# Strategic Research and Innovation Agenda 2022

Networld Europe

# Technical Annex

Version for public consultation

*Note: Some editorial aspects will only be finalized at the final version stage, after consultation process.*

## Table of Contents

# 1. Introduction

This annex to the Strategic Research and Innovation Agenda 2022 is an integral part of the white paper, but that is focused towards a more technically oriented audience. It discusses concepts and technologies essential for developing innovative services. The diversity of technological domains required for future communication infrastructures highlights the relevance of multiple innovation domains for European Research. In the white paper, a simplified version is presented, but in this annex multiple detailed aspects are covered. We have nine different chapters in this annex, which include:

- System Services aspects – overall system tradeoffs that need to be considered for the future, posing the stage for technology development
- System Architecture – analyzing the evolution of systems towards dynamically composed, multi-stakeholder environments, with an increasing softwarization and intelligence of the whole system, and the accompanying challenges.
- Network and Service security – discussing the paths on the increasingly relevant aspects of security in our infrastructure
- Software technologies – addressing the software related challenges of the ongoing network softwarization, the increasing system complexity, and the enabling of adaptive and customized services.
- Radio technology and Signal Processing – where the challenges and potential solutions perceived for the future wireless (and mostly cellular) communications are discussed
- Optical networks – a critical component of the backbone (amongst other potentialities) and its perceived evolution is detailed in this chapter.
- Non-terrestrial networks and Systems – discusses the upcoming closer integration of 3D networks into the overall communication system
- Devices and Components – tackles the unavoidable challenges at the fundamental element level, which will constrain and limit all system developments.
- Future Emerging Technologies – is a final chapter discussing promising technologies that may bring structural changes across all the current communication concepts. Some of these technologies are already being researched, but have not yet a clear path (if ever) to the transformational impact it is expected by their wide adoption.

Given the specificities of the different technologies, some slight structure differences exist across chapters, but the structure remains essentially the same across chapters, with the exception of the Future Emerging Technologies. Overall, this document was based in the previous version of the Strategic Research and Innovation Agenda [C1-01] as a baseline, which went through a long development process. The SRIA 2022 discussion started during the pandemia, with a public event in Lisbon, in November 2021, in the *Visions For Future Communications Summit*. This was followed by a long period of discussion inside the Expert Group of Networld Europe, where contributions collected from hundreds of experts in Europe, and where different key innovation stakeholders were directly addressed to provide comments. In a final stage, a public consultation was issued, and its comments properly reflected inside the final text. Overall, this SRIA has been the result of the work of a set more than 150 volunteers, coming from more than 85 entities. The Networld Europe community is in debt to all for their selfless efforts. The full list of technical editors and contributors is included in the last chapter of the annex.

# 2. System services and their description

*Editor: Holger Karl*

## 2.1 Introduction

### 2.1.1 Chapter goal

The technical annexes as a whole set out to identify research questions for future 6G (and related) systems: how best to provide the services that are expected from these systems? But to answer this question, it seems useful to lay out the expectation about these services: What are they, how can such services be described, how can they be used in a real system, about which of their properties should the system or the services themselves be able to give account? What level of services are we considering: from simple pre-built, pre-configured connectivity services up to a "meta service" to execute arbitrary services? Also, what is actually "the system" in such discussions? Is there a conventional system-versus-end user device dichotomy, with "the system" being under some sort of organization control? Will there rather be many systems, from the very large (like conventional mobile operator networks) to the very small (two or three devices that communicate directly without any support from other systems), but with the additional ability to smoothly merge and split into bigger and smaller systems? If so, what would corresponding services look like?

This chapter does not try to answer this extremely broad question, nor is it yet another attempt to list many requirements why a 6G system has to provide such-and-such an improvement factor in a particular metric – many papers have already provided many speculations on those properties of future 6G systems [C2-01 up to C2-08]. Instead, this chapter tries to reflect on the level of services – beyond simple link-layer or cell-level services – that should make sense for such a 6G system. This has implications for the scope of the overall system, possible basic architectural approaches, and interface styles. We collect this discussion in the first research theme (Section 2.2) of this chapter.

In doing so, we point out that we perceive a 6G system not only, or not even primarily, as a mere managed communication system. The push towards integrating communication, computation, storage, and sensing continuous to gain strength. And it is not limited to just integrating that but opening it up to a broader user community, not just for telco-internal operations or for a few select "verticals". Hence it seems advisable to indeed start from the old idea of "the network is the computer". This entails that we not only, and not even primarily, need to talk about communication services, but also about cloud-like services, with ideas from edge computing and related fields, when speculating about 6G systems. We discuss these and related aspects in Section 2.3.

We also point out that we expect the notion of non-functional properties of a service to gain more and more relevance; examples for non-functional property are trust and security, resource requirements, resulting performance, but also properties like governance.

to the point where they will become explicit parts of service descriptions. Section 2.4 deals with this theme.

Overall, this chapter will have a slightly different emphasis compared to the following ones. It will be less technical; it will not necessarily identify concrete research topics for all the discussed aspects. But we hope that it will provide a useful preface to the following considerations and act as a background for them. Before diving into the different research themes themselves, however, we want to elucidate some possible axes along which the following discussions can take place.

### 2.1.2 Discussion axes

Before delving into the actual research themes of this chapter, we want to point out several axes of discussions that are relevant for the following text.

#### *2.1.2.1 Systems and services*

Telecommunication systems have, historically and conventionally, been entirely asymmetric: there is a "telecommunication system", well regulated and well maintained, that provides a range of services to a "user". A user, on the other hand, would never offer a service to a system. There are multiple systems (e.g., multiple mobile network operators) that can offer services to users, but at any one point in time, a user usually only uses services from a single such system.

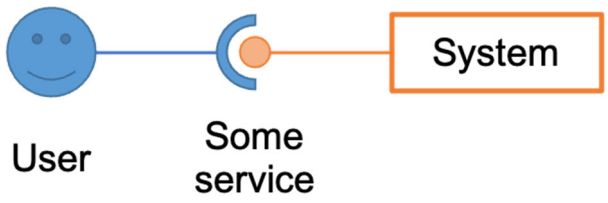



*Figure 2-1: User accessing a system's service*

Multiple systems might cooperate and provide and use services to and from each other; but usually, this happens on longer time scales and with more complexity than in a system-user interaction. These services can be of many different types: connectivity as such, ability to access computational/storage resources, delegation of responsibility to execute a given service in another system, etc.


*Figure 2-2: System accessing another system's service*

This distinction between interfaces between user and system/network on one hand and system-system interfaces on the other hand goes back at least to the days of ATM and ISDN (User-Nework Interface UNI and Network-Network Interface NNI) and is present in many system architectures. Some approaches (e.g., [C2-09, C-10]), however, questioned whether this is the only possible option. An alternative idea could be to just use a single type of interface, conceiving of even a single end user device as a system as well, and then thinking in terms of systems/networks merging and splitting.


*Figure 2-3: Mutually providing and accessing services between systems*

In the past, this idea has not taken root; it was considered complex with marginal benefits, blurring legal boundaries and responsibilities, and the division of labour between user and network was, after all, entirely clear. But we position that this is a promising approach that deserves to be reconsidered, in particular given current developments in universal sensing, isolated networks working without access to infrastructure in an ad hoc manner and still able to integrate with said infrastructure once in contact, or the growing capabilities of end-user devices (think of a car with substantial computational power for autonomous driving that goes idle when the car is parked [C2-11]). In the long run, we believe that it is worthwhile evolving our understand of systems (or resources) uni-directionally providing services to users towards services (or resources) being provided and used between different participants on a more equal footing.

Such a more generalized understanding of service provisioning should also have benefits for the typical use case of so-called over-the-top providers offering services to end users. Clearly, this happens by a network providing both OTT and end user with services (connectivity, in-network caching …). But evolving towards a more equitable footing should make it easier for network, end user and OTT to mutually profit from their resources and services (e.g., a user can provide sensing services to an OTT, an OTT could provide ML training services to a network). All these options exist today, in principle, but often feel awkward and hot fixes at best, rather than a consistent architectural approach.

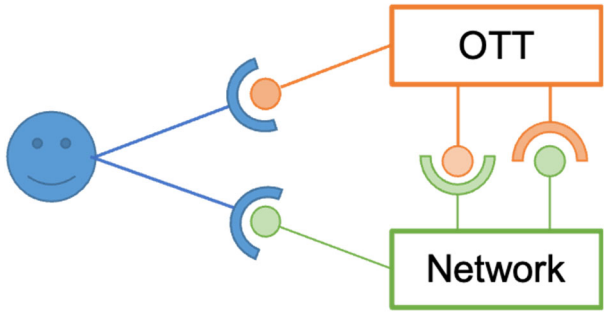



*Figure 2-4: Users and various systems offering and accessing services*

#### *2.1.2.2 Metrics*

Few terms are less precisely defined and yet often used as *metric*. By a metric, we mean a *qualitative or quantitative* figure of merit of a system, a service, a component, etc. We juxtapose the term metric with that of *parameter*: a property of a system or component under consideration whose value influences the metrics. Some parameters are fixed (e.g., speed of light clearly influences a metric like latency, but cannot be controlled), others are under our control. Often, the research task of the subsequent sections, which deal with the design of the system and its mechanisms, will be to find parameter settings that optimize or at least improve metrics (sometimes, parameters that are changed during such an attempt are called *factors* to differentiate them from unchanged parameters). This discussion is sometimes complicated by the layered nature of our systems, where

one layer's parameter might be an underlying, i.e. realization, layer's metric. When focusing on a concrete layer or component, the differentiation is, however, usually unproblematic.

For a metric to make sense, we require them to have typical basic characteristics. First, a metric must be valid: does it measure what we are actually interested in (e.g., CPU utilization is a notoriously invalid metric as it is easy to utilize a CPU with nonsensical work, driving utilization to 100%)? Second, it must be *reliable* in the sense that it must measure in a consistent and stable manner, do we get the same result if we measure the metric twice, possibly within bounds of statistic uncertainty? (Obviously, this is not the same thing as reliability as a metric, in the sense of the probability of a system being continuously operational in the interval [0, t] provided it was operation at time 0 [C2-12]). Third, a metric must be objective, i.e., the metric value must not depend on who executes the measurement. Fourth, a metric must be usable in that that is must be practical and feasible to actually measure it.

Since not all metrics have equal relevance, the term *Key Performance Indicator (KPI)* has become popular to emphasise some metrics are particularly important; however, we stress that the decision which metrics to regard as key in this sense highly depends on scenario and circumstances, on business models, or can perhaps even be just a matter of taste. The use of KPI is also not uniform; sometimes, non-quantitative metrics are ruled out as KPIs, but we do not follow that narrow interpretation. Also, we do not regard KPIs as a relative term ("20% improvement") but as an absolute term, which can of course be used to determine relative improvements if the KPI as such is quantitative.

We do point out that is usually a body of knowledge and established set of metrics and terminology in multiple fields (e.g., dependability [C2-12]) that is often ignored, and metrics are reinvented under different names (e.g., common words as dependable/available/reliable are commonly misused due to different interpretations). Obviously, this leads to needless confusion.

#### *2.1.2.3 Trade-offs*

Usually, a single metric will not be able to express all aspects of interest. We need to consider multiple metrics, and often, it will not be possible to convert multiple metrics into a single one, e.g., by just computing a weighted sum (what would be the weighting factor for the metrics "total number of users served" vs. "reliability of a single connection"?). Instead, in such a *multi-objective optimization problem*, we need to consider the *Pareto front* of all solutions to figure out the *trade-off between metrics*: How much merit in one (or several) metric(s) do we have to sacrifice to obtain an improvement in some other metric?[1]

Related, but often confused, is the notion of *trading off parameters* against each other: When a metric depends on multiple parameters, it can be possible to obtain the same level of merit by different parameter combinations, effectively substituting one resource against another (e.g., energy efficiency could be maximized by trading off transmission power against code rate). To emphasize again, confusions between these two trade-off concepts often arise here when parameters of one layer are metrics of an underlying layer, bringing extra complexity to the simply

[1] Weighted sums of metrics only find the full Pareto front if both objective functions and decision space are convex.

definition of terms. This means that it is necessary to be very precise in identifying which layer is under discussion, lest we carelessly mix aspects of different layers or wantonly move perspectives.

We shall do our best, in all these annexes, to make it as clear as possible which type of trade-off we are discussing.

#### *2.1.2.4 Level of risk and level of surprise: Evolution vs. revolution*

A simpler question is the level of surprise: does a problem require evolutionary or revolutionary work? Can it perhaps be solved by either approach? If a revolutionary approach is necessary, how deep does the revolution go; is an entire *clean-slate* approach across multiple system levels necessary or can the revolution be contained?

Closely related is the level of risk: Revolutionary approaches might hold greater benefits if they succeed, but the risk of failure might also be larger. Obviously, public funding for research, and research in general, plays a natural role in supporting revolutionary, high-risk/high-reward approaches compared to incremental, evolutionary technology development.

Related to this question is the question of backward compatibility: Even if a system is realized in a revolutionary manner, as long as this happens in a black-box style without harming external interfaces, it might not really matter. The true challenge then is whether or not a revolutionary approach can or should be backward-compatible and when does it become advisable to rethink old interfaces and cut out dead wood.

This discussion, of course, departs from mere technical decisions, requiring complex analysis of business perspectives and economic feasibility.

Similarly, the following chapters and their research themes might not all be on the same time horizon. Some work will have fairly quick turn-around into products, other work is decidedly basic research with a long-term perspective.

#### *2.1.2.5 Level of contestation*

In the following chapters, we will see discussions of many different ideas to fulfil a broad range of requirements and achieve many different goals, driven by many different stakeholders. It is only natural that such a collection is not uniformly agreed upon, least of all that everybody agrees on relative priorities. There is a natural *level of contestation* for requirements, where some are easily agreed to by everybody (“networks should use less energy”) and others must be regarded as niche scenarios from one perspective and as highly relevant from another (“mobile networks must provide 100 Gbit/s to every user”). Often, it is a societal negotiation process to find a consensus on priority rankings, as well as an economic one (where economic success is, often, just another such requirement that must be balanced against sustainability). We absolutely do not claim to have the answers to such a negotiation process at hand, or even to know how to best organize it in many cases; but we will try to point out crucial areas where we perceive high levels of contestation.

#### *2.1.2.6 Area of impact*

While often relatively clear, we shall also point out the areas that are impacted by a particular research theme: is it mostly a technological question, does it have regulatory impact, does it pertain to political or legal decisions, or could it have broader societal interest? Typical aspects in this context are security, privacy, a government’s desire to sacrifice an individual’s security and privacy

expectations for some supposedly greater common good (e.g., for so-called lawful intercept or for censorship), etc. It is worth noticing that there is a trend in wireless communications towards privacy solutions without caveats, e.g., not enabling any possibility of tracking a user by design. Example of such approaches can be found in the recently created IEEE 802.11bi group on Enhanced service with Data Privacy Protection.

After these preliminary remarks, without any further ado, let us start the discussion of research themes for system services and their description.

## 2.2 Research Theme 1: Basic assumptions

### 2.2.1 Aspect: Scopes

A system like 6G has to support a wide range of scopes. This entails, of course, the well-known ones from 5G, i.e., enhanced mobile broadband (eMBB), ultra-reliable low-latency communication (URLLC), and massive machine-type communication (MMTC). While it is tempting to now prefix all of these and similar scopes with “ultra” or to arbitrarily increase requirements like going to very high user densities, this is vague, unspecific, and at best intended to intimidate the reader – at the very least, a more precise description of quantitative changes to requirements is required for a well-founded discussion.

We believe that such mere quantitative changes to requirements are not sufficient to warrant an extensive research effort for a next-generation system; they could very likely be covered by evolutionary innovation. Besides, there seems to be a significant risk of increasing such metrics for their own sake, without clear justification from real, end-user requirements. Instead, we think that qualitatively different scopes are more promising, and we also see substantial chances here in a simplification rather than complication.

Consider a scenario of two users communicating directly with other, in near vicinity. While there is support for this today even in 5G (device-to-device), it still requires support from a core network to establish such communication. Here, 5G and cellular networks are clearly not competitive with local communication technologies like WLAN.

But just closing such a competition gap is probably not sufficient justification, either. We intend to extend that core scenario (Figure 2-5) – enabling the communication of two devices – to an approach that can incrementally grow such network islands, merge them together, split them if needed, and provide services (not just communication services) to their members. A key architectural enabler could be *emergent networks of networks* (“emergent networks”, for short): networks that work without dedicated, pre-decided or pre-deployed core components, that can provide whatever is needed as “core-like” functions on their own, and merge that with other networks, possibly even with a network using a conventional core. Challenges here are how to identify relevant functionality to be available on individual devices, how to deal with inconsistent state, or how to deal with identity management in such a scenario. This idea is not limited to core functions, of course. Similar ideas would apply to RAN functions, with approaches like control/user plane separation pointing the way towards many forms of disaggregation, and with CRAN already paving the road towards a disaggregation-by-design approach (e.g., thinking of a Reconfigurable Intelligent Surfaces system in a disaggregated context). Down the road, we could think of end terminals which, supported by some level of edge computing, may deploy a set of supporting pre-

defined functions to enable new modes of communication in a dynamic, ad hoc way. In the end, disaggregation and functions-on-demand could be distinguishing features of a 6G system; they require some of the facilities discussed below.

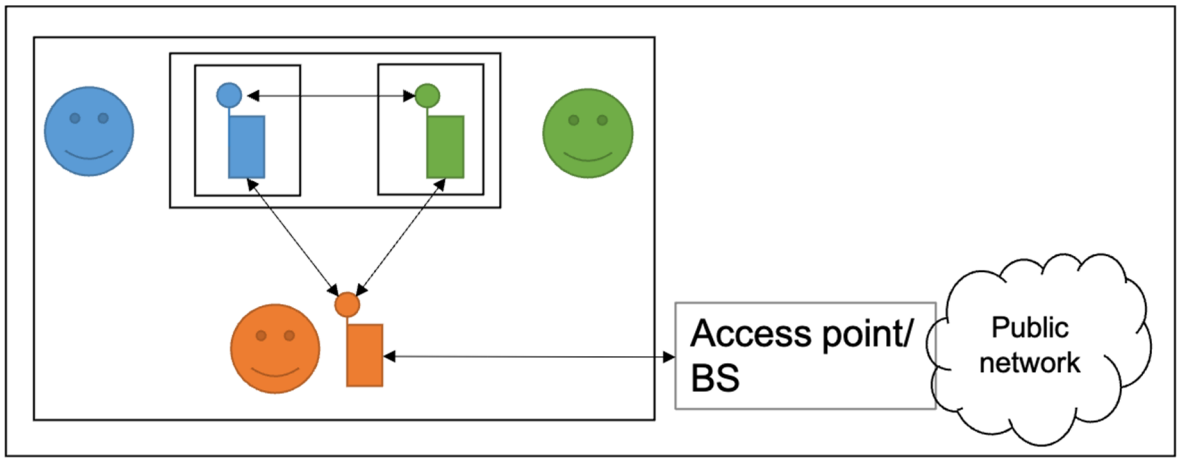



*Figure 2-5: Bootstrapping networking communication from two devices; note the absence of a core (figure to be made nice, of course)*

In addition to pure communication tasks, we foresee a 6G network to strengthen the trend to deeper integration of communication and computation – the old "the network is the computer" will become truer. This will straddle and overcome the fog-edge-core cloud distinction, removing those silos and turning them in a *cloud continuum* (see below), and resulting in a single, pervasive computation/storage infrastructure that can provide services anywhere, with a consistent API. This continuum will extend to end user terminals and to constrained devices on their surrounding area, extending the cloud continuum to account for locality of data and contextual information. Important also to consider is the current trend towards the integration of sensing on communication networks. Through these new technologies, the network extends its senses into the physical word, being able to obtain direct information of it. This enables not only new services, but also the capability of the network to interact with the real world. For example by the control of appliances, the network may be able to interact with the human reality even modifying human behaviours by interaction with their context and observing the result.

Given trends in modern software engineering, we expect that most services to be deployed will be in the form of microservices (graphs of individual components), very similar to network function chains. As soon as these microservice components are multi-tenant capable, we need to think about supporting services for entire user populations and not just single users. Typical challenges here are dealing with multi-domain multi-user situations, charging models when services are shared between different users, or dealing with potentially widely varying resource requirements different components (e.g., consider a component of a user application that occasionally runs machine-learning training algorithms inside a network). Section 2.3.3 will go into more details.

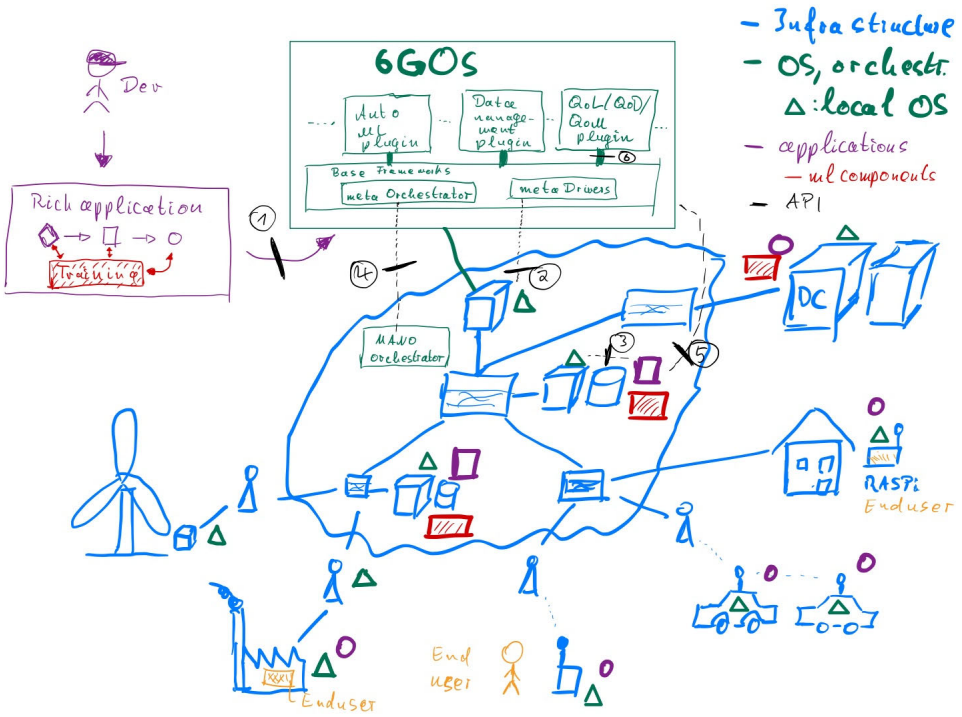



*Figure 2-6: Deploying services to a cloud continuum*

In summary, we see questions of scopes appearing with respect to at least functional scopes, application scopes, geographic and administrative scopes.

### 2.2.2 Aspect: Architectural tenets

The past generations of mobile cellular networks as well as the Internet, data centres, and probably most IT systems have suffered from a common problem: their monotonously increasing complexity. It seems doubtful that we are on the right path if we simply intend to make 6G and even more complex system as 5G. To combat this trend, we suggest to keep a few basic architectural tenets in mind:

- **Simplicity is better than complexity**, less is more. That pertains to all phases of a lifecycle, from conception, development, deployment, and operation. But accept that there is likely a tradeoff between simplicity and the desire to support new business models/scenarios.
- **Resist feature creep**. Simply adding more features to a system does not necessarily make a system better – in fact, it might be that now is a time for a "**no new features**"[2] moment in the evolution of networks, fixing our attention on the fundamentals on top of which all our systems live. Instead, it seems promising to identify *orthogonal feature sets* that are independently develop-able, deployable, operable, and usable. For example, the idea of emergent networks from above is such a case: devices that create a network on their own have all the features that are needed, in many scenarios. Adding features must not increase complexity for unrelated feature sets.
- **Latency and energy matter**: While obvious, it is worthwhile to stress these two metrics explicitly as they create friction, possibly even contradictions to the previous two tenets.

[2] Apple keynote during the introduction of the Snow Leopard operating system.

- **Explicit accountability** means that a system can give account for its actions and, for example, explain the resources used to provide a certain service (e.g., energy consumed or CO2 produced to stream a video) or that a network can explicitly tell about its coverage rate, the cost it charges, etc. This is tightly coupled with the next point.
- **Open interfaces and automation** are well established ideas but not well supported in practice. There should be automatic interfaces to enquire a network about where it provides coverage at which typical rates, or where a regulator can allow private campus networks at which cost, for which duration. This needs to be machine-readable and automation friendly. The following section provides more details.
- **Ownership, functional dis-/aggregation, and implementation are orthogonal** issues. While sometimes intermingled in previous architectures, we point out that the way functions of a network are designed and organized, their practical implementation and execution, and their legal ownership (and responsibility) are three (almost) orthogonal issues. That means that (1) it could be an architecture decision to provide a certain functionality by single or a few functions or be a finely defined set of such functions, breaking down said functionality into very small pieces; (2) that in either case, these functions could be implemented in a very centralized manner, in one extreme integrating a lot of functions into a single executable running on a single, central machine and in another extreme, running a large number of executables widely spread out, and (3) that the question of which commercial entity owns and is responsible for which functions and executables can also be decided differently. Obviously, some combinations of these three design issues are easier to imagine than others, but we should be very careful not to mix up questions of functional design, implementation, and ownership just because legacy architectures perhaps did not carefully distinguish between them.
- **Ownership and control are orthogonal issues**. Similar to above, we point out that who owns a system (e.g., single end device, network element, software component, or even an entire network) is not necessarily the one who controls it. Ownership might entail an initial right to control, but even on very low protocol levels, control is often ceded immediately, e.g., in a schedule-based TDMA MAC protocol. The idea of relinquishing control over an owned resource is at the heart of many concepts like virtualization, slicing, virtual networks, etc. We hence believe that future architectures should be more explicit in how control is ceded or retained, either explicitly or implicitly. This might mean that interfaces for such control establishment, federation, delegation, fusion, release, etc. might become an integral part of the system service understanding.
- **System evolvability** is required: it must be possible to add, at development/deployment or runtime, missing components to a system. As long as they are orthogonal to other feature sets, they should not interfere with them. Similarly, it must be possible to roll back such evolutions and to go back to an earlier state. Cloud applications have successfully embraced very rapid rollouts of such evolutions (in the spirit of continuous integration/continuous deployment); perhaps, the high dynamics of these environments does not directly apply to network environments, but there are surely some lessons to be learned even for telecommunication systems.
- **Separation of concerns** is necessary since it appears impossible to find one solution that optimizes all possible requirements and takes all possible constraints into account. While simple, robust, general-purpose solutions are clearly attractive (and IP is the prime example for the success of that approach), it seems questionable whether it is possible to create that single, perfect solution for all possible niche cases. The architectural and

mechanism/protocol design question to solve here whether a single solution using all resources is better or worse than specialized solutions using subsets of resources – specialized solutions should be better on the same set of resources, but statically subdividing resources is a clear disadvantage even from just queueing theoretical and diversity arguments (the core problem of the "slicing" idea). This question is driven by the ongoing extension of the simple, uniform mobile access use cases of the past to the incorporation of potentially widely different requirements of various verticals. In consequence, we do not subscribe to the idea of **one size fits all solutions**, but we are also keenly aware of the downsides of resources divisions and we think research on approaches to balance these pros and cons is needed.

- **Resist early partitioning of functionality**. It is tempting to early on divide up the functions a system should provide and put them into certain categories - a good example could be to early on label functions as "core" or "access" and use that to also partition devices into being "core" or "access" devices, with clear and simple binding who does what. This approach is time-tested, but with the continuously increasing complexity of demands and scenarios, it might come to its limits. We do encourage resisting such temptations and to keep an open mind about overcoming such functional and structural silos. For example, it might turn out to be beneficial to not distinguish devices along the core/access axis, but rather to think about which functions are needed, where, and when; possibly, we might no longer need such a differentiation at all but are, of course, still able to provide all required functions without early on deciding which functions are provided in a centralized, core-like manner. Only further research and deployment experience will tell.

As a consequence of these considerations, we question whether it is indeed conducive to research 6G as a single architecture. We raise the question whether it does not seem more promising to think of it as a **lean architecture toolset**, rather than a single, bloated, monolithic architecture. Such a toolset needs to encompass all essential phases of a lifecycle, at the least design/development/deployment/runtime. The later chapter on system architecture will consider the consequences from this observation.

It makes a 6G system a much more fluid entity and incorporates the evolution into the system design – in a sense, getting closer to the idea of "the last generation" that was already popular at the start of the 5G discussion. Clearly, this is contentious, and it is indeed not clear whether this is advantageous. Which means that it is a research question.

### 2.2.3 Aspect: Interface and interaction styles

A lesson to be learned from the cloud computing success story is that there is beauty in simplicity for interface and interaction design. Key insights there seem to be that a simple interaction style is key, with Create-Read-Update-Delete (CRUD) and the corresponding representational state transfer (REST) approach being a primary example. Finer-grained versions of such architecture concepts (e.g., Command-Query-Responsibility-Segregation, CQRS) appear quite amenable to the needs of a flexible, toolset-oriented 6G system.

It is worthwhile to point out that approaches like REST are often equated with a very overhead-rich transfer syntax, e.g., JSON transported over HTTP for simple API invocations. But this is only an implementation option; it is not mandated by the architectural style as such. We hence argue that it is promising to stick to such successful interaction styles instead of trying to invent new ones just for the sake of being different.

Having embraced such interaction and interface styles, opening up such interfaces should become much simpler than sticking to proprietary or telco-specific interfaces. That does not imply, of course, that these interfaces are open in the sense that everybody is allowed to make all calls; checking access rights is of course an inherent part of all such architectures.

We encourage work on 6G to also embrace standard interface definition standards and popular toolchains (e.g., Swagger) for their own definitions. Again, the reuse and ease of integration in existing automation systems should vastly pay off in efficiency gains.

### 2.2.4 Research Challenges

| Research Theme | Basic assumptions | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Scopes | Short | Decision whether to think of 6G as just a (collection of) radio access networks/technologies or a broader service execution infrastructure<br><br>Decision whether to pursue incremental improvements or radical reorientation to different scopes | Clearer orientation of research and development efforts |
| | Medium | Emergent networks of networks; networks that work without, e.g., core functions when possible but incorporate them when necessary. | Simpler and faster to deploy, less management overhead, easier to grow. |
| | | Continuum of resources: Radical concept to overcome resource silos (no edge vs. cloud; no end device vs. core). | Higher resource efficiency, simpler systems. |
| | | Coherent approach to sensing, controlling, storing, and acting; feasible to use via microservice chains/chains of network functions. | Simpler application development; opening up new markets by new application types |
| Architectural tenets | Long | Provides explicit accountability for resource usage and achieved metrics. | Improve competition, give consumers informed choices. |
| | | Embrace open interfaces with ownership/control separation, improve automation. | Reduce friction in system deployment and operation |
| | | Find minimal deployable architecture sets and means to extend them as necessary. | Reduces overhead, reduces time to market for new scenarios. |
| Interface and interaction styles | Short | Apply lessons learned from cloud systems to telecommunication systems | Broadens workforce, streamlines overall development |

## 2.3 Research Theme 2: System services

### 2.3.1 Aspect: Forming and emerging a network

Prior to using a network, it needs to be formed. To simplify the exposition, discussion, we first distinguish real and virtual networks (of course, when virtualization is done properly, such a

distinction should usually not be necessary). Forming a real network can start from as simple a scenario as getting two devices to exchange data (see above, emergent network) and growing it by incorporating more and more devices and functions. During that process, rules for data exchange are explicitly or implicitly negotiated: medium access, routing (=computing routing tables) and forwarding (=using existing routing tables to send a packet onwards on the next hop), managing identities, roles for membership, access rights to resources, providing resources for storage or computation to the network, etc. Much of this has been a very traditional research topic; much has not really been considered, e.g., running services inside such a network.

A more interesting case happens when two already existing networks come in contact with each other; in fact, when we conceive of a device as a degenerate network, this case is already covered. Historically, this idea has been around at least since the early 2000s when it was considered in a project called Ambient Networks; perhaps it is time to revisit it again and going beyond the mere connectivity issues that were considered in Ambient Networks (irrespective of which use cases might be considered under the related term of "ambient networking"). While establishing mere packet forwarding is comparatively easy to achieve, harmonizing and integrating, e.g., membership management or access rights, ensuring that certain services run exactly once inside a network, etc. are not at all clear (and might even be theoretically impossible in the strict sense, requiring to figure out pragmatic, workable compromises). Another example could be mobility management: once the network grows, this function likely will be required, but it is not really clear how and where to provide that. A conventional approach might be to assume that there is a "core" that provides these functions, but in the scenario considered here, such a core would not exist but such functions are still needed. Various options then exist, e.g.: a) on-demand self-assemble a core (e.g., by dynamically electing some devices into that role) to mimic how functions are provided in conventional networks, or b) provide required functions in a "coreless" manner, without fixing the locus of some functions to a subset of devices. Figuring out which functions actually are needed in such a case is an additional challenge, specifically when this is not possible to do beforehand but must be figured out at runtime. We emphasise again here one of our architectural tenets: how to provide certain functions, where to provide, and who owns them are three orthogonal issues, which becomes clear in this example scenario.

The key takeaway of this discussion is that the **required functional set should not be hard-wired into an architecture** (and even less into specific roles, assigned to specific devices), but that it should be decided and realized based on requirements and concrete situation. We consider this a natural evolution of the generalization of functions initiated during 5G; non-public networks in 5G were a first step into that direction.

Turning our attention to virtual networks, we find that this is indeed a very similar idea and process. We need to manage a network, on top of some resources, and these resources need to be somehow managed, but that is really the only key difference. Issues like installing and locating services are largely identical inside real and virtual networks, if these concepts are set up right. This captures the creation and management of slices as a special case. Again, if done right, we believe that this could result in considerable simplification of our current, highly complex architectures, improving their resilience and maintainability on the way.

The distinction between real and virtual is, of course, not either-or. Actual systems are likely to have both real and virtual resources. In a sense, this is as should be: a virtual resource's interfaces, after all, should be indistinguishable from its real counterpart's interfaces. On the other hand, it is quite likely that dynamics in real and virtual resources happen on different timescales; e.g., changing a virtual topology can happen in seconds, a real topology change might take weeks or months. This is aggravated by those changes potentially happening inside subsystems (e.g., inside a local operator or at an Internet exchange) without full visibility or upfront warning to using systems but must not harm the services running on-top.

In addition to forming a network, the idea of "re-forming" or adapting a network is natural. While a lot of research exists and is ongoing on network adaptation at multiple layers (e.g., adapting topology of optical wavelength-division multiplexing networks, adding drones to a cellular network on demand) and while it is even common practice to do so today on long time scales (e.g., on demand adding basestations to locations of large-scale events), there is basically no work available on concrete, practical descriptions of what kind of adaptation a network supports, how it could be requested, how to pay for it, what additional capabilities of the network would result, what might have to be sacrificed, how long an adaptation would take, how much energy it would consume, etc. Basically, there is no commonly agreed adaptation API, across multiple layers, for either network-internal or network-external use.

### 2.3.2 Aspect: Providing access

We distinguish here between an administrative/contractual aspect and a technical aspect of providing access.

Once there is a notion of a "network", the notion of "client" or "user" emerges as well. Then, the question of how, whether, and to where to provide "access" to/via this network to a given user arises. Conventionally, this is well understood by establishing contracts (e.g., between mobile operators and private end users or between multiple operator networks via exchange points). Based on such contracts, secrets (shared, private, ...) of multiple kinds are used to conduct authentication, authorization, and accounting for such network access.

While this process is well established (with plenty of research on all kinds of these stages), it seems insufficient for more advanced scenarios as discussed here. We will need processes to figure out the identity of a network as such, how to negotiate contracts automatically, to automatically describe what access means and to what and in which quality, and how to police and monitor the fulfilment of such contracts and their payments in scenarios where networks themselves take on a more fluid identity than is typically the case today.

The technical aspect of providing access is, in an xG network, closely associated with the idea of wireless access. Obviously, there is plenty of research on advanced wireless techniques ongoing and planned in the 6G context; regarding the technologies involved, we refer the reader to several of the remaining annexes of this document. Examples include, but are not limited to, questions of terahertz (THz) communication, visible light communication (VLC), multiple access, coding, cell-free massive multiple-input multiple-output (CF-mMIMO) zero-energy interface, intelligent reflecting surface (IRS).

The service description aspect of technical access for such a broad range of technologies is, nevertheless, challenging as well. We foresee the need for a network to give better account of its access capabilities than what is common practice today. We would expect a network to provide an API that, at the very least, can express:

- At a link level: Which technologies are available, supporting which types of scenarios (e.g., movement speed)? What link metrics are supported (data rate, latency, packet errors, …). Typically (but not exclusively), we expect link-level access APIs to be used for instantaneous access and decisions. An example question that could be answered is how far a beam tracking-based link layer could profit from voluntarily provided movement prediction of a mobile terminal.
- At a network level: Which targets are reachable, at what quality (data rate, latency, packet errors, …) and at what cost, using which types of transport? Which technologies are available, supporting which types of scenarios (e.g., movement speed)? What guarantees are provided for (virtual) private networks or "slices"? How many additional users can the network handle; what is available capacity? Typically (but not exclusively), we expect such network-level access APIs to be used for mid- to long-term planning of network access.
- At a system level: where is which coverage available, which low-level services (e.g., compute, storage, ...) are available at which cost, are high-level services provided as well (e.g., authentication, ... - roughly analogous to IaaS vs. PaaS vs. SaaS) how to get subscribed, which partners are accepted, etc.
- What kind of energy supply and total power efficiency is achieved (compare below)?

Clearly, this is based on existing ideas for SLA negotiation, but we expect the need to more explicitly differentiate between link- and network-level access APIs, as well as between instantaneous vs. longer-term APIs.

Also, we need to consider a basic fact of current technologies: basically all technologies are incompatible with each other. It would be desirable to have a common set of technologies that can interact with each other, adapting to a minimum set of negotiated capabilities. This is need for a truly integrated network.

### 2.3.3 Aspect: Transporting bits and transporting data

Transport bits is certainly the bread-and-butter activity of a network, and many ways to describe these transport services have been investigated in the past, ranging from plain old sockets to modern access abstractions like 0mq (https://zeromq.org/). Similarly, various ways to express quality of service and quality of experience requirements have been investigated. We do not foresee any specific needs to go beyond these APIs over what has been discussed in the previous section on access provisioning.

In addition to the mere transport of bits, transporting more meaningful bits or data is also a basal function of many networks. Examples include carrying voice calls, support for IMS, transporting data for timing, localization, or AAA services as well as providing these services as such (clearly, this starts to form a form a gray area with the following aspect). For all these types of data of services, well-established APIs seem in place.

### 2.3.4 Aspect: Network as a Computer – Execute arbitrary services

Unlike mere data transport, the idea of using a network as a service execution and service provisioning platform is (although nearly as old) much less settled. In research so far, the description of such a "service execution service" has focussed on how to describe a service as an artefact comprising multiple, individually executable components, connected into a graph (cmp. IETF SFC model); how to describe the deployment units of such components; and how the concrete details of onboarding, monitoring, using, … such a service should look like. Provided such services to users is tightly related to required service quality metrics such as dependability (security, privacy, reliability, etc.) and performance (latency, throughput, ...), intrinsic to the service definition. QoS requirements must be properly stated to realize or at least monitor them.

#### *2.3.4.1 Services and components*

Sometimes, careful distinctions are being made between components that work on lower layers of a network stack (often called network-facing services or virtual network functions) vs. higher layers (often called application-facing services or microservices, depending on the heritage of the particular discussion). We stipulate that this distinction can be useful when thinking about concrete execution techniques and planning resources (e.g., packet forwarding inside a switch as well as some forms of in-network machine-learning would profit from a P4 environment, whereas a general-purpose web server component would not[3] [C2-13]), but from a service description perspective, we question the necessity of differentiating between these aspects for a service as a whole.

Rather, we suggest to think of this distinction as a property of a component, annotated as meta data that gives information about what kind of resources that component can run on top of. Possibly, a component can use multiple different resources and comes with multiple, corresponding deployment units, as has been suggested in the context of multi-version services.

But on a service level (in the sense of an entire graph of components), we believe that it is rather useful to think of a service as comprising both kinds of components – e.g., low-level components like packet-based load balancers as well as high-level components like an application server. This will enable an application to package all its own components as well as its entire network stack (where needed) and customize networking as well. It is then up to an orchestrator to properly manage such mixed-level services.

#### *2.3.4.2 From silos to a continuum of resources*

Despite mentioned above already, we believe that it warrants repeating: resources should not be divided up prematurely. Currently, there are various "resource silos": the cloud outside the network operated by hyperscalers, some in-network clouds owned and run by mobile network operators, some edge devices possibly owned by a factor owner, and finally resources on end devices. This is a convenient structure well matched to existing business models, but it also jeopardizes optimal solutions and is likely inferior on many metrics, e.g., latency, energy consumption, overall cost. We believe that instead, it makes sense to move on from such silos to a "continuum of resources" (computing, storage ...) [C2-14] where the best resources are used for each service.

[3] Canini, Switch ML,https://sands.kaust.edu.sa/project/switchml

Realizing a resource continuum has aspects ranging from design time (where to provide additional resources?) to runtime (which service to run where, a classical orchestration problem), contractual (how to share responsibility between owners of different parts of that continuum), legal, or application-specific (e.g., ability to acquire and apply knowledge using context awareness; working within and across different vertical sectors like IoT)?

#### *2.3.4.3 From stateless functions to state management*

A key differentiator for such services is their handling of state. In the simplest case, they are stateless: they do not need to remember anything, all information necessary for a processing step is included in the data unit currently available. Depending on the level, this data unit can be a packet, an entire data flow, or an application unit (e.g., a video frame). The key point is that no management of state is necessary.

Such statelessness significantly simplifies orchestration, management, operation of components; it is one of the reasons behind the popularity of corresponding frameworks like Amazon Lambda Functions. But it is also a significant restriction of expressiveness: Some scenarios or applications simply make little sense without managing state.

Once state is present, however, it needs to be managed: it needs to be kept consistent when multiple updates happen in parallel from multiple writers, it might have to be kept persistent, it might have to be migrated when the place of function execution changes, etc.

For individual network functions, state management has been researched for a number of years. For larger graphs of functions, as well as for the collection of multiple such services, we believe there is still work to be done. This entails the design of such concepts, where inspiration can be drawn from systems-of-systems approaches from software engineering, or questions of managing storage, which has somewhat different usage modalities than CPU or link capacity, which is typical focus of attention in much orchestration work.

#### *2.3.4.4 From single user to user populations*

Orchestration and in-network service work has, to a large degree, focused on the idea of a service being instantiated separately for each individual user. While that is attractive in many circumstances (and might be a good approach to deal with some privacy aspects), this idea is challenged once the notion of a "user" is no longer exactly the same as an individual entity but, e.g., could represent an entire network itself. Also, it is often useful to share state across multiple instances of the same service, e.g., when provisioning a cache – a cache typically only makes sense if and when we can reuse activities of other users.

To support such scenarios, the notion of *multi-tenant-capable* services/components exist: realizations of components that can deal with multiple users at the same time, have sessions open to multiple users concurrently. In fact, this is the normal mode of operation for many components; just consider a typical web server as an example. What seems to be missing, however, is a clear idea how to identify which information is shared across users. Moreover, the orchestration of multi-tenant instances is much more challenging than dealing with an individual instance per user. A mixed-mode setup where different services comprise both single-tenant and multi-tenant-capable components, some of which are shared with different subsets of services, becomes particularly interesting and actually seems like the most natural description of a typical software ecosystem.

Closely related to multi-tenant capability, but actually a separate issue is the question of what we assume about our user population. Do we consider each user separately, and provision service access for each user separately (possibly using multi-tenant capable components)? Or do we additionally assume that we have some knowledge about the entire population of users of a particular service? E.g., do we know that a particular service is used mostly in the evening hours, but with a world-wide distribution of users (and usage shifting following a regional diurnal pattern)? Such additional information about user populations holds the promise of better resource utilization and faster, even predictive reaction to changes at runtime, but clearly, this is a more complex model to deal with and needs to consider state at runtime not just for an individual user but for an entire population of users.

Typically, such knowledge about the usage patterns of a population can only be stochastic, but even deterministic models might be available in some specific cases (e.g., industrial automation). In either case, the opportunity would be to take better

So far, we have largely ignored that opportunity of leveraging such advance, user population-wide information for the orchestration of services. Advance information would allow advance action, e.g., to provision deployment units were they are needed, to provision computational or storage resources where needed (e.g., by migrating other data to remote storage to free up capacity where needed), or even to adapt networks (see above). Instead, so far, we have refused this proactive orchestration approach and have embraced a reactive-only approach, migrating and provisioning only when needed. We conjecture that this was partially driven by cloud infrastructures, which are indeed mostly reactive, but there, the inherent latencies are quite different from what happens when executing services in a widely dispersed network. We believe that it is time to work out whether such proactive orchestration is worth the effort in return for an effectively shorter and more agile service provisioning. This entails work on corresponding mechanisms, figuring out whether population-wide predictions are accurate enough, and work on suitable APIs to provide such descriptions for a service, e.g., during onboarding of a service.

#### *2.3.4.5 Geolocated services*

A key motivator for providing services inside a network in the first place (as opposed to doing so conventionally in a remote cloud) is to lower latency. For that, a service must execute at the right location. But what if at the desired location no infrastructure for service execution is available (and also cannot be dynamically provisioned), but only users and service demand? And users move around, and the group of users at a given location constantly changes?

A promising idea is to think of that volatile group of users, with constantly changing composition, nevertheless as an infrastructure that can run services (compare e.g. the V-Edge idea). Clearly, issues like service discovery or state management become much harder, as does fairness, payment, etc. – and it is also not really clear how to describe such a service ("this service needs to be provided in this geographic position, with an availability of x% and a maximum latency of %ms, assuming that in this location, user population evolves according to the following model"). We speak here of *geolocated services* (not to be confused with geolocating services).

#### *2.3.4.6 Conventional and ML-based components*

As a last aspect of providing services inside a network, it makes sense to consider what type of services there are; more precisely, what type of work the components of a service do?

In the simplest case, a component executes a conventional program. It could work on a packet stream, work on an HTTP request, or even on a symbol stream inside the signal processing of a CRAN-style BBU hotel setup. For such components, we already have a reasonably good understanding how to describe them and how to orchestrate them.

But what if these components use machine-learning techniques? Again, a simple case is a component that only uses inference at runtime, using a precomputed model to, e.g., classify traffic. For such components, orchestration techniques are likely very similar to today's approaches.

The real challenge occurs when a component starts to also train a model, inside a network. Then, the workload will likely have drastically different characteristics. Training could happen in batches, with large spikes in required computational resources; it could be distributed, requiring significant amounts of data to be transported; it could be triggered by the accuracy of a model falling below an acceptable threshold (even determining that threshold and detecting its violation is difficult). While there has been significant progress in distributed machine-learning techniques (e.g., federated learning), there does not appear to be a concerted effort to give a proper description of such learning-based services, a description that would be amenable to be used in orchestrating and managing such services, pertaining to their computational, communication, and data storage needs. Basically, today, we are unable to orchestrate such services inside a network or to even decide whether to run a service inside or outside a network, or in hybrid manner, efficiently distributing work between edge, in-network, or remote cloud resources. We believe that this gap needs to be closed, urgently.

We emphasize that this question pertains both to services that a network runs on behalf of its customers as well as to services that a network uses for its own internal operation (e.g., to orchestrate its own resources, making the problem a challenging case of introspection and recursion). We are, however, in this chapter not concerned with the specific machine-learning techniques that can be applied to various scenarios; this is to be discussed in the following chapters. We also emphasize that, despite ML currently enjoying a substantial wave of attention and popularity, we do not subscribe to the belief that ML will be the cure to all needs or that it is the unavoidable solution; we are convinced that a healthy dose of scepticism is still the hallmark of good science and that conventional techniques and approaches can and will outperform ML in many situations; it is for upcoming research to figure out which approach is superior, when.

### 2.3.5 Aspect: Diversity and Convergence

As alluded to already, there is a trend to endow 6G systems with even more radio access technologies, ranging from body-area networks to satellite networks – sometimes called *SuperConvergence*. This raises the challenge how to harness this diversity in access; how to make sure that a usefully integrated communication experience emerges, rather than the impression of hopping from one specialized access point to another, which might only happen to be sold under a single brand name? In other words, how to ensure that the radio technology diversity actually converges?

It even raises the question *whether* this is indeed a promising and useful proposition, whether it is indeed necessary to, say, converge body-area networks and satellite networks into a single system.

Of course, different levels of integration are possible, each of them with different trade-offs and benefits. Deciding the level of integration rests on many trade-offs – costs and benefits of integrating an ever-wider range of diversity – that cannot and should not be answered beforehand. We strongly believe that it should be a research *result* rather than an *axiom* whether to go for such super-convergent systems. This requires bringing together research from technological, use case, and business model backgrounds.

### 2.3.6 Research Challenges

| **Research Theme** | **Basic assumptions** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Forming and emerging a network | Medium | Concepts to bootstrap a network out of components and subnetworks | Simpler operation of new scenarios; increased robustness (in particular, recovery) |
| | | Architectures without hard-wired functional sets or roles | Easier management, update, innovation. |
| | | Make adaptation or reconfiguration a first-class citizen with proper interfaces | Easier management, update, innovation. |
| Providing access | Medium | Network architecture that does not distinguish between user vs. network. | Simpler operation of new scenarios; increased robustness (in particular, recovery) |
| | | Simplified service interface for access links; easier specification of required link capabilities and automatic selection of useful technology (or negotiation of alternatives). Similar for entire networks (connectivity to where, under what conditions). | Incorporate new technologies easily, without having to reinvent service access points. |
| | | Make energy consumption transparent. | Prerequisite to achieving carbon reduction. |
| Network as a Computer – Execute arbitrary services | | Richer service descriptions, from hardware requirements (from P4 or FPGA bitstreams up to VM images) or execution patterns (simple stateless firewall to complex ML training application). Overcome pre-segregated resource silos. | Realize continuum of resources and services; new markets for telcos; better service quality in new applications types. Better resource efficiency by always using the best resources, rather than being confined to pre-established silos (“at the edge”). |

| | | | |
|---|---|---|---|
| | | Explicit support for state management across all system layers and tiers, state management as a service. | Simplifies and integrated development of network and application functions. |
| | | Make user populations a first-class abstraction in 6G, allowing to think of providing services not just to individual users but to whole populations. | Increase resource efficiency and application service quality; make service quality more predictable. |
| | | Geolocated services: pinpoint a service to certain areas where it should be accessible, even if the set of devices changes. | Realizes new application types. |
| Diversity and convergence | | Converge across multiple dimensions (hardware, platforms, location, layer). | Simpler deployment with better QoS/QoE and better resource utilization, reducing cost and energy consumption. |

## 2.4 Research Theme 3: Non-functional aspects

The previous theme was mostly looking at a functional perspective: which functionality should be made available, accessed how, described how? Some of the aspects were already non-functional, in the sense that we touched upon typical performance metrics like throughput, latency, error rates, …. This research theme revisits some of the aspects mentioned above, and adds additional ones, but focuses on a *non-functional* perspective: Which aspects of services beyond mere functionality and simple performance metrics should be made available as part of a service description and ensured by service execution? In doing so, we omit a list of typical performance metrics in this section (like throughput and latency); these metrics are on one hand well understood and on the other hand, it is difficult to predict how the requirements for these metrics will develop (lacking concrete requirements, all we currently have at our disposal is stating “ultra” or “hyper” requirements, which still need to be made more concrete).

### 2.4.1 Aspect: Accountability and meta-data

Conventionally, networks as well as cloud systems were designed from an asymmetric perspective: the user was considered dangerous, but the network/cloud was considered trustworthy and beyond reproach. Clearly, this idea no longer holds up to scrutiny today, and it will be even less tenable with the flexible ideas of how networks could be formed, as outlined above.

Instead, we expect any entity in a network to at least provide some explicit accounting of itself and its operational principles. For example, a network or an over-the-top application should be able to answer questions like 1) do I follow BGPSec, do I only accept secure routes? 2) how much censorship is applied to content travelling in my network, or 3) what is my coverage area?

Obviously, the answers to such questions as such must not be considered to be trustworthy, either, but it forms an interesting research problem (e.g., drawing on incentive design from game theory) to make sure that networks/applications to answer such questions at least semi-truthfully (building

on existing incentive mechanism work). Nonetheless, that should not be misinterpreted as networking becoming trustworthy; clearly, every system is trust-unworthy, and the key research challenge really is how to build semi-trustworthy systems out of such components.

And equally obviously, the interpretation of such answers will depend on the legal and regional context in which they are given, on the region. As a blunt example, a mobile operator in an autocratic country should have no other choice but to answer “yes” to the question “censorship?”; today, there are no means to even formulate the question. We endeavour to provide means to at least ask such questions.

### 2.4.2 Aspect: Guarantees and flexibility

Providing guarantees – on the maximum latency of a data communication, on the deadlines of a service execution – is a time-honoured problem of communication and computation engineering. Approaches like Integrated Services or Differentiated Services, with many descendants, have been introduced and have seen varying degrees of success. A recent iteration has been the introduction of “slices”, which – among other properties – attempted to provide such guarantees. In its simplest form, it did so by statically allocating resources to a given set of users and their communication relationships. Slicing did, at its core, leverage the fact that guarantees for an aggregated user base have the potential to scale, where guarantees to individual users would not (compare the scaling shortcomings of IntServ). But it is a well-known result of basic queuing theory that such static allocations are very costly in the amount of resources necessary to provide any such guarantee. And since under realistic assumptions, it is anyway only possible to give stochastic guarantees, it is much more cost-effective to exploit *stochastic multiplexing* to reuse those resources across multiple users, slices, etc.

Slicing, in this example, indeed evolving from its first form in 5G of functional packaging (even without providing at least fixed resources) to an approach that embraces stochastic multiplexing, making it a much more tenable proposition. However, it is only one such example: We need to make this trade-off between level of guarantees and flexibility explicit and allow a customer much better control over what type of guarantees, on what aggregation level these guarantees should be given, and which flexibility trade-offs are acceptable. This should, as has been correctly approached by slicing, happen for communication, computation and storage in an integrated manner.

### 2.4.3 Aspect: Resources and performance

When approaching guarantees, it is useful to understand what resources are required to provide a certain guarantee, as well as how often these guarantees will be needed. The second point has already been discussed in the context of the service descriptions (see above), arguing that at least a stochastic understanding of how often a service will be requested, from where, would be beneficial.

Understanding what resources are needed to provide certain guarantees is necessary for strict guarantees, but even for loose guarantees (e.g., soft real-time behaviour with relaxed requirements), it is useful for up-front provisioning of a service. For example, when a stochastic load description tells us that it is likely that the request rate is soon going up (e.g., in the evening for video streaming), it is straightforward to provision additional service instances; but to know how many instances to provision, we need to know how many video streams one instance can handle

when running on a particular type of hardware. The alternative – reactive provisioning – is perhaps acceptable in a cloud environment but not in wide-area networks with its inherently limited performance or latency.

Such information can be made available for concrete components; they map arrival load and resource type (e.g., which type of CPU) to achievable performance (e.g., average latency per video frame for a transcoding service). This information is often called a *performance profile*. Figuring out such profiles is difficult and resource consuming (e.g., via test benches), and it is not clear how to provide such profiles during an onboarding process to an orchestrator or why an orchestrator should trust them. Alternatively, it is conceivable to update such profiles in a continuous learning process, collecting data from ongoing service execution and continuously improving the accuracy of such profiles. Once available, these profiles need to be properly used in orchestration processes that can deal with their inherent inaccuracies and with the inaccuracies of predicted or currently measured load. First ideas here exist, but we do not have a full lifecycle approach from service design to profile-aware orchestration at hand. Also, a careful comparison against reactive-but-overprovisioning based approaches is missing.

### 2.4.4 Aspect: Energy consumption and climate footprint

The energy consumed and the carbon dioxide produced for computation and communication purposes has become an increasing concern. We start a discussion of this aspect by some basic observations:

- Energy consumption as such might quickly turn into a misused indicator for climate footprint, with cloud providers heavily pushing into all-green-energy, if only for image purposes. Even though, energy consumption certainly still stays important form a cost perspective.
- Reducing energy consumption and reducing climate gas production are not necessarily the same thing. It is easy to conceive examples where solution A consumes less energy (e.g., by computing locally) but has a big greenhouse gas production (by using non-renewable power supply), whereas solution B uses more energy (e.g., by having to transport data to a remote data centre) but its greenhouse gas impact is smaller (by supplying only relatively power-modest optical communication from non-renewable resources and powering the remote data centre from all-renewable resources). Obviously, these two metrics need to be considered together and cannot replace each other and simply knowing all local energy mixes does not answer the question whether solution A or B is preferable (indeed, a typical example for two different Pareto-optimal solutions).

It is not clear how to balance these two (exemplary) aspects practically (and needs exogenous aspects to decide); the conceptually best approach is of course to report both metrics separately and then let the metrics' user decide how to interpret that. This is related to ongoing work in EC's Sustainable Product Initiative (SPI)[4] and the Digital Product Passport[5], with natural extensions to digital services becoming necessary.

The following question is then about what to report such energy metrics (compare SPI's information model)? A first idea could be to report on the execution of a single service request: How much

[4]https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12567-Sustainable-products-initiative_en
[5] https://ec.europa.eu/commission/presscorner/detail/en/ip_22_2013

energy was used to execute it? That is more or less available from operating systems via suitable system calls, but there is no plausible approach to deal with such meta-data becoming available at runtime. Thinking, for the sake of example, in terms of remote procedure calls or REST interfaces, each function result would then be accompanied by such energy-related meta-data. This is conceivable, but we have no conceptual or tooling support for that at hand.

The problem, however, is compounded by at least two aspects. First, collecting energy data just for a single service (and all its components) is likely not good enough. We need to take into account all the background services that are needed to execute the actual service at hand. For example, this could pertain to collecting and analysing data in a Kafka-style system. It gets aggravated once we are using machine-learning systems: the training overhead needs to be amortised over the service invocations that profit from that trained model, but how to do that while we do not know for how long, for how many service invocations that model will last? There is no straightforward idea how to do that.

Additionally and similarly, a second aspect to consider is that also hardware needs to be amortized. Simply put, we not only use energy while executing a service, it also needed energy to produce the involved hardware – similar to CAPEX and OPEX, but for energy. And this hardware energy up-front investment also has to be amortized over the hardware lifetime and the service requests that it can execute (e.g., such a total energy perspective over system lifetime questions the wisdom of shutting down hardware to save energy; they are still writing off their production cost even while they are inactive).

The description of this aspect has focused on service execution. The same ideas apply, *mutatis mutandis*, to networking. In addition, we will need more networking-specific metrics (e.g., expressed as Joule/bit/m or Joule/bit/m/s). And we need to work out how to integrate that into the operation of a system – e.g., can energy efficiency or “greenness” become a routing metric; could it be expressed, e.g., via BGP?

And similar to above, it makes sense for a network (or a cloud) to be able to give account of itself: for example, on average, how much CO2 is produced by this network to transport 1 Mbit of data between its customers? Averaging here can take place over various user population, time, etc. – here, a lot of research and requirement identification will be necessary to identify truly meaningful ways of expressing such properties.

Finally, certification will become a relevant question here. It is easy for a network or a cloud to claim all kinds of things, but how to ascertain that these claimed properties are actually true? It seems worth investigating whether existing “energy labels” can be extended to cover such aspects. The ongoing European Sustainable Product Initiative to provide such information to all products has a realistic chance of happening and being extended to digital products/services as well. Hence, it is prudent to investigate means to realize such information in early-on research.

### 2.4.5 Aspect: Trust and values

The issue of trust has been alluded to already in previous aspects. We want to emphasize a key understanding here: All *components* are inherently not trustworthy. No end device, no router, no base station, no edge server, no piece of such deserves, as such, any particular level of trust. It is

then the challenge to build a *system* that still can be trusted to a reasonable degree, with all due caveat and care nevertheless still being applied.

Techniques for achieving that are discussed in a later chapter. In the context of service descriptions, it is in fact difficult to come up with a good idea how to describe trust in a service. Clearly, have a network/cloud/service itself claim that it is trustworthy per se is a laughable idea (notwithstanding the fact that this is more or less what we are currently doing). Also, requiring certain trust levels from a system is questionable without a means to policy and monitor any such promises (e.g., asking a network NOT to censor content posted on it will only result in an emphatic "of course" as an answer, with censorship going on nevertheless).

In addition to these more technical aspects, we also have to recognize that trust/ trustworthiness is intimately linked to values. And while we might, in principle, agree on a certain very limited set of values, the fact of the matter is that values have become a highly subjective and regional property – one country's values are another country's anathema. Hence, since values are no longer a cross-border property, neither is trust. By extension, this is no longer the case between domains or other actors. All this reinforces the basic premise: no system is trustworthy.

### 2.4.6 Aspect: Negotiations and integrated billing

All the points above - functional aspects in general as well as non-functional aspects like accountability, climate footprint or trust - come with a price. Today, that price is essentially fixed from an operator to a customer, and there is no automated negotiation going on (there is at best manual negotiation on long time scales). While the idea of automated negotiating contracts for SLAs is an old one indeed, it is conceivable that this will take on a more pressing need in an 6G network-of-networks, where the resources required to proved a service are dynamically composed. Correspondingly, the price for such a service should be dynamic as well, and a range of options should be made available to an application to choose from. Possible choices could include between "best transport" vs. "cheapest service", "most dependable" vs. "least climate footprint" etc.

The real challenge here will be to compose not just the individual services, but also the individual bills, ending up in a single, integrated bill towards an application, comprising all resources from all parts (e.g., geographic) and of all kinds (e.g., network, server, energy).

### 2.4.7 Aspect: Governance

Related to trust is the question of governance of a network (as the rules used to regulate a system). Internet governance has received considerably more attention in recent years, for example, with the tussle around ICANN and the different perspectives on regulation rules from different political backgrounds. While this is, so far, more an Internet-centred discussion, we foresee a chance that this discussion will spread to 6G as well once the realization sinks in that 6G is not only (or not even predominantly) a mobile access network but a much broader service provisioning infrastructure. Political developments here are difficult to predict, but it seems necessary to closely monitor developments in this area.

### 2.4.8 Research Challenges

| Research Theme | Basic assumptions | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Accountability and meta-data | Medium | Networks can give account of their behaviour, properties, guarantees, and regulations via explicit interfaces. Incentives for truthful accounts are necessary; level of trusts in such accounts need to be assessed. | Improve competition by better transparency; open up market mechanisms. Improve trust, foster basic values in communication. |
| Guarantees and flexibility | Short | Tuneable options to select guarantees vs. levels of statistical multiplexing (e.g., at the granularity of "slices") | Better cost control for users, better resource efficiency for customers. |
| Resources and performance | Short | Proactive understanding of required resources and how they depend on load patterns (as opposed to reactive-only approaches). | Shorter reaction times to load/system changes, improving QoS/QoE. |
| Energy consumption and climate footprints | Medium | Explicit "bill of energy" and "bill of greenhouse gases" for any service execution | Important mechanism to assist in saving the planet. |
| Negotiations | Medium | Develop negotiation and billing mechanisms for a network-of-network, extending to all types of resources and different cost types | Simpler yet more powerful means for applications to request broad range of services |
| Trust and values | Long | Work with systems that are inherently not trust-worthy; make trade-offs and risks more explicit. Deal with incompatible value systems. | Education aspects, also for laypersons. |
| Governance | Long | Evolve and possibly converge governance models of different networks (Internet, cellular); find compromises between different value systems. | Contribute to keeping a world-wide communication system possible, despite political tousles. |

## 2.5 Conclusion: List of service families

We conclude this chapter by identifying certain families of services/interfaces that can be offered and consumed by a system – be the system an entire network or a simple end device. We emphasize again our idea that overcoming the asymmetry between networks and devices might be a pivotal step forward to a richer yet simpler overall system, with a lot of special cases disappearing into a few, uniform interfaces. Hence, the following list will not use a distinction between, e.g., "provider services" and "subscriber services".

Clearly, some services are very conventional, e.g., exchanging data between two end points; we only briefly hint at those. Also, it is of course not possible to go into details for all aspects of these service

families, or even to specify these services in detail. We hope this will serve as an incentive for future work!

- Namespace management and address bindings
  - Create a namespace and obtain a name for the namespace itself, with rules how to obtain a new name (e.g., uniqueness required in what scopes, if at all?)
  - How to map names between different namespaces by (a) creating addresses by binding a name in one namespace to another name in another/the same namespace) and (b) looking up addresses (obtain the mapped name). There should be multiple options to map between the same pair of namespaces; it should be possible to explicitly provide and deploy these mapping functions; rules which functions to use when and where should be configurable as well.
  - Multiple namespaces should be attachable to networks.
  - Closely tied to authentication and authorization
- Network creation services
  - Create a network, starting from a single device, grow by adding more devices.
  - Create networks at a given location/area, for given duration, with limited or public accessibility, with configurable rules how an end device can join a network, etc. Compare campus networks/non-public networks.
  - Generalization: Join two networks (with end device joining a provider network being an important special case). Questions in addition to all of the above: How to deal with state? Which state needs to be merged, how to deal with inconsistencies, make rules for state management deployable as well.
  - How to split a network into two networks. Similar questions: How to deal with state? Which parts of the state space need to be copied, to where?
- Connectivity and communication
  - Provide all typical types:
    - Uni-/multi-/broadcast, group communication with various ordering constraints, etc.
    - Streaming, datagram, dependable or not, congestion controlled or not, with different QoS parameters (maximum delay or jitter, minimum data rate)
    - Use inspiration not only from old-fashioned interfaces (e.g., sockets) but also modern APIs (e.g., 0mq); provide richer, easier to use communication and connectivity APIs
  - Scope: Locally vs. globally, communication possible even without connectivity to a core infrastructure; merge ad hoc and cellular communication in a single API. Overcome 4G D2D / 5G sidelink constraints.
- Introspection and inspection
  - From a device perspective (not only end-user device): Which networks does a device belong to?
  - Network introspection: who belongs to me, where are they, what are my traffic statistics, what can I reach at what cost/trust/… levels, network telemetry. Aggregate this information in variable fashion (hierarchy of introspection).
  - Inspection of a network from the outside: what is your coverage? How expensive are you? what QoS can you provide (connectivity, computation, storage)? What type of power supply do you have (battery, black/brown/green power, energy mix of … %); which parts of that power are you willing to share? What type of accelerators are

available, where (topology and location)? Again, aggregation in variable, hierarchical form.
    - Notes:
        - Introspection and inspection clearly overlap and it might be a matter of taste to which area a particular interface is mapped.
        - With a device being a special case of a network, all these introspection and inspection services apply to individual devices as well.
        - Obviously, authentication/authorization checks apply.
        - Twins might be a promising implementation technique for these services, but that is outside the scope of a service description.
- General storage capability
    - Capacity and cost: How much storage can be provided, at which cost models; this would be analogeous to an IaaS model?
    - Types: bulk, streaming, object, file, …;
    - Suitability: for virtual machine images, for machine-learning training data, for video, …
    - Speed: required read/write speeds, or speeds for other access models (e.g., random access speeds)
    - Location: Globally, localized, with delay or locality as a requirements parameter. Control over location (topological or geographical) or freedom of choice (with delay or locality as a parameter, at what relative costs?)
    - Dependability requirements: what guarantees are provided (think availability, reliability, …), e.g., how much redundancy is applied, is there an automatic failover, what is the concurrent access model (note CAP theorem)?
    - Other QoS parameters like a function stating expected size as a function of write rate
    - Specialized semantics, like KV storage
    - APIs: Offer storage, request storage, acquire, release, grow, shrink, migrate, open/read/write/update (depending on storage type).
- General compute capability
    - Think: network is a cluster/hyperscaler that can host and run different deployment units, for different types of workloads; this would be analogeous to an IaaS model.
    - Flexible with respect to deployment units: Virtual Machine, Docker container-like artefacts, functions-as-a-service e.g. provided as a WebAssembly, even up to bitstreams for FPGAs
        - With usual security properties: signature for integrity of the file, …
    - APIs: offer capacity, upload, start, stop, pause, reset, upgrade, migrate, …
        - Offer and request can/should be parameterized by type (general-purpose CPU, GPGPU accelerations, … ) and speeds. Similar to existing cloud-style APIs (comp. e.g. Terraform).
        - Explicit knowledge about type of an execution artefact (e.g., batch processing, streaming requests, ML training, …)
    - Globally, localized, with delay or locality as a parameter
    - With QoS parameters as usual (speed, memory, data rate, …), described as functional relationships if known (for a data rate if r requests/sec, I need 5*r + 10 cpu cycles)
    - Dependability requirements: what guarantees are provided (think availability, reliability, …), e.g., how much redundancy is applied, is there an automatic failover, …?

  - Control over location (topological or geographical) or freedom of choice (with delay or locality as a parameter, at what relative costs?)
  - Note: relevant special cases here are (a) computational offloading from end device to edge cloud / core cloud or (b) running jobs on mobile devices moving through an intersection, smoothly moving the job and its state from one mobile device to another
- Service deployment
  - Combine general storage and general compute into a more meaningful abstraction: Deploy service function chains (analogous to moving from IaaS to PaaS/SaaS).
  - Basic APIs: upload, start, stop a service, push deployment units where needed; place components, route traffic between components; migrate, scale out/up/in/down individual component functions; deal with state migration when scaling out/in.
  - Go beyond existing ideas by looking at population of users; rich workload descriptions ("ML training job that can tolerate the following model inaccuracies, measured as follows: …; it uses the following algorithmic data exchange pattern")
  - Dependability requirements as above.
  - Note: If done correctly, no need to differentiate between a B2B service ("Netflix running inside an MNO") and an end-user device offloading processing to an edge cloud!
  - Closely related to sensing and actuation, which could figure as special functions inside a service chain, could nicely integrate IoT.
- Sensing and Actuation
  - What kind of sensing capabilities are available?
    - E.g., precise localization for end device, …
    - E.g., Radar-style based on THz communication
    - Open issue: taxonomy to describe that? Is it just another namespace, with multiple, mutually independent namespaces coexisting (harmonizing to a single, universal sensing namespace seems impossible); how then to translate between these namespaces for sensed data (if that is even necessary)? Is de facto standardization for typical cases good enough?
  - Where should the resulting sensed data be sent to? What is typical size, data rate, latency requirement? Energy consumed for a sensing task? To which service does such a sensing task belong (comp. service deployment above). Who owns which data; who can be trusted with which sensed data?
  - Similarly for actuation: What actions can be taken, where (e.g., open/close a radiator valve); what are dependability requirements, how to check whether an action has actually been performed ("never trust an actuator"); what are fallback actuators/actions for a given one? Is it possible to turn this into a service execution problem, or are additional mechanisms necessary?
  - Closely related to introspection/inspection and security services (in particular, AAA).
- Security services
  - Authentication and authorization for all of the above
  - Possession proofs for data
    - And proofs of deleting data? Possible?
- Auxiliary services
  - Negotiations and billing: Provide means to negotiate for SLAs, covering networks of networks owned by independent entities, covering all types of resources (network,

radio, server, storage, ...) and cost categories (money, energy, ...). Provide an integrated bill covering all resources owned by any entity.

- Location (necessarily geo-location?)
  - Position proving service (system gives you a time-limited certificate that the terminal was at least within this circle at this time)
    - How to ensure privacy? How to protect that from secret services or criminal police?
- Time: Provide a common time reference, possibly with adjustable accuracy (e.g., relaxed requirements in pure ad hoc modes, stringent ones in a cellular mode of operation).

# 3. System Architecture

*Editor: Artur Hecker*

## 3.1 Evolution of Networks and Services

Distributed computing has taken a significant step forward with the development and utilization of the Internet in many industries, pushing the digitization of processes and opening opportunities for creating or improving many business-to-business (B2B) and business-to-customer (B2C) processes. It does so, however, on the back of an Internet, whose core design started in the 1970s on very basic assumptions of an end-to-end connectivity between two remote machines, usually denoted as *client* and *server*. Inter-domain connectivity, enabled through the overall IP suite, allowed for reaching any machine through a multi-tier architecture of autonomous systems (ASs). This basic principle, unchanged to this day, had to shoulder the burden of *service routing*, i.e., associating a request to an instance of a service name, supported by newer innovations such as content delivery networks (CDN) albeit still relying on separate indirection architectures to the basic IP packet delivery. Some of these limitations are currently being addressed in the evolution of the future of the IP protocols, with different protocol innovations being pursued in different frameworks (e.g. [C3-1][C3-2][C3-3][C3-4][C3-5][C3-6], among many others)[6].

While unchanged in principle, many things have evolved from this basic picture of Internet connectivity. In the following, we differentiate three aspects, namely the *nature of communication* over the Internet, the *nature of services* (and their relation) and the *nature of provisioning* in the serving endpoints that are being reached via the Internet.

The *nature of communication* over the Internet has changed significantly from the single-client-single-server model. Today, many such servers are hosted in large-scale *data centres*, exposing services via a data centre's internal routing mechanisms to the wider Internet – here, the client communicates to the data centre (over the Internet) rather than the server directly, said data centre serving as a *point of presence* (PoP), enabling a service provider to host the service without having to own or operate their own resources. In recent years, those PoPs have been moved closer to end users in an attempt to reduce costs (e.g., for inter-domain transfer) as well as latency (by being closer located to the relevant users), particularly for services such as over-the-top (OTT) video or social media. This move has been driven by large-scale service providers, such as Google and Facebook, but also by *content delivery networks* (CDNs). These companies have deployed their own PoPs and, by selling excess capacity, have established themselves as large cloud players. By pushing data centres towards the network edge, communication in the Internet has significantly concentrated on the customer access networks with, for instance, an estimated 61% of Asia Pacific Internet traffic expected to being served through CDNs alone by 2021 [C3-07]. Netflix's estimated 15% share of the Internet traffic is mostly served through localized PoPs [C3-08]. Extrapolating this to other content platforms (e.g., Amazon, Disney+, as well as country-specific platforms such as BBC iPlayer), we can project the amount of traffic originating and terminating in customer access networks to be easily around *90% of the overall generated traffic* downstream to end users. In

[6] We expect that the increased impact of vertical (e.g. society) requirements will further constrain the evolutions on the Internet protocol.

essence, **the nature of communication has moved from servers towards services, the realization of which, in turn, moves closer to the end-user.**

When it comes to the *nature of services*, advances in software engineering broke up monolithic code blocks that served services with a single locus of consistency into smaller, independent pieces of cooperating *microservices*. Hence, the centralized client/server model has evolved into a *chains of (collaborative) transactions,* with typical challenges like *atomicity*, combined *resource management*, and *execution correctness* of the transactions. This, in turn, has created the desire to extend the basic DNS+IP service routing in place today by network support for such chaining, as witnessed by the ongoing Service Function Chaining (SFC) work in the IETF [C3-09]. This application-level trend goes hand-in-hand with the realization that a network cannot just limit itself to blindly forwarding packets; it needs to take an active role in, e.g., providing security (firewalls), assist in service routing (load balancing, redirecting), or traffic shaping. All this is, essentially, software that needs to operate on a stream of packets, just like many application services do. In consequence, this increasingly establishes application- and network-level services at an equal footing with utilizing the increasing *in-network processing & computation* capabilities. However, at present, a proper control framework for such in-network processing is still missing – while IETF ANIMA [C3-10] establishes a virtually separate control plane, it hides compute resources behind application functions. Some work has started, e.g., the recently established IRTF COIN (Computing In-Network) research group [C3-11] or IETF FORCES [C3-12] (separation of forwarding and control elements). Overall, **the nature of services has moved from monolithic services towards chains of collaborating microservices, at both application- and network-service level.**

Along with changes in the nature of services, the third aspect are changes in the *nature of service provisioning.* While microservices (networking or application-level) can be provisioned directly on bare metal, *virtualization* has opened up new opportunities. Since a long time, it has been driving the hosting model in clouds and PoPs; the evolution towards more lightweight virtualization approaches, e.g., through containers or unikernels, has increased the dynamicity of serving instances on a pool of available compute resources. Large-scale services, such as Gmail, YouTube and others, use this approach by dispatching service requests at the DC ingress to dynamically created micro-services, which in turn are based on container-based virtualization. The 5G community has realized the power of such flexibility and enabled its 5G Core specifications to use service-based architecture (SBA), which adopts the micro-service model for realizing vertical industry specific control planes over a cloud-native infrastructure, within a so-called *telco cloud*. *Service routing* becomes key here for the dispatching of service request, e.g., to establish a data traffic session quickly to the right service instance in the data centre of the mobile operator. Given proper service routing, the data centre can easily be distributed, giving mobile operators a decisive competitive advantage over conventional cloud operators in localizing services, as already observed above as a trend in the Internet. We observe that **the nature of service provisioning has changed towards virtualization, for both application services and network services.**

Many major Internet players, such as Google, have long recognized this trend and focused their attention on improving service access in the customer access network (to their POPs hosting their services). QUIC [C3-13], as an example, initially was implemented in the Chrome browser on top of UDP as a differentiator for Google services; standardization in the IETF only followed the initial

deployment in millions of Chrome browsers. The intention here was clear, namely, to improve the invocation of services that support the (initially proprietary) extension, with the access network becoming even an opaque pipe and utilizing service end points instead for everything from name resolution to service invocation.

Complementing virtualization of service elements, *network programmability* has enabled programmatic changes of forwarding operations post-deployment. In consequence, programmability enables the functionality of all/some network elements, network functions and network services to be dynamically changed in all segments of the network infrastructure (i.e., wireless and wired access, core, edge and network cloud segments. Therefore, network programmability supports different and multiple execution environments at the forwarding plane level, those execution environments enabling the creation, composition, deployment, the actual execution and management of network services and/or network functions.

The *digitization of processes* has been proliferating in many industry branches, significantly diversifying the use cases for communication technologies beyond the often consumer-oriented focus of typical Internet services (such as social media or OTT video). Communication technologies have penetrated manufacturing, supply chains, vehicular engineering, health technologies and governmental services, among others. The Internet-of-Things (IoT) has created a vibrant industry sector with a plethora of service scenarios well beyond the consumer-oriented Internet. This has broadened the scope of services and both functional and extra-functional service requirements. The questions are a) if the existing networking model, with its one-size-fits-all approach, can support this mix of services, and b) whether custom-tailored, in-network service provisioned as in-network service chains are a superior model. These questions go well beyond the addition of a small set of QoS parameters to different data flows or the usage of network slices as isolated parts – it considers the whole set of resources and service semantics. As a trend, **new service types are realized by integrating application and network services and their provisioning, across all types of networks.**

Another key aspect is the assumed *service invocation model*. While we already discussed the transition from pure client-server to collaborative model, the 'language' chosen for the transactions performed in said collaborative chains also varies. Although arguments have been presented that HTTP/REST may be seen as the new waist of the Internet [C3-14], the reality of many service invocation frameworks and protocols persists. Those range from request-response models (such as in HTTP), over pub-sub models (with HTTP/2 enabling some functionality) and message passing abstractions to remote memory access models (to create the abstraction of a large yet distributed computer with shared local memory). Similarly, there is an abundance of service discovery protocols (Bonjour, UPnP, …), none of which are interoperable, and few of which are applicable outside very specific environments. We can observe from this situation that *distributed computing has not converged* onto a single universal invocation framework that can be used to connect to any other compute resource. Furthermore, each service invocation framework usually comes with its particular lower layer protocols onto which to map the service invocation itself (e.g., HTTP->TCP->IP), often leaving IP as the only common denominator. Therefore, **services choose the best means of interacting with each other, while relying on basic means to route service requests.**

A final aspect is the changing *nature of the relationships* between the entities providing these services. Currently, systems providing services are mostly assumed to be trusted (or not), and

reliable (with occasional faults), but the overall trend we are witnessing is to an increasingly more complex environment, where multiple providers compete with different (albeit similar) offers, with not exactly the same levels of guarantees and trust. Hence, the overall system can only provide *trustworthy end-to-end services* by relying on high system dynamicity to adapt to variable trust relationships across the different system components. **A service environment of determined trustworthiness needs to be set up by dynamic and intelligent methods over subsystems or micro-services of variable trustworthiness.**

The key takeaway from these trends is that collaborative services in the Internet have moved on significantly since devising the key fundamentals of network forwarding that underpin the transfer of bits over the Internet.

## 3.2 System Architecture Vision: Towards Smart Green Systems

With the general move towards collaborative services in the general ICT domain, the main problem is to overcome the traditional yet obsolete separation of the entire compute-and-communicate infrastructure into separate domains (logic: network vs. application; business: telcos vs. clouds; silos: automotive vertical vs. manufacturing vertical; …), while providing better quality of service (more performance, less latency, adjustable, verifiable trustworthiness levels, etc). Chiefly, if the original Internet was about inter-networking, i.e., best effort bit transport between different networks, **future research must address inter-computing**, i.e., service execution between different systems, potentially deployed and operated between and by different stakeholders yet accounting for the respective service expectations within the whole chain, including potentially higher value goals like, e.g., trustworthiness or sustainability.

In particular, the same applies to mobile communication systems, which have become a crucial part of the overall Internet ecosystem with the tremendous success of the mobile Internet (cf. smartphone revolution). Indeed, to shorten the paths (and latencies), to reduce general infrastructure involvement (better greenness, risk reduction) and to keep the local operations/data local (better governance), these systems exhibit a unique positioning in terms of standardized omnipresence, best possible locality and realizations already involving both compute and networking resources. However, to achieve this target, their ongoing transformation from single authority domain, mere access networks to dynamically aggregated arbitrary service execution platforms must continue.

With more and more intelligence and computing power available per resource, in the future, the resources of these systems, configurable and orchestratable dynamically (i.e., also reprogrammable in runtime), do not have to be limited to particular predefined roles and can be used both to deploy/support new services (both network and end-user services) and to better match the requirements of services running over the infrastructure – again, potentially accounting for requirements not necessarily stemming from the service logic per se, like energy consumption reduction, some form of confinement, etc. With this however, *unlike 5G,* 6G will be not only more flexible in both its services and in its realization but will also exhibit much higher dynamics, in service types/loads but also in its own topology. With that *higher dynamics* and the seamless *co-existence of virtual and physical entities*, the currently physically separate islands of 5G and prior systems will often overlap in resources in 6G. This applies both to different domains of one single network

(Terminal/RAN/Core), just as it applies to several networks (e.g. run by different MNOs) and to entirely different systems (mobile networks and clouds).

Using the offered large variety of novel challenging ICT services, a massive number of devices will be served by these systems generating, exchanging and treating very large quantities of data. The infrastructure that supports society (IoT, cyber-physical systems) will be integrated with the Internet, which will help improve the effectiveness and efficiency of both. Useful insights can be generated based on the automatic analysis of all that data (e.g., using machine learning methods, ML, and artificial intelligence, AI). Beyond the analysis, AI/ML can also be used to optimize deployment, adaptation, reconfiguration and other decisions or to create better-suited system modularizations and novel entities better suitable for the overall required processing. Hence, *it is paramount to approach AI/ML systemically to correctly assess the relevant trade-offs*: AI/ML instrumentations per se require massive data transfers, are computation-intensive and, ultimately, might consume massive amounts of energy. *Relying on siloed solutions and dedicated implementations limits the usefulness of AI/ML, while it increases both its costs (resources) and the cybersecurity risks (attack surface).*

The postulates above imply that the future network technology will have to support the general Internet economy and the particular needs of the cyber-physical infrastructure, like those encountered in the production industry, alike. It will have to work with virtual objects and remote objects, the density, distribution, longevity and interconnection of which in any area can vary a lot. It will have to integrate local and remote objects and different connectivity modes seamlessly. It will have to handle its own constituting nodes and services of transient nature, which can disappear and reappear, possibly at a different location and in zero time, be multiplied and shrunk without notice, etc. At the same time, this future network will be expected to operate as a facility: it will be relied upon by private users, businesses, critical branches and governments. Therefore, it will have to be resilient to both failures and security threats, in a world, where autonomic operations for both services and infrastructures, and in particular AI/ML techniques, will be widely used. Open standards will be required, while governments will want to impose limits and regulations on the operations on all the data required to drive these new systems. In this context, overcoming the digital divide will be a key driver for technology evolution, and personal freedom and rights will need to be assured across all media.

Here, flexible provisioning and elastic execution on a dynamic and changing resource pool emerge as key challenges for the future system architecture. Flexible provisioning refers to the generality of the infrastructure and its capability to on-board and execute essentially any ICT service. The generality of the infrastructure, as opposed to the reliance on service-dedicated components, is important to increase *infrastructure sustainability* in time and *degrees of freedom for multiplexing gains*. Execution elasticity refers to an efficient adaptation before, during and after the execution, i.e., in particular in runtime, and supports the selection of best suitable links and components, to preserve the expected service properties while limiting overprovisioning. In particular, elasticity, as the capability of adjusting resources used in service execution, is key to enable truly green networking, as it allows to redirect requests to resources with better ecological sustainability and to limit the overall resource footprint while preserving the service throughput. Given the resource mix, we have to assume that elasticity and flexibility also apply to infrastructure resources. Hence,

working with individual resources is limiting and not sustainable; rather, allocations and executions should refer to the resource pool as a whole. This in turn requires pervasive, resilient resource control.

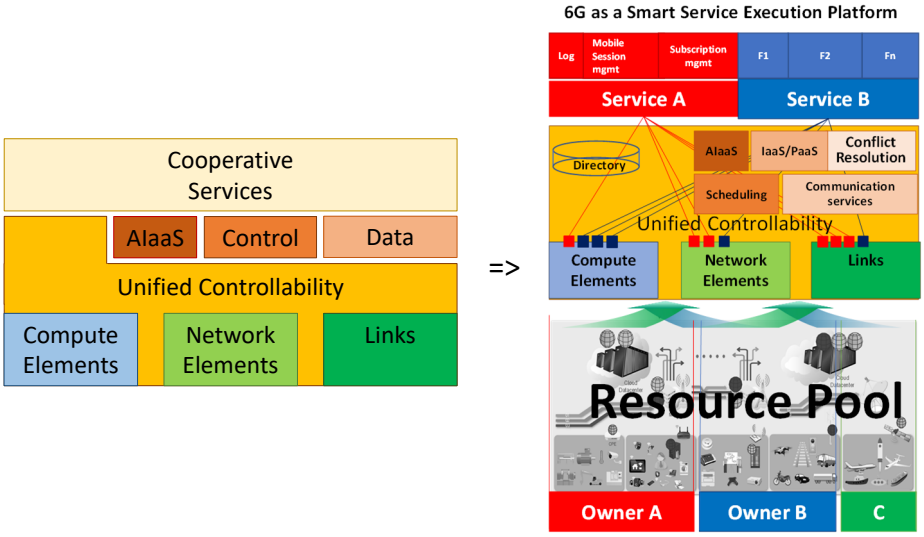



*Figure 3-1: The general Smart Green Networks Concept and its projection to 6G*

Overall, we envision a Smart Green Network as a programmable system based on a common, unified controllability framework spanning all resources to provide each authorized tenant with the required capability to control respectively her resources regardless of their location, type and nature, i.e., including from previously separate and heterogeneous domains, e.g., enterprise and telecom networks, virtual and physical, data centres and routers, satellites and terrestrial nodes, etc. The unified controllability framework allows a tenant to glue such disparate resource islands to one system of that tenant supporting smart flexible instantiation and adaptive, elastic and correct execution of any service on her resources (Figure 3-1). For 6G in particular, the resources will stem from all system players, typically from mobile network operators, but also from cloud providers, non-public network providers and might include terminals, where suitable, while tenants could be mobile system operators, particular service providers, vertical industries, enterprise networks, IoT services or private subscriber networks. Interestingly, 6G will have to architecturally embrace the fact that system resources used for service execution might, per se, be provided, e.g., as services, i.e., that the service and its control in general cannot be limited to the strict boundaries of the authority domain of the service operator only, nor to any particular layer. Rather, in this vision of 6G, all system participants are *potentially* both *resource providers* and *service consumers*. In this situation, the properties of the service must be, in general, enforced regardless of (or even in spite of lacking) assurances at the resource layer.

Hence, the key challenges that the Smart Green Network controllability layer must solve are: the aspects of control over multiple general-purpose, distributed, network control operating systems; the availability of powerful abstractions from resources to services; new naming schemes for virtualised resources; dynamic and automated discovery; structurally adaptive logical interconnection; multi-criteria routing in networks of different densities; intent-based open APIs and highly configurable policies to control the resource and service access as well as dynamics; isolation of application's execution environments and performances; efficient scheduling of

requests to resources; a high degree of automation and support of self-* principles (*self-driving networks*); and distributed yet trustworthy ML instrumentations.

In addition to time-proven algorithm design approaches in order to provide provable and understandable behaviour, the Smart Green Networks concept requires both the existing AI/ML algorithms as well as new, *network-suitable, distributed AI/ML*, to implement data-driven closed control loops that can enable cognitive and comprehensible system behaviour. The training and validation of such technologies require the availability of *cross-technology and cross-sectorial datasets*. The networking research community needs to build those datasets, agreeing how they are generated, accepted and accessed.

Overall, it is imperative to:

- Allow dynamic pooling of resources from diverse participating systems, devices and objects;
- Enable seamless allocation of complex IT objects (ranging from atomic modules to complex services, with the possibility of reusing allocated objects) over some selected as well as over panoplies of such objects, i.e. APIs, interfaces, control hooks, etc.;
- Integrate autonomics to enable both self-organized, resilient programmability and elastic, correct execution of such IT objects;
- Offer programmable analytics and cooperative machine learning to the service layer through open interfaces.

Keys to the realization of this vision are discussed in the next sections of this document:

- programmable infrastructures composed of versatile devices and subsystems (3.7),
- integration of AI/ML at the system level (3.6),
- efficient yet correct runtime resource allocations and their execution (3.5),
- extensible and flexible data plane solutions (3.4),
- and pervasive operational control solutions (3.3).

It should also be noticed that it is important that the directions captured in this section are **accompanied by the appropriate economic and policy work** in future research to make way for the envisioned new services that go beyond current 5G.

## 3.3 Virtualised Network Control for Increased Flexibility

### 3.3.1 Programmability is Control

Future infrastructures must be extremely flexible in operations and elastic in resource usage. Programmability of resources is the only way to achieve this. However, different from configuration management, **programmability requires runtime resource control**, i.e., a way for a program executed somewhere to receive some infrastructure event and to possibly tell to a given resource what to do, both proactively and reactively, including in runtime. The requirements on any control plane are classically intrinsically linked to the requirements on the data plane. Yet, with programmability (such as that being explored currently with P4), *any* data plane becomes possible, and hence, both functional and extra-functional requirements on the control plane are enormous.

For a control plane used for software-defined infrastructure operations, network structure, the available functionality, transported payloads, data rates for the latter, the latencies of exchanges, the resilience and the security are difficult to predict.

**A programmable system must provide an autonomic programmability** after deployment. There are several pragmatic reasons for that: first, setting up such a versatile and resilient control plane manually is not a skill readily available in any environment; second, this approach would be delicate, as one would need to predict future needs correctly. The main reason however is fundamental: *autonomic organization is imperative to support infrastructure dynamics, which programmability as such creates*. Any programmability solution not able to self-organize or adapt is, therefore, incomplete [C3-10][C3-15]. Network and system control cannot rely on rigid approaches, as any such approach would only be suitable for particular environments (e.g., centralistic control, particular hierarchies, etc). Instead, *novel solutions are required capable of organizing control flows and control-related processing dynamically among all controllable system elements*, i.e., across multiple domains, systems and layers. This includes initial self-organization, self-preservation during runtime facing external and internal events and *structural adaptation*. Modern ICT infrastructures need to provide dynamic resource management to fulfil different SLAs and to achieve E2E service assurance. Rigidity in any aspect limits the degrees of freedom and, hence, limits the optimality.

With infrastructure programmability (often referred to as “network virtualization” or “network slicing”, not to be confused with the “5G slicing” concept), the decoupling of the platform delivering the service and the service elements reaches a new level. While IP networking has decoupled services from network infrastructure by putting all services on the same technological foundation (the TCP/IP suite) and by pushing the service logic to the edge, network virtualization brings additional degrees of freedom in flow processing and combines edge and network in one logical entity: it is possible to have different flow processing logics active at the same time within the same physical infrastructure, usually in the form of software elements (different configurations, different active modules) deployed on top of more generically capable hardware resources (typical technologies: OpenFlow SDN, IETF ForCES, ONF P4). Whereas legacy networks rely on specific flow processing machines (e.g., IP routers or Ethernet switches), whose flow processing capabilities are intrinsically linked to the purpose of the device, network virtualization breaks this barrier by allowing to define different flow treatments on the same network node and by concurrently reusing any given link for flows of different “slices” or services requiring different assurances. The same applies to the compute nodes (typical technologies include virtual machines, containers, different host virtualization techniques and industrial frameworks such as ETSI NFV).

### 3.3.2 Separation of control/controllability

The discussion above immediately raises a completely new question of a *service-independent control of resources per se*: as all infrastructure capacities are, in principle, service-independent, we need novel means to make sure that the execution of any service-specific element on an infrastructure element is sustainably possible. In other words, while a legacy hardware router routes and a network switch switches, and there is hardly anything to verify about that, programmability allows to tell a node (how) to route, while this same node was not a router before, yet has had other roles and tasks. It must be verified that it routes correctly over time despite possible task overlap in logic (allocated tasks could result in contradictory operations) or resources (allocated tasks could

get an insufficient resource share). Classically, control was always integrated in a particular solution logic (on the respective OSI layer or abstraction level) and directly projected to resources dedicated to realize (a part of) that solution. Previously, as the existence and the function of a node used to be the same, so was their control. With programmability however, this changes drastically. **We need to understand resource control as a new, paramount domain:** since node and links generally do not have single predefined functions, **there is a new requirement to allocate, monitor, migrate and execute/run several service elements on a shared, per se service-agnostic, infrastructure**. To separate that new notion from the task- or service-specific control, we call it *controllability.*

As a side note, controllability is also different from the notions of management or administration as well, as management / administration rather refers to a) who is in charge (humans, management platform), b) what is being done (management model) and c) different timings. In contrast, programmability could be between devices; it is typically employed for tasks radically different from the classical management (in particular, not OAMP, not FCAPS, but rather related to some particular function realization) and executed and adapted in runtime, exhibiting time-criticality to the running service. As an example, consider OpenFlow SDN, where SDN controller and SDN switches in principle implement the control plane for traffic switching, whereas the management of the whole domain is done independently of that, e.g. over the so-called north-bound interface and classical SDN switch management.

Additional complexity arises from the insight that, generally, an allocated function does not translate to a single infrastructure element, but can be sustained by resource capacities distributed over the infrastructure. Due to scalability and availability requirements and geographical distribution, most network functions rely on distributed realisations, causing the allocation, extension, monitoring or migration of a network function to be much more challenging than the question of copying a software state from one node to another.

### 3.3.3 Multi-Tenancy and Ownership

Network virtualization is resource sharing. Therefore, service footprints, projected to physical resources involved into the execution, are expected to overlap, constituting multi-tenancy in the overall system.

Multi-tenancy in management and control is generally hard, as it contributes to a so-called “split brain” problem: conflicts are likely to happen at the resource level, when several independent owners assign tasks to a shared resource (pool). Such conflicts can be in resource capacities (e.g., two tenants trying to book 2/3 of the resource each), or they can be of semantic nature (e.g., “close port” followed by “listed on that port”). In control, multi-tenancy is harder to resolve, because of the potential time-criticality of the commands. **This calls for autonomic, system-integrated, runtime mechanisms for either conflict resolution or conflict avoidance, both in allocations and execution**. Candidate mechanisms per se should cater for multi-tenant operations and the expected system dependability and size. In particular, they cannot rely on single entities or centralistic approaches. This makes the design of such mechanisms generally harder and optimality as a goal questionable. Besides, while trying to provide service guarantees, such mechanisms should not sacrifice system availability and be aware of energy efficiency.

In spite of its expected pervasiveness, resource control solutions need to respect and maintain boundaries of the responsibilities, power and rights for each stakeholder in the ecosystem, as these are key for a secured, guaranteed SLA enforcement. The problem is that, with network virtualization, tenants can change their control scopes dynamically. Therefore, the classical notion of ownership is not well adapted to the problem space. Instead, the notion of **ownership through controllability** seems better suitable. This notion extends classical ownership through resources obtained through dynamic allocations, booking, and "leasing". For instance, while resource limits of a virtual machine are up to the owner of the executing host, the definition of processes within the virtual machine is up to the owner of that virtual machine. Suitable control solutions should enforce this principle, also in the sense of (secure) isolation.

Known Unknowns

To support different realizations for semantically identical entities and to hide implementation complexity, a general key challenge is to separate enforcement (the "how" part) from the decision (the "what" part). Given multi-tenancy and dynamicity, it is necessary to investigate the ways, in which the control boundary evolves between the objective (e.g., a number of decisions at a given point in time) and its realisation (e.g., considering the operational limits of realising any decision being made, the actually available resources, etc.).

Insisting on perfect knowledge in the described environment will often be in contradiction to the operational reality. Therefore, **solutions should be prepared to work with some degree of "fuzziness"**, i.e., with incomplete data, with data of different freshness, with unreliable postulates. That is why **adaptation is more important than optimality** in this regard. Generally, decision modules need intrinsic flexibility and call for software control elements, realising an adaptive control over the resources they manage. Changes in control objectives are reflected in the existing software, which, in turn, can establish additional software elements in order to react to changes in the control objectives. The enforcement, e.g., of flow handling or computation instalment, is realised by the resource owner, possibly self-constrained by objectives imposed by the physical infrastructure and its operational environment. With all this, the overall system will nevertheless need to fulfil the service requirements.

### 3.3.4 Self-Preservation

Given the importance of the controllability framework for the overall operations and its central position in the architecture, it is crucial to devise dependable, i.e., reliable and secure, solutions. In particular, the roles with respect to the programmability (controllability) and service operations (control) should be verifiable, and necessary protections must be applied to both control channels and control end-points, acknowledging decentralization, multi-tenancy and known unknowns, i.e., also dynamics in the overall span of the control plane and dynamics in the available infrastructure resources.

A running control framework must be able to adapt to such changes, e.g., include and remove resources, adjust its own resource usage yet still protect its own integrity. Besides, the execution of its constituting parts in possible remote, virtual objects on devices physically owned by other tenants calls for either trustworthiness verifications of such executing devices or for systemic approaches to mitigate dependency on any particular component.

The self-preservation solution must also counter so-called *self-inflicted errors* inherent to programmability: a running "program" of a tenant could have negative impact on the resource control framework per se. For instance, it could overload crucial control elements (e.g., putting controller under high load leading to timeouts), influence control transport channels (redirecting traffic) or the control plane structure (e.g., blocking control plane traffic to and from nodes and disconnecting controlees from controllers, etc). **Establishing system integrity and self-preservation in runtime for a distributed, dynamic resource control sub-system is one of the research challenges**.

### 3.3.5 Research Challenges

Challenges on resource control in Programmable Infrastructures include:

- Resource control emerges as an initial glue that first allows operators to program their infrastructures, i.e., as an initial new service that allows to allocate, monitor, execute and remove service elements on/from sets of nodes and links. To avoid vendor lock-in and to allow truly end-to-end slicing, it is exactly this glue that requires standardisation, and not any domain-specific management interface.
- Resource control must be able to reach out to all resources controllable by a tenant and be capable to check the states and operations of all service- or slice-specific elements on those resources. Besides, the realisation of the resource control itself should follow the insights from above, i.e., it must be distributed over all controllable nodes and must support elasticity of itself (reaching out to new elements, adaptability, including in structure, self-preservation, conflict awareness).
- Resource control needs to be able to handle questionable data quality, how to proceed when data is not good or biased, developing either fall-backs, or safe modes. Solutions need to be self-stabilizing, and able to address all these potential uncertainties.

Because of the novel degree of decoupling of service elements from the infrastructure, the central problem of programmability is not to make a blueprint, but to be able to execute any requested blueprint on top of a shared, distributed infrastructure composed of different capacities, occupied by loads from other executed services or slices. Such a distributed guaranteed execution under contention and with concurrency is extremely challenging and, currently, can only be solved on very small scales.

| **Research Theme** | **Virtualised Network Control for Increased Flexibility** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Pervasive Resilient Autonomic Resource Control | Mid-term | Highly scalable, distributed, self-organizing routing protocol to provide in-band connectivity between all resources (zero-conf, zero-touch). Should work across a variety of different topologies (sparse, dense, changing), should support mobility and multi-homing of nodes and avoid to create traffic concentrations. Zero-conf and zero-touch are essential so that no configuration errors can break the connectivity, and that failures be | Flexibility and universality<br>Sustainability (in time)<br>Trustworthiness |

| | | | |
|---|---|---|---|
| | | auto-corrected. Solutions should notably feature progress as per time axis.<br>Final solutions delivered validated to achieve:<br>- high scalability (number of nodes n>100.000)<br>- high success ratio (delivery of packets to random non-isolated destinations >99.9%)<br>rapid (re)convergence to the expected success ratio under dynamics (<10s for up to 20% random link or node failures) | |
| Separation of Controllability and Control | Mid-term | Emergence of resource control as a separate domain from service-related control:<br>- Confirm architectural and operational principles for suitable solutions for multi-tenant concurrent control in relevant environments.<br>- Define APIs with authorization from the resource control plane to the service layer, so as to avoid repetitive control and management traffic (e.g. for monitoring, command piggy backing, etc). | Trustworthiness<br>Dependability<br>Better governance<br>Multi-tenancy support |

### 3.3.6 Recommendations for Actions

| **Research Theme** | **Virtualised Network Control for Increased Flexibility** | |
|---|---|---|
| Action | Pervasive Resilient Autonomic Resource Control | Separation of Controllability and Control |
| *International Research* | Need agreement and standardization | |
| *Cross-domain research* | X | X |

## 3.4 Re-Thinking the Data & Forwarding Planes Towards Compute Inter-Connection (CIC)

As outlined in Section 3.1, Internet communication has evolved from its original design towards the CDN-assisted provisioning of services in many of today's deployments. However, the role of the network is still largely defined by the original E2E argument, postulated by Salzer et al [C3-16], as communication across network elements between communication **endpoints**. This argument has driven much of the development of key Internet technologies in recent decades.

But as also outlined in Figure 3-1, compute elements are seen as not just crucial elements in future system architecture but as also driving the methods used at the data and forwarding plane. Instead of purely being data-oriented but opaque to the purposes of the communication, following thereby the core of the E2E principle, future data planes must become intrinsically aware of computation and the necessary service and compute-specific information that can improve on the communication resource usage, i.e., the performance of the network.

This intertwining of compute and communication resources is captured in the evolution from today's data (and forwarding) planes to Compute Inter-Connection (CIC) fabrics, where CIC embodies this strong linkage of both key aspects, namely the computation at the communication ends, the communication itself but also the serialization of individual communication relations to more complex computational tasks, i.e., the chaining of service functions.

Any future data plane design MUST enable the interconnection of compute resources across one or many computational tasks that are often jointly evaluated in their overall performance.

While the strength of the Internet is its agreement on key building blocks, most notably its Internet Protocol (IP) with key aspects like addressing and best effort packet delivery, the concept of Limited Domains [C3-17] has long recognized the power of the more service and compute-rich network edge in driving network technology innovations that accommodate the specific stakeholder requirements that those deploying innovative edge solutions bring to the table. As argued in [C3-18] this innovation vehicle through limited domains has driven the larger overall innovation in Internet technologies, while relying on the Internet's reach to ensure the global impact that its development has brought to the wider society.

We assert that many of the developments pertinent for evolving today's data planes into full CIC fabrics will be driven by the same limited domain notion.

Through this, we will often see development and deployment of stakeholder-specific (private) solutions, remaining interconnected through the Internet as it evolves. This assertion reconciles the desire for vertical-specific innovation with the need for sustainable evolution of the overall system.

In the following, we discuss a number of aspects for designing future solutions that embody such evolved CIC as a progression to today's data and forwarding planes of the Internet.

### 3.4.1 Design Considerations for an Evolved CIC Fabric

The original design of the IP-centric data plane of the Internet focused on three key fundamentals (i.e., principles – key design choices), namely ensuring *global reachability* through a *robust* packet forwarding mechanism that would provide a *best effort* service to higher layers [C3-19]. Those higher layers would complement the basic mechanisms through aspects of, e.g., reliability, error control, but also support for specific service invocation models.

From the discussion in Section 3.1, we derive a number of design considerations for data plane solutions that would ensure a continued support for the evolved services and interactions we have been seeing in the Internet, depicted in **Figure 3-2**.


*Figure 3-2: Dataplane Evolution - Design Considerations*

We exclude from our considerations the approach to *deployment of the solutions*, therefore not specifically addressing the possible *evolutionary* vs *clean slate* nature of re-thinking the data & forwarding plane in order not to constrain the research albeit pointing out that the *feasibility of solutions* will ultimately need to consider the evolutionary nature of any deployment in existing infrastructures. For instance, P4 is currently being explored as a tool for implementation of data plane programming, but this is effectively simply a technology tool.

It is important to note that evolved data & forwarding plane solutions do not need to necessarily address all considerations and we can already see examples for proposed solutions [C3-01][C3-02][C3-03][C3-04][C3-05][C3-06] considering certain aspects described here:

1. **Dynamicity:** As observed in Section 3.1, many relationships are bound to become ever shorter lived, driven by *virtualization* approaches, with the possibility of network resources to appear and disappear frequently. This introduces aspects of *dynamicity* into the relations that significantly depart from the long-lived locator concept that underpins IP, which assumed a long-lived relation between a client and a *portal* of information in the Internet. Instead, the assignment of forwarding relationships must align with the ability of the corresponding SW component to change relationships, or else the data plane will only inadequately support the advances we see in complex SW systems utilizing the Internet, e.g., through container-based micro-services. This raises a number of challenges to the existing off-path (indirection) architecture of the current Internet, utilizing a combination of DNS resolution with following data plane communication between IP locators. If we were to move to a world where communication endpoints (in ever more serendipitous compute elements of a future 6G world) can be rapidly established but also torn down, an explicit resolution step, incurring around 15 to 100ms latency (according to [C3-20]) in modern point-of-presence based deployments, seems to be inadequate for highly dynamic relationships in use cases such as those found in mobile edge computing, distributed AI and others. Instead, approaches to on-path resolution or runtime traffic scheduling, as suggested in [C3-21][C3-22], accommodate the required dynamicity at line-speeds and high frequencies, albeit questioning the placing of traditional application layer functions nearer to the data packet processing.
2. **Green efficiency:** While we recognize that many of our considerations can be and partially have been realized through a myriad of *add-ons*, *extensions* to and *overlays* on top of Internet protocols, we strongly believe that *green efficiency* is a consideration that must be added to the design for an evolved data plane, even to the point where the selection of

suitable mechanisms ought to include an *energy efficiency KPI* at the same level of today's focus on performance KPIs such as throughput or delay. Overprovisioning and the aforementioned overlaying of solutions to improve on otherwise limited designs have played too long a role in communication networks for it to continue in the light of the increasing policy trends to fight against climate change, such as Europe's Green Deal [C3-23]. While providing a flexibility in change (through yet another overlay), it has also led to complexity in management and the *inefficiencies* caused through indirections over many shim layers that make up the final communication relation. This not only stands in the way of achieving true high throughput and low latency communication, required by many emerging services, but also drives the ratio of ICT in the energy consumption [C3-24]. Furthermore, communication without suitable insights into the usage of compute resources inevitably leads to situations where those compute resources are inefficiently used, e.g., when today's often practised shortest path routing leads to an otherwise overloaded compute element. Instead, approaches that are compute- and therefore also possibly energy-aware, such as proposed in [C3-21][C3-22], are key to building a CIC fabric that moves from the pure communication centrism to smarter routing and forwarding actions that take the overall computational task into account with the aim to reduce overall energy consumption of the joint compute and communication system.

3. **ualitative Communication:** Relationships will not only become more dynamic in nature but also more complex in terms of *inter-dependencies*. The current model in the Internet treats relationships at the application or session layer, realized through independent connections, managed through protocols like TCP and others, with separate resource management schemes. This leads to inefficiencies in cases where one sub-relationship is transferred well compared to the other, spending efforts on, e.g., error control, for a sub-relationship that is reduced in value due to reduced performance of another sub-relationship. The result is often overall loss of end user experience that ultimately decreases the value of the communication. This qualitative communication is crucial to be taken into account when designing data plane solutions in order to be able to optimize the use of resources spent on the overall relationship rather than the sub-parts of it. Leaving this handling purely to the application or session layer leads to inefficiencies of resource usage, which can be avoided through application awareness, e.g., additional in-packet metadata at a lower part of the data plane, expanding on existing concepts such as service function chaining (SFC) [C3-09] albeit for parallel not sequential transactions.
4. **Security** plays an important part in data plane mechanisms and the current Internet has well recognized this with security considerations having become essential in every protocol solution standardized, for instance, in the IETF. However, the fundamental of building *security on top of an otherwise unsecured packet forwarding* has not changed, therefore focussing efforts on end-to-end security of the application-level content, but not the *security nor the privacy of the packet forwarding operation* itself (who is talking to whom, compared to what is talked about). Consequently, this has enabled for long mechanisms such as IP geo-tracing as well as enabling spoofing and therefore denial of service attacks. Mitigating methods deployed are add-ons to the otherwise unsecured IP, require extra effort rather than basing themselves on an *intrinsically secure* design per se where security of end points and networks alike is ensured together with the *privacy of the interaction* between communicating end points, striking the right balance between accountability and anonymity. Decoupling "security appliances" from the analysis of events and policy-based decision-making is another aspect to consider. Tiny security-handling functionalities embedded into

virtual entities should monitor events, collect information and transfer it to suitable functions (possibly based on AI/ML) capable of more powerful analysis and anomaly detection, which in turn would enforce policy based-decisions back to the local actuators.

5. **Precision delivery:** the best effort nature of the current IP suite does not suffice for a number of the new emerging services, e.g., for Beyond 5G. Therefore, it will need to be extended in order to capture new demands for specific performance characteristics, such as *strict delay and latency bounds* for system control, human interaction and many other services as well as *on-time bounds*. This requires the control loops involved to ensure the specified performance requirements of various applications, particularly for access networks with widely varying performance characteristics such as wireless. Those control loops will also need to enable trading off latency at the control level against the necessary operations at the data plane.
6. **Diverse Addressing:** While the universality of higher layer service concepts over a single addressing scheme has been praised as key for the Internet protocol, we assert that the support for *diverse addressing* will need to replace this aspect of the current Internet in order to improve on efficiency when supporting the many new services, while still *ensuring the global reachability* that the current Internet has achieved. This could lead to solutions for *optimized Internet-of-Things* communication (with *smaller identifiers* being used for efficiency purposes), while preserving inter-domain access to the IoT resources. As another example, instead of relying on an interaction between DNS and IP routing, adding initial latency to the service exchange (and leading to problems in future service invocation if service relations might dynamically change), research in, e.g., routing on labels [C3-25], information-centric networking [C2-26] and solutions on *semantic addressing* [C3-27] have shown that those latencies can be significantly reduced through name-based addressing, pushing name information to the far edge of the network as a trade-off (which can be accommodated through increasing availability of storage, even in mobile devices), while still scaling to significant network sizes, particularly in the recognition that much Internet traffic is being localized, as discussed in Section 3.1. In addition, changes in named relations become merely an *ingress routing decision*, being removed as a burden from the DNS, for instance, therefore significantly *increasing flexibility* in routing when the service instance serving a named relation is changing in the light of virtualization of service endpoints, as discussed in Section 3.1.

The aforementioned considerations for *designing suitable packet delivery solutions* need to furthermore consider the following aspects when *being realized for and deployed* in the emerging communication infrastructure:

7. **Manageability:** All the above characteristics will require suitable instrumentation to monitor and validate the delivery of promised assurance levels. Furthermore, telemetry capabilities, i.e., the process of measuring, correlating and distributing network information, are required (and will need to be enabled at the data and forwarding plane level) to gain the visibility of network behaviour to improve operational performance over conventional network Operations, Administration, and Management (OAM) techniques to enable full network automation.
8. **Programmability:** As per Section 3.3, respective owners (e.g. service providers) will need to be provided with the methods to dynamically govern all resources incl. the forwarding plane in order to rapidly and easily introduce new network services or to adapt to new enhanced and modified contexts. A higher programmability of the forwarding plane could be achieved, e.g., through insertion of programmable metadata in packet headers traversing the network.

Such programmability particularly aims at providing the desired overall green efficiency by moving from HW to SW upgrades, including executable code injected into the execution environments of network elements in order to create the new functionality at runtime (in network compute) with the required characteristics (e.g., security). Furthermore, what is handled in-network or in the data plane needs to be assessed.

9. **Slicing:** Resource management is discussed in details in Section 3.5. In the forwarding plane, it needs to promote easy and efficient execution of multiple and different types of delivery mechanisms, possibly each with different guarantees for KPIs/QoS/ stringent non-functional requirements of network services at a given time on the same infrastructure but across separate subsets of resources in the shared resource pool for realization of the desired functionality. Such "slices" may offer uniform capability interfaces to entities and network functions, abstracting the autonomous slice components, which may be loosely coupled, with different functional and non-functional behaviour. A challenge to address is the realization of large-scale and multi-domain data plane deployments in sliced environments, including aspects of identifying the participating resources being used.

### 2.2.1 Key Research Questions

The following research questions are not purely limited to the data and forwarding planes but address wider holistic systems aspects, leading to the following research challenges:

1. **Which layering in which part of the network?** To cater to the often starkly different 'scopes' of communication, ranging from localized sensor communication over POP-based access to OTT services to truly global communication, the question on layering is crucial in the light of an *efficient/green* implementation of the overall system. With the desire to support *diverse addressing* of the data plane, the question needs consideration as to *what layer best realizes the semantically different forwarding operation(s) most efficiently, taking into account not only the individual service itself but also the overall system efficiency from the perspective of resources that provide that service.* Note that this does not preclude combinations of different layers.
2. **What is the role of soft architecting?** With the proliferation of software-centric approaches to networking, allowing for a much higher degree of post-production as well as post-deployment programmability (cmp. Section 3.2), the question arises *what the deployed architecture really is* or *if everything manifests its own (soft) architecture?* Assuming such soft-architecting, as discussed in Section 3.3, the desire to agree on a common substrate, on top of which all such (soft) architectures reside, still remains, similar to the origin of the Internet protocol albeit with a possibly different answer. Instead of the commonality being that of a common postal system between locations, such *commonality* could be the interconnecting bus-like system between resources, where resource control becomes fundamental, while global transport and global routing degrade to applications, many of which can run in parallel. Any answer to that common substrate, however, should still provide the right set of fundamentals among those outlined in Section 3.4.1 that align with the services at hand. *In other words, while soft-architecting is a promising evolution path, ultimately, the considerations above need to be applied to and solved by the global "glue" at the resource layer, be it a control bus or the delivery system itself.* A possible advantage of a solution based on a resource control is a clear set of and a better understanding of the requirements of the latter.
3. **What are the tussle boundaries of the overall system?** Tussles [C3-28] are caused by interactions of players as defined through the interfaces of the overall systems, with each

player often pursuing their individual interest. Understanding the boundaries of tussles, the mechanisms to express them and those to resolve them, is crucial for the overall working of the system. Much has been done to study the tussles of the Internet (and its main players) but postulating a system of high *dynamicity* also postulates one of changing relations, particularly when it comes to *trusted* relations. Enforcement through trusted third party is often a mechanism that will not do in such often ad-hoc relationships and *solutions will need to realize more suitable, equally dynamic and ad-hoc mechanisms to ensure an otherwise trustworthy execution of the overall system, while also preserving the privacy and ensuring the security of the individual participants.*

4. **What (meta)data is required to make the data plane work (well)?** Any data plane solution, including existing ones, works on a set of metadata, such as identifiers, as well as state, such as link data. While much of this data is vital for the basic operations realized in the data plane itself, it is also required for *control plane* decisions (e.g., for load-depending resource allocations across the network) and for realizing *management goals* (e.g., matching long term demand to supply information). With this in mind, data plane solutions must not focus solely on hitting the key fundamentals outlined in Section 3.4.1 but also enabling a fruitful interaction with the corresponding parts of the overall system that ensure the working beyond the pure transport of relationship information.

| **Research Theme** | | **Re-Thinking the Data & Forwarding Planes Towards Compute Inter-Connection (CIC)** | | |
|---|---|---|---|---|
| # | Research Challenges | Time line | Key outcomes | Contributions/Value |
| 1 | **CIC architectural frameworks** to enable dataplane evolution that is economically and technically sustainable | Mid-term | - Architectural blueprints/ frameworks with key roles and interfaces<br>- Economic models to show viability of new communication services<br><br>Proposed solutions should:<br>-**Confirm architectural concepts** and principles for suitable and novel data plane solutions in relevant environments.<br>- **Define relevant APIs** to enable key exchanges in multi-stakeholder deployment models | - **Sustainability** through innovation models for evolving data plane<br>- Improved **innovation** capability<br>- new services<br>- Facilitation and identification of **new market entrants**<br>- faster deployment |
| 2 | **Runtime CIC (resource) scheduling** mechanisms that allow for broad communication and service semantic support, while continuing to adhere to net neutrality | Mid-term | Protocols, algorithms, architectures and solutions for dynamic, runtime assignment of resources to tasks, leading to efficient steering of resulting communication traffic between resource endpoints. Specifically needed are<br>- Overall resource-, i.e., compute- and network-aware, **service and network path selection algorithms** for steering resource requests at runtime, including **energy-aware** mechanisms | - **Sustainability** through energy efficiency improvements through reducing overprovisioning of resources or, equivalently, increasing system throughput<br>- Better **governance** through service- |

| | | | | |
|---|---|---|---|---|
| | | | - **Multi-optimality routing** solutions with higher convergence rate than existing protocols<br>- Protocols to allow for **service-specific** traffic steering mechanisms<br>- Solutions that investigate the use of novel **data plane capabilities** for realizing runtime scheduling (cmp. Section SA.5)<br>- **Privacy mechanisms** to allow for app-aware traffic steering in presence of net neutrality<br>Final solutions delivered validated to achieve:<br>- **high scalability** (number of services n>100.000 and number of ingress nodes aligned with expected 6G deployments)<br>- **reduction of service competition time** for best effort services for better user experience and acceptance by at least 50%.<br>- **increase of resource efficiency**, measured in either an increased service goodput or in a decreased resource consumption for the same goodput, by at least factor 3<br>- **rapid (re)convergence** to dynamically updated service metric information under dynamics (<1s for large number of updating services) | specificity of traffic steering solutions<br>- **Trustworthiness** in solutions through privacy mechanisms built into CIC |
| 3 | **Addressing for an evolving dataplane**, building on and suitably extending existing IPv6 | Mid-term | - **Emergence of semantic addressing** that allows for wide and specifically novel communication semantics<br>Final solutions should:<br>- **Confirm addressing concepts** and approaches in relevant environments.<br>- **Dataplane implementations** based on available or emerging packet processing HW<br>- Define **new packet formats** with engagement in relevant SDOs for adoption | - **Sustainability** by reducing overhead of unused address space in transmission<br>- **Innovation** in new services at the data plane level through flexibility in addressing<br>- **Resilience** through avoiding cross-solution interaction through joint and extensible addressing approach |
| 4 | CIC Operation and management (OAM) frameworks | Mid-term | - New OAM approaches that **reconcile network and service operator policies** while adhering to the commercial boundaries<br>Final solutions should: | - **Trustworthiness** in the overall system through suitable OAM solutions<br>- **Faster deployment** of novel solutions through |

| | | | | |
|---|---|---|---|---|
| | | | - **Confirm architectural concepts** and principles for suitable and novel data plane solutions in relevant environments;<br>- **Define relevant APIs** to enable key exchanges in multi-stakeholder deployment models;<br>- **Define protocol, mechanism and operational principles** for suitable solutions to maintain stability of the executed systems under the dynamicity models. | appropriate OAM solutions |
| 5 | Cross-flow and cross-endpoint data plane operations | Mid-term | - New **mechanisms and protocols** for cross-flow resource and timing (latency) control<br>Final solutions delivered validated to achieve:<br>- **high scalability** (number of services n>100.000 and number of ingress nodes aligned with expected 6G deployments)<br>experimentally verifiable **improvement on KPIs**, such as reduction of service competition | - **Sustainability** through improved energy efficiency by dynamically managing resources across flows and endpoints<br>- **Societal values** by enabling new interactive and human-centric services at high efficiency and with great precision |

### 3.4.2 Recommendations for Actions

The following list are suggestions for important actions towards realizing the research agenda for DP/FP evolution, not claiming to be exhaustive:

1. *Call for internationalized efforts*: given the challenge to evolve the data /forwarding planes, European efforts should liaise or even directly collaborate in internationalized research efforts, i.e., in the creation of solutions not just the exploitation in standards or OS communities. This could be realized through targeted **international calls** (e.g., EU-China, EU-US, …) on data/forwarding planes technologies as well as through the creation of **international expert groups**, e.g., in coordination and support actions.
2. *Call for experimentation*: although strong theoretical foundation is desired for any change of fundamental data/forwarding planes functionality, strong **experimental evidence** and **large-scale open testbeds** are crucial to show feasibility but also foster adoption through the operational community. This could be realized through an evolution of the original FIRE efforts or a similar trial phase as in 5GPPP. Open experimentation data/forwarding facilities are required for a large number of **third-party experimenters** of promising solutions and possibilities for looking, e.g., beyond 5G - an Internet of experiments (IoE).
3. *Call for data/forwarding planes research repository*: in order to foster the adoption of evolved data plane technologies, experimentation (see item 2) will need to ensure **replicability** in other, possibly pre-commercial or otherwise research, settings. This could be ensured through making evidence **data and code base availability** mandatory for certain aspects of data plane research (e.g., for certain TRLs upwards), including **migration solutions** that will allow legacy IP-based applications and IP-Services to be used with the new enabled forwarding plane capabilities.

4. *Call for clean slate research*: following the argumentation in other efforts, such as FP7 FI, NSF FIND, the evolution of core Internet technologies requires a combination of an **evolutionary and revolutionary** approach. This could be achieved through setting aside specific **clean slate** or greenfield funds for testing more revolutionary approaches to the data plane evolution.
5. *Call for funding data / forwarding planes research in solutions* along the considerations discussed in Section 3.4.1, such as those providing precision delivery in extension to existing best effort. Examples for such research aspects are
    a. Precision packet delivery (with QoS) to extend/complement best effort delivery;
    b. Intrinsically secure, i.e., authenticated and accountable, packet delivery;
    c. Semantic routing, extending current endpoint-based routing for lower latency and higher flexibility delivery of service requests;
    d. Deployment on tenant-specific (in-)network service functions;
    e. Inter-connection of compute/storage resources at Layer2, with focus on customer access networks while interconnecting to Internet-based clouds;
    f. Programmability of forwarding under control triggered by management.

| **Research Theme** | **Re-Thinking the Data & Forwarding Planes Towards Compute Inter-Connection** | | | | |
|---|---|---|---|---|---|
| Action | Challenge 1 | Challenge 2 | Challenge 3 | Challenge 4 | Challenge 5 |
| *International Calls* | Needed for possible concerted standardization efforts | | Needed for possible concerted standardization efforts | Needed for possible concerted standardization efforts | |
| *International Research* | | Encouraged due to international expertise beyond Europe | | | Encouraged due to international expertise beyond Europe |
| *Open Data* | Encouraged to broaden insights into markets beyond Europe | Encouraged for replicability of research | | Encouraged for broadening best practises to many deployment scenarios and markets | Encouraged for replicability of research |
| *Large Trials* | | | | | X |
| *Cross-domain research* | Encouraged to gain the needed economic viability insights | Encouraged to develop possible strong AI-based solutions | | Encouraged to develop possible strong AI-based solutions | Encouraged to develop possible strong AI-based solutions |

## 3.5 Efficiency and Resource Management

Efficiency in terms of managing the resource pool of a communication system is essential for controlling costs and therefore OPEX in offering communication services. With Total Cost of

Ownership (TCO) becoming a major design target and the push for *sustainability* of telecommunication infrastructures, the role of efficient resource management will increase significantly in future deployments. This translates to several new problem spaces, currently unaddressed, underestimated or completely overlooked in both the industry and academia.

### 3.5.1 Network Slicing and Infrastructure Programmability vs. Network Capacity Planning

As network slicing promises a sheer endless customisation of network-spread functionality, it becomes difficult to plan the capacity of network infrastructures in the same way as today. Whereas operators currently use their combined empirical knowledge regarding both infrastructure and the expected service (and its prices), network slicing turns this principle upside-down: while the infrastructure operator remains neutral to the service, the slice owner is expected to translate the *service to capacity requirements* onto the infrastructure capabilities, an exercise that lacks a reliable general methodology. Incapable of correctly translating service to capacity requirements, slice owners are likely to engage in a cloud-like operation model: start small, then expand or reduce contracts as you go. The *elasticity of the slice therefore is a central requirement*. This fact together with the required radical reduction of the service creation time (from 90 days to 90 minutes, as, e.g., per 5GPPP KPIs) underlines the upcoming shift from planning of the infrastructure to continuous (and likely dynamically adapting) runtime operations on the latter. In simple terms, network planning and network slicing are misaligned, as the former, driven by the presumed physical deployment, operates within completely different time frames than the latter, which exhibits on-demand elasticity.

Hence, while the initial planning provides the larger operational bounds, within which slicing can operate, it is the runtime (continuous, real-time, hot) management and control that determines the efficiency and therefore the costs of the provided service. If network slicing in particular or, more generally, challenging, beyond best effort services, are to be successfully provided by the telecom infrastructures, the employed technologies must embrace this change and use mechanisms and practices that feed runtime control over a longer timeframe back into the planning and investment cycle for network infrastructure.

Independently of scale, slicing renders the infrastructure usage and occupation much more diverse and more dynamic. This emphasises the requirement for continuous operation of the real-time management or control, while infrastructure control and management are required to handle the dynamics in a new, currently unsupported manner. This includes handling node and service element loads, departures, additions, errors and the like.

Runtime management and control ultimately still drives the longer-term planning that we can see today in networks. Following our cloud analogy, the longer-term demand and supply pattern emerging from the many tenants of a data centre still drives the planning, and therefore investment patterns, for sufficient build-out of the cloud. Similar feedback must exist for slicing-based network infrastructure albeit situated in a many point-of-presence nature of resources, utilised over a possibly huge area of requirements on those resources.

The change towards network and infrastructure programmability results in an even more radical shift towards the need for runtime optimizations. As programmability advocates a reactive change

of the behaviour of the infrastructure in runtime, the efficiency of such changes must be also addressed in runtime: the information on what to optimize is simply not yet available at deployment time, and hence capacity planning is hard to apply.

### 3.5.2 Slicing and Programmability Require Conflict Resolution

To better support multi-tenancy and to allow efficient resource sharing, especially at bigger scales or facing known unknowns, **consistency and concurrency of allocations and executions of the latter should be addressed** at the systemic scale in runtime [C3-29]. Indeed, concurrent resource-competing or semantically contradictory requests at either allocation time or during (elastic) execution must be dealt with to avoid partial operation (e.g., of a slice), generally being useless and, hence, waste of resources, while requiring **novel mechanisms for networked garbage collection** to free up any erroneous resource allocation during such conflict resolution.

While mechanisms exist for handling concurrency at individual component/node level, guaranteed slice allocations would require novel, system-wide mechanisms. Herein, fundamental systemic limits are to be properly addressed at large scale, since strong consistency of allocations (e.g., through consensus, atomic commit protocols with locking, etc.) might otherwise lead to a decrease of availability (starvation effects) and therefore reduce the supported dynamics in slice allocation and elasticity.

Inspired by distributed database management systems and distributed Internet services, **novel research should consider multi-level guarantees for services and service-level redundancy**. In spite of the similarity, the central insight here is the difference in the definition of consistency for databases and systemic allocations: while databases treat replica of the same object (which makes concurrent writes to replica R1 and R2 problematic), systems work with redundant, independent objects (e.g., concurrent allocations on two equivalent yet different paths are non-problematic). Given the observed increase in systemic redundancy (e.g., network density, trend to regional data centres), this insight promises better scalability of guaranteed allocations without sacrificing availability. Hence, **novel approaches could explore the suitability of concurrency-preserving schedules** (e.g., with commitment ordering) **for programmable networked IT systems** [C3-30]**.**

### 3.5.3 Elasticity: Efficiency Requires Runtime Scheduling

When addressing efficiency, *Total Cost of Ownership* KPI and *green ICT* become important aspects to consider. For instance, given a slice blueprint, one must find suitable resources in the infrastructure and make a reasonable long-term allocation of the blueprint on the selected resources (as per slice lifecycle). This topic has received a considerable attention and is often referred to as *"virtual network embedding",* with both simplified greedy solutions and optimised heuristics (with tuneable sub-optimality bounds) being available. However, the overall resource allocation problem of network slicing is twofold, and the second part is still unsolved, relating to the question of *elasticity of slices*. Indeed, to achieve slice properties not readily provided in the serving infrastructure (e.g., elasticity, but also availability, resilience, latency guarantees, etc.), slice embedding will be usually broader than the purely functional requirements of the blueprint. Therefore, for every entering flow, a simplified, yet more dynamic and online question of the resource allocation problem will arise: **which of the suitable function-equivalent infrastructure resources should be involved into the treatment of an incoming flow**? [C3-21][C3-22] Note that this cannot be solved within the slice, if the infrastructure owner promises (and sells) extra-

functional properties of the allocated slice; in other words, such provisioning will be done in the infrastructure, transparently to the slice owner.

The answer to this question of **runtime service scheduling** [C3-21] is paramount to address the TCO KPI, as solutions to this problem would allow to overprovision slices, without the need to overprovision the underlying infrastructure. The runtime service scheduling therefore is the answer to the questions of elastic and dynamic allocations, currently unsolved. Moreover, if an efficient solution to this problem can be found, network slices can – and, for efficiency reasons, should – be implemented as dynamically scheduled entities rather than exclusively reserved (and, therefore, possibly wasted) resource pools for tenants.

### 3.5.4 Towards Green ICT

In recent years, the *ecological conscience* has generally increased in Europe. Backed by political and economic initiatives both by the Commission (e.g., Renewable Energy Directive, Green Deal) and the Member States (e.g., German *Energiewende*), the main trend is *to reduce the dependency on conventional energy sources* (nuclear, fossil) *to the advantage of renewable energy supply* (wind, photovoltaics, hydroelectricity). Given the decreased flexibility in the energy production of the latter, *this shift must be accompanied by smart energy demand management functions*, resulting in a strong push for Smart Grids in the energy sector. That is where ICT is generally regarded as an important enabler (e.g., using 5G MTC and network slicing). However, swapping power sources does not address the power consumption of the consuming infrastructure as such.

Given the increased reliance of the society on ICT infrastructures, these have emerged as essential consumers. For instance, while 5G is 10 times more energy efficient than 4G in transmission, recent studies suggest that, by 2025, 5G alone can increase the anyhow growing energy demand in the data centres by up to 3.8 terawatt hours (TWh) in addition [C3-31]. Even though this effect is due to the increased “popularity” and not to a shortcoming of 5G per se, undeniably, **energy efficiency of distributed compute facilities emerges as a central preoccupation for resource management**. This is not limited to individual data centers, but should be considered for distributed compute at large, i.e., for edge computing, mixed virtualized/physical networking and data centers together. While overprovisioning is a simple and popular method in networking (e.g., in fibre optics, it is a simple mechanism for both network development and service quality increase), crude overprovisioning is not a valid approach for the computing domain. Rather, modern DCs reduce the required compute power and energy for the same load, e.g., using DC-internal schedulers (e.g., Apache Mesos, Kubernetes K8) and many other means [C3-31][C3-32][C3-33].

Novel methods are required to overcome the limitation to a single DC and should embrace path and compute allocations together, in order to exploit infrastructure diversity. **Future research should explore and develop approaches to elastic resource management** in addition to the current trends somehow limited to green energy power supply (eco-current) for data centres (potentially using smart grid’s demand management) and the “recycling” of waste heat from the DC cooling systems. Such novel approaches could generally rely on elasticity mechanisms, i.e., runtime redirection of incoming service requests to best suitable infrastructure components with the goal of increasing the throughput on the same resource footprint. Preferred redirection to eco-powered components can be integrated into runtime service scheduling.

This theme translates to the overall ICT sector and ICT infrastructures in that **green or sustainable ICT cannot be achieved without a profound consideration for resource management**. Given the steady increase in the dynamics and the diversity of services, pre-planning and fixed allocations of any kind (dedicated devices, pre-provisioning, long-term configurations, mapping to particular nodes, single points of failure) are doomed to overprovisioning, which, for the same service load, requires more resources to be deployed, maintained and powered up in the infrastructure. This wastes energy and is ultimately not sustainable.

Among suitable techniques, some of which already adopted in the presence of non-virtualized equipment, dynamic adaptation (like adaptive rate and low power idle) and smart sleeping need to be further investigated to extend them to the virtualized environment, where chains of VNFs are implemented by Virtual Machines or Containers. Power states that are immediately identifiable in physical hardware (as defined by the ACPI industrial standard [C3-34]) do not map straightforwardly to virtual CPUs, as the interaction with hardware is mediated by software artefacts and hypervisors. Likewise, defining interfaces to convey energy management capabilities of virtual devices, in order to abstract their internals and expose energy-related parameters that can be manipulated by control strategies in the control and management planes, by extending the so-called Green Abstraction Layer (GAL), needs further refinements, although efforts have been already undertaken by ETSI and ITU [C3-35][C3-36] in this direction.

### 3.5.5 Research Challenges

On the opportunity side, programmable ICT infrastructures increase the degrees of freedom in service-to-infrastructure mapping and, therefore, could yield more sustainability both in time (flexibility) and in resources or energy (elasticity). On the challenges side however, the mixed compute/ storage/ networking environments, even under the assumption of pervasive controllability, require suitable solutions with respect to resource management: the heterogeneity of resources makes it harder to rely on single mechanisms, as different domains apply their own approaches internally, and often do not exhibit this knowledge externally. Also, a given unique approach will likely not fit the requirements of different resource types. Besides, the scale of the overall infrastructure makes it hard to rely on any consistent, up-to-date picture of the current consumption vs. load, as described above.

Challenges in this area can be summarised in the following:

- The question of runtime service scheduling in programmable ICT systems is paramount, as it permits both to provide superior extra functional properties of the supported allocations (“slices”) and to lower the Total Cost of Ownership. Indeed, the TCO of a slicing implementation using only fixed-quota assignments (meaning that the sum of the resources consumed by all slice instances will define the necessary infrastructure resource footprint) would be horrible, roughly comparable to hardware slicing. **The dynamic resource assignment problem,** as a quest for a more efficient infrastructure sharing, including computing, networking and energy resources, **is difficult because of heterogeneity, partial or outdated information, its runtime nature and the absence of any central party or mechanism** (like ordering or synchronized clocks).
- The answer to the job scheduling in large networked systems requires a lot of fundamental research, such as leveraging existing solutions from data centre research and applying them at network scale with multitenancy and concurrency. Suitable **conflict handling mechanisms**

**are required** here, especially if guaranteed execution is required. Utilizing insights from distributed systems research, **the major goal should not be optimality, but rather improved efficiency**: given the size of the infrastructure, *1 % efficiency increase might translate to hundreds of millions of Euros/Watts/additional users*/etc. Given the assumption of sub-optimality, novel **mechanisms for networked garbage collection can be considered**.

- The elasticity of slicing has to propagate towards subscriber level and even application level. For instance, an application could use different slices during its session in order to best utilise the network as well as to provide superior quality of service with respect to slice offerings. In a view similar to application-driven networking, an application could also explicitly ask for a "slice" suitable to its needs. This rules out any pre-provisioning and can only be reasonably implemented in public infrastructures like the telecommunication networks, if the provision of the slices is highly dynamic yet resilient. Thus, **application requirements need not only signalling but also suitable translation to constraints, under which the slicing control can operate to meet the applications' needs**. Separation of concerns between Vertical/Network Application Orchestration (VAO/NAO), in terms of organization of micro-services and their interconnection, and Network Functions Virtualization Orchestration (NFVO) will be a facilitator to implement this framework, where slice intent configuration can be conveyed through the mediation of the Operations Support System (OSS).

| **Research Theme** | **Efficiency and Resource Management** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Runtime Service Scheduling (RSS) | Mid-term | Protocols, algorithms, architectures and solutions for dynamic, runtime assignment of resources to tasks, such that the executing system handles each task successfully under that task's specific constraints while explicitly accounting for the resources used by the solution per se and its novel, added constraints. It must achieve overall improvements in:<br>- either resource usage (energy, capacity, etc) for the same throughput<br>- or in terms of successful service throughput on an unchanged system<br>Final solutions should achieve improvement of:<br>- Min. 3 times for non-critical services;<br>- Min. 5 times for services with high guarantees;<br>within systems composed of up to 100k nodes with high task arrival rates from tens of thousands of system nodes. | - Reducing overprovisioning in resources (entities, nodes, links, energy) or, equivalently, increasing system throughput<br>- Improving greenness and sustainability |
| Conflict Avoidance/ Resolution (CAR) | Mid-term | - Protocols, algorithms, architectures and solutions for conflict avoidance or resolution facing concurrent, uncoordinated usage of resources, where the executing system strives to handle tasks successfully under that task's specific constraints, while the throughput of | - Reducing resource waste<br>- Improving greenness and sustainability |

| | | | |
|---|---|---|---|
| | | successfully executed critical tasks (i.e., goodput) increases compared to a situation without the said mechanism in place on a fixed resource footprint. Final solutions should demonstrate sublinear scaling of required resources with the increasing load.<br>More generally, solutions providing:<br>- Sublinear increase of the required system resources under the increasing goodput requirement (including the resources required for the proposed solution as such);<br>- Where this performance improvement is maintained across systems composed of up to 100k nodes. | - Improving multi-tenancy and governance |
| Networked Garbage Collection (NGC) | Mid-term | - Protocols, algorithms, architectures and solutions for freeing unused, stalled, crashed or incorrect (e.g. partial) allocations on resources.<br>Final solutions capable of:<br>- Doing this without perceptible negative impact on running task performance, throughput and correctness;<br>- Freeing, over its runtime, in average more resources than the proposed solution per se uses within the system;<br>- Supporting systems composed of up to 100k nodes. | - Reducing resource waste<br>- Improving greenness and sustainability |
| Articulation of needs and provisions from the system to the user/applications (EXP) | Long-term | - Protocols, algorithms, architectures and solutions for user-to-system interface, i.e. exposing available resources and capabilities to the user applications and getting requirements from user applications explicitly or implicitly.<br>Final solutions:<br>- Supporting legacy applications as much as possible, including solutions for support of applications that use traffic encryption;<br>- Reusing (extending, integrating, mimicking, maintaining compatibility to) existing methods, where applicable, such as network exposure, ETSI MEC, etc.<br>- Allow runtime negotiations with compatible, novel applications. | - Improving sustainability and universality<br>- Improving transparency, trustworthiness, governance |

### 3.5.6 Recommendations for Actions

| Research Theme | Efficiency and Resource Management | | | |
|---|---|---|---|---|
| Action | RSS | CAR | NGC | EXP |
| *International Research* | | | | X |
| *Cross-domain research* | X | X | X | X |

## 3.6 A self-learning, AI-Native, Service Provisioning Infrastructure

Utilizing knowledge gained over a longer time is well-established in the industry. OTT services have long been using AI/ML techniques, albeit operating largely on data sets derived from the services and their users directly. At the level of improving network operations, self-* solutions have advocated the use of operational insights to adapt network functionality without intervention from either human operators or users.

Given the vast amount of data available in complex network environments albeit in a distributed fashion, AI/ML is well suited to produce new insights into emerging behaviour patterns in such distributed environments. To this end, suitable AI/ML techniques are applied, provided as a service capability towards (a) *operations of networks* and (b) *improvements of service provisioning and functionality* itself. In other words, we see a strong evolution of future networks from a mere communication and computing infrastructure to an integral part of the overall knowledge pool that can be used to improve functioning of networks and services alike; *AI-as-a-Service (AIaaS)* provides this capability in a prosumer-centric notion.

### 3.6.1 Proliferation of AIaaS in Network Operations

We foresee AI/ML playing an increasingly important role in network management**,** with the aim of reducing costs, increasing productivity, deriving more value, and improving customer experience. A range of learning techniques can be used to predict the behaviour of the network and its users to better provision resources by avoiding today's typical over-dimensioning. In terms of OPEX optimization, where energy consumption is one of the major cost items for network operators, AI/ML will leverage ***"data lakes"*** to analyze performance and optimize energy consumption versus quality of service. We furthermore see a strong alignment with the move towards fully virtualized network functions, where AI/ML capabilities are utilized to ensure reliable controllability in a fully automated manner, specifically to:

- Instantiate a complete end-to-end network that includes, e.g., the RAN, mobile core, other forms of access networks (DSL, etc.), transport network, as well as the Data Network, edge and beyond-edge devices' resources, expected to become integral components of the network infrastructure in, e.g., 6G systems as per Section 3.2.
- AI usage in runtime optimization of service provisioning and coordination mechanisms, or as a replacement to those approaches, providing a dynamic allocation of resources to the services from the overall resource pool as per 3.2, improving the resource multiplexing and hence the OPEX and energy cost of the scheduled services in the system.
- Deploy and provide network services to other operators and/or service providers when requested, or dynamically scaling up and down existing network service based on varying application demands and service requirements, via open interfaces. This way, other operators and/or service providers can re-sell/extend the provided network services.
- Realize fast lifecycle management (LCM), automatically triggered based on vendor-independent FCAPS management.
- Instantiate new components into a live production network in a plug-and-play manner.
- Terminate one or more network slices or service(s).

AI/ML-based network control – as a way to implement fully automated Smart Green Networks – seems like a must for future networks rather than a nice-to-have. To wit, the scale of deployments

made possible by function virtualisation, the extreme split in micro or atomic functions and the proliferation of more and more functions at the edge create network deployments of unprecedented complexity, challenging to manage and control with current decision support tools. Down the road, we see a need to overcome the current juxtaposition of conventional *model-based* approaches (which have, after all, driven the Internet for decades) with still untested but promising *data-driven* approaches and come up with integrated, hybrid solutions. Possibly, data-drivenness could compensate for fuzziness and uncertainty while model-driven approaches could provide a solid operational foundation.

- The system challenge here is to develop a future network with *Full Automation*, which reduces and tries to eliminate any human intervention. In principle, such automation can be achieved, once exact behaviour of all components is understood and expressed in a suitable model. In practice, however, for the highly complex and interwoven system outlined here, such a full-model description is not feasible, rendering *model-driven* automation and control impractical. For such situations, data-driven approaches leveraging powerful AI/ML systems might come to the rescue. One challenge here is to determine which data to use for what control aspect, using which AI algorithm. For example, there is a challenge that AI/ML is seamlessly applied to network control, to run automated operations of network functions, network slices, transport networks, in an end-to-end scope. As another example, we can consider the use of AI and machine learning for coverage hole detection and outage detection (AI and Machine learning can be used to predict the coverage hole based on MDT data and self-organise the networks, even, e.g., triggering deployment of UAV base stations to improve reliability of the networks).
- Another challenge is to develop robust AI solutions, which can adapt and adjust to the changes in the environment, with capability of transferring learning from past experiences, also detecting and preventing wrong decision making, when an anomaly occurs in the system. An anomaly in a distributed and complex system, such as a carrier network, could occur due to many reasons, and it is not clear how to detect and react to such anomaly. Should we trust the decision made by the trained policy, or should we fall back to some backup, safer, policy? And how do we detect anomalies in decision making to enable such fallback mechanism? Note that, in some scenarios, we might not be able to detect an anomaly directly but only through its impact on the fitness of decision made by the trained policy. Besides, specific attention should be also paid to the challenges when distributing the learning, e.g., how to enable an efficient learning having access to partial state (environment) information only. This is specifically important if we are deploying deep reinforcement learning techniques.
- In devising control and management strategies (often implying closed-loop control over different time scales), unless applying AI/ML techniques specifically meant to bypass the issue of modelling (in other words, acting as blackboxes, where both system dynamics *and* control are functionally approximated), we need to model a VNF in terms of consumption and performance versus load and configuration [C3-42]. In any case, attempting to model system dynamics does not prevent the application of AI/ML to the synthesis of complex control strategies. A possible approach would be to use models, e.g., at different levels and with different granularity – packet, flow, discrete- (queueing models) or continuous-state (fluid models), where available and feasible, to describe the dynamics of the system, and AI/ML to parametrize the functions expressing optimal control strategies as the solution of Infinite Dimensional Optimization (IDO) – i.e., *functional* – problems (see, e.g., [C3-37]).

Moreover, a thoroughly integrated AI scheme would open up new venues, how to think about operating a network in general. For example, suppose good to very good predictions (load, failures) were available. Then, the possibility arises to implement predictive behaviours in the network, to make available a network control intelligence capable of mitigating failures, the usage load, etc. and quickly adapt network configuration to be always available at the target performance levels requested by the applications. Basically, we could switch from closed-loop control to open-loop control (or, at least, to receding horizon control – also sometimes termed "open-loop feedback control" – where the optimizing control strategy is recomputed in the light of new observations leading to new predictions over a forward shifting time horizon).

### 3.6.2 AIaaS Proliferation in Service Provisioning

Beyond the use of AI/ML for improving on network operations directly, AI/ML will enable innovative features when provisioning future digital services for homes, businesses, government, transport, manufacturing smart cities and other verticals. At the same time, we expect a significant increase in the amount of machine-to-machine (sensor) communication monitoring smart cities, Industry 4.0, smart energy, etc. These changed traffic patterns will drive the move of computational and memory/storage resources from huge data centres towards the edge of the network, therefore impacting network designs to support this move. New services powered or parametrized by or even dynamically conceived through AI/ML may also bring significant socio-economic impacts together with improved sustainability models for Network Operators and might constitute a breakthrough in the service development. Indeed, for future platforms and services, AI/ML could be used to dynamically develop suitable network and service functions (NF, SF), e.g., from general stubs or derive best-suitable service-function chains (SFC) both for the operators of the platforms as such and for their users/subscribers. For prosumers, we foresee the proliferation of *personal data platforms* that are tightly connected with network services, and the development of tools allowing any involved principal to control their data and the models established from the latter. This is illustrated in Figure 3-1, where AIaaS is shown as a general block at the new fusion layer between the heterogeneous resources and the cooperative services: in this vision, the service layer is dynamically programmable over a composable infrastructure, with AIaaS kept at the novel abstraction layer. Here, AIaaS could be used to actively change or create service types and instances alike.

Key to reaping these benefits lies in utilizing the knowledge derived from the vast pool of network data in the services provided over the future telecommunication infrastructure but also utilizing the *highly distributed processing capability* that an AIaaS offering would provide. It is crucial to also understand the impact of M2M traffic, generating e.g., smart city data, etc., which will shape system designs. Both aspects drive the *provisioning of data into the system* as well as *complementing processing capability* of the network with service-level ones. With this, we see AIaaS capabilities of the infrastructure merge with those capabilities at the data and processing level that vertical customers will bring to the table. Consequently, we see an *emerging data marketplace* that goes beyond raw data (such as location traces) but is lifted to knowledge and insights provided by network operators to their service provider customers. For instance, radio measurements at the deep level of small-cell base stations can provide insights on physical objects that in turn can be

utilized by service providers for consumer-facing services that would have otherwise required dedicated hardware deployments or other means of realization. However, key to making an AIaaS useful for service providers, clear and open interfaces, both for data provisioning but also the reasoning logic, are required. Furthermore, *control over the distribution of, and access rights* to, both data and processing is crucial for the alignment with privacy considerations that both network operators and service providers will adhere to and may contribute to future regulatory aspects.

### 3.6.3 Resource-aware AI services

Although numerous AI and ML solutions have been proposed during the last decade for wireless systems, so far, those solutions focus solely on optimizing one or more operational or management aspects and procedures of the networked system. However, as specified above, a necessity can already be identified for a radical evolution for networks in order to introduce a novel paradigm, which will not use AI/ML only for optimizing certain network operations, but will ultimately integrate them as structural enablers of the system itself. To this end, AI will be an integral capability of diverse, heterogeneous network entities, spanning from the core, down to the deep edge; those AI-enabled network entities will be participating with their own heterogeneous types of resources (computation, communication, storage, energy, etc.) in complex operations, both via contributing as well as via consuming resources, towards an AI-as-a-Service (AIaaS) concept [C3-38].

These AI services need to be performed and coordinated in a distributed, or at least decentralized, manner. In a national scale networking context, performing centralized learning is simply too costly, as data would need to be collected and sent all over the network to that centralized entity. Also, such data collection and central training rapidly create bottlenecks and single points of failure in and around the central location. Moreover, if training uses private data, then a centralized learning might not be feasible/authorized.

To this end, an AIaaS operation control framework will be designed and developed, responsible for dynamically switching and selecting among available heterogeneous, and possibly unreliable, resources that will participate in the required AI operations in a distributed manner, targeting to optimize the trade-off that results from the respective resources' selection. Particularly, deep edge devices have usually uncertain and time-varying communication capabilities and are also constrained in computation and storage resources. Besides, distribution of data gathered across these devices could be non-i.i.d (independent and identically distributed). A big challenge therefore is how to perform a reliable learning over such heterogeneous and unreliable set of resources and how to deal with data heterogeneity. [C3-38][C3-39]

Energy is another aspect that should be taken into account. Training a deep neural network model is known to be a very energy-hungry process. Distributing such training on deep edge devices that have limited energy resources or that are running on green energy resources, which are by nature uncertain and time varying, is not very straightforward. How to adjust training complexities to the energy availability at a device? How to benefit from green energy resources? What is the accuracy vs green trade-off?

Finally, the most recent cloud computing programming models should be considered. In particular, Function-as-a-Service (FaaS) is very popular because the applications are realized as a composition of short-lived and stateless function calls, which is ideal to deploy following a serverless computing

paradigm. Due to the absence of local state and in combination with a flexible container-based virtualization infrastructure, services can be up-/down-scaled in a fast and easy manner and offer fine-grained billing granularity. However, AI/ML applications are very data-intensive, especially during the training phases, which leads to a high communication overhead and unpredictable tail latencies. Can NFV play a role in addressing cold start effects? Where to keep the application's data in a heterogeneous and fast-changing network architecture?

### 3.6.4 Research Challenges

While such AI/ML-driven or self-driving networking can start using existing AI and ML protocols, algorithms and approaches, it will gradually require network-specific adaptations in several regards. Below are some of the challenges we can identify in pursuing an AIaaS vision:

- One aspect is the **availability of network-typical and network-characteristic datasets** for training and validation. There is no commonly agreed reference dataset to use in research or development to compare different approaches against each other, nor is there a good understanding which data is actually needed to drive an AI/ML scheme, which features need to be extracted from an operational network.
- Similarly, current experience shows that the **procedures to train and validate** AI/ML algorithms and the architectures they use are mostly focused on static pattern recognition (e.g., images, sounds, diagnostics of fixed analysis data) and are therefore not well adapted to the nature of dynamic networks. We need schemes suitable for changing environments, changing number of users, changing topology, etc. – properties not typically found in popular ML algorithms.
- Even with suitable datasets and algorithms in place, **there is the need to extend the currently mostly centralized AI/ML algorithms to be distributed** to accommodate the distributed deployment in (often multi-domain and multi-technology) networks. This, in turn, will introduce challenges to ensure *scalability, consistency, consensus* and *convergence* of both data as well as decision making and reasoning in such distributed environment, providing auditable solutions that may foster future regulations. Complementing this need for supporting the distributed realization of AI/ML (which seems critical for general AIaaS) is the opportunity provided by the move towards *Edge and Fog Computing* that we can already see in 5G. This opens the opportunity to complement the resources of cloud computing data centres to analyse the expected vast amount of network data; it could even do so while better adhering to privacy demands through *localizing the processing* of raw data.In such a scenario, there are trade-offs between data volume to be transported vs. localized or distributed energy consumption and computational capacity; latency for training vs. latency for action; questions about ensemble learning when locally learned insights should be merged and generalized. For both learning and control in ML, we need a *meta-control* that allows for deciding, which data is fed into a learning scheme, where and which learned models are distributed to which place in the network for taking control decisions. This is similar to provisioning micro-services in general. However, it might have quite different data-rate/computational/latency/resiliency requirements compared to an application-level microservice. In other words, *AIaaS will need its own control plane logic* built upon the control plane capabilities of the infrastructure itself.
- Reliable learning over a pool of either transiently available or generally constrained resources, including specifically deepest edge resources, remains one of the biggest challenges, going beyond the question of full distribution of AI/ML towards full distribution with expendable resources.

- Network Functions (NFs) based on AI/ML: In future networks, systems may functionally operate using AI/ML techniques, rather than only traditional methods. As mentioned above, beyond managing a set of running NF instances, AI/ML can be used for automated parameterization of given and even programming of novel NFs. One can imagine code generated in CI/CD pipelines based on observations from NF use, to create more effective versions of NFs, then pushed to deployments to fundamentally improve behavior based on observations over time. In simple cases, configurations could be generated and pushed to production and model updates, where AI/ML engines reside inside the NFs. Rethinking the execution of NFs and code for them, are important areas going forward.
- Meta-control immediately raises the question of self-application: **can ML be used to decide on ML**? This idea is currently gaining ground in the Auto-ML community, where ML is used to learn hyperparameters of ML. Here, we need ML to learn, how to apply ML to a network. Clearly, there is considerable risk of oscillations, feedback loops, etc.
- The **scope of AI/ML schemes will also need to be investigated.** One possible, perhaps naïve approach is to have one set of AI functions/data sets that is applied only to a segregated, intra-service based scheme ("*sliced AI*"), which is easy to realize and ensures data privacy, but squanders possible optimization potential. Removing redundancy and going to a cross-service, cross-network, integrated AI/ML ("*integrated AI*") scheme is promising, yet fraught with complex design choices.
- Given the increasing multi-domain and multi-technology deployment of infrastructure, AIaaS will require the capability for *multi-domain orchestration* of distributed processing, meaning end-to-end interoperability is a must (cmp. Sections 3.2, 3.3). This requires greater standardization efforts and further progress in the functional architectures.
- Furthermore, **aspects related to security** beyond the conventional application of AI as a tool, e.g., ensuring data flow provenance and distribution within the system, and dealing with AI-enhanced (-amplified or even -rooted) attacks are essential. With the emergence of high-bandwidth and low latency requirements of applications for Immersive User Interfaces such as Wearable Cognitive Assistance (e.g., Google Glass, Microsoft Hololens), private 5G networks, and IoT appliances, the edge clouds or cloudlets are becoming ubiquitous. The security and performance of such private cloud datacenters is of paramount importance. Development of automatic verification systems to assess the performance and security of edge clouds by leveraging Open Source solutions like Central Office Re-architected as Datacenters (CORD) and operationalizing the results is another interesting avenue for the researchers from academia and industry.
- The topic of energy consumption of AI/ML algorithms themselves has started to be investigated. Recent work [C3-40][C3-41] contains some preliminary indications on how to compute the energy consumption during one cycle of inference. Further investigation will be needed, addressing both in-operation and training/adaptation power requirements.
- Novel and better solutions are needed to introduce **serverless computing and functional programming aspects** in communication and networking architectures, thereby providing the advantages of infinite scalability and flexible orchestration also to AI/ML services. The use of external services, abundant in the cloud but scarce and inefficient in edge/mobile systems, must be redesigned following a **frugal approach**, for which we need new development patterns and deployment models.

| **Research Theme** | **A self-learning, AI-Native, Service Provisioning Infrastructure** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Distributed, generic AI/ML (platform) | Mid-term | System as distributed joint AI/ML platform for various AI applications, enabling AIaaS, which can create and execute AI services on the fly.<br>Design of APIs for this platform, which would be usable by end-users (e.g., IoT, industry, subscribers), by mobile network operators, and even by the network functions of the system per se, to access, configure, and execute AI services while:<br>- requiring no exchange of raw data (potentially private, raw data stays at its source)<br>- reducing resource consumption (communication, computation, memory, energy, etc)<br>- scaling to systems of up to 100K nodes.<br>Relevant proposals should:<br>- Design a generic, system-integrated AI/ML platform capable of using up to 100k diverse nodes (different types, capacities, stability profiles) without introducing any single point of failure.<br>- AIaaS: Design a generic API suitable for both system internal functions (self-optimization) and for different system users (e.g. MNOs, subscribers, verticals) alike.<br>- Demonstrate that this system can orchestrate and execute AI services over the platform on the fly and in a runtime manner, without requiring human intervention.<br>- Demonstrate how this system-integrated AI/ML implementation can reduce resource consumption (resources used for learning and inference) on average, i.e. for a set of arbitrary AI/ML tasks, compared to a monolithic centralized AI/ML platform. | - Better locality and governance<br>- Better trustworthiness<br>- Universality of infrastructure<br>- Sustainability of infrastructure |
| AI/ML computing with and on transient/limited resources | Mid-term | Development of techniques and mechanisms for reliable and robust learning over limited, heterogeneous, and volatile resources that are available at edge and deep edge, which can:<br>- enable Few-Shot Learning through improving transfer learning capabilities<br>- scale to networks of 100k nodes | - Further impoving locality and governance<br>- Contribute to digital self-determination, potentially better privacy and data governance<br>- Faster deployment |

| | | | |
|---|---|---|---|
| | | guarantee fairness among different participants (devices, users, etc).<br>Relevantcontributions should:<br>- Show a system achieving stable execution of any AI/ML tasks in spite of potential instability of any used resource, supporting up to 100k nodes and average loads comparable to achievable loads with similar tasks on a system of a comparable stable resource size plus size-independent constant.<br>- Demonstrate that the system can provide both fairness of participation and fairness of contribution among the participants, guaranteeing that we learn from all data, even those gathered/restored in users with limited resources. | |
| Net zero (distributed) AI/ML | Long-term | Development of techniques, protocols and of a new architecture to promote the utilization of renewable energy resources (deployed and distributed all over the network, in the EU often available by energy regulation at the deep edge) for the learning process, so as to:<br>- enable a reliable and robust learning, while relying on uncertain and time varying energy resources<br>- provide an efficient greenness-accuracy tradeoff<br>- scale to large networks<br>- preserve privacy<br>Relevant solutions need to:<br>- Demonstrate a system achieving stable execution of any AI/ML task while relying only on green energy resources that are available and distributed over the network, achieving then the net-zero goal.<br>- Demonstrate that the system can support large numbers of users and large number of energy domains, and hence is scalable to real world settings. | - Greenness, sustainability<br>- Potentially better privacy |
| AI/ML driven system updates and evolution (self-updates) | Long-term | Evolutive and automatic adjustment and development of Network and Service Functions.<br>Final solutions should demonstrate a creation of a novel NF type from a stub or e.g. by combining modules from an existing library by AI/ML methods based on the expected features and KPIs. | - Sustainability, universality |

### 3.6.5 Recommendations for Future Actions

Based on the challenges above, we recommend research into the following aspects:

- making available network-characteristic datasets for training and validation;
- agreed procedures to **train and validate** the AI/ML algorithms;
- **distribution of AI/ML algorithms** instead of using centralized AI/ML algorithms, in order to apply AI/ML to a network, considering placement and distribution of AI/ML functions within a network;
- **meta-control procedures** applied for learning and control in AI/ML to decide which data is fed into a learning scheme;
- **integration with AI/ML features provided at the edge of the network**, e.g., when provisioning future digital services for homes, businesses, government, transport, manufacturing smart cities and other verticals;
- devise architectures, approaches and algorithms for **sliced vs. integrated AIaaS*;***
- **development of use cases** for new services powered by AI/ML at the network and service provider level;
- development of network management techniques embracing the AI/ML predictions;
- **support performance analysis and optimization** methods for energy consumption versus quality of service analysis, e.g., through an AI/ML enabled "data lake" approach;
- support for new AIaaS services and applications that require, e.g., **multi-domain orchestration** of distributed processing and end-to-end interoperability;
- address of **security and privacy challenges** and provide information for future regulation;
- usage of machine learning for **network anomaly/vulnerability pattern identification** and, thus, proven useful in identifying persistent threats/bugs/vulnerabilities.
- support the **provisioning of data** required for AI/ML learning phases, particularly from network infrastructure functions;
- address the **scaling requirements**, e.g., through partitioning mechanisms, to enable efficient AI/ML data processing to provide timely responses required by AIaaS solutions.

| Research Theme | A self-learning, AI-Native, Service Provisioning Infrastructure | | | |
|---|---|---|---|---|
| Action | Distributed, generic AI/ML | AI/ML on transient / limited resources | Net zero (distributed) AI/ML | AI/ML-driven system updates and evolution |
| *International Research* | | | | X |
| *Open Data* | | | X | X |
| *Large Trials* | | | X | X |
| *Cross-domain research* | X | X | X | |

## 3.7 Deep Edge, Terminal and IoT Device Integration

Architecturally, the 'deep edge' with its IoT as well as end user or vertical industry devices well integrates into the vision of Section 3.2 by becoming part of the common resource pool, provided as a non-decomposable set of resources by some edge entity, such as an end user, industrial site owner, or a building owner. Following the 'ownership through control' mantra, described in Section 3.3.1, we therefore envision tenant-specific resource usage to expand into the deep edge with the

same control and data plane considerations, as discussed in Sections 3.3 and 3.4 respectively, and resource management considerations, as discussed in Section 3.5, applying to all those resources. In other words, in principle, we see aspects of controllability of those edge resources to equally apply together with the general programmability for the realization of compute tasks as well as for data and forwarding plane operations through those resources.

However, some edge resources might not directly fit into this vision. For instance, IoT will introduce particular, service-dedicated, possibly intelligent yet resource-constrained components (micro-electronic, battery driven components), which will need a particular consideration for the integration with the rest of the system. Indeed, such IoT components and devices might impose additional requirements on, e.g., volatility and longevity, punctual presence at any moment, persistence, generality, capacities, connectivity, interfaces and APIs from/towards the system. Hence, they might not support direct integration and require particular solutions instead (e.g., gateways or subsystems).

Generally, edge resources often provide human- or task-centric input and output capabilities, expressed in a plethora of sensory capabilities, situational awareness, quality of experience perception, which make these resources very useful for integration into the overall vertical application. This yields a *richness* of resources that is challenging, when being integrated into a common resource worldview. Unlike the emerging COTS (customer-off-the-shelf) platform basis in other parts of the communication system, e.g., in the core, the edge provides a more *diversified and heterogeneous environment* with many device platforms and their supported local connectivity technologies (e.g., WiFi, BT, LiFi, and others), all of which are provided through a plethora of programming environments. **Future research will need to develop a suitable common model of system-wide representation akin to 'device drivers' in existing computing platforms.**

In this regard, novel forms of dynamic edge resource discovery, management and orchestration are required, allowing service provisioning to exploit on-premises deep edge devices as "on-demand" extensions of resources provided from the core or the edge. In this framework, novel resource control schemes, balancing between autonomy of devices and the overall optimization and control of the network by the operator(s) will be required, thus innovating the existing collaboration models between different network service providers. This will also allow to take in better account users' context, exploiting the typical co-location of users with on-premises devices and, sometimes, their very tight physical bound. In this sense, this approach will allow designing network services in a more user-centric way. **Future research will need to develop advanced schemes for adaptable and dynamic service provisioning involving deep-edge devices resources, taking into account their dynamic and sometimes hardly predictable availability (moving from deterministic to partly stochastic virtualized platforms).**

### 3.7.1 Massive Heterogeneous Edge Resources

This resource richness at the edge, however, often comes with a **limitation in capability**, e.g., in terms of available processing cores in smartphones that can be utilized in the common resource pool. Given that devices at the edge exhibit a high heterogeneity ranging from a simplistic sensor and IoT devices to edge data centres, other typical limitations include energy/battery, form factors, human-machine interface, storage, physical security. This stands in stark contrast to the perceived limitless resource capabilities in data centres as well as core networks and, therefore, impacts the

*decomposition* of computational tasks over a resource pool that is geographically and physically limited. As a consequence, the aforementioned *controllability* will need to be ensured through the realization of a suitable *control agent* that integrates the (edge) resource pool into the larger system but also interfaces with the (edge) resource pool to adequately govern the resource usage in the light of the resource-specific characteristics in terms of constraints and dynamicity. Here, research into the *minimal requirements* in terms of processing and communication needs and the realization of those requirements as *novel control agent realizations* will need to ensure that integration into the overall control fabric of the larger system to align with our vision of a smart network as laid out in Section 3.2. Furthermore, *resource scheduling* requires extra consideration in the presence of potential resource scarcity, particularly when combining specific input/output capabilities into the scheduling decision. Scarcity may be increased when utilizing specialized resources, such as GPUs or NPUs, rather than general purpose ones. We may also find that *locality of the resources* becomes crucial when applying policies for, e.g., localized processing for privacy reasons. Scheduling solutions are required that provide suitable trade-offs between moving data to functions or vice versa, possibly under locality constraints. Ultimately, a scheduling decision in favour of one tenant may result in detrimental performance of another, calling for solutions to resource scheduling that likely extend beyond those operating on a large pool of resources with uniform capabilities. **Future research will need to address these edge-specific constraints through suitable scheduling mechanisms that take those constraints into account, while relying on edge-specific control agents enabling the enforcement of the policies underlying the scheduling solutions.**

### 3.7.2 Dynamicity of Edge Resources

The *dynamicity* of (edge) resources is another aspect to deal with as an edge-specific constraint. While edge infrastructure, such as in an industrial site, can obviously be very well managed and long-lived, we also foresee edge resources of a much higher *volatility*, particularly when considering end-user provided resources, therefore creating a *limitation in availability* in contrast to, e.g., long-lived data centres. Those resources could be switched off, temporarily disconnected or simply become unavailable, e.g., if linked to human behaviours or policies (such as “do not make my phone available, if battery drops below 15%”). From a control perspective, *maintaining the basic control fabric* needs to take such dynamicity into account, while the *scheduling* will need to react to disappearing and reappearing resources alike to operate at a defined optimum of resource usage. From a data plane perspective, volatile resources need consideration when *routing packets* but also when *establishing in-network state* for forwarding operations. While volatility of resources and dynamics are already covered by the controllability framework presented in Section 3.3, **future research will be required to delve into the systems of systems aspect of such controllability**, given that individual subsystems might not be fully independent.

### 3.7.3 Governance of Edge Resources

Furthermore, *governance* of edge resources (and their provisioning through entities like individual users and localized industries) differs vastly from the often long-lived contractual relationships we can identify in the core network business. Instead, the addition and usage of resources with such volatile and temporary nature requires means for *contractual management*, including methods for billing, accounting as well as authorization of use that align with the dynamicity of the envisioned relationship. *Distributed ledger technologies* and *eContracts/ smart contracts* will likely lend themselves to being applied in this world of (possibly highly) ephemeral resource utilization with

the appropriate means to keep the resource owner (e.g., the end user) in the loop in order to preserve *digital sovereignty* but also enable *participation in the digital market,* akin to the changes in the energy market but likely much more dynamic. An important challenge for entering contractual relations is the *advertisement of resource capabilities*. While today's solutions are mainly focussed on the pure ability to communicate (e.g., through advertising a radio bearer), solutions are required that expand the negotiation towards clearly articulated *demands* beyond 'just communication' that can be dynamically matched against the *supply*. For instance, attaching to a WiFi access point is futile, if connectivity to particular backend services is not enabled at this edge resource. *Efficiency* is key here, avoiding unnecessary signalling between components. Particular consideration must also be given to *security*, both towards the tenant utilizing the resources and those providing them. With tenant-specific instructions eventually being executed on what are possibly end-user provided devices, *accountability* for this usage is key for accepting such usage in the first place, complementing (edge) platform capabilities such as secure enclaves to ensure trustworthy execution at the level of the computational instructions themselves. **Through research in this space, we foresee future solutions to enable an edge resource market that would allow for auctioning the availability of resources to tenants very much like the bidding for white space on a webpage as we know today, basing all interactions on a trusted, auditable, and accountable basis that caters to the dynamics experienced at the edge.** For this edge resource market to emerge, policy descriptions with their rules and constraints will need to be specified in a form that can be enforced by the infrastructure on the services, since direct human oversight is not feasible at this scale. **This will require research into novel programming models and (e.g., policy) languages that not only support all of these services, applications and deployments but also cater to the expected dynamics of the market itself**. Deploying and managing a large set of distributed devices with constrained capabilities is a complex task. Moreover, updating and maintaining devices deployed in the field is critical to keep the functionality and the security of the IoT systems. To achieve the full functionality expected of an IoT system, research should be done in advanced network reorganization and dynamic function reassignment. **Research is needed for providing new IoT device management techniques that are adapted to the evolving distributed architectures for IoT systems based on an open device management ecosystem.**

Decentralisation of IoT edge systems has been discussed in several publications, see e.g., [C3-43]. In particular, with the exponential rise in the number of devices, IoT is applied together with edge-centric computing to offer high bandwidth, low latency, and improved connectivity. Moreover, the cloud-centric platforms offer a high amount of data storage, but with deteriorated bandwidth and connectivity that affect the quality of service. The edge-centric Internet of Things-based technologies, such as Multi –Access Computing (MEC) and fog computing, offer distributed and decentralized solutions to resolve the drawbacks of cloud-centric models. However, in order to realize these distributed edge-centric models in the context of 6G, it is needed to realize the concept of decentralized distributed IoT Edge Systems, which should at the same time incentivize all the participants in the 6G value chain to share their edge resources.

### 3.7.4 Edge-Specific Architectural Considerations

The continued growth in video applications including augmented reality (AR) and virtual reality (VR) required by, among others, the emerging applications (cmp. Chapter 2), requires new architectural approaches and solutions. Surveillance and monitoring further complicate the space, as will the

growth in real-time sensor data e.g., for industry and smart cities. The ongoing shift of TV distribution from broadcast to the Internet will accelerate, requiring at least a 10x increase in video traffic volume with increased performance and resolution. The implications on application-level networking are tremendous: we will need to integrate video services with the web content framework, delivery model and APIs, with effective use of ultra-dense and diverse wired and wireless networks. Video provenance will become a key issue to combat ”fake news” and the effects of AI/ML-generated video that can subvert legitimate content. Strong security and integrity of applications, network transport and in-network processing will be required. A future key development in the system architecture can be the deep integration of application and service functionality pervasively within the network, as discussed in this document. To cope with that, this document introduces a highly dynamic system architecture (cmp. Sections 3.2-3.6).

This architecture will need to be supported by the nodes that constitute it (i.e., devices, elements, subsystems, etc. or whatever nature). Hence, at the node level, an active entity (e.g. an agent) becomes necessary, capable of a) offering runtime access to node-local resources and to all executed allocations and b) acting as part of a dynamic system, i.e., establishing and maintaining it. In its first controlee aspect, this entity is an entry point to the internal organization and realization of the node (e.g. of a whole subsystem). Exporting a common set of protocols and an API, it can hide the complexity of the internal organization through its own implementation and allow independent evolution of the node-internal and systemic organizations. In its second systemic aspect, this entity must autonomically and continuously construct, maintain and preserve the control plane considering the requirements in Sections 3.3-3.6. In other words, beyond service provisioning, management and security, which are critical to effectively manage billions of devices, ensuring they are suitably configured, running appropriate software, kept up-to-date with security updates and patches, and run properly authenticated and authorized applications, this entity must ensure system integrity and resilience of the programmable environment per se, while taking into account the available resources of the node that it represents. Chiefly, the agent must assure that both intrinsic and situational capabilities of its node (e.g. secure boot, local secure hardware modules, secure enclaves; input/output capabilities e.g. positioning, sensing usable for discovery of other potential nodes; topological position of the node, e.g. its connectivity degrees and its centrality; but also the available generic compute and networking capacity) correspond to the role, tasks and the topological position in the overall system in both directions. Therefore, a balance between agent commitments towards the system vs. resources required by the agent itself is required.

Locally, the agent must consider additional considerations. For instance, in addition to classical contractual models, micropayments might become a key part of the system as the infrastructure to support in-network services and applications is not free. Privacy and data management, and the location of processing and data to match legal and moral restrictions on data distribution, access and processing, will be increasingly important. Many of the services and applications will operate on, process and deal with personal data that is increasingly (and rightly) subject to strict regulation, control and limitation. Strong tools do not exist to describe in human language, legal language or code how data can be processed, located and distributed. Policy descriptions, rules and constraints will need to be specified in a form that can be enforced by the infrastructure on the services, since direct human oversight is not feasible at this scale. In addition, novel programming models and languages are required to support all of these services, applications and deployments.

### 3.7.5 Service Execution on Edge Resources

Processing at the edge in the architecture is essential for ultra-low latency and reliability, while the AI processing is already today often transferred to the mobile device. Research challenges in this area cover open distributed edge computing architectures and implementations for IoT and integrated IoT distributed architectures for IT/OT integration, heterogeneous wireless communication and networking in edge computing for IoT, and orchestration techniques for providing compute resources in separate islands. In addition, built-in end-to-end distributed security, trustworthiness and privacy issues in edge computing for IoT are important, as well as federation and cross-platform service supply for IoT.

Similarly, distributed service provisioning will extend also even beyond the edge, i.e., to on-premises devices such as Industrial IoT devices, robots, AGVs, connected cars. Novel forms of dynamic resource discovery, management and orchestration are required, allowing service provisioning to exploit on-premises devices as “on-demand” extensions of resources provided from the core or the edge. In this framework, novel resource control schemes, balancing between autonomy of devices and the overall optimization and control of the network by the operator(s) will be required, thus innovating the existing collaboration models between different network service providers. This will also allow to take in better account users’ context, exploiting the typical co-location of users with on-premises devices and, sometimes, their very tight physical bound. In this sense, this approach will allow designing network services in a more user-centric way.

### 3.7.6 Edge AI

Authors of [C3-44] estimated that 850 ZB of data were generated by people, machines and sensors at the network edge in 2021. The physical proximity between the data and the computational resources provided by the edge computing represents a promising marriage, the so-called edge intelligence or edge AI [C3-45]. Moreover, the recent booming of deep learning has been achieved thanks to the innovations in hardware, which allows to manage neural networks of many layers. However, these networks need more data in order to learn the huge number of parameters they are composed of. Moving these data toward a centralized cloud can be very inefficient in terms of delay, cost and energy. Therefore, in order to efficiently exploit data on the edge, the scope of edge AI is twofold: run AI models (inference) and train AI models (training).

For what concerns the training, the main problem of a distributed solution is the convergence of a consensus, i.e., whether and how fast the training can be considered finalized. This problem is related on how the gradient is synchronized and updated. Several solutions have been proposed in this respect, the most promising one being represented by federated learning [C3-46]. In this solution, the server is in charge of combining the results of the training of a shared model. Specific gradient methods have to be used, like the Selective Stochastic Gradient Descent [C3-47], which however is not optimized for working with unbalanced and non-i.i.d. (independent identically distributed) data. The frequency of the updates of the model at the central server is also an open issue. Too frequent updates allow to relax the hardware constraints of the edge, but imply more risks for the unreliable network communications. An interesting approach to overcome this issue is the Blockchain Federated Learning [C3-46], which allows to work without a central server by performing the updates via blockchain. Another interesting solution is the Knowledge Transfer Learning, where a teacher network is trained with general data and then student networks are

retrained on a more specific local dataset. This allows to reduce the resource demand at the edge devices.

For what concern the inference, the main problem is the limited resources of the devices at the edge. In this case the solutions try to relax the computational requirements of the model when performing the inference. In model compression, some of the weights can be pruned according to a specific policy, e.g., their magnitude [C3-48], the energy [C3-49]. In model early-exit the inference is performed only with a subset of the network, according to the latency requirements. On the other hand, to reduce the computational complexity on the device, *model partition* and *input filtering* represent interesting solutions, which rely on pre-processing the data on the device and perform the inference at the edge. When considering the processing of the original data, another technique that edge AI will need to investigate accurately on is data curation, which is the process of selecting the subset of data that is really valuable, especially when it comes from heterogeneous sources.

Particularly in the perspective of distributed services at the edge and beyond, edge AI is also a technique to keep data local to devices of their legitimate owners for privacy or ownership reasons such as, for example, data related to manufacturing processes in industrial environments. Moreover, keeping models “close” to edge devices might be the only viable solution to guarantee stringent time constraints. A research challenge is therefore how to guarantee accuracy and efficiency of both the training and inference phases given specific constraints in terms of where data can be moved inside the network.

AI will play an important role also for providing solutions to the resource management problem in edge computing, the so-called AI for the edge, which is complementary to the problems above, where the issue is how to carry out the AI process on the edge (AI on the edge). Typical examples are radio resource management in wireless networking, computation offloading strategies and services placement and caching. In this case the challenges are on the model definition, which often has to be defined as a tractable Markov decision process, on the algorithm deployment, since it has to work on-line and, consequently, a trade-off between optimality and efficiency has to be found.

A possible distinction regarding in particular the application of AI parametric approximation models adopted for control and resource allocation purposes on the edge may be between “function approximation” and “parametrized infinite dimensional (or “functional”) optimization” [C3-50]. The (functional) solution of many complex control and decision problems can be approximated by families of fixed structure parametrized functions, where parameters also appear within the basis functions themselves (e.g., one- or multiple-hidden-layer networks). If a family of approximating functions can be found that allows avoiding the so-called “curse of dimensionality” (the growth in the dimension of the parametrization with increasing number of variables the function to be approximated depends on), the optimization problem might be solved “off-line” (e.g., in the background in the cloud), whereas the “local” implementation of the decision strategies can be performed at almost negligible computational cost at the edge, over time frames within which the parameters do not vary. However, a possible problem to consider in this case would still be the transfer of big amounts of data to be processed. In this respect, techniques of local data aggregation and pre-processing, redundancy reduction, importance sampling and the like are worth investigating in this context. On the other hand, distributed computational methods for the local coordinated execution of parametric optimization techniques are also of interest, to perform the

strategy approximation over limited computational resources in the edge. It is also worth noting that parametric approximations of infinite dimensional (functional) optimization problems can be based on sound problem formulations, which can help understanding their algorithmic behaviour.

AI techniques and methods are necessary for IoT in an edge computing environment to provide advanced analytics and autonomous decision making. AI encompasses various, siloed technologies including Machine Learning, Deep Learning, Natural Language Processing, etc. In future IoT applications, AI techniques and methods will be increasingly embedded within several IoT architectural layers to strengthen security, safeguard assets and reduce fraud. Research challenges overlap with topics identified earlier in this document but it is worth mentioning AI-IoT integration subjects at the "edge" such as new energy- and resource-efficient methods for image recognition, edge computing implementations (neuromorphic, in-memory, distributed), distributed IoT end-to-end security, swarm intelligence algorithms, etc.

Finally, in the design of AI solutions it will be crucial to consider the energy consumed, as suggested in Section 3.6. The high energy requirements of deep learning solutions suggest that both industry and academia promote the research of more energy efficient AI algorithms [C3-48]. Moreover, all the new proposed AI solutions should be presented with their training time and computational resources required, as well as model sensitivity to hyperparameters. Examples of such analysis are the characterization of tuning time, which could reveal inconsistencies in time spent tuning baseline models compared to proposed contributions. To this respect, tools like Machine Learning Emission Calculator [C3-51] and Green Algorithms [C3-52] should be used to analyse, audit and report the carbon footprint of novel solutions proposed.

### 3.7.7 Research Challenges

Research challenges in this area include:

- **delivery model and API**s, with effective use of ultra-dense and diverse wired and wireless networks (cmp. Sections 3.3 and 3.4);
- need of effective management of billions of devices, ensuring they are suitably configured, running appropriate software, kept up-to-date with security updates and patches, and run only properly authenticated and authorized applications. Such management and terminal devices in general are to be integrated with the general architecture vision as per Section 3.2, e.g. as a potential IoT specific "allotment" including both hardware and software objects.
- **need of privacy and data management,** and the location of processing and data to match legal and moral restrictions on data distribution, access and processing.
- **need of policy descriptions, rules and constraints**; to be specified in a form that can be enforced by the infrastructure on the services (cmp. Section 3.4).
- IoT architectures applied in 6G, considering the requirements of distributed intelligence at the edge, cognition, artificial intelligence, context awareness, tactile applications, heterogeneous devices, end-to-end capabilities. This is also to be put in the context of the general architecture in Section 3.2.
- Research on distributed intelligence at the edge, cognition, context awareness, tactile applications and integration of heterogeneous devices. Autonomies and distributed intelligence in IoT towards the Internet of Autonomous Things. This a crucial topic for the AI/ML research, and is therefore already mentioned in Section 3.6.

- Need for open distributed edge computing architectures and implementations for IoT and integrated IoT distributed architectures for IT/OT integration, heterogeneous wireless communication and networking in edge computing for IoT, and orchestration techniques for providing compute resources in separate islands. Dynamic, partly-deterministic and user-centric virtualization, extending service infrastructure to deep-edge devices.
- Need for built-in end-to-end distributed security, trustworthiness and privacy issues in edge computing for IoT, as well as federation and cross-platform service supply for IoT.
- Need for novel resource control schemes, balancing between autonomy of devices and the overall optimization and control of the network by the operator(s) will be required, thus innovating the existing collaboration models between different network service providers.
- Need of deriving specific architectural requirements for distributed intelligence and context awareness at the edge, integration with network architectures, forming a knowledge-centric network for IoT, cross-layer, serving many applications in a heterogeneous network (including non-functional aspects such as energy consumption) and adaptation of software defined radio and networking technologies in the IoT.
- New AI techniques and methods are necessary for IoT in an edge computing environment to provide advanced analytics and autonomous decision making. AI encompasses various, siloed technologies including Machine Learning, Deep Learning, Natural Language Processing, etc. See relevant Research Challenges in Section 3.6.
- New AI-IoT integration challenges at the “edge” arise, e.g., new energy- and resource-problems with image recognition at the edge, edge computing implementations (neuromorphic, in-memory, distributed), distributed IoT end-to-end security, swarm intelligence algorithms, etc. See relevant Research Challenges in Section 3.6.
- Need for the design of AI solutions it will be crucial to consider their consumed energy. See Research challenges in Section 3.6.

| **Research Theme** | **Deep Edge, Terminal and IoT Device Integration** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| IoT architecture | Long-term | A suitable architecture to be executed within the general resource pool as per Section 3.2, customized for the particular needs of IoT, providing:<br>- Not only individual management of millions of heterogeneous often constrained devices, balancing the needs of the respective organization (efficiency, security, governance) and concerned users/objects (privacy), but also management of collaborative services and tasks executed by the latter.<br>- Efficient, adaptive, runtime communication environment for particular ultra-dense wireless environments with a capable, multi-modal delivery model.<br>- Efficient, adaptive, runtime edge computing and swarm intelligence. | |

| | | | |
|---|---|---|---|
| Dynamic, partly-deterministic, user-centric virtualisation | Mid-term | Extension of resource provisioning through dynamic inclusion of deep-edge devices' virtualized resources, to be assessed by:<br>- extensibility of resources provisioned for a certain CAPEX;<br>- speed of (re-)configuration of virtualized resources facing change of demand.<br>Final solutions should support:<br>- Management of dynamic resource provision by deep-edge devices in spite of instability and time limitation of devices' connection<br>- Working protypes of communication environments, where virtualized resources include deep-edge devices, and planning of their availability takes into account predictions of their future (short-term) availability via stochastic models for resource provisioning<br>- Large scale systems including management of deep-edge devices are ready for deployment | Flexibility, resource efficiency |
| Edge intelligence | Mid-term | Particular AI/ML mechanisms suitable for:<br>- The transient nature of resources in the IoT domain (links, compute resources), cf. Section SA.6.<br>- The constrained nature of devices (constraints of compute power, of energy, of time)<br>- Achieving guaranteed convergence of the insight in the swarm environment, i.e. facing the availability of many yet individually weak agents.<br>Design, conceive and demonstrate particular AI/ML mechanisms suitable for:<br>- The transient nature of resources in the IoT domain (links, compute resources), cf. Section 3.6.<br>- The constrained nature of devices (constraints of compute power, of energy, of time)<br>- Achieving guaranteed convergence of the insight in the swarm environment, i.e. facing the availability of many, yet individually weak, agents. | Universality, sustainability, flexibility, potentially better privacy and data governance |

### 3.7.8 Recommendations for Actions

| **Research Theme** | **Deep Edge, Terminal and IoT Device Integration** | | |
|---|---|---|---|
| Action | IoT architecture | Dynamic, partly-deterministic, user-centric virtualisation | Edge intelligence |
| *Large Trials* | X | | |
| *Cross-domain research* | X | X | X |

# 4. Network and Service Security

*Editor: Emmanuel Dottaro*

## 4.1 Introduction

The Universal Declaration of Human Rights [C4-1], Art.3 states that “Everyone has the right to life, liberty and the security of person”. By many aspects 6G Systems and Services cross security matters. This is actually the case, at least in European Fundamental Rights, for natural persons with respect to their personal data as stated in GDPR [C4-2]. Far beyond, 6G, following 5G enlarged scope, stands as a foundation of Digital Transformations involving natural, legal and up to national security issues. Whereas 6G is expected to be deployed in essential and critical sectors (private or public) of the society, Holistic security must be provided to mitigate inherent risks and ever growing number of Cyber-Attacks [C4-3] [C4-4].

## 4.2 Vision

As a general principle, Systems & Services evolutions (Architectures, Technologies, Operations, Usages) mandates concomitant evolution of the Cybersecurity. 6G aims at being the enabler of an unprecedented number of use cases encompassing large diversity of architectures (Cellular, Cell-less, Edge, IoT, 3D, mesh, Adhoc, Digital Twins, ...) with massive usage of AI (increasing the Attack Surface) and new technologies. If one consider the diversity of expectations up to Mission Critical but also Human-Centric, the Cybersecurity set of challenges can not be limited to classical hardening of some components or even obsolete perimetric and static approaches.

Last but not least, 6G is crossing multiples digital-related fields such as AI, Data, Micro electronic, Cloud, HPC, Quantum, Sustainability,...from security point of view, it results in a diversity of necessary regulations, knowledge sharing and transverse actions.

Among those actions it is worth to mention existing work achieved in regulation [C4-5][C4-6][C4-7], outcomes from the ENISA developed in a set of reference documents:

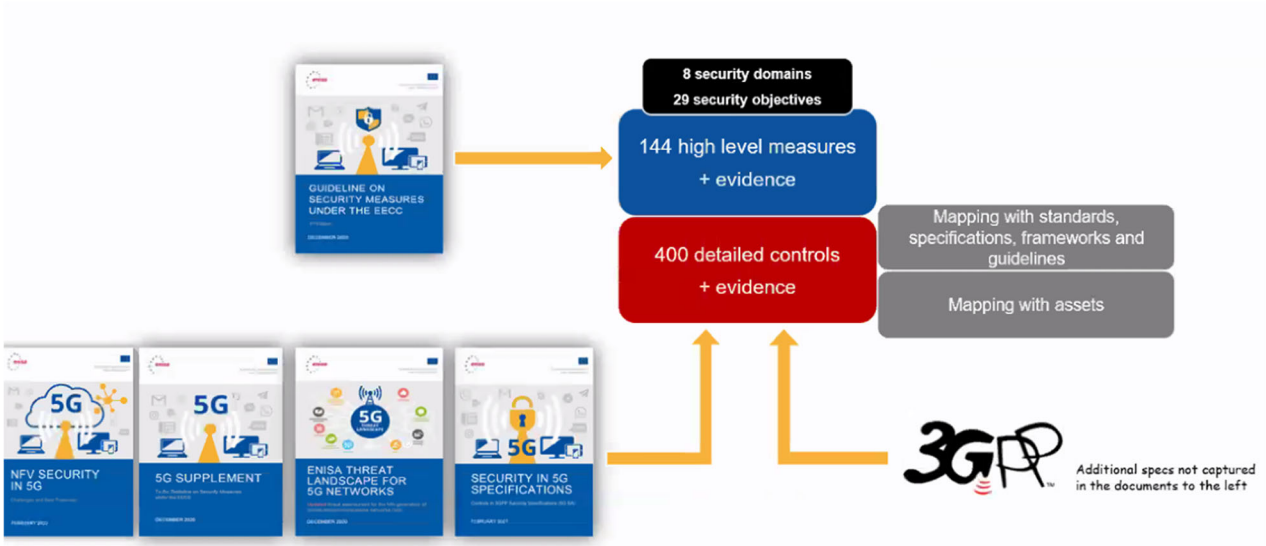



*Figure 4-1 Figure from ENISA presentation*

The State of the Art is ever growing on security and specific of security applied to 6G. Useful references should be found (updated on a regular basis) from associations, among others in [C4-8][C4-9][C4-10], as well as a recurrent topic in literature [C4-11][C4-12][C4-13]

- **Intrinsic nature of Systems & Services: Holistic & Metamorphic**
    - Basically inherited from the Middle-Age, the previous generations of cybersecurity suffer from more and more decoupling with the Systems & services evolutions. While security only makes sense from an End-to-End point of view, the challenge here is to deal with over-complex, highly distributed architectures. Multiplying fragmented multi-lateral, multi-layer, multi-party, (micro-)services "Black Boxes" perimeters with numerous interfaces and exchanges requires to re-think the **Holistic Distribution of Security** in all its phases (protection, detection, response).
    - Besides spatial distributions challenges, the transient nature of 6G Systems & Services configuration combinations is raising in turn a large set of security challenges. This requires holistic but also **adaptive security** fitting metamorphic properties of the systems. Among others, yet unsolved, challenges encompass, continuous/predictive assessment of security conditions, incremental certification, (Hybrid)AI-controlled time line of operations,...
    - Moving from Cloud-Native towards **AI-Native 6G** is opening a specific area for security. On the one hand the whole AI life cycle has to be secured. In particular, a direct consequence is to reinforce **data centric security** fueling 6G AI, together with xAI and assurance for models and behaviors predictability. On the other hand security will make full use of the AI power enabling enhanced 6G Cyber Threat Intelligence (i.e. OSINT for 6G), smart protection (from physical layer to services), Smart distributed and collaborative Attack Detection, Smart remediation and Response.
    - Security is horizontal and applicable to all Systems & Services parties. But one can wonder what is the global security level resulting from such heterogenous combination, made of more or less opaque (Black Boxes) segments. With well-known benefits (scalability, up-to-date features, consistency,...) of the XaaS paradigm, 6G should integrate **Security-as-a-Service** (SecaaS) provided by dedicated security vertical pure players.
- **Intrinsic need of diversity**
    - 6G is aiming at covering such a diversity of usages that it should be obvious that a "one size fit all" security model should either be too (dangerous, de facto trust in providers) basic or too (costly) high grade. It comes to the paradigm of **Differentiated Security (DiffSec)** Mimicking the multiple QoS-based attributes defining a Digital Service, Security Service Level Attribute (SSLA) should be one of them, based on Quality of Security (QoSec) criteria.
    - Once 6G will be capable to deliver DiffSec, users (Human or Machines) will be empowered to have a smart usage of Digital Services. Some challenges attached to this simple requirement consist in awareness of Services attributes through adhoc exposure, . request through security policies (ranging from Natural Language, Contextual specific syntax to Intent-based/semantic elicitation) and actual mapping into concrete combined provisioning some vertical applications may require formal proof).

The multiple 6G security challenges introduced above should be addressed through the following 7 areas :

- **Directions**
  - **Architectures & Strategies**: Both Protection and Detection **End-to-End Security Distribution** (beyond perimetric), encompassing the evolving diversity of 6G architectures (Edge, 3D, Mesh,..). Integration of Security Services including **Security as a Service** . **Differentiated Security** architectures. **Cooperative Holistic Security** across domains, layers, stakeholders. As unavoidable weaknesses will remain, in particular at termination points, **Root of Trust** distribution and **BlackBoxes Tolerant** architectures.
  - **Data Centric**: Data is key for privacy issues but also in control and management as fuel of AI-Native 6G. Beyond (lightweight) Post Quantum Encryption, processing of Data in 6G should be driven by dedicated approaches such as Sticky Policies (Data policy self-support), metadata, binding with hardware processing,...
  - **Hardware and Physical layer:** 6G is bringing new hardware in the picture such as Intelligent Surfaces. More stringent requirements on Clocks for Time Sensitve Networks. Although 3D or Cell-less architectures changing the attack surface. From Trusted Execution Element on terminals, hardening of newly introduced 6G components (incl. side channel attacks), to jamming and eavesdropping attacks 6G security research landscape must embed systematically security considerations for Components and Physical layer. This should also cover the supply chain and any operations before entering in production.
  - **Software & Virtualization:** Identified in 5G threat landscape as a main source of vulnerabilities Software —from Safe Source Code to the entire life cycle (incl. static/dynamic code analysis, OTA updates, Access Management & privileges) remain a strategic concern. Despite commonalities with the IT domains, 6G being more and more software predominant and disaggregated the question must be addressed with the complexity and authority fragmentation inherent to 6G. Virtualization tools and operations (OS, hyperware, APIs, slice controlers/orchestrators) are key to 6G and should participate to the Holistic approach with secured distributed interactions.
  - **AI-based Operational Security**: the overall goal is the application of the Zero Touch paradigm to security. It results on multiple research direction for smart deployment of Security in such complex 6G architectures. Protection is already massively complex solving policies elicitation and combination. But Detection and response is even more challenging. 6G AI-Native should not only take benefit of atomic AI-based function (xDR) but being able to integrate it in a holistic way with all subsequent interactions between suppliers, providers, users. 6G AI-based security should also encompass Lawful Interception issues as well as Root cause analysis and identification.
  - **Quantification, Evaluation:** There won't be trustworthiness if the Quantification and Evaluation are not there. As introduced above there is a need to define QoSec, provide approaches to evaluate it and maintain this information available from service request to the decommissioning. Thus continuous assessment is a challenging objectives mixing somehow certification complexity with E2E perimeters and dynamicity. One should note that forensic, liabilities and major societal impacts depends on the future capability to

evaluate the security quality (requires models, data lakes, potentially Digital Twins, friendly Hacking) and expose it to users,

- **Governance**: security is based on Standards, Open Source Communities all of it under multiple regulations. From education (Research Platforms as Cyber Range) to CTI sharing. Research actions may contribute to build a safe and secure ecosystem and make undoubtedly 6G acceptable from societal , industrial, strategic point of views.


*Figure 3-2. Numerous Sources of information relevant to 5G-6G security (ENISA)*

## 4.3 6G Security Architectures

Developing security architectures for providing E2E security assurance across the heterogeneity and dynamicity of technologies and architectures envisaged in 6G is a major challenge. The meaning of security from user perspective can't be anything else than End-to-End, hence leading to holistic distribution of security in all its aspects.

The solutions should be able to handle a diversity of 6G scenarios (cell, cell-less, IoT, 3D, private, public environments, integrated Digital Twins) as well Cyber Physical (CPS) scenarios integrating sensing as termination points of the systems.

Beside the complexity of the architectures, the ownership, the control & authority scopes are intrinsically fragmented with increasing interworking requirements and consequences on features distributions & liabilities sharing.

By its multiplicity of architectures patterns, flexible configurations and usage targets 6G is not addressing a single security level. IT makes no sense to provide 6G systems delivering high grade security when not required (for obvious economic reasons) as it makes no sense to underestimate security requirements. Recognizing that one-size-fit-all drives immediately architecture approaches where the resources and features are adequately provided as per the actual needs (time and space). We will talk in the following of Differentiated Security (DiffSec) by analogy with 20$^{th}$ century concept of packet technologies.

Security and trust are close companions, considering evolutions from Cloud-native to AI-native 6G, the question is raised with renewed intensity. Who can we trust an AI-driven system without xAI and security applied in AI deployment? From data collection for statistical AI to Control Loops and distributed smart multi-agent collaboration, securing both AI for 6G and 6G for AI is a pre-requisite of 6G advent. The societal perception of multi-Agents may be negative understanding the intrusive power into the privacy area. Trust in AI would need to be built on Human-Centric usage of the technologies remaining under well-defined boundaries and empowerment of the users.

Open access to digital services should be enabled by awareness of the security conditions associated with those services. Either for Business-to-Customers or Business-to-Business there is a fundamental need to provide integrated architectures qualified with security attributes.

### 4.3.1 Security Distributions in 6G Architectures

This aspect should address time and spatial distribution of protection, detection and response security capabilities across realistic horizontally and vertically fragmented architectures (multi-layer, multi-provider). This should notably encompass protocols and interfaces for E2E adaptive security delivery (inter-orchestrator, agent-based distributed convergence) ensuring multi-tenancy (e.g., verticals) remediation strategies with regard to business objectives (although vertical specifics).

Architecture patterns considered should be representative of 6G systems and technologies. As such the diversity of use cases will address Cellular, cell-less, Edge, IoT, 3D (NTN), Public, Hybrid, Private, Mesh, fully distributed D2D/V2V, Adhoc, Vertical specific architectures and Regulations, Distributed Ledger Technologies, Quantum-based architectures

Cooperative holistic approaches involving multi-layers/stakeholders authorities (including compute/network/security service providers) will be key together with smart distribution of root of trust AI capabilities to be used in AI-based Operational Security (SecOps)

Security distribution is both spatial and time and the following research challenges cover the four dimensions:

- End-to-End, Multi-lateral (stakeholders), Multi-layer Security functions distribution
- End-to-End Security Policies Decision and Enforcement distribution
- Hardware and Software Root of Trust distribution and adaptation along life cycle
- Digital Twins integration with massive remote management, sensing, monitoring

### 4.3.2 Differentiated 6G Security

There are no uniform requirements in security level. Three main axis are potential candidates to constitute the basis of resources and features segregation. The first one is driven by the usages and is based on business objectives. For instance, a business depending on time precision would requires assurance to the availability and precision of time and phase. Another one would like to protect its confidentiality even forbidding activity detection or analysis. A third one may have critical needs to escape from extra-territoriality of foreign countries. As illustrated above, the security differentiation may go far beyond the traditional Confidentiality, Integrity and Availability criteria. Sticking to industrial reality and geopolitics, both sectorial and regional regulation may apply and participate to the adequate 6G systems and Services provisioning.

Once some security differentiation exists and resources being limited, the question of the applicability of priorities and precedence will raise automatically. The limitation will obviously depend on the type of 6G architecture concerned i.e. Drone Swarms (adhoc network) don’t deliver same bandwidth and service availability as fixed cellular network. This covers:

- Multi-level Security and architecture profiles, plan-based (user, control, management, service) segregation, resource profile and isolation
- Priority & Precedence policies and mechanisms for security objectives

### 4.3.3 Secure Artificial Intelligence (statistical, hybrid) for 6G

Before contributing to the security of 6G for AI or AI-based 6G security (in other sections) we discus here the need of securing massive usage of AI for 6. AI has certainly caught tremendous attention has it has the potential to significantly change the operations of the network. But AI itself is subject to adversarial attacks that require security preserving the integrity (attacks degrading the AI models and models functionality), Availability (attacks interfering with expected operations).

The intrinsic security of the AI process depends on capabilities to prevent adversary to attack the models by input data poisoning, use response to queries in order to steal personal data or learn cyber defence parameters.

Beyond defending AI, societal concerns such as potential biased usage of AI should be purpose of research to guarantee legitimate use of the technology and build trust in 6G and services enabled by 6G. This covers:

- 6G-AI models security toolbox including
    - AI environment (training, development, production) evaluation
    - Vulnerability assessment of AI models and their applications
    - Protection measures along life cycle of AI models
    - Specific measures for constrained environment using frugal AI (embedded AI)
- xAI including both Statistical and Symbolic AI building Trust in AI usage in 6G

### 4.3.4 Human-Centric Multi-Agent & Federative Learning

A specific focus is given to multi-agents and federative learning as promising approaches fitting the distributed nature of 6G. Positions ranging potentially from typical far Edge form factors up to cooperative interdomain applications, the multi-agents & Federative learning should be developed with the intent to limit information spreading taking operations needs but also privacy and Human-Centric protection. This addresses:

- Private, secure close-to-the-source learning frameworks
- From Edge-AI to the Cloud 6G-enabled secure interworking

### 4.3.5 Service-Based Architectures

6G Systems and Services users, either Humans or Machines, should be empowered to choose their Digital Services as a function of security information and transparency associated to those services. In turn, Service providers investing in better security should be able to expose their differentiators to users. Through APIs, Catalog of services, market places or any service delivery workflows, security attributes should be exposed. The first research challenge in this domain is the ability to define and expose trustable security attributes. The attributes will then enable smart usage of 6G services and purpose of service agreement, composition End-to-End where security matters.

Either in public or private 6G architectures integrating Security as a Service (Managed or co-managed by Security Service Providers, MSSP) remains an interesting perspective allowing best of

bread security maintaining up-to-date knowledge (Cyber Threat Intelligence), benefiting from mutualized Security Operating Centers platforms and tools. Examples of existing Security services are Identity and Access Management as a service, Key management as a service, emerging xDR/monitoring services, etc… Nevertheless, these approaches are often limited to a specific organization, and as such are not capable to provide integrated holistic understanding of security conditions. Vertical integration (from physical layer to Service layer) and Horizontal (multiple domains E2E)

The service-based architecture consists in the two following independent topics:

- Security Attributes exposure & smart usage
- Security as a Service (vertical & Horizontal) Integration

### 4.3.6 Research Challenges

| **Research Theme** | **6G Security Architectures** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| **Security Distribution in 6G**<br>(4.3.1)<br>Holistic Security distribution enabling AI-based Security Operations across multiple layers/stakeholders in 6G architectures (cell/cell-less/3D/…) | Short-term | Reference architecture & framework for security deployment.<br>Granular interfaces (from micro-services to organization wide) and protocol exchanges (data models, syntax, semantic) for cooperative E2E security. | - Reaching constancy between security and system nature and complexity.<br>- Pre-requisite for adoption by most of vertical sectors and final users.<br>- Fostering Europe competitiveness in Security aspects of 6G. |
| **Mission-based Security Distribution in 6G**<br>(4.3.2)<br>Covering exhaustive security functions in exhaustive 6G architecture patterns with integration of Sensing and Digital Twins as Termination points and Root of Trust for Mission-based operations (including response) | Mid-term (with optional future long-term as per 6G Architectures evolutions) | Beyond obsolete perimetric approaches, integrating evolved 6G architectures with AI at the Edge and Root of Trust for smart remediation.<br><br>[KPI] Reference architecture & Framework coverage<br>2025: main 6G Architectures<br>2028: most 6G architectures with cooperative operations<br>2031:exhaustive applicability with Mission-Oriented applications | - Actual security transformation mitigating increasing attack surface and system architecture evolution, all with final mission objectives. |

| | | | |
|---|---|---|---|
| **Differentiated 6G Security**<br>(4.3.3)<br>Definition of actionable Security level criteria & associated control mechanisms | Mid-term | Resource segregation and security profiling in 6G architectures<br>Priority and Precedence mechanisms applications.<br><br>[KPI] Multi-Level Security 6G resource profiling and associated mechanism<br>2025: plan-based segregation (user/control/management/service)<br>2028: flexible profiling of 6G resource as per security level and associated control | - Delivery of relevant security level/features through Differentiated services level and user awareness |
| **Secure AI for 6G**<br>(4.3.4)<br>Protection of AI usage in 6G | Short-Term<br><br>Mid-term | 6G-AI Toolbox including:<br>-6G-AI life cycle evaluation<br>-Vulnerability assessment & risk analysis<br>-Protection measures<br><br>-application to constained (frugal) environment<br>-adversarial attack detection<br>-xAI & Trust<br><br>[KPI] AI Tool Box Availability & Enforcement<br>2025: recommendations for most of 6G-AI protection<br>2028: coverage of constrained use cases<br>2031: Regulation as per evolution of AI Act, xAI | Pre-requisite, thus enabler of massive AI usage societal acceptance in 6G |
| **Human-Centric Multi-Agents & Federative Learning**<br>(4.3.5)<br><br>focus on distributed approaches enabling Human-Centric protection | Mid-Term | Private & Secure learning frameworks & Federated Learning applicability from Edge to the Data Center | |
| **Service-Based Architecture**<br>(4.3.6)<br>Integration of Security either as attributes to digital services in general or as Security services; | Mid-Term | Security Service Level Attributes exposure & usage (awareness, E2E composition)<br>Security Services integration (SecaaS)<br><br>[KPI] Security Integration in Services<br>2025: available SSLA template for most of service delivery workflows.<br>2028: Digital Service life cycle integrating provisioning of Security as a Service for most common usages<br>2031: International mutual recognition of SSLA | Digital decade must come with security but security delivered in similar agile mode as the rest of digital services components. It should also leverage skills and competitiveness of EU security industry |

### 4.3.7 Recommendations for Actions

| Research Theme | 6G Security Architectures | | |
|---|---|---|---|
| Action | DiffSec (4.3.2) | AI-based 6G (4.3.3) | SBA (4.3.5) |
| *International Research* | Regional angle considerations for global applicability | | Global applicability |
| *Cross-domain research* | | Sharing with AI community, in particular for xAI | |
| *Regulation/NSA at least ENISA* | May rely on numerous standards and region-based regulation | | Mutual recognition and composition rules |

## 4.4 Strategies and paradigm shift

### 4.4.1 Beyond perimetric strategies

There is a mismatch between the previous generation security strategies, basically inherited from middle-age (even defense in depth) and the nature of 6G systems. Some promising directions have emerged in the last period for disruptive strategies. Most of them are taking benefits of flexible technologies and are enabled by the technology evolution. 6G research Roadmaps for some of them may be relatively short term (i.e. “Zero Trust”/confidential computing paradigm) others are longer term perspective.

Among known strategies of interest, we can mention:

- Deception aiming at luring the attacks with fake emulated systems,
- Moving Target Defense (MTD) and its potential derivations taking benefits of 6G flexibility may change continuously system morphology to prevent attacks,
- Spatial fragmentation of data, processing, routing making difficult or even impossible re-assembly of the information. This type of approaches being also potentially used for recovery with given N:M replicates of data fragments.

Listed here as one topic, each strategy may be subject to a set of research challenges, under the scope of:

- Innovative and disruptive strategies

### 4.4.2 Black-Boxes and new attack Tolerant Architectures

The postulate here is simple: we will de facto never have homogeneous security levels (or just understanding/trust) across the technologies and architecture of 6G systems. Thus, instead of seeking for security grade elevation, it should be more efficient to study integration of weak or unknown parts with appropriate countermeasures. Most probably those countermeasures will have to be complex, smart, active, detecting, filtering attacks, misbehavior, anomalies in a consistent way as per the overall service objectives. A simplistic example may fit into IoT gateways, filtering data and detecting attacks (DDoS for instance) from low-security objects. Topic:

- Countermeasure integration for untrusted sub-systems and Services

### 4.4.3 Recovery strategies

Similar to Disaster Recovery philosophy, 6G critical systems should anticipate massive and sophisticated attacks able to create a Nation-wide blackout. Obviously, the strategies and mechanisms to be envisaged here must be self-protected. Research challenges may encompass minimal set of resources protection preserving identified critical missions, graceful remediations minimizing service interruptions with scheduled service resilience. This can be handled as:

- Mission-critical aware degraded modes and graceful remediations

### 4.4.4 Per vertical specific security profile

A major application of Security differentiation should be the purpose of vertical sectors specific profile. On a case by case basis 6G should demonstrate its capability to satisfy key requirements in terms of security and security assurance. Topic to consider:

- Specific requirements and solutions as per vertical needs (e.g. synchronization, formal proof,;..)

### 4.4.5 Research Challenges

| **Research Theme** | **Strategies and paradigms shift** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| **Beyond Perimetric Strategies**<br>(4.4.1)<br>What is the future strategies for 6G Cybersecurity.Has the current perimetric already failed ? Pushing Zero Trust, Confidential Computing, Deception, Moving Target Defense, quantum-based or any disruptive strategies enabling the expected fundamentals of security (CIA) | (short to) Long-term | Innovative and Disruptive strategies applied to or leveraging new technologies and 6G architectures. Examples of such approaches are given in challenges description.<br>The approaches should be promising in terms of capabilities, functional and/or non-functional gains.<br><br>[KPI] Security Capabilities<br>2025: selection of most promising directions as per preliminary capability gains estimates<br>2028: some gains from strategies demonstrated and validated and/or first deployments.<br>2031: Strategies mostly adopted for actual 6G deployments. | Decoupling between systems & technologies and security is a challenge and a risk for the Digital transformation itself.<br>Digital future is dependent on security evolution and security is a valuable differentiator and investment in sustainable democratic way of life.<br>Investment in security research for 6G is paving the way for a safe future. |

| | | | |
|---|---|---|---|
| **Black-Boxes and new attack Tolerant Architectures**<br>(4.4.2)<br>Integration of blackboxes and new attack surfaces in 6G conditioned by relevant mitigation strategies keeping security level. | Short-term | Despite unavoidable weaknesses and/or black areas in the systems and services, the target here is to develop strategies (policies, solutions) to provide expected security levels.<br><br>[KPI] 6G Risk Tolerant Architectures<br>2025: IoT risks mitigation by means of filtering/monitoring<br>2028: Most of termination points vulnerabilities mitigations<br>2031: End-to-End continuum assurance through advanced strategies for most of 6G architectures | |
| **Recovery Strategies**<br>(4.4.3)<br>Anticipating hierarchical recovery strategies for Mission-Critical services (6G dependent) | Mid-Term | Scheduled plan for 6G dependent service recovery as per critical needs (up to nation/Region wide)<br><br>[KPI] Recovery plans maturity<br>2028: most of essential and critical services performing dependability analysis and recommendations including degraded modes for recovery phase.<br>2031: recovery plans concomitant deployment with continuous adaptation. | Lack of stability worldwide, remind every day the need to anticipate recovery plans as our society is more and more depending on Digital systems and services. |
| **Vertical Specific Security Profiles**<br>(4.4.5)<br>Completion of KPI set per vertical applications. This should encompass security levels and specific attack surface. | Short-Term | Providing matching between specific verticals and 6G security capabilities (potentially any functional and non-functional requirement expected from 6G.<br><br>[KPI] 2025: >80% of 6G applications forecast covered.<br>2028: >90% coverage | Expected value is to enable extended applicability of 6G (and beyond) to specific vertical needs. Thus a clear enabler for economic development but also essential services such as Health or Public safety. |

### 4.4.6 Recommendations for Actions

| **Research Theme** | **Strategies and paradigms shift** |
|---|---|
| Action | Vertical Specific Security Profiles |
| *Cross-domain research* | Liaison with related programs |

## 4.5 Data Centric Security in 6G

The digital world is data centric, so 6G is. While all Data Centric security issues are not specific to 6G, an endogenous set of challenges deserve both fundamental understanding and shorter-term applications.

**Intra-6G Data protection**, often based on advanced cryptographic technologies, is key for privacy and confidentiality. Similar to trackers covering a majority of website, (too) invasive massive data collection and information can be retrieved from behaviours and 6G usages. This constitutes a societal risk for citizens but also an economic risk from Economic Intelligence and its illicit version industrial espionage.

6G Systems and Services will provide much more than “pipes”, In-Network or Edge computing has evolved into a standard component. As such 6G will integrate data processing in numerous architecture patterns. **Intra-6G Data processing** should be secured and should deal with the generalized ciphering. Data may self-described (Sticky Policies) with whom it can be shared, who can simply access, what functions are applicable… In the longer term this raise some Data/Hardware binding Secured optimizations (metadata, neuromorphic designs..).

6G is expected to be AI-native, and Artificial Intelligence capabilities are depending on data quality, integrity and availability. Either for Users' data or System’s data, Data security becomes a major concern of **AI for 6G**. The three classical dimensions of security must be satisfied Confidentiality (multi-party architectures), Integrity (data poisoning for instance in sensing/communications fusion), Availability (Federated, Distributed AI). Massive usage of AI in 6G control and management is directly increasing the Attack Surface and trustable 6G.is in turn a pre-requisite for trustable AI.

Data Centric security in 6G, beyond the integration of AI needs to be assured with respect to the 6G-enabled architecture diversity. We already mentioned In-network and Edge computing but most of the architectures comes with stringent requirements on Data security. This the case for Autonomous Systems (based in turn on Zero Touch 6G), Digital Twins (in particular used for operations of critical systems), Sensor-driven systems, 6G for AI for smart X (X being a city, a building, an industrial system remotely managed,…)

### 4.5.1 3.5.1 Intra-6G (All type) Data Protection

The main focus is here, the privacy/confidentiality of all type of Data. The data typology in scope is not limited to User’s payload data, but also glocalization, mobility, behaviour, usages of 6G. The topics

The related aspects of Data protection in 6G considered is listed as follows:

- Post Quantum Cryptography, algorithms, management, Energy Efficient designs
- IoT, embedded form factors, Data security under Size, Weight and Power (SWaP) constraints, secured Data mutualization -multiple Access to sensors)
- Beyond point-to-point paradigm: Distributed Ledger, Tor-like data overlay connectivity
- GDPR conformance and anomaly detection
- Anonymisation, pseudonymisation, counter-measures against inappropriate learning
- Lawful Interception in 6G and antagonism to data confidentiality measures.

### 4.5.2 Intra-6G Data Processing

Basically, Intra-6G Data processing is addressing 6G In-Network computing capabilities. This research area is following two main directions:

- Data Security Sticky policies (owner defined): open challenge to attach to the Data the necessary information (self-sufficient) usable for its processing in the 6G architecture. Beyond the definition or even standardisation of the syntax and the semantic of this information, the field may provide advanced and sophisticated ontologies for access and usage, identity-based solutions, traceability (for instance by means of watermarking).
- The second ambition is to re-think the Data Plane to make it, by design respectful of Data processing security. This binding between Data and Hardware may benefit from evolution of cryptographic technologies such as Full Homomorphic Encryption (FHE), Multi Party Computation (MPC), neuromorphic designs, metadata processing,…

### 4.5.3 Data powering 6G AI

As AI applications in 6G covers multiple different scope, the Data security will result in various level of requirements. One may for instance consider differently the local data for massive MIMO-implementations contributing to future modulation and coding schemes compared to data poisoning risk as root of AI for 6G End-to-End control and management. It can address societal concerns such as potential biased usage of AI and includes both the threats directly applicable to user data traffic, and their control and management.

The 6G scope is including a specific interest in federated learning architectures and platforms close to the edge, to enhance data protection, improve inference reliability, and increase autonomy of end clusters.

- Integrity and Availability of Data enabling smart control, management, service
- Data and/or models inter-domain (End-to-End, Multi-layer, Multi-party) distribution secured and trustable approaches. Benefits from promising technologies such as Distributed Ledgers, blockchains or zero Knowledge Proof (ZKP)
- Resilience and recovery responding Data failure
- Abnormal Data detection

### 4.5.4 Data security (CIA) in relation to exogenous impacts

The data centric security in 6G is a criterion for applicability of System architectures evolution forecasted. According to the usages of Digital Twins, Remote management, robotics, autonomous systems or even indirect contribution to sustainability, the requirements for Data Centric security in 6G will mandate specific security level. One should notice that integrity and availability is becoming more and more important for Data.

- Data security for 6G-enabled from sensing to Digital Twins or remote management including real time availability
- Data security for 6G-enabled autonomous systems
- Data security for 6G-enabled sustainability optimization

### 4.5.5 Research Challenges

| Research Theme | Data Centric Security in 6G | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| **Intra-6G Data Protection** (4.5.1) Set of data protection topics applied within 6G systems including cryptographic protection, anonymization, DLT and GDPR conformance check. | Short-Term | Protection of all type of Data within 6G (users' payload, usage data extracted from system, 6G control and management data.<br><br>[KPI] all Type of Data protection<br>2025: > 90% completeness<br>2028: 100% completeness | In line with general data protection and its (among other) privacy goals, 6G must demonstrate exemplarity in Human-Centricity |
| **Intra-6G Data Processing** (4.5.2) Data processing in 6G driven by sticky policies and confidentiality-preserving technologies | Mid-Term | Solutions providing strong guarantees for data processing in 6G, including user's privacy and policies<br><br>[KPI] Secure Data processing in 6G<br>2028: availability of solutions for most pof 6G use cases<br>2031: scalable applications with power constraints | In line with general data protection and its (among other) privacy goals, 6G must demonstrate exemplarity in Human-Centricity |
| **Data powering 6G-AI** (4.5.3) Secure Data and its distributions for 6G-AI | Short-Term | Security of Data (AI-feed) life cycle in 6G architectures, including inter-domain and federated learning.<br><br>[KPI] 6G-AI data security<br>2025: integrity assurance solutions for most of 6G applications<br>2028: full secure life cycle with anomaly detection | As the foundation of digital services, in turn serving numerous verticals, 6G must ensure that its behaviour will not be at risk with biased data and models. |
| **Data Security in relation to exogeneous Impact** (4.5.4) Enabling system evolutions such as Digital Twins, Autonomous systems, Sustainable systems by fully secured data transport | Long-Term | Assurance for systems depending on 6G systems and Services enabling smart and innovative applications.<br><br>[KPI] Dependability on 6G data transport<br>2028: partial mapping of exogenous systems requirements onto 6G<br>2031: 6G systems & services commitment to most of required guarantees | Most of promising digital services , DT, immersive applications will be strongly dependent on data transported by 6G systems and associated Key performances (delays, availability, integrity,..) |

### 4.5.6 Recommendations for Actions

| Research Theme | Data Centric Security in 6G |
|---|---|
| Action | Intra-6G Data Processing |
| *Cross-domain research* | Solutions strongly dependent on cryptographic technologies from cybersecurity domain and microelectronic Hardware capabilities |

## 4.6 Hardware for 6G security & Physical Layer issues

Expected to rely massively on virtualized infrastructures, the directions enabling deployment of security functions in the hardware layers is at least two folds:

The first one, is related of the implementation and distribution of **Root of Trust** and Secured environment in the overall system complexity. It encompasses Trusted Execution Environment (**TEE**), (micro)Hardware Security Modules (HSM) in a scalable way from CPU-limited terminals to Data Centers. It may notably contribute to the so-called "**Zero Trust**" paradigm. Specific security appliances virtualized remain part of any network security architecture and may deserve specific integration issues with the evolution of 6G architectures.

Another direction gaining traction, is to recognize the performances and sustainability issues raised by the load in VMs and containers induced by some infrastructure functions, including filtering and other network security functions. These is illustrated by the Open Programmable Infrastructure/IPDK project aiming at offloading Network, security, learning functions.

Smart usage of security hardware is dependent on two main constraints:

- Although true for non-security specific components the control of the supply chain from design to delivery remain part of the tools required to ensure security.
- Along the life cycle the management of **secret elements** (keys, attack patterns,…) is a fundamental as weakness on this side may ruin any security level expected from the security hardware.

Still related to 6G & hardware security, there is constant need to take into account the security aspects of any hardware component integrated in the architecture. 6G is coming with new specific components which deserve attention. This is the case in general with sensors involved in 6G being for instance dedicated to sensing/communication fusion or **Intelligent Surfaces** involved in communication capabilities and such part of the Attack Surface.

Clustered here with the Hardware issues, physical layer and particularly bearer protection is the source of some challenges and a threat landscape on its own. The jamming issues have to be re-considered considering the diversity of architectures (including 3D and cell-less) and critical needs that can occur for vertical applications.

Waveform sophistication and smart but complex access protocols may be source of exploits and malicious usage. Security considerations have to be taken here with as most as possible "by desgin" protections.

Intention for Eavesdropping/IMSI catcher type of attacks are not going to disappear with 6G. Topics not covered by data protection should be also in 6G scope.

### 4.6.1 Network Security Hardware

- Root of Trust Distribution and life cycle.
- Security functions Offloading in virtualization infrastructure (from Edge to cloud)
- Trustable Degraded modes, Graceful 6G (private/public) recovery
- Secret Elements distribution management

### 4.6.2 Securing Network Elements

- Secure LIS/RIS
- Clocks, time/phase distribution
- Flexible Hardware secured deployment
- Advanced Supply chain Assurance, authentication and traceability

### 4.6.3 Bearer protection

- Anti-jamming and counter measure in 6G
- Protection against intentional source of light (optical domain)
- Authentication, IMSI catchers,…

### 4.6.4 Research Challenges

| **Research Theme** | **Hardware for 6G security & Physical Layer issues** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| **Network Security Hardware** (4.6.1) Design and Run of HW-based security in 6G | Short-Term to Mid-Term | Open Framework for security function offloading in virtualization infrastructure. Trustable environments and their usages.<br>[KPI] Availability of trustable infrastructure framework<br>2025: System Overarching with security offloading and root of trust.<br>2028: > 70% Applicability on 6G use cases | Security is a complex combination where secured hardware should contribute to high level of assurance. Added vale from HW is both in interest of control of HW supply chain and overall security. |
| **Securing Network Elements** (4.6.2) Security of newly introduced 6G hardware | Short-Term | Secured 6G components and their supply chain | Security considerations can't be an option but should be mandatory for any 6G component participating to the 6G architectures. |
| **Bearer protection** (4.6.3) Denial of Service and eavesdropping mitigations | Mid-Term | Affordable Anti-jamming solutions and Attack Detection | The risk of attacks against bearers is quite Asymmetric…jamming being quite easy and cheap. (critical) services availability and privacy are dependent on bearer protection, raising value to mutualized infrastructure. |

## 4.7 Softwarization

Research related to softwarization is occurring in multiple aspects in the architectures, AI-based control and management or confidential computing. The focus addressed here is a subset of software related topics, basically safe code, remaining open research on virtualization enablers and virtualization of security functions.

It starts with the source of any softwarization that is the safe code. Beyond safe code, which is an entire cybersecurity area, the whole Software life cycle is driving the security conditions. Updates, upgrades, exchanges, workflow, privilege management...the path towards system control is much more at risk than previously with quasi-unique path from OSS to EMS. Looking at the emergence of 5G, first security analysis such as the one done by ENISA [C4-4] shown the tremendous importance of the software and its life cycle. Next generation and evolution are clearly re-enforcing the trend and the needs.

Many threat Intelligence aspects are already addressed in the literature including a comprehensive survey [C4-12] and numerous softwarization enablers feed on a regular basis the Common Vulnerabilities and Exposure (CVE) list. The architectural 6G trends tend to accentuate the potential issues with more distributions, more interfaces leading to more interfaces/exchanges and thus wider Attack Surface.

From security point of view, software-based and virtualized functions have been considered weaker than those based on hardware. Nevertheless, it would make no sense to block all the system flexibility without mitigating the issues raised by virtualization. Security is thus purpose of evolutions in virtualization areas:

- secured “hyperware”: Operating Systems, binding with Hardware platforms, Hypervisors and containers. The issues related to integration of security considerations here are not limited to functional aspects. There are direct impacts on overall performances and sustainability as well as challenge evolution linked to the architecture evolution itself.
- virtualization of the security functions for various platforms from objects, terminals towards cloud servers. This includes a large diversity of functions from classical firewalls, towards Detection & Response systems (xDR) with potentially sophisticated smart protection and detection inside.

Security is still on a long and complicated path towards fully secure and resilient 6G. Softwarization is dependent of the diversity of the platforms and their specifics (Edge, IoT, space. Softwarization is expected to follow and adapt to the evolving 6G architectures, although matching a wide range of security level expectations.

### 4.7.1 6G Safe Code life cycle

The agility and operational gains enabled by the softwarization on going since 5G has to be handled with the counterparts in terms of cybersecurity. As any software those taking part of the 6G systems and Services start with code. Drawing a dependence chain, it becomes obvious that software has at least as critical impact as hardware on the overall security.

Safe code is a common concern for the whole ICT domains, integration of advanced research in this area is one of the topics but it should be complemented with specific 6G aspects. A fundamental

difference with some IT similar issues may be found in the distribution of the system with multiples authority perimeters, multiple layers, multiple platforms, disaggregated microservices,… Combined with the dynamicity of the systems and the intrinsic Code temporal distribution (design, supply chain, runtime, updates, upgrades, patches, traceability…) it leads to identify 4 distinct research aspects:

- Code analysis, beyond static code analysis. Mapped into 6G architectures it is critical to understand complex interactions and potential effects of vulnerabilities and malicious. Similar and actually not decorrelated from AI a focus on federative approaches should enable high levels of assurance as well as automation of the operations.
- The second critical research aspect is the code timeline with the Updates/upgrades critical phases. Most probably maintaining code security is not scalable without development on incremental analysis or even certification.
- A safe code life cycle should build upon a trusted supply chain. For the benefit of operators, system integrators or verticals users a stable and trusted source of elementary or even integrated (certified) codes would contribute to secure deployment of 6G. It should be understood that multiple verticals deploy systems with very long life cycle/generation (several decades) and would have difficulties to sustain too fast, unchecked code integration.
- Last code specific aspect is traceability, forensic and technologies for identification. On one hand approaches such as watermarking or equivalent may give some guarantee of the sourcing, on the other hand, methodologies based on code morphology (AI-based) should help to identify what is actually at stake in the large protocol stacks and code running in the systems.

### 4.7.2 Full Security for 6G virtualization

Ensuring fully secured virtualization enablers may appear as a solved problem as large communities use to address those issues during the last period. Unfortunately, literature is listing year after year specific and comprehensive survey [C4-12] on virtualization software generic or implementation-specific components vulnerabilities. Subscribing to Security Alert from Computer Emergency Response Team (CERT) may be sufficient to realize how long and frequent is the Common Vulnerabilities and Exposure (CVE) list. It even concerns on a regular basis well-known Operating Systems, open or closed.

Sometimes, a simplistic vision is to believe that problem is solved but limited to a single authority, benefitting from a “god view” on its system. 6G is clearly going in the other direction. The intrinsic nature of networks is to interconnect, whatever the granularity (micro-service, sub-system, high level service, service provider,…) of entity interconnected overall security is only meaningful End-to-End and is not a simple sum of local properties. 6G is often coming with visions of Inter-computing, where computing may result in blackboxes (from system view) running on non-standard features and implementations. Another typical challenge is to actually control and manage security concomitantly to the other digital components of the architecture (communication, compute, storage, AI,..). This may be considered as the more complex, dynamic 6G version of the historical issues of NOC/SOC relationship (different clearance levels, different tools/view scope, different control/management scope).

Security in virtualization is often seen as an isolation/confidentiality issue. By full security here the intention is to cover all the security dimensions: confidentiality, integrity and Availability. The later is a stringent requirement for most of critical verticals while available failover mechanisms or current virtualization rely on lower layers for protection and restoration.

Last but not least, controllers and orchestrators are Policy Decision (PDP) and Enforcement (PEP) points. That is critical asset in the infrastructure to maintain sovereignty in the sense of guaranty on system behavior with respect to expectations from legitimate policy maker.

From the above statements, two research aspects are emerging as contributing to 6G security and use case coverage:

- Resilience, Protection & Restoration in virtualization layers
- Resource scheduling and performances under security (collaborative Inter-Controllers & Inter-Orchestrators policy-based operations)

### 4.7.3 Virtualized Security Functions

In softwarization domain, a sub-class is made of the necessary virtualized security functions. The availability and integrability of such protection, detection & response functions is just mandatory to be the companion of conventional network functions. As mentioned in previous section, performances, interactions (common syntax, semantic, APIs) across various perimeters and granularities constitute a set of research challenges in order to address evolving cyberthreats (up to Zero-Days vulnerabilities and Advanced Persistent Threats). Research aspects are:

- Scalable Cryptographic and filtering functions
- Convergent knowledge sharing, syntax & semantic for Detection & Response. (xDR including NDR)

### 4.7.4 Research Challenges

| **Research Theme** | **Softwarization** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| **Safe Code Life Cycle** (4.7.1)<br>Code Analysis, updates/upgrades, traceability | Short-Term | Framework and process enabling full life cycle mastering<br><br>[KPI] Uncontrolled code ratio<br>2025: <10% code in production<br>2028: Zero unsafe code in critical applications | Reality of softwarization is strongly dependent on the ability to control code/software life cycle. Impact and consequences of safe code is here at the dimension of 6G usages. |

| | | | |
|---|---|---|---|
| **Security of 6G virtualization**<br>(4.7.2)<br>Protection and restoration mechanisms for virtualization components<br>Collaborative, policy-based scheduling | Mid-Term | Fast failover and cooperative protocols for resilient by-design controllers and orchestrators<br><br>[KPI] Response time of protection and restoration mechanisms<br>2025: <1s (most of)<br>2028: <50ms (availability) | Virtualization capabilities in terms of security should come with stringent requirements for protection and restoration. State of the Art today in this area is often seconds or even 10's of seconds response time and not compliant with many vertical applications. |
| **Virtualization of Security functions**<br>(4.7.3)<br>Frugal cryptography<br>xDR | Mid-Term | Development of deployable security functions contributing to CPU, energy consumption reduction<br><br>[KPI] Sustainability performance<br>2028: <15% supplementary cost (CPU, Energy) in E2E chain | Integration of security in 6G. As any network function should be available as virtualized function keeping its robustness properties and adding sustainability criteria. |

## 4.8 AI-based Operational Security

### 4.8.1 Security Policy Life Cycle

Assuming that the architecture is providing the relevant Policy Decision Points and Policy Enforcement points distributions, the protocols and APIs. Multiple challenges remain to feed the system with the required intelligence and make it operational.

The first set of challenge, and the starting point of security policies life cycle, is to achieve semantic extraction, provide elicitation of user-driven requirements potentially expressed as security policies. It may also be based on Natural Language Processing (NLP) and put in Intent-based perspective. It should be noticed that talking security policies at the interface of users and providers may come with obligations and consequences on liabilities.

The second set is mapping of the user's requirements into systems policies, configuration, orchestration. As mentioned, many times, 6G is not a monolithic system under a single authority, thus the solutions will rely on various distributions and combinations End-to-End. This type of issues is already difficult with past systems using powerful solvers in bi-lateral relationship. The challenge is therefore difficult to take up and will be even harder considering that some vertical applications may require formal proof to validate the overall policy implementation.

Two priority sets of challenges towards users/providers matching in security policies:

- Intent-based, User requirements semantic extraction
- Solving distribution & combination

### 4.8.2 Zero touch, autonomic and multi-agent

A promising 6G application is to enable Autonomous Systems for instance Drones Swarms, Robotics, vehicles, etc… Those systems will hardly rely on communications which are not themselves autonomously controlled and managed. Zero touch is the paradigm for 6G and security must follow the same path to avoid blocking 6G while maintaining the expected level of security.

Far beyond self-configuration of security component, hybrid AI will be massively used to provide adaptive protection but also adaptive attack detection and response. The Ai capabilities are intended to be the solution in order to face the extremely large number of events and variations of the problem. The ideal case (long term) being intrinsic reasoning capabilities to deploy response to unknown attacks based on zero days vulnerabilities.

This where Digital Twins may participate to the security operations through modelling of the system, reasoning and validation features allowing smart responses.

Last but not least, coordination of security features should be done in a cooperative holistic approach to cover the scope and complexity of the systems.

The research topics are listed as follows:

- Adaptive protection, Hybrid AI
- Adaptive detection & response, Hybrid AI
- Digital Twins for reasoning and validation
- Cooperative holistic security

### 4.8.3 Root cause and Identification

Attacks may basically be classified into three categories.

The first one, Criminal is the most known as numerous visible occurrences of those attacks such as Ransomware are subject to communications and impact many users…

More critical, espionage category may have higher societal impact but is less known that the first one. This is also not the same level of sophistication attacks and take sometimes years to be discovered. This category can already involve states or industrials mandated by states with significant number of skills and efforts. This type of attacks is less frequent but is increasing.

The last category is nation-wide or even military grade attacks, sophisticated with the potential to create major blackout in the essential interests of a nation.

As baseline for digital transformation and enabler of numerous verticals, 6G will definitively be in the targets of the three categories and will definitely be a vector for the attacks ! In this context and at least for the most dangerous categories, the legitimate authorities must be able to conduct investigations towards identification of attackers. This is particularly complex mixing public/private means, potential fakes made on purpose to accuse someone else,… AI-based security is an indispensable means to overcome these challenge. The research topic may be called:

- AI-based security for root cause and identification

### 4.8.4 Research Challenges

| Research Theme | AI-based Operational Security | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| **Security Policies Life Cycle** (4.8.1)<br>Intent-based User to System/services security policies mapping<br>Solving E2E policies combinations | Mid-Term to Long-Term | Fluid interface between needs and solutions respectful of security policies<br><br>[KPI] Automation of policy life cycle ratio<br>2028: >50%<br>2031: >75% | Simple, easy to,use and accurate translation of user policies into systems operations and configuration. |
| **Holistic Zero Touch** (4.8.2)<br>Adaptive protection<br>Adaptive Detection/Response<br>Digital Twins for reasoning<br>Cooperative Holistic Security | Mid-Term to Long-Term | AI-based security control and Management I 6G Architectures<br><br>[KPI] Solution availability<br>2028: >75% Protection, xDR<br>2031: >50% Reasoning & Cooperative Holistic applicability | Squeezing the most of AI technologies for autonomous operation of security in 6G. |
| **Root Cause and Identification** (4.8.3)<br>Multi-Source: forensic, OSINT,…)<br>AI-based Identification | Long-Term | Defense against ever growing sophistication of all types of attacks | Strategic capabilities |

## 4.9 Security Quantification and Evaluation

Trustworthiness is not going to be built on declarations (potentially under conflict of interest) basic static measures. A major challenge for 6G is to develop an entire framework allowing definition, evaluation of security levels and awareness/usage of those levels.

### 4.9.1 Quality of Security (QoSec) and relation to Security Service Level Attributes (SSLA)

The closest analogy to understand the QoSec is the quality of Service in general. Service security is not a new concept and was already mentioned in the ITU-T E.800 in late eighties. Nowadays, criteria definition is the first step and means to evaluate it the purpose of research challenges. The path towards QoSec is potentially inspired by certification history, but needs to integrate 6G complex scopes and also societal aspects such as immunity against foreign laws.

QoSec should be the roots of SSLA and basic enabler of exchanges within Service-based architectures (APIs, Intent-based, Service Catalogues, marketplaces, B2B/B2C interfaces...)

The related research topic is then:

- QoSec definition and evaluation

### 4.9.2 Continuous assessment of security conformance along life cycle

The security monitoring is facing new challenges as stationarity of the security conditions (resources, features, perimeters, updates variations in time) are not verified in virtualized and flexible 6G architectures. It leads to develop on the fly means to evaluate the security conformance

providing continuous assessment. It also the benefit of security to check security much frequently and detect potential anomalies.

Within assessment scope, there is also a challenge to demonstrate E2E provable security with the two dimensions: composition by combining multiple, lateral sub-systems and services, and incremental to deal with change in time of the configuration/composition of the systems/service. Research topics:

- Continuous assessment of 6G security
- E2E provable security including composition and incremental methodologies
- Representative Platforms and data lakes

### 4.9.3 Research on Economic and Societal impacts, liabilities

Security evaluation is expected to participate to regulation, liabilities, insurance issues and perception of 6G impact in general. At the crossroad of Human Sciences, legal laws and technologies, research topics targeted here are:

- Usage of security quantification and evaluation in 6G societal impacts

### 4.9.4 Research Challenges

| **Research Theme** | **Security Quantification and Evaluation** | | |
|---|---|---|---|
| Research challenge | Timeline | Key outcomes | Contributions/Value |
| **Quality of Security** (4.9.1) Definitions, roles (vs. Risks and Trust), Evaluation | Mid-Term | Integration of Multi-Level Security in Users/service & service/service interfaces. Evaluation framework in relation to Regulation. | Before limited to certification to limited scope, concepts of QoSec should enable knowledge, differentiation, added value and awareness of 6G systems & services. User empowerment. |
| **Continuous Security assessment** (4.9.2) On the fly means to capture security conditions variations as per transient system states | Long-Term | Trust by monitoring of security boundaries up to provable conditions | Evaluation capabilities should evolve as systems are evolving. State of the Art is somehow late and poor today compared to the need. Any service should come with means to monitor the service quality, security attribute is one of them. |
| **Economic & Societal impacts** (4.9.3) Convergence of 6G security and Societal Impacts | Mid-Term | Using quantification and Evaluation as a Factor of 6G success. Actual figures usable by economic (insurance, industries,…) and society as a whole | Vertical specific regulation or simply cyber-insurance issues, smart usage of secured digital services and means to achieve that have a massive societal impact. |

### 4.9.5 Recommendations for Actions

| Research Theme | Security Quantification & Evaluation | | |
|---|---|---|---|
| Action | QoSec | Economic & Societal | Continuous Security |
| *International Research* | Global consensus & Standards | | |
| *Cross-domain research* | | Human Science (Legal Law, Economy, Sociology) | X |
| *Regulation Authorities* | Liaison and relationship with Certification/labelling framework | | |

## 4.10 Security Governance

There is a potential EU perspective for cooperation among private and public organizations to develop a dedicated 6G Cyber Threat Intelligence (CTI) as well as building cooperative response to incident. This type of logic is already at stake in other ICT domains and fully make sense for 6G. The topic should ne:

- Security knowledge sharing for 6G developments

### 4.10.1 Research Challenges

| Research Theme | Security Governance | | |
|---|---|---|---|
| Research Challenge | Timeline | Key outcomes | Contributions/Value |
| **6G-CTI**<br>Building common knowledge on 6G security | Mid-Term | CTI platforms and policies across EU or wider) | Strengthening 6G security posture and response, at least starting and reinforcing European cohesion. |

### 4.10.2 Recommendations for Actions

| Research Theme | Security Governance |
|---|---|
| Action | 6G-CTI |
| *Cross-domain research* | EU initiative in liaison w/ ECCC, ENISA and NSA |

# 5. Software technologies for telecommunications

*Editor: Josef Urban*

## 5.1 Introduction

Software technologies are one of the fundamental enablers of telecommunication networks and they increasingly shape network architectures and capabilities. For example, 5G has adopted a service-based architecture (SBA) for its core functions providing flexibility and scalability. Standardized APIs (e.g. Network Exposure Function – NEF) have been introduced to provide applications with access to network resources and data in a controlled manner. Network slicing is a key concept of 5G, built on NFV, SDN, and the flexible SBA of the 5G core allowing the dynamic creation of multiple virtual end-to-end networks across the same physical infrastructure and offering network services tailored to specific use cases. DevOps approaches are applied to develop, integrate and deploy network services and software updates in an agile way. Open source software has become increasingly important to be witnessed for example in case of open software for the RAN [C5-1]. Open source has been proven to be a successful model allowing competitors to work together towards common platforms and de-facto standards not only in the telecommunication industry, but more general in software-intensive business ecosystems.

## 5.2 Vision

The network softwarization will continue. Future networks are envisioned to be built over heterogeneous federated ”clouds”, whose resources are homogenously managed by a unified control and orchestration framework. This framework will form a computing continuum in which network functions and services will be created, deployed on demand, subsequently scaled, and seamlessly moved across the federated cloud infrastructure. The computing continuum will be able to optimize autonomously service performance and possibly off-load computation based on in-depth knowledge about the capabilities and resources exposed by the federated clouds. Service meshes and workloads will be based on stateless and serverless functions, microservices running in containers and virtual machines, as well as new advanced concepts that allow to optimize the use of specialized high performance resources including quantum computing resources. The concept of the computing continuum will impact architectures, interfaces, and the disaggregation of networks.

The network softwarization will also include the extensive use of artificial intelligence and machine learning models throughout the network and even at the radio level, and digital twins will allow to simulate and test networks. Software-based capabilities of smart networks will play a significant role for the commercial success of SNS ecosystems by addressing the digital needs for automation, adaptive and customized services, as well as the needs for agility in delivering complex, but reliable and trusted software and services.

The network softwarization requires research to get answers on questions such as how to manage the lifecycle of AI/ML components and to assure access to and the trustworthiness of data required by AI/ML, how to guarantee that a self-adapting AI/ML component will behave within its design parameters, or how to engineer and integrate such a software-intensive system in general so that the growing system complexity can still be managed. It has to be explored how the software needs

to be architected so that it is best adapted to the distributed 6G system and benefits most from the capabilities offered by the ICT continuum across devices, edge, and cloud. Even quantum computing resources will become available as part of the ICT continuum raising the question how to integrate and use these special compute resources in 6G. We also need to understand better how the non-functional requirements of sustainability, human-centricity, and resilience will impact the software architecture of 6G systems and applications.

The following software research themes have been identified and will be further outlined in the following sections:

- **AI-powered edge cloud compute continuum**. It explores the role of AI and federated learning to enable the edge for data protection, improved inference reliability, and supporting autonomous device clusters as well as the role of the network to aggregate and share insights from and among distributed IoT edge clusters. It also covers the use of AI-based cognitive capabilities to manage the distributed resources of the edge cloud compute continuum in a predictive and efficient way.
- **Automated and agile software engineering**. A key question in this area will be how future CI/CD pipelines and DevOps approaches will be designed to support a fully automated software lifecycle bundling software, AI components, APIs, and security into one development, delivery and deployment process as well as the enhanced integration with business processes on the network operator side. Also, low code platforms are considered as an easy-to-use development environment for applications and services.
- **Enablement of digital services.** This area covers approaches and software technologies supporting the creation and provisioning of SNS based services including services that can help to achieve sustainability goals of verticals.
- **Engineering complex, software-intensive, and self-adaptive SNS systems.** Managing the software complexity of a system of systems is becoming an increasingly challenging task and requires new operational concepts based on self-adaptation models and relying on AI algorithms as well as new SW engineering approaches for software intensive systems.
- **Software architectures.** This research theme explores software architectures and mechanisms that support offloading of computation-intensive tasks. Also, the use of quantum computing in the telecom context is covered in this theme: what are the telecom specific complex problems and algorithms suitable for quantum computing and how those algorithms can be implemented and integrated with classical computing.
- **Human centricity and digital trust.** SNS-based services should follow a service model that enhance human-centricity, meaning that services need to be trustworthy and ergonomic including easy to use and easy to access. In consequence, the development of software and services need to follow a “human-centricity-by-design” approach.
- **Digital twins in the SNS context.** Networking and the computing continuum will be important enablers providing the infrastructure and basic services required for the implementation of digital twins. Digital twins will be also used for monitoring and augmenting SNS systems.

## 5.3 Metrics, KPIs and benchmarks

The proposed research themes aim at improvements of various software related aspects such as the performance in service delivery, the human-centricity of network services, or the reliability of network systems. Metrics, KPIs and benchmarks will allow to measure the progress and the

improvements that research will achieve with regard to those aspects. There are existing and well-established measurement approaches in software engineering and software deployment that can be used to also measure the achievements of related research. However, in other areas such as human-centricity those metrics might still need to be developed in the context of the research.

KPIs that might be applicable in the context of software-related research for future networks include

- KPIs to measure the DevOps performance[C5-2]: deployment frequency (how often releases and updates to production are done; lead time for changes (the time it takes to get a commit into production); change failure rate (percentage of deployments causing a failure in production; time to restore (time it takes to recover from a failure in production)
- Measuring the accuracy of machine learning models and the efficiency and failure rate of MLOps
- Sustainability related KPIs that can be defined on the basis of the Energy Efficiency Directive [C5-3]
- Benchmarks and KPIs to assess security by design and secure cloud deployment environments [C5-4]
- Monitoring of network management functions following the FCAPS model

## 5.4 AI-powered Edge Cloud Computing Continuum

### 5.4.1 Federated Learning and AI for IoT Edge, applied in 6G infrastructures

Typically, federated learning involves training AI statistical models over different types of equipment, including data centres, end clusters (e.g., end IoT/IIoT devices, on-premises servers, etc.) or remote devices, while keeping data as much as possible localized. In particular, federated learning brings AI models close to the edge to enhance data protection, improve inference reliability, and increase autonomy of end clusters. The cloud (data centre) plays a federation role for aggregating insights from different IoT edge distributed clusters to generate a federated model shared with each individual cluster. Training in potentially large and heterogeneous networks, such as 6G cellular networks, introduces novel challenges that require a fundamental change compared to standard approaches for large-scale machine learning and distributed optimization. In order to enable this shift distributed IoT models, applied in 6G, with embedded AI are recommended.

### 5.4.2 Proactivity of the future network

Adding proactivity to the future network means providing the network with the capability to implement autonomic decisions, thanks to the use of AI/ML. The objective of such decisions may aim at improving how the network operates, how it performs in the services delivered to the customers and how the network itself is managed, configured and healed when issues arise. For such reasons, 6G is expected to deliver a network with built-in AI capabilities, where analytics functions are potentially collocated in every network function instance, take part in the business logic and decision-making and provide highly distributed optimizations, potentially within any of the existing and future defined network functions and procedures.

The Service Based Architecture of the fifth-generation mobile network already contemplates the integration of data analytics functions paving the way for the future developments and making the network more proactive. However, in a bigger picture to deliver its promise, 6G networks must be complemented with solutions such as low power wide area networks, transparent end-to-end TSN

communication with wired and wireless devices or new interactions with three-dimensional networks, such as the proposed Space-Air-Ground Integrated Networks (SAGIN). It adds a new level of complexity to the expected 6G networks, which, thanks to these built-in AI/ML capabilities, will be able to provide optimization to end-to-end services, even proposing an integrated digital twin of the heterogeneous network itself, supported by these AI/ML capabilities. Achieving an end-to-end and coordinated proactivity over the heterogeneous 6G ecosystem remains a big challenge to be addressed.

To make it a bit more complex, as the computing-oriented service will become popular in various public places and enterprises, their needs and requirements has to be considered when we talk about the proactivity. It means besides a coordination between various data transport domains, a coordination and harmonization with the application layer is needed. Diversified edge computing nodes will appear, and be publicly available, with wide deployment of edge computing nodes in different forms, with edge capability fitting specific purposes. In such an environment where various forms of edge computing nodes are deployed in a distributed manner, an essential requirement for the distributed edge computing nodes is to cooperate and collaborate in a federated manner. To dynamically configure and adjust network and computing resources in a federated and continuum edge environment, a coordination that can control and orchestrate resources based on an AI technique is an important requirement. The number of federated edge computing nodes in a distributed environment may have influence on the size of machine learning models and the frequency of overall coordination and control of distributed edge nodes, leading to inefficiency in performance and energy consumption. Therefore, it is required to construct and deconstruct federated edge nodes in a flexible manner while allowing their control and coordination to be efficient.

### 5.4.3 Research challenges

| **Research Theme** | **AI-powered Edge Cloud Computing Continuum** | | |
|---|---|---|---|
| Research Subtheme | Timeline | Key outcomes | Contributions/Value |
| Th1 - Collaborative work required in the context of federated learning for IoT devices and services discovery, applied in 6G | Short-term (finished in 3y) | Federated learning Architecture and Platform, including interfaces, data model, orchestration<br>Federated learning will bring AI models close to the edge to enhance data protection, improve inference reliability, and increase autonomy of end clusters. | AI enablement of 6G |
| Th2 -Challenges-workflow in (1) standardization, (2) on interfaces edge/cloud, (3) the orchestration, (4) model contamination, and (5) on the pipes for handling the distributed traffic | Mid-term (finished in 5y) | Work on open source and standardisation aspects on federated learning architecture, interfaces for edge/cloud, API, data models, orchestration and open source contributions<br>See Research subtheme 1: federated learning will bring AI models close to the edge to enhance data protection, improve inference reliability, and increase autonomy of end clusters. | AI enablement of 6G |

| | | | |
|---|---|---|---|
| Th 3 - AI as native feature for proactive networking | Mid-term (finished in 5y) | Supporting diverse and novel applications, building a new ecosystem, and delivering user-centric service experience through fully leveraging the built-in communicating, computing, and sensing capabilities of networks<br>Adding proactivity to the core framework of future networks provides the network with the capability to implement autonomic decisions, thanks to the use of native and embedded AI/ML features at 6G core. | AI-based cognitive and context-aware resource federation and optimization of edge compute continuum |
| Th 4- Next generation network data analytics as the fuel of the future AI-enabled networking | Short-term (finished in 3y) | The AI support based on the Network Data Analytic Function will be extended with AI plane and AI Functions to optimise control and user plane operations in scenarios requiring deterministic latencies and energy efficiency<br>Support of a fully distributed and collaborative AI approach across the various RAN, edge and core domains. | AI-based cognitive and context-aware resource federation and optimization of edge compute continuum |
| Th 5- A pervasive, extremely capable resource control | Short-term (finished in 3y) | Autonomic and distributed conflict resolution, correctness enforcement and distributed resource scheduling schemes empowered with the advanced resource allocation to support context awareness and cross-layer design for time sensitive 6G applications allowing several user equipment (e.g. AGVs, connected cars, IoT systems) to have specific resources without any conflic*t*<br>Distributed AI approach to avoid single points of failure, to bring learning closer to the event sources and to be able to harness the available yet scattered compute power. | AI-based cognitive and context-aware resource federation and optimization of edge compute continuum |
| Th 6- Autonomous AI-empowered control and management plane | Mid-term (finished in 5y) | Autonomic and distributed conflict resolution, correctness enforcement and distributed resource scheduling schemes empowered with the advanced resource allocation to support context awareness and cross-layer design for time sensitive 6G applications allowing several user equipment (e.g. AGVs, connected cars, IoT systems) to have specific resources without any conflict<br>Distributed AI approach to avoid single points of failure, to bring learning closer to the event sources and to be able to harness the available yet scattered compute power. | AI-based cognitive and context-aware resource federation and optimization of edge compute continuum |
| Th 7- Integration and AI/ML optimization of heterogeneous networks integrated with 6G SBA | Short-term (finished in 3y) | Federated learning for heterogeneous architectures and different network resources orchestration | AI enablement of 6G |

### 5.4.4 Recommendations

| Theme | AI-powered Edge Cloud Computing Continuum | | | | | | |
|---|---|---|---|---|---|---|---|
| Action | Th 1 | Th 2 | Th 3 | Th 4 | Th 5 | Th 6 | Th 7 |
| *International Collaboration* | common and global federated learning architecture | common international standards on federated learning | | | Aligning distributed AI approaches | Aligning distributed AI approaches | Aligning distributed AI approaches |
| *Open Data /Open Source* | Federated learning requires in several scenarios open data | Federated learning requires in several scenarios open data | | AI requires proper data access | | | |
| *Large Trials* | to deploy federated learning solutions in different vertical industry scenarios within different demos sites and trials | to deploy federated learning solutions in different vertical industry scenarios within different demos sites and trials | | | | | |
| *Cross-domain research* | To deploy federated learning solutions in different vertical industry scenarios within different demos sites and trials | to deploy federated learning solutions in different vertical industry scenarios within different demos sites and trials | Impact of AI based decisions on service experience and applications | | | | |

## 5.5 Automated and agile software engineering

### 5.5.1 From cloud-native to continuum-native software

Cloud-native architectures[C5-5], based on technologies such as microservices, containers, and serverless architectures, have flexible and easy-to-scale structures to maximize the exploitation of elastic cloud resources.

Today's cloud computing will evolve into a computing continuum providing a programmable computing infrastructure across devices, fog, edge, and central clouds and, by that, offering a

significantly broader range of choices how to deploy and run microservice-based applications and network services. The choices will be influenced by the profiles of the available hosting compute locations. These profiles include context information for example about hardware architecture, the capacity, available network connectivity, its geolocation, and more.

One of the main challenges will consist of distributing the components of an application across the computing continuum so that requirements of the application in terms of, for example, performance, security, or usability, will be met in an optimal way. This is also true for IoT applications as illustrated in Figure 5-1. Today's microservice architectures might need to be reviewed and further developed to support the distribution across the computing continuum in flexible ways. Software design approaches that allow to defer the decision about the software componentisation as long as possible could be one path to achieve the flexibility. Abstraction mechanisms that allow for loose coupling between applications and infrastructure and support dynamic placement optimization could be another way. In any case, the computing continuum will impact the architecture, the interfaces, and the disaggregation of future network functions.

Specific research challenges to be addressed in this context include:

- Research on addressing the end-to-end software capabilities (continuum) of technologies across sensors, connectivity, gateways, edge processing, robotics, platforms, applications, AI, and analytics, including underlying technologies like optical, wireless (cellular and non-cellular) and satellite communications.
- Research on supporting the continuum of intelligence (e.g., ability to acquire and apply knowledge using context awareness) and other edge capabilities, e.g., computing, connectivity, processing, sensing, privacy, security, see Figure 5-1.
- Research on supporting the continuum of edge applications within and across vertical sectors and seamless integration.

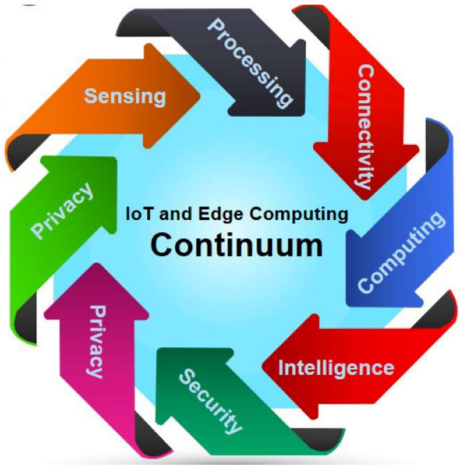



*Figure 5-1 Continuum-native software of IoT applications [C5-6]*

### 5.5.2 Low-code and no-code platforms

Low-code platforms provide an easy-to-use development environment for creating and configuring software applications. They require only little coding skills, promise high productivity in developing applications, and accelerate time to market. The idea behind low-code platforms is in line with recent developments in the area of network programming and configuration, meaning the

transformation from SDN to intent-based networking. In the future, low-code platforms will enable the “intent-driven” use of the services offered by SNS.

Innovations in modelling and code generation technologies are needed to enable the transition to low-code and eventually no-code platforms. In addition, novel technologies in the direction of human-centricity and AI will help address fundamental challenges of the low-code approach.

Modelling technology is the basis of low-code platforms. Domain-specific modelling languages have been successfully applied in many domains [C5-7], but there is still a lack of effective methodologies on how to extract the commonalities among applications in the domain and design the modelling languages or interfaces that fit the domain experts. Further improvement is still needed to allow the domain-specific integration of network services, using orchestration languages or interfaces that represents the convention and vocabulary in the target domains, rather than general purpose service definitions.

Code generation connects the high-level models of a low-code platform with the executable code, usually in general-purpose programming languages. One of the main challenges is the trustworthiness of the generation engine, meaning that users trust that the generated code represents their intents and that no vulnerabilities have been introduced during code generation. Automatic testing, as well as the analysis and inspection of test cases, is important for developing trustworthy code generation solutions, but a remaining challenge is how to make model users understand and trust the test cases. Software Language Engineering [C5-8] is a promising research area for developing methods, tools, and use cases on designing domain-specific modelling languages and developing code generators.

Human centricity is strongly relevant to low code platforms, as the ultimate objective is to provide an easy-to-use development environment for a wide range of people. When designing the low-code modelling language, it is important to understand what the target users really need, and to design the modelling language and interface based on the concepts and vocabularies that fit the target group. Research methods from the social science[C5-9] are useful to understand the real user needs. Human-centric programming methods and tools may also provide an alternative direction for low-code platform. For example, the experimental code completion tool, GitHub Copilot, can generate large body of code snippets, comments, and test cases from small hints. This significantly reduces the technical knowledge and effort for developers, without a need to design a new modelling language.

The state of the art of low code platforms is based on the assumption that the application functionality and logic is embedded in deterministic code, and the low-code language provides a higher level abstraction of such logic. With AI embedded in an application, some parts of the logic will not be any more deterministic. In these cases, low-code platform will be more like a Human-AI interface that allow users to express what the intended behaviour of the AI-based application should be.

### 5.5.3 Integrated lifecycle management: DevOps and CI/CD pipelines

DevOps culture and the use of a continuous integration and delivery pipeline [C5-10] have become common practice in the software lifecycle management, also in the telecommunication industry.

While regular IT CI/CD pipelines and workloads usually operate in a homogeneous and fixed cloud environment managed by a single organisation, telco CI/CD pipelines are split across different organisations – ie. multiple telecom software vendors and communication service providers – and operate in more complex and heterogeneous environments. In telco CI/CD pipelines development, building, and testing are in the vendor part of the pipeline, and deployment and operation in the operator part. All involved parties need to align on the delivery and feedback method across the organizations, covering for example the intervals of delivery by multiple vendors, the timing of integration of the software from multiple vendors, as well as the testing of the integrated software system before going into production. Further challenges include the adaptation of the software to the various hybrid cloud environments of telecom networks and performing software updates without violating stringent SLAs. Effective automation to enable service creation and management on large scale of multi-, hybrid- and edge-cloud systems is needed [C5-11].

In the future, CI/CD pipelines have to deliver software into the more heterogeneous and dynamic infrastructure of a computing continuum and have to provide effective support for the entire lifecycle of complex service chains. CI/CD pipelines will need to deliver not only pure software but also AI and ML components, as well as APIs. And all that in a secure way.

As a result, research is needed on the following aspects:

- Adequate abstraction mechanisms that allow to design software in a cloud infrastructure agnostic way.
- AI-based enhancements of Infrastructure as Code techniques for the automatic deployment of software components so that new deployment arrangements can be learned from historical and simulated deployment experiences.
- Automatic generation of test cases and the simulation of test environments making use for example of digital twin technology or chaos engineering.
- Extending the CI/CD tools and techniques to allow for bundling software, AI components, APIs, and security into one development, delivery, and deployment process
- Developing standards to enable interoperability in multi-vendor CI/CD scenarios

### 5.5.4 Integration of DevOps with business processes

DevOps aligns and automates the collaboration between development and operations teams. As a result, the software development process has become tightly interwoven with the operation of the software and, by that, with the business processes implemented by this software. This accelerates the introduction of new features, the adaption of business processes to changing user and market requirements as well as creates additional opportunities for the network automation to achieve efficiency gains. Network management and orchestration, network slicing, security workflows, or the exposure of network capabilities via APIs are all examples of network functionality and processes that can benefit from a close integration with a CI/CD pipeline.

Feedback loops being integral part of a CI/CD pipeline play an important role in this kind of network automation. Feedback loops provide a constant flow of information that allow to gain insights into the health of the deployed software and related business processes and might trigger automated or semi-automated responses to identified issues. There are two types of feedback loops: development loops along the entire CI/CD pipeline with humans involved, and adaptation loops that

monitor system changes and make automated adaptation in real-time. The development loops are running at a slower pace, continuously tuning the adaptation loops. More feedback loops and layers of feedback loops may be added depending on the business processes and applications running on top of the network and using for example network slices.

Research challenges in this context include:

- The investigation of use cases and business processes that can benefit from a close integration with a CI/CD pipeline focusing not only on the technical aspects, but also addressing organizational questions such as how to setup cross-organizational teams and collaboration to make the CI/CD pipeline work
- The hierarchy and interworking of the feedback loops as well as their integration into network management and orchestration processes
- The secure exchange of data via the feedback loops which might span across multiple administrative domains and organizations

### 5.5.5 Networks and data

5G, and even more 6G society will result in huge amounts of data, and the analysis and application of these massive amounts of data promote the informatization and digitization of society. The focus on data requires, then, important considerations, not only because of the shift toward a data-centric, AI-based O&M that is essential for telco operators per se, but it is quite evident that data will be more and more essential for the whole society.

Data needs to be collected, cleaned, processed and stored in the end-to-end network. Even though it is possible to process data centrally, the integration of different networks in 6G scenarios leads to a more distributed data processing logic. More and more data will be used for algorithm training purposes, and this will lead to the stringent necessity to engineer software AND data operational cycles.

### 5.5.6 Research challenges

| **Research Theme** | **Automated and agile software engineering** | | |
|---|---|---|---|
| Research subtheme | Timeline | Key outcomes | Contributions/Value |
| Infrastructure-aware microservices model (SNS-native software) | Mid-term (finished in 5y) | Extended microservices model and container technologies with explicit requirement or preference on the context of the hosting resource | Computing continuum alignment |
| Continuum-native network functions and applications (SNS-native software) | Short-term (finished in 3y) | Extending the current container and orchestration platforms to cover network functions and services | Computing continuum alignment |
| Low-code integration language and platform for SNS (low code platforms) | Short-term (finished in 3y) | User-friendly integration languages with domain-specific knowledge | Consumability of network services |

| | | | |
|---|---|---|---|
| Human-centric methods to power the design of low code platform | Mid-term (finished in 5y) | Methods and tools to understand end user needs to feed the design of low code platform | Human-centricity |
| Human-AI interface for application development (low code platform) | Mid-term (finished in 5y) | New languages and interfaces for human developers to influence, understand and trust AI in conducting business tasks | Human-centricity |
| Auto testing on simulated SNS (DevOps and lifecycle management) | Mid-term (finished in 5y) | Digital twins of the whole SNS environment and the surrounding physical context to automate the execution of test cases | Computing continuum alignment |
| Abstraction and AI-powered deployment models (DevOps and lifecycle management) | Short-term (finished in 3y) | Cloud agnostic software and automated deployment models for the computing continuum | Computing continuum alignment |
| Common CI/CD engine for SW, AI, APIs, security, etc | Mid-term (finished in 5y) | Concepts and tools for extended CI/CD engines | Computing continuum alignment |
| CI/CD automation use cases incl. organizational issues (Integration with business process) | Mid-term (finished in 5y) | CI/CD enabled network automation and organizational optimizations | Computing continuum alignment |
| Hierarchical DevOps loops (Integration with business process) | Mid-term (finished in 5y) | Novel DevOps practices where many different processes with different time scales and focuses are coordinated | Computing continuum alignment |
| Aligned network software and data lifecycle | Short-term (finished in 3y) | Lifecycle management approaches keeping software and data consistent across networks | Increasing dependency on AI and data |

### 5.5.7 Recommendations

| **Research Theme** | **Automated and agile software engineering** | | | | |
|---|---|---|---|---|---|
| Action | SNS-native software | Low-code platforms | DevOps and lifecycle management | Integration with business process | Networks and data |
| *International Collaboration* | Fundamental research to investigate the innovative software paradigm | Research and innovation towards new languages and new ways for developer-SNS interaction | | | |

| | | | | | |
|---|---|---|---|---|---|
| *Open Data / Open Source* | | Open data to better understand how different stakeholders interact with the SNS | | | Keeping data aligned with network software |
| *Large Trials* | | | Large trials to test and promote the adoption of DevOps in SNS | | Keeping data aligned with network software at scale |
| *Cross-domain research* | | | | Joint research with other domains to investigate the integration with business processes | |

## 5.6 Enablement of digital services

### 5.6.1 Time guarantees on virtualization and containerization

In an edge without time guarantees, virtualization and containerization have no constraints, and share the host resources without limitations. When addressing time sensitive/time engineering applications, such differentiation is needed, and applications should run in virtualized environments capable of having different priorities between services and with pre-emption techniques. Such techniques will enable a deterministic network and edge, capable of processing critical requests within specific time windows (on-time processing). Recent research [C5-12] has been conducted to understand the extent to which time-sensitive payloads can be virtualized. These mostly encompass containerized execution environments resorting to specific Linux kernel fine-tuning or co-kernels to improve determinism. Lastly, the devices deployed in this environment support distinct hardware architectures, in which the orchestrator should be aware. The integration of fine-tuned execution environments with suitable lightweight edge orchestration frameworks, are envisioned as a possible way to tackle the challenge of building a time awareness orchestration framework, suited for constrained edge computing devices. Specific orchestration scheduler is also needed to support the time-sensitive applications.

### 5.6.2 Network compute fabric supporting passive IoT

Today's computing continuums are advancing away from central data centres to the very edge of the network. It enables time-critical and massive-data hungry use cases. As the use cases continue to grow in complexity and criticality, new forms of integrated computing fabric will inevitably be needed to complement today's edge solutions and bring the rest of the computing continuum closer together.

Future deterministic networks that ensure bounded latency, jitter, and packet-loss and out-of-order packet delivery are complex where next-hop forwarding requires service protection along an end-to-end resource reserved explicit routes that can be achieved basically through packet replication, duplicate elimination and/or network coding. Some topics have to be considered:

- Sequencing Information expressed in terms of sequence number or time stamp
- AI/ML-powered software-defined regenerative payload creation
- Enhancement on Packet Replication Function (PRF), Packet Elimination Function (PEF) and Packet Ordering Function (POF) in addition to Network Coding Function (NCF)

In addition, a large-scale deployment of sensors often requires cost-effective solutions fuelling the introduction of passive IoT devices. These devices are different that they do not possess batteries. A minimal impact that such a device has on the mobile communication system is that the network needs to identify those UEs that are of passive IoT type. Following issues are important to consider:

- New UE type and such information has to be included as part of UE Subscription data
- Geographical location a network can communicate with a given passive IoT device depending on renewable source availability
- New Location Management Function (LMF) feature
- Making transparent the network compute fabric to users/developers who want to use the heterogeneous edge resource embedded in the network to better support advanced analytics and coordination mechanism

### 5.6.3 Use case and ecosystems driven service development

Research towards 6G is mainly sustained by technology enablers and, particularly, technology enablers of use cases and the so-called verticals. In particular, the current trend is that only the application should matter, keeping the technology as a secondary asset.

This can be observed, for instance, in the Open Networking Foundation (ONF) recent activities, which have shifted from focus on standardization, recommendations, and development of platforms (like the well-known SDN framework, ONOS) towards a considerable emphasis on what they defined as Reference Designs (RDs) [C5-13]. RDs represent a particular assembly of components (acting as “blueprints”) required to build and address specific use cases. For instance, Converged Multi-Access and Core (COMAC) is an RD, whose architecture involves, at the same time, components from use case-based horizontal projects like SEBA or CORD, also promoted by the ONF. Another example is the recently created one6G association [C5-14], defined as an “innovation hub” and currently working on different verticals.

In this regard, it can be noticed that ecosystems flourishing from those use cases mainly rely around these open source communities and organizations, many of them funded by big telcos and manufacturers, together with other research entities and universities. Therefore, an interesting initiative would be to foster additional collaborations, from academia and industry, even considering different levels of expertise (as they would help as well as brainstorming to gather different perspectives and ideas) and merging effort towards a common ecosystem.

Additionally, not only the definition of use cases and components should be bolstered, but also other ecosystem towards the building of digital services like data ecosystems for training AI/ML algorithms and testbeds. One example of this is the Softwarised Network Data Zoo (SND Zoo) [C5-15], but more effort is required in this field.

### 5.6.4 6G enabling sustainability in vertical industries

6G is also aiming, by digitalisation, to provide significant support in achieving the EU Green Deal and the United Nations Sustainable Development Goals (SDGs). That means that not only 6G itself should be sustainable, but 6G should also enable other industry sectors to achieve the EU Green Deal objectives and the SDGs. In this context 6G is supposed to enable applications for use cases that address aspects such as resource-efficiency, safety, and inclusiveness.

Green Digital Transformation or Green Digitalization is the adoption of digital technology to transform services or businesses, through replacing non-digital or manual processes with digital processes or replacing older digital technology with newer ones, aiming for sustainable production. An example of such transformation is Industry 4.0 (I4.0) that refers to a global method in industrial production, based on digitalization and automatization of the entire production process. The method is based on efficient distribution of data and processes across industry, turning passive physical objects into comprehensive systems of interconnected and interdependent elements. I4.0 defines a new and highly productive value chain, utilising smart services to provide data-driven connectivity between suppliers, distributors, customers and production lines, which in turn makes the manufacturing process more efficient, cost-effective, flexible and sustainable. Such an enhanced process is capable of producing smart products with higher quality, lower price, shorter lead time and minimum carbon footprint.

The above mentioned items will expose changes on the way the 6G network has to be designed and delivered. The main goal is to foster adoption of private industrial networks as a major solution for vertical sectors. Enhanced wireless 6G connectivity along with edge cloud continuum are key enablers to finally realise the active information continuum in our future digital society and economy, a concept that has been introduced for example in the framework of I4.0 but never been deployed for real due to the lack of technology readiness. Active information continuum not only enables simple data exchange among different parts of a processing chain but also provides distributed intelligence on all sides. It enables key use cases such as safety in human robot interaction, remote quality of production control, AR/VR use for predictive maintenance of elements, on demand and customised production, configurable and flexible production line, etc. It adds to the automation, reduces the cost and CO2 footprint, improves the energy efficiency and provides enough means to introduce new concepts such as virtual factories. We have already identified some key needs that could be good first steps to prepare 6G networks according to the vertical market needs:

- Ensuring adequate consideration of special requirements of 6G-enabled cyber physical system of systems, such as time sensitive and deterministic networking, real time operation, reliability, etc.;
- A novel zero-touch network management model that simplifies the operation of the private networks;
- Support of Asset Administration Shell (AAS) and OPC Unified Architecture (OPC UA) to seamlessly integrate the 6G mobile communication system within the vertical domain assets.

Carbon footprint tracing of 6G communication infrastructure based on updated AAS (in line with the EU digital product passport promoted by Orgalim [C5-16]) to make production green. The network itself will be treated as an asset in the near future being described and managed by AAS. The digital

twin of user ends and network gears will help to forecast potential network loads in given industrial applications and to negotiate the quality of service parameters needed in the actual application, i.e. latency and reliability of 6G connection. We should target extending and demonstrate 6G AAS structure & sub-models, modelling and building digital twins that are seamlessly integrated with their real-time intelligence, hence allowing "the first green production" by including carbon footprint of communication infrastructures, physical assets and services, encompassing the whole value chain of a networked vertical scenario.

Additionally, focusing on the heterogeneity of 6G networks, enhanced programmability and network softwarization can foster the optimization of IoT, low-power and constrained devices computing capabilities. In this regard, these types of devices could benefit from the installation or integration of light software plugins to build federated intelligence systems, so that groups of devices could make local decisions and compute services on their own and, hence, overall communications would be reduced, saving energy as a consequence. This is particularly challenging on battery-less devices (e.g., powered by solar panels), which should delegate computing capabilities based on different parameters (including weather).

Finally, 6G softwarization could also enable alternative sustainable paradigms in vertical industries like, for example, in agri-food environments. Substituting generalized chemical fertilizers, insecticides, etc., by fine-grained intelligent sensor control could bolster sustainability. Furthermore, many tasks in agri-food are still performed manually or by visual inspection, and federated networks of sensors could increase control and reduce the need for transport to proceed with these manual checks, particularly in large extensive fields.

### 5.6.5 Research challenges

| **Research Theme** | **Enablement of services and business models** | | |
|---|---|---|---|
| Research subtheme | Timeline | Key outcomes | Contributions/Value |
| 1. Time guarantees for containers | Short-term (finished in 3y) | Time aware orchestration frameworks for time-sensitive applications | Service and business enablement |
| 2. Abstraction mechanisms for the network compute fabric to support passive IoT | Mid-term (finished in 5y) | Programmatic framework for accessing network and compute resources for the use case "passive IoT" | Enabling energy efficiency and sustainability |
| 3. Software engineering for computation offloading | Short-term (finished in 3y) | SW engineering and QoS frameworks for computationally intensive applications | Service and business enablement |
| 4. Use case driven service research | Short-term (finished in 3y) | New services ideas + requirements on SNS services | Service and business enablement |
| 5. SW based features of private networks to support sustainability in verticals | Mid-term (finished in 5y) | Seamless integration of 6G into vertical domain assets | Enabling energy efficiency and sustainability |
| 6. Optimize energy consumption across end devices | Mid-term (finished in 5y) | Federated intelligence systems for devices | Enabling energy efficiency and sustainability |

### 5.6.6 Recommendations

| Theme | Enablement of services and business models | | | | | |
|---|---|---|---|---|---|---|
| Action | Subtheme 1 | Subtheme 2 | Subtheme 3 | Subtheme 4 | Subtheme 5 | Subtheme 6 |
| *International Collaboration* | Common standardized solutions | Common standardized solutions | Common standardized solutions | | | Common standardized solutions |
| *Open Data / Open Source* | Open Source implementations | Open Source implementations | Open Source implementations | | | |
| *Large Trials* | PoCs | | PoCs | | PoCs | PoCs |
| *Cross-domain research* | | | | Developing service ideas | Domain knowledge of verticals needed | |

## 5.7 Engineering complex, software-intensive, and self-adaptive systems

### 5.7.1 Managing the software complexity of a system of systems

As described in [C5-17] smart networks will be a highly distributed and decentralised system of systems comprising countless heterogeneous physical and virtual entities and supporting a broad range of services and applications with divergent requirements. All mapping from services to network slices and then to virtual resources will be completely elastic and flexible. There will be no direct relationship between the lifecycle of a service and the lifecycle of the assigned virtual resources. The countless entities need to be managed throughout their specific lifecycle and to have their parameters configured and adapted to a dynamically changing environment. The services have to be of high quality and provided in flexible ways to meet users' demanding expectations, whilst consuming network resources as efficiently as possible to minimize cost. Managing the resulting system complexity will become increasingly challenging and will require new operational concepts based on sophisticated self-adaptation models and relying heavily on AI algorithms.

AI is used to evaluate the current resource status and current service status, and more importantly, to predict any future problems. AI is used to decide how to react to current or predicted certain level of future status changes. AI is also used to optimise the delivered service to avoid unprofitable waste of resources. Although closed loop automation, with limited human intervention brings all the promise of a fully automated business, one of the main ongoing debates is to consider a role for humans in the loop. The issue arises from the concerns about trusting a fully automated system. Humans in the loop can make or break AI success and have to be addressed in a human-centric way.

Developing, debugging and testing such complex cognitive systems will be challenging. The conflicting requirements of extreme flexibility, dynamic adaptation and optimized resource utilization are hard to reconcile in a distributed system of autonomous AI-based subsystems, and the resulting overall system behaviour might be unexpected. To avoid unexpected and even

hazardous effects, what is needed is predictable governance for self-adapting AI-based software systems including the traceability of AI.

This implies that AI/ML algorithms have to be made aware of changes in the surrounding system that impact the learned model, and would require re-training. It also involves explainable AI, for transparency of how an AI-based system works and responsibility for the resulting output. Access to high-quality, trustable and sufficient training data needs to be assured, so that no undetected biases find their way into AI systems.

Techniques such as deep neural networks may achieve a high performance in terms of speed and accuracy, but they are generally seen as black-box models due to lack of explanation associated with their outputs. In the context of SLA/Service Level Specification (SLS) management and network orchestration, the features of Explainable AI (XAI) become very desirable in terms of transparency of the model with respect to the intended outcome. Current research contemplates two main approaches aiming at achieving explainable AI models: the first one involves designing models that are inherently interpretable (which means they can easily explain how specific decisions can influence achieving specific objectives); a second approach is to complement the AI black box with the aid of external complementary models. Different levels of model transparency (or explainability) may apply to different categories of AI applications. As a first level, models can explain how conclusions are reached by the AI system in order to improve future decision making, decision understanding and trust from human users and operators. As a second level the external model can allow inspection and traceability of actions undertaken by the AI systems. Traceability will enable humans to get into AI process loops. Depending on the use cases and market/feasibility priorities, pros and cons of various characteristics of human plus AI has to be considered and investigated, i.e. assisted intelligence (improving decisions and actions of people); augmented intelligence (enabling humans to do more than before); autonomous intelligence (adaptation of various situations, acting without human assistance).

### 5.7.2 Engineering software intensive systems

The discussions in [C5-17] identify the need to expand the scope of software engineering to encompass the full range of possible deployments from embedded devices to the cloud, and the full lifecycle of the software including automated operation of self-adapting software intensive systems. This unification of operational and business aspects is not supported by adequate software tools today. Software engineering methods such as UML modelling cannot handle situations where interconnected services are not known in advance, and cannot easily model consequences that may have a legal or ethical dimension. Some aspects previously considered the domain of programmers such as the composition of resources and services will also need to be handled autonomously at run-time. Therefore, we need new engineering approaches which can be applied over the lifecycle of software services and data (including design, implementation and testing); respond to agile changes in self-adapting systems; handle ethical and legal aspects; and support purposeful sharing.

The use of AI presents both challenges and opportunities for software engineers. As previously noted, operating software on a large-scale, distributed, heterogeneous and smart infrastructure requires new approaches such as AI. Can AI also be used to support the evolution of DevOps

methods for software design and development? How will those methods enable the design and development of smart components that use AI, and ensure that those components meet ethical, legal, social and economic requirements as they evolve in the presence of new input data? The approaches developed must be able to handle requirements and constraints not only in the use of AI, but also of other novel technologies such as specialised (including quantum) processing devices, and novel modes of human-computer interaction.

### 5.7.3 Research challenges

| **Research Theme** | **Engineering complex, software-intensive, and self-adaptive systems** | | |
|---|---|---|---|
| Research Challenge | Timeline | Key outcomes | Contributions/Value |
| 1. Testing of self-adaptive systems | Mid-term (finished in 5y) | Testing approaches and frameworks for self-adaptive systems | System resilience |
| 2. Predictive governance for self-adapting AI-based software systems | Mid-term (finished in 5y) | Governance framework for monitoring behaviour of AI-based systems | System resilience |
| 3. New SW engineering approaches | Mid-term (finished in 5y) | Overall SW architecture and design approaches for complex systems | Managing the complexity |
| 4. AI-assisted software design | Mid-term (finished in 5y) | AI-based approaches and tools for the design of software intensive systems | Managing the complexity |

### 5.7.4 Recommendations

| **Research Theme** | **Engineering complex, software-intensive, and self-adaptive systems** | | | |
|---|---|---|---|---|
| Action | Subtheme 1 | Subtheme 2 | Subtheme 3 | Subtheme 4 |
| *Cross-domain research* | SW engineering applied to the telco domain | SW engineering applied to the telco domain | SW engineering applied to the telco domain | SW engineering applied to the telco domain |

## 5.8 SW architectures

### 5.8.1 Edge and embedded computing

Location-based applications or applications for monitoring the environment might comprise tasks with large computational requirements such as pattern recognition. Sensors and field computation units involved in this kind of applications often do not have the compute resources to carry out computationally intensive and delay-sensitive tasks. Edge computing offers the capability to offload those workloads from the end device to an edge node to reduce latency and bandwidth bottlenecks.

These computational units or edge nodes have been specified by 3GPP to be directly integrated in the 5G architecture, in order to improve reliability and round-trip latency. This integration is defined in Technical Specification (TS) 23.501 and is expected to be developed towards the next generation of cellular communications (6G). Examples of industrial applications making use of this

architecture are described in [C5-18] and [C5-19] showing that satisfactory results in terms of latency can be achieved.

In order to ensure QoS using edge computing, the use of offloading computation algorithms must be integrated into Beyond 5G and 6G communications. This integration will provide better support for the use of edge computing in mobile scenarios such as drone or autonomous vehicle-based use cases. In [C5-20], several algorithms are studied to manage the offloading in terms of computation cost and energy savings. Research challenges are about application software architectures that support the offloading of subtasks and the interworking with AI based mechanisms that help to decide about when and where a task should be offloaded and considering the impact that the offloading will have on the overall end-to-end QoS.

### 5.8.2 Integration of Quantum computing

The paper [C5-21] describes the need and the research challenges to integrate quantum computing with classical computing.

Quantum computers operate on qubits which can represent a combination of both zero and one at the same time by exploiting the quantum phenomenon of superposition. Combining qubits into a larger system enables quantum computers to perform multiple calculations with multiple inputs simultaneously. This allows to achieve an enormous - in some cases exponential - speed-up in executing certain algorithms for solving specific problems. For example, multivariable problems could be solved in a significantly more efficient way in quantum computers than in classical computers, in applications such as complex optimization in network planning, large system simulations, and quantum machine learning.

Quantum computers and classical computers will coexist, and hybrid algorithms utilize both. Both quantum computing and classical computing resources will be accessible from the computing continuum and should be usable in an integrated way, similar to the way today's High Performance Computing (HPC) is used. However, the full software lifecycle and software stack for quantum computers are different to those in classical computing. At the lowest machine level, qubits and their interfacing need to be controlled and optimized. At the next level up the software stack, it is about the coding of the algorithm by quantum circuits combing multiple qubits. On top of the stack, the algorithm execution is integrated with other parts of the application software.

Research challenges in this context include:

- The problem categories that are suitable for quantum computers in the telecom domain need to be better understood. Increased interdisciplinary research - including quantum information, computer science, and especially mathematics – is needed to develop domain- and problem-specific quantum algorithms, and to improve the efficiency of existing ones.
- Research into the programming abstractions (e.g. abstract machines, compilers, libraries, programming languages, APIs, etc.) that facilitate the design of hardware-independent quantum programs which nonetheless benefit from the specific features of the underlying quantum hardware.
- More research emphasis on the integration and orchestration of classical and quantum computing. This includes the interplay at the level of algorithms and the lower level of pre- and post-processing of a quantum computing. It also should cover the access, orchestration

and integration of quantum computