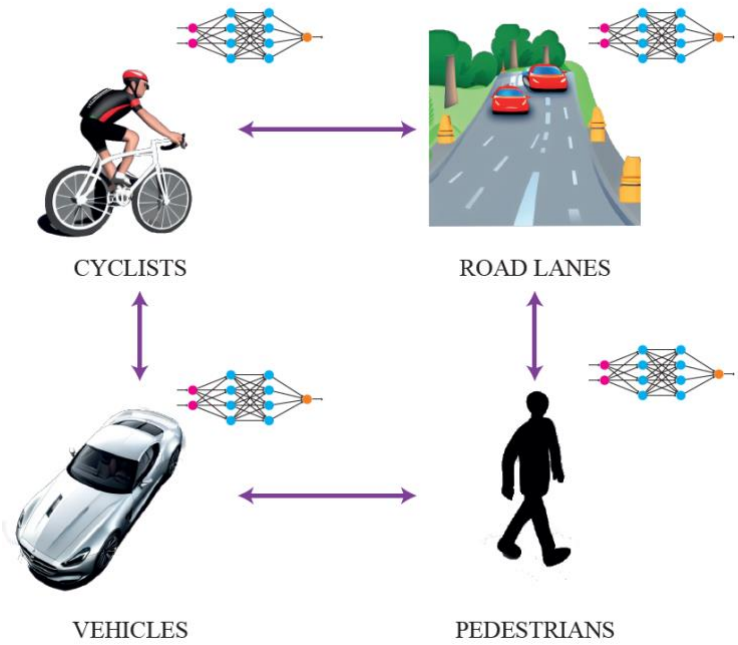



*Figure 20: Autonomous navigation is an example use case of distributed learning.*

## 7.2 State-of-the-art

### 7.2.1 Standardization & Related Initiatives

ML techniques can be applied to improve the performance, smartness, efficiency and security of SDN, as extensively surveyed in [50]. It has been largely recognized that AI and ML techniques can play a pivotal role to enable intelligent and cognitive orchestration of 5G network functionalities [51]. At the standardization level, 3GPP has introduced in the 5G architecture the Network Data Analytics Function (NWDAF) [1], enabling NFs to access to the operator-driven analytics for different purposes, including intelligent (i.e., ML-enabled) slice selection and control.

In another 3GPP study [2], the aspects to be considered in relation to the potential extension of the current NWDAF functionalities towards distributed operations are described. Several proposals are made, among which some considerations include a) NWDAF being responsible for data analytics generation based on a model, which can be trained using different machine-learning algorithms and training data sets; b) to study whether NWDAF functional split is required, and identify the NWDAF functionality that can be separated or placed in a different NF/NF Service. Additionally, key issues are highlighted in this work, such as trained data model sharing between multiple NWDAF instances. Furthermore, some questions raised are the following: Are there use cases, where NWDAF instances can generate data analytics reusing a model trained by other NWDAF instances? How does a NWDAF instance provide the trained data model to other NWDAF instances? Which interactions should have standardized interfaces with NWDAF architecture? To such questions, the same study identifies the employment of Federated Learning as one of the most promising solutions, which could also handle issues related to data privacy and security, model training efficiency, etc.

In ETSI, the Experiential Networked Intelligence Industry Specification Group (ENI ISG) [3] is defining a Cognitive Network Management architecture, using AI techniques and context-aware policies to

adjust offered services based on changes in user needs, environmental conditions and business goals.

ITU's Focus Group on Machine Learning for Future Networks including 5G [4] has generated a number of relevant outputs, such as an architectural framework for machine learning in future networks, use cases' identification for machine learning, framework for evaluating intelligence levels, data handling, etc., as well as a number of deliverables, including requirements, architecture and design for machine learning function orchestration, ML-based end-to-end network slice management and orchestration, vertical-assisted network slicing based on a cognitive network, etc. In the proposed "Architectural framework for machine learning in future networks including IMT-2020", high-level architectural requirements for ML in future networks are introduced, such as enablers for cross-layer data correlation, enablers for ML architectures deployment, namely points of interaction between ML functions and technology-specific underlay network functions, flexible placement of ML functionalities in the underlying network functions, plugging in and out new data sources or configuration targets to running ML environments, ML models' and training/testing data transfer among ML functionalities on different levels, etc.

In addition to the aforementioned initiatives, leading mobile network operators have established the O-RAN Alliance in 2018, with the intention to empower the open RAN with AI technologies, thus making mobile networks more intelligent, open, virtualized and fully interoperable [5].

### 7.2.2 H2020 research projects

Many recent projects also promote the design of AI-driven, cognitive network orchestration and resource management mechanisms.

In 5G-Monarch [6] network resource efficiency is increased through AI-based resource scaling mechanisms for elastic VNF deployment.

Similarly, 5GZORRO [7] uses distributed AI to implement cognitive network orchestration and management with minimal manual intervention.

5G-CLARITY [8] targets to develop a management plane featuring SDN/NFV components together with an AI engine to automate network management by receiving high level intent policies from the network administrator.

5GROWTH [9] introduced a novel architecture, which features an AI/ML core component, namely the 5Growth-AI/ML platform that realizes the concept of AI/ML as a Service (AIMLaaS). The target of the AIMLaaS is to address the needs for AI/ML models for fully automated service management, network orchestration, and resource control within the 5Growth architecture.

DAEMON (Network intelligence for aDAptive and sElf-Learning MObile Networks) project [10] introduces a vision for a Network Intelligence (NI) framework, which will operate across all different "micro-domains", from the core of the network to the (far) edge, and mainly aims at fulfilling the vision of zero-touch network and service management in mobile systems.

MonB5G (Distributed Management of Network Slices in beyond 5G) project [12] focuses on a novel autonomous slice management and orchestration framework that aims to facilitate the mass deployment of slices, relying on state-of-the-art mechanisms based on data-driven AI. The network management system is hierarchical, fault-tolerant, automated, data-driven, and with adaptive, hands-off services.

Hexa-X [13] project is aimed at predictive orchestration and service management, fragmentation without over-provisioning, elasticity of segmentation according to traffic conditions, and real-time,

hands-off automation using multi-action AI mechanisms to achieve intelligent end-to-end orchestration.

## 7.2.3 Academic literature

### 7.2.3.1 From the cloud to the edge: towards distributed AI

As the number of networking applications increases, significant progress has been made in the performance and accuracy of AI-based solutions. However, AI integration into decision-making systems and critical infrastructure still requires assuring end-to-end quality. The common approach to build an AI-enabled system is to stream all the data from source devices (e.g., IoT objects) to the cloud and perform the model building/training as well as the inference there.

As an alternative, the interplay of edge computing and AI, also referred to as "edge intelligence" or "edge AI", recently gained momentum [14], as testified by the flourishing literature as well as by high-end chips commercially available, e.g., Google Edge TPU, that are getting smaller and cost-efficient, hence making feasible to fit them easily in mobile devices or even IoT devices, in agreement to the tinyML philosophy [15]. By pushing AI/ML close to the end-users and where data is produced, edgeAI and tinyML bring advantages to the aforementioned fields mainly in terms of more responsive and privacy-preserving decision making.

However, restricting AI algorithms to run only on edge devices may not be a practical and efficient solution. The most pursued and straightforward approach is to perform the model training into the cloud and run inference at the edge. More sophisticated approaches should instead be pursued, so that the hierarchy of IoT devices, edge and cloud nodes are leveraged to optimize the performance.

Recently, a lot of research initiatives have been devoted on how to adapt (e.g. compress, split) AI/ML models, to fit within edge/end-devices with tighter constraints with regard to the cloud, and on where to place them [14]. The early work in [16] proposes a scheduling strategy that splits a multi-layer DL network into several tasks, with different sizes of intermediate data and computational overhead, to be distributed among edge and cloud nodes.

### 7.2.3.2 Federated learning

Federated learning (FL) is a recently introduced decentralized approach, where learning is performed in a distributed manner on each terminal/device/entity, enabling them to share knowledge without any need to exchange raw data [45]. Thus, one important characteristic of FL is that, since the user generated data are kept locally, it can preserve data privacy and security by design given of course that certain design guidelines are adopted. On the other hand, training in such heterogeneous settings e.g. smartphones, data centres etc. creates new challenges for large scale machine learning and associated optimisations. For example, in [80], a distributed mode of deployment of NWDAF was proposed, in order to optimize network resource consumption while ensuring data security. In particular the following critical aspects are identified: a) the communication cost since a vast number of devices may need to communicate, b) the heterogeneous nature of the systems, c) the heterogeneity of the statistical properties of the collected data and d) the preservation of the privacy of the exchanged data [47]. Thus, the research community should devote efforts to address these challenges in order to achieve the envisaged FL gains fully.

For example, a recent work aims at reducing the communication cost and impact of the heterogeneity of the data generating distributions from different users by enhancing the so called over-the-air (OTA) Federated Learning approach [48]. In this approach all users simultaneously transmit their updates as analog signals over a multiple access channel, and the server receives a superposition of the analog transmitted signals. The authors propose the Convergent OTA FL (COTAF) algorithm which enhances the common local stochastic gradient descent (SGD) FL algorithm and it is shown that can alleviate the channel noise that affects the optimization

procedures in OTA FL. This is achieved by introducing a time-varying precoding and scaling scheme that leads to an effective decay of the noise contribution.

In the field of intelligent transportation systems, FL techniques with heterogeneous model aggregation have been applied and two distributed layers are used to leverage the capabilities of the central cloud to achieve better training efficiency and higher accuracy results [17] the computation and communication energy requirements can be optimized under a latency constraint that affects the learning accuracy as shown in [49]. An iterative algorithm with low complexity, for which, at each iteration is proposed, and closed-form solutions for computation and transmission resources are derived that reduce significantly energy consumption compared to conventional FL.

#### 7.2.3.3 Network for AI

Distributed AI approaches raise challenges concerning the way distributed AI components devoted to training/inference tasks are connected and spread over the cloud continuum. Recently, communication techniques enabling distributed ML have been designed but are mostly limited to the wireless edge [18]. In [19] radio resources are allocated to edge devices depending on the importance of data provided for model training. Similar approaches for retransmission and for joint data selection and radio resource allocation are devised in [20] and [21]. Departing from the radio segment, in [22] the role of network topology in distributed ML is investigated and in [23] the one which speeds up Federated Learning training is selected according to several performance metrics.

The network affects distributed AI performance. For instance, in [24] it is recognized as the bottleneck for distributed Reinforcement Learning (RL) due to the huge data exchange over multiple rounds. Hence, networking capacity should be considered jointly with other resources; besides, it can actively contribute to learning and inference. Despite the interest for pushing in-network computation [25][26], in-network support of distributed AI/ML techniques is still poorly investigated. This issue is early discussed in a recent work [46], where the new network design requirements and challenges for supporting AI applications are preliminarily presented. There, it is emphasized that since emerging AI and ML applications require to transport not only raw data but also information and knowledge, new communication primitives are required, including multipoint-to-multipoint and in-network processing/aggregation due to control and latency constraints. Networks need to manage not only the bandwidth and buffer, but also computing and caching resources. With similar motivations, the need to shift from a network of information to a network of intelligence is also argued in [27].

The work in [28] leverages programmable switches to build an accelerator which conducts computation on packet payloads to facilitate gradient aggregation for distributed RL training. In-network inference is also deemed promising in [29]. In [30], the idea of inference delivery networks (IDN) is proposed as networks of computing nodes that coordinate to satisfy inference requests achieving the best trade-off between accuracy, latency and resource-utilization, adapted to the requirements of the specific application. Conceiving a network enabled distributed AI also passes through proper, even revolutionary, approaches [31]. Through native in-network caching and name-based forwarding, information-centric networking (ICN) can facilitate the orchestration of distributed AI components as early argued in [32][33] and recently investigated also in [34]. There, the interplay between SDN and ICN is suggested to support a network of intelligence. Overall, how to re-engineer the network to support AI needs to be fully unveiled.

The aforementioned trends are well described in the recent survey in [71]. There networking solutions in the literature, ranging from the radio segment to the core, aimed at improving the performance of distributing intelligence throughout the cloud-to-things continuum are critically scanned and guidelines are provided for the design of a “future network for distributed Intelligence".

#### 7.2.3.4 AI-as-a-service

The majority of edge solutions embedding AI capabilities rely mainly on complex Systems on Chips (SoCs), with sophisticated architectures, requiring dedicated hardware accelerators and huge

memory requirements. Alternatives are GP-GPUs, FPGA's, or powerful multi-core devices [35]. Although the number and variety of these accelerators are increasing, they are typically designed for specific AI algorithms, hence introducing additional complexity for platform abstractions. Moreover, AI algorithms are typically tightly coupled to the application that exploits them, so hindering the provisioning of the same offered service to other applications.

Unlike centralized deployments, distributed AI solutions may suffer from interoperability issues, due to fragmented and mainly application-specific solutions [15]. To circumvent this issue, it is crucial to set up mechanisms to identify and discover AI components and build intelligent applications upon them as, for instance proposed in [36]. There, a virtualization layer is designed which is hosted at the network edge and is in charge of the semantic description of AI service requirements needed for augmenting their cognitive capabilities. In such deployments the provision of AI services could be done by third party entities on demand alleviating the need of in-house AI component building capabilities.

#### 7.2.3.5 AI Towards Green Communications

Over the recent years, energy efficiency in wireless communication networks has been the primary focus for a number of research studies, addressing the problem from different perspectives through the exploitation of resource orchestration, computation offloading, as well as load balancing strategies.

A subset of research works is tackling the problem of energy efficiency, considering generic (non-FL related) computation tasks [84]-[88]. In [84], the authors propose novel user cooperation approaches, considering two computation offloading schemes for Mobile Edge Computing (MEC) systems. The proposed approaches improve the energy efficiency for latency-constrained computation, by jointly optimizing the available computation and communication resources. Zhang et al. in [85] introduce an improved SAC (ISAC) algorithm towards a joint optimization of partial task offloading and resource allocation decisions, among various Industrial Internet of Things (IIoT) devices. Another work [86] presents a knowledge plane-based MANO framework, using a twin-delayed double-Q soft Actor-Critic (TDSAC) method towards energy consumption and Virtual Network Function (VNF) instantiation cost minimization. In [87], the authors propose a Q-learning and Double Deep Q Networks (DDQN)-based method to determine the joint policy of computation offloading and resource allocation in a dynamic multi-user MEC system. AlQerm et al. in [88], propose a novel intuitive online reinforcement learning methodology, in order to determine the most energy efficient traffic offloading strategy.

Additionally, there is a number of proposed solutions that consider an FL process as a use case, targeting to address energy-related challenges [89]-[97]. The work in [89] presents a novel framework that targets to minimize the overall energy consumption of a Federated Edge Intelligence (FEI)-supported IoT network, through the joint optimization of multiple key parameters, using the Alternate Convex Search (ACS) algorithm. The authors in [94] propose a joint resource allocation scheme for efficient FL in IoT, aiming at the minimization of the system and learning costs by jointly optimizing bandwidth, computation frequency, transmission power allocation and sub-carrier assignment. Mo et al. [93] focus on minimizing the total energy consumption of a federated edge learning system, subject to a maximum training delay constraint, by optimizing both the communication and computation resources, using the techniques from convex optimization. Another work [90], proposes energy-efficient strategies for bandwidth allocation and scheduling so as to reduce the energy consumption while warranting learning performance in a federated edge learning (FEEL) framework. In [95], the authors formulate a joint learning and communication problem as an optimization problem for FL and a bisection-based algorithm is proposed, whose goal is to minimize the total energy consumption of the system under a latency constraint.

Zhang et al. [91] present an energy-efficient FL framework for Digital Twin (DT)-enabled IIoT that exploits a DDQN, in order to jointly optimize training strategies and resource allocation, considering a dynamic environment. Another work [92], through the use of a Proximal Policy Optimization (PPO)-based actor-critic method, targets the energy efficiency improvement of FL, by jointly minimizing the learning time and energy consumption. The authors in [96] exploit the merits of

Multi-Agent Deep Deterministic Policy Gradient (MADDPG), in an attempt to address the challenges posed by adopting FL into the Internet of Vehicles (IoV) scenario. They design a mobility supported FL participant decision algorithm, which is followed by a joint resource allocation problem targeting to optimize the FL computation and communication costs. Nguyen et al. [97] propose a Deep Q Learning (DQL) algorithm concerning the resource allocation, in a mobility-aware FL network, which aims to maximize the number of successful transmissions, while minimizing the energy and channel costs.

## 7.3 Ongoing research

### 7.3.1 Distributed Privacy Aware Learning Framework for QoS Prediction

Beyond 5G networks bring a new era in system automation, by introducing new and demanding, in terms of Quality of Service (QoS) use cases and applications. Predicting the QoS for end users in a timely manner and enabling service adaptation methods to react in advance in case of QoS degradation is of high importance, especially for safety-critical applications such as in vehicular communications. In parallel, the use of AI has been recognized as a key enabler in future wireless networks by both Industry and Academia [52]. So far, mainly centralized AI approaches [53][54][55][56][57][58] have been employed for QoS prediction, requiring data transfers from source devices to centrally located data centres and, thus, suffering from data privacy, computational complexity and operational costs. Alternative AI solutions have also recently been emerged (e.g. Split Learning, Federated Learning), distributing AI related tasks of reduced computational complexity (e.g. local training) to a number of (in general heterogeneous) devices, while exploiting locally stored datasets and avoiding any raw data exchanges [59].

The benefits of distributed learning are well understood but can be gained as long as some significant challenges are addressed, related to distributing AI in wireless networks. One of the major challenges towards an effective learning is factoring in device heterogeneity, in terms of data types, data distribution, and computational capabilities; model heterogeneity may also be present due to possible variations in model architectural options. The challenge in this case is to strike a good balance between the level of integration of heterogeneous model architectures and learning effectiveness. The latter necessitates the investigation for model aggregation techniques capable of producing robust and well performing models, especially in cases of critical services and applications with stringent QoS requirements. In addition, ways to cope with the anticipated increased burden on network resources are needed: distributed learning typically requires continuous model update transmissions, comprising millions or even billions of model parameters, a burden that increases proportionally to the number of devices participating in the distributed AI mechanism. In view of the limitations of centralized QoS prediction approaches and the challenges in deploying effectively distributed learning, it becomes apparent that new and innovative solutions are required for QoS prediction in future mobile networks. Such solutions should enable a scalable distributed learning execution across heterogeneous network entities and model architectures, while contributing to privacy preservation and ensuring a robust model performance.

Motivated by these open challenges, a novel privacy-aware distributed learning approach is proposed, designated as DISTINQT [61], based on sequence-to-sequence autoencoders, capable of providing accurate predictions to end users for a future time horizon. DISTINQT contributes to data privacy preservation by distributing (among distinct nodes) a number of Neural Network (NN)-Encoders, capable of encoding input data into a non-linear latent representation before any transmission. DISTINQT also allows different architectural options for each NN-Encoder, depending on the data types collected at the different nodes. This flexibility enables the incorporation of diverse knowledge and model architectures into a sole learning process that will enhance the robustness and generalization capabilities of the final QoS prediction model.

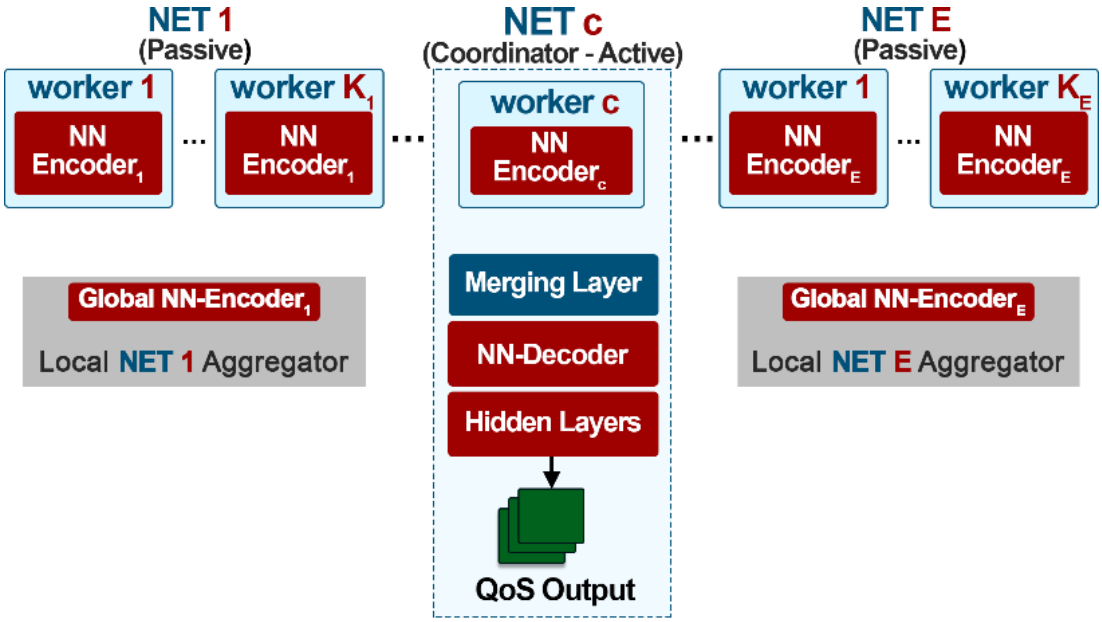



*Figure 21 DISTINQT's Architecture Overview*

More specifically, Figure 21 depicts a high-level overview of the DISTINQT's NN architecture that will be distributed among all involved nodes of different Network Entity Types (NETs).

Role Assignment: The DISTINQT framework is realized in a synchronized manner. At the beginning of the learning phase, each NET is assigned with one of the following roles based on data samples availability.

- Active NET: An active NET holds data samples related with the learning task, as well as the set of ground-truth data (i.e. QoS KPI). Only one NET can be assigned this role.
- Passive NET: A passive NET holds data samples related with the learning task, without any ground-truth data. Multiple NETs can be passive.
- Coordinator NET: A coordinator NET coordinates the learning process in a synchronized manner, by contributing to the learning process with its own data samples, while also orchestrating the communication of the involved NETs.
- Local NET Aggregator: A local aggregator is assigned to each NET by collecting and transmitting all the necessary inputs, aggregating and updating the trainable weights of the involved nodes of a specific NET.

The structure of the DISTINQT's NN architecture is as follows:

- NN-Encoder: Each node is equipped with one NN-Encoder that may be comprised of multiple encoding layers. An NN-Encoder processes an input sequence and encodes it into a fixed-length vector, named as context vector, while applying a non-linear activation function (e.g. the rectified linear activation function - ReLU). This preserves data privacy, since the mapping of the input sequence to a non-linear latent representation is of higher complexity and as such cannot be easily reversed by any other worker or snooper [60]. The nodes of the same NET employ identical NN-Encoders, in terms of number and type of encoding layers, as well as number of neurons per layer. The NN-Encoders used by distinct NETs may differ in their NN structure, depending on the use case, data types (e.g. visual, aural, numerical), and computational capabilities of the nodes. DISTINQT's flexibility in architecture configuration could enable and support a multi-modal learning process. The rest of the NN architecture is located at the coordinator NET and includes a Merging Layer, an NN-Decoder, as well as a number of Hidden Layers.

- Merging Layer: This layer is responsible for applying a merging function (e.g. summation, concatenation, average) to the context vectors produced by all nodes, including the coordinator. The merging is performed between interconnected nodes resided on different NETs.
- NN-Decoder: The NN-Decoder is responsible for constructing the future sequence over the selected time horizon, based on the output of the Merging Layer. It should be noted that the NN-Decoder despite its name, does not attempt to reconstruct any initial input sequence.
- Hidden Layers: This part of the DISTINQT's architecture includes a set of hidden layers (e.g. fully connected), where their number and type could vary, depending on the prediction task.
- Global NN-Encoder: When all NN-Encoders of a NET are updated (trainable parameters) during the learning process, the respective Local NET Aggregator produces a global model update, named as Global NN-Encoder, by aggregating the updated trainable parameters of the corresponding NN-Encoders. The Global NN-Encoder can integrate collective and heterogeneous knowledge from the participating nodes, enhancing the performance, convergence speed and the generalization capabilities of our proposed framework.

The DISTINQT framework motivated the design and integration of a QoS prediction function in a distributed manner, transferring the functionality from the core network to the edge, where edge network entities and end users could be involved in the QoS prediction process in a collaborative manner. Since the core network is decoupled from the QoS prediction process, the signalling overhead as well as the network delays can be significantly reduced. The proposed distributed scheme for QoS prediction, Figure 22, realizes a distributed LSTM-based architecture, considering several distinct NETs (e.g. UE, BS, MEC, Cloud), where each type of NET is responsible of collecting and processing a subset of features related with the prediction task.

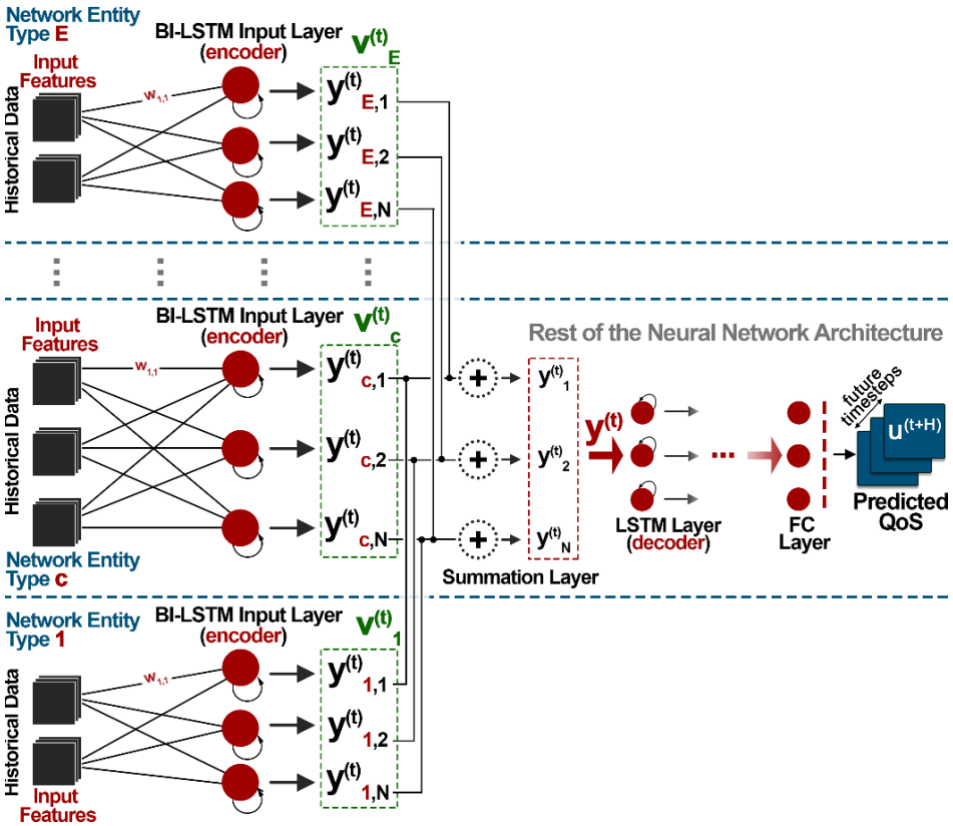



*Figure 22 Distributed QoS prediction scheme*

The objective of the proposed scheme is the prediction of the QoS that a NET experiences over multiple time steps. To achieve this, each NET processes its own historical sequence of data with length equal to a predefined time window. This is achieved with the use of a Bidirectional LSTM (BiLSTM) Auto-encoder that encodes the historical data sequence into a fixed length vector, known as context vector, applying the rectified linear activation function (ReLU). It is assumed that the data privacy can be preserved by using the ReLU function [62]. One of the available NETs is selected to aggregate the context vectors (Aggregator NET) of the distributed BiLSTM Auto-encoders (located at the different NETs), to merge the received vectors with its own, to execute the rest of the Neural Network (NN) architecture and to provide the QoS prediction sequence. The proposed distributed architecture ensures that there are no raw data exchanges between the involved network entities, as only the encoded vectors are transmitted.

The performance of the proposed distributed scheme was evaluated using a simulated environment and over different test case scenarios, exploiting the discrete event simulator (NS3). The evaluation methodology chosen for this study is based on the 3GPP's guidelines. Ten evaluation scenarios, differing on background traffic load and the mobility patterns are exploited for the evaluation results. The performance of the distributed architecture was compared to an LSTM-based centralized architecture [63] and against another two well-known centralized state-of-the-art solutions.

The results prove the effectiveness and feasibility of the distributed QoS prediction scheme, proving a statistically similar performance to the centralized solution, while also preserving end user's privacy. Additionally, it results in a significant reduction of the signalling overhead compared to the centralized solution. Finally, evaluation results show the outperformance of the distributed scheme over the two existing state-of-the-art solutions. The future work includes an investigation towards performing distributed training, as well as enhancements of the proposed scheme's architecture.

### 7.3.2 Decentralized offloading decisions

V2X applications are computation intensive (e.g. inferring or training a large neural network), characterized by huge number of users, dynamic nature, and diverse QoS requirements. These applications currently reside in the onboard units (OBU) of the vehicles. However, these units are limited on computation capabilities. Besides, post-production OBU upgrades for higher on-board computation power are typically not commercially viable. The ability to offload the V2X application to edge/cloud via multi-access edge computing (MEC) devices improves the performance, protecting also vehicles against IT obsolescence, considered as a key technique for future V2X scenarios [64].

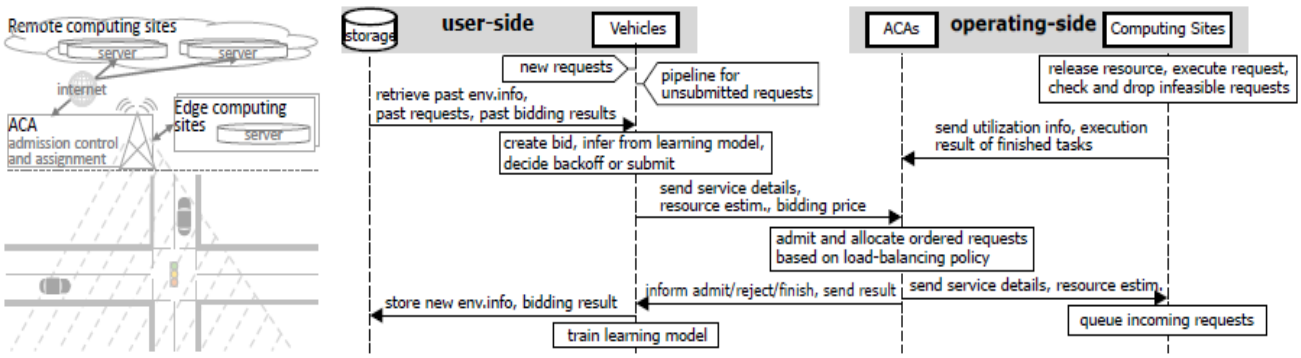



*Figure 23: System model; right: an example topology, left: message sequence.*

Currently, computation offloading decisions are made centrally at the MEC. This however requires centralized data about the state of the network and the traffic requests and preference and requirements imposed by each vehicle. Gathering such data imposes communication and latency overhear. Besides, Centralized modelling is too complex, e.g., in a fast-changing dynamic

environment with many vehicles with different objectives. A centralized solution would be therefore computationally slow/costly, failing also to adapt to the changes in the environment in the runtime.

In a series of work, published recently [41], [42], and [65], we study the application of decentralized Multi-Agent Reinforcement Learning (MARL) techniques to these problem. The system adopts the classic edge cloud computing architecture: user-side vehicles request services; operating-side admission control and assignment (ACA) units (e.g., road side units or base station) control admission of service requests and assign them to different computing sites, which own resources and execute services (Figure 23).

Multi-agent systems use agents to represent individual interests and model complex interaction between players. When a centralized modelling problem is broken down into local, individual models, it reduces model complexity and data requirements. In a MARL system, each agent interacts with the environment to obtain rewards, which allows the agent to learn highly-rewarded behaviour. At each state, each agent takes an action, and the actions of all agents together determine the next state of the environment and reward of agents. Therefore, one major challenge is how to learn in a non-stationary environment when the static properties of the environment is changing over time.

Besides, in such distributed system, each agent can observe partial state information (part of full environment state information available to the agent) and local reward information (as it cannot see the impacts of its actions on the system level) only. Therefore, the second challenge is finding an algorithm that efficiently learns from partial information with just enough feedback signals, keeping information-sharing at a minimum. The third challenge is how to incentivize vehicle's behaviour such that they willingly align their individual goals to the system, long-term, goals. The vehicles can learn about these system goals through delayed and sparse reward signals received from the operators.

We proposed a MARL algorithm which enables each vehicle to learn the best offloading strategy, having access to the local, partial, state information only. So, a vehicle does not know about the state of other vehicles in the environment, or the number of vehicles in its vicinity. It learns its offloading strategy based on the reward signal it received from the MEC, which determines if its bid for offloading a task was successful and the price to pay (MEC applies a second auction method to determine the winners of a bidding round):

- is it better to bid for offloading a task and what is the price to bid for the service; or,
- is it more beneficial to back off the request to future, with the hope that it can get a better and cheaper service price as the network might be less loaded.

Beside the outcome of a biding, the MEC also informs the vehicles times by times about the average traffic load in its computing sites and also some other system metrics such as fairness. Vehicles integrates this delayed sparse signal reward information into their learning, to enhanced their predictive power, as well as better alignment between individual, short-term, and system, long-term, goals.

Evaluation results show that the proposed solution enable the vehicles to align private and system goals without sacrificing either user autonomy or system-wide resource efficiency, despite the distributed design with limited information-sharing. The results also indicate significant performance enhancement achieved using this solution, compared with the state of the art. More details can be found in [65].

The proposed solution however assumes that all players have a single objective, although many real-world problems are multiobjective in nature. One open problem then still remains on how to adapt the proposed framework to multi-objective settings, with different vehicles aiming at optimizing different objectives.

### 7.3.3 Autonomous Navigation

Traditional approaches to solving autonomous navigation require the use of high-tech sensors in order to have a high-quality map and accurate localization to allow the vehicle to choose an optimal trajectory from the starting point to the destination point. To reduce the level of technology required, many researchers focus on using AI techniques to perform path planning without using a map. For this reason, a novel method was proposed in [66], where the authors combined the benefits of Experience Repetition, Heuristic Knowledge, and Genetic Algorithms to obtain a behavioural policy based on Double Deep Reinforcement Learning. The proposed algorithm defines the probability of choosing the action to be executed using directed, autonomous or hybrid knowledge. This means that the agent chooses the action using a routing and avoidance algorithm, by the policy trained, or by the combination of the routing algorithm and the probability of choosing the second-best action based on the Q-value (See Figure 24).

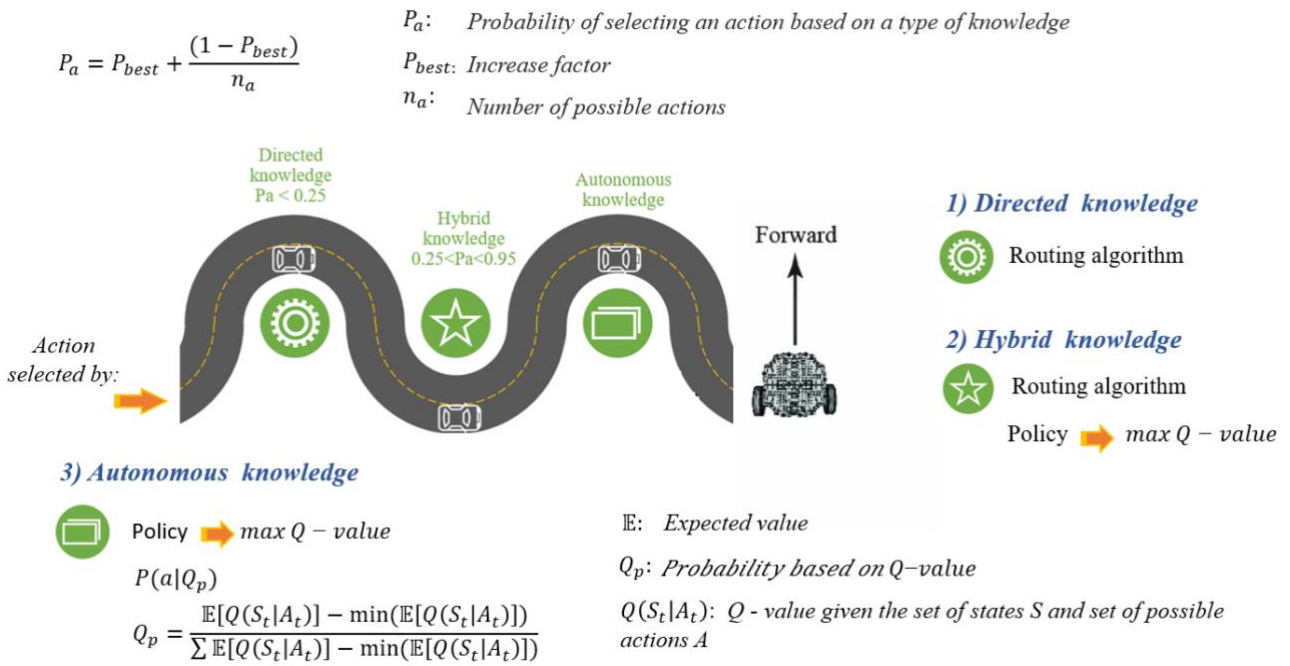



*Figure 24: Description of the transition process from exploration to exploitation through three types of knowledge*

In addition, three memories are defined to store the samples. The main memory stored all the samples. The second stores the best sequences of actions that reach the highest reward, the third stores temporarily the samples of each episode to calculate the best sequence to be stored in the second memory and copy n times in the main memory in order to increase the probability of sampling them during training phase (see Figure 25). In this phase, the batch used to feed the neural network is the result of the combination of the best sequences of actions (second memory) and random samples from the main memory. As a result, the convergence of the network is faster compared to other approaches. Another important fact is since the training was performance using multi-target and multi-source points the generalization problem is approached. Summarizing, the algorithm shows high performance and robustness, as it needs less time to train a policy than other algorithms, even using only the CPU. This demonstrates that it does not require equipment with a high computational capacity. Moreover, its efficiency is demonstrated since it is able to optimize the path chosen by the agent both in terms of time and distance. Therefore, the paths traced are smooth and short.

