(one6G)

Taking communications
to the next level

6G TECHNOLOGY OVERVIEW

WHITE PAPER

Second Edition - November 2022

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 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), distributed and federated artificial intelligence (Section 5), intelligent user plane (Section 6) and flexible programmable infrastructures (Section 7). 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 seven 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, 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 below 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]: **fixed point-to-point applications**, 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 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 which 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 centers**, 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, and more compact and efficient. Also, the absence of wired connections improves their re-configurability 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, tables, or hard disks. This use case enables the realization of Wireless Personal Area Networks (WPANs), where personal devices, such as smartphone, PCs, and monitors, can automatically interconnect without user intervention.

- Complementary **wireless links in data centers** [10][11][12] enable a faster reconfiguration of data centers, avoiding the deployment of large amount of fiber 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 fiber 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], **beam forming** [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 centers	[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]
	Vehicle-to-X communications	[21][22]
	Space communication	[23][24]
	Factory automation	[36]

Use case group	Use cases	Reference
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 systems. 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 modeling 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 analyzes 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 analyzers 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 needs 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 perform 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 centers
[45]	300 GHz	Broadband channel sounding	Wireless links in data centers
[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. Modeling 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 modeling 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 centers, [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 center	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

Reference	Scenario	Notes
[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 to a few centimeters	Three-dimensional space-time-frequency non-stationary geometry-based stochastic model
[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. Modeling 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 pathloss 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 which have to be taken into consideration, such as mutual coupling and near-field effects.

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

1.3. THz device technology

While front end at E-band (71 – 86 GHz) are already available in the market, the technology for links above 100 GHz is still at 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 constrains 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 problem of sway in case of wind, large footprint and cost due to the required high fabrication accuracy.

To note that the THz high attenuation is also advantageous because it permits an effective spatial division, frequency reuse and low interference expanding the potential of this technology.

In recent years, numerous THz wireless front end were presented up to 400 GHz with different technologies and performance [35][65]. Substantial advancements are reported in developing chipset based on different processes such as CMOS, InP, GaAs, Si Ge and GaN, mostly for low power electronic to support multi-Gb/s transmission. The data rate exceeded 10s Gb/s in most of the prototypes. However, all these systems provide short tens of meters of range, with the need to use very high gain antennas, due to lack of the required high transmission power. GaN is the most promising 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 the parts with tight tolerances. At 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/km2 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) for point-to-point transport. The key target is a low cost per bit, competitiveness with the fiber and long range. The novelty of these project 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.

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*, 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.

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) that 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. The 3GPP recently defined the survival time as the maximum time a cyber-physical control system may continue its operation without receiving an anticipated message [34]. 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. Instead, value of information (VoI) and semantics of information rely on harnessing the fact that data is communicated towards a particular end goal [16][26]—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.

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 [16], but also to schedule the users and allocate resources optimally.

Smart metering

This use case focuses on exploring and exploiting the influence of space and time in the behavior 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 neighbors and, hence, sets the bases 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-specific access mechanisms.

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. Thus, leading 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 [35]. 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 [38].

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 [40], usually referred to as puncturing [2][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) [41] 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 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 [41]. 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 [39]. 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 [39]) 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 fulfill 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 [42].

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 [1], 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 [2], the slicing of the resources is performed assuming that, if an URLLC is scheduled with sufficiently low latency, the reliability aspects are covered. Hence, [2] focuses on the importance of the loss model for the punctured eMBB traffic, considering a linear, a threshold-based, and a quadratic loss model as a function of the punctured resources. Furthermore, [46] 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 [57] comparing the energy consumption of using OMA and NOMA as scheduling solutions for eMBB and URLLC traffic. In particular, the schemes proposed in [57] 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 [57] 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 guaranteeing 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 [31].

The performance of NOMA and OMA mechanisms in the uplink have been studied in AoI-focused scenarios with homogeneous users [21]. 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 [7][19]. Even though the collision channel model is not favorable 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 [7][19] were recently extended in [58] 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 [25] 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 [18]. 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) [37][33]. Therefore, a significant amount of research has been performed on grant-free RA mechanisms [24] and on understanding the nature of packet transmissions in the finite-block length regime [27][45].

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) [43]. 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 [9]. 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 [30].

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 [10][14]. 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 [44]. 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 [4]. An algorithm for the link scheduling optimization that maximizes the network throughput for multi-connectivity in millimeter-wave cellular networks is proposed in [32]. The

considered approach exploits a centralized architecture, fast link switching, proactive context preparation and data forwarding between millimeter-wave access points and the users.

Furthermore, a specific type of multi-connectivity where different interfaces are used is termed interface diversity. In [12], 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 [8], 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.

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 [11], 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 [1] and resource allocation. Furthermore, lightweight ML techniques could be deployed at the user side for the individualized adaptation of the users' behavior (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 [5] and the impact of the knowledge and fundamental assumptions about the environment, for example, the channel [3].

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 to capacity exists [17]. One of the reasons for this loss is due to the non-optimal probability distribution of transmitted symbols. In [6] and [20] probabilistic shaping and geometric shaping schemes, respectively, are introduced that avoid this so called 'shaping loss'. It was shown in [48] 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 [29] with '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.

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 [46], 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 [22][23]. Polar codes with large kernels were shown to asymptotically achieve m=2[13]. Several constructions of polarization kernels are available [53][54]. 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 [49]. 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 [50][51]. Furthermore, it is possible to implement code length adaptation by employing a family of shortened polarization kernels [52], 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 [55], or to explicitly optimized the code structure for belief propagation decoding [56].

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 [47]. 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 [15].

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, [59] 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.

2.4. References

[1] 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.

[2] 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.

[3] 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.

[4] 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,

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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

[14] 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.

[15] 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.

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

[17] F. R. Kschischang and S. Pasupathy: “Optimal nonuniform signaling for Gaussian channels,” IEEE Transactions on Information Theory, 39(3), pp. 913-929, 1993.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

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

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 3GPP, "Service requirements for cyber-physical control applications in vertical domains," TS 22.104 V16.5.0, 2020.

[35] 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.

[36] 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.

[37] 3GPP, "Study on RAN Improvements for Machine-Type Communications," TR 37.868, Jul. 2011.

[38] 3GPP, "NR; Physical channels and modulation," TS 38.211 V17.0.0, Dec. 2021.

[39] 3GPP, "Physical layer procedures for control," TS 38.213 V16.3.0, 2020.

[40] 3GPP, "NR and NG-RAN Overall Description; Stage 2. TS 38.300 V15.3.1, Oct. 2018.

[41] 3GPP, "5G; NR; Medium Access Control (MAC) protocol specification," TS 38.321 V16.3.0, Jan. 2021.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] P. Trifonov. Design of Multilevel Polar Codes with Shaping. In Proc. Of Int. Symp. On Inf. Theory, 2022.

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

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

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

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

[53] 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.

[54] 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, October 2015.

[55] 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.

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

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

[58] 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.

[59] 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 23rd International Workshop on Signal Processing Advances in Wireless Communication (SPAWC), 2022, pp. 1-5.

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. In the following we give an overview of the state-of-the-art and promising trends in next generation multi-antenna technologies.

3.1. Fundamental MIMO Benefits

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 or not. Also, the design of MIMO solutions crucially depends on the particular scenario under consideration. There is not a "one fits all" MIMO solution. When designing MIMO systems, one always needs to consider practical constraints. It is also important to consider the tradeoff between the basic MIMO gains [1], which are as follows.

Array gain – By coherently combining 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. This gain is also known as beamforming gain.

Multiplexing and multiple access gain – Another 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 many 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. With an M-antenna array one can separate up to M users without interference. It should be noted, however, that there is a fundamental tradeoff 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 by combining multi-antenna processing with scheduling.

Antenna diversity and channel hardening – 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 the other antennas. This is exploited by transmitting or receiving redundant information in parallel over multiple antennas. Therefore, antenna diversity makes the channel robust against fading.

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 time synchronization requirements 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 from 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 cooperatively 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" [2]) or distributed over the service area. In this context, full-dimension (FD) MIMO was introduced in LTE Release 13. In 5G NR, 3GPP has specified 32 antennas in Release 15, which will rise to larger number in future releases (massive MIMO). The distributed MIMO (D-MIMO) technology has been discussed for 15 years (see e.g. [3]) and gained recently renewed attention within cell-free massive MIMO [4]. For 6G, the concept of multiple D-MIMO arrays, termed modular massive MIMO (mmMIMO) [5] 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 [5].

3.2. Channel modeling aspects

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 modeling of THz and sub-THz channels >100 GHz poses another challenge due to the special propagation properties in the THz spectrum, which behaves similar to optical channels. This is discussed separately in Section 1.

- 3GPP [6]: 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 [7]. The model also supports mobility and spatial consistency.
- METIS [8]: 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 [9]: 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 [10]: 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 [11]: 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 [12]: 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 [13]: 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. Circuit theoretical methods have been developed [14], [15] to take mutual coupling into account, leading to a systematic framework [16] to achieve the best performance taking mutual coupling into account. Various aspects of such a system were analyzed, including reciprocity [17]. The design of decoupling and matching networks (DMNs) for such systems is essential, as the number of elements to realize a DMN grows quadratically with the number of antennas in general [18]. 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 [19], [20]. Two new classes of power amplifiers, class M and class N, were introduced in [21], [22] to improve their power efficiency by energy recycling. Prototypes consisting of 3 to 4 antennas with antenna spacings smaller than half a wavelength and a wideband DMN were demonstrated in [23].

Although massive MIMO is a promising technology, it comes with a cost of increased number of circuit components and hence increased energy consumption and signaling 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 mmW bands [24]). Such techniques, for example, are based on low-resolution data converters, use of constant envelope transmit signals and hybrid beamforming. 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 [25]. The analysis of a communication system with low-resolution ADCs is extended to the power efficiency afterwards [26]. At the transmitter side, the power amplifiers (PAs) are the most power consuming elements and constant envelope transmit signalling is optimal for the PA's efficiency. One way to implement constant envelope transmit signalling is to employ 1-bit digital-to-analog converters (DACs). The distortions due to the quantization at the DACs are started to be taken into account in precoder design [27]. 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 [28], [29], some on providing nonlinear and linear precoders for constant envelope signalling [30] and some have offered linear and nonlinear precoding techniques for the systems with uniform DAC quantization [31]. Also, works such as [32], [33] 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

As modern wireless communications systems shift towards higher frequency bands, such as millimeter wave and THz bands, the use of multiple antennas becomes necessary due to the propagation characteristics and corresponding path loss [34]. As the channel at higher frequencies is mainly dominated by the line-of-sight (LoS) path, the deployment of multiple antennas at the transmitter (Tx) and receiver (Rx) can result in rank-one MIMO channels, thereby eluding an increase

of the spatial degrees of freedom. In such LoS MIMO channels, however, spatial multiplexing gain can still be extracted, by optimizing the placement of the antennas at the Tx and Rx depending on the Tx-Rx distance [35]. Consequently, if the antenna separation at the Tx and Rx arrays satisfies certain design criteria, full spatial multiplexing gain can be extracted at a specific Tx-Rx distance. In particular, there have been results for the optimum antenna placement for uniform linear arrays (ULAs) [36], uniform planar arrays (UPAs) [37] and uniform circular arrays (UCAs) [38]. However, as the optimum antenna placement depends on the distance between the Tx and Rx arrays, LoS MIMO systems are susceptible to variations in the Tx-Rx distance, leading to a decrease in performance when operating at different distances. Nonetheless, many applications require LoS MIMO systems operating over a range of Tx-Rx distances. As a result, there have been recently proposals for enabling high rank LoS MIMO systems over a wide range of Tx-Rx distances, e.g. by antenna selection techniques [39]. Furthermore, the design of the optimum antenna placement of LoS MIMO systems assume that the Tx and Rx arrays are aligned, i.e. they are facing each other. In practice, however, there can be misalignment between the arrays, which also lead to a performance degradation. For this purpose, schemes have been proposed to alleviate the misalignment, e.g., via precoding/combining [40].

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. But for general interference channels 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 cooperation of the nodes, the availability and exchange of CSI, the casting type (unicast or multicast [41]), 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 the power allocation and the scheduling. The scheduling ensures that the users are orthogonalized in frequency and time domains to avoid interference, while the power control ensures that certain 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 [42].

3.6. Cell-free massive MIMO

Network densification is an important factor in realizing 6G services with extreme area spectral efficiency and low latency [43]. 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 [44]. 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. Cell-free schemes can provide nearly 5-fold improvement in 95%-likely per-user throughput over the small-cell scheme, and 10-fold improvement when shadow fading is correlated [45], [46].

In order to allow scalability, the user-centric dynamic cooperation clustering (DCC) scheme was introduced [47], where each UE can connect to nearby APs in a fully flexible manner. The performance analysis of cell-free massive MIMO systems for different fading channel models [48], [49] shows the general conclusion that it can achieve a great performance 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 [50], [51]. 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 [52]. In the context of cell-free MIMO, well-established system components, 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 [53], [54].

3.7. Intelligent reflecting surfaces

Intelligent reflecting surfaces (IRSs) are emerging as a potential key technology for beyond 5G systems [55]. The term "IRS" refers to low-cost wireless planar structures that are able to reflect and transform electromagnetic waves, either actively or passively. So-called holographic IRS ("Holographic MIMO") integrates a massive (virtually infinite) number of tiny antenna elements on a compact surface: [56], [57].

There are already numerous works aimed at optimizing the parameters (amplitudes, phase shifts) of an IRS according to the desired use case. Some use cases include coverage extension [55], signal-to-noise ratio (SNR) maximization for the single-user case [55], [58], leveraging the IRS beamforming capabilities for base station (BS) transmit power minimization in the multiuser case [59], and channel rank improvement [60].

In the aforementioned works, perfect channel state information (CSI) is assumed, i.e. the CSI of user(s) to IRS links and IRS to BS links are assumed known. This is unfeasible in practice. The works [61], [62] tackled the channel estimation problem for IRSs; nonetheless the required training overhead scales proportionally to the number of IRS antenna elements, limiting the applicability of these schemes in practice. This is due to the fact that an IRS is very likely to have 100s or even 1000s of antennas. Later works, e.g., [63] , [64], [65] have developed channel estimation schemes with limited overhead. In particular, the works [64] and [65] propose to group IRS antennas together and to use fixed DFT coefficients for each IRS antenna group. This has the effect of reducing the effective IRS channel dimension. While the authors in [64] pursue a least-squares approach to estimate the resulting IRS channels, the authors in [65] use linear MMSE estimation based on channel spatial correlations.

Going beyond the work in [65] which uses fixed IRS parameters, we propose in [66] to optimize the IRS parameters to minimize the channel estimation MSE or sum MSE in a system with multiple IRSs. The optimization exploits channel spatial correlations as side information and is based on an alternating optimization and projected gradient descent framework. The final result is the forming of IRS beams that direct the user pilots into the IRS channel spatial correlation matrices' eigenspaces containing the most power. This is a desired result in the low-training overhead regime, as shown in numerous previous works. Indeed, numerical results show superior channel estimation and data rate performance to the method in [65].

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 at Terahertz bands 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 [67] 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 [68], 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 [69]. 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 are approached in recent studies. An algorithm to generate pairwise secret keys in massive multi-user MIMO networks is derived in [70], which exploits the spatially correlated structure of the underlying massive MIMO channels. In [71], 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 is a key design principle for future MIMO implementations (see e.g. [72]). 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 requires less signaling 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) [73]. 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, 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, 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 [74]. 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 [75], [76], that are required to support the huge number of devices in the future. Also, rate-splitting approaches, which are well-known from information theory [77], have recently regained significant interest for MIMO downlink communication [78] 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 [79], [80].

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 Integrating Sensing and Communications in Section 4.

3.11. References

[1] L. G. Ordóñez, D. P. Palomar and J. R. Fonollosa, "Fundamental diversity, multiplexing, and array gain tradeoff under different MIMO channel models," 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 3252-3255.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

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

[10] 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.

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

[12] "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.

[13] R. J. Weiler, et al., “Quasi-deterministic millimeter-wave channel models in MiWEBA,”EURASIP Journal onWireless Communications and Networking, vol. 2016, no. 1, Mar 2016.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

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

[22] B. Lehmeyer, A. Mezghani, and J. A. Nossek, “Electronic amplifier for amplifying an input signal,” WO 2018/024756 A1.

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] E. Torkildson, H. Zhang and U. Madhow, "Channel modeling for millimeter wave MIMO," 2010 Information Theory and Applications Workshop (ITA), 2010, pp. 1-8.

[35] F. Bohagen, P. Orten and G. E. Oien, "Design of Optimal High-Rank Line-of-Sight MIMO Channels," in IEEE Transactions on Wireless Communications, vol. 6, no. 4, pp. 1420-1425, April 2007.

[36] M. H. Castañeda Garcia, M. Iwanow, and R. A. Stirling-Gallacher, "LoS MIMO design based on multiple optimum antenna separations," in 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall), 2018, pp. 1–5.

[37] F. Bohagen, P. Orten, and G. E. Oien, "Optimal design of uniform planar antenna arrays for strong line-of-sight MIMO channels," in 2006 IEEE 7th Workshop on Signal Processing Advances in Wireless Communications, 2006, pp. 1–5.

[38] L. Zhou and Y. Ohashi, "Performance analysis of mmWave LOS-MIMO systems with uniform circular arrays," in 2015 IEEE 81st Vehicular Technology Conference (VTC Spring), 2015, pp. 1–5.

[39] M. Palaiologos, M. H. C. Garcia, R. A. Stirling-Gallacher and G. Caire, "Design of Robust LoS MIMO Systems with UCAs," 2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall), 2021, pp. 1-5.

[40] R. Chen, H. Xu, M. Moretti and J. Li, "Beam Steering for the Misalignment in UCA-Based OAM Communication Systems," in IEEE Wireless Commun. Letters, vol. 7, no. 4, pp. 582-585, Aug. 2018.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, ”Cell-free massive MIMO versus small cells,” IEEE Trans. Wireless Commun., vol. 16, no. 3, pp. 1834-1850, Mar. 2017.

[46] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson and T. L. Marzetta, "Correction to “Cell-Free Massive MIMO Versus Small Cells” [Mar 17 1834-1850]," in IEEE Transactions on Wireless Communications, vol. 19, no. 5, pp. 3623-3624, May 2020.

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

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," in IEEE Commun. Mag., vol. 58, no. 1, pp. 106-112, Jan. 2020.

[56] C. Huang et al., "Holographic MIMO Surfaces for 6G Wireless Networks: Opportunities, Challenges, and Trends," in IEEE Wireless Communications, vol. 27, no. 5, pp. 118-125, October 2020.

[57] Yuanwei Liu, Xiao Liu, Xidong Mu, Tianwei Hou, Jiaqi Xu, et al.. Reconfigurable Intelligent Surfaces: Principles and Opportunities. IEEE Communications Surveys & Tutorials, 2021. ffhal-03357962.

[58] E. Björnson and L. Sanguinetti, "Power Scaling Laws and Near-Field Behaviors of Massive MIMO and Intelligent Reflecting Surfaces," IEEE Open Journal of the Communications Society, vol. 1, pp. 1306-1324, 2020.

[59] Q. Wu and R. Zhang, "Intelligent Reflecting Surface Enhanced Wireless Network via Joint Active and Passive Beamforming," in IEEE Transactions on Wireless Communications, vol. 18, no. 11, pp. 5394-5409, Nov. 2019.

[60] M. A. ElMossallamy et al., "On Spatial Multiplexing Using Reconfigurable Intelligent Surfaces," in IEEE Wireless Communications Letters, vol. 10, no. 2, pp. 226-230, Feb. 2021.

[61] 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.

[62] 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.

[63] 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.

[64] B. Zheng and R. Zhang, "Intelligent reflecting surface-enhanced OFDM: Channel estimation and reflection optimization," IEEE Wireless Commun. Lett., vol. 9, no. 4, pp. 518-522, April 2020.

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

[66] 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.s

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

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

[73] 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.

[74] 3GPP, TR 38.812, "Study on Non-Orthogonal Multiple Access (NOMA) for NR", Dec. 2018.

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

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

[77] 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.

[78] 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.

[79] 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.

[80] 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.

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 (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]. Machine learning-based approaches based on soft information for 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 map, and UE profile. 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 point-to-point communications such as vehicular networks (with great applications to autonomous driving) and complicated mobile/cellular networks, which can potentially revolutionize the current communication-only mobile networks [4].

ISAC at mmWaves may reach significant performance levels 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]. Exploiting THz technologies, the potential gains to sensing are enormous in terms of resolution and accuracy, especially in 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 signaling 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.

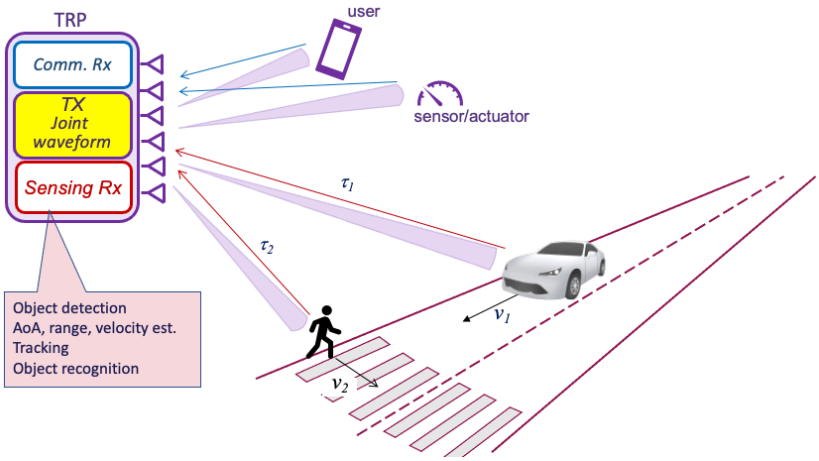


Figure 2: Joint sensing and communication in a vehicular scenario (monostatic configuration).

4.2. Use cases

4.2.1. 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 application; for instance, they can assist the driver when executing safety-related maneuvers. 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.

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 2. 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 3. 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.

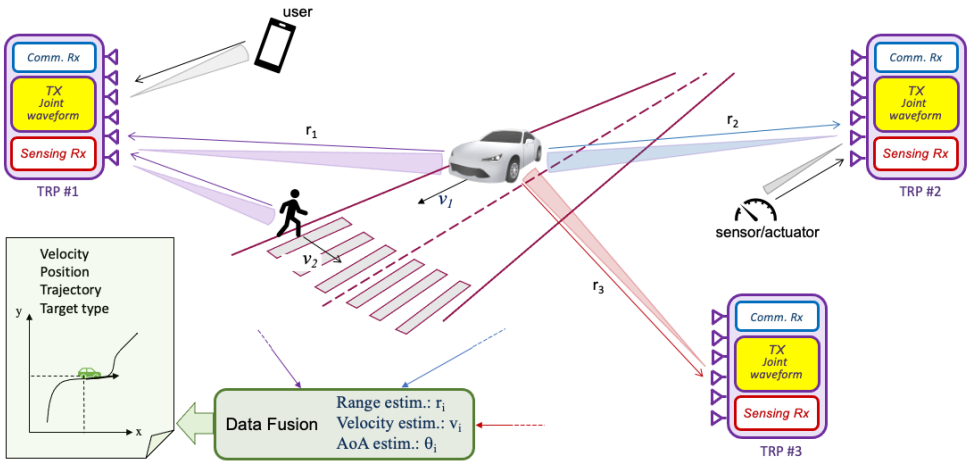


Figure 3: Joint sensing and communication in a vehicular scenario (multi-static/distributed configuration).

While general communications can deal with hundreds of ms delays, autonomous vehicle applications require delays in the order of tens of ms [19][20][21]. In such scenarios, sensing should provide robust, high-resolution obstacle detection in the order of a few decimeters.

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 2 with a monostatic base station (BS) or in Figure 3 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.

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. Ongoing research and open problems

There are numerous directions on which ISAC research is focusing. The impact of bistatic configuration on sensing capabilities is investigated in [22], where a multi-beam JSC system analysis in terms of achievable sensing coverage is performed. The design of new waveforms that combine OFDM and OTFS to enable JSC has been recently proposed in [23]. In IoT scenarios, communication is mainly packet-based, and where the transmission can be sporadic, the impact of packet length on the sensing/communication trade-off is studied in [24], where a JSC system is investigated under the finite block-length regime. Regarding sensor fusion strategies (as depicted in Fig. 3), in [25], tracking algorithms are used to perform a fusion of local target position estimates obtained by multiple monostatic sensors. In this context, the benefits of data fusion can be used to increase sensing performance or maintain the same performance (of a single sensor) but release radio resources for communication.

4.4. 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] 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.

[20] 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.

[21] 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.

[22] 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.

[23] L. Rinaldi, D. Tagliaferri, F. Linsalata, M. Mizmizi, M. Magarini, and U. Spagnolini, “Dual domain waveform design for joint communication and sensing systems,” preprint arXiv:2111.12339, 2021.

[24] 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.

[25] 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.

5. 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 analyzed 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 fulfill 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.

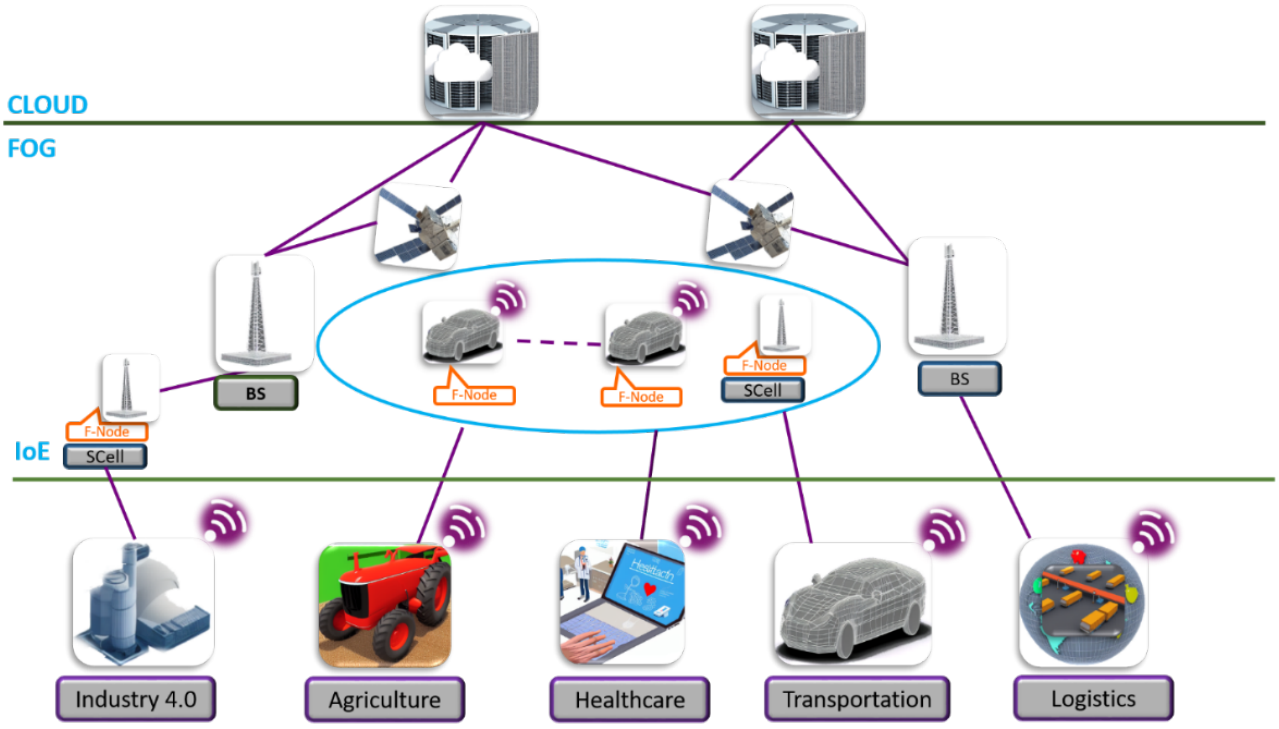


Figure 4: A carrier network along with connected terminals forms a distributed platform.

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 4. 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.

5.1. 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 behavior 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 5.3.

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 take into account nearby vehicles, pedestrians, cyclists, road lanes, among others (see Figure 5). 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].

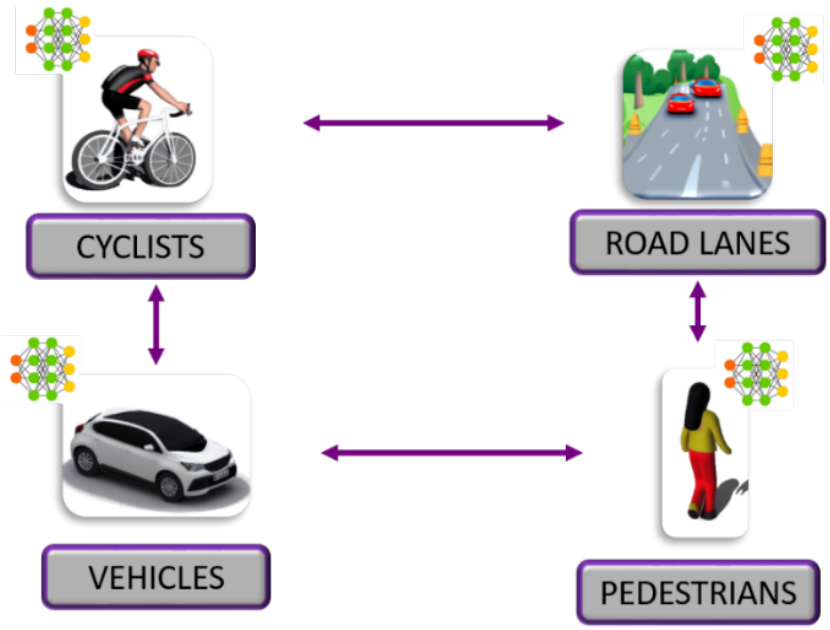


Figure 5: Autonomous navigation is an example use case of distributed learning.

5.2. State-of-the-art

5.2.1. 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, virtualised and fully interoperable [5].

5.2.2. 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.

5.2.3. Academic literature

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.

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 centers etc. creates new challenges for large scale machine learning and associated optimisations. 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.

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.

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.

5.3. Ongoing research

5.3.1. Distributed 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. Current state-of-the-art approaches propose solutions towards the identification of potential QoS deterioration in a centralized manner. However, centralized solutions may raise privacy issues, since sensitive user information may need to be transmitted to communication network entities for processing and analysis. Other practical challenges of centralized solutions may also arise, such as the computational bottleneck and the fast increase of signaling overhead with number of end users.

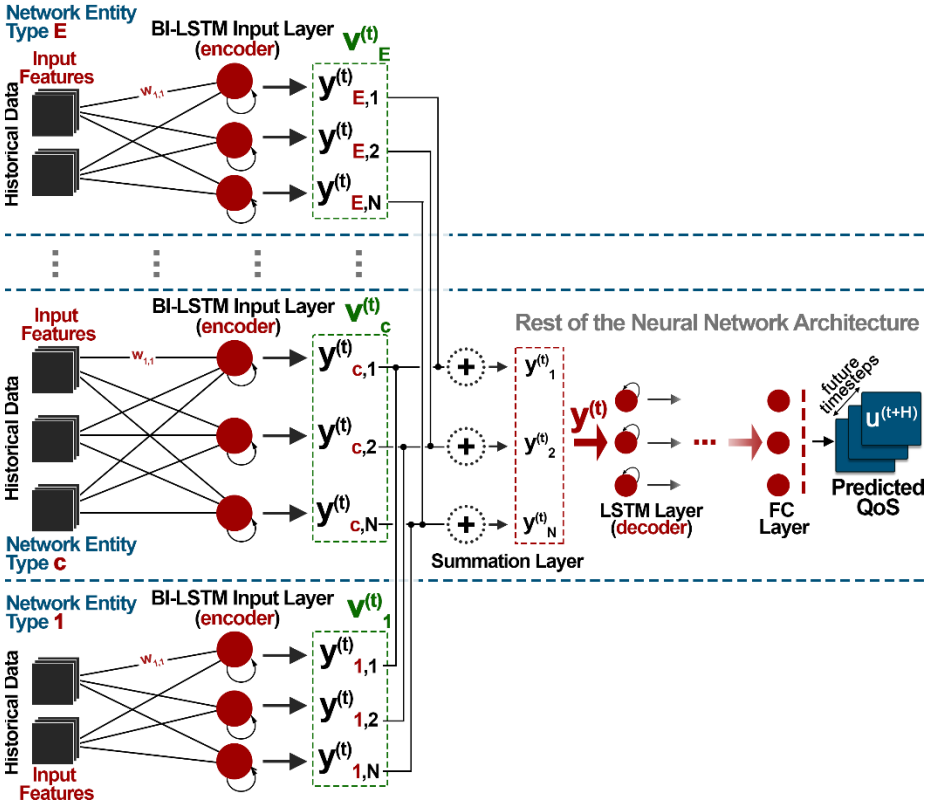


Figure 6: Distributed QoS prediction scheme

The aforementioned concerns motivate the design and integration of the 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 signaling overhead as well as the network delays can be significantly reduced. The proposed distributed scheme for QoS prediction (Figure 6) realizes a distributed LSTM-based architecture,

considering several distinct types of network entities (e.g. UE, BS, MEC, Cloud), where each type of network entity (NET) is responsible of collecting and processing a subset of features related with the prediction task.

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 [52]. 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 [53] 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 signaling 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.

5.3.2. 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 [54].

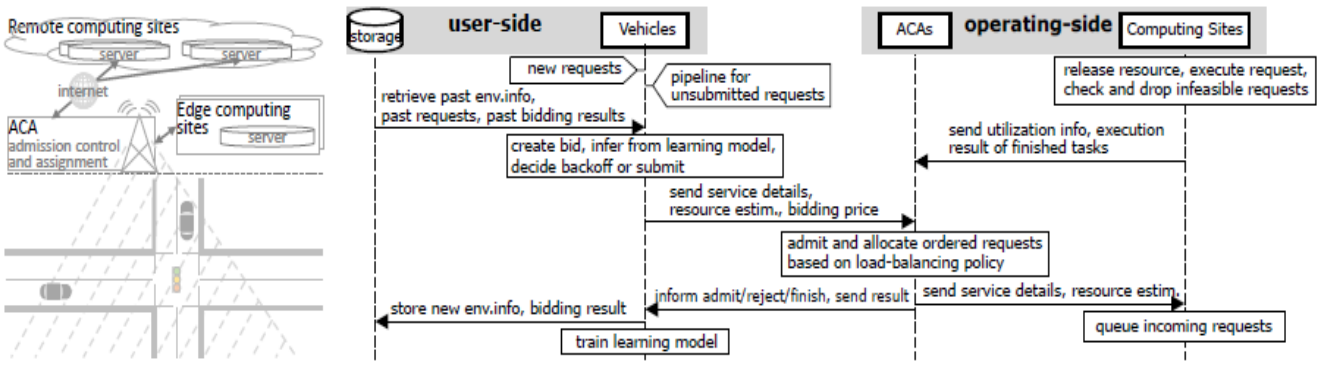


Figure 7: system model; right: an example topology, left: message sequence.

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 vehicles. Gathering such data imposes communication and latency overhear. Besides, Centralized modeling 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 [55], 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 7).

Multi-agent systems use agents to represent individual interests and model complex interaction between players. When a centralized modeling problem is broken down into local, individual models, it reduces model complexity and data requirements. In a MARL system, each agent interact with the environment to obtain rewards, which allows the agent to learn highly-rewarded behavior. 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 behavior 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 backoff 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 these 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 [55].

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.

5.4. 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] H. Ning, Cyberspace Data and Intelligence, and Cyber-Living,Syndrome, and Health. Springer, Singapore, 2019

[53] 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.

[54] C. J. Bernardos et al., “European vision for the 6g network ecosystem,” The 5G Infrastructure Association, 2021.

[55] 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.

6. Intelligent User Plane, In-Network Computing

6.1. 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 [33], 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 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 [34].

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.

6.2. 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 [1]: 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 6.3.

6.3. 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” [2][3] 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.

6.3.1. Remote haptic operation

This use case is related to haptic tele-operation in a dynamic environment. In [3], "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.

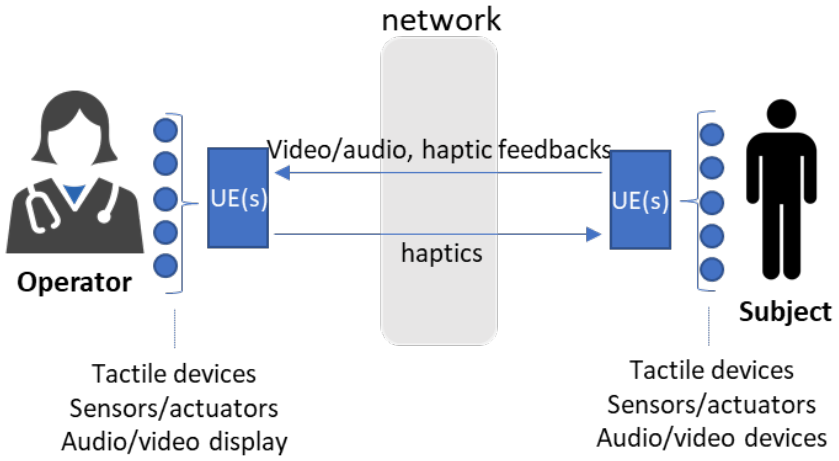


Figure 8: Remote haptic operation

Figure 8 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 4 recalls the communication requirements according to [3]. 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 4 - 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) 10-100 (dynamic environment)	1-4k pkt/s (P) (w/o compression) 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) 100-500 (w/ compression)

*DoF: Degree of Freedom (i.e., the number of joints in the human body) [4]
(P): Periodic

6.3.2. 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" [3].

Figure 9 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.

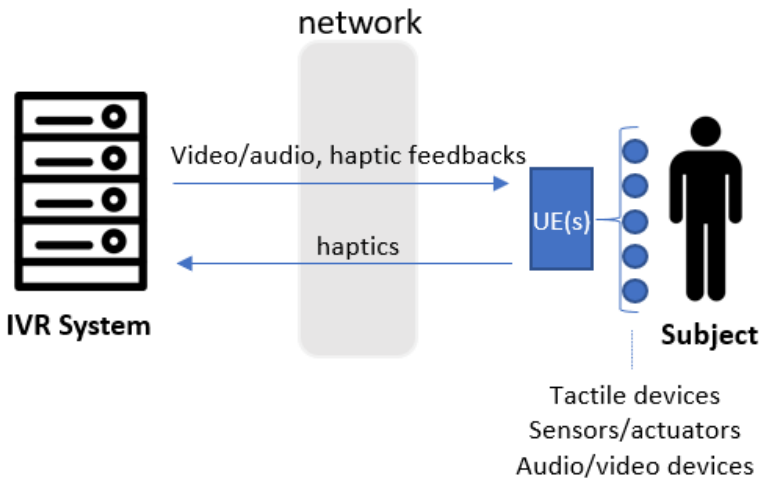


Figure 9: Immersive Virtual Reality (IVR)

The communication requirements according to [3] and [6] are reported in Table 5 and they can be summarized as: device to server communication with latency <5ms and five-9 reliability.

Table 5 - 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.) 99.999% (w/ comp.)	<5	1-4k pkt/s (P) (w/o compression) 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.) 99.999% (w/ comp.)	1-50	1-4k pkt/s (P) (w/o compression) 100-500 (w/ compression)

6.3.3. 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" [3]. 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 10 (which is an elaboration of Fig. 4 in [3]) 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)" [3] and they may be physical/virtual embodiments of digital twins updated in real-time.

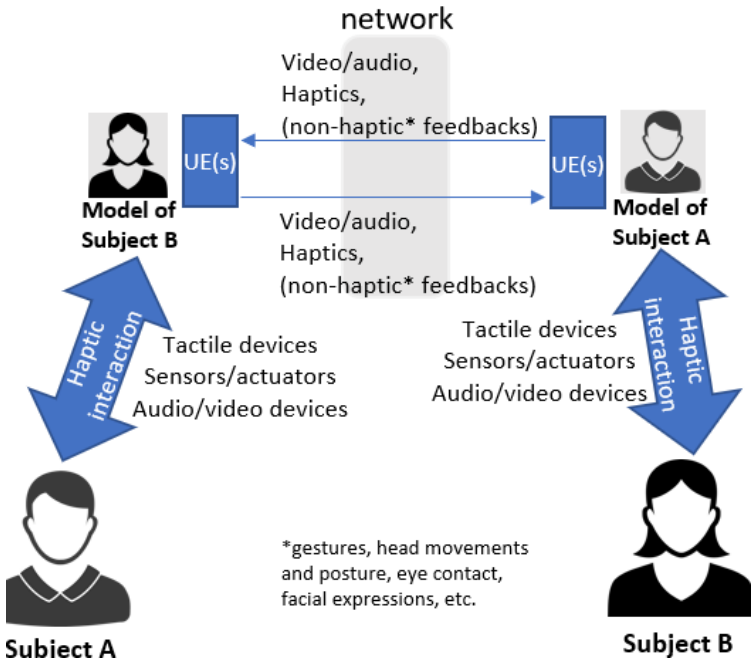


Figure 10: Haptic Interpersonal Communication (HIC)

The communication requirements are reported in Table 6 and can be summarized as: device to device communication with latency <10ms and five-9 reliability.

Table 6 - 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)

6.3.4. 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, localised 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) [5]. Additionally, as shown in Figure 11, 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.

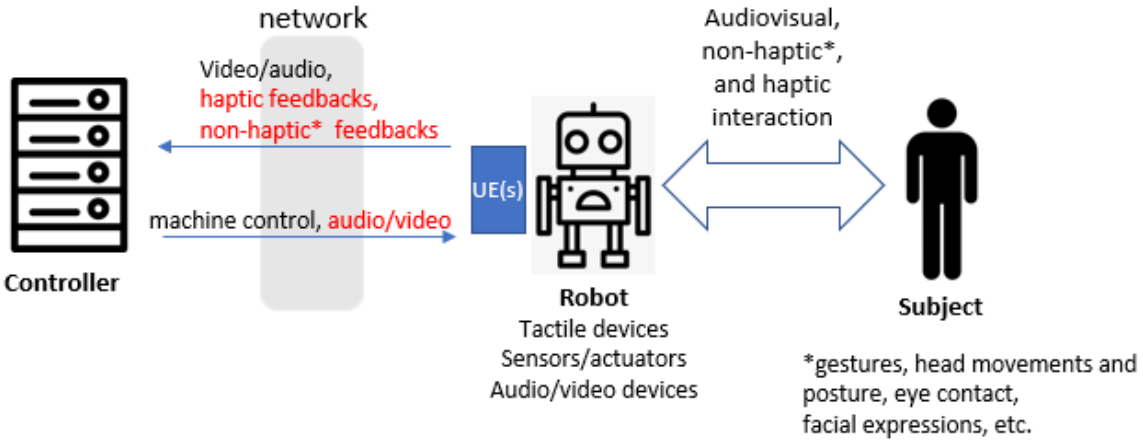


Figure 11: AI-based customer services

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.

6.3.5. 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.

6.4. SotA and discussion topics

The use cases and requirements discussed in the previous section bring to re-thinking the user plane architecture based on new requirements. 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

6.4.1. 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 be still 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 12 points out flatter, e.g. mesh, topologies as future direction. 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.

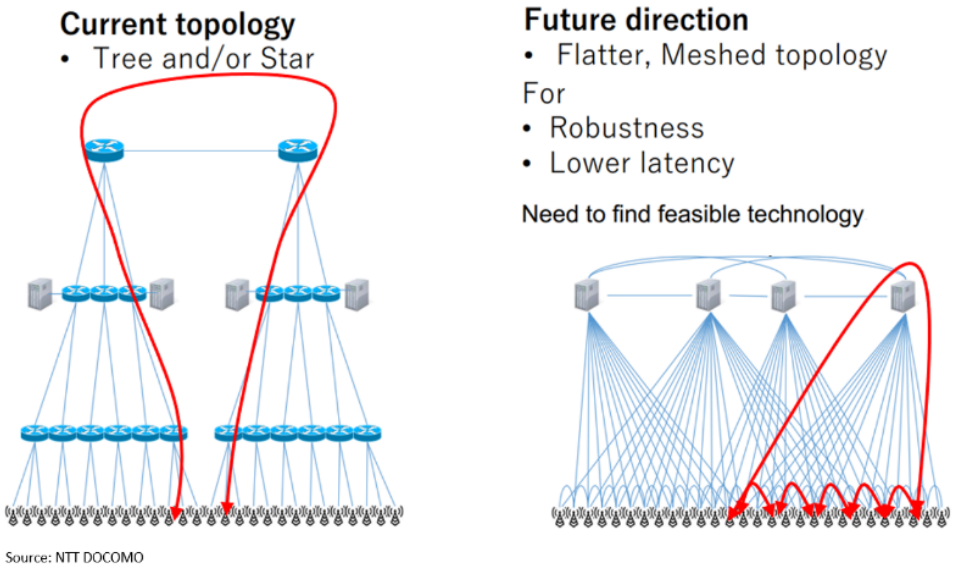


Figure 12: Current network topology and future direction towards flatter topology

6.4.2. 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 13 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.

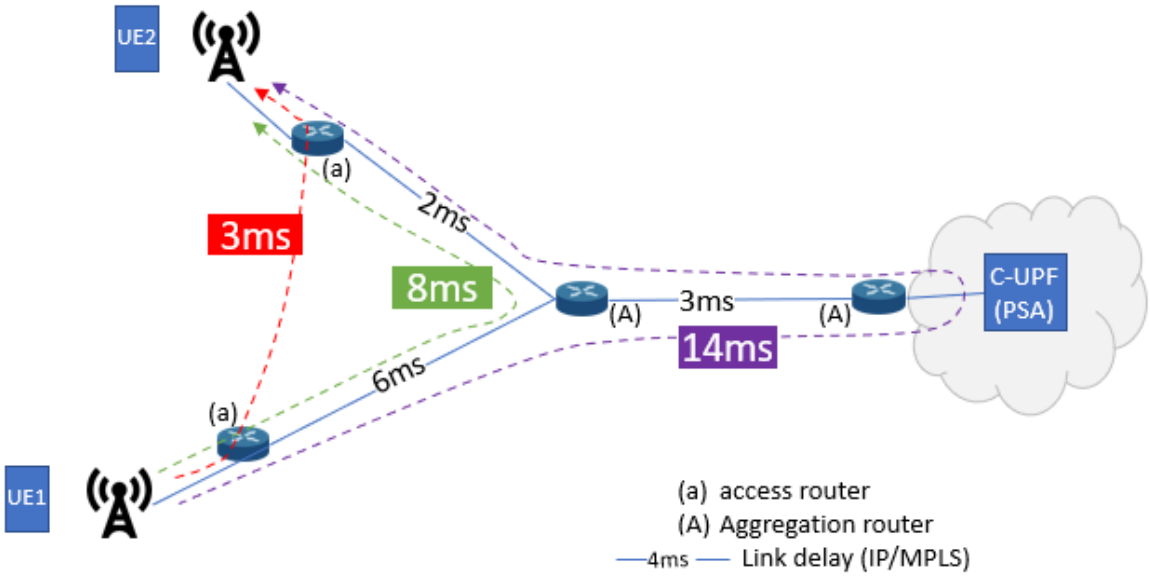


Figure 13: Achievable latency targets in 5GC and beyond

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 lower latency target may be expected for the next generation radio access.

6.4.3. Wide area synchronization and deterministic communication

The 5G System (5GS) 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 [5][7] 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 [10] in the future.
- These use cases set the requirements for IEEE TSN in 3GPP Rel-16 [11] and are further improved in in Rel-17 [12].

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 [12].

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 [1] 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.

6.4.4.In-network computing

In-network computing describes the paradigm of delegating application-layer processing functions to the data plane [13]. The concept is not a novelty with the advent of softwarized networks, but has already been proposed decades ago as active networks [14][15]. 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 [13] in a network device, and with the advancements of network programming languages, in-network computing evolved as an important research field again. It allows to process the traffic as it is transmitted, which reduces the computing resource- and energy-consumption of devices located at end hosts.

According to [16], the key potentials in-network computing can bring are (1) a significant reduction 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 5. Some works aim at using in-network computing to enhance the efficiency of applications. A brief overview is given in the following.

In-Network DNS

P4DNS [17] presents an in-network solution so to reduce the latency until a DNS response is obtained. 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 [18]. 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 and 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 [16], a proof of concept system for data aggregation, which is a common task in several distributed data center 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 [19]. The proposed architecture consists of a payload analyzer, 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 [13]. While prior works aim at placing computational tasks into existing networking hardware, the authors proposal a general computing platform which takes two tasks at the same time: Performing general computations and acting as a network node. As – 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 [20]. 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 analyze 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 [21]. The authors consider a setup, where a robot is interacting with a human. Thus, any type 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 analyzed. 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 reduced, thus enhancing the safety of the environment.

6.4.5. 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 whist being cost-efficient. Although the prior arts do not consider the intelligence being placed directly

within the user plane, these mechanisms can give a direction towards an intelligence user plane design.

Dynamic Scaling

The work in [22] 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.

Optimized Placements

Not an optimized scaling, but an optimized placement of UPFs (and MEC servers), is the target of several prior works [23][24][25]. More specifically, [23] 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, [24] 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 [23]. 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 [25] 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.

6.4.6. 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 [26]. 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 [27]:

(1) Switches have a high performance and are capable of achieving a latency in the order of hundreds of nanoseconds per packet [28] and are thus faster than high-end ML accelerators [29].

(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 center, 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.

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 [27][30]. 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 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 7: 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

Reducing Complex ML Logics to Simple Match-Action Rules

The authors of [27] 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.

Distributed Machine Learning

The work in [31] 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 14.

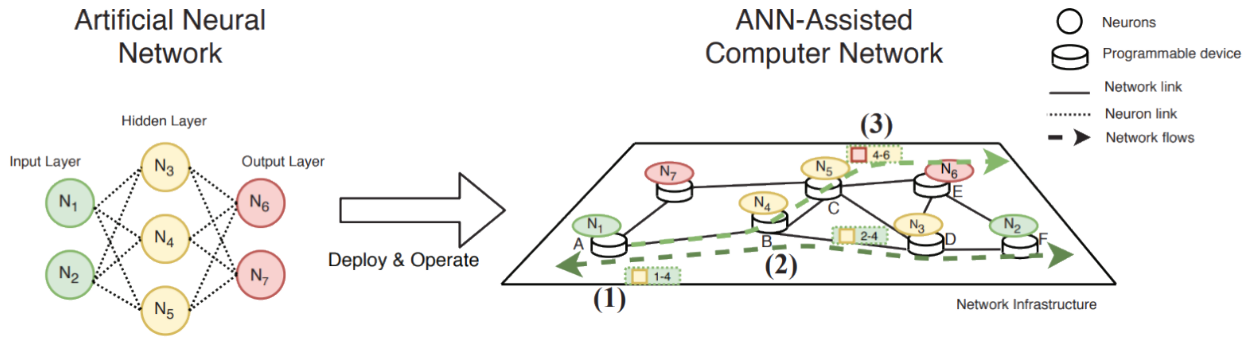


Figure 14: Distributed ANN [31]

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-packing 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.

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 of 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 in [32] 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 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 5. Building on top of this architecture, Taurus [30] 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 center 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.

6.5. Technology trends: summary

Section 6.4 presents different solutions for an evolution towards an 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.

6.5.1. 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 15 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.

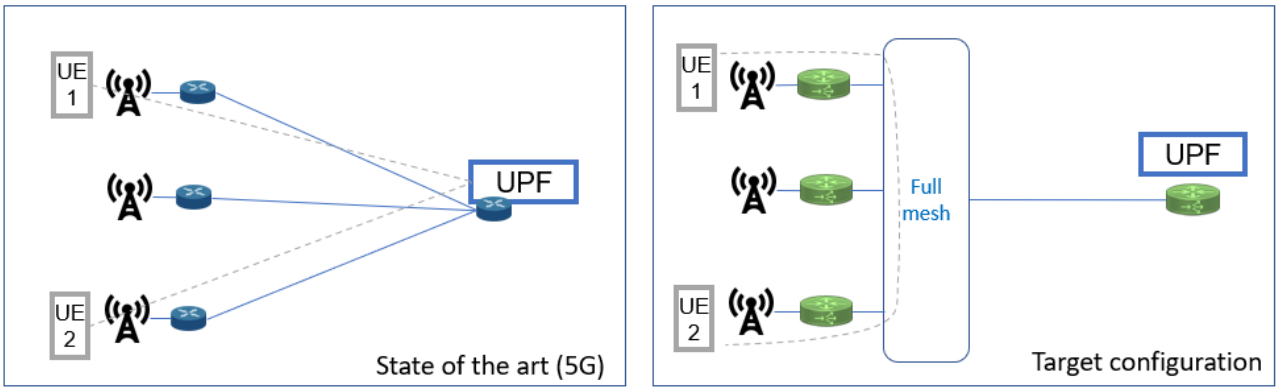


Figure 15: UE-to-UE communication in legacy and future architecture

In comparison with the 5G system architecture depicted in Figure 16, Figure 17 describes the architecture changes required to realize the target configuration in the 3GPP User Plane by delegating functionalities to the transport layer.

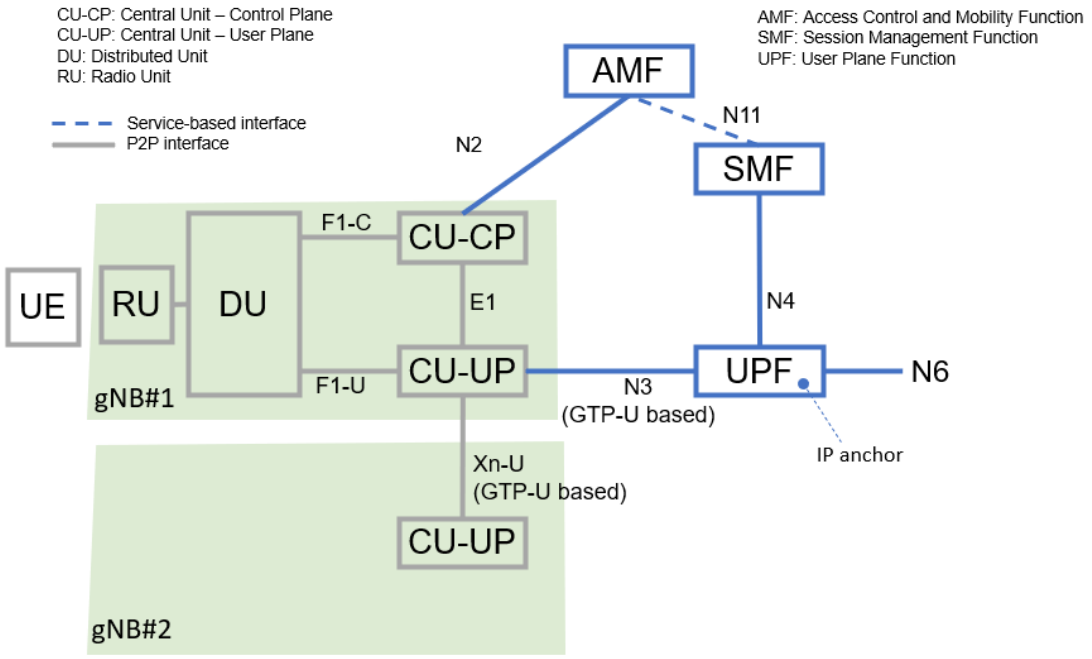


Figure 16: 5G state of the art architecture (simplified view)

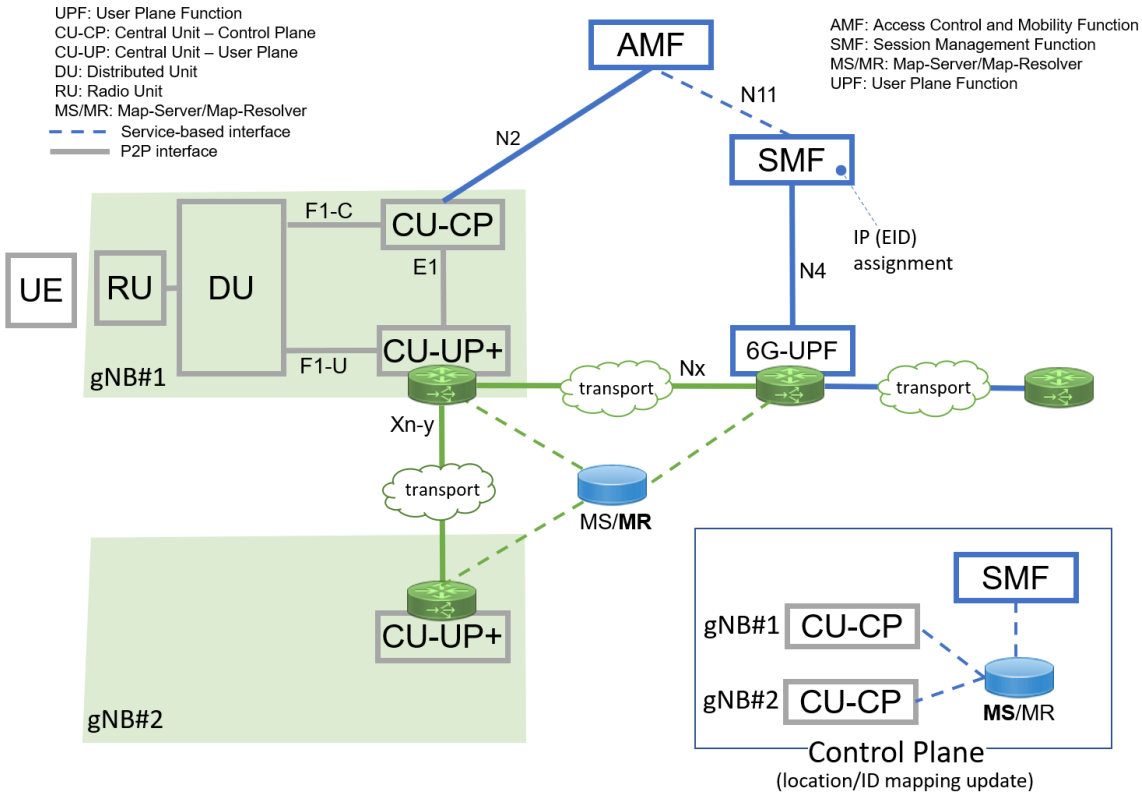


Figure 17: proposed 6G User Plane architecture to delegate UP functionalities to the transport network

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 16) 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.

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.

6.5.2. Leveraging In-network Computing in the User Plane

In the following, we present a solution stream leveraging in-network computing for 6G networks. 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.

AR end devices have conflicting design goals. This involves high wearing comfort, being wireless (to support high freedom of movement to the user), as well as being powerful in terms of their computational resources so to perform the complex application-related tasks (e.g. video rendering or object/pattern detection). However, AR devices cannot be equipped with large batteries, as the devices would otherwise become too heavy. On the other hand, the complex tasks they have to complete are computing-, and thus, energy-consuming and can also lead to heating up the devices.

Figure 18 illustrates an exemplary AR scenario, highlighting the fast battery drainage. The first gray box on the left denotes the AR devices and the applications running on them ①. It shows a haptic glove ②, a racket ③ and ball, and an AR glass with integrated headphones ④. The haptic glove performs an intelligent filtering of sensory information and the racket can detect whether the ball was hit or missed. The AR glasses act as a designated device for synchronizing the flows belonging to the different modalities of the user (haptic feedback of glove and racket, audio, and video). This can be done, for example, via tethering, where the AR glasses sets up a hotspot to which the glove and the racket connect. The AN ⑤ connects the devices with the core network ⑥. Via the data network ⑦, the data – pre-processed at the AR devices - is transmitted to the application server and vice-versa ⑧. Due to the computations carried out on the end-user equipment, their batteries drain fast.

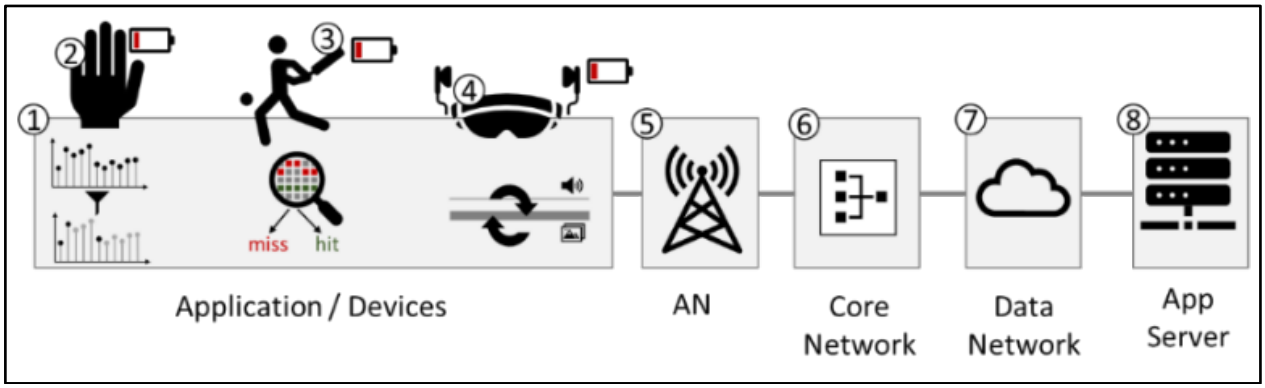


Figure 18: The AR application tasks are executed in the AR devices (glove, racket, ear-phones and AR glasses), leading to fast battery drainage.

This issue can be overcome by leveraging INC in the 6G user plane, such that application- and device-specific tasks can be offloaded to the UP entities. Such a scenario is outlined in Figure 19. The

user plane entities (i.e. AN and UPFs) can be programmed to execute the AR application tasks in the network. The first gray box shows the involved AR devices ①. Compared to the scenario denoted in Figure 18, the computations are not performed at the end-devices themselves. Instead, by means of the programmable UP ②, the involved UP entities (the AN ③ and the UP function(s) in the core network ④) carry out the information filtering, pattern detection, and flow synchronization while simultaneously forwarding the data. The processed data is then sent via the data network ⑤ to the application server ⑥. The UP entities and the UE(s) are programmed via a controller, e.g. via the AF ⑦, to define the packet treatment.

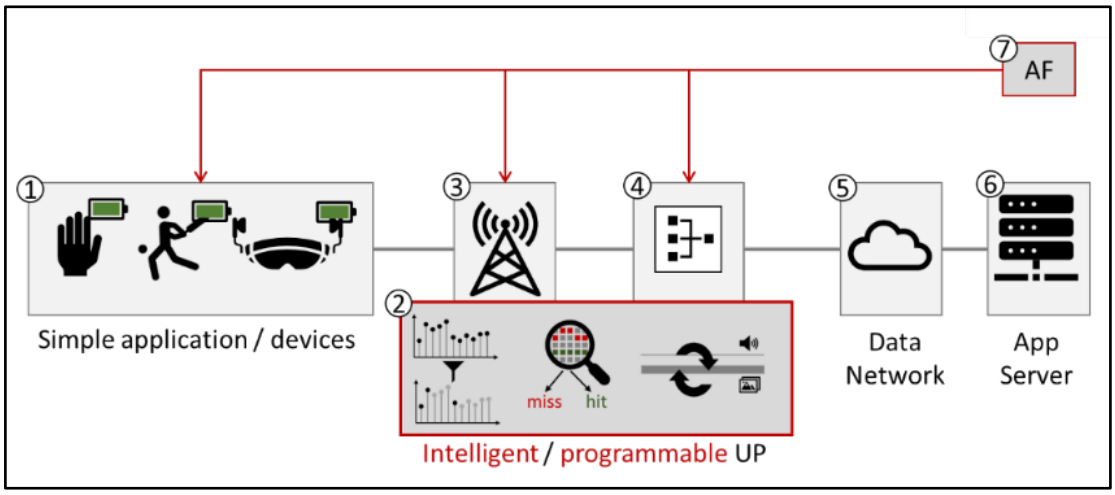


Figure 19: Offloading application tasks to the UP nodes: The AR application tasks are executed in network devices, reducing the complexity of the AR devices and thus reducing their energy consumption. The UP is programmed via a controller.

Such a solution would mean a paradigm shift from pure communication flows or QoS Flows to communication and compute flows. For their realization, the UP entities would need to be enhanced by sufficient (virtual) computational resources and memory to carry out the computations. To define the computations in the UP nodes, two options could be feasible. That is, (1) a pre-defined (potentially standardized) set of generic micro-services that (nearly) every UP entity supports off the shelf and (2) very specific micro-services, defined for example by an AF, supported by a subset of UP entities.

6.6. Conclusions on Intelligent User Plane

Section 6 elaborates 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 and AR. 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 20. 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.

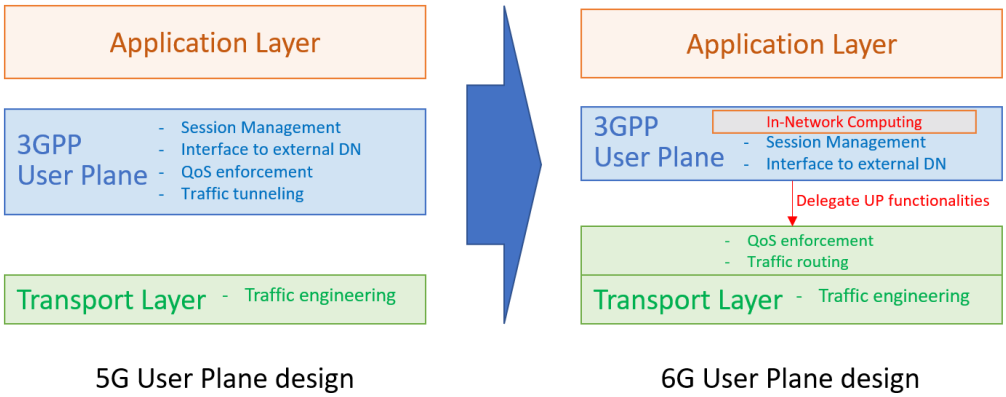


Figure 20: potential 6G User Plane design in comparison to 5G state of the art

6.7. References

[1] NTT DOCOMO White Paper on "5G Evolution and 6G", Feb. 2021 (Version 3.0). [Online]. Available: https://www.nttdocomo.co.jp/english/binary/pdf/corporate/technology/whitepaper_6g/DOCOMO_6G_White_PaperEN_v3.0.pdf

[2] IEEE P1918.1 Tactile Internet Working Group. (2017) Tactile internet working group project website. [Online]. Available: https://standards.ieee.org/develop/project/1918_1.html

[3] O. Holland et al., "The IEEE 1918.1 "Tactile Internet" Standards Working Group and its Standards," in Proceedings of the IEEE, vol. 107, no. 2, pp. 256-279, Feb. 2019

[4] K. S. Kim et al., "Ultrareliable and Low-Latency Communication Techniques for Tactile Internet Services," in Proceedings of the IEEE, vol. 107, no. 2, pp. 376-393, Feb. 2019

[5] 3GPP TS 22.104, "Service requirements for cyber-physical control applications in vertical domains", Release 16. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3528

[6] 3GPP TR 22.847, "Study on supporting tactile and multi-modality communication services" Release 18 V0.3.0 (2021-07)

[7] 3GPP TS 22.804, Study on Communication for Automation in Vertical domains (CAV), on-line at https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3187

[8] Stream Control Transmission Protocol, IETF RFC 4930, on-line at https://datatracker.ietf.org/doc/html/rfc4960

[9] Multiprotocol Label Switching Architecture, IETF RFC 3031, on-line at https://datatracker.ietf.org/doc/html/rfc3031

[10] IEEE 802.1 Time-Sensitive Networking (TSN) Task Group, on-line at https://1.ieee802.org/tsn/

[11] 3GPP TS 23.501 V16.11.0 (2021-12), System architecture for the 5G System (5GS); Stage 2 (Release 16), available on-line at https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144

[12] 3GPP TS 23.501 V17.3.0 (2021-12), System architecture for the 5G System (5GS); Stage 2 (Release 17), available on-line at https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144

[13] 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.

[14] K. Psounis, "Active networks: Applications, security, safety, and architectures," IEEE Communications Surveys, vol. 2, no. 1, pp. 2-16, 1999.

[15] 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.

[16] 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.

[17] J. Woodruff, M. Ramanujam and N. Zilberman, "P4DNS: In-network DNS," in Symposium on Architectures for Networking and Communications Systems (ANCS), 2019.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] Arista, "Arista 7170 Multi-function Programmable Networking.," 2018.

[29] 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.

[30] T. Swamy, A. Rucker, M. Shahbaz and K. Olukotun, "Taurus: An intelligent data plane," arXiv preprint arXiv:2002.08987, 2020.

[31] 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.

[32] 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.

[33] 3GPP TS 23.501, "System architecture for the 5G system." V15.12.0 (2020), Available online: www.etsi.org/deliver/etsi_ts/123500_123599/123501/15.12.00_60/ts_123501v151200p.pdf

[34] 3GPP TS 23.558, “Architecture for Enabling Edge Applications”, V17.2.0 (2021), Available online: https://www.3gpp.org/ftp/Specs/archive/23_series/23.558/23558-h20.zip

7. 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 flavor of Flexible Programmable Infrastructures.

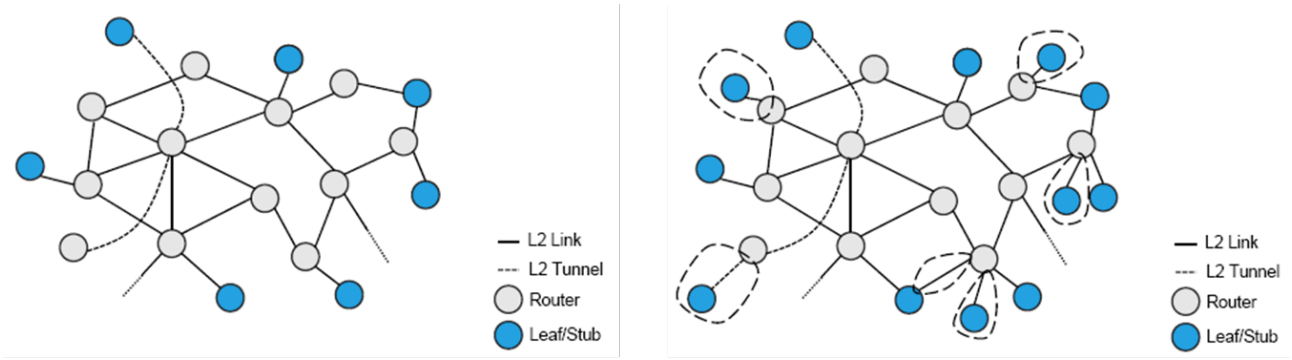
7.1. Scalable resource control

Figure 21a 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 21b. One leaf node gets an additional link and is multi-homed then. In Figure 21c 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.

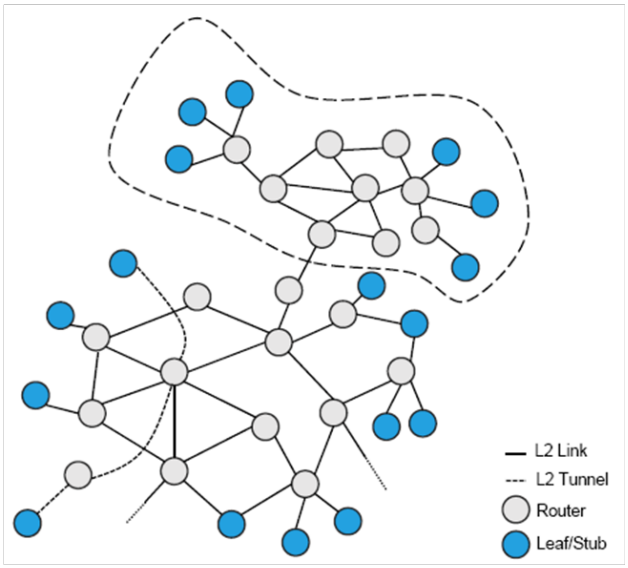
In our opinion, an essential precondition to realize this use case is resilient interconnection of all resources. This is to say, inclusion cannot come as a conventional, management activity that requires manual configuration or the added resources. It has to be dynamic, control-oriented, with each added 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.



a) Initial Topology, Stub/Leaf Nodes Do Not Forward Packets

b) Added Leaf Nodes and Links



c) Added a Complete Network, Leaf Node converted to Router

Figure 21: 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.

7.2. 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.

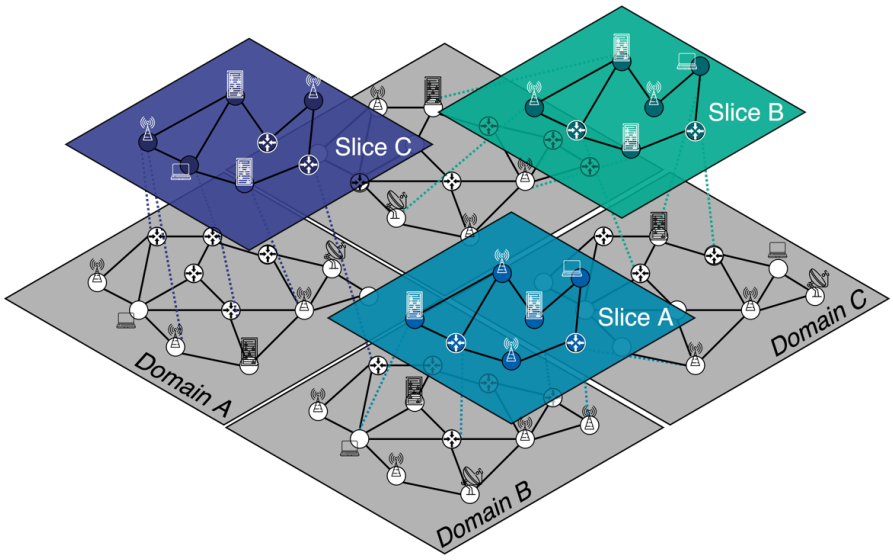


Figure 22: Next generation network slicing[1]

[1] This image has been designed using resources from Flaticon.com

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 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.

7.3. 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, in a distributed system entities 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 23, 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].

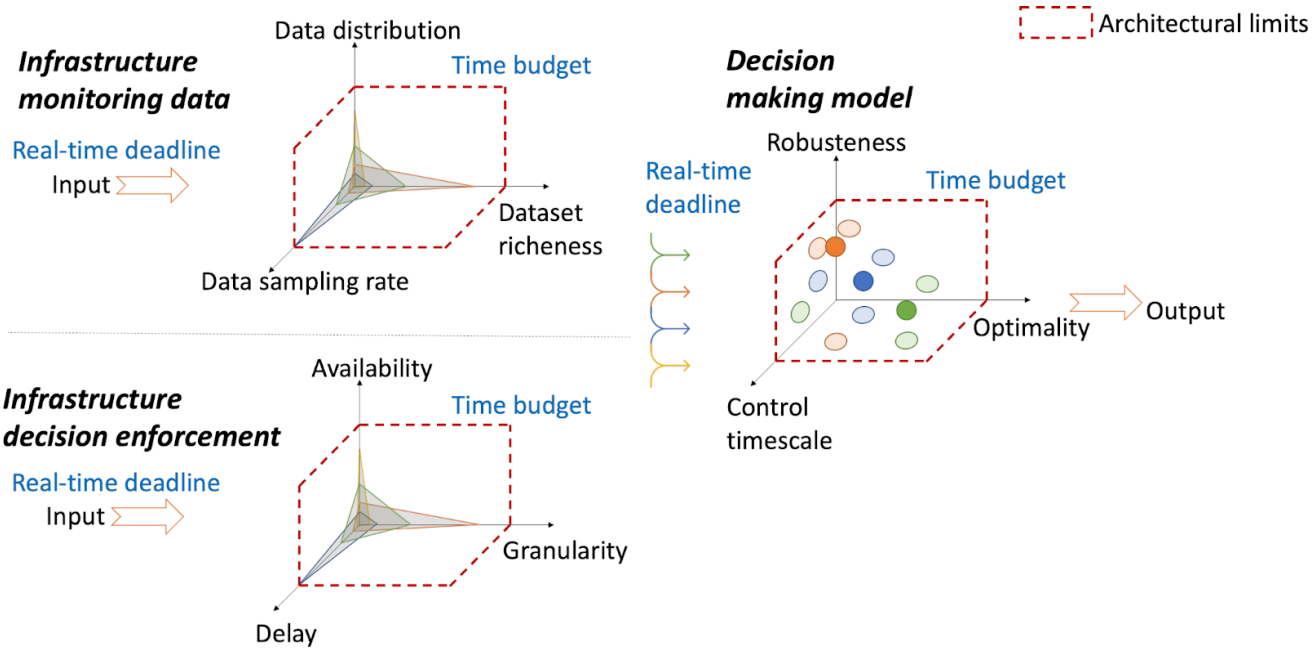


Figure 23: Network Intelligence Degrees of Freedom [24]

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.

7.4. 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 more dense, highly connected networking topologies is largely driven by the advent of data center 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 to R2/Kad 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 center 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 continuum, since developers require to go through a complex and cumbersome process of patch working and integrating different solutions [26].

7.5. 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 24. 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.

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.

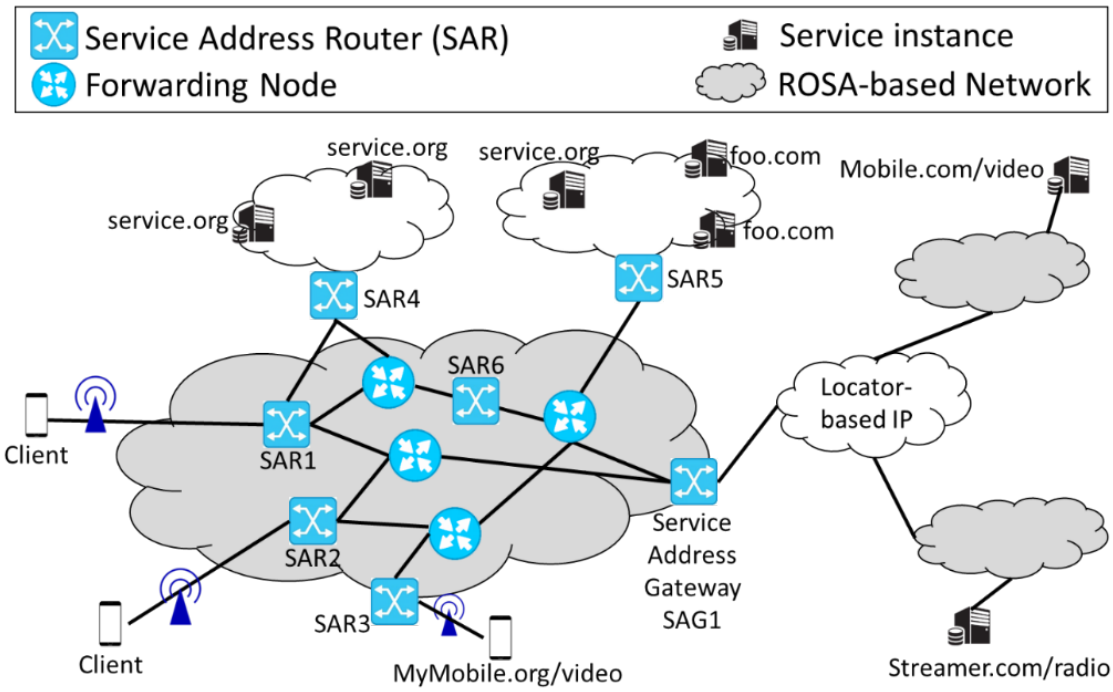


Figure 24: Service routing system overview

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 25.

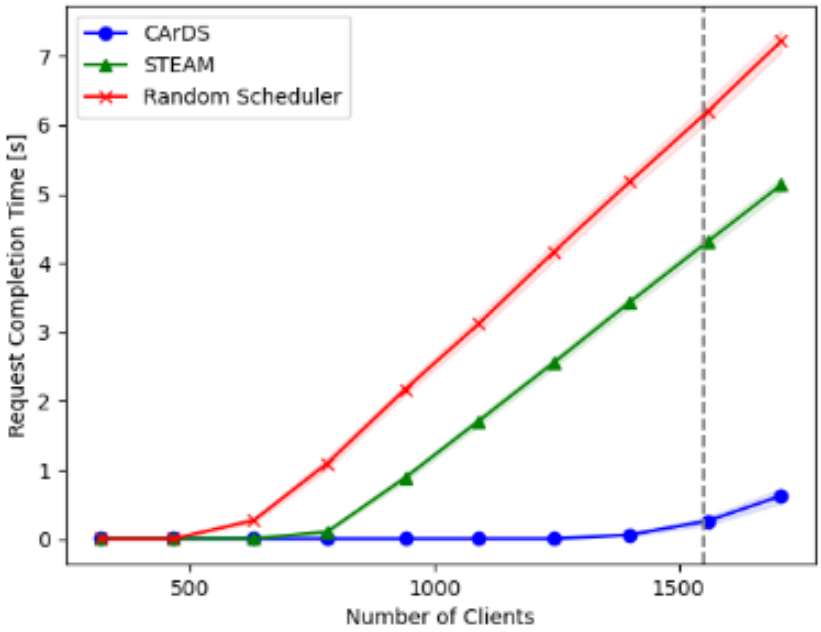


Figure 25: Runtime scheduling performance

7.6. 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.

8. 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 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 the user plane enhancements, critical to realize the 6G vision. Finally, network programmability and the need for holistic network architecture was addressed as the direction toward desired flexibility.

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 *(Huawei),* Claudio Paoloni *(Lancaster University).*
- **6G Radio Access:** Israel Leyva Mayorga *(Aalborg University),* Nikolaos Pappas *(Linköping University),* Najeeb Hassan *(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, Samer Bazzi, Tobias Laas *(Huawei).*
- **Integrated Sensing and Communication:** Andrea Giorgetti *(University of Bologna),* Richard Stirling-Gallacher *(Huawei).*
- **Distributed Federated AI:** Ramin Khalili *(Huawei),* Sokratis Barmpounakis, 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, Riccardo Guerzoni *(Docomo Euro-Labs).*
- **Flexible programmable Infrastructures:** Carlos Guimarães, Luca Cominardi *(ZettaScale Technology SARL),* Aitor Zabala Orive *(Telcaria),* Artur Hecker, Dirk Trossen, Zoran Despotovic *(Huawei).*

info@one6g.org

@One6GGlobal

one6g.org