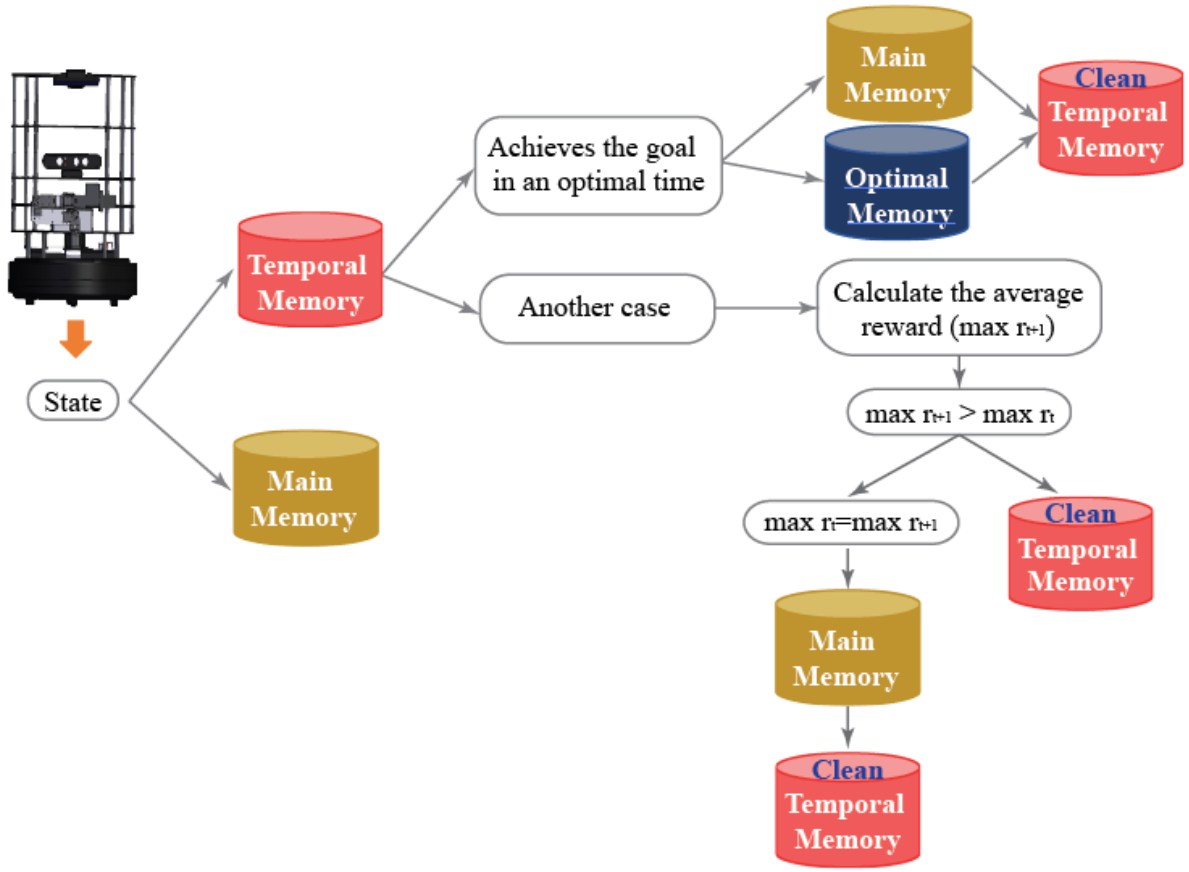



*Figure 25: Description of the process of storing data collected by the agent in three different memories depending on the trajectory performed by the agent*

An alternative for trajectory planning of one agent is proposed in [67], where a multi-layer Q-learning approach uses global and local information to train a lower and an upper layer for local and global planning, respectively. In contrast [68], combines a black hole field strength with a deep Q-learning network to learn an optimal behavioural policy. However, when there is more than one agent in the environment, they will have to interact, which is a major challenge; therefore, in [69] it was proposed that the agent learns a high-level policy to select the target to reach and the action to perform if there is no obstacle. In case it finds an obstacle, it executes a low-level policy that determines the turning angle to avoid collisions. Another interesting approach is proposed in [17] where the action to be performed is chosen by visual observation, the movement intention of local neighbours, and a local policy that is the result of shared knowledge among all agents.

Once an **agent has been provided with autonomy,** the next step is to implement a multi-agent navigation system. This process presents numerous challenges, not only related to network requirements but also in terms of fostering cooperation among agents. Many authors use Independent Learning (IL) to teach agents in a multi-agent system. Where each agent assumes that the other agents are static elements of the environment (see Figure 26). Consequently, each agent may require more time to determine its optimal policy, as it relies only on its own experiences.

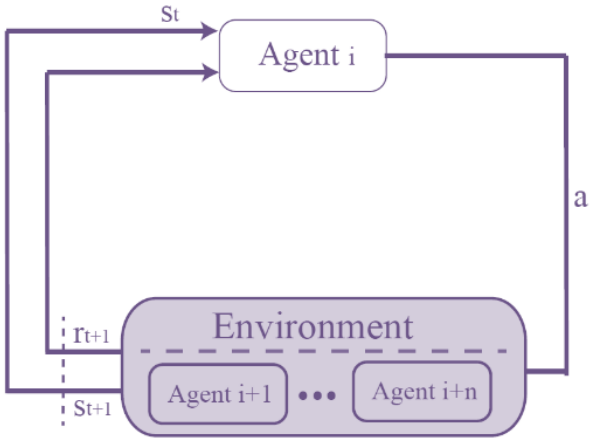



*Figure 26: Description of Independent Learning: each agent selects its action based on the assumption that the other agents represent a static part of its environment*

Conversely, Figure 27 illustrates how an agent learns by considering its actions and those performed by other agents, it is called Joint Action Learning (JAL). This approach allows finding the optimal behavioural policy more quickly, however the computational cost increases as the number of agents in the system increases. On the other hand, IL has the advantage of avoiding the computational complexities associated with JAL, i.e., it offers a scalable solution.

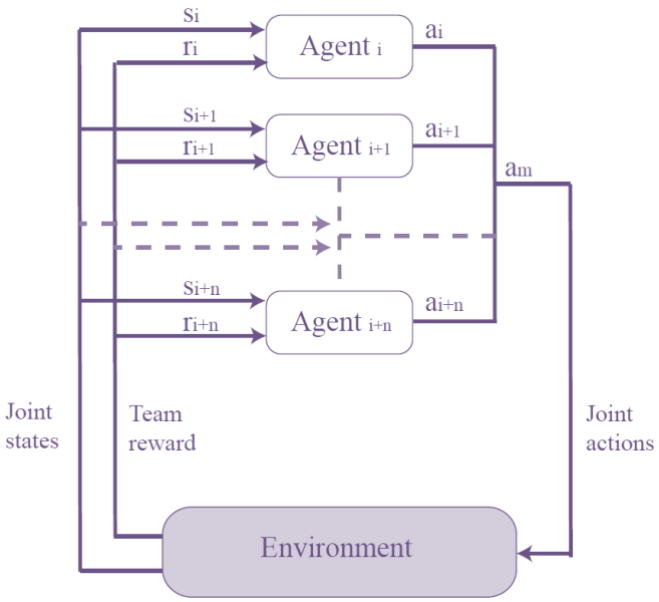



*Figure 27: Description of Joint Action Learning. Each agent learns by using its own actions as well as actions performed by other agents.*

Recognizing the trade-offs between IL and JAL, [101] proposed a novel approach that combines the benefits of both methods. They use the concept of JAL to search for the optimal behavioural policy, but instead of relying on the actions performed by other agents, these actions are provided by a self-advice module. This innovation allows the benefits of IL, maintaining a constant system complexity even as the number of agents increases.

This approach is based on a semi-decentralized architecture, where all the data collected by each agent is sent to a common memory located in the cloud. Therefore, all the data collected by each of the agents will be used for training the behavioural policy, but these data will be treated as if they came from a single agent. As a consequence, an optimal behavioural policy will be derived in a way that guarantees a simple learning model that does not increase the computational complexity with

the number of agents. The authors simplify their algorithm by limiting the number of actions to two: one chosen by the agent's current policy and the other suggested by a self-advice module. This keeps the system manageable and efficient.

In addition, this approach reduces third-party data manipulation, as agents send sensor data to the cloud and the information is only accessible within the cloud, which enhances security and privacy. Regarding the self-advising module, it has two components: the first one is in charge of building a 2D map in the cloud using data from all agents, and the second one involves a routing and obstacle avoidance algorithm, which suggests the optimal action to help the agent avoid collisions. This innovative approach offers a promising solution to the challenges of multi-agent navigation and cooperation.

A collision controller is also designed to evaluate whether the expected future distance between the agents will result in a collision. If a potential collision is detected, the controller determines which agent has the priority tag and applies the k-nearest neighbours (kNN) algorithm to select an action that moves the agent away from the collision points (see Figure 28).

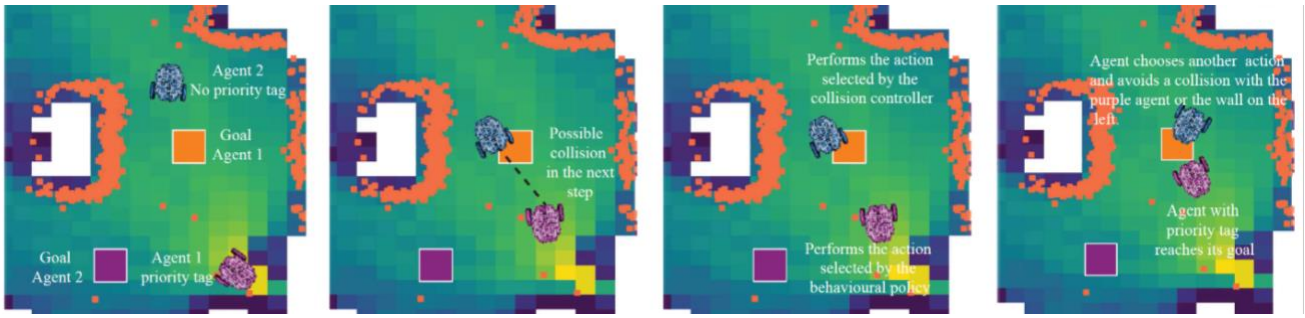



*Figure 28: Explanation of the process performed by the collision controller to minimize even more the number of collisions*

Within this module, each agent calculates its future position according to the action chosen by its behavioural policy. For this calculation, the authors considered that each action has both linear and angular velocities associated with it. Thus, future positions are calculated considering the current position, the expected action and the respective linear and angular velocities.

Summarizing, the cooperative navigation algorithm proposed in [101] has been shown to have high performance and low computational complexity. This algorithm generates a global policy based on data from all agents, where each agent is treated as an independent entity and only two actions are required for each agent, regardless of the number of agents in the system. The action chosen by the global policy and the action of the self-advising module. Within this module, each agent calculates its future position according to the action chosen by its behavioural policy. For this calculation, the authors considered that each action has both linear and angular velocities associated with it. Thus, future positions are calculated considering the current position, the expected action and the respective linear and angular velocities.

The above approach assumes uniformity among the agents, since they are all equipped with the same type of sensor. However, what happens when the agents have heterogeneous systems, i.e., each agent is equipped with a different sensor? To address this challenge, [102] introduced a homogenization algorithm and a data estimation technique.

The authors considered for the homogenization process that the neural network input requires a vector of 24 data. These data correspond to the distances detected by a laser with 360-degree vision. Since the objective is to train a global policy on a system of heterogeneous agents, the data obtained by each of the agents with cameras will be transformed from Point Cloud type data to Laser Scan type data. In simpler terms, the 3D points that are captured by the camera will be converted into 2D data as if they had been captured by a Laser Scan.

It is also considered that when an agent is equipped with multiple sensors, a synchronization process will be executed to ensure that data are collected not only in the same format but also simultaneously. Depending on the field of view of the sensor, the angle will be divided by considering a minimum angle, a maximum angle and an angle of increment. With this increment angle the number of individual scans and the position within the input data will be determined. To calculate the remaining data, an estimation process is employed using an arithmetic approach. Is important to note that this process varies according to the last action performed by the agent. Finally, for those missing data that could not be estimated by geometry, a linear interpolation is used to estimate them (see *Figure 29*).

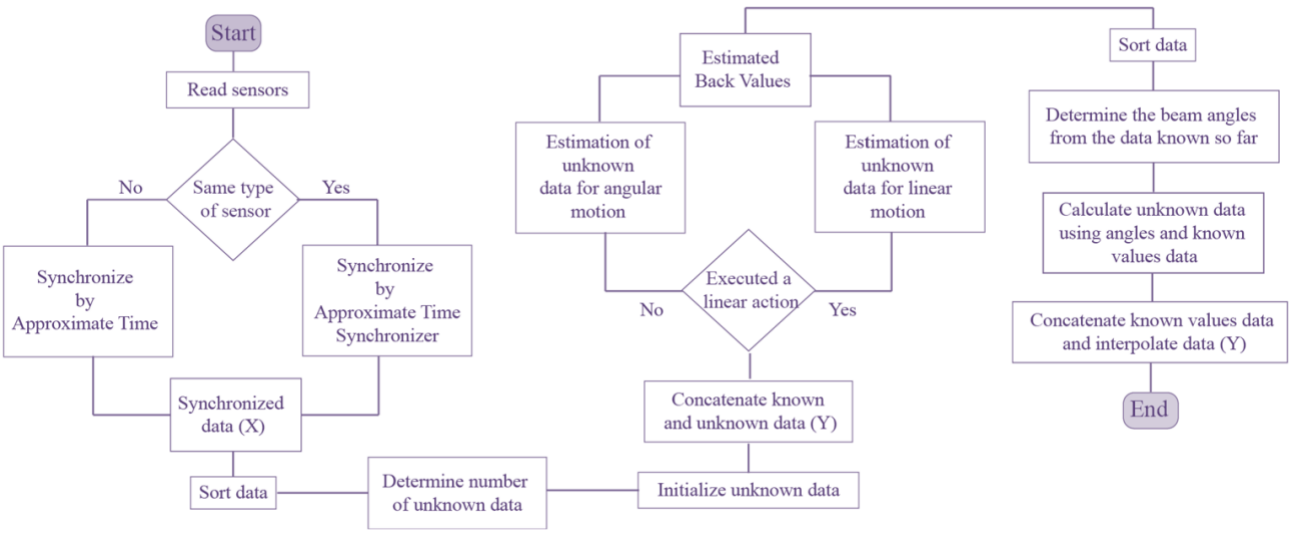



*Figure 29: Process of data homogenization, synchronization and estimation proposed in [102]*

### 7.3.4 Enhanced Client Discovery and Client Selection in Federated Learning

The FL performance is highly dependent on effective selection of client devices participating in distributed model training. Some of them may have low computing/battery resources, making it difficult to perform heavy training tasks; others may have poor datasets that compromise the accuracy of the model.

A further problem that can affect the FL process is the possibility that the scarcity of resources in the network segments crossed by the data streams exchanged between clients and servers negatively affects the overall obtainable performance in terms of convergence rate. The latter one does not represent a big issue for applications that exploit FL for example to make long-term predictions (in the field of e-health, or in the environmental field, etc.) but it is in more dynamic contexts in which applications can be sensitive to the delay in the learning process.

In the light of the above, it can be stated that, if randomly selected by the FL server, some clients can both delay the training procedure and negatively affect the accuracy of the aggregate model.

Starting from this perspective, research is being carried out aimed at enhancing the FL client selection process to support next-generation intelligent applications by acting on different levels.

1. The first one is to investigate how clients (and the relevant capabilities) can be discovered to identify the most suited ones for the learning task, which protocol to use for interactions between the FL server and potential (heterogeneous) clients, and what is the network load caused by the aforementioned procedures.

2. The second one is to understand the potential of an approach that exploits the Software-

Defined Networking (SDN) paradigm to maintain the distributed learning process at high levels of effectiveness and efficiency even in the presence of clients, whose model updates may experience poor network conditions over the path towards the server.

3. Utilizing timing and value of information aspects will lead to more efficient spectrum utilization for client scheduling by considering only the most critical clients, thus, having lower network energy consumption and faster convergence rates.

#### 7.3.4.1 Studies on MQTT-based Client Discovery

Client discovery is a crucial topic in FL. However, the majority of literature solutions focus on the definition of criteria to select clients [72] and not on the procedures needed to efficiently enable a judicious client discovery. To fill this gap, a study that is being carried out in the context of the topics described in the first item is reported in [73] which aims to define:

- An edge-based FL architecture aligned with the European Telecommunications Standards Institute (ETSI) Multi-Access Edge Computing (MEC) specifications to make the most of natively provided MEC services (see grey boxes in Figure 30) in retrieving information about clients. In particular, position and link quality information are retrieved by the FL server to make, with the support of other parameters, a more conscious client selection.

- The use of the Message Queue Telemetry Transport (MQTT) publish/subscribe protocol, suitably configured to enable the FL Server to interact with potential clients to discover their capabilities, e.g., in terms of computing, battery, datasets. The MQTT broker (in Figure 30) decouples the clients and the FL server to allow a greater scalability, because the FL server does not need to keep connections with all the clients.

- The use of the Open Mobile Alliance (OMA) Lightweight Machine-to-Machine (LwM2M) semantics in the MQTT topics to pursue low communication overhead and interoperability in the presence of heterogeneous devices which describe their capabilities.

By departing from conventional networking approaches, the proposal in [74] goes well beyond the usage of MQTT and investigates the Named Data Networking (NDN) paradigm, an ICN instantiation building upon in-network caching, native multicasting, name-based location-independent data forwarding, as an enabler of FL client discovery. Although promising in terms of communication footprint, the viability of implementing NDN besides green-fields deployments is still under debate and deserve further investigation.

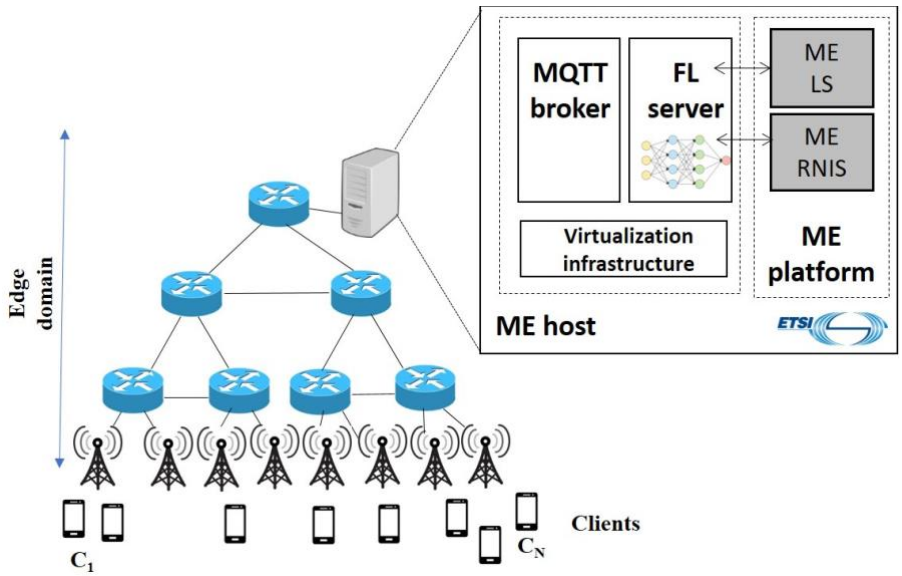



*Figure 30: Reference architecture for research on Edge-assisted MQTT-based Client Discovery [73]*

#### 7.3.4.2 Studies on SDN-assisted FL Client Selection

With reference to the topics described in the second item, the approach that is being followed starts from the very recent literature dealing with the topic of the "Network for AI" (already discussed in previous sections) to propose an SDN-assisted Federated Learning mechanism, in which the possibilities offered by the recent network softwarisation techniques are put at the service of FL processes. In an initial study [75] it has been demonstrated how the use of SDN techniques can reduce training times while maintaining the same accuracy performance. An ongoing study is instead investigating how to implement an SDN-assisted FL solution that leverages a closer collaboration between the Network Controller and an FL Orchestrator and the FL Server for a dynamic selection of clients assisted by the SDN controller during the various rounds of the training process (see Figure 31).

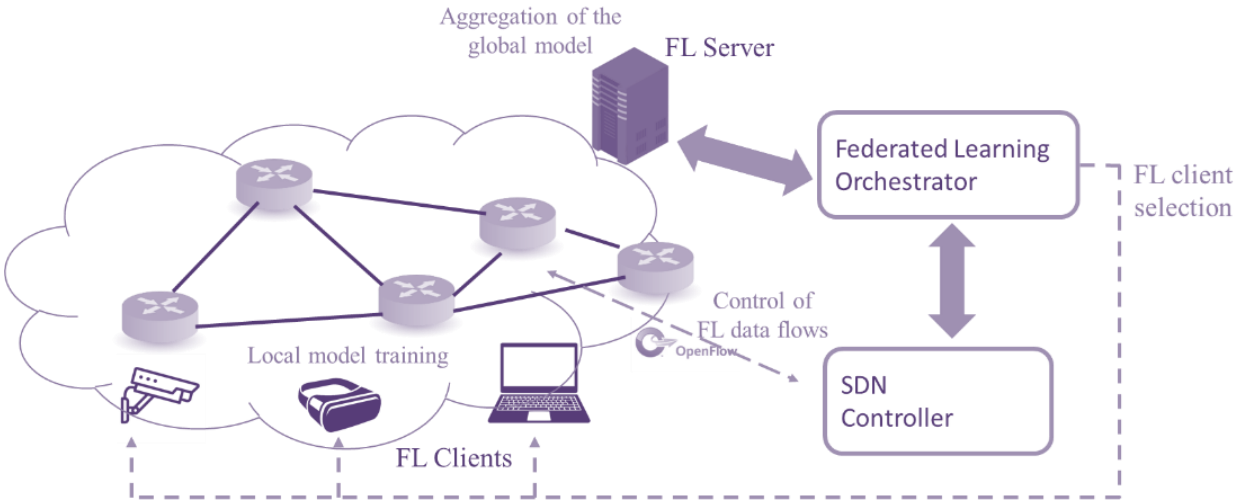



*Figure 31: Possible reference scheme for research on SDN-assisted FL Client Selection.*

### 7.3.5 Resource-aware Federated Learning

Federated learning has proven effective in large-scale systems. However, training of a model, especially a deep NN, is a very resource-hungry (in term of computation, energy, time, etc.) process, and thus it is unrealistic to assume that all the devices in the system can perform all types of training computations all the time. Distributed systems often consist of a heterogeneous set of devices with differing constraints in terms of computational capability (e.g., through the presence of an accelerator), available memory, power (e.g., because of thermal constraints), energy (e.g., for battery-driven devices or devices running on green energy), and etc. This is specifically the case if the learning is distributed over deep edge devices as suggested in 6G systems. Also, to guarantee that a device can deliver its main tasks without any interruption, only free resources available on the device should be allocated to the training procedure. Besides, some devices may use green energy resources, such as solar-cells, which are by nature time variant and thus unreliable. Hence, heterogeneity is not only between the devices but also over the time, as the amount of resources available on a device for training may vary over the time (see Figure 32).

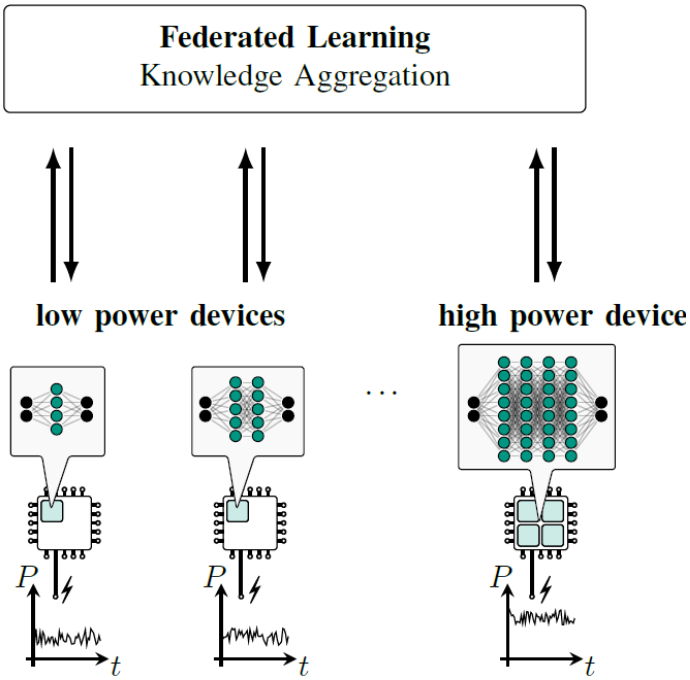



*Figure 32: The availability of resources at devices can vary over devices and over time. These heterogeneities should be considered when applying FL.*

## DISTREAL

A distributed on-device learning mechanism should take all these types of heterogeneity into account. While several works study the problem of heterogeneity across devices [47], time-varying resource availability has so far been neglected. To address these shortcomings, [76] proposes a distributed, resource-aware, adaptive, on-device learning technique, DISTREAL, which enables the system to fully exploit and efficiently utilize available resources on devices for the training, dealing with all these types of heterogeneity. The objective is to maximize the accuracy that is reached after limited training time on devices, i.e., convergence speed. To fulfil this goal, we should make sure that C1) The available resources on a device are fully exploited. This requires fine-grained adjustability of the training model on a device, and a method to instantly react to changes; and C2) The available, limited, resources on a device are used efficiently, to maximize the accuracy improvement and hence the overall convergence speed.

A dropout technique is proposed in [76] to adjust the computational complexity (resource requirements) of training the model at any time. Thereby, each device locally decides the dropout setting which fits its available resources, without requiring any assistance from the server, addressing C1. This is different from the state-of-the-art techniques, such as [77][78][79], where the server is responsible for regulating resource requirements of training for each device at the beginning of each training round, which may take several minutes. DISTREAL is therefore able to reacts in a runtime manner to the changes in the resource availability at the devices participating in the training process.

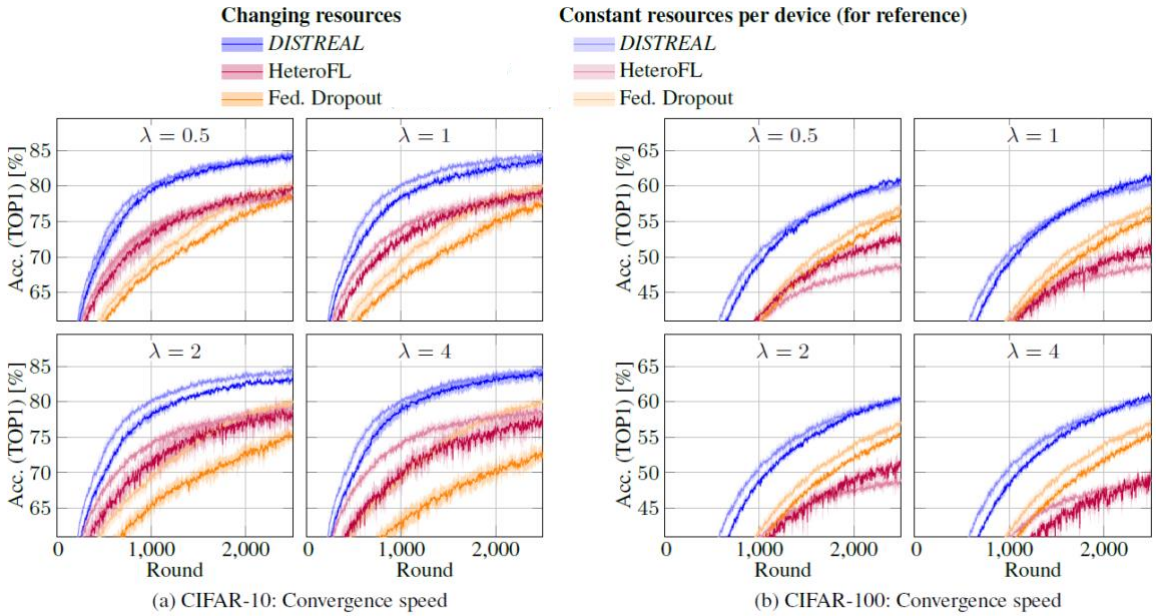



*Figure 33: Convergence with CIFAR-10 and CIFAR-100 on heterogeneous devices where resources availability changes randomly over the time with rate λ. DISTREAL achieves a higher convergence speed than HeteroFL [77] and Fed. Dropout [78], and the highest final accuracy.*

The evaluation results show that using different per-layer dropout rates achieves a much better trade-off between the resource requirements and the convergence speed, compared to using the same rate at all layers as the state of the art, addressing C2. A DSE technique is applied to automatically find the Pareto-optimal dropout vectors at design time. DISTREAL is implemented an FL system, in which the availability of computational resources varies both between devices and over time. It is shown through extensive evaluation that DISTREAL significantly increases the convergence speed over the state of the art, and is robust to the rapid changes in resource availability at devices, without compromising on the final accuracy (e.g. see Figure 33). Note that DISTREAL is scalable, as it does not require any assistance by the server.

## CoCoFL

Note that DISTREAL focuses only on computation constraints at devices. However, in a wireless network, the training devices are at deep edge, communicating with the FL server (which resides in an edge computation centre close to e.g. a base station) through wireless links. These links are limited in terms of throughput and their qualities are varying over time. To enable an efficient distributed training over these devices, we should also consider the communication constraints at these deep edge devices.

In [98], we propose an adaptive mechanism to enable an efficient training of a deep Neural Network over such setting, tuning the computation, memory, and communication overheads of learning tasks to the capabilities of participating edge devices. We propose a novel technique, Communication and Computation-Aware Federated Learning (CoCoFL), that allows all devices to calculate gradients based on the full model, irrespective of their capabilities, through partial freezing and quantization of the model at constrained devices. We show that quantizing frozen layers but keeping trained layers at full precision results in a large reduction in resource requirements, while still enabling efficient learning at devices. Figure 34 shows an overview of the proposed solution.

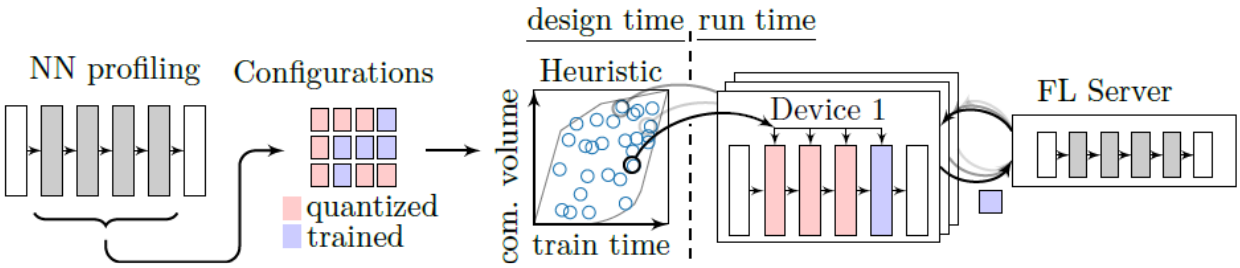



*Figure 34: Overview of CoCoFL. At design time, different configurations of frozen/trained layers are profiled w.r.t. communication, computation, and memory in training. At run time, a heuristic selects a suitable configuration on each device w.r.t. the device's constraints.*

Freezing layers reduces the required gradient computations, the storage of intermediate activations, and the size of the parameter update, while quantization further speeds up the computations of frozen layers. Partial freezing and quantization open up a large design space, where each layer can be frozen or trained on each participating device. The selection of trained layers has a significant impact on the required resources and on the accuracy. We introduce a heuristic that allows for server independent selection of layers with respect to the local resource availability at run time, based on design-time profiling of the performance of devices.

We evaluate performance of CoCoFL in an FL system. We consider three types of devices, strong (with sufficient resource to train the full model and transmit the data to server in time), medium (with 2/3 of communication, computation and memory resource of strong devices), and weak devices (with 1/3 of resource of the strong devices). For each experiment, we distribute the data from the datasets CIFAR10/100, FEMNIST, CINIC10, XChest, IMDB, and Shakespeare over devices. We evaluate ResNet, DenseNet, MobileNet, and Transformer NN models.

We study several data split scenarios:

- An iid case, where data is randomly distributed to all devices, and hence, every device has about the same number of samples per class.
- A non-iid case, where we vary the non-iid-ness with the value of $\alpha$ of a Dirichlet distribution, similar to [99]. Hereby, the number of samples per class varies between devices.
- A rc-non-iid case, where data is resource correlated non-iid. This means that information about certain classes is only available on specific device groups, increasing the necessity to include them in the FL process.

We compare CoCoFL to several baselines: state-of-the-art HeteroFL and FjORD [100], which are the closest to our technique. Additionally, we compare to a theoretical bound and a straightforward baseline that drops all but the strong devices from FL training:

- Centralized: All data is centralized (on one device), serves as a theoretical upper bound.
- FedAvg (full resources): FL is applied, but all devices have full (homogeneous) resources. This baseline serves as a theoretical upper bound.
- FedAvg: Devices that cannot train the NN (e.g., due to limited memory) are dropped from the training (therefore also their data). This serves as a naive baseline, used in production use cases.

*Table 11: Accuracy (Top 1) in % for DenseNet, MobileNet, ResNet18, ResNet50, and Transformer. For XChest the F1 macro score is given (unbalanced data).*

| Topology | DenseNet | | | MobileNet | | | | | ResNet18 | |
|---|---|---|---|---|---|---|---|---|---|---|
| Setting | CIFAR10 | | | CIFAR10 | | CIFAR10 (w. GroupNorm) | | | FEMNIST | |
| Dirichlet $\alpha$ | – (iid) | n.-iid@0.1 | rc@0.1 | – (iid) | rc@0.1 | - (iid) | n.-iid@0.1 | rc@0.1 | – (iid) | rc@0.1 |
| Centralized | | 87.8±0.2 | | | 87.0±0.7 | | 82.5±1.1 | | | 88.1±0.0 |
| FedAvg (f. res.) | 84.3±0.1 | 75.6±1.5 | 74.9±3.3 | 84.9±0.2 | 77.4±2.7 | 76.1 ± 0.1 | 69.4 ± 0.5 | 72.9 ± 1.4 | 86.2±0.1 | 82.9±1.2 |
| CoCoFL (ours) | **82.0±0.2** | **71.9±1.8** | **68.8±4.6** | **83.2±0.3** | **72.4±2.9** | **71.3 ± 0.1** | **61.2 ± 1.4** | **63.6 ± 3.5** | 85.0±0.1 | **81.5±0.6** |
| FjORD | 73.7±0.1 | 60.4±2.1 | 48.8±6.8 | 79.1±0.3 | 51.9±7.3 | 64.4 ± 0.6 | 42.9 ± 1.7 | 47.0 ± 5.0 | 85.5±0.0 | 69.3±8.3 |
| HeteroFL | 76.4±0.3 | 64.0±2.4 | 51.2±7.4 | 79.5±0.2 | 53.0±7.6 | 64.8 ± 0.1 | 55.2 ± 0.3 | 47.5 ± 5.0 | 85.9±0.0 | 70.9±5.8 |
| FedAvg | 76.5±0.1 | 60.4±4.2 | 50.9±7.5 | 78.1±0.4 | 49.9±8.7 | 56.2 ± 0.8 | 52.8 ± 1.1 | 45.9 ± 4.7 | **86.1±0.1** | 64.9±7.8 |

| Topology | ResNet50 | | DenseNet | | | MobileNet(large) | | | TF | TF-S2S | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Setting | CIFAR100 | | CINIC10 | | | XChest | | | IMDB | Shakespeare | |
| Dirichlet $\alpha$ | – (iid) | rc@0.1 | – (iid) | n.-iid@0.1 | rc@0.1 | – (iid) | n.-iid@0.5 | rc@0.5 | – (iid) | – (iid) | –rc (Leaf) |
| Centralized | 61.6±0.4 | | 80.5±0.2 | | | 94.2±0.2 | | | 84.7±0.7 | 52.9±0.7 | |
| FedAvg (f. res.) | 57.0±0.3 | 53.0±0.6 | 77.2±0.1 | 53.9±2.3 | 65.1±1.1 | 94.1±0.3 | 85.9±1.8 | 93.2±0.2 | 82.6±0.4 | 49.1±0.1 | 49.4±0.1 |
| CoCoFL (ours) | **52.5±0.2** | **41.8±2.5** | **73.6±0.1** | **53.5±4.3** | **52.4±7.3** | **91.3±0.3** | **73.0±6.4** | **87.3±3.8** | **82.5±0.5** | **49.3±0.3** | **49.1±0.3** |
| FjORD | 43.6±0.8 | 29.6±4.3 | 65.1±0.7 | 49.2±2.2 | 41.1±6.9 | 66.3±0.9 | 52.7±3.9 | 62.4±0.8 | 78.5±0.7[1] | 42.9±0.5 | 43.0±0.3 |
| HeteroFL | 45.9±0.7 | 31.0±2.3 | 69.4±0.2 | 50.4±2.4 | 43.4±7.2 | 69.4±1.0 | 65.0±0.9 | 65.4±1.6 | 79.2±0.3[1] | 44.1±0.2 | 44.1±0.2 |
| FedAvg | 35.2±0.2 | 23.7±0.4 | 67.7±0.4 | 48.3±2.5 | 42.4±7.2 | 68.2±1.0 | 67.0±0.6 | 66.8±1.4 | 78.5±0.6 | 40.5±0.3 | 40.3±0.1 |

The results are presented in Table 11. We observe that CoCoFL reaches higher accuracies in almost all presented scenarios. Additionally, CoCoFL preserves fairness (accuracy parity) by enabling constrained devices to contribute to the global model. We attribute this large accuracy gap with respect to the baselines to the fact that CoCoFL allows any device to calculate gradients based on the full NN, while still reducing required resources, as opposed to state-of-the-art techniques that calculate gradients on subsets of the filters of the NN.

### 7.3.6 Need for understanding the potential of distributed AI in various use cases for 6G

Distributed intelligence solutions have been proposed previously for solving 6G network management problems, as an effective competing alternative to centralized solutions, and has been investigated extensively for several key use cases. One of these reference use cases is autoscaling CPU resources in a network slice, shown in Figure 35. The network slice is composed of several Network Functions (NF), where each NF is deployed as software on an individual Virtual Network Function (VNF). VNFs share a pool of virtualized computing (CPU) resources. Slice admission control ensures that if, admitting a user would increase the CPU utilization of any VNF beyond the admission threshold, the service request to the user would be denied. VNFs are able to scale up (add CPUs), scale down (release CPUs) or take no action in order to reach a pre-defined target VNF CPU utilization $u_T$, according to the resource efficiency considerations stipulated by the infrastructure proprietor. Moreover, a conflict situation may occur when multiple VNFs attempt to scale up at the same time and the available CPUs in the resource pool is not sufficient to satisfy the combined demand. The conflict causes disproportionate resource allocation to the VNFs, thereby degrading system Key Performance Indicators such as the served load of the slice. Therefore, the goal of autoscaling is to ensure that for a given incoming slice load of users, the load served by the slice is maximized, while keeping the CPU utilization of VNFs close to $u_T$.

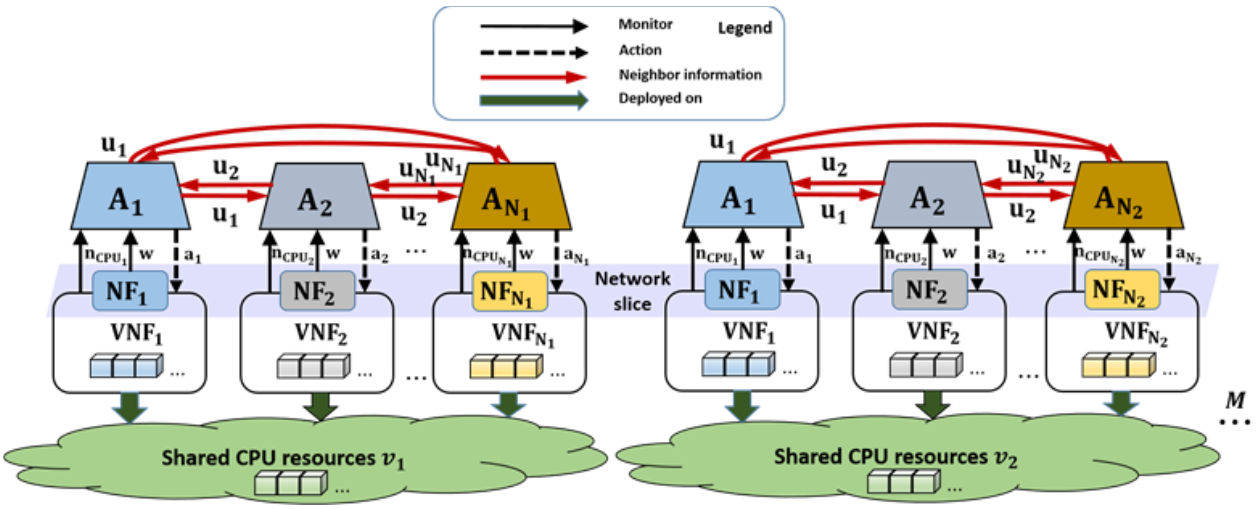



*Figure 35: The Distributed Intelligence Framework for the Autoscaling System*

To achieve efficient autoscaling, we proposed a distributed framework consisting of multiple Intelligent Agents (IAs) [81]. Each Intelligent Agent manages its own NF-VNF pair – monitoring its local variables (no. of users in the slice and the no. of CPUs allocated to its VNF), performing decision-making based on the monitored variables and taking corresponding autoscaling actions. IAs sharing a resource pool are defined as Neighbour Intelligent Agents (NIAs). Since actions taken by an IA are capable of impacting those of its neighbours and vice versa, we empower each IA to monitor additionally the CPU utilization of its neighbours, as highlighted by the red arrows in Figure 35. Monitoring neighbour information provides a greater degree of awareness to the IA for its own actions, hence ultimately encouraging cooperation with its neighbours.

To evaluate the performance of distributed intelligence, we proposed IAs based on Reinforcement Learning – Q-Learning (QL) and Deep Q-Networks (DQN). Simulation results showed that the performance (served load) of these distributed intelligence algorithms for the autoscaling use case deviates at most 6% away from a centralized optimum benchmark based on a Mixed Integer Optimization formulation, for any number of IAs, confirming the performance scalability of distributed RL [82]. Particularly, in terms of accuracy of the Q-values, QL is more accurate than DQN, as the former is a tabular approach, calculating and updating the Q-values iteratively, while DQN approximates a function that represents the Q-values. Consequently, we observed that DQN is at least 37% faster in converging to near-optimum values, compared to QL, when the number of IAs in the framework increases [82]. While the combined convergence time of IAs increases linearly with the number of IAs in DQN, it increases exponentially in QL, severely limiting the scalability of QL in terms of convergence time. Hence, there is a performance-convergence trade-off in the usage of these algorithms.

However, we highlighted a critical issue w.r.t. adopting DQN for distributing intelligence – its training instability [82]. We observed that in some scenarios, when DQN IAs were continued to train after they learned the near-optimal performance, they moved away from the convergence point or diverged. This behaviour of DQN is attributed to the overestimation bias, typically encountered in DQN-based solutions, leading to divergence hence sub-optimal system performance. Hence, this unstable behaviour demonstrated by DQN hinders distributed AI scalability, especially when the number of IAs in the framework increases.

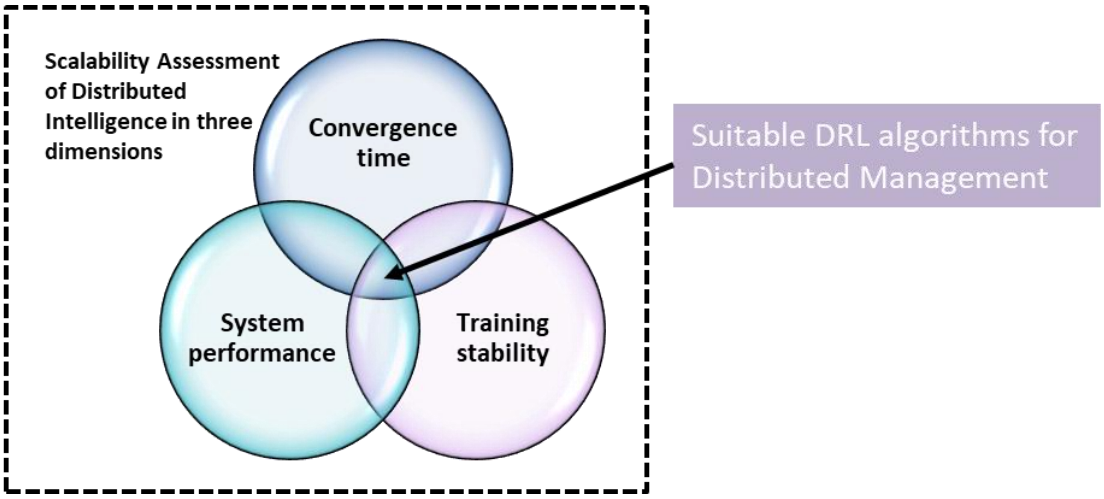



*Figure 36: Three Identified Dimensions of Scalability Assessment of Distributed Intelligence*

Therefore, towards ensuring widespread adoption of RL for 6G network management, the following dimensions relating to the scalability of the distributed algorithms (when the number of IAs increases) need to be addressed in a holistic manner (see Figure 36): 1) system performance, that needs to be near-optimal, 2) convergence time, that needs to be short, and 3) training stability, that ensures that distributed AI does not destabilize network operations. To this end, we designed an algorithm improved Double Deep Q-Network (IDDQN) [103], that is specifically tailored to the scalability requirements of distributed intelligence. In IDDQN, we combined Double Deep Q-Network (DDQN), that addresses the overestimation bias of DQN, and reward scaling, that further stabilizes the training process. Simulation results show that IDDQN eliminates the instability issue encountered in DQN and is at least two times faster than QL, while demonstrating close-to-optimum system performance.

The in-depth investigation of QL, DQN and IDDQN in [103] unveils a number of important lessons learned for distributing intelligence. However, a network operator will be more inclined to embrace RL for 6G network management, if the benefits and potentials of RL are generalizable and can be justified across a number of different use cases or problem scenarios. To this end, it is imperative to explore the applicability of these distributed intelligence solutions, i.e. IDDQN, on other use cases. Presently, we are investigating the application of IDDQN to another use case of interest: Dynamic Task Migration in the 6G User Plane [104]. Additionally, we are exploring a DRL alternative to IDDQN, Advantage Actor Critic (A2C), that has demonstrated good performance and convergence in the literature [105]. Our aim is to understand if A2C brings performance gain over IDDQN, or if additional implications could be derived from distributing intelligence for this use case.

Finally, given the lessons learned from the scalability analysis of distributed intelligence, a concrete analysis of the overall effort needed to integrate these algorithms in 6G, in terms of their impact on the 6G architecture design, is still missing in literature. Hence, deeper investigation is required in this direction, that would ultimately provide hints towards when, or in which use cases, distributing intelligence is worth the effort in 6G, and which other AI algorithms are better candidates.

## 7.4 References

[1] 3GPP, TS 23.288, v16.2.0, Architecture enhancements for 5G System (5GS) to support network data analytics services. Rel. 16. Dec. 2019

[2] 3GPP, TS 23.700, v17.0.0, Study on enablers for network automation for the 5G System (5GS); Phase 2, Rel.17, Dec. 2020

[3] ETSI Experiential Networked Intelligence Industry Specification Group (ENI ISG), https://www.etsi.org/technologies/experiential-networked-intelligence

[4] ITU-T, FG-ML5G, https://www.itu.int/en/ITU-/focusgroups/ml5g/Pages/default.aspx

[5] O-RAN Alliance White Paper, "O-RAN: Towards an Open and Smart RAN," Oct. 2018

[6] 5G-MONARCH, https://www.5g-monarch.eu/

[7] 5GZORRO, https://www.5gzorro.eu/

[8] 5G-CLARITY, https://5g-ppp.eu/5g-clarity/

[9] 5GROWTH, https://5growth.eu/

[10] DAEMON, https://h2020daemon.eu/

[11] AI@EDGE, https://aiatedge.eu/

[12] MONB5G, https://5g-ppp.eu/monb5g/

[13] HEXA-X, https://hexa-x.eu/

[14] Z. Zhou, et al. Edge intelligence: Paving the last mile of artificial intelligence with edge computing. Proceedings of the IEEE, 107(8), 1738-1762, 2019.

[15] H. Doyu, et al. Bringing Machine Learning to the Deepest IoT Edge with TinyML as-a-Service. IEEE IoT Newsletter, 2020.

[16] H. Li, K. Ota, M. Dong. Learning IoT in edge: Deep learning for the Internet of Things with edge computing. IEEE Network, 32(1), 96-101, 2018.

[17] X. Zhou, W. Liang, J. She, Z. Yan and K. I. -K. Wang, "Two-Layer Federated Learning With Heterogeneous Model Aggregation for 6G Supported Internet of Vehicles," in IEEE Transactions on Vehicular Technology, vol. 70, no. 6, pp. 5308-5317, June 2021, doi: 10.1109/TVT.2021.3077893.

[18] D. Gündüz, et al. Machine learning in the air. IEEE JSAC, 37(10), 2019.

[19] D. Liu, et al. Data-importance aware user scheduling for communication-efficient edge machine learning. IEEE Trans. on Cognitive Communications and Networking, 2020.

[20] D. Liu, et al. Wireless data acquisition for edge learning: Data-importance aware retransmission. IEEE Trans. on Wireless Communications, 2020.

[21] Y. He, Importance-Aware Data Selection and Resource Allocation in Federated Edge Learning System. IEEE Trans. on Vehicular Technology, 69(11), 2020.

[22] G. Neglia, et al. The role of network topology for distributed machine learning. IEEE INFOCOM 2019.

[23] O. Marfoq, et al. Throughput-Optimal Topology Design for Cross-Silo Federated Learning, 2020.

[24] Z. Zhang, et al. Is network the bottleneck of distributed training? Workshop on Network Meets AI & ML, 2020.

[25] G. Bianchi, et al. Back to the Future: Towards Hardware" Netputing" Architectures (position paper). IEEE MedComNet'20.

[26] IETF Computing in the Network Research Group (coinrg) https://datatracker.ietf.org/rg/coinrg/about/

[27] F.R. Yu, From Information Networking to Intelligence Networking: Motivations, Scenarios, and Challenges. IEEE Network, 2021

[28] Y. Li, et al. Accelerating distributed reinforcement learning with in-switch computing. ACM/IEEE ISCA 2019.

[29] Z. Xiong, et al. Do Switches Dream of Machine Learning? Toward In-Network Classification. ACM HoTNet Workshop, 2019.

[30] T.S. Salem, G. Castellano, G. Neglia, F. Pianese, A. Araldo, Towards Inference Delivery Networks: Distributing Machine Learning with Optimality Guarantees. 2021, arXiv preprint arXiv:2105.02510.

[31] ITU FG-NET2030 – Focus Group on Technologies for Network 2030, Additional Representative Use Cases and Key Network Requirements for Network 2030, June 2020.

[32] D. Aguiari, et al. C-Continuum: Edge-to-Cloud computing for distributed AI. IEEE INFOCOM 2019 Workshops.

[33] C. Campolo, M. Amadeo, G. Lia, G. Ruggeri, A. Iera, A. Molinaro, Towards Named AI Networking: Unveiling the Potential of NDN for Edge AI. In International Conference on Ad-Hoc Networks and Wireless 2020

[34] X. Li, R. Xie, F.R. Yu, T. Huang, Y. Liu, (2021). Advancing Software-Defined Service-Centric Networking Toward In-Network Intelligence. IEEE Network.

[35] J. Tang, D. Sun, S. Liu, J.L. Gaudiot, Enabling deep learning on IoT devices. Computer, 50(10), 92-96, 2017.

[36] C. Campolo, G. Genovese, A. Iera, A. Molinaro, Virtualizing AI at the distributed edge towards intelligent IoT applications. Journal of Sensor and Actuator Networks, 10(1), 13, February 2021.

[37] ITU-T Y.3170-series – Machine learning in future networks including IMT-2020: Use cases.

[38] S. Schneider, R. Khalili, A. Manzoor, H. Qarawlus, R. Schellenberg, H. Karl, A. Hecker, “Self-Learning Multi-Objective Service Coordination Using Deep Reinforcement Learning”, in IEEE Transactions on Network and Service Management (TNSM), 2021.

[39] S. Schneider, H. Qarawlus, and, H. Karl, “Distributed Online Service Coordination Using Deep Reinforcement Learning", in IEEE ICDCS 2021.

[40] C. J. Bernardos et al., “European vision for the 6g network ecosystem,” The 5G Infrastructure Association, 2021.

[41] J. Tan, R. Khalili, H. Karl, A. Hecker, “Multi-Agent Distributed Reinforcement Learning for Making Decentralized Offloading Decisions”, IEEE INFOCOM 2022.

[42] J. Tan, R. Khalili, H. Karl, “Learning to Bid Long-Term: Multi-Agent Reinforcement Learning with Long-Term and Sparse Reward in Repeated Auction Games”, AAAI 2022 RLG workshop.

[43] Stone, P., Veloso, M. Multiagent Systems” A Survey from a Machine Learning Perspective”, Autonomous Robots 8, 345–383 (2000)

[44] M. Bertogna, P. Burgio, G. Cabri and N. Capodieci, "Adaptive Coordination in Autonomous Driving: Motivations and Perspectives," 2017 IEEE 26th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)

[45] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in 20th International Conference on Artificial Intelligence and Statistics, 2017, pp. 1273–1282.

[46] J. Pan, L. Cai, S. Yan, X.S. Shen, Network for AI and AI for Network: Challenges and Opportunities for Learning-Oriented Networks. IEEE Network, 2021

[47] T. Li, A. K. Sahu, A. Talwalkar and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," in IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 50-60, May 2020, doi: 10.1109/MSP.2020.2975749.

[48] T. Sery, N. Shlezinger, K. Cohen and Y. C. Eldar, "Over-the-Air Federated Learning From Heterogeneous Data," in IEEE Transactions on Signal Processing, vol. 69, pp. 3796-3811, 2021, doi: 10.1109/TSP.2021.3090323.

[49] Z. Yang, M. Chen, W. Saad, C. S. Hong and M. Shikh-Bahaei, "Energy Efficient Federated Learning Over Wireless Communication Networks," in IEEE Transactions on Wireless Communications, vol. 20, no. 3, pp. 1935-1949, March 2021, doi: 10.1109/TWC.2020.3037554.

[50] J. Xie, et. al. A survey of machine learning techniques applied to software defined networking (SDN): Research issues and challenges. IEEE Communications Surveys & Tutorials, 21(1), 393-430, 2018.

[51] D.M. Gutierrez-Estevez, et al. Artificial intelligence for elastic management and orchestration of 5G networks. IEEE Wireless Communications, 2019, 26.5: 134-141.

[52] A. Kaloxylos, A. Gavras, D. Camps Mur, M. Ghoraishiand H. Hrasnica, "AI and ML – Enablers for Beyond 5G Networks". Zenodo, Dec. 01, 2020. doi: 10.5281/zenodo.4299895.

[53] R. Zhohov, A. Palaios and P. Geuer, "One Step Further: Tunable and Explainable Throughput Prediction based on Large-scale Commercial Networks," 2021 IEEE 4th 5G World Forum (5GWF), Montreal, QC, Canada, 2021, pp. 430-435, doi: 10.1109/5GWF52925.2021.00082.

[54] D. Minovski, N. Ögren, K. Mitra and C. Åhlund, "Throughput Prediction Using Machine Learning in LTE and 5G Networks," in IEEE Transactions on Mobile Computing, vol. 22, no. 3, pp. 1825-1840, 1 March 2023, doi: 10.1109/TMC.2021.3099397.

[55] B. Sliwa, H. Schippers and C. Wietfeld, "Machine Learning-Enabled Data Rate Prediction for 5G NSA Vehicle-to-Cloud Communications," 2021 IEEE 4th 5G World Forum (5GWF), Montreal, QC, Canada, 2021, pp. 299-304, doi: 10.1109/5GWF52925.2021.00059.

[56] M. A. Gutierrez-Estevez, Z. Utkovski, A. Kousaridas and C. Zhou, "A Statistical Learning Framework for QoS Prediction in V2X," 2021 IEEE 4th 5G World Forum (5GWF), Montreal, QC, Canada, 2021, pp. 441-446, doi: 10.1109/5GWF52925.2021.00084.

[57] A. Kousaridas et al., "QoS Prediction for 5G Connected and Automated Driving," in IEEE Communications Magazine, vol. 59, no. 9, pp. 58-64, September 2021, doi: 10.1109/MCOM.110.2100042.

[58] A. Palaios et al., "Effect of Spatial, Temporal and Network Features on Uplink and Downlink Throughput Prediction," 2021 IEEE 4th 5G World Forum (5GWF), Montreal, QC, Canada, 2021, pp. 418-423, doi: 10.1109/5GWF52925.2021.00080.

[59] Omar Nassef, Wenting Sun, Hakimeh Purmehdi, Mallik Tatipamula, Toktam Mahmoodi, 'A survey: Distributed Machine Learning for 5G and beyond', Computer Networks, Volume 207, 2022, 108820, ISSN 1389-1286, https://doi.org/10.1016/j.comnet.2022.108820.

[60] H. Ning, Ed., Cyberspace Data and Intelligence, and Cyber-Living, Syndrome, and Health. Singapore: Springer, 2019. doi: 10.1007/978-981-15-1922-2.

[61] N. Koursioumpas, L. Magoula, I. Stavrakakis, N. Alonistioti, M. A. Gutierrez-Estevez, and R. Khalili, "DISTINQT: A Distributed Privacy Aware Learning Framework for QoS Prediction for Future Mobile and Wireless Networks," arXiv preprint arXiv:2401.10158, 2024.

[62] H. Ning, Cyberspace Data and Intelligence, and Cyber-Living, Syndrome, and Health. Springer, Singapore, 2019

[63] S. Barmpounakis, L. Magoula, N. Koursioumpas, R. Khalili, J. M. Perdomo, and R. P. Manjunath, "Lstm-based qos prediction for 5g-enabled connected and automated mobility applications," in 2021 IEEE 4th 5G World Forum (5GWF), 2021, pp. 436–440.

[64] C. J. Bernardos et al., "European vision for the 6g network ecosystem," The 5G Infrastructure Association, 2021.

[65] Jing Tan, Ramin Khalili, Holger Karl, Artur Hecker, "Multi-agent reinforcement learning for long-term network resource allocation through auction: A V2X application", in Elsevier Computer Communications Journal, Volume 194, October 1, 2022, pp. 333-34.

[66] E. Palacios-Morocho, S. Inca and J. F. Monserrat, "Multipath Planning Acceleration Method With Double Deep R-Learning Based on a Genetic Algorithm," in IEEE Transactions on Vehicular Technology, doi: 10.1109/TVT.2023.3277981.

[67] Cui, Zhengyang, and Yong Wang. "UAV path planning based on multi-layer reinforcement learning technique." IEEE Access 9 (2021): 59486-59497.

[68] Q. Yao et al., "Path Planning Method with Improved Artificial Potential Field—A Reinforcement Learning Perspective," in IEEE Access, vol. 8, pp. 135513-135523, 2020, doi: 10.1109/ACCESS.2020.3011211.

[69] Y. Jin, S. Wei, J. Yuan and X. Zhang, "Hierarchical and Stable Multiagent Reinforcement Learning for Cooperative Navigation Control," in IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 1, pp. 90-103, Jan. 2023, doi: 10.1109/TNNLS.2021.3089834.

[70] Z. Liu et al., "Visuomotor Reinforcement Learning for Multirobot Cooperative Navigation," in IEEE Transactions on Automation Science and Engineering, vol. 19, no. 4, pp. 3234-3245, Oct. 2022, doi: 10.1109/TASE.2021.3114327.

[71] C. Campolo, A. Iera, A. Molinaro. Network for Distributed Intelligence: A Survey and Future Perspectives. IEEE Access, 2023.

[72] S. AbdulRahman, H. Tout, A. Mourad, C. Talhi. FedMCCS: Multicriteria client selection model for optimal IoT federated learning. IEEE Internet of Things Journal, 8(6), 4723-4735, 2020.

[73] G. Genovese, G. Singh, C. Campolo, and A. Molinaro. Enabling Edge-based Federated Learning through MQTT and OMA Lightweight-M2M. IEEE Vehicular Technology Conference (VTC2022-Spring) (pp. 1-5).

[74] M. Amadeo, C. Campolo, A. Iera, G. Ruggeri, A. Molinaro. Client Discovery and Data Exchange in Edge-based Federated Learning via Named Data Networking. IEEE International Conference on Communications (ICC) 2022.

[75] A. Mahmod, G. Caliciuri, P. Pace, and A. Iera. Improving the quality of Federated Learning processes via Software Defined Networking. International Workshop on Networked AI Systems (NetAISys'23), June 18, 2023, Helsinki, Finland.

[76] M. Rapp, R. Khalili, K. Pfeiffer, J. Henkel, "DISTREAL: Distributed Resource-Aware Learning in Heterogeneous Systems", AAAI 2022

[77] E. Diao, J. Ding, and V. Tarokh, "HeteroFL: Computation and Communication Efficient Federated Learning for Heterogeneous Clients" IEEE International Conference on Learning Representations (ICLR), 2021.

[78] S. Caldas, J. Konecny, H. B. McMahan, and A. Talwalkar, "Expanding the Reach of Federated Learning by Reducing Client Resource Requirements", arXiv preprint arXiv:1812.07210.

[79] S. Horvath, S. Laskaridis, M. Almeida, I. Leontiadis, S. I. Venieris, N. D. Lane, "FjORD: Fair and Accurate Federated Learning under heterogeneous targets with Ordered Dropout", NeurIPs 2021.

[80] Y. Jeon, H. Jeong et al., "A Distributed NWDAF Architecture for Federated Learning in 5G," in 2022 IEEE International Conference Consumer Electronics (ICCE), pp. 1–2.

[81] S. Majumdar, R. Trivisonno and G. Carle, "Understanding Exploration and Exploitation of Q-Learning Agents in B5G Network Management," 2021 IEEE Globecom Workshops (GC Wkshps), 2021.

[82] S. Majumdar, L. Goratti, R. Trivisonno and G. Carle, "Improving Scalability of 6G Network Automation with Distributed Deep Q-Networks," GLOBECOM 2022.

[83] Walia, Jaspreet Singh, et al. "A virtualization infrastructure cost model for 5g network slice provisioning in a smart factory." Journal of Sensor and Actuator Networks 10.3 (2021): 51.

[84] X. Cao, F. Wang, J. Xu, R. Zhang, and S. Cui, "Joint computation and communication cooperation for energy-efficient mobile edge computing," IEEE Internet of Things Journal, vol. 6, no. 3, pp. 4188–4200, 2019.

[85] F. Zhang, G. Han, L. Liu, M. Martinez-Garcia, and Y. Peng, "Deep reinforcement learning based cooperative partial task offloading and resource allocation for iiot applications," IEEE Transactions on Network Science and Engineering, pp. 1–1, 2022.

[86] F. Rezazadeh, H. Chergui, L. Christofi, and C. Verikoukis, "Actor- critic-based learning for zero-touch joint resource and energy control in network slicing," in ICC 2021 - IEEE International Conference on Communications, 2021, pp. 1–6.

[87] H. Zhou, K. Jiang, X. Liu, X. Li, and V. C. M. Leung, "Deep reinforcement learning for energy-efficient computation offloading in mobile-edge computing," IEEE Internet of Things Journal, vol. 9, no. 2, pp. 1517–1530, 2022.

[88] I. AlQerm and B. Shihada, "Energy efficient traffic offloading in multi- tier heterogeneous 5g networks using intuitive online reinforcement learning," IEEE Transactions on Green Communications and Network- ing, vol. 3, no. 3, pp. 691–702, 2019.

[89] Q. Wang, Y. Xiao, H. Zhu, Z. Sun, Y. Li, and X. Ge, "Towards energy- efficient federated edge intelligence for iot networks," in 2021 IEEE 41st International Conference on Distributed Computing Systems Workshops (ICDCSW), 2021, pp. 55–62.

[90] Q. Zeng, Y. Du, K. Huang, and K. K. Leung, "Energy-efficient radio resource allocation for federated edge learning," in 2020 IEEE Inter- national Conference on Communications Workshops (ICC Workshops), 2020, pp. 1–6.

[91] J. Zhang, Y. Liu, X. Qin, and X. Xu, "Energy-efficient federated learning framework for digital twin-enabled industrial internet of things," in 2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), 2021, pp. 1160–1166.

[92] Y. Zhan, P. Li, L. Wu, and S. Guo, "L4l: Experience-driven compu- tational resource control in federated learning," IEEE Transactions on Computers, vol. 71, no. 4, pp. 971–983, 2022.

[93] X. Mo and J. Xu, "Energy-efficient federated edge learning with joint communication and computation design", Journal of Communications and Information Networks, vol. 6, no. 2, pp. 110–124, 2021.

[94] J. Ren, J. Sun, H. Tian, W. Ni, G. Nie, and Y. Wang, "Joint resource allocation for efficient federated learning in internet of things supported by edge computing," in 2021 IEEE International Conference on Com- munications Workshops (ICC Workshops), 2021, pp. 1–6.

[95] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," Trans. Wireless. Comm., vol. 20, no. 3, p. 1935–1949, mar 2021.

[96] G. Wang, F. Xu, H. Zhang, and C. Zhao, "Joint resource management for mobility supported federated learn- ing in internet of vehicles," Future Generation Computer Systems, vol. 129, pp. 199–211, 2022.

[97] H. T. Nguyen, N. Cong Luong, J. Zhao, C. Yuen, and D. Niyato, "Resource allocation in mobility-aware federated learning networks: A deep reinforcement learning approach," in 2020 IEEE 6th World Forum on Internet of Things (WF-IoT), 2020, pp. 1–6.

[98] K. Pfeiffer, M. Rapp, R. Khalili, J. Henkel, "CoCoFL: Communication- and Computation-Aware Federated Learning via Partial NN Freezing and Quantization", in Transaction on Machine Learning Research (TMLR), 2023.

[99] T.-M. H. Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification", arXiv:1909.06335, 2019.

[100] S. Horvath, S. Laskaridis, M. Almeida, I. Leontiadis, S. Venieris, and N. Lane, “Fjord: Fair and accurate federated learning under heterogeneous targets with ordered dropout”, in Advances in Neural Information Processing Systems, volume 34. NeurIPS, 2021.

[101] E. Palacios-Morocho, S. Inca and J. F. Monserrat, "Enhancing Cooperative Multi-Agent Systems With Self-Advice and Near-Neighbor Priority Collision Control," in IEEE Transactions on Intelligent Vehicles, vol. 9, no. 1, pp. 2864-2877, Jan. 2024, doi: 10.1109/TIV.2023.3293198

[102] E. Palacios-Morocho, P. López-Muñóz, M. A. Costán and J. F. Monserrat, “Data Homogenization Method for Heterogeneous Sensors Applied to Reinforcement Learning,” in IEEE Access, vol. 11, pp. 77347-77358, 2023, doi: 10.1109/ACCESS.2023.3298602

[103] S. Majumdar, S. Schwarzmann, R. Trivisonno and G. Carle, "Toward Massive Distribution of Intelligence for 6G Network Management Using Double Deep Q-Networks," in IEEE Transactions on Network and Service Management, vol. 21, no. 2, pp. 2077-2094, April 2024, doi: 10.1109/TNSM.2023.3333875.

[104] Majumdar, S., Schwarzmann, S., Trivisonno, R., & Carle, G. (2023). Distributed Intelligence for Dynamic Task Migration in the 6G User Plane using Deep Reinforcement Learning. Authorea Preprints.

[105] Addad, R. A., Dutra, D. L. C., Taleb, T., & Flinck, H. (2021). Toward using reinforcement learning for trigger selection in network slice mobility. IEEE Journal on Selected Areas in Communications, 39(7), 2241-2253.

# 8. Intelligent User Plane, In-Network Computing

## 8.1 User Plane enhancements for the next generation network

Leveraging on Control Plane/User Plane separation and the possibility of controlling distributed User Plane (UP) functions by a centralized control plane function, the 5G system provides fundamental enablers to support services that require distributed connectivity. In the 3GPP 5G system Rel-15 [1], a user equipment can have:

- Simultaneous UP sessions with a central UP anchoring point and local UP anchoring point dedicated to specific services.
- A single UP session that allows offloading locally part of the UP traffic, e.g. the traffic related to application servers deployed in an edge hosting environment, while the rest of the traffic is forwarded to a central UP anchoring point.

3GPP 5G Rel-17 further enhanced the support of Edge Computing deployments by introducing DNS functionalities as well as specific edge computing fields in the traffic influence API to support the selection of the appropriated traffic offload point based on the location of the user equipment. Moreover Rel-17 introduces 5G core network features to support seamless Edge Application Server relocation [2].

The trend towards supporting distributed network connectivity will be further developed as 5G evolves and, with more radical changes in the User Plane, in the 6G network. One6G WI 207 analysed the use cases issued by the social development towards the 2030s to identify the requirements that may be considered for designing the next generation User Plane architecture.

## 8.2 Social development towards the 2030s

The on-going discussion on 5G Evolution and 6G identified some trends of social development towards the 2030s, which may result in new use cases for the mobile communication networks in the 6G era. Great importance is given to the development of “human augmentation” services [3]: the technology evolution of wearable devices (including personal sensors and actuators, tactile devices, audio/video devices, etc.) leads to developing a new generation of wireless connected devices, enabling services to support and enhance the human abilities (physical strength, perception, cognition, presence). The requirements associated to human augmentation services are analysed in Section 8.3.

## 8.3 Reference use cases

Human augmentation services can be typically described as real-time sensory services involving multisensory communication for e-health, sport training or generic personal assistance for “well-being”. Additionally, also shared virtual experiences, such as engaging in virtual collaboration in the cyberspace, may be part of this category of use cases.

Scenarios of real-time sensor services are described by IEEE 1918.1 “Tactile Internet WG” [4] [5] with reference to Tele-operation services, Immersive Virtual Reality (IVR) services and Haptic Interpersonal Communication (HIC). Moreover, the ongoing 3GPP SA1 Rel-18 study [6] on tactile and

multi-modality communication services (TR 22.847) describes uses cases for Immersive Multi-modal Virtual Reality and Immersive VR games.

The remainder of this section describes some exemplary use cases and identifies the communication requirements.

### 8.3.1 Remote haptic operation

This use case is related to haptic tele-operation in a dynamic environment. In [5], "tele-operation allows human users to immerse into a distant or inaccessible environment to perform complex tasks". Reference applications can be tele-examination, tele-rehabilitation, and possibly tele-surgery.

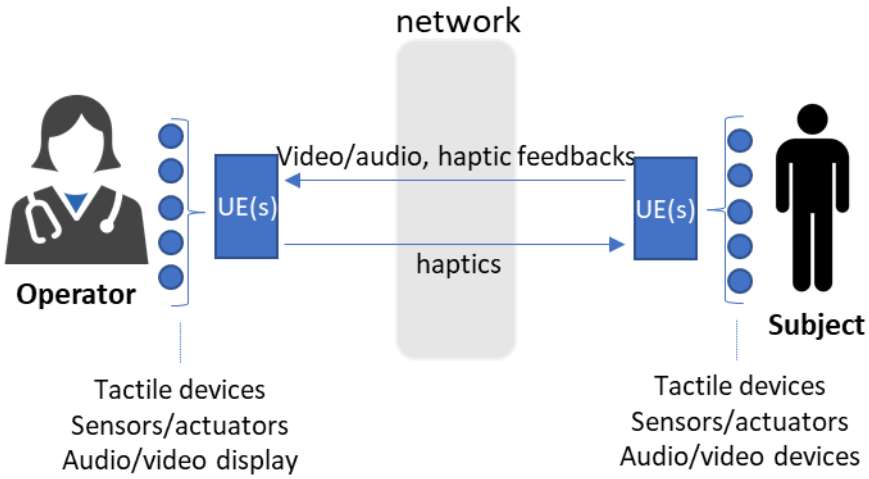



*Figure 37: Remote haptic operation*

Figure 37 illustrates the use case from the communication point of view. A human operator performs haptic interaction with a remote human subject by controlling remote tactile devices, sensors and actuators. The operator receives the haptic feedbacks from the subject, as well as synchronized audio/video streams.

Table 12 recalls the communication requirements according to [5]. They can be summarized as: device to device communication with latency <10ms and five-9 reliability. Scheduled (periodic) communication is required, as well as synchronization of the different traffic types in order to avoid simulator sickness and motion to photon delay.

*Table 12 - Communication requirements for remote haptic operation use case*

| Traffic direction | Traffic types | Burst size | Reliability | Latency (ms) | Avg data rate |
|---|---|---|---|---|---|
| Operator → Subject | Haptics | 2-8 B per DoF* | 99.999% | 1-10 (high dynamic environment)<br>10-100 (dynamic environment) | 1-4k pkt/s (P) (w/o compression)<br>100-500 (w/ compression) |
| Subject → Operator | Video | 1.5 kB | 99.999% | 10-20 | 1-100 Mbps |
| | Audio | 50 B | 99.9% | 10-20 | 5-512 kbps |
| | Haptic feedback | 2-8 B per DoF | 99.999% | 1-10 | 1-4k pkt/s (P) (w/o compression)<br>100-500 (w/ compression) |

**DoF: Degree of Freedom (i.e., the number of joints in the human body. ) (P): Periodic*

### 8.3.2 Immersive Virtual Reality (IVR)

IVR refers to "the case of a human interacting with virtual entities in a remote environment such that the perception of interaction with a real physical world is achieved" [6]. Reference applications can be "VR Video, VR Gaming, education, health care and skill transfer such as training drivers, pilot and surgeon" [5].

Figure 38 illustrates the use case from the communication point of view. A human subject interacts with a remote IVR System by using tactile devices, sensors/actuators controlled by the IVR System and audio/video devices that reproduce synchronized audio/video streams transmitted by the IVR System. The IVR System receives haptic feedbacks from the devices used by the human subject.

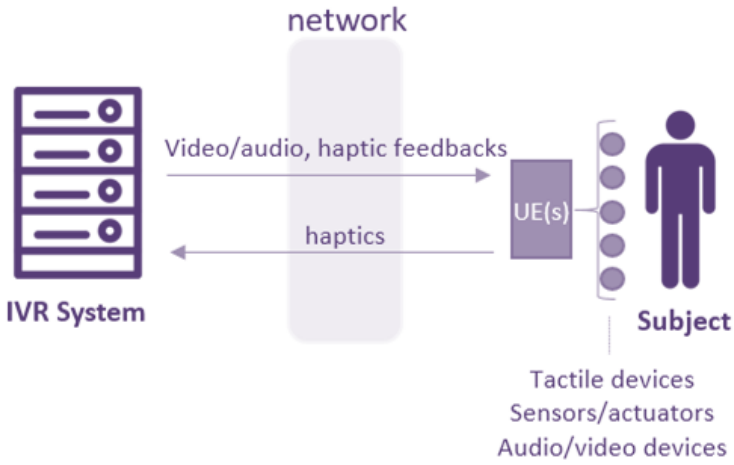



*Figure 38: Immersive Virtual Reality (IVR)*

The communication requirements according to [5] and [6] are reported in Table 13 and they can be summarized as: device to server communication with latency <5ms and five-9 reliability.

*Table 13 - Communication requirements for IVR*

| Traffic direction | Traffic types | Burst size | Reliability | Latency (ms) | Avg data rate |
|---|---|---|---|---|---|
| Subject → IVR system | Haptic feedback | 2-8 B per DoF | 99.9% (w/o comp.)<br>99.999% (w/ comp.) | <5 | 1-4k pkt/s (P) (w/o compression)<br>100-500 pkt/s (w/ compression) |
| | Sensing data (e.g. positioning) | | 99.99% | <5 | <1 Mbit/s |
| IVR system → Subject | Video | 1.5 kB | 99.9% | <10 | 1-100 Mbps |
| | Audio | 50 B | 99.9% | <10 | 5-512 kbps |
| | Haptic feedback | 2-8 B per DoF | 99.9% (w/o comp.)<br>99.999% (w/ comp.) | 1-50 | 1-4k pkt/s (P) (w/o compression)<br>100-500 (w/ compression) |

### 8.3.3 Haptic Interpersonal Communication (HIC)

HIC "aims to facilitate mediated touch (kinesthetics and/or tactile cues) over a computer network to feel the presence of a remote user and to perform social interactions" [5]. Reference applications can be social networking, gaming and entertainment, education, training, health care. The immersive VF game in [6] is an example of HIC applications.

Figure 39 (which is an elaboration of Fig. 4 in [5]) illustrates the communication scenario. Two users A and B perform haptic interaction with a model of the other user. Additionally, the models reproduce audio/video streams received from the respective user. The "models can be either a physical entity (e.g., social robot) or a virtual representation (e.g., virtual reality avatar)" [5] and they may be physical/virtual embodiments of digital twins updated in real-time.

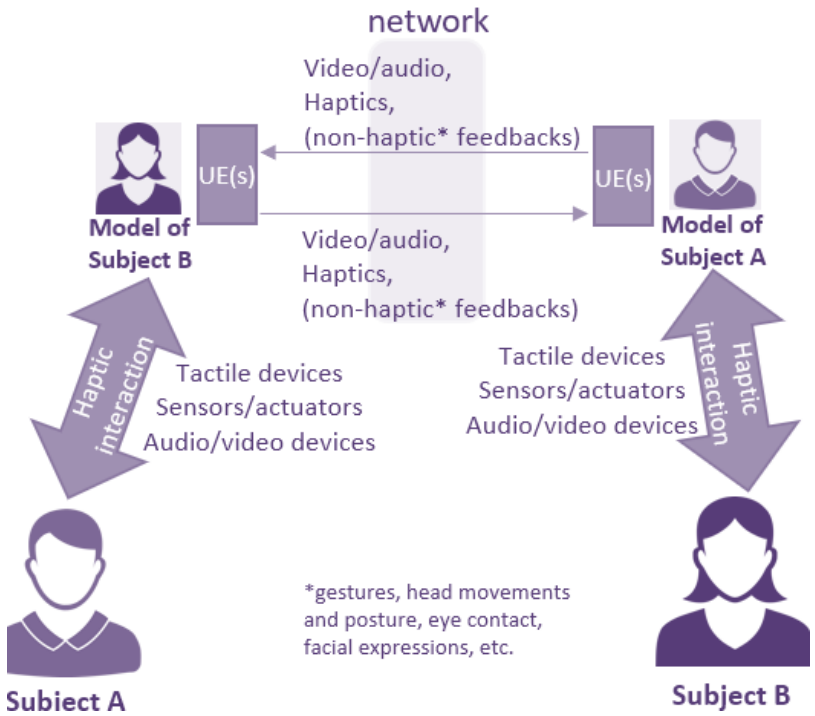



*Figure 39: Haptic Interpersonal Communication (HIC)*

The communication requirements are reported in Table 14 and can be summarized as: device to device communication with latency <10 ms and five-9 reliability.

*Table 14 - Communication requirements for HIC*

| Traffic direction | Traffic types | Burst size | Reliability | Latency (ms) | Avg data rate |
|---|---|---|---|---|---|
| Subject A → Subject B | Video | 1.5 kB | 99.999% | 10-20 | 1-100 Mbps |
| | Audio | 50 B | 99.9% | 10-20 | 5-512 kbps |
| | Haptics | 2-8 B per DoF | 99.999% | 1-10 (for interaction) | 1-4k pkt/s (P) (w/o compression) 100-500 (w/ compression) |

### 8.3.4 AI-based customer services

Interactive customer service in unmanned shops, where robotics remotely controlled by AI deal with customers, may arise as a relevant use case. This use case can be considered an extension of the human-robot coexistence currently studied for industrial applications, with a substantial difference: the target environment is not limited to the factory floor (controlled environment, localized in industrial premises), since potentially any public shop may adopt such type of service.

This use case can be modelled as a mobile robot use case (3GPP TS 22.104 section A.2.2.3) including cooperative motion control and real-time streaming (video/audio) from the robot (3GPP TS 22.104 section A.2.2.3 use cases 1 and 4) [7]. Additionally, as shown in Figure 40, the communication involves traffic of haptic and non-haptic feedbacks from the robot to the controller, as well as audio/video streams to the robot.

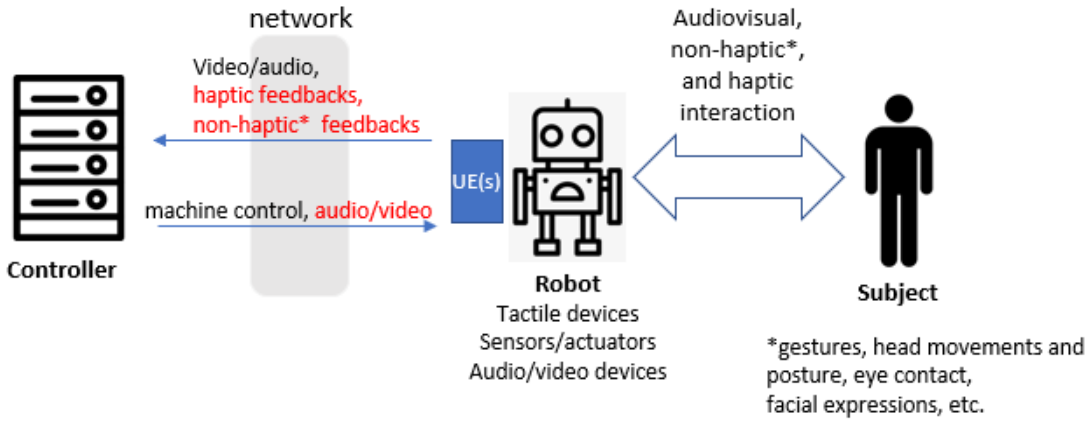



*Figure 40: AI-based customer services*

In 3GPP TS 22.104 section A.2.2.3 use case 1, periodic communication for the support of machine control requires latency targets between 1 ms and 10 ms.

### 8.3.5 Summary of the requirements

The requirements derived from the real time sensory services described so far can be summarized as:

- Low latency (5-10 ms) and high reliability (99.999%) “everywhere”, applicable both to device-to-server and device-to-device communication
- Support (periodic) time-sensitive communication for consumer applications everywhere
- Wireless network communication with high capacity to carry synchronized audio/video streams.

## 8.4 SotA and discussion topics

The use cases and requirements discussed in the previous section bring a re-thinking of the user plane architecture. Some directions of investigation are described in the following subsections:

- Flat network topology
- Core network transmission control supporting low latency

- Wide area synchronization and deterministic communication
- In-network computing
- Intelligent scaling and placements in the UP
- Leveraging ML in the Data Plane

### 8.4.1 Flat network topology

The 5G system architecture supports features to enable low latency communication. Nevertheless, these features have been designed with reference to use cases that target hot spot locations, such as the factory floor for industrial applications or an event location for Audio/Video production use cases. New tactile Internet applications issue tight latency requirements that need to be supported in wide areas, possibly "everywhere", i.e. in any house, enterprise premise, shop, hospital, etc.

Specifically:

- Tactile Internet applications like "Immersive Virtual Reality" or "AI-based customer services", which involve application servers implemented in edge/fog compute nodes, require "low latency everywhere" for the communication between the end device and the application server.
- Tactile Internet applications like "Remote haptic operation" or "Haptic Interpersonal Communication", which involve device-to-device communication implemented in edge/fog compute nodes, require "low latency everywhere" for the communication between two or many end user devices.

In 5G, like in 4G, the 3GPP network topology is an overlay on top of the transport network. In the 5G network topology, the User Plane data may need to traverse via a far-end 3GPP user plane function even if two UEs are in adjacent gNBs and the tranport network supports direct connectivity between those gNBs. Tree and star topologies will still be used in the future public networks. However, it would be necessary to consider new topology options to enable shortest path communication in a wide area as required by the above-mentioned applications. Figure 41 points out flatter, e.g., mesh, topologies as future directions. Flat topologies seem more suited to tackle "low latency everywhere" requirements.

With reference to the state of the art of the 3GPP 5G RAN and Core Network architecture, there are two complementary research directions to enable flatter topologies for the next generation mobile network:

- In the RAN architecture: remove the limitations set by using SCTP-based communication [8] between access nodes, which requires preconfigured persistent associations between node pairs. This would allow extending the direct connectivity between node pairs, currently limited to neighbouring nodes, to large areas. The target is to enable full mesh of access nodes in a wide area.
- In the Core Network architecture: study enablers for shortest path communication. This topic is further elaborated in the following section.

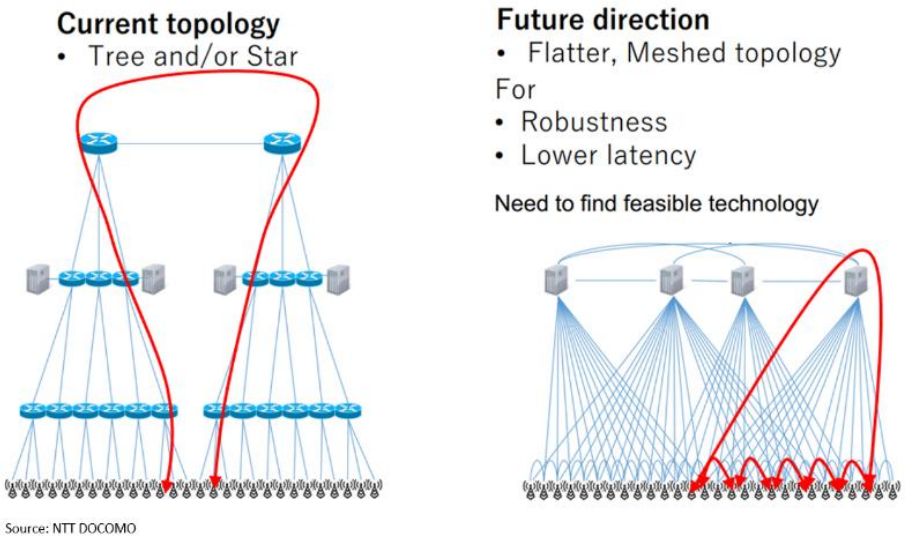



*Figure 41: Current network topology and future direction towards flatter topology*

## 8.4.2 Core network transmission control supporting LLC

5G has never attempted to reduce the latency in consideration of any (i) transmission paths actually installed and (ii) actual switching equipment in the transport network. Figure 42 shows examples of end-to-end latency targets achievable for device-to-device communication with reference to an extremely simplified IP/MPLS [9] transport network topology.

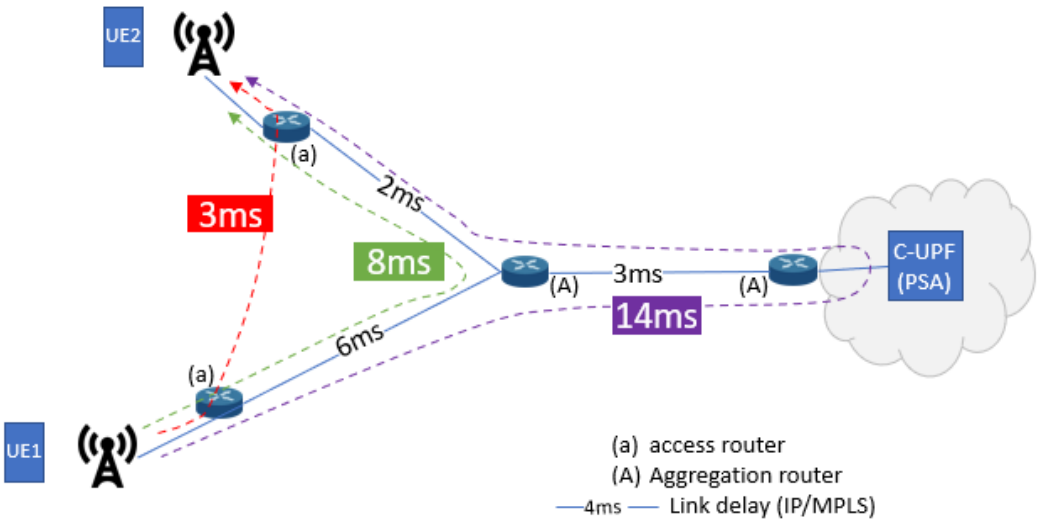



*Figure 42: Achievable latency targets in 5GC and beyond*

In the 5G system up to Rel-17, the latency achievable for device-to-device communication depends on the placement of the 5GC User Plane Function (UPF) where the UP sessions of the end user devices are anchored. If the anchoring UPF is located in the central office of the mobile network, the achievable latency for the end-to-end communication (purple path in the figure) is in the order of tens of ms. By placing the anchoring UPF at the aggregation router near the UEs, the 5G system allows to achieve latencies lower than 10ms (green path in the figure). Nevertheless, placing local UPF in the aggregation sites may require significant infrastructure investments. Similar latency targets may be achieved with lower expenditures if the traffic forwarding could be performed by

the aggregation router without involving co-located 3GPP UPF functionalities in the aggregation sites.

This scenario could be implemented by delegating the path selection to the transport layer once the 3GPP UPF has set appropriate quality target for the user plane traffic of the application session.

Finally, even lower latency could be targeted if traffic forwarding (red path in the figure) could be performed directly between the access routers co-located with the User Plane components of the radio access nodes. This scenario is based on deploying a full mesh transport network topology among the access routers. In this scenario the contribution of the core network to the latency budget would be lower than 1ms. The exemplary latency of 3ms indicated in the figure is based on assuming that each radio access node introduces a latency of 1ms, but a lower latency target may be expected for the next generation radio access.

### 8.4.3 Wide area synchronization and deterministic communication

The 5G System supports time sensitive communication among devices in a local network. Basic support of IP-based time synchronization was introduced by the 3GPP 5G core Rel-17, but accuracy for synchronization of widely distributed devices may still be an issue.

The use cases targeted by the 3GPP 5G System specifications [7] [10] to determine time sensitive communication are:

1. Industry 4.0, connected factory
    - Time-sensitive networking is crucial due to the critical nature of manufacturing processes.
    - Today's applications use Industrial Ethernet, that will evolve to IEEE TSN [11] in the future.
    - These use cases set the requirements for IEEE TSN in 3GPP Rel-16 [12] and are further improved in in Rel-17 [13].
2. Professional audio/video production
    - AV (Audio/Video) production applications require a high degree of reliability, since they are related to the capturing and transmission of data at the beginning of a production chain. Time synchronization is crucial since the target may receive streams from multiple sources (e.g. microphones, cameras, etc) and the streams need to be synchronized.
    - The support for these use cases is provided by Industrial IoT (IIoT) in 3GPP Rel-17 [13].

The 5G System does not support (up to Rel-17) wide range deterministic communication and IP-based deterministic networking. Supporting these capabilities is important to enable the creation of new services involving traffic scheduling and synchronization in a wide area. Some enhancements are under evaluation for 5G evolution (starting from Rel-18) but more aggressive changes to the user plane architecture may be required for the next generation network.

Docomo's 6G White Paper [3] suggests to selectively use multiple transmission paths specialized for the data transmission of different traffic characteristics. For instance, using dedicated transport paths for synchronization signalling, while time sensitive traffic is forwarded through dedicated transport paths with specific traffic scheduling capabilities controlled by the 3GPP core network.

### 8.4.4 In-network computing: Areas of application, prospects, and challenges

In-network computing (INC) describes the paradigm of delegating application-layer processing functions to the data plane [14]. The concept is not a novelty with the advent of softwarized networks,

but has already been proposed decades ago as active networks [15] [16]. Despite the promising solutions related to them, active networks did not achieve great success, mainly due to the limited processing capabilities of chips back then.

To date, with the availability of high-performance programmable ASIC chips, which leverage processing of up to 10 billion packets per second [14] in a network device, and with the advancements of network programming languages, in-network computing evolved as an important research field. It allows to process the traffic while it is transmitted, which reduces the resource- and energy-consumption of devices located at end hosts.

According to [17], the key potentials in-network computing can bring are: (1) a significant reduction of latency and an increase of throughput for certain operations, which is especially of interest in performance-oriented contexts; and (2) a reduction of the network load. However, the use-cases in which in-network computing can sensibly be exploited is limited by a number of requirements: (1) A significant reduction of the network traffic load should be achievable; (2) no significant changes on application-level should be required; and (3) the correctness of the overall computation must be obtained. Several approaches have been presented in literature, among others to support AI, as discussed earlier in Section 7. A brief overview of different areas of application is given in the following. Afterwards, we highlight some key challenges of INC.

#### 8.4.4.1 Areas of application

##### In-Network DNS

P4DNS [18] presents an in-network solution to reduce the latency from a DNS response. A parser inspects all incoming packets and extracts their headers up to the DNS header. If the packet is a DNS query and the respective answer is available in the DNS cache table, the network device actively responds to the query. This is done by simply modifying the packet in the sense that source and destination address are swapped and the DNS response header is appended. The authors evaluate their proof of concept implementation and show that P4DNS is capable of drastically reducing the latency and increasing the throughput.

##### Data caching

Another approach for in-network computing focuses caching and is presented as NetCache [19]. It proposes a novel key-value store architecture, making use of programmable network devices to cache data in the network for dynamic load balancing. The designed packet processing pipeline is capable of detecting, indexing, storing, and serving popular content in the data plane. The prototype implementation using Barefoot Tofino switches and commodity servers shows a 3-10x throughput improvement and significant latency improvement.

##### Traffic aggregation

The in-network computing concept is exploited in the scope of DAIET [17], a proof of concept system for data aggregation, which is a common task in several distributed data centre applications. The authors study a use-case, where aggregation functions should be used within the network for performance improvement. More specifically, their work considers a machine learning approach, where hand-written digits are identified on several machines in a distributed manner. The distributed approach leads to the fact that information exchanged between the machines can overlap and high overlaps mean that an aggregation of updates could significantly reduce the network load. To detect and aggregate duplicates, the network devices are equipped with the following information: (1) an aggregation tree ID, (2) the associated output port to forward the traffic to the next node in the tree, and (3) the specific aggregation function that should be performed. By means of preliminary evaluations, the authors show that a load reduction of 80% could be achieved compared to any of the considered baselines. A similar approach, also considering in-network data aggregation logics, is presented in SwitchAgg [20]. The proposed architecture consists of a payload

analyser, multiple processing elements for traffic forwarding and aggregation at line rate, and a two-level memory hierarchy. By means of a prototype implementation, the authors show that in-network traffic aggregation tasks can be performed at line rate and that the completion time of a simple word count job can be reduced by about 44% with the support of SwitchAgg, while the CPU utilization is reduced by 50%.

### Energy efficiency

In-network computing is exploited with the goal to increase energy efficiency in 6G networks [14]. While prior works aim at placing computational tasks into existing networking hardware, the authors propose a general computing platform which takes two tasks at the same time: Performing general computations and acting as a network node. Compared to traditional network devices, the proposed unified operating platform allows for running also more complex application tasks through hypervisors and containers. A network controller is scheduling the tasks which are executed on the network node as the traffic passes through.

### Out-of-control sensor signal detection

The usage of INC for intelligent process monitoring and for detecting out-of-control sensor signals is discussed in [21]. The authors consider an industrial robot, connected via an intelligent switch to a control and monitoring server. The presented approach uses INC to distinguish between the three process phases of a fine-blanking system. For instance, while traversing the network node, the sensor signals are clustered using an ML model, allowing to detect whether the process is in a ramp-up phase, or running in control, or if the process is running out-of-control. Less critical phases (i.e. in-control), do not require to analyse each and every sensor signal with high frequency. Accordingly, their INC-capable switch discards or aggregates those packets, which contain non-critical sensor information, leading to a reduced overall traffic volume. As soon as anomalies are detected in the switch data plane, all traffic is again forwarded to the control server for further analysis.

### Low latency industrial robot control

Another work leveraging the INC concept in industrial scenarios is presented in [22]. The authors consider a setup, where a robot is interacting with a human. Thus, any type of critical situation, e.g. collisions potentially resulting in human injuries, has to be avoided. The robot is connected to a controller via a P4 switch, which is capable of detecting position threshold violations of the robot. To achieve this, the TCP communication between robot and controller is parsed and analysed. In case of critical robot positions, the INC-enabled switch autonomously sends an emergency stop signal back to the robot. By short-cutting, i.e. the switch sending the stop signal instead of the controller, further critical movement of the robot can be avoided, thus enhancing the safety of the environment.

#### 8.4.4.2 Key challenges

While the INC concept has been shown to bring significant benefits to many different domains, we also want to note that it still faces a couple of unresolved challenges, as briefly discussed based on [23] and further prior art.

### Traffic encryption

Nowadays, most Internet traffic is encrypted. This makes INC impracticable outside of vertical networks, where the external network is a non-trusted party. A possible solution to the problem can be Homomorphic Encryption (HE) [24], a technique allowing to compute on encrypted traffic. This would require that: (i) The source encrypts the flow's data according to a specific HE scheme; (ii) the network element holds a transformed version of the compute task which conforms to the applied HE scheme (to enable the computation on the encrypted traffic), and (iii) the destination decrypts

the flow's received processed data according to the specific HE scheme applied. To enable this, the application (which is sending and receiving the data) and the network element carrying out the computation must be aligned in terms of the used HE scheme and the specific parameter setting (e.g., the used polynomial modulus). While HE would allow to leverage INC despite traffic encryption, one obstacle still remains: it is too slow for practical usage. Definitely, extensive research and development will be necessary to make HE efficient enough, e.g., by means of advanced algorithms and hardware acceleration [25] [26] or in later stages by utilizing quantum computing solutions.

### Inter-operability with existing protocols

The usage of unreliable protocols, such as UDP, is limited for INC, as dropped packets could contain relevant computations. On the other hand, using TCP, which adapts its sending rate according to the RTT, can lead to problems due to the RTT increase resulting from additional computation time. In addition, it needs to be clarified which entity (the source node versus an intermediate INC node) would be responsible for re-transmissions in case of packet loss with reliable protocols such as TCP. Definitely, further studies are needed to examine in how far existing protocols can support INC, but it is expected that dedicated protocols will be necessary [27].

Furthermore, segment routing [28] is another technique that could be suitable in this context. Using this technique, an end-host is able to include an ordered list of instructions in the packet headers. The intelligent user plane could then use this information to forward the traffic to the appropriate location.

### Trust issues and ensuring the correctness of the computation

In order for the network to carry out the computation as intended, user and application provider trust are required [29]. It further needs to be ensured that the conducted computations are correct. Thereby, verifiable computing [30] will also be a concept to be exploited.

### Interoperability with QoS

The UP entities in 5G are responsible for fulfilling the QoS requirements of a flow while routing its traffic through the network. That is, e.g., to ensure a certain guaranteed bitrate or to keep a specific delay budget. When introducing INC to the 6G UP, it needs to be clarified how to ensure both requirements at the same time, i.e., the QoS of the flow, as well as the correct computation. Fulfilling critical QoS requirements will be more challenging, due to the additional latency for computing on the flow's packets.

### Disruption of the end-to-end principle

The end-to-end principle generally states that information pushed on the sending side of the connection should be received without modification at the receiver side. Intermediary nodes should only connect the two explicitly addressed end points of the communication. This pure end-to-end communication might no longer be suitable with INC [29]. Besides the payload modification with INC, several proposals [18] [22] allow the network entity to even actively respond, although it is not directly addressed by the sender. The concept of having one dedicated sender and one dedicated receiver, as well as existing transport layer solutions, will need to be reconsidered [31].

### UP path setup

The complexity of determining an appropriate UP path is increased when additional constraints – relating to the computations – come into play. That is, besides factors such as the location and link properties, we need to consider the computational resources and supported tasks of the respective UP nodes. Solving the problem of an optimal embedding is not straightforward.

### Dynamic and optimized computation allocation

Sophisticated mechanisms are required to determine an optimized allocation of the compute tasks among the involved entities (UE, MEC server, UP entities), given a current set of dynamic conditions. In [32] a multi-criteria edge-computing-enabled live service migration procedure is optimized considering different types of migration costs and benefits.

### User mobility

An important aspect is to ensure that the computation is ''moved'' with the user. Otherwise, the computation results might no longer be meaningful with the user in a new connectivity context. Depending on the magnitude of movement, a couple of entities need to be re-selected, e.g., Access Node (AN), UPFs, Access and Mobility Function (AMF), or Session Management Function (SMF). This entails aspects associated with computational and user context transfer, that need to be present during the associated handover processes. This mobility also requires that INC computation functions must be dynamically deployed across the data path nodes which can pose a challenge since most programmable devices require a reset to change their source code.

### Memory capacity

Typically, programmable networking devices can process data at very high rates (Tbps) and thus, the memory available in these devices must be suitable for such speeds. This means that these devices have a very fast memory but in limited amounts. Furthermore, these devices run on a time budget principle, limiting the memory access time window as well as the number of memory accesses allowed per stage [33]. Due to these limitations, INC is a strong candidate to run stateless functions.

### Dependability

Finally, the proposed communication and compute system gives rise to challenges related to dependability. In contrast to current mechanisms that typically only consider static configurations, more sophisticated approaches will be necessary. Particularly, operational states will need to be taken into consideration when dealing with failures in order to maintain correct operation.

## 8.4.5 Intelligent Placement and Scaling in the UP

Several efforts have been made towards an intelligent placement of computational resources or content as well as towards an efficient scaling of UPF instances so to satisfy QoS requirements whilst being cost-efficient. Although the prior works do not consider the intelligence being placed directly within the user plane, these mechanisms can give a direction towards an intelligence user plane design.

### 8.4.5.1 Dynamic Scaling

The work in [34] aims at dynamically scaling active UPF instances according to the number of PDU sessions. The proposed approach assumes that CP NFs (e.g. AMF or SMF) provide information on the current state of the UPFs. That is, the number of active and booting UPFs, the number of currently active PDU sessions, as well as an approximation of the PDU session arrival rate. A reinforcement learning agent decides based on the current state information if up-, or down-scaling or no action is needed. It actively learns the implications of its actions and is thus capable of optimizing its decisions. The ultimate goal is to run as few UPF instances as possible, whilst being capable of satisfying all PDU sessions QoS requirements.

#### 8.4.5.2 Optimized Placements

Not an optimized scaling, but an optimized placement of UPFs (and MEC servers), is the target of several prior works [35] [36] [37]. More specifically, [35] formulates the problem for a joint placement of both UPF instances as well as MEC servers, within 6G networks so to minimize latency. The NP-hard problem is simplified by means of studying the placement relationship between UPF and edge servers. The proposed algorithm outperforms benchmarks on a real-world data set and on edge network emulations. Similar to that, [36] addresses the placement of UPFs and MEC servers, but with the constraint that UPFs can only be initiated at MEC servers, thus relaxing the complexity of the problem compared to [35]. The goal is to minimize the operational service costs for providing enough resources and a sufficiently low latency. The proposed framework for joint placement of edge nodes and UPFs relies on integer linear programming and heuristic solutions to optimize the trade-off between costs and latency reduction gains. The follow-up work presented in [37] additionally considers dynamicity when deciding about the placement of the instances, i.e., the proposed approach allows to adapt to changes in user locations while ensuring QoS. By means of a scheduling technique relying on Optimal Stopping Theory (OST), the UPF placement is dynamically orchestrated depending on observed latency violations. To avoid too frequent re-configurations of the UPF placement, the authors also study the best re-computation time.

### 8.4.6 Leveraging ML in the Data Plane

Machine Learning offers a huge and diverse set of potential use-cases for being used in the context of communications networks. The most prominent ones include traffic classification, load forecasting, active queue management, detection of anomalies or malicious traffic, or path computation and routing or switching [38]. The outcome of the decision is often triggering (near-) real-time actions in the network, and hence, has strict requirements in terms of the speed with which a decision is made. As most current approaches for using ML in the network consider the logic being deployed in the control plane, these strict requirements cannot be met, due to the delay incurred for communicating with the control plane. As a consequence, the practical usage of ML-triggered actions in the network is limited. To overcome the issue, ML logics can directly be deployed in the data plane. Apart from meeting the stringent delay requirements, this brings several additional benefits [39]:

(1) Switches have a high performance and are capable of achieving a latency in the order of hundreds of nanoseconds per packet [40] and are thus faster than high-end ML accelerators [41].

(2) In addition to that, switches have a lower power consumption compared to most accelerators.

(3) For distributed ML use-cases, where the performance is bounded by the duration for transmitting data between notes, the fact that switches can classify with the same rate as they can carry packets to nodes, is beneficial. Potentially, they can even outperform approaches relying on/in a single node.

(4) When used outside a data centre, classification within network devices can reduce the load on the network (due to early data termination) and provide scalability over time.

(5) Given that a switch can support both, networking and ML operations, it is often the cheaper solution, as no additional hardware needs to be installed. All in all, deploying ML in the network, and specifically in the data plane, is more power efficient, can reduce the load and delay, and consequently reduces costs while enhancing the user satisfaction.

#### 8.4.6.1 Limitations and Potential Solutions

Despite the promising benefits, the deployment of ML in switches is not used in production environments and scarcely discussed by the research community. This is due to several practical obstacles, which are hard to overcome. Thereby, the most significant ones are the scarce availability of memory and the limited capability of switches in terms of performing complex mathematical operations. The following table summarizes some obstacles, consequential limitations, potential

solutions, and drawbacks of the solutions [39] [42]. In the following, three distinct solution directions from literature are presented. The first one tackles the challenges by translating high complex ML models to simple match-action pipelines. The second one follows the approach of distributing an Artificial Neural Network (ANN) over (geographically) distributed network nodes, where each node is taking care of the computations carried out by one neuron. The third one proposes a domain-specific enhancement of switch architectures to support in-network ML.

*Table 15: Obstacles of leveraging ML in the data plane*

| Obstacle | Resulting limitations for in-network ML | Potential solution | Solution drawback |
|---|---|---|---|
| Pipelined architecture of switches and finite amount of resources | Complex operations, such as polynomials or logarithms, cannot be performed | Look-up tables to store pre-computed results | Very finite size of table that can be stored makes the solution unpractical |
| Finite amount of memory | Limited size of lookup tables (potentially overcoming the obstacle of not being able to perform complex operations) | Reduced number of entries in the look-up table (e.g. by grouping similar values) | Reduced reliability / correctness of the results |
| Limited number of stages per pipeline (typically 12 to 20 per switch) | Support of only partial de-capsulations and packet processing. Number of extractable header information (features) limited by number of stages. | Packet re-circulation: Fragmentation of packet into header-sized units and iterative processing in the pipeline | Adjustments needed to maintain the metadata information, degradation of throughput, only applicable in networks with low utilization |
| | | Concatenation of multiple pipelines to increase number of stages and supported operations per packet | Reduction of throughput, metadata cannot be used anymore to share information between stages between pipelines |

### 8.4.6.2 Reducing Complex ML Models to Simple Match-Action Rules

The authors of [39] overcome the drawback of a limited set of simple operations that can be performed in networking hardware by mapping trained ML models to low complex match-action pipelines. They consider the use-case of classifying IoT device types based on the packet header information (11 features in total) given in the IoT traffics' packets. More specifically, they propose different solution to resemble the decision logic of four different state-of-the-art classifiers: Decision Tree, SVM, Naïve Bayes, and K-means. For the two examples Decision Tree and SVM, the translation to the way simpler match-action pipelines works as follows:

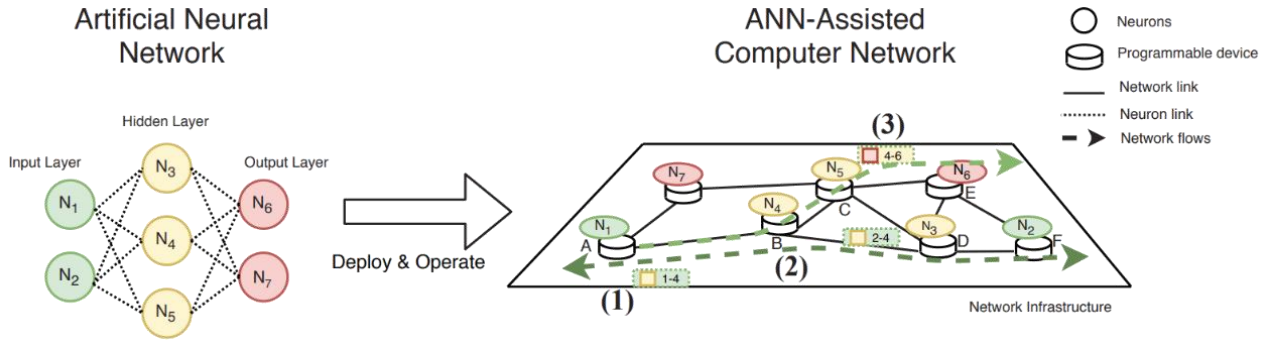
- Decision Tree: The number of stages implemented in the pipeline equals the number of used features plus one. In every stage, one feature is matched with all its potential values. The result, i.e., the branch taken, is encoded into a meta data field and at the last stage of the pipeline, the coded fields of all features are matched and the value is mapped to the resulting leaf node.
- SVM: Implementation of "m" tables, where each table is dedicated to a hyper-plane which indicates the side of a given value. The set of features is the key for the match-action table and the action is a "vote", a simple flag denoting whether the input belongs within or outside the hyper-plane. After the input has passed all "m" tables (i.e. hyper-planes) the votes are counted and the input is classified accordingly.

The proposed prototypes, both implemented in hardware and software, are capable of classifying the traffic of real-world traces at line rate. The most reliable classification with an accuracy of 0.94 could be achieved by the decision tree. Reducing the number of features, and thus the tree depth as well as the number of per-packet operations, to five, still an accuracy of 0.85 can be achieved.

### 8.4.6.3 Distributed Machine Learning

The work in [43] proposes a distributed ANN approach, where network nodes at different locations perform the computations of a single neuron of that ANN and where the network links represent the connections between the neurons. An illustration is given in Figure 43.

*Figure 43: Distributed ANN [43]*




The illustrated ANN on the left side consists of an input layer with two neurons (N_1,N_2), a hidden layer with three neurons (N_3, N_4, N_5), and an output layer consisting of two neurons (N_6, N_7). Each of the input layer neurons is connected to each of the hidden layer neurons and each of them is again connected to all of the neurons of the output layers. The ANN-assisted computer network depicted on the right-hand side shows the distribution of the ANN over the network devices. Each node in the ANN-assisted network is responsible for the tasks carried out by one of the ANN's neurons. Thus, the ANN's logic is distributed over the network's devices. Information sharing between the neurons is done by piggy-backing data on existing flows.

The authors motivate their approach with the use-case of smart network telemetry (whereas the approach is generic enough to be mapped to other use-cases such as real-time flow classification or smart traffic engineering). With network telemetry, as done today, the switches embed the telemetry data to the packets as they traverse the switch. Dedicated switches, typically those nearby the network edge, collect the telemetry embedded in the packets and send the information to the control plane. With increasing line speed and the constantly growing load on networks, this approach could however saturate or even overload the communication channel between control- and data plane. To overcome this, telemetry data could be sent intelligently to the control plane in a selective manner, e.g. only when unexpected events occur. Hence, instead of relying on hard-coded collection rules, the authors propose to rely on an approach which intelligently, via the distributed ANN, decides whether telemetry information should be shared with the CP.

The proposal of distributing an ANN's neurons is, however, coupled to a variety of challenges and a valid mapping of neurons to network nodes has several pre-requisites that need to be fulfilled. If neurons are connected in the ANN, the respective network nodes need to be connected as well. In addition to that, it needs to be guaranteed that flows with sufficient capacity are active between these nodes, to enable picky-packing of information. The paths between the neurons should be as short as possible to avoid synchronization issues and runtime timeouts. The authors formulate an optimization problem to minimize the distance between the neurons, accounting for all involved constraints.

The authors implement their approach using P4 and show that it is capable of reducing the amount of data shared between CP and DP (5 times fewer data reported) while achieving a precision of 91% in terms of correctly determining non-expected events that should be reported.

#### 8.4.6.4 Switch Architecture Enhancements

In order to deploy ML in the data plane, a certain degree of flexibility is needed in terms of the match-actions to deploy. However, the speed and flexibility are conflicting goals. While programmable off the shelf hardware can cheaply be adapted according to the current needs, they are much slower compared to dedicated hardware implementations. However, dedicated hardware, which is capable of providing high speed, is very expensive to replace if a different architecture, e.g. of the switch pipeline, is needed. The paramount goal of the work presented [44] is to meet the trade-off between programmability and speed. For instance, it is desirable on the one hand to allow as much flexibility as possible and the usage of a wide range of commands and actions. On the other hand, packets should be processed as fast as possible, which can be achieved with dedicated hardware. This conflicts with the goal of flexibility, as replacing the hardware with each new feature that should be provided, results in immersive costs. The match-action hardware proposed in the scope of Protocol Independent Switch Architecture (PISA) allows just enough re-configuration in the field, such that new rules for packet processing can be implemented and enforced at run-time. In general, it allows to re-configure the data plane in the following four ways:

1. Definition and re-definition of fields
2. Specification of number, topology, widths, and depth of match tables
3. Definition of new actions
4. Placing arbitrarily modified packets in specified queues

This proposal is hence suitable to (partially) address the limitations as denoted in Table 13. Building on top of this architecture, Taurus [42] presents a domain-specific architecture for switches (and NICs) to perform per-packet ML in the data plane at line rate. It adds a new compute block which is based on parallel-patterns abstraction, i.e., MapReduce. The new block introduced is a grid of memory units (MUs) and compute units (CUs), implementing a spatial single-instruction-multiple-data (SIMD) architecture. The control-plane of a Taurus-enabled data centre shall obtain a global view of the network and train ML models, which are then used by the data plane to perform optimized, per-packet decisions.

## 8.5 Technology Trends: Summary

Section 8.4 presents different solutions for an evolution towards an Intelligent User Plane (IUP) in 6G systems. We can aggregate the solutions in two technology trends:

- Delegating 3GPP User Plane Functionalities to the Transport Layer
- Leveraging In-network Computing in the User Plane.

This section draw conclusion on a potential architecture approach to implement these two technology trends.

### 8.5.1 Delegating 3GPP User Plane Functionalities to the Transport Layer

This technology trend aims at:

- Achieving UP latency performance gains by leveraging on shortest path communication in the transport network
- Distributing UP routing functionalities at the access nodes of the 3GPP network without incurring in the significant increase of deployment costs that would require deploying full-fledged 3GPP UPF at each access site

- Simplifying the 3GPP User Plane components to reduce their cost, possibly leveraging on general purpose IT technologies (e.g., standardized by IETF) instead of designing 3GPP-specific functionalities.

With specific reference to UE-to-UE communication, Figure 44 depicts the 5G state of the art (the communication path is implemented by two PDU sessions anchored to CN UPF(s)) and the target configuration that could be achieved in the 6G architecture (direct user plane path between the gNBs). The target configuration assumes that all the gNBs in the service area are interconnected by full mesh topology in the transport network.

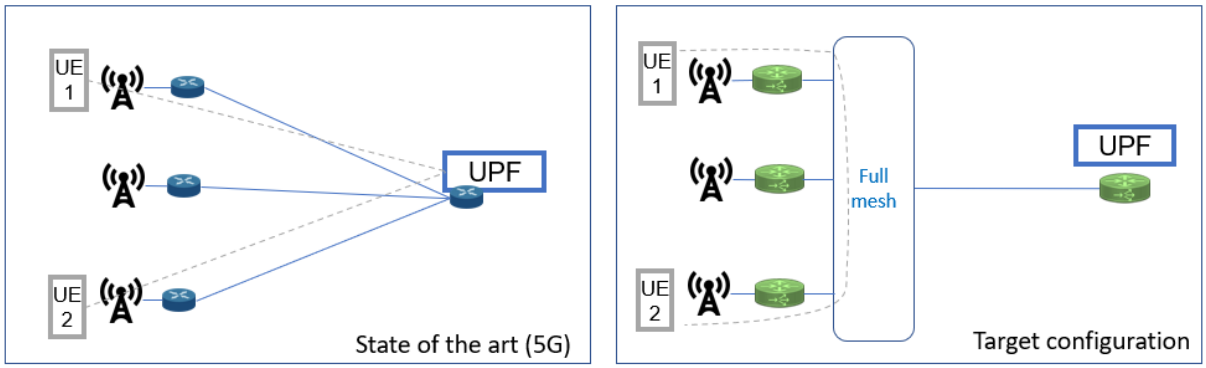



*Figure 44: UE-to-UE communication in legacy and future architecture*

In comparison with the 5G system architecture depicted in Figure 45, Figure 46 describes the architecture changes required to realize the target configuration in the 3GPP User Plane by delegating functionalities to the transport layer.

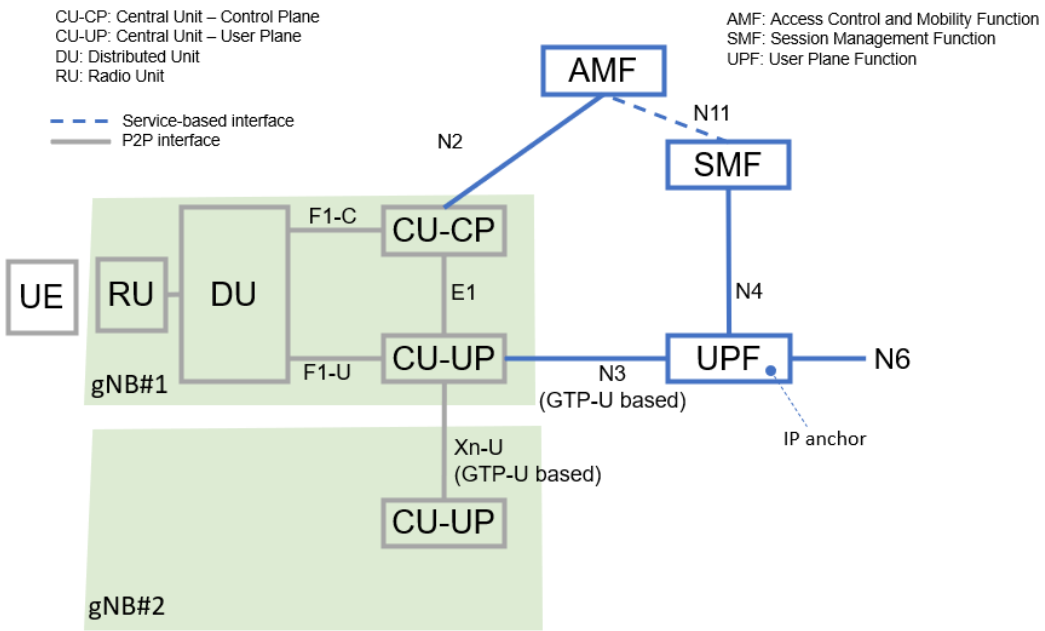



*Figure 45: 5G state of the art architecture (simplified view)*

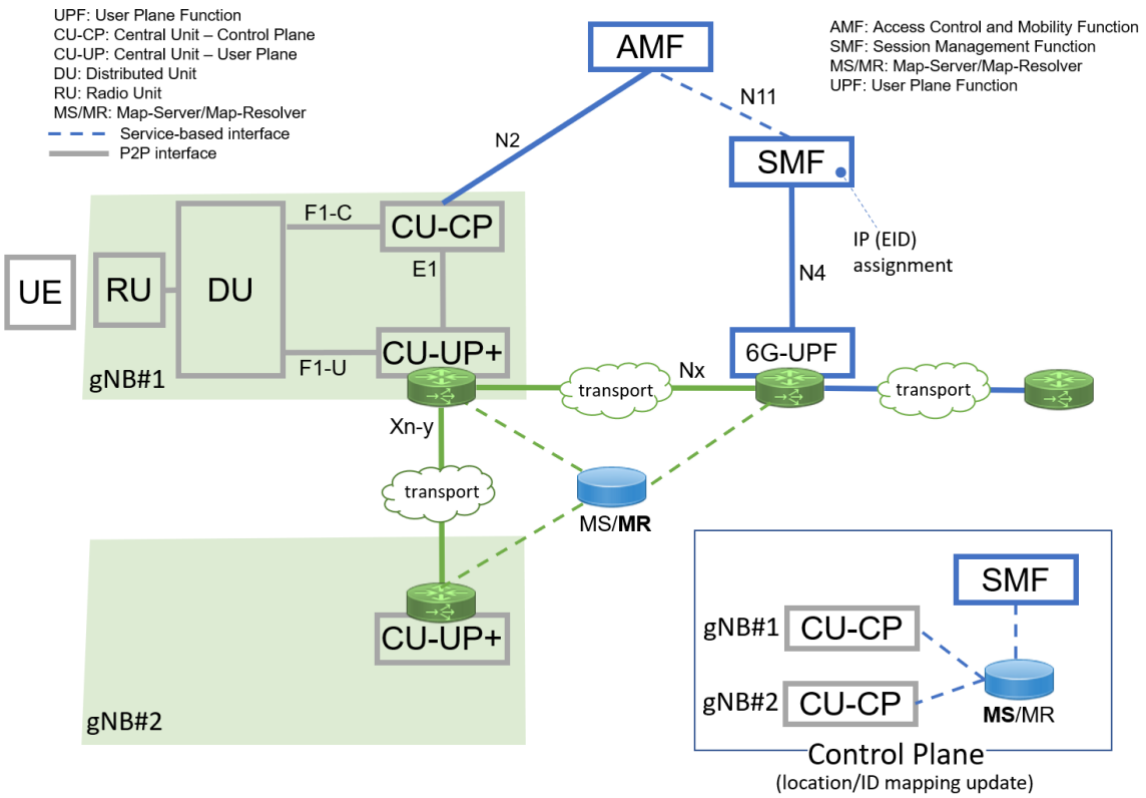



*Figure 46: Proposed 6G User Plane architecture to delegate UP functionalities to the transport network*

#### 8.5.1.1 User Plane aspects

### 1) Shortest path user plane connectivity

Shortest path user plane connectivity between gNBs is realized by leveraging on SR-enabled access routers integrated with the CU-UP+ component of the gNB. The access routers of the gNBs are connected by full mesh topology in the transport layer.

### 2) 3GPP User Plane reference points

User Plane traffic may be routed over:

- The reference point Xn-y for UE-to-EU communication within a wide area where direct user plane path between gNBs is possible.
- The reference point Nx for UE-to-UPF communication with a Data Network or UE-to-UPF-to-UE communication when establishing a direct user plane path between gNBs is not possible.

### 3) CU-UP+

The 5G CU-UP is extended with a set of functionalities implemented by the access router integrated with the 6G CU-UP+:

- Perform Locator/ID resolution with the MS/MR Control Plane function for UL packets

- Act as QoS enforcement component (see point 5).

The reference point Xn-y could be used also as evolution of the legacy Xn-U (GTP-U based, see Figure 45) for data forwarding between the gNBs.

### 4) 6G UPF

The 6G UPF implements much simplified functionalities in comparison with the 5G UPF; the functionalities are implemented by the access router integrated with the 6G UPF:

- Perform Locator/ID resolution with the MS/MR Control Plane function for DL packets; this can be implemented by IETF LISP methods as proposed in Solution #1.
- Act as QoS enforcement component (see point 5).

### 5) QoS enforcement

QoS enforcement is performed by the routers either at the CU-UP+ (for Xn-y or Nx communication) or at the 6G UPF (for Nx communication) or at the IUF (for Xn-Y communication). The QoS enforcement component:

- Enforces the uplink bitrates based on the QFI marked by the UE in the UL packets
- In DL direction, in addition to the DL bitrate enforcement, marks the packets with the corresponding QoS Flow Identifier (QFI) that are used in the target gNB for radio resource scheduling.

The same QoS enforcement component may perform the QoS-functions for both directions for a packet.

NOTE: it is for further study how the QoS enforcement component is made aware of the session and QoS policy. In particular, it is for further study how to enforce the same QoS bi-directionally for UE-to-UE communication if the UEs have different subscriptions or capabilities.

#### 8.5.1.2 Control Plane aspects

Locator/ID resolution is supported by the MS/MR (Map-Server/Map-Resolver) Control Plane function. The MS/MR handles two reference points implemented by service-based interfaces:

- The MS/MR exposes a service-based interface to the routers integrated in the gNB's CU-UP+ and CN's 6G UPF to perform Locator/ID resolution
- The MS/MR exposes a service-based interface to the Control Plane functions in CN (SMF) and RAN (CU-CP) that the CP functions invokes to store Location/ID mapping in the MS at PDU session establishment and modification.

## 8.6 Enabling the Intelligent User Plane for 6G

In the following, we describe the solution presented in [23] for integrating In-Network Computing into 6G networks by means of an Intelligent User Plane (IUP).The key idea is a programmable User Plane, capable of carrying out computations on packets as they traverse the network devices, instead of purely performing (QoS-aware) store and forwarding actions on them. By means of an AR/VR use-case, we show how applications and involved devices can benefit from offloading application tasks to the 6G network.

### 8.6.1 The Intelligent User Plane for supporting VR/AR use-cases

Along several research associations, INC is seen as a key enabler for future immersive media, such as VR and AR [45]. We also motivate the 6G IUP by means of a mobile AR gaming use-case, showing that the IUP is capable of (i) simplifying the mobile end-devices and (ii) reducing the latency as compared to, e.g., MEC-based solutions. The upper part of Figure 47 shows a 5G scenario, where several devices are involved in a game. Via the AN and the core network (CN) UPF, the devices establish a connection to the Data Network (DN), and thus to the remote application server. We assume that each device has a dedicated connection to the AN, i.e., acts as independent user equipment (UEs).

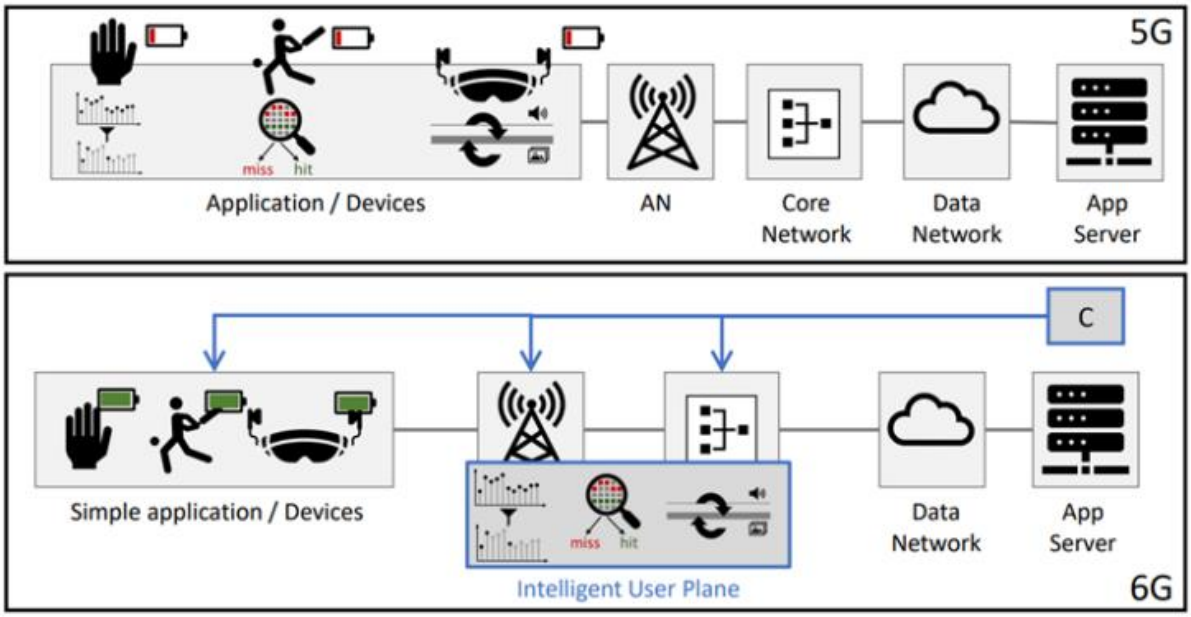



*Figure 47: Intelligent User Plane to offload application-specific tasks to the UP elements*

Each device performs different game-relevant tasks: The glove sends haptic information captured by its sensors, the racket detects if the ball was hit or missed, and the AR glasses render video objects and detect objects in the real environment. The problem with these (complex) computations running at the mobile devices is the clash with their vital design requirements and characteristics: Mobile devices should run without external power supply to allow a high degree of freedom in terms of movement to the user. To achieve a high wearing comfort, the integrated batteries should be as small as possible. Further, the equipment shall not become too warm, even after a long usage period. However, a high computational load can heat up the device and lead to fast drainage of the small batteries. Indeed, the issues of battery lifetime recently challenged the development of AR glasses and led to delays in commercial releases of end-consumer grade equipment.

The lower part of Figure 47 denotes the same scenario, but with the envisioned 6G IUP. Now, the devices can be kept simple, because the tasks are offloaded to the 6G UP entities, i.e., the AN and/or the UPF. A controlling entity (denoted as C) programs the AR/VR end-devices and the involved UP entities for INC usage. Further, the controlling entity allocates the compute tasks among UP entities. This requires a certain level of interaction between application, devices, and the 6G system. Hence, the Application Function (AF), or the respective 6G counterpart providing information exchange with third-parties, can be a well-suited candidate to be involved in the process. Please note that the controlling entity is not necessarily a single entity, but can be composed of multiple NFs and their interactions.

### 8.6.2 Benefit as compared to current MEC solutions

Other concepts, such as mobile code offloading [46], have also shown a great potential for reducing the energy consumption of mobile devices by means of migrating processor-intensive tasks to resource-rich surrogates. But such solutions are prone to exceed the stringent delay requirements of novel use-cases, such as mobile AR gaming. Figure 48 highlights the benefit for bringing computations into the network, when aiming at both simultaneously, high energy-efficiency and low application latency.

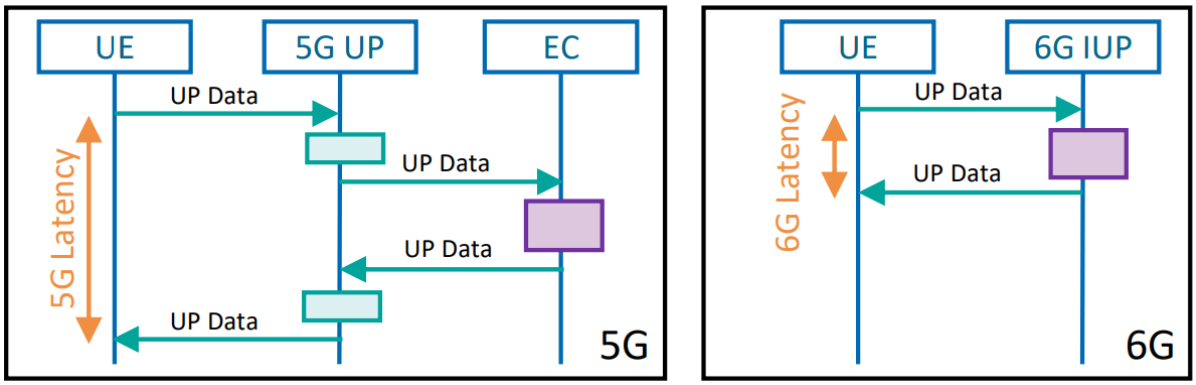



*Figure 48: Latency reduction with the IUP for 6G*

The left part of Figure 48 denotes the 5G scenario, where the UP data is sent via the 5G UP (both AN and CN) to an edge computing (EC) server. There, the data is processed and then sent back to the UE via the 5G UP.

Having the proposed IUP in 6G networks allows to process the UP data directly within the mobile 6G system. This allows to drastically reduce the latency as compared to the depicted EC solution.

### 8.6.3 Proposed key enablers

In the following, we describe four key enablers for realizing the envisioned IUP for 6G networks.

#### Key Enabler 1: Computation-enabled User Plane Entity (CUPE)

Compared to state-of-the-art UP entities (i.e. AN and UPF), dedicated to solely perform communication-related tasks, the Computation-enabled User Plane Entities (CUPEs) are significantly enhanced in terms of their computational resources. This includes among others CPU, GPU, RAM, and storage. We envision two realizations of a CUPE:

1. A general-purpose CUPE: It may run on virtualized infrastructure and can be scaled dynamically to the current needs, e.g., the number of active flows requiring computation or the complexity of the conducted compute tasks. This realization has the benefit that it can carry out any arbitrary task. However, it is potentially slow because computations are carried out in software.

2. A specific-purpose CUPE: This CUPE realization is equipped with dedicated hardware-acceleration. This brings the key benefit of being able to process packets very fast, even at line-rate. On the other hand, however, this comes with the limitation of being applicable only to very particular tasks and inducing higher costs.

### Key Enabler 2: Compute Service (CS)

With the term Compute Service, we refer to computations carried out in the CUPEs. That is, a CS unites all computational instructions and precisely describes the actions to be executed on a flow's packets. We envision two types of such compute services:

1. Pre-defined CS: Those represent generic computations, which can be useful for a wide range of applications. As they are foreseen to be frequently used, all - or at least a large set of - CUPEs would support these compute services off the shelf.

2. Customized CS: Those represent highly specific computations, e.g., well-tailored to the particular needs of a given application. While such customized CSs offer a high degree of flexibility, they cannot be supported by a large set of CUPEs per se. The deployment of a customized CS at the CUPEs in charge, i.e., those along the application flow's UP path, can be initiated via the AF. An application provider can communicate the desired CS, e.g., in the form of an execution script, to the 6G system's orchestration and management (OAM). If the request is accepted, the new CS can be deployed and the application's packets are processed in the network as specified by the application provider.

### Key Enabler 3: Communication and Compute Flow (CC Flow)

To allow computation management on a per -flow level, we introduce the concept of Communication and Compute Flows (CC Flows) in the 6G system. CC Flows may be understood as an evolution of QoS Flows. The key advancement they introduce is that – besides treating the flows in the specified QoS-aware manner – the CUPEs carry out computations on CC Flows' packets, according to the CSs associated to that flow. The packets' payload data may hence be modified along the way from sender to receiver. This is a key break with the flow concept as known today.

### Key Enabler 4: Communication and Compute Control Entity (CCCE)

As described earlier using Figure 47, a control unit is necessary to determine how the compute tasks shall be allocated among the UP entities and to program them accordingly. For that, we introduce the Communication and Compute Control Entity (CCCE). It is responsible for determining the appropriate CUPEs to use for a CC Flow, allocating the computations among the involved entities, and for any further tasks related to the CC Flow setup. The CCCE is aware of the CSs supported by the different CUPEs, as well as their current load. The control entity is not necessarily a single entity, but can be realized in a distributed manner, such that logical tasks are distributed among different entities, i.e., CP NFs.

## 8.6.4 Communication and Compute Flows

We understand the CC Flows as the key enabling concept for the 6G IUP. For this reason, we provide more details based on Figure 49.

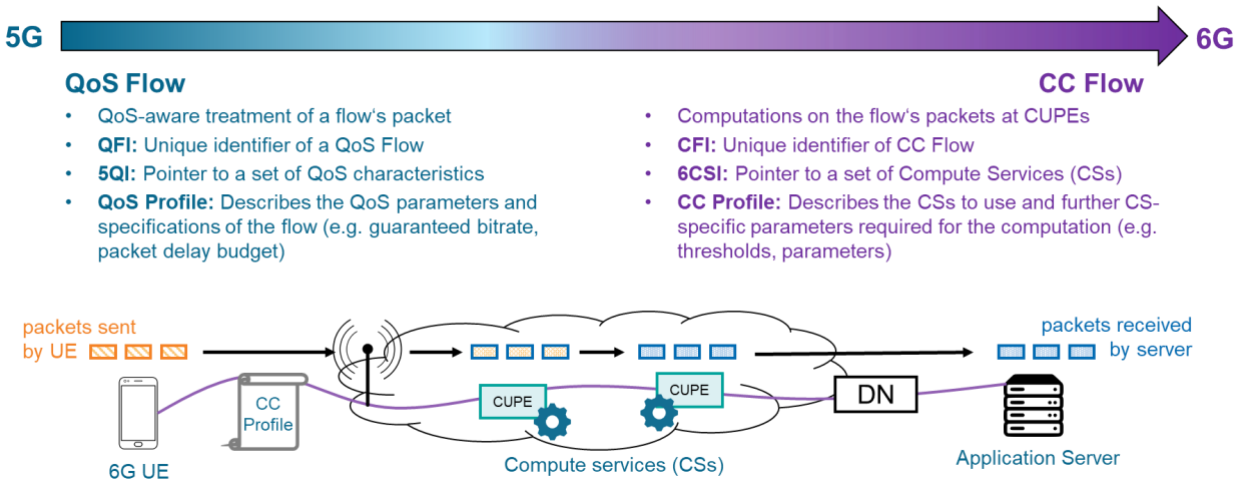



*Figure 49: The concept of CC Flows. By making use of the CUPEs and their CSs, packets are processed while being transmitted from the 6G UE to the DN via the 6G network.*

Considering the CC Flows as an evolution of QoS Flows, we map and extend existing concepts from QoS Flows to CC Flows. For example, the QoS Profile in 5G describes the QoS parameters and respective specifications (e.g. the guaranteed bitrate (GBR), the packet delay budget). Accordingly, the CC Profile associated to a 6G CC Flow describes the Compute Services (CSs) to be used for that flow and provides further specifications, such as specific threshold values to be applied to the computation carried out).

The illustration in Figure 49 further denotes a 6G UE which has a CC Flow established to a remote application server. The UE is sending packets via the AN to the first CUPE, in this case a UPF. The Compute Service(s) running on that CUPE modify the packets while forwarding them (indicated as the modified border colour). The modified packets are forwarded to another UPF, also acting as a CUPE, where the packets are once more modified. Finally, the UE's packets are transmitted via the DN to the remote application server. The packets received by the server differ from the packets originally sent by the UE.

### 8.6.5 Considerations of the Architectural Impact

In order to realize the envisioned IUP for 6G as described above, and specifically the concept of CC flows, major modifications as compared to the 5G architecture are necessary. Besides the UP enhancements, the capabilities of the relevant involved control plane NFs must be enhanced. New capabilities of the AF include for example the definition and deployment of new, customized CSs at the CUPEs or the initiation of the CC Flow setup. The SMF (or the respective 6G counterpart), when selecting the UP path, needs to additionally consider the set of supported CS at the UPFs as well as their computational load.

A 6G access node acting as a CUPE may perform capillary traffic offloading to a local access to the DN, resulting in an enhanced utilization of the underlying transport network and in reduced latency. Any use-case requiring high throughput and low latency everywhere in wide areas would benefit from such a solution. However, it requires to merge or co-locate the AN with some networking functionalities which are so far only implemented by the UPF (e.g. uplink/downlink QoS enforcement, downlink traffic notification, traffic forwarding, etc.), to allow operating on higher, i.e. PDU, layer packets at the AN. The consequence is a break with the functional split between access and core network – one of the key design principles of 5G systems. To summarize, the required functional and architectural changes that are necessary for realizing the envisioned Intelligent User Plane can only be implemented at a generation shift towards 6G [23].

## 8.7 In-Network Computing as an Enabler for Split-AI

The increasing use of artificial intelligence (AI) and machine learning (ML) techniques in mobile applications is revolutionizing our world. AI/ML technology is being widely utilized across various industries. For instance, in mobile communication devices such as smartphones, robots, and cars, AI/ML models are utilized to perform complex tasks like speech recognition, image recognition, and video processing.

The 3GPP Technical Report [47] analyses the types of AI/ML operations that can be supported on top of 5G System, one of which is AI/ML operation splitting where the AI/ML operation/model is divided into different parts depending on the task and environment. For example, the computationally and energy-intensive parts can be offloaded to the network endpoints while the privacy-sensitive and delay-sensitive parts may be left on the end device. The intermediate data is sent to the network endpoint for further processing, and the results are sent back to the device.

### 8.7.1 Split-AI Use-Cases

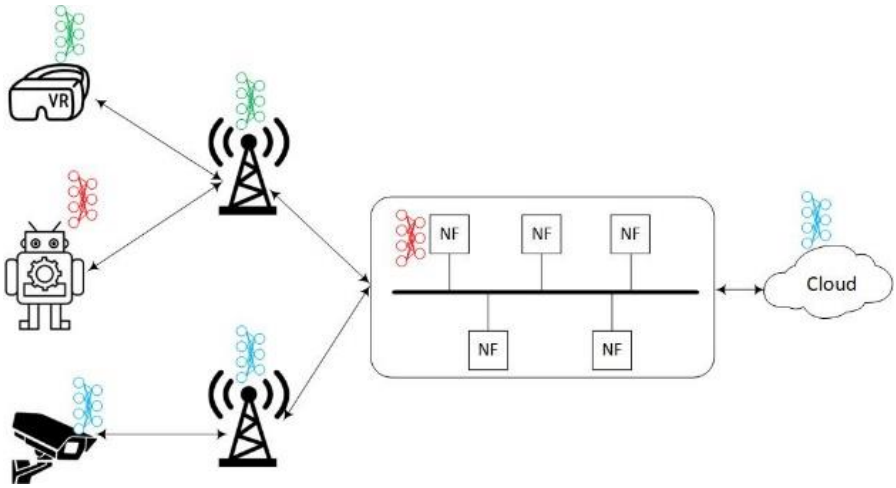



*Figure 50: Example Use Cases*

Image recognition, which encompasses a wide range of applications including surveillance, security, robotics, and automotive driving, is a critical area where AI/ML model/operation splitting plays an important role. As image and video content account for more than 70% of daily Internet traffic, it is essential to consider the computational load, data rate, and privacy requirements at Internet Data Centres [48]. By dividing the AI/ML inference/model, used for image recognition, between the mobile device and the network server, we can effectively reduce latency and energy consumption, improve accuracy and privacy, and relieve resource pressure on both sides. In this approach, the device handles privacy, and delay-sensitive tasks, while the network server takes care of the computation-intensive and energy-intensive parts of the Convolutional Neural Network (CNN) model, as illustrated in Figure 50. Moreover, these techniques can be applied for media quality enhancement purposes. Imagine being able to fully immerse yourself in a virtual reality game with high-quality videos, such as 8K per eye at 120 frames per second. Even if hardware, load, and networking limitations only allow for 4K video streaming, low-resolution video can be upscaled using AI/ML techniques to 16K content, enhancing user experience to a different level. Additionally, splitting the model provides the aforementioned benefits in this particular use case.

Another interesting area to leverage the benefits of AI/ML model splitting is mobile robots which have become increasingly important in scenarios such as warehouses, disaster rescue, and smart factories thanks to their high mobility. However, to enable these robots to perform their tasks in ever-changing environments, they need to have fast and reliable sensing, planning, and control capabilities. However, achieving these capabilities can lead to increased computation requirements and power consumption if performed on the robot. The lightweight form required for mobile robots

working in real-world environments prevents them from being equipped with a large number of CPU/GPU units and large-capacity batteries. To address this, researchers have studied the offloading of computations from robots to the cloud [49]. However, designers of autonomous mobile robots also need to consider scenarios where the robots must have local processing capacity to ensure low-latency responses during periods of varying network access quality. The resulting split control system is different from fully remote-controlled robot systems, which rely on cloud computing for planning and control while the robots only report sensing data and receive control commands. Split control allows for more efficient processing, as the complex and less latency-critical tasks can be offloaded to the cloud or edge control server, while the low-complexity and latency-critical tasks can be performed locally by the robot. A study showed that a robot completely controlled by a cloud server cannot finish a walking task if the round-trip latency exceeds 3ms. However, with split control, the robot can still perform the task with a worse-case round-trip latency of 25ms [50].

### 8.7.2 Split-AI Inference Models

The 3GPP deployment options (modes) introduced in [47] can be used for both AI/ML training and Split-AI inference. Modes a) and b) are the traditional approaches where the AI/ML inference is handled entirely on one endpoint. Modes c) through g) aim to divide the AI/ML inference or model into multiple parts based on the current task and environment.

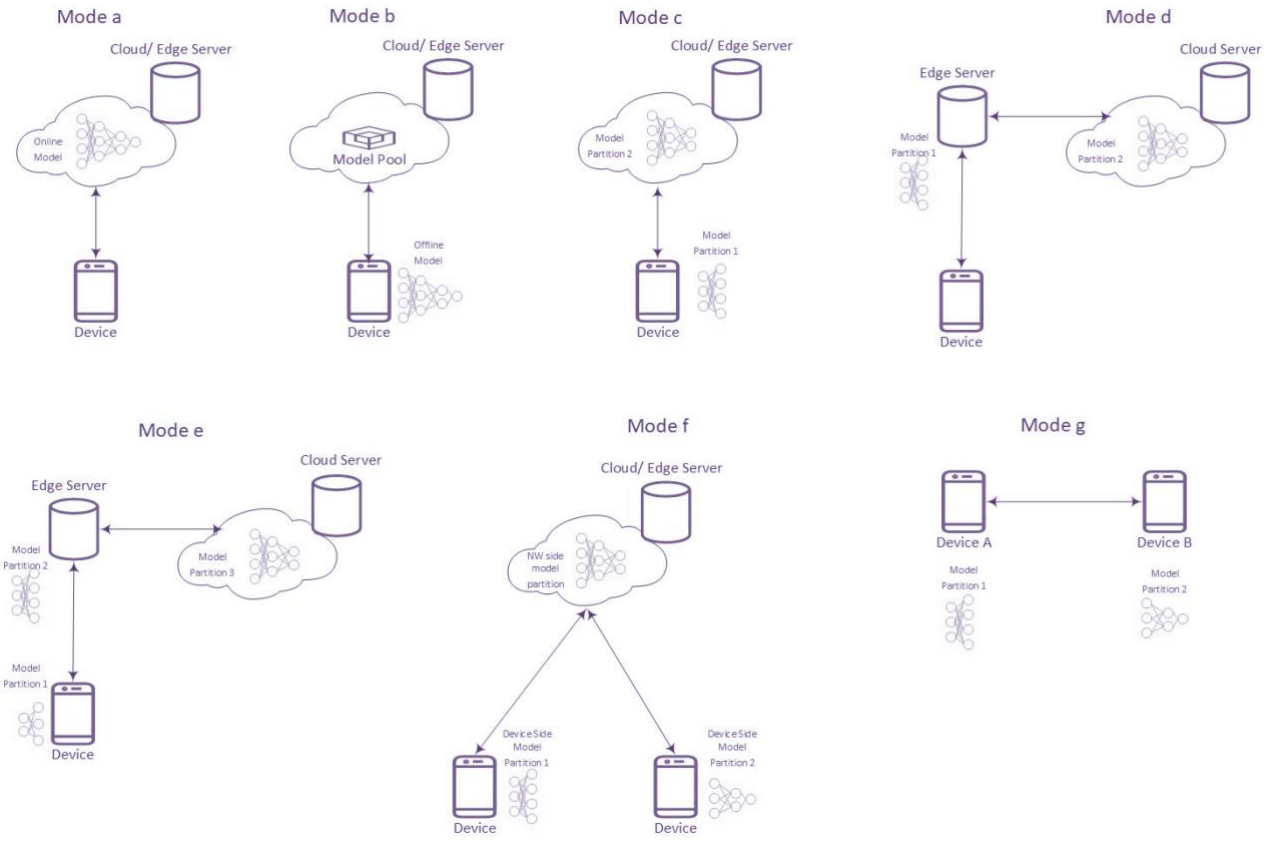



*Figure 51: AI Splitting Modes [47]*

There are several modes of AI/ML model inference to consider as shown in Figure 51. In Mode a), cloud/edge-based inference, the model inference is carried out in a cloud or edge server, and the device only reports sensing/perception data to the server. While this limits device complexity, inference performance depends on communication data rate and latency, which can be

challenging to guarantee in 5G systems. Additionally, privacy protection measures are necessary due to the disclosure of sensitive data to the network.

In Mode b), device-based inference, the AI/ML model inference is performed locally on the mobile device, preserving privacy at the data source. However, this can potentially impose excessive computation, memory, and storage requirements on the device. In some cases, the device may need to obtain the AI/ML model from the edge cloud/server, requiring corresponding downloading data rates from the 5G system.

Mode c), device-cloud/edge split inference, offers more flexibility and is robust to varying computation resources and communication conditions. The inference operation or model is split between the device and server based on current system environmental factors. The device executes the AI/ML inference or DNN model up to a specific part or layer and sends intermediate data to the server which executes the remaining part/layers and sends the inference results to the device.

In Mode d), edge-cloud split inference, the DNN model is executed through edge-cloud synergy, with latency-sensitive parts or layers performed at the edge server, and computation-intensive parts or layers offloaded to the cloud server. The device only reports sensing/perception data to the server.

Mode e), device-edge-cloud split inference, combines Mode c) and Mode d). The AI/ML inference operation or model is split over the mobile device, edge server, and cloud server. The computation-intensive parts/layers can be distributed among the cloud and/or edge server, while the latency-sensitive parts/layers can be performed on the device or edge server, preserving privacy-sensitive data at the device.

Mode f), device-device split inference, provides a decentralized split inference where a group of mobile devices can perform different parts of an AI/ML operation or different DNN layers and exchange intermediate data with each other.

Finally, Mode g), device-device-cloud/edge split inference, combines Mode c) or e) with a de-centralized execution of the device part split over different mobile devices. The intermediate data can be sent from one device to the cloud/edge server, or multiple devices can send intermediate data to the cloud.

Considering the deployment options defined in 3GPP architecture, it is observed that there is still a risk of huge amounts of data being transmitted to endpoints for processing, which might cause congestion and degradation in network performance. Here, the remedy is using In-network computing, which offers numerous benefits when compared to traditional methods. In-network computing decreases energy consumption by allowing computations to be executed in the network user plane devices, resulting in less data being transmitted to the application server in the cloud. In addition to the benefits related to traffic volume and energy efficiency, in-network computing can also improve network security. Analysing data in real-time during transmission can identify and prevent security threats, thereby thwarting attacks before they cause any damage.

In the current state of the art, the split of computation between the end user and network endpoints is already discussed in order to deliver effective results for the user. These tasks may include machine learning and other complex computations that require input data and models, e.g., rendering augmented reality with limited computational resources in a headset. In the 6G system, with the vast potential it provides, AI splitting on in-network elements is also a promising approach. The decision of how to divide the computation task between the UE and other resources will depend on the network conditions and available computational resources in the UE and within the network. It's important to note that the decision of how to split computation is not discussed under the context of 3GPP, and it is considered an implementation detail.

### 8.7.3 Key challenges of INC-assisted Split-AI

In the following, we refer to INC-assisted Split-AI as the case, where an NN is vertically split into multiple parts (each part comprising one or more NN layers) and the resulting parts are allocated to different compute nodes. In the following, we assume that a 6G AN, the 6G UPF, as well as the

UE and the application server can serve as a compute node, i.e., AN, UPF, UE, and server may execute some layers of the given NN. The execution of NN layers in AN and UPF is enabled through INC, hence, we refer to the term **INC-assisted Split-AI** (INCaS-AI) in the following. Further, we assume that the Communication and Compute Control Entity (CCCE), introduced as key enabler 4 in Section 8.6.3, is in charge of determining the optimal split, the allocation of NN layers to compute nodes, and delegating the layers to execute to the involved UP entities.

Introducing INC-assisted Split-AI as a native feature into 6G networks is coupled with a number of challenges that need to be solved [51]. This includes among others the following issues:

1. Variety of potentially conflicting optimization goals: Fulfilling optimization goals for a combination of communication performance metrics, compute performance metrics, and AI-based application layer's performance metrics is extremely complex. These optimization goals may include, e.g., minimizing the UE energy consumption, minimizing the overall energy consumption, reducing the generated traffic volume, or keeping the end-to-end latency of the application as low as possible. Further, to save on compute resources and traffic volume (i.e., if capped by the user's subscription plan), it could be essential to minimize the compute load in the network and minimize the traffic that would be billed to the user. The KPIs to optimize might even relate to various possible stakeholders of the e2e service (e.g., end user, operator, and the application provider). Thus, the question is which of the many possible optimization targets set by different stakeholders (end user, application provider, mobile network operator, cloud service provider) should be prioritized to benefit from Split-AI operations. Some of the optimization goals may even be conflicting with each other. For example, if the UE's energy consumption should be minimized, an obvious approach would be to offload as much compute load as possible to the network or the application server in the cloud. This would therefore reduce the UE energy consumption with the cost of increasing traffic and compute load in the network and the cloud. Evidently, those trade-offs must be carefully considered when designing the end-to-end system.

2. Complexity of finding the optimal split: Even in the simplest case, where one single optimization target is set (e.g. minimize end-to-end inference latency), it is still challenging to find the optimal NN split and the respective compute node allocation. This is due to the amount of degrees of freedom (DoF) when it comes to allocating NN layers to compute nodes: i) the number of partitions of the NN and the split positions, ii) the selection of UP entities (UP path selection), iii) the allocation of NN layers to compute nodes. Let us assume a simplified case, with the UP path being short and already set: UE-AN-UPF-Server, i.e., we eliminate DoF ii). Further, let us consider a very simple NN with only three layers. This creates four different split options (no split at all, 2 possible options with one split, splitting at each layer). This simplified consideration already results in 17 different options for allocating the NN layers to the four compute nodes (UE, AN, UPF, server). The three NN layers must be executed in order. The inclusion of just one alternative UPF that could be used for NN execution would lead to the double amount of possibilities, i.e., 34.

3. Large amount of information to be collected and processed: Not only the complexity of determining the optimal split is a challenge, but also the amount of information needed. The optimal NN split and allocation of NN layers to compute nodes depend on a multitude of factors. These factors can be related to the end user, the application workload (the NN to be applied), as well as the network conditions (radio access and transport network). Figure 52 summarizes the key information (third column) that would need to be exchanged in order to fulfil the required novel capabilities (second column). As it can be seen, the information required in the 6G CP is very comprehensive, comprising various metrics that should be provided by from application layer, end user, UE, and UP.

| | *Required Capability* | *Needed Information Exchanged* |
|---|---|---|
| 6G CP | - Determine number and locations of splits of the NN<br>- Determine UP path and NN layer allocation<br>- Charging (user/App provider to pay for INC service) | **Application-related**: Requirements (e.g. inference time), per-layer info (complexity, output size) as it determines the processing latency and traffic generated between compute nodes)<br>**User-/UE-related**: Privacy requirements (e.g. user consent to share info with network), energy-related info (e.g. current battery level), user preferences (e.g. minimize INC usage as much as possible to save monthly cap limit according to subscription plan)<br>**UP-related**: Communication-related information (UL/DL volume, delay), computation-related information (static, e.g., available CPU/GPU and dynamic, e.g., current CPU/GPU utilization) |
| CP ↕ UP | **Efficient signaling**:<br>CP → UP: Enforcement of NN layers to be executed<br>UP → CP: Various monitoring information | **CS (NN-layer info:)** Activation Functions, weights, biases (possibly compressed NN model)<br>Provide UP-related information needed at the CP on a given time-granularity |
| 6G UP | - Execution of NN layers as an INC CS<br>- Support required level of dynamicity | Compute requirements (e.g. the CUPE's time budget for executing the assigned NN layers, provided by the CP via the CC Profile) |

*Figure 52: Novel 6G capabilities and signalling information required for INC-assisted Split-AI [51]*

4. Higher level of signalling overhead: INCaS-AI operations will generate additional signalling overhead, as the communication stack and computing stack require coordination. All information denoted in the table above needs to be acquired by the CCCE, which is assumed to be residing in the 6G CP. The split and allocation decision need to be propagated from the CCCE to the entities in charge. Enhanced interaction is required between the 6G system and the application provider, e.g., via the AF and NEF (Network Exposure Function). The two most important enablers to mention here are (i) signalling for providing the NN-related metadata information from the application provider to the 6G system and (ii) signalling for providing the information on how the NN layers are scheduled to the different execution nodes, e.g. to the UE, network and the cloud, as the cloud and the network need to know which parts they have to execute. Further, enhanced interaction between CP and UP is needed. This refers to the collection of monitoring information (e.g. UL/DL bit rate at the radio interface or compute load of AN and UPF(s) involved), as well as exchange of control messages to inform the UP entities about the layers they need to execute. One important aspect in the context of scheduling and associated signalling complexity relates to scalability of the whole system. Since NN models can be huge, intelligent mechanisms are needed to efficiently handle the representation of a whole or partial NN within the 6G system. During CC Flow setup, each UP entity is informed about the layers it needs to execute. High dynamicity is required, because the layers to execute can be different for each flow. Thus, the signalling overhead would grow exponentially if for each CC Flow setup huge NN models must be shared between CP and UP.

5. Business relations and trust between multiple systems: Given the data being an asset for the application providers, as well as users' privacy policies, opening the NN model over the network infrastructure, with different business owners, is not a done deal. Despite the willingness of mobile operators to host storage or compute for applications at their access network, over the past two decades, such services did not become a reality. Although recent years have shown a great deal of convergence between different involved sectors, there is still no clear model for such business relationship.

### 8.7.4 Key requirements for the 6G architecture to support INC-assisted split-AI

From the challenges denoted above, we derive the required enhancements needed in the 6G architecture (specifically UP and CP) so to practically realize the INCaS-AI. In general, it can be stated that 6G networks must be significantly enhanced to support this increased level of complexity for Net4AI and the specific case of Split-AI. The major architectural innovations as compared to the legacy 5G system can only be realized with a generation shift towards 6G. In the following, we summarize what the next generation of mobile networks should at least provide:

- Support for INC in the User Plane with the UP nodes reporting rich information about their capabilities e.g. hardware capabilities (including accelerators), available resources per time unit and location, support for partitioned NN processing (software);

- Support for fast analytics generation, allowing to collect vast amounts of information about the whole network and cloud state;
- Fast optimization engines that can harness KPIs set by network, UE, end-user, cloud and application layer. Further being capable of resolving trade-offs between conflicting targets, and derive optimal solutions even for complex problems;
- Enhancements of signalling to enable exchange and collection of information needed at the CCCE to determine the optimal split. Further, the 6G system should provide new information spreading techniques from the CCCE to the UP, so to enforce the execution of the NN layers allocated to the UP entities. For instance, there must be capabilities for the CCCE (located in the CP) to instruct the UP entities and the application layer to execute specific NN layers. It must provide means to share such information as efficiently as possible, as NN models (and parts of them) can become very large in size and the parts to be executed by one compute entity can differ for each flow in the network;
- Novel solutions for exchanging and representing NN-related information within the 6G system and between the 6G system and 3rd party applications. This is required, among others, for efficiently signalling the assigned NN layers from CP to UP, and for the application layer to provision the NN partitions (partially to be executed by the network) to the 6G system, e.g. via the AF;
- An enhanced session management entity (e.g. an evolution of the 5G SMF) integrating CCCE functionalities. That is, it is equipped with the necessary logic and sufficient compute power to be able to solve the complex problem of determining the optimal partition of the NN. Further, it is capable of determining a suitable UP path as well as allocation of the resulting partial NNs (i.e., the NN layers) to the involved compute nodes (AN, UPF(s), UE, and cloud);
- Targeted interaction between the application layer and the 6G system to enable the collaborative computation of NNs between the application and the network.

## 8.8 Conclusions on Intelligent User Plane

This section elaborated on technical solutions and their architectural impacts for 6G networks. More specifically, it discussed different solutions for an Intelligent User plane design so to meet the challenging requirements of future real-time sensory services and applications like the Metaverse, AR, and AI-based applications. We presented a broad set of ideas, covering solutions related to a flatter network design and segment routing, as well as in-network computing and machine learning. Indeed, we have indicated how the different proposals can reduce both, the network load and latency and thus support future Internet applications.

We condensed the presented ideas into two technology trends that enable an IUP, as depicted in Figure 53. The first one by delegating functionalities to the transport layer, the second one by leveraging in-network computing. For both the technology trends, we focused on their architectural impact by elaborating on their potential control and user plane implications. For the latter, the adaptations for the 6G architecture may include (but not be limited to) the integration of new reference points to support shorter paths along the user plane, as well as an extension of the CU-UP functionality. Furthermore, a simplified UPF design as compared to 5G UPFs, in the sense that functionalities are implemented in the access routers integrated with the 6G UPF. Finally, we detailed on embedding INC features into the UPF, which can be achieved by means of UPF programmability and a logical entity, allowing to perform computations on the flows as they traverse the UPF. This logical entity could be realized, for example, by means of a new computational layer integrated into the UP stack.

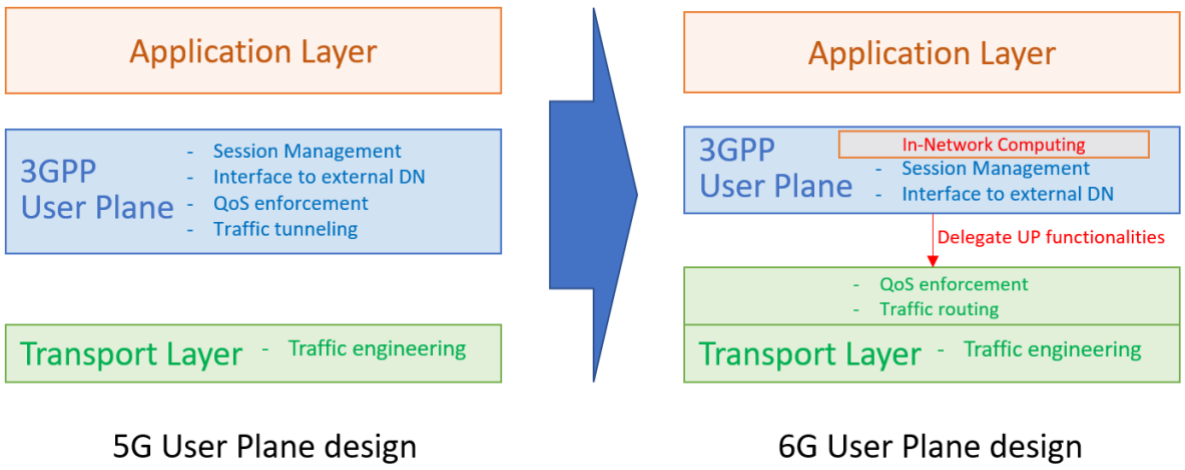



*Figure 53: Potential 6G User Plane design in comparison to 5G state of the art*

## 8.9 References


[1] 3GPP, "TS 23.501 - System architecture for the 5G system (V15.12.0)," 2020. [Online]. Available: http://www.etsi.org/deliver/etsi_ts/123500_123599/123501/15.12.00_60/ts_123501v151200p.pdf.

[2] 3GPP, "TS 23.558 "Architecture for Enabling Edge Applications" (2021), V17.2.0," 2021. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/23_series/23.558/23558-h20.zip.

[3] NTT Docomo, "White Paper on "5G Evolution and 6G"," Feb 2021. [Online]. Available: https://www.nttdocomo.co.jp/english/binary/pdf/corporate/technology/whitepaper_6g/DOCOMO_6G_White_PaperEN_v3.0.pdf.

[4] M. Behringer, T. Carpenter, L. Eckert, L. Ciavaglia and J. Nobre, "A Reference Model for Autonomic Networking," RFC Editor, Fremont, CA, USA, May 2021. [Online]. Available: https://www.rfc-editor.org/rfc/rfc8993.txt.

[5] O. Holland, E. Steinbach, V. Prasad, Q. Liu, Z. Dawy, A. Aijaz, N. Pappas, K. Chandra, V. Rao and S. Oteafy, "The IEEE 1918.1 "Tactile Internet" Standards Working Group and its Standards," Proceedings of the IEEE, vol. 107, no. 2, pp. 256-279, Feb 2019.

[6] 3GPP, "TR 22.847 "Study on supporting tactile and multi-modality communication services" (Release 18), V0.3.0," 07 2021. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/22_series/22.847/22847-030.zip.

[7] 3GPP, "TS 22.104 "Service requirements for cyber-physical control applications in vertical" (Release 16),"[Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specification.

[8] "Stream Control Transmission Protocol, IETF RFC 4930," [Online]. Available: https://datatracker.ietf.org/doc/html/rfc4960.

[9] "Multiprotocol Label Switching Architecture, IETF RFC 3031," [Online]. Available: https://datatracker.ietf.org/doc/html/rfc3031.

[10] 3GPP, "TS 22.804 Study on Communication for Automation in Vertical domains (CAV)," [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specification.

[11] IEEE, "802.1 - Time-Sensitive Networking (TSN) Task Group," [Online]. Available: https://1.ieee802.org/tsn/.

[12] 3GPP, "TS 23.501 "System architecture for the 5G System" ; Stage 2 (Release 16)," Dec 2021. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specification.

[13] 3GPP, "TS 23.501 "System architecture for the 5G System (5GS)"; Stage 2 (Release 17)," Dec 2021. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specification.

[14] N. Hu, Z. Tian, X. Du and M. Guizani, "An energy-efficient in-network computing paradigm for 6G," IEEE Transactions on Green Communications and Networking, pp. 1722-1733, 2021.

[15] K. Psounis, "Active networks: Applications, security, safety, and architectures," IEEE Communications Surveys, vol. 2, no. 1, pp. 2-16, 1999.

[16] D. Tennenhouse, J. Smith, D. Sincoskie, D. Wetherall and G. Minden, "A survey of active network research," IEEE communications Magazine, vol. 35, no. 1, pp. 80-86, 1997.

[17] S. Amedeo, I. Abdelaziz, A. Aldilaijan, M. Canini and P. Kalnis, "In-Network Computation is a Dumb Idea," in Proceedings of the 16th ACM Workshop on Hot Topics in Networks}, 2017.

[18] J. Woodruff, M. Ramanujam and N. Zilberman, "P4DNS: In-network DNS," in Symposium on Architectures for Networking and Communications Systems (ANCS), 2019.

[19] X. Jin, X. Li, H. Zhang, R. Soule, J. Lee, N. Foster, C. Kim and I. Stoica, "NetCache: Balancing Key-Value Stores with Fast In-Network Caching," in Proceedings of the 26th Symposium on Operating Systems Principles, 2017.

[20] F. Yang, Z. Wang, X. Ma, G. Yuan and X. An, "SwitchAgg:A Further Step Towards In-Network Computation," arXiv preprint arXiv:1904.04024, 2019.

[21] I. Kunze, P. Niemietz, L. Tirpitz, R. Glebke, D. Trauth, T. Bergs and K. Wehrle, "Detecting out-of-control sensor signals in sheet metal forming using in-network computing," in IEEE 30th International Symposium on Industrial Electronics (ISIE), 2021.

[22] F. E. R. Cesen, L. Csikor, C. Recalde, C. E. Rothenberg and G. Pongracz, "Towards low latency industrial robot control in programmable data planes," in 6th IEEE Conference on Network Softwarization (NetSoft), 2020.

[23] S. Schwarzmann, R. Trivisonno, S. Lange, T. E. Civelek, D. Corujo, R. Guerzoni, T. Zinner and T. Mahmoodi, "An Intelligent User Plane to Support In-Network Computing in 6G Networks," in International Conference on Communications (ICC), Rome, 2023.

[24] Y. Xun, R. Paulet and E. Bertino, "Homomorphic encryption," in Homomorphic encryption and applications, Springer, 2014, pp. 27-´46.

[25] N. Samardzic, A. Feldmann, A. Krastev, S. Devadas, R. Dreslinski, C. Peikert and D. Sanchez, "F1: A fast and programmable accelerator for fully homomorphic encryption," in Annual IEEE/ACM International Symposium on Microarchitecture, 2021.

[26] B. Reagen, W.-S. Choi, Y. Ko, V. Yeongil, H.-H. Lee, G.-Y. Wei and D. Brooks, "Cheetah: Optimizing and accelerating homomorphic encryption for private inference," in Intl. Symposium on High-Performance Computer Architecture, 2021.

[27] B. Stephens, D. Grassi, H. Almasi, T. Ji, B. Vamanan and A. Akella, "TCP is Harmful to In-Network Computing: Designing a Message Transport Protocol (MTP)," in Proceedings of the 20th ACM Workshop on Hot Topics in Networks, 2021.

[28] P. L. Ventre, S. Salsano, M. Polverini, A. Cianfrani, A. Abdelsalam, C. Filsfils, P. Camarillo and F. Clad, "Segment routing: A comprehensive survey of research activities, standardization efforts, and implementation results," IEEE Communications Surveys & Tutorials, vol. 23, no. 1, pp. 182-221, 2020.

[29] I. Kunze, "Transport protocols in the age of In-Network Computing," 04 11 2020. [Online]. Available: https://blog.apnic.net/2020/11/04/transport-protocols-in-the-age-of-in-network-computing/. [Accessed 21 06 2023].

[30] R. Gennaro, C. Gentry and B. Parno, "Non-interactive verifiable computing: Outsourcing computation to untrusted workers," in Annual Cryptology Conference, Springer, 2010, pp. 465-482.

[31] I. Kunze, D. Trossen and K. Wehrle, "Evolving the End-to-End Transport Layer in Times of Emerging Computing In The Network," in Proceedings of the 1st Workshop on New IP and Beyond, 2022.

[32] A. Radwan, H. R. Chi, D. Corujo, J. Quevedo, R. Silva, D. Santos, R. Aguiar, O. Abboud and A. Hecker, "Multi-Criteria Modeled Live Service Migration for Heterogeneous Edge Computing," in IEEE Global Communications Conference (GLOBECOM), 2022.

[33] A. Sivaraman, A. Cheung, M. Budiu, C. Kim, M. Alizadeh, H. Balakrishnan, G. Varghese, N. McKeown and S. Licking, "Packet transactions: High-level programming for line-rate switches," in Proceedings of the 2016 ACM SIGCOMM Conference, 2016.

[34] H. Nguyen, T. Van Do and C. Rotter, "Scaling UPF Instances in 5G/6G Core With Deep Reinforcement Learning," IEEE Access, vol. 9, pp. 165892-165906, 2021.

[35] Y. Li, X. Ma, M. Xu, A. Zhou, Q. Sun, N. Zhang and S. Wang, "Joint Placement of UPF and Edge Server for 6G Network," IEEE Internet of Things Journal, vol. 8, no. 22, pp. 16370-16378, 2021.

[36] I. Leyva-Pupo, A. Santoyo-Gonzalez and C. Cervello-Pastor, "A framework for the joint placement of edge service infrastructure and user plane functions for 5G," Sensors, vol. 19, no. 18, p. 3975, 2019.

[37] I. Leyva-Pupo, C. Cervello-Pastor, C. Anagnostopoulos and D. Pezaros, "Dynamic scheduling and optimal reconfiguration of UPF placement in 5G networks," in Proceedings of the 23rd International ACM Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems, 2020.

[38] F. Musumeci, C. Rottondi, A. Nag, I. Macaluso, D. Zibar, M. Ruffini and M. Tornatore, "An overview on application of machine learning techniques in optical networks," IEEE Communications Surveys & Tutorials, vol. 21, no. 2, pp. 1383-1408, 2018.

[39] Z. Xiong and N. Zilberman, "Do Switches Dream of Machine Learning? Toward In-Network Classification," in Proceedings of the 18th ACM workshop on hot topics in networks, 2019.

[40] Arista, "Arista 7170 Multi-function Programmable Networking.," 2018.

[41] N. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden and A. Borchers, "In-datacenter performance analysis of a tensor processing unit," in Proceedings of the 44th annual international symposium on computer architecture, 2017.

[42] T. Swamy, A. Rucker, M. Shahbaz and K. Olukotun, "Taurus: An intelligent data plane," arXiv preprint arXiv:2002.08987, 2020.

[43] M. Saquetti, R. Canofre, A. Lorenzon, F. Rossi, J. Azambuja, W. Cordeiro and M. Luizelli, "Toward in-network intelligence: running distributed artificial neural networks in the data plane," IEEE Communications Letters, vol. 25, no. 11, pp. 3551-3555, 2021.

[44] P. Bosshart, G. Gibb, H.-S. Kim, G. Varghese, N. McKeown, M. Izzard, F. Mujica and M. Horowitz, "Forwarding metamorphosis: Fast programmable match-action processing in hardware for SDN," ACM SIGCOMM Computer Communication Review, vol. 43, no. 4, pp. 99-110, 2013.

[45] 3GPP, "TR 22.874, Study on traffic characteristics and performance requirements for AI/ML model transfer in 5GS," 2022.

[46] 3GPP, "TS 22.261, Service requirements for the 5G system," 2022.

[47] B. Kehoe, S. Patil, P. Abbeel and K. Goldberg, "A survey of research on cloud robotics and automation," IEEE Transactions on automation science and engineering, vol. 12, no. 2, p. 398–409, 2015.

[48] Z. Huaijiang, M. Sharma, K. Pfeiffer, M. Mezzavilla, J. Shen, S. Rangan and L. Righetti, "Enabling Remote Whole-body Control with 5G Edge Computing," in IEEE/RSJ International Conference on Intelligent Robots and Systems, 2020.

[49] M.-J. Montpetit, "In Network Computing Enablers for Extended Reality," in Internet Engineering Task Force (IETF), 2019.

[50] J. Benedetto, V. Valenzuela, P. Sanabria, N. Neyem, J. Navon and C. Poellabauer, "MobiCOP: A scalable and reliable mobile code offloading solution," Wireless Communications and Mobile Computing, 2018.

[51] S. Schwarzmann, T. Erkilic Civelek, A. Iera, D. Corujo, G. Karetsos, R. Guerzoni, O. Abboud, A. Meseguer Valenzuela, R. Trivisonno, M. G. Spina, T. Zinner and T. Mahmoodi, "Native Support of AI Applications in 6G Mobile Networks via an Intelligent User Plane," in 6G-Arch: The 3rd Workshop on 6G Architecture (in conjunction with IEEE WCNC 2024), Dubai, 2024.

# 9. Flexible programmable infrastructures

6G should provide full-service support (i.e. going beyond just network connectivity) that would transform the notion of the "session" from a "connectivity session" to a full service execution. For this to be possible, scalable resource control is needed, i.e. a coherent, holistic control of a running network, which includes controlling access, routing, compute and storage nodes at the same time. Under these conditions, it also becomes possible to transform the network from a static entity, rooted in careful dimensioning and pre-planning, to a more dynamic entity with runtime assignment of resources for specific tasks such as flows, processing requests, etc. This dynamicity, based on efficient request scheduling, is known to lead to benefits for both the end user and the network operator (e.g. reducing the total cost of ownership, the network footprint, etc.).

Dealing at scale with the amount of monitoring information that will be generated in 6G systems, both from the infrastructure and the services requires embracing a more distributed paradigm for data distribution and storage. Furthermore, location-transparent access to data should be provided in 6G systems to facilitate the consumption of information throughout the system, regardless the location of the storage or the data consumer.

These changes should be accompanied by an advance in the adopted programmability model. Instead of network focused, low level programmability mode, as available now, 6G systems should adopt a more generic, declarative programmability model, where the desired status is stated instead of the set of corrective actions to reach it. That model should focus on all parts of the infrastructure, including the access, transport network, higher level network processing, storage, etc. Taken together, these traits give the future infrastructure the flavour of Flexible Programmable Infrastructures.

## 9.1 Scalable resource control

Figure 54a shows an excerpt of a resource set controlled by a hypothetical network operator. Some of the depicted resources are physical devices (e.g. switches or routers), owned by the operator, whereas many of them can be virtual, e.g. virtual machines hosted by a local, or even global, cloud provider. Some resources act as routers and forward control or data packets between different interfaces, others are leaf or stub nodes that are only source and sink of messages but do not forward them between different interfaces. However, stub nodes can be multi-homed nevertheless. This resource set may be expanded by additional resources, e.g., smartphone resources that extend the 6G control and/or data plane, as shown in Figure 54b. One leaf node gets an additional link and is multi-homed then. In Figure 54c a complete network of resources gets connected, i.e., it is included into the existing resource set. The latter two illustrate scenarios in which the operator expands (and later possibly shrinks) its network in terms of geographical footprint, capacity, quality of service under increased load, etc.

This expansion, extension of resources may be due to an increased capacity demands, e.g. due to events such as fairs, demonstrations or sport event. The original, physical resources may be owned by the event organizer or a third party (e.g. cloud provider). In any case, their quick and precise inclusion in the resource set of the operator and their subsequent usage to support existing or enable running new services is what is of interest in this scenario.

In our opinion, an essential precondition to realize this use case is resilient interconnection of all resources. This is to say, inclusion cannot come as a conventional, management activity that requires manual configuration or the added resources. It has to be dynamic, control-oriented, with each added resource being available to use almost instantaneously after it gets connected to at least one other resource already owned by the operator. This is why we see protocols to interconnect a dynamically changing set of resources the most critical contribution at this stage.

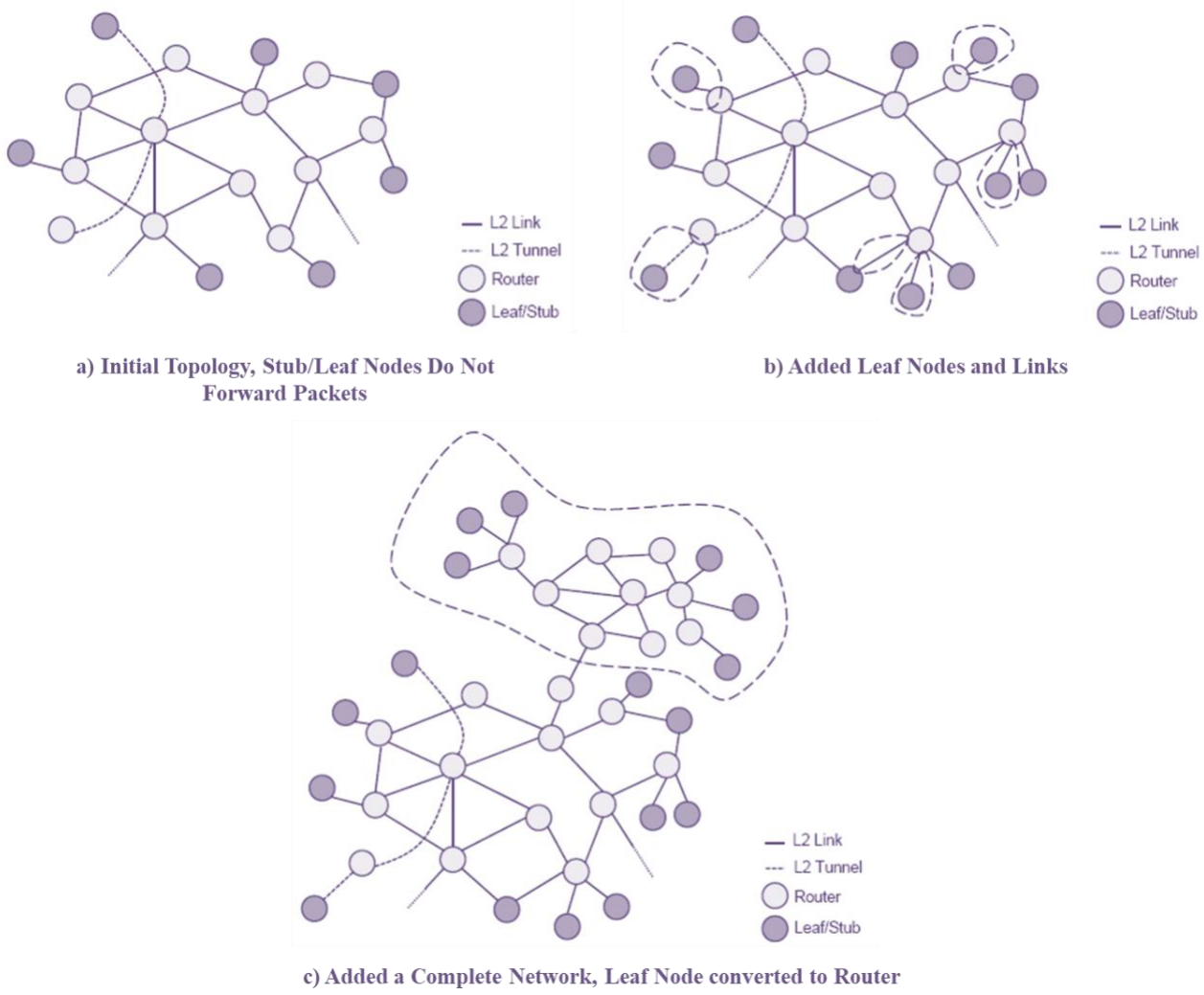



*Figure 54: Large Scale Operator Network Scenario*

There are plenty of requirements for the said interconnection protocol that can be seen relevant based on the depicted scenario, such as:

1. Scalability to a large set of resources. Large operator networks comprise several 100 000s network resources. Additional user equipment can easily scale this up to several millions of resources. Scalability should consider the size of the routing tables and the number of exchanged protocol control messages in dependence of the number of resources. This includes both initial connectivity establishment as well as the protocol's reaction to dynamic failure situations.

2. Support for continuous dynamics in the resource set due to different workload in the control plane. One can expect that the control plane has to inherently scale with the overall load and size of the system, i.e., when more users generate more services requests, the control plane resources must be scaled accordingly and (ideally) automatically. Therefore, the routing protocol must be able to converge quickly enough in order to allow for a stable and reliable connectivity among the currently active resources.

3. Supporting high dynamics (churn) of devices at the edge/access of the network. This churn stems from a large number of potential users as well as from their mobility and multi-homing possibilities.

4. Instantaneous change of a large number of resources. Nowadays an additional larger amount of resources can be provided ad-hoc by using cloud-based virtual resources, e.g., when a larger telecommunication provider includes a large set of virtual resources. Another scenario that may cause a large number of changes can occur when virtual mobile network operators lease a large set of resources from vertical telecommunication providers.

5. Support for heterogeneous topologies, i.e., the overall topology may possess a power law property, but some parts of the topology may have denser properties, e.g., resource subsets stemming from cloud-based virtual resources.

## 9.2 Next-generation programmable network infrastructures

In next-generation mobile networks, connectivity is going to be present everywhere. So there exists the need to be able to manage these extremely complex networks that expand multiple domains (even reaching the extreme edge) in an intent-based approach or in a more declarative way, i.e. with declarative interfaces to manage the whole heterogeneous network of resources. As the number of devices participating in the networking topology increases, the complexity of its management and control also increases proportionally. The main objective of such an interface is to be able to automatically translate from user requirements into network deployment and operation strategies. In order to achieve such objective, one needs to i) explore the use of big data to collect network operation and user performance data, required for the automation and acquire visibility, ii) develop AI models that are capable of predicting user experience and network performance, to be able to determine if changes in the network could impact performance, and ii) explore network programmable data planes (P4 and OpenFlow) and their respective network controller's integration into next-generation networks, to further understand how these technologies can enable an intent-based interface to control the network infrastructure.

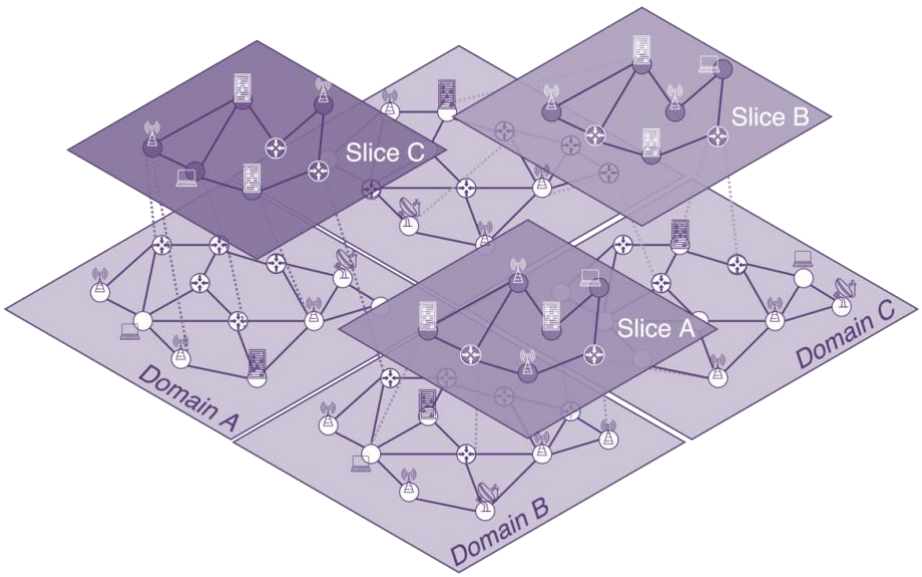



*Figure 55: Next generation network slicing*[1]

Moreover, aside from dynamically supporting and interconnecting a wide variety of devices that can dynamically be scaled based on the capacity demand and supporting declarative MANO interfaces, another use case that this WI will explore is the ability to guarantee performance isolation in such heterogeneous and decentralized networks. The network should be capable of being sliced into

[1] *This image has been designed using resources from Flaticon.com*

smaller performance segments that can guarantee delay, traffic isolation, and/or capacity for a subset of the communicating services. To realize this performance isolation one needs to explore, i) smart traffic performance detection algorithms to predict traffic patterns in microscale, and ii) dynamic network resource allocation control mechanisms, that reprogram the networking resources, based on real-time metrics, and predicted traffic, and iii) network digital twin, providing network simulation, planning, and replay capabilities.

## 9.3 Decentralised and Distributed Data Fabric

6G leverages on the shift already introduced by 5G on which the monolithic architecture of the previous generation moved towards network and service-oriented architectures. Consequently, it is pushing towards even more decentralised and distributed architectures for the data, control and management planes, while expanding it further in several technological domains targeting more disaggregated and flexible programmable infrastructures. This not only means that its architecture will be more decentralised, distributed and heterogeneous, but it will also accommodate a more open and collaborative paradigm between different stakeholders (e.g., bringing an even closer engagement of the vertical stakeholders) and interfaces towards decision-making entities either for telemetry or increased AI-based automation. Thus, programmable network infrastructure will be able to dynamically add or remove resources from different domains on-the-go and in a plug-and-play fashion way, without the burden of complex reconfigurations.

The underlying network infrastructure will then evolve from mere data pipes to an actual decentralised and distributed data fabric that is able to understand data and eventually its semantics, and therefore transparently applying additional processing to improve the overall efficiency of the network. Let's assume the following examples for a better understanding of its importance:

1. Location abstraction: In a decentralised network infrastructure, it is paramount that the operation of many of its core components are decoupled from any location requirements. In this way, nodes can e.g. dynamically join and leave or move across different location without impacting the configuration and operation of the decentralised system as a whole. To facilitate such goal, data shall become independent of its location, application, storage or any means of transportation. While in traditional data pipes the end points of these called pipes must be known à priori, a data fabric supporting the network infrastructure is expected to improve efficiency, security, and provide better scalability and robustness for decentralised systems.

2. Plug-and-play functionalities: the capability to introduce new functionalities is paramount for the so-called programmable network infrastructures, thus they must be incorporate in a seamless way that does not disrupt the entire system. For example, network storages can be easily plug-in anywhere in the cloud-to-things continuum, making them ubiquitously.

3. Fencing-based data governance: if part of the infrastructure is shared between different parties, or exposed to third-party users, it is paramount to isolate each tenant and ensure that information does not leak to unauthorised parties. The same applies when data shall not cross a given regional or global boundary. In this case, the data fabric itself must be able to ensure the previous cases, releasing the infrastructure managers, service providers from such burden.

Altogether, the scale at which data will be generated and consumed by 6G systems, both from infrastructure and services, will highly increase, imposing innovative and distributed paradigms for data distribution and storage that can simultaneously preserve data privacy and anonymity. In the following are presented several characteristics that should be considered in the design of a data fabric:

1. Communication paradigms: Data can be consumed in by following different communication

paradigms: push or pull. On one hand, the push communication paradigm data is sent towards the destination without having a prior request. On the other hand, both poll and pull communication paradigms require an explicit request by the consumer prior to the sending of data. Moreover, pull communication paradigms can be subcategorized into publish/subscribe or request/response, which differ on how data is requested and delivered. While the former only needs to request the data once (i.e., subscription) which is then sent to whenever an update occurs, the latter needs to send individual requests for each data retrieval.

2. Cardinality: Producers and consumers of data might define different relationships between one another by defining different communication associations and scopes, which can be classified as unicast (1-to-1), broadcast (1-to-many), multicast (1-to-many, many-to-many), or anycast (many-to-few. many-to-1). Each differs on how data is first grouped among communication entities and later transmitted and forwarded within the network.

3. Data Consistency: Keeping data uniform as it moves between communication entities is paramount for a consistent view of a distributed system state. It can be categorised into point in time consistency, transaction consistency, and application consistency. Moreover, different consistency models can be considered as a balance between consistency, availability and partition tolerance (i.e., CAP theorem) [24].

4. Location transparency: Although the origin of data is always tied to a specific location, entities in a distributed system intend to fetch data about the overall system (or of a specific function) without explicitly stating the location of such data. As such, host-oriented approaches introduce bigger overhead and complexity when compared to data-oriented approaches. In the latter, producers and consumers address the data itself instead of the host serving it, thus relying on the network infrastructure to forward the requests towards the entities containing such data and retrieve back the response to the consumers.

5. Storage: In general, that data is produced and consumed on the spot and at the time it was generated, thus putting the burden on the consumer side in case it needs to access past data. The capability to store any data not only increases the flexibility of the data distribution but also decouples the producer and consumer in time, facilitating the integration of new entities and mechanisms. For example, AI/ML decision-making entities to build datasets for their training stages as well as to create more complex workflows where the input might vary according to intermediary decision outputs.

6. Operation Modes: Data exchanges must support different operation modes both dependent and independent from the network infrastructure. In a dependent mode, communication entities rely on brokers or rendezvous points to match and forward data between them. In an independent mode like peer to peer or ad hoc modes, communication entities can organize themselves in different network topologies (e.g., mesh network) while being able to not only forward data between directly connected entities but also route data in a multi-hop fashion.

Nevertheless, the decentralised and distributed data fabric supporting a 6G system will have to accommodate a multitude of usage scenarios and application since they might differ in terms of requirements and/or data distribution needs. Therefore, it must be flexible to support the widest range of capabilities to support the current and envisioned requirements but at the same time to be extensible to support non-expected requirements that might arise in the future.

Since network intelligence is one of upcoming aspects to be embedded in 6G systems, which highly rely on data for its operation, see Figure 56, as an example, the trade-offs to be tackled by a network intelligence native framework in what concerns its data fabric needs. These are intrinsically different, mostly driven by the monitoring and enforcement elements in the network (as depicted in the left-hand side of the figure). As the main driver for decisions made by a network intelligence is time, time is either directly or indirectly bounding the quality and selection of selected algorithms. When associated with a given task for network intelligence, these trade-offs must be taken into

consideration in order to meet any task deadlines (as depicted in the right-hand side of the figure). Data collection and/or data processing depend on the underlying infrastructure and its capabilities to complete a given action, affecting directly on the time needed to execute the network intelligence algorithms. Two main contextual environments define the conditions upon which the network intelligence algorithms must be executed: (i) the infrastructure monitoring data; and (ii) the infrastructure decision enforcement. Finally, once timing constraints are set, the network intelligence algorithm must consider further degrees of variability that must be articulated by means of selection and implementation of its decision-making model [23].

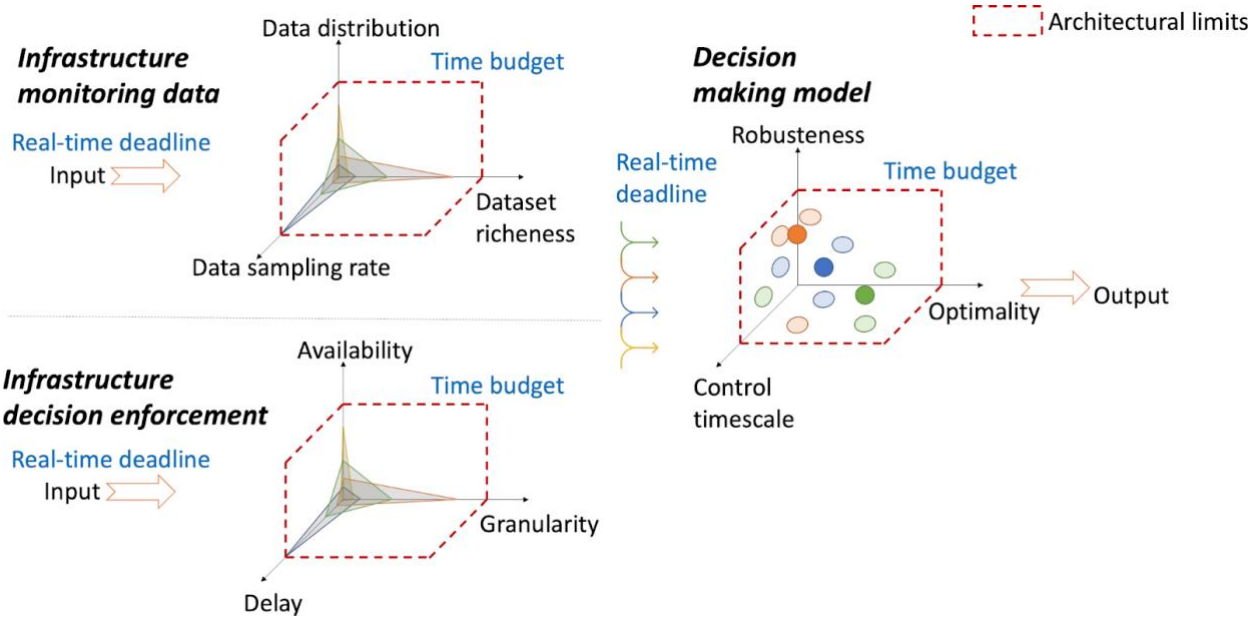



*Figure 56: Network Intelligence Degrees of Freedom [24]*

Therefore, it must be flexible to support the widest range of capabilities to support the current and envisioned requirements but at the same time to be extensible to support non-expected requirements that might arise in the future.

## 9.4 State of the art

The increased use of software-based and virtualized network infrastructures promises the provision of more flexible and elastic network services, but also inflicts new challenges for Operations, administration, and Maintenance (OAM) of the corresponding network resources: their large number, the higher dynamics, and the denser meshed topologies. A separate infrastructure for out-of-band OAM (which would also require its own setup, configuration, and OAM) has prohibitive costs and scaling limitations. Thus, a highly reliable and stable in-band control plane connectivity between the network resources for OAM is required [8][9][2].

Network topologies have changed a lot in the last decades. While network links were expensive and network topologies were sparse in the past, the trend toward denser, highly connected networking topologies is largely driven by the advent of data centre fabrics (e.g., Leaf-Spine, Clos or Fat Tree topologies [1]) and use of virtualization technologies in networking (software-based switches and network function virtualization). Virtualized network infrastructures imply not only an increased number of connected nodes as virtual instances can be easily created in larger numbers, they also entail higher dynamics, i.e., topology changes, due to their on-demand supply feature (elasticity).

Legacy routing protocols do not scale well in such highly connected denser topologies and require modifications [15][4][20]. Traditionally, scalability was considered by subdividing networks into separate areas (e.g., as in introduced in OSPF [17]). However, this required careful manual configuration and is unfeasible for large scale highly virtualized elastic infrastructures: Forthcoming

network infrastructures must be easier to manage or must be even able to manage themselves [5]. Current approaches try to reduce overhead and configuration effort [15] in specific topologies, but do not change the basic flooding nature of the protocol or are limited to specific topologies [20][22].

Example uses for control plane fabrics are the control connections between a controller and its controlled switches of SDN-based networks or between the Virtualized Infrastructure Manager and its resources in the NFV-MANO framework [10], the Network Virtualization Authority to interconnect network virtualization edge switches as considered in the NVO3 Architecture [3], or cluster networking (networking solutions for Kubernetes). For the latter two, BGP with reflectors is often used, which also requires configuration.

DISCO [18] is a distributed compact routing scheme for ID-based routing that scales with $O(\sqrt{n \log n})$ and provides a worst case stretch guarantee. It is not a genuine ID-based protocol since it uses topological addresses and two mapping systems for ID to locator resolution. The approach is more complex as it requires four protocols: synopsis diffusion for estimating the number of nodes, a path vector protocol within a node's vicinity and to all landmarks, a distributed hash tables (DHT) protocol across all landmarks, and a DHT combined with a distance vector protocol for the sloppy group maintenance. Since traffic is routed via landmarks, one can expect some traffic concentration at landmarks. The dynamics of the protocol (i.e., reaction to link failure and link repair) have not been designed or evaluated in [18]. A closely related approach is UIP (Unmanaged Internet Protocol) [11][12] that also uses ID-based addressing and a Kademlia-based routing table. Its main goal is to provide connectivity between otherwise unconnected domains and subnetworks. The efficiency of the routes themselves was not in focus (e.g., no PNS or PR used) rather than efficiently finding some route. Moreover, dynamics besides network split and network merge were not considered or evaluated. Similarly, the original Kademlia scheme [16] does not provide any route rediscovery mechanism: non-reachable nodes get merely evicted after some time. Virtual Ring Routing (VRR) [6] is also a DHT-inspired routing protocol that is ID-based. It does not consider route efficiency so some routes may incur high stretch. VRR sets up virtual links along the path instead of using source routing to route between ID-wise neighbours. This requires to establish forwarding state in intermediate nodes thereby reducing scalability: some nodes have to establish lots of forwarding entries and may have to forward lots of traffic. In contrast to KIRA's PathID scheme, paths setup by VRR are not aggregable. VRR was designed in the context of wireless ad-hoc networks and was evaluated in small topologies (200 [6] – 1 024 nodes [18]) only. VIRO [14][7] proposes a virtual ID-routing scheme based on an additional mapping layer that employs a Kademlia like structure. However, changes in the physical topology require changes in the virtual ID space, i.e., the virtual IDs are topology dependent. The protocol requires address space construction, address assignment as well as a publish and query mechanism for address mappings.

The RPL routing protocol [21] was chosen as routing protocol for the Autonomic Control Plane [9] and is said to be scalable as it aims to connect 100 000s of IoT devices. RPL is a distance vector protocol that creates a tree-like structure (DODAG), but requires configuration of DODAG roots and creates heavy traffic concentration near DODAG root node [19]. The routing table can be extremely small as it only needs to store a few paths towards the root. Its inherent treelike structure enforces routes along the DODAG, leading to high stretch and inefficiency in certain topologies. Using more efficient routes comes at the cost of additional entries or also additional route discovery overhead [13]. Recent efforts try to mitigate the scaling issues of link-state protocols in denser data centre topologies [20]. However, some of the solutions possess inherent scaling limitations, because they still use flooding and state, while other solutions (e.g., RIFT) are designed for specific topologies only.

Modern systems operating across the Cloud-to-Things continuum highly rely on information exchange between devices and applications while operating across a network. Usually, to support such approach information exchange at scale is provided by cloud computing, which provides on-demand availability of computer system resources, especially data storage and computing power, without direct active management by the user. For example, Function as a Service (FaaS) services [25], such as Azure IOT Edge, Amazon Greengrass, Google's Cloud IoT, Apache OpenWhisk, or OpenFaas, are used to support the Things segment through widely-adopted platforms for stream processing. Although it provides a level of abstraction of the continuum, OPEX reductions, or

functions scalability, developers still depend on lower-level communication frameworks in a very segmented environment. As an example, different solutions are used to address data-in-motion (e.g., Apache Kafka, DDS, MQTT, NATS, øMQ) or and data-at-rest (e.g., MariaDB, OpenZFS, Amazon S3, InfluxDB), which usually are not efficient or design to different segments of the continuum. Lastly, the things segment has always been hidden from such integration through points of contact, like gateways or brokers, that intermediate the communication between the IoT devices and the Edge and Cloud, as a way to decouple solutions from the requirements of low-power and constrained networks widely used in the IoT domain (e.g., Bluetooth, LoRA, Zigbee, Thread, etc). Although such solution abstracts the constraints and requirements from the IoT domain, it implies a centralization of communications that introduces single point of failures and attacks, and limits the potentially to use the resources available in last segment of the continuum.

As a consequence, this diversity and heterogeneous of programming frameworks and models hinders the efficient development of any decentralized and fully distributed data fabric that exploits all the resources in the continuum, since developers require to go through a complex and cumbersome process of patch working and integrating different solutions [26].

[30] is a potential solution to the problems outlined so far. It is a highly scalable a robust routing protocol that provides connectivity for extremely large networks (e.g. up to or even more than a hundred of thousands of nodes) under extreme conditions such as frequent node failures, mobility, etc. The approach taken (named KIRA and R2/KAD, KIRA being the entire forwarding/routing framework, R2/KAD being the underlying routing protocol) is based on a flat ID-based addressing scheme to easily support self-organization and zero-touch as well as mobility and multi-homing. ID-based routing has the advantage of providing IDs as stable addresses to upper layers. Thus, in case (virtual) resources are moved within the topology, any control connection to them stays alive. KIRA is a genuine ID-based scheme, because it does not use topological addresses at all and thus does not require any additional identifier-locator mapping (increased risk of non-consistency) and associated protocols (additional overhead and convergence time).

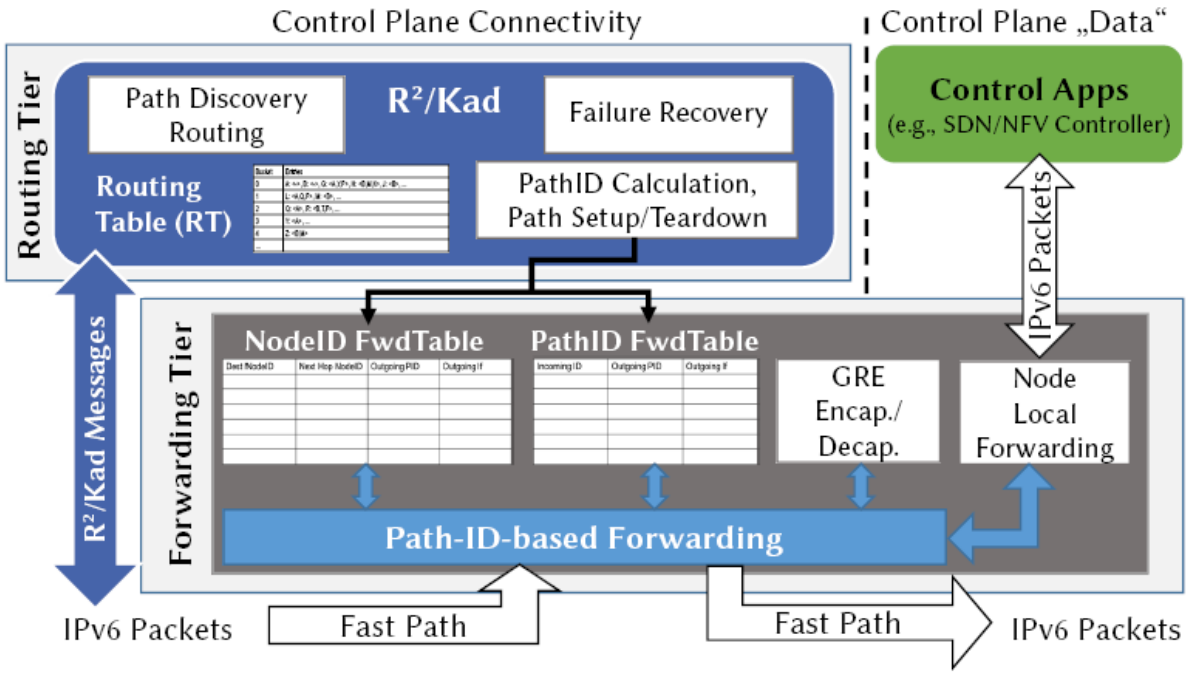



*Figure 57: KIRA Architecture [30]*

Figure 57 depicts the architecture of KIRA. It consists of essentially two parts: Path-ID-based Forwarding Engine and R2/KAD routing protocol. R2/KAD messages are forwarded using source routing. In order to avoid the per-packet overhead of source routing for KIRA's data packets the Forwarding Tier (extension and adaptation of the approach taken in [31]) employs Path-ID-based forwarding for the learned paths. It thus enables a more efficient and fast forwarding of data

packets. A Path ID denotes a certain path uniquely and is used as path label. It replaces the source routing path. Some Path IDs are distributedly computed a priori, others need to be setup in intermediate nodes by R2/KAD on demand. KIRA's Routing Tier does not depend on the Forwarding Tier and also works without it as its routing messages bypass it (cf. Figure 57). R2/KAD constructs the forwarding tables and includes the path setup procedure to install Path IDs along certain paths. Figure 57 also shows that R2/KAD messages and data packets use IPv6 with IDs as addresses. The Forwarding Tier uses IPv6 GRE (Generic Routing Encapsulation) for packets with Path IDs as destination.

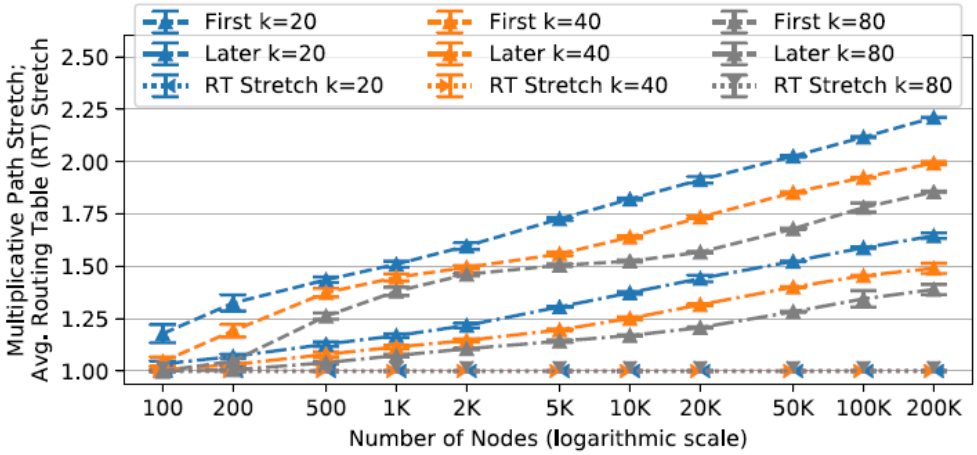



*Figure 58: KIRA performance [30]*

We refer to [30] for details about the detailed operation of the routing protocol and the forwarding engine. Instead, we now give a glimpse of the KIRA performance. Figure 58 shows path stretch to randomly chosen destinations for first and later packets as well as average routing table stretch (i.e., path stretch for contacts) in dependency of different k values and network sizes n. Stretch of response packets is between that of first and later packets. For k=40 path stretch of later packets is below 1.25 for sizes up to 10K nodes and below 1.5 even for 200K nodes. Stretch for first, response, and later packets grows with O(log n), caused by the average number of required overlay hops that increases in the same way with n. The results highlight the efficacy of eliminating cycles for response packets and applying shortcuts for later packets. Stretch of later packets is reduced by about 25% compared to first packets. Most importantly, the average RT stretch is close to 1 irrespective of n.

## 9.5 Runtime request scheduling

In an environment as dynamic as 6G is expected to be, in which a plethora of diverse services are supposed to coexist and efficiently share the underlying infrastructure, proper scheduling of incoming service request (in the broadest possible sense of the word) becomes critical. In the following we briefly describe the approach taken by [27] to solve the problem.

The working environment of [27] is shown in Figure 59. Service request scheduling refers to the selection of the 'best' service instance, within the set of active service instances, to serve a service request at runtime of the overall system. This scheduling decision is performed only at the ingress SARs, which receive the service requests from the clients, while the remaining SARs illustrated act as forwarding nodes. The scheduling decision is interpreted as a service request routing problem at the data plane level.

The implemented and evaluated runtime scheduler is one that takes the computing capabilities of the service instances in the network into consideration for the scheduling decision. It is therefore referred to as the Compute-Aware Distributed Scheduler, or CArDS. The objective of CArDS is to maximize the system's processing throughput by minimizing the (service) request completion time (as the sum of the delays at SARs and instances, together with network propagation delays) for individual service requests.

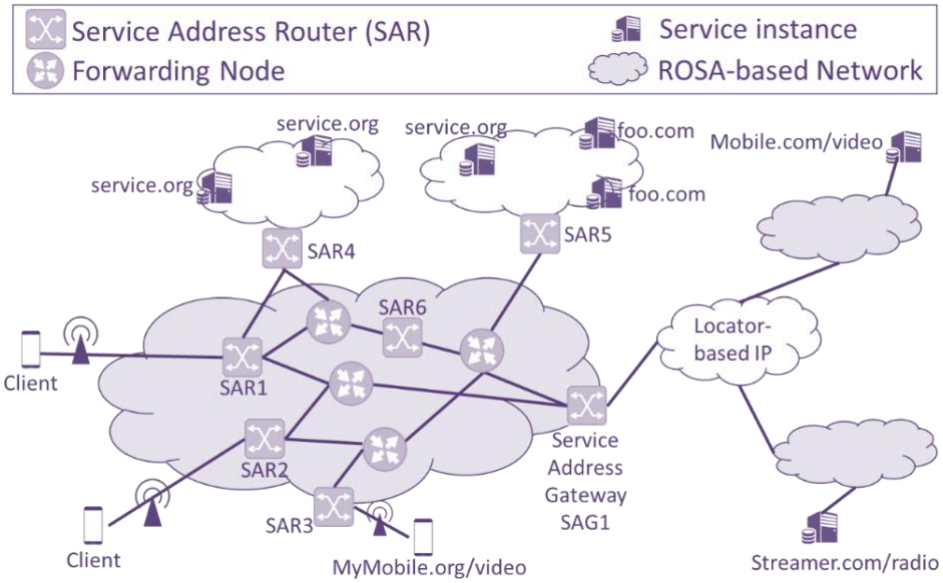



*Figure 59: Service routing system overview*

The forwarding is realized as a two-stage process. First, the ingress SAR determines all outgoing interfaces along which an incoming service request could be sent. It then selects the appropriate interface to be used by implementing the scheduling decisions, which is elaborated in the following paragraphs. In essence, the SAR performs an on-path resolution of the service identifier provided in the service request to (a direction towards) a possible service instance; with this, the SAR has taken over the role of the DNS albeit utilizing the compute awareness in the scheduling decision to forward packets. A forwarding SAR then simply forwards the request to the next hop of a SAR, utilizing suitable encapsulation techniques.

Key to the compute-awareness of our solution is the mapping of compute units onto suitable routing constraints that can be taken as input during the ingress forwarding decision, i.e. the scheduling decision. This routing constraints are used for scheduling a packet at an ingress semantic router to one of the possible many service instances.

For this, we assume the integration of the compute metric assignment in placement methods and service orchestration operations. In order to turn the compute unit assignments into routing constraints, the service orchestration flattens and joins the service instance-specific compute units into a compute vector for a specific service identifier that represents a set of service instances. The needed information for each service identifier, containing each SI's locator together with the number of compute units allocated per instance, is expressed as lower and upper sub-interval bounds in the compute vector. The reasoning behind the use of the interval-based method in the compute vector is further explained in the scheduling mechanism in the following paragraphs.

The compute vector then needs distribution to the network ingress points to perform suitable scheduling operations together with the respective locator information for each service instance for the given service identifier. Key here is that this vector is seen as being rather stable since it is part of the overall service deployment and placement of service instances. Hence, any change will likely happen infrequently only, if at all during the service lifetime. As a consequence, extensions to existing routing protocols, to distribute the computing vector among all routers, will unlikely cause much additional overhead to the routing protocol performance. As an alternative, a service management system may directly signal the routing information to the ingress semantic routers only.

Once an ingress SAR receives a service request, after checking for a routing table entry for the service identifier provided in the request, the suitable next hop (or service instance destination) is

selected through a weighted round robin, with the weights being the compute unit for the service instance in the compute vector of the service identifier.

In order to avoid the need for implementing multiplications for the weights (i.e. compute units) at the scheduling decision at ingress SA, we assume that compute units are distributed as sub-intervals instead, with the total interval length being the sum of the compute units (each sub-interval equals one compute unit) of all the available service instances for the service identifier. This flattening of the weights into a vector allows for realizing the weighted round robin through a simpler counter, k, that cycles through that interval for any new service request that arrives at the semantic router. For every new increment of the counter, or wrap-around once the end of the complete interval vector is reached, the scheduling operations retrieve the next hop, i.e. service instance destination information, for the current counter and stores its new value in the routing table to be used for the next arriving request. Each semantic router chooses a random initial value for k, therefore increasing the randomness between individual semantic routers.

The needed scheduling operations are limited to a routing table lookup and a cycling of a counter over an interval (stored as part of the routing table). Technologies such as P4 [28] can be used for realizing such operations at line speed. Using structured binary names for the service identifier in our system allows for utilizing existing longest-prefix match operations to determine the suitable interval in our operations, while increment operations over such interval can be directly realized through P4 operations.

[27] provide comprehensive evaluations of the proposed solution, focusing in particular on comparing it with alternative viable scheduling solutions, such as random scheduling and the solution presented in [29]. Without discussing the details of the experimental evaluation setup, it suffices here to mention that CArDS brings benefits by significantly reducing request completion times in high load settings, e.g. those with above 1300 clients. This is shown in Figure 60.

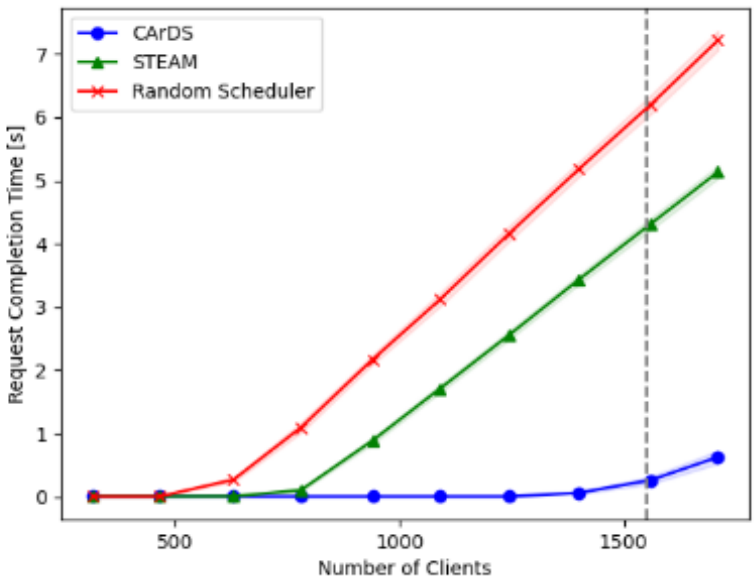



*Figure 60: Runtime scheduling performance*

## 9.6 References

[1] M. Al-Fares, A. Loukissas, and A. Vahdat, “A Scalable, Commodity Data Center Network Architecture,” in Proceedings of the ACM SIGCOMM 2008 Conference on Data Communication, ser. SIGCOMM ’08. New York, NY, USA: ACM, 2008, pp. 63–74. [Online]. Available: http://doi.acm.org/10.1145/1402958.1402967

[2] M. Behringer (Ed.), B. Carpenter, T. Eckert, L. Ciavaglia, and J. Nobre, "A Reference Model for Autonomic Networking," RFC 8993 (Informational), RFC Editor, Fremont, CA, USA, May 2021. [Online]. Available: https://www.rfc-editor.org/rfc/rfc8993.txt

[3] D. Black, J. Hudson, L. Kreeger, M. Lasserre, and T. Narten, "An Architecture for Data-Center Network Virtualization over Layer 3 (NVO3)," RFC 8014 (Informational), RFC Editor, Fremont, CA, USA, Dec. 2016. [Online]. Available: https://www.rfceditor.org/rfc/rfc8014.txt

[4] R. Balay, D. Katz, and J. Parker, "IS-IS Mesh Groups," RFC 2973 (Informational), RFC Editor, Fremont, CA, USA, Oct. 2000. [Online]. Available: https://www.rfc-editor.org/rfc/rfc2973.txt

[5] M. Behringer, M. Pritikin, S. Bjarnason, A. Clemm, B. Carpenter, S. Jiang, and L. Ciavaglia, "Autonomic Networking: Definitions and Design Goals," RFC 7575 (Informational), RFC Editor, Fremont, CA, USA, Jun. 2015. [Online]. Available: https://www.rfceditor.org/rfc/rfc7575.txt

[6] M. Caesar, M. Castro, E. B. Nightingale, G. O'Shea, and A. Rowstron, "Virtual Ring Routing: Network Routing Inspired by DHTs," SIGCOMM Comput. Commun. Rev., vol. 36, no. 4, pp. 351–362, Oct. 2006. [Online]. Available: https://doi.org/10.1145/1151659.1159954

[7] B. Dumba, H. Mekky, S. Jain, G. Sun, and Z.-L. Zhang, "A Virtual Id Routing Protocol for Future Dynamics Networks and Its Implementation Using the SDN Paradigm," Journal of Network and Systems Management, vol. 24, no. 3, pp. 578–606, Jul. 2016. [Online]. Available: https://doi.org/10.1007/s10922-016-9373-0

[8] T. Eckert (Ed.) and M. Behringer, "Using an Autonomic Control Plane for Stable Connectivity of Network Operations, Administration, and Maintenance (OAM)," RFC 8368 (Informational), RFC Editor, Fremont, CA, USA, May 2018. [Online]. Available: https://www.rfceditor.org/rfc/rfc8368.txt

[9] T. Eckert (Ed.), M. Behringer (Ed.), and S. Bjarnason, "An Autonomic Control Plane (ACP)," RFC 8994 (Proposed Standard), RFC Editor, Fremont, CA, USA, May 2021. [Online]. Available: https://www.rfc-editor.org/rfc/rfc8994.txt

[10] ETSI Industry Group Specification, "ETSI GS NFV 002 V1.2.1 Network Functions Virtualisation (NFV); Architectural Framework," Dec. 2014.

[11] B. Ford, "Scalable Internet Routing on Topology-Independent Node Identities," Massachusetts Institute of Technology, Tech. Rep. Technical Report MIT-LCS-TR-926, Oct. 2003. [Online]. Available: https://dspace.mit.edu/handle/1721.1/30432

[12] B. Ford, "Unmanaged Internet Protocol: Taming the Edge Network Management Crisis," SIGCOMM Comput. Commun. Rev., vol. 34, no. 1, pp. 93–98, Jan. 2004. [Online]. Available: http://doi.acm.org/10.1145/972374.972391

[13] M. Goyal (Ed.), E. Baccelli, M. Philipp, A. Brandt, and J. Martocci, "Reactive Discovery of Point-to-Point Routes in Low-Power and Lossy Networks," RFC 6997 (Experimental), RFC Editor, Fremont, CA, USA, Aug. 2013. [Online]. Available: https://www.rfceditor.org/rfc/rfc6997.txt

[14] S. Jain, Y. Chen, Z.-L. Zhang, and S. Jain, "Viro: A scalable, robust and namespace independent virtual id routing for future networks," in 2011 Proceedings IEEE INFOCOM. Piscataway, NJ, USA: IEEE, Apr. 2011, pp. 2381–2389.

[15] T. Li, P. Psenak, L. Ginsberg, T. Przygienda, D. Cooper, L. Jalil, and S. Dontula, "Dynamic Flooding on Dense Graphs," Internet Draft draftietf-lsr-dynamic-flooding-09, Jun. 2021, work in progress. [Online]. Available: https://datatracker.ietf.org/doc/draft-ietf-lsr-dynamic-flooding/

[16] P. Maymounkov and D. Mazieres, "Kademlia: A peer-to-peer information system based on the xor metric," in Peer-to-Peer Systems, P. Druschel, F. Kaashoek, and A. Rowstron, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 53–65.

[17] J. Moy, "OSPF Version 2," RFC 2328 (Internet Standard), RFC Editor, Fremont, CA, USA, Apr. 1998, updated by RFCs 5709, 6549, 6845, 6860, 7474, 8042. [Online]. Available: https://www.rfceditor.org/rfc/rfc2328.txt

[18] A. Singla, P. B. Godfrey, K. Fall, G. Iannaccone, and S. Ratnasamy, “Scalable Routing on Flat Names,” in Proceedings of the 6th International Conference, ser. CoNEXT ’10. New York, NY, USA: ACM, 2010, pp. 20:1–20:12. [Online]. Available: http://doi.acm.org/10.1145/1921168.1921195

[19] J. Tripathi (Ed.), J. de Oliveira (Ed.), and J. Vasseur (Ed.), “Performance Evaluation of the Routing Protocol for Low-Power and Lossy Networks (RPL),” RFC 6687 (Informational), RFC Editor, Fremont, CA, USA, Oct. 2012. [Online]. Available: https://www.rfc-editor.org/rfc/rfc6687.txt

[20] R. White and M. Aelmans, “Recent Developments in Link State on Data-Center Fabrics,” Internet Protocol Journal, vol. 23, no. 2, pp. 2–19, sep 2020, https://ipj.dreamhosters.com/.

[21] T. Winter (Ed.), P. Thubert (Ed.), A. Brandt, J. Hui, R. Kelsey, P. Levis, K. Pister, R. Struik, J. Vasseur, and R. Alexander, “RPL: IPv6 Routing Protocol for Low-Power and Lossy Networks,” RFC 6550 (Proposed Standard), RFC Editor, Fremont, CA, USA, Mar. 2012, updated by RFCs 9008, 9010. [Online]. Available: https://www.rfc-editor.org/rfc/rfc6550.txt

[22] Y. Wei, Z. Zhang, D. Afanasiev, P. Thubert, T. Verhaeg, and J. Kowalczyk, “RIFT Applicability,” Internet Draft draft-ietf-riftapplicability-06, May 2021, work in progress. [Online]. Available: https://datatracker.ietf.org/doc/html/draft-ietf-rift-applicability-06

[23] H2020 Daemon, "D2.1 - Initial report on requirements analysis and state-of-the-art frameworks and toolsets", June 2021: https://doi.org/10.5281/zenodo.5060978

[24] F. D. Muñoz-Escoí, R. de Juan-Marín, J. García-Escrivá, J. R. González de Mendívil and J. M. Bernabéu-Aubán, "CAP Theorem: Revision of Its Related Consistency Models," The Computer Journal, vol. 62, no. 6, June 2019.

[25] M. Gramaglia, P. Serrano, A. Banchs, G. Garcia-Aviles, A. Garcia-Saavedra and R. Perez, "The case for serverless mobile networking," 2020 IFIP Networking Conference (Networking), 2020, pp. 779-784.

[26] Luca Cominardi, Robert Andres, Kilton Hopkins, Frédéric Desbiens. From devops to edgeops: A vision for edge computing. Tech. Rep. Eclipse Foundation, Edge Native Working Group, April 2021.

[27] Karima Saif Khandaker, Dirk Trossen, Ramin Khalili, Zoran Despotovic, Artur Hecker, Georg Carle. CArDS: Dealing a New Hand in Reducing Service Request Completion Times. IFIP Networking 2022.

[28] "P4 Language and Related Specifications". Available online https://p4.org/p4-spec/docs/P4-16-v1.2.0.html, Retrieved 2 December 2019.

[29] Blöcher, M., Khalili, R., Wang, L. and P. Eugster, "Letting off STEAM: Distributed Runtime Traffic Scheduling for Service Function Chaining", IEEE INFOCOM 2020 - IEEE Conference on Computer Communications, DOI: 10.1109/INFOCOM41043.2020.9155404.

[30] Roland Bless; Martina Zitterbart; Zoran Despotovic; Artur Hecker, “KIRA: Distributed Scalable ID-based Routing with Fast Forwarding”, 2022 IFIP Networking Conference (IFIP Networking).

[31] H. T. Kaur et al., “BANANAS: An Evolutionary Framework for Explicit and Multipath Routing in the Internet,” in Proc. of the ACM SIGCOMM FDNA Workshop. ACM, 2003.

# 10. Sustainability

In its 2030 agenda for sustainable development, UN proposed 17 integrated Sustainable Development Goals (SDGs) spanning sustainability across social, economic and environmental aspects [1]. UN also considers mobile telecommunication systems a fundamental part of the society and notably expects it to play a key role in attaining its SDG goals. The number of connected devices is a good indicator for this: by 2030, the number of connected IoT devices is expected to be 24.1 billion, raising the resulting revenue from the IoT market by up to 1.5 trillion [2]. For the UN, mobile networks will play a crucial role for enabling global coverage, low latency high data-rate transmissions, promoting clean energy usage, developing sustainable cities, and infrastructures and efficient resource usage to ensure responsible consumption and production. Generally, ICT sector can contribute towards decarbonization by improving the overall network energy consumption, use of renewable energy, energy-efficient equipment, infrastructure and services, circular economy, waste management, and help develop mechanisms for environmental impact analysis, etc. [5].

A similar initiative has recently emerged in the European Union. Green Deal / Sustainable Product Initiative [3] is the recent set of initiatives of the European Commission that defines challenging political targets for the EU to reach climate neutrality by 2050. It introduces several action plans, e.g., regarding clean energy (energy efficiency and renewable energy production), sustainable industry, biodiversity, building renovation, etc. It also introduces the principles of Circular Economy, aimed at waste reduction, reparability, reuse and recycling promotion. The Circular Economy involves all actors along supply chains of products sold in the EU, going towards "product-as-a-service" principle, hereby assigning each vendor the responsibility for the whole lifecycle of the offered products. At the heart of the Circular Economy lies the principle of Product Ecodesign. In this vein, as part of its Sustainable Product Initiative (SPI), the EC has already established its "Digital Product Passport" (DPP), meant to gather ecologically relevant data on all products sold in the EU along their respective supply and value chains. Mobile telecommunications and ICT in general play an important role in the European Green Deal too. The key to understand this is the role of telecom operators as providers of ICT solutions that contribute to both energy and material efficiency [4]. Potential scope of these solutions being almost endless, it suffices here to mention just a number of them: smart logistics (where appropriate technological solutions may enable important fuel consumption reduction, e.g. through optimal route management) vehicle maintenance and driver behaviour, delivering cuts in fuel consumption of up to 30%), grids (in which smart meters enable both energy provider, authorities and households to optimally use the energy and thus reduce its consumption) [4], farming (sensors giving farmers the power to measure and record data such as physical, chemical, microbial soil analysis, pest presence, satellite sensing information and data from IoT sensors such as soil moisture probes.

The two mentioned initiatives were followed by a set of national or regional initiatives with similar objectives and got translated into concrete work items defined by governments, project funding institutions, various industry associations (e.g. GSMA [4], NGMN [5]), etc. An obvious question of interest for this white paper is how are all these initiatives relevant to the information and communication technology in general and 6G in particular? Although we hinted at answers above, let us once again address it and state the answers clearly. We can look for them along two directions. The first one, we name it "sustainability of ICT", targets decarbonization of ICT and mobile networks as such, i.e. creating networks that are more energy efficient and deliver services with a lower carbon footprint posture. Given that, currently, the ICT sector contributes to 1.4% of the global greenhouse gas emission [12], which primarily stems from the electricity consumed by resources during the end-to-end service operation [13], one can ask if further ICT decarbonization can bring us any closer to the UN set goals. However, keeping in mind the above described role of ICT in the foreseen future, potential reductions of greenhouse gas emission within ICT will lead to much higher global effect, far beyond 1.4%. This is where the other direction, we name it "ICT for sustainability" comes into play. Through the continuation of the current trend of "digitization" of conventional services and the role of ICT in the future society (remember the figures from the beginning of this section) we expect that mobile telecoms will become more and more pivotal to the carbon footprint reduction of our

society at large. It is thus paramount to pursue the research on these aspects within ICT, i.e. investigate sustainability of ICT, even though the current effect of reduction appears relatively small. In this sense, a recent study [11] can be seen as an attempt to quantify this claim. According to [11], it is expected that by 2030, about 15% of the greenhouse gas emission can be reduced with 5G only. Keeping in mind that 5G was not built with explicit energy-efficiency principles (although many requirements are being introduced and features added), we can expect stronger and farther-reaching effect of 6G, if that technology is built from the outset as energy aware and efficient.

## 10.1 Use Cases

In this section, we first make a quick illustration of what methods can be used in 6G to enable energy consumption reduction. We do it with help of two carefully selected use cases. The first one introduces changes in how services are offered to the consumers by explicit inclusion of energy consumption in the SLAs. The second one illustrates opportunities offered to the operators if they monitor energy consumption of their infrastructure real-time and optimize usage thereof.

### 10.1.1 Use case 1: Exposing Dynamic Energy Consumption to Subscribers

MNO networks are the closest hop for end-users during end-to-end service delivery. Hence, application service providers such as e.g., a streaming service provider like NetFlix, YouTube, etc., would benefit by leveraging the edge-cloud offered by an MNO. In addition, the external 3rd parties also want to optimize/control/report the energy consumption of their service offerings to their service consumers. Therefore, this use case considers a scenario where an external service provider has deployed his service-instances in an MNO's edge-cloud and as a result would like to know the energy consumption of his service-instances running in the MNO's network.

This is important for external service providers for not only exposing the EC information to their end-users but also to be able to estimate their own ecological impact, most notably in terms of CO2e production. This, however, depends on the actual server location, server's electrical power-mix and wireless usage. In other words, the business partner would like to know the energy consumption (EC) of service-related sessions end-to-end (E2E).

On the other hand, through exposing the required EC information, MNO's do not feel responsible for the (additional) ecological impact caused by this particular service-instance and can discount this from their own energy-footprint.

To safeguard their respective ecological footprint targets, the operations related to this service can be subject to energy-based limits like "maximum amount of energy to be consumed" per session, per user, per day, by a service instance, by the whole service, etc. When the energy consumption of the service increases beyond the set limit, the throughput and availability of the service can be affected.

The power-mix of the particular server could also change over time, because of external factors (availability in the electrical grid, weather, wind). Eco-aware subscribers who prefer services with the lowest CO2e footprint should be able to make informed decisions through the EC information exposure.

### 10.1.2 Use case 2: Energy-aware Service Execution

As we know, future wireless mobile networks like 6G will provide various services to billions of globally distributed users with sophisticated (future) mobile end-devices from applications like Artificial Reality (AR)/Virtual Reality (VR), holographic communications, etc., with stringent performance requirements. Hence, multiple different services will co-exist in 6G. Since different services have different underlying service requirements, one or more services could end up

monopolising the available resources. As a result, different entities in the network will experience different types of traffic, resulting in fluctuations in their energy consumption.

Therefore, energy-aware traffic distribution over all available links and participating networking nodes is crucial for improving energy efficiency of networks; in addition to improving the Quality of Experience (QoE) of users and resource efficiency of networks. Traffic steering employed in todays' networks focus towards achieving a more uniform distribution of traffic on all available links in order to avoid/address congestion. Today, many different types of traffic steering mechanisms exist, usually based on the characteristics used for traffic distribution like link quality, load balance, bandwidth capacity, etc. Moreover, SoTA Quality of Service (QoS) and Policy frameworks also utilize Traffic Engineering to improve the latency and reliability of the network. However, the traffic engineering solutions should be enhanced to incorporate energy consumption of the network as a key additional feature during traffic steering decisions to ensure energy is consumed efficiently in the network and the overall energy consumption of the network is controlled. Moreover, the services must be scheduled on resources in the network by considering their energy profile and their current operating conditions to ensure all resources are operating at optimal capacity under required energy limits.

## 10.2 Industry initiatives and standardization

### 10.2.1 NGMN

NGMN was among first industry associations to understand the importance of the sustainability topic and deliver concrete work on it. Running through three phases as of now, the "Sustainability/Green Networks" project of NGMN started in 2021 and resulted in a number of important white papers. [5] sets the stage by introducing the subject and explaining its context (e.g. UN SDG and EU Green Deal). It also touches upon the important related background concepts such as sustainability Key Performance Indicators (KPIs), role of renewable energy, etc. [6] narrows down the focus to energy efficiency of mobile networks and discusses ways to potentially even further increase the efficiency of future mobile systems by utilizing various energy saving features and more efficient hardware and network architecture.

[7] promotes real-time energy consumption metering as an essential feature that will lead to more energy efficient networks. Metering is seen not only as a tool to better understand the energy usage in the network, but also as an enabler of more sophisticated techniques saving such as optimal usage of the available energy mix (e.g. use of green energy where it is most needed). [8] discusses challenges related to sustainability of supply chains in mobile networks, paying particular attention to the emerging regulations and standards. The focus thus departs from mobile networks in operation to the equipment suppliers and their sustainability practices in making and delivering the equipment. A checklist of best practices in the industry is provided to support operators in developing a sustainable procurement strategy. [9] outlines a set of KPIs, along with a method for merging the KPIs into an overall measure. Two sets of KPIs are seen critical, throw related to energy and the conventional ones, related to quality of service (or experience). The operators' success in delivering services is measured both in terms of how well each end every individual KPIs is addressed, but also what trade-off are made when some KPIs cannot be made.

[10] discusses further potential approaches to saving energy in mobile networks. They are organized into three broad categories: 1) process optimizations, 2) engineering optimizations, and 3) new technologies. Process optimizations include approaches (many of them being AI/ML driven) to optimally operate radio sites, which are seen as the main energy consumers in the mobile network. Engineering optimizations go deeper into the architecture of radio access sites and propose new methods to make more efficient radio units, antennas, etc. New technologies include various emerging research initiatives with promising energy related performance, e.g. distributed MIMO or reflective intelligent surfaces (RIS).

### 10.2.2 ETSI

ETSI (European Telecommunications Standards Institute) with its Technical Committee (TC) Environmental Engineering (EE) since long (before the UN SDG and EC Green Deal) defines standards related to "environmental aspects of telecommunication infrastructures and equipment". Whereas the focus of this technical committee is wide and involves as various and detailed matters as environmental conditions (e.g. climatic and thermal) in which equipment operates or measurements are performed or power interface specifications for ICT equipment, we are here interested primarily in its standards related to "environmental matters associated with mobile Information and Communications Technologies (ICT) devices".

Of highest relevance to this white paper are the following ETSI standards: ES 203 228 – Assessment of mobile network energy efficiency [19], ES 202 706 – Metrics and measurement method for energy efficiency of wireless access network equipment. The latter comes in two parts, Part 1: Power consumption - static measurement method [20] and Part 2: "Energy Efficiency - dynamic measurement method [21].

ES 203 228 defines metrics for energy efficiency of mobile networks, parameters that play a role in determining those metrics, as well conditions in which concrete measurements should be performed. Three energy efficiency metrics were defined: 1) data Energy Efficiency (EEMN,DV), 2) coverage Energy Efficiency (EEMN,CoA), and 3) latency Energy Efficiency (EEMN,L). Data Energy Efficiency is defined as the ratio between the data volume (DVMN) and the Energy Consumption (ECMN) when assessed during the same time frame:

$$EE_{MN,DV} = \frac{DV_{MN}}{EC_{MN}}$$

where ECMN is the mobile network Energy Consumption (ECMN) and is calculated as the sum of the energy consumption of each equipment included in the mobile network under investigation. DVMN is the data volume delivered by the equipment of the mobile network under investigation during the time frame of the energy consumption assessment. Thus, ECMN and DVMN are the parameters relevant for the calculation of the data Energy Efficiency metrics. EEMN,DV is expressed in bit/J.

Similar approach applies to coverage EEMN,CoA and EEMN,L. For example, mobile network coverage Energy Efficiency (EEMN,CoA) is the ratio between the area covered by the mobile network under investigation and the energy consumption (ECMN) when assessed during one year.

As for the conditions and procedures for energy efficiency measurements, ES 203 228 provides many details on timing and setup of the measurements.

ETSI ES 202 706 (both Part 1 and Part 2) starts with the assumption that the energy consumption of the access network is dominating the energy consumption of other subsystems of the wireless telecom networks and defines the measurement method for the evaluation of base station power consumption and energy consumption. ETSI ES 202 706-1 (Part 1) [20] focuses on defining methods for average power consumption of BS equipment under static test conditions, i.e. when the BS is loaded artificially in a lab environment with three levels of load, namely low, medium and busy hour. All radio access technologies currently in operation are treated in the document. ETSI ES 202 706-2 (Part 2) [21] repeats this exercise for dynamic loads and all radio technologies but 5G. ETSI TS 103 786 [22] fills in the missing piece, namely 5G under dynamic load. Various (dynamic) data traffic models are used to describe the loads. Under a specific data traffic model, a set of UEs are downloading files of various sizes, while roaming around a base station at three different locations (near, middle and far range). Three levels of load, low, medium and busy hour (named just as in the static case) are determined by an appropriate combination of parameters in the traffic model (e.g. waiting time between two consecutive downloads, total number of UEs, etc.).

### 10.2.3 3GPP

Energy Efficiency (EE) efforts in 3GPP have been initially undertaken by SA5 (System Architecture Working Group 5), which specified EE features aiming at monitoring and improving energy efficiency of the 5G System by impacting the operations of the system. 3GPP SA5 introduced EE KPIs in TS 28.554 Release 17 and PEE (Power, Energy and Environment) measurements in TS 28.552 Release 18.

In December 2021, 3GPP adopted Energy Efficiency as a guiding principle for the design of the 3GPP 5G System Architecture in Release 18, the first release for 5G-Advanced [23].

The biggest contribution to energy consumption in a mobile network is provided by the RAN (Radio Access Network) and specifically by the radio transmission/reception. 3GPP RAN Working Group 1 (RAN1) addressed network energy savings in Rel-18 with a study item on "Network Energy Savings for NR" followed by a work item [24]. During the study phase, RAN1 defined a base station energy consumption model along with necessary evaluation methodology and KPIs. Using the model and evaluation methodology, RAN1 studied and evaluated various techniques in time, frequency, spatial, and power domain. Based on the outcome of the study item, during the normative phase RAN1 specified enhancements for network energy savings in spatial/power domain and enhancements on cell DTX/DRX.

In Release 19, use cases and requirements for "Energy Efficiency as service criteria" were consistently addressed in 3GPP SA WG1 by a stage 1 study that was completed in December 2023. Requirements on energy efficiency as service criteria, captured in TS 22.261, include enabling the 5G System to:

- Support subscription policies that define a maximum energy credit limit for services without QoS criteria, as well as means to associate energy consumption information with charging information based on subscription policies.
- Support different energy states of network elements and network functions, as well as dynamic changes of energy states of network elements and network functions.
- Support energy consumption monitoring per network slice and per subscriber granularity.
- Expose information on energy consumption to 3rd parties, subject to operator's policy and agreement with the 3rd party.

Based on these stage 1 requirements, stage 2 Working Groups in 3GPP started study/work items in 2024:

- SA WG2: Rel-19 "Feasibility Study on 5GS Enhancement for Energy Efficiency and Energy Saving - FS_CNEES" [26]. After completion of the study phase in June 2024, the normative phase should be completed by end-2024. The main objectives of this study include:
  - Studying a framework for network energy consumption exposure. This will include whether and what information is exposed, how it is exposed (e.g., charging) and at what granularity, e.g., at RAN level, Core Network level, network slice level, UE level, PDU session level, and/or QoS flow level.
  - Studying enhancement for subscription and policy control to enable network energy savings as service criteria.
  - Studying 5GS enhancements for network energy saving (e.g., energy usage adjustment for NF from CN aspect, energy saving related decision making, NF selection leveraging NF energy states), including 5GC NFs and NG-RAN interactions, analytics, etc.
- RAN WG1: Rel-19 WI "Enhancements of network energy savings for NR" [27]. The objectives of the work item are:

  - Specify procedures and signalling method(s) to support on-demand SSB SCell operation for UEs in connected mode configured with CA, for both intra-/inter-band CA. Specify triggering method(s): select from UE uplink wake-up-signal using an existing signal/channel, cell on/off indication via backhaul, SCell activation/deactivation signalling.
  - Study procedures and signalling method(s) to support on-demand SIB1 for UEs in idle/inactive mode.
  - Specify adaptation of common signal/channel transmissions, such as: adaptation of SSB in time domain (e.g. adapting periodicity), adaptation of PRACH in time domain, study adaptation of PRACH in spatial domain.

- SA WG5: Rel-19 "Study on energy efficiency and energy saving aspects of 5G networks and services - FS_Energy_OAM_Ph3" [28]. This study item addresses:
  - Energy consumption measurement/estimation at various granularities
  - Feasibility of mapping measured or estimated energy consumption measurements to carbon emissions.
  - Support of different energy states of network elements and network functions
  - OAM support to new Rel-19 RAN energy saving features.

While Release 19 stage 2 study/work items are progressing, SA WG1 started working on Release 20 requirements for 5G Advanced in the "Study on Energy Efficiency as Service Criteria Ph2 - FS_EnergyServ_Ph2" [29]. Expected completion date is March 2025. The objectives are:

- Information exposure of energy-related characteristics of the network for the communication service (i.e. energy consumption, energy supply mix, carbon footprint, energy capacity and availability conditions) to authorized users or authorized 3rd parties.

- Potential dynamic adjustments of the delivered communication service from 5G system perspective (including service performance adjustments) resulting from the changes of energy-related characteristics of this service.

### 10.2.4 ITU

ITU as a whole and its Telecommunication Standardization Sector (ITU-T) and Radiocommunication Sector (ITU-R) have a long history of defining directions of future development of telecommunications and creating concrete standards and recommendations that mark important milestones in this development. What is of interest to us in this section are, to a lesser degree, the work of so-called Study Group 5 (SG5) of ITU-T on "electromagnetic fields (EMF), environment, climate action, sustainable digitalization, and the circular economy"2 and, to a greater degree, a number of recent research studies of ITU that provide valuable insights into the problems and potential solution related to ICT sustainability.

ITU-T organizes its recommendations and standards in series, L series being the most relevant to us ("L series: Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant"). In particular, the range of recommendations L.1300-L.1399: "Energy efficiency, smart energy and green data centres" as well as L.1400-L.1499: "Assessment methodologies of ICTs and CO2 trajectories" and L.1700-L.1799: "Low cost sustainable infrastructure" are most important. Whereas detailed descriptions of all these recommendations are neither possible nor necessary here, we describe just a couple of them as examples.

L.1470 [14] provides detailed trajectories for GHG emissions for the ICT sector and sub-sectors that are quantified for the baseline year 2015 and estimated for 2020 (the year of publishing of L.1470),

[2] *https://www.itu.int/en/ITU-T/studygroups/2022-2024/05/Pages/default.aspx*

2025 and 2030. Further, L.1470 presents a long-term goal for 2050. All trajectories and the baseline have been derived in accordance with Recommendation ITU-T L.1450 [15]. L.1450, thus, defines the methodology for calculating the ICT sector footprint with respect to life cycle greenhouse gas emissions.

The ITU-T Recommendation L.1333 [17] defines a key performance indicator (KPI) called network carbon intensity energy (NCIe) and elaborates on ways to calculate it for different types of networks, such as public telecom network (PTN), non-public network (NPN) and enterprise network. The standard only considers the network operation phase and the energy efficiency metric defined in the Recommendation ITU-T L.1331 [16]. (L.1331 coincides with ETSI ES 203 228 [19], i.e. they are the same.)

[18], a common recent publication of the World Bank and ITU, brings further insights into GHG emissions of the ICT sector. First, it raises the estimate of the total ICT sector GHG emission to 1.7%. (based on more comprehensive data than [12]). Further, it breaks down this emission to various technological and geographical areas sectors, offering thus a better view where the focus of future initiatives for reduction should lie. First, the emissions related to equipment use are dominating the emissions related to production thereof. On the equipment use side, telecommunications operators are the strongest GHG contributors, dominating the data centres. However, whereas the GHG emissions of the telecommunications operators seem stable over the three measured years (2020 – 2022), that of the data centres records a substantial increase (around 30%). On the consumer side, personal computers are the leading GHG contributors, whereas smartphones are far behind. If we are to summarize the data, the following three recommendations apply: 1) reduction of the absolute GHG emission in the telecommunications sector and in the 2) PC use, as well as 3) reverting the trend of data centre emission growth.

Geographically, Asia is by far the highest contributor of GHG emissions, contributing more than twice as much as Americas and the EMEA (Europe, Middle East and Africa) region.

## 10.3 Research Initiatives

### 10.3.1 EU Projects

#### BeGREEN

BeGREEN [33] is an SNS–JU Phase I project with an exclusive focus on the radio network. The project vision is to evolve radio networks by explicitly considering the power consumption as a radio network design factor. It starts by questioning the energy efficiency of 5G radio and considers bringing in the latest cutting-edge technologies, such as AI, as explicit engineering choices for 6G radio design.

BeGREEN plans to evaluate a range of mechanisms to help reducing power consumption. The most important are: 1) Massive MIMO RAN design, with the hope to achieve flexible and efficient connectivity and spectrum utilization; 2) Various radio-unit controlling schemes; 3) Integrated sensing techniques to provide a better estimate of the impact of the radio channel.

The project also targets improvements at the system level, i.e. how the network as a whole operates with respect to energy and power consumption. AI-based procedures to adapt the energy consumption of softwarised network functions are at the core of the approach taken towards this.

#### 6GREEN

The ultimate goal of the 6GREEN project [34] is to enable reduction of the carbon footprint of 5G or 6G networks by a factor of 10 or more. As the main means towards achieving this goal, the projects will focus on cloud-native technologies and appropriately extend the B5G Service-Based Architecture with new cross-domain enablers. Algorithmic improvements of the energy

consumption and usage will be based on sophisticated Artificial Intelligence mechanisms that will allow propagating the environment impact to any players acting on virtual domains, and triggering zero-touch operations according to customizable energy- and carbon-aware policies, allowing to optimally use renewable energy sources. The project will deliver three future-proof use-case applications that will provide a diverse awareness of the green capabilities of the 6Green platform.

### Exigence

Exigence [35] is an SNS-JU project with the focus on a number of novel methods to enable energy consumption reduction. First, it puts energy metering in focus, i.e. direct energy consumption measurements. This is to contrast with indirect measurements, which have been so far established as a conventional method of energy consumption attribution. These assume that a user's portion of energy consumed in a physical node (e.g. a base station) is derived by simply measuring that total energy consumed by the node and the portion of data volume transferred for the user by that physical node. Exigence sees this as a source of inaccuracies as the actual energy consumption is a more complex function of various QoS parameters of the traffic handled by that node. Second, equipped with accurate energy consumption data that pertain to users, their flows, services, etc., Exigence plans to develop incentive mechanisms that will internalize the energy consumption and lead to its further reduction. An example is introducing energy caps in the service offerings and energy certificates that can be traded in an open market, in which unused energy claims can be sold. As its third pillar, Exigence sees real-time (runtime) energy consumption optimizations, e.g. driving green energy to the nodes in which it will bring the highest energy savings, etc.

### HEXA-X-II

Hexa-X-II [30] is an SNS–JU 6G Flagship project with a broad industry and academia participation. It aims at defining and to an extend realizing the 6G vision. Its focus is broader than just sustainability of mobile telecommunications networks, i.e. it tries to both investigate the enabling technologies that might be relevant to 6G and define (at least relevant pieces of) the system architecture. As of now, the main results of Hexa-X-II related to sustainability (which we could find in deliverables D1.1 [31] and D1.3 [32] of the project) are in form of recommendations for a holistic treatment of sustainability in the 6G design. Whereas D1.1 sets the stage by defining "sustainability guidelines for 6G design", D1.3 comes up with a little bit more technical details by analysing a set of relevant use cases (relevant in the sense that they present a consensus of a broad 6G community as of now) and presenting potential environmental implications thereof.

## 10.3.2 Academic Research

This section presents an overview of academic research, including a number of studies that incorporate the energy efficiency aspect for intent-driven mobile networks. As the network technology evolves, so does the density and complexity of mobile networks. Currently, mobile networks comprise a growing number of heterogeneous devices, including IoT sensors, smartphones, and vehicles. Concurrently, there is an ever-increasing demand for higher computational capacity and faster communication. In this complex and demanding environment, the integration of Artificial Intelligence (AI) has a significant role to play, as it is able to analyse the environment and adapt to the rapid changes, optimizing network performance, while accounting for the various limitations of different devices. Energy efficiency in wireless communication networks has been the main focus of a number of research studies, which have approached the problem from different angles by exploiting resource orchestration, computation offloading, and load balancing strategies.

A subset of research works is tackling the problem of energy efficiency, considering various computation tasks. In [36], the authors consider two computation offloading schemes for Mobile Edge Computing (MEC) systems, improving the energy efficiency for latency-constrained computation by optimizing the available resources. Zhang et al. in [37] introduce an improved Soft

Actor Critic (SAC) Reinforcement Learning (RL) algorithm towards a joint optimization of partial task offloading and resource allocation decisions. Another work [38] presents a knowledge plane-based MANO framework, using a twin-delayed double-Q SAC method towards energy consumption and Virtual Network Function (VNF) instantiation cost minimization. In [39], the authors propose a Q-learning and Double Deep Q Networks (DDQN)-based methods to determine the joint policy of computation offloading and resource allocation in a dynamic MEC system. AlQerm et al. in [40], propose an online RL methodology to determine the most energy efficient traffic offloading strategy.

There is a number of proposed solutions targeting to address energy-related challenges considering Federated Learning (FL) process as a use case, where devices perform a model training in a collaborative manner, exploiting locally stored datasets and avoiding any raw data transmissions. Several critical challenges have been raised related to the application of FL to wireless networks, significantly affecting the overall energy consumption of the involved devices. The authors in [41], [42] propose a Genetic Algorithm and a SAC RL approach, respectively, targeting the minimization of both the overall energy consumption of an FL process and any unnecessary resource utilization, by orchestrating the computational and communication resources of the involved devices, while guaranteeing a certain FL model performance target. A penalty function is also introduced that penalizes the strategies that violate the constraints of the environment, ensuring a safe learning process. The work in [43] presents a framework that targets to minimize the overall energy consumption of a Federated Edge Intelligence-supported IoT network, using the Alternate Convex Search algorithm. The authors in [44] propose a joint resource allocation scheme for FL in IoT, aiming at the minimization of the system and learning costs by jointly optimizing bandwidth, computation frequency, transmission power allocation and sub-carrier assignment. Mo et al. [45] focus on minimizing the total energy consumption of an FL system, by optimizing both communication and computational resources using techniques from convex optimization. Another work [46], proposes energy-efficient strategies for bandwidth allocation and scheduling, so as to reduce the energy consumption, while warranting learning performance in an FL framework. In [47] a bisection-based algorithm is proposed, whose goal is to minimize the total energy consumption of an FL system under a latency constraint. Zhang et al. [48] present an energy-efficient FL framework for Digital Twin-enabled Industrial IoT (IIoT) that exploits a DDQN, in order to jointly optimize training strategies and resource allocation, considering a dynamic environment. Another work [49], through the use of a Proximal Policy Optimization-based actor-critic method, targets the energy efficiency improvement of FL, by jointly minimizing the learning time and energy consumption. The authors in [50] exploit the merits of Multi-Agent Deep Deterministic Policy Gradient in resource allocation, addressing the challenges of FL in an Internet of Vehicles (IoV) scenario. Nguyen et al. [51] propose a Deep Q Learning (DQL) algorithm concerning the resource allocation, in a mobility-aware FL network, which aims to maximize the number of successful transmissions, while minimizing the energy and channel costs. The work [52] targets to optimize the performance of an FL model, using multi-agent RL, while considering energy-related requirements. The authors in [52] propose an RL solution for device selection at each communication round, towards FL performance optimization.

At the same time, a subset of research works, are expanding the scope of energy minimization by also including renewable sources in their solutions. Currently, there are several ways in producing energy through renewable sources. However, solar and wind energy, are the two most common sources of energy. In [53] a green MEC system with energy harvesting devices is investigated and an effective computation offloading strategy is developed. An Execution Cost Minimization (ECM) problem is composed and a Lyapunov Optimization-based Dynamic Computation Offloading (LODCO) algorithm is suggested to minimize the total cost derived from the execution delay and task failure. In [54] the target is to guarantee stable and effective services for numerous devices, under highly intermittent renewable generations and variable computation energy demands. The authors achieve system cost minimization by finding the optimal number of MEC servers, energy storage units, optimal size of local and offloaded tasks, the amount of battery (dis)charging rate and the utilized renewable energy of the micro Base Station (BS).

## 10.4 Power and Energy Measurement

Existing power and energy measurements approaches, recommended by standardization organizations, mainly address energy consumption and energy efficiency of individual domains (e.g., radio access network, fixed network, core network, data centres, etc.) rather than delve into finer granularity of each system. The latter would require attention to hardware and software components level, e.g., services and applications. The exercise may become even more complex in the case of measuring power and energy consumption for the entire end-to-end flow of each single service active in the system. The approach of measuring power and energy consumption (and efficiency) at the domain level may seems good enough for addressing power and energy consumption related challenges in general, while it would require much more sophisticated methods in case the question is about per service or per application energy consumption in virtualized environments (e.g., cloud-edge continuum). Additionally, an economic viewpoint perspective suggests that energy costs for ICT services are nowadays more or less externalities, as the end user is paying for the service without any significant dependence on how it is provided in terms of energy consumption. While there are strong initiatives proposing more granular approach to power and energy measurements in virtual environments, e.g., for 6G systems [34][35], it may be concluded that in order to reach a full environmental sustainability of the 6G and cloud-edge continuum systems in general, the real usage of resources would need to be considered [55].

Multiple power and energy measurement methods and tools are already available, including direct measurement of the power draw from the electric outlet, as well as more sophisticated methods measuring (or estimating) power/energy consumption on the level of processes in the virtualized environments. Within this context, we may mention following methods:

- power meter as an external device: usually plugged into the power outlet, thus directly measuring power drawn from the electric grid and monitoring total consumption of the hardware device. It is obvious that this method cannot provide energy measurements on the service or application level,

- interfaces provided by chipset or other components manufacturers: reports power/energy consumption of various power domains such as memory, CPU, GPU, entire chipset socket, etc. For that purposes, modern processors commonly provide Running Average Power Limit (RAPL) interfaces (e.g., Intel [56], AMD [57]), while Nvidia provides power/energy related data via its NVIDIA Management Library (NVML) monitoring and managing interface [58].

- power and energy modelling for virtualized environments: by taking into account several parameters, including the above mentioned RAPL- and NVML-based parameters (e.g., CPU power consumption, CPU utilization, thermal design power, DRAM utilization), power models calculate the power/energy consumption at multiple hardware and software levels. Commonly used methods for modelling power consumption include classical regression analysis and emerging methods like reinforcement learning [59].

Power and energy measurement tools that enable measurements at the application level are software based and they usually feature RAPL/NVML interfaces, power/energy models or combination of both approaches. Measurement tools supporting virtualized environments include solutions such as Scaphandre [60], Kepler [61] and PowerAPI [62].

In virtualized environments, there are, besides (customer) application processes, also processes that are required to enable servers and virtualized environment to run properly. Energy consumption of these processes can be seen as an indirect contribution to the total energy consumption of a respective application, which leads us to a problem of estimating the distribution of the total indirect energy consumption among virtualized applications. Since this is not the only challenge in virtualized environments it is, for the evaluation purposes, therefore necessary to know the methodology utilized by the tool used, as certain differences can be observed when comparing results measured by different tools. An extensive reference on software-based power measurement tools and comparing them can be found in [63].

## 10.5 Challenges

With regards to the above discussion, this section aims to summarize the challenges for sustainable operation of future telecommunication networks like 6G.

Separation of concerns: in today's 5G networks E2E EC and EE information cannot be retrieved dynamically for an ongoing session from the network, since energy-related data is collected from management layer, which does not manage sessions in the Core CP. In order to have any real-effect, the energy-related controlling responsibilities must be performed in the control plane.

Granularity of measured information: so far, in mobile networks, the management layer collects EC information. However, the granularity of the collected information is very large e.g., at slice or RAN-site level. This may be useful for fulfilling certain types of requests e.g., accounting at the end of a billing period. The challenge however is to switch the granularity of EC information to a level that corresponds to the nature of the request and pertaining to the ontology causing it.

Accuracy & Verifiability: for accounting, accurate measurements are necessary instead of approximate (e.g. model-based, statistics-based) estimations. When the EC data belongs to different authorities (e.g., different network domains, roaming cases, core network provisions through 3rd party platforms, etc.), the verifiability and auditability of the collected data becomes critical for accountability.

Eco-accounting: the energy consumption in the network is important however, the nature of the power source is also equally important. Measuring the EC and the resulting CO2e based on the energy-mix at the time of measurements requires new mechanisms that can measure and expose EC information at various granularity and time frames based on subscribers request and accurately show the correlation of EC and CO2e.

Volume & effort: real-time, per-session/per-flow EC data measurements across all network devices and domains could result in increased efforts for exchanging non-negligible volumes of data.

Over-exposure and privacy of the providing entity: the EC data measurements could over-expose details of the providing entity, considered confidential, such as details of the internal implementation.

Dynamicity: fluctuations in operating conditions such as increasing load, congestion, node failures, etc., also affect the energy consumed in the network. However, the subscribers have no way of knowing the network's EC state and the resulting EC they will be held accountable to.

## 10.6 References

[1] UN SDG Report 2019, Accessed: Apr. 12, 2023. [Online]. Available: https://unstats.un.org/sdgs/report/2019/The-Sustainable-Development-Goals-Report-2019.pdf

[2] Transforma Insights Research. (May 2020). Global IoT Market Will Grow to 24.1 Billion Devices in 2030, Generating $1.5 Trillion Annual Revenue. Accessed: Apr. 12, 2023. [Online]. Available: https://transformainsights.com/news/iot-market-24-billion-usd15-trillion-revenue-2030

[3] The European Green Deal, Accessed: Apr. 18, 2024. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en

[4] GSMA Position paper on the European Green Deal, Accessed on April 20, 2024. https://www.gsma.com/gsmaeurope/wp-content/uploads/2020/06/GSMA-position-paper-on-Green-Deal-June_2020.pdf

[5] NGMN Green Future Networks project. White paper: "Green Future Networks: Sustainability Challenges and Initiatives in Mobile Networks" v1.0, July 2021. Accessed: April, 2024. [Online]. Available: https://www.ngmn.org/wp-content/uploads/210719_NGMN_GFN_Sustainability-Challenges-and-Initiatives_v1.0.pdf

[6] NGMN Green Future Networks project. White paper: “Network Energy Efficiency”, December 2021. Accessed: Apr. 12, 2024. [Online]. Available: https://www.ngmn.org/wp-content/uploads/211009-GFN-Network-Energy-Efficiency-1.0.pdf

[7] NGMN Green Future Networks project. White paper: “Metering for Sustainable Networks”, January 2022. Accessed: Apr. 12, 2024. [Online]. Available: https://www.ngmn.org/wp-content/uploads/220125-GFN-Metering-White-Paper-v1.0.pdf

[8] NGMN Green Future Networks project. White paper: “Telco Supply Chain Sustainability”, January 2023. Accessed: Apr. 12, 2024. [Online]. Available: https://www.ngmn.org/wp-content/uploads/230117-NGMN-GFN_Supply-Chain-Sustainability_v1.0.pdf

[9] NGMN Green Future Networks project. White paper: “KPIs and Target Values for Green Network Assessment”, Febuary 2023. Accessed: Apr. 12, 2024. [Online]. Available: https://www.ngmn.org/wp-content/uploads/230117-NGMN-GFN_Supply-Chain-Sustainability_v1.0.pdf

[10] NGMN Green Future Networks project. White paper: “Network Energy Efficiency Phase 2”, October 2023. Accessed: Apr. 12, 2024. [Online]. Available: https://www.ngmn.org/wp-content/uploads/NGMN_Network_Engergy_Efficiency_Phase2.pdf

[11] J. Falk et al., Exponential Roadmap v1.5, Future Earth, Sweden, September 2019. Accessed: Apr. 12, 2023. [Online]. Available: https://exponentialroadmap.org/wp-content/uploads/2019/09/ExponentialRoadmap_1.5_20190919_Single-Pages.pdf

[12] Andrae, A. S. G. (2021). Internet’s handprint. Engineering and Applied Science Letters, 4, 80-97. Accessed: Apr. 12, 2023. [Online]. Available: https://pisrt.org/psr-press/journals/easl-vol-4-issue-1-2021/internets-handprint/

[13] Hannah Ritchie, Max Roser and Pablo Rosado, “Our World In Data: $CO_2$ and Greenhouse Gas Emissions”, August 2020. Accessed: Apr. 12, 2023. [Online]. Available: https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions

[14] ITU-T Recommendation L.1470: “Greenhouse gas emissions trajectories for the information and communication technology sector compatible with the UNFCCC Paris Agreement”. Accessed: Apr. 12, 2023. [Online]. Available: https://www.itu.int/rec/T-REC-L.1470-202001-I/en

[15] ITU-T Recommendation L.1450: “Methodologies for the assessment of the environmental impact of the information and communication technology sector”. Accessed: Apr. 24, 2024. [Online]. Available: https://www.itu.int/rec/T-REC-L.1450-201809-I/en

[16] ITU-T Recommendation L.1331: “Assessment of mobile network energy efficiency”. Accessed: Apr. 24, 2024. [Online]. Available: https://www.itu.int/rec/T-REC-L.1331-202201-I/en

[17] ITU-T Recommendation L.1333: “Carbon data intensity for network energy performance monitoring”. Accessed: Apr. 24, 2024. [Online]. Available: https://www.itu.int/rec/T-REC-L.1333-202209-I

[18] ITU and the World Bank: “Measuring the Emissions & Energy Footprint of the ICT Sector: Implications for Climate Action”. Accessed: Apr. 24, 2024. [Online]. Available: https://www.itu.int/hub/publication/d-ind-clim-2023-01/

[19] ETSI ES 203 228: "Environmental Engineering (EE); Assessment of mobile network energy efficiency".

[20] ETSI ES 202 706-1: "Environmental Engineering (EE); Metrics and measurement method for energy efficiency of wireless access network equipment; Part 1: Power Consumption – Static Measurement Method".

[21] ETSI TS 102 706-2: "Environmental Engineering (EE); Metrics and measurement method for energy efficiency of wireless access network equipment; Part 3: Power Consumption – Dynamic Measurement Method".

[22] ETSI TS 103 786: "Environmental Engineering (EE); Measurement method for energy efficiency

of wireless access network equipment; Dynamic energy performance measurement method of 5G Base Station (BS)"

[23] 3GPP SP-211621: "LS on Energy Efficiency as guiding principle for new solutions)"

[24] 3GPP RP-223540: "New WID on Network energy savings for NR"

[25] [3] 3GPP S1-221232, “Study on Energy Efficiency as service criteria”

[26] 3GPP SP-231192, “Feasibility Study on 5GS Enhancement for Energy Efficiency and Energy Saving”

[27] 3GPP RP-240170, “Enhancements of network energy savings for NR”

[28] 3GPP SP-231723, “Study on energy efficiency and energy saving aspects of 5G networks and services”

[29] 3GPP SP-240494, “Study on Energy Efficiency as Service Criteria Ph2 - FS_EnergyServ_Ph2”

[30] HEXA–X–II Project, https://hexa-x-ii.eu/

[31] HEXA–X–II Project, Deliverable D1.1, Accessed: Apr. 24, 2024. [Online]. Available: https://hexa-x-ii.eu/wp-content/uploads/2023/07/Hexa-X-II_D1.1_final-website.pdf

[32] HEXA–X–II Project, Deliverable D1.3, Accessed: Apr. 24, 2024. [Online]. Available: https://hexa-x-ii.eu/wp-content/uploads/2024/03/Hexa-X-II_D1.3_v1.00_GA_approved.pdf

[33] BeGREEN Project, https://www.sns-begreen.com/

[34] 6GREEN Project, https://www.6green.eu/

[35] Exigence Project, https://projectexigence.eu/

[36] X. Cao, F. Wang, J. Xu, R. Zhang and S. Cui, "Joint Computation and Communication Cooperation for Energy-Efficient Mobile Edge Computing," in IEEE Internet of Things Journal, vol. 6, no. 3, pp. 4188-4200, June 2019, doi: 10.1109/JIOT.2018.2875246.

[37] F. Zhang, G. Han, L. Liu, M. Martínez-García and Y. Peng, "Deep Reinforcement Learning Based Cooperative Partial Task Offloading and Resource Allocation for IIoT Applications," in IEEE Transactions on Network Science and Engineering, vol. 10, no. 5, pp. 2991-3006, 1 Sept.-Oct. 2023, doi: 10.1109/TNSE.2022.3167949.

[38] F. Rezazadeh, H. Chergui, L. Christofi and C. Verikoukis, "Actor-Critic-Based Learning for Zero-touch Joint Resource and Energy Control in Network Slicing," ICC 2021 - IEEE International Conference on Communications, Montreal, QC, Canada, 2021, pp. 1-6, doi: 10.1109/ICC42927.2021.9500265.

[39] H. Zhou, K. Jiang, X. Liu, X. Li and V. C. M. Leung, "Deep Reinforcement Learning for Energy-Efficient Computation Offloading in Mobile-Edge Computing," in IEEE Internet of Things Journal, vol. 9, no. 2, pp. 1517-1530, 15 Jan.15, 2022, doi: 10.1109/JIOT.2021.3091142.

[40] I. AlQerm and B. Shihada, "Energy Efficient Traffic Offloading in Multi-Tier Heterogeneous 5G Networks Using Intuitive Online Reinforcement Learning," in IEEE Transactions on Green Communications and Networking, vol. 3, no. 3, pp. 691-702, Sept. 2019, doi: 10.1109/TGCN.2019.2916900.

[41] L. Magoula et al., "A Safe Genetic Algorithm Approach for Energy Efficient Federated Learning in Wireless Communication Networks," 2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Toronto, ON, Canada, 2023, pp. 1-6, doi: 10.1109/PIMRC56721.2023.10293863.

[42] N. Koursioumpas et al., "A Safe Deep Reinforcement Learning Approach for Energy Efficient Federated Learning in Wireless Communication Networks," in IEEE Transactions on Green Communications and Networking, doi: 10.1109/TGCN.2024.3372695.

[43] Q. Wang, Y. Xiao, H. Zhu, Z. Sun, Y. Li and X. Ge, "Towards Energy-efficient Federated Edge

Intelligence for IoT Networks," 2021 IEEE 41st International Conference on Distributed Computing Systems Workshops (ICDCSW), Washington, DC, USA, 2021, pp. 55-62, doi: 10.1109/ICDCSW53096.2021.00016.

[44] J. Ren, J. Sun, H. Tian, W. Ni, G. Nie and Y. Wang, "Joint Resource Allocation for Efficient Federated Learning in Internet of Things Supported by Edge Computing," 2021 IEEE International Conference on Communications Workshops (ICC Workshops), Montreal, QC, Canada, 2021, pp. 1-6, doi: 10.1109/ICCWorkshops50388.2021.9473734.

[45] X. Mo and J. Xu, "Energy-Efficient Federated Edge Learning with Joint Communication and Computation Design," in Journal of Communications and Information Networks, vol. 6, no. 2, pp. 110-124, June 2021, doi: 10.23919/JCIN.2021.9475121.

[46] Zeng, Qunsong & Du, Yuqing & Leung, Kin & Huang, Kaibin. (2019). Energy-Efficient Radio Resource Allocation for Federated Edge Learning.

[47] Yang, Zhaohui & Chen, Mingzhe & Saad, Walid & Hong, Choong Seon & Shikh-Bahaei, M.. (2020). Energy Efficient Federated Learning Over Wireless Communication Networks. IEEE Transactions on Wireless Communications. PP. 1-1. 10.1109/TWC.2020.3037554.

[48] J. Zhang, Y. Liu, X. Qin and X. Xu, "Energy-Efficient Federated Learning Framework for Digital Twin-Enabled Industrial Internet of Things," 2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Helsinki, Finland, 2021, pp. 1160-1166, doi: 10.1109/PIMRC50174.2021.9569716.

[49] Y. Zhan, P. Li, L. Wu and S. Guo, "L4L: Experience-Driven Computational Resource Control in Federated Learning," in IEEE Transactions on Computers, vol. 71, no. 4, pp. 971-983, 1 April 2022, doi: 10.1109/TC.2021.3068219.

[50] Wang, Ge & Xu, Fangmin & Zhang, Hengsheng & Zhao, Chenglin. (2021). Joint resource management for mobility supported federated learning in Internet of Vehicles. Future Generation Computer Systems. 129. 10.1016/j.future.2021.11.020.

[51] Nguyen, Huy & Nguyen, Cong & Zhao, Jun & Yuen, Chau & Niyato, Dusit. (2020). Resource Allocation in Mobility-Aware Federated Learning Networks: A Deep Reinforcement Learning Approach. 1-6. 10.1109/WF-IoT48130.2020.9221089.

[52] Wang, Hao & Kaplan, Zakhary & Niu, Di & Li, Baochun. (2020). Optimizing Federated Learning on Non-IID Data with Reinforcement Learning. 1698-1707. 10.1109/INFOCOM41043.2020.9155494.

[53] Y. Mao, J. Zhang and K. B. Letaief, "Dynamic Computation Offloading for Mobile-Edge Computing With Energy Harvesting Devices," in IEEE Journal on Selected Areas in Communications, vol. 34, no. 12, pp. 3590-3605, Dec. 2016, doi: 10.1109/JSAC.2016.2611964.

[54] Y. Liu, S. Xie, Q. Yang and Y. Zhang, "Joint Computation Offloading and Demand Response Management in Mobile Edge Network With Renewable Energy Sources," in IEEE Transactions on Vehicular Technology, vol. 69, no. 12, pp. 15720-15730, Dec. 2020, doi: 10.1109/TVT.2020.3033160.

[55] R. Bolla, R. Bruschi, C. Lombardo and B. Siccardi, "6G Enablers for Zero-Carbon Network Slices and Vertical Edge Services," in IEEE Networking Letters, vol. 5, no. 3, pp. 173-176, Sept. 2023, doi: 10.1109/LNET.2023.3262861.

[56] Intel, Running Average Power Limit Energy Reporting, 202, Accessed: Jun. 7, 2024. [Online]. Available: https://www.intel.com/content/www/us/en/developer/articles/technical/software-security-guidance/advisory-guidance/running-average-power-limit-energy-reporting.html

[57] AMD, uProf User Guide, January 2024.

[58] NVIDIA, NVML Reference Manual, March 2024.

[59] W. Lin, F. Shi, W. Wu, K. Li, G. Wu, A.-A. Mohammed, “A Taxonomy and Survey of Power Models and Power Modeling for Cloud Servers,” in ACM Computing Surveys, Volume 53, Issue 5, Article No.: 100, pp. 1–41, 2023, https://doi.org/10.1145/3406208.

[60] Scaphandre documentation, Accessed: Jun. 7, 2024. [Online]. Available: https://hubblo-org.github.io/scaphandre-documentation/

[61] P. Singham, H. Chen, "Introducing Kepler: Efficient power monitoring for Kubernetes," Red Hat, Accessed: Jun. 7, 2024. [Online]. Available: https://next.redhat.com/2023/08/22/introducing-kepler-efficient-power-monitoring-for-kubernetes/

[62] PowerAPI, Accessed: Jun. 7, 2024. [Online]. Available: https://powerapi.org/

[63] M. Jay, V. Ostapenco, L. Lefevre, D. Trystram, A. -C. Orgerie and B. Fichel, "An experimental comparison of software-based power meters: focus on CPU and GPU," 2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing (CCGrid), Bangalore, India, 2023, pp. 106-118, doi: 10.1109/CCGrid57682.2023.00020.

# 11. Conclusions

6G is expected to make a breakthrough on how the mobile network services are delivered and consumed. It is supposed to enable a myriad of novel use cases, many of which come with a whole set of, often conflicting, requirements. Besides just enabling these use cases, 6G is supposed to be green, i.e., contribute to reducing overall energy consumption and achieving environmental sustainability. In addition to being user- and environment-friendly, 6G should also be operator-friendly, make possible easy development and deployment of new services, extensions of running ones, as well as interventions in the network itself.

This white paper reviewed a number of technologies that may play a pivotal role in 6G. Latest advances related to radio access and its accompanying technologies such as extensions in the range of higher frequencies, next generation MIMO, non-terrestrial networks and ISAC were discussed, state of the art ideas introduced and explained, open problems stated. Potential roles of artificial intelligence and machine learning in the context of networking were reviewed in a comprehensive and elaborate manner, as well as and advanced methods to realize novel use case, such as multimodal sensing and communication and various user plane enhancements, critical to realize the 6G vision. Network programmability and the need for holistic network architecture was addressed as the direction toward desired flexibility. Finally, latest developments related to sustainability initiatives and technical progress taken as a response to them were discussed too.

We view this set of technologies as a nucleus from which the 6G will emerge. We, however, do not limit 6G to these technologies only. Future versions of this white paper will follow advances in these technologies, just as they will report on development of other enabling technologies that are yet to emerge and come into 6G focus.

# Contributors

- **THz Frequencies:** Thomas Kürner, Tobias Doeker *(TU Braunschweig),* Mate Boban, Tommaso Zugno, Mengfan Wu, Lutfi Samara *(Huawei),* Claudio Paoloni *(Lancaster University).*
- **6G Radio Access:** Israel Leyva Mayorga *(Aalborg University),* Nikolaos Pappas *(Linköping University),* Najeeb Hassan, Miguel Angel Gutierrez Estevez *(Huawei),* Peter Trifonov *(ITMO).*
- **Next generation MIMO:** Danaisy P. Prado Alvarez *(Universitat Politècnica de València),* Eduard A. Jorswieck *(TU Braunschweig),* Ferhad Askerbeyli, Mario Castañeda, Martin Schubert, Michail Palaiologos, Ronald Böhnke, Malte Schellmann, Tobias Laas, Wen Xu *(Huawei).*
- **Integrated Sensing and Communication:** Andrea Giorgetti *(University of Bologna),* Richard Stirling-Gallacher, Tobias Laas, Michail Palaiologos *(Huawei).*
- **Non-terrestrial Networks:** Malte Schellmann, Sripriya Srikant Adhatarao, Henri Alam, Antonio De Domenico *(Huawei),* David López-Pérez *(Universitat Politècnica de València), ),* Riccardo Marini, Fabio Patrone, Alessandro Guidotti *(CNIT).*
- **Multimodal Sensing, Computing, Communication, and Control for 6G remote operation:** Mona Ghassemian *(Huawei),* Ana Garcia Armada *(UC3M),* Toktam Mahmoodi *(KCL),* Dejan Vukobratović *(UNS),* Albena Dimitrova Mihovska *(CGC),* Lina Magoula *(NKUA),* Andrés Meseguer Valenzuela *(ITI),* Peizheng Li, Adnan Aijaz *(Toshiba*), Nikos Bartzoudis *(CTTC),* Periklis Chatzimisios *(IHU),* Fatemeh Golpaygani *(UCD),* Firooz B. Saghezchi *(RWTH),* Christos Papadopoulos *(IHU),* Aryan Kaushik *(university of Sussex).*
- **Distributed Federated AI:** Ramin Khalili, Sayantini Majumdar *(Huawei),* Lina Magoula, Nikolas Koursioumpas *(NKUA),* Claudia Campolo, Antonio Iera, Antonella Molinaro *(CNIT),* Elizabeth Palacios *(Universitat Politècnica de València),* George Karetsos *(University of Thessaly).*
- **Intelligent User Plane:** Susanna Schwarzmann *(Huawei)*, Jari Mutikainen, Tugce Erkilic Civelek, Bahador Bakhshi, Riccardo Guerzoni *(Docomo Euro-Labs),* Rui Miguel Morais Silva, Daniel Nunes Corujo *(Universidade de Aveiro).*
- **Flexible programmable Infrastructures:** Carlos Guimarães, Luca Cominardi *(ZettaScale Technology SARL),* Aitor Zabala Orive *(Telcaria),* Artur Hecker, Dirk Trossen, Zoran Despotovic *(Huawei)*
- **Sustainability:** Riccardo Guerzoni, Bahador Bakshi *(Docomo Euro-Labs),* Rudolf Sušnik *(Internet Institute),* Nikolaos Koursioumpas *(NKUA),* Sripriya Srikant Adhatarao, Zoran Despotovic *(Huawei).*

**Published by:** one6G Association in September 2024



info@one6g.org

@One6GGlobal

one6g.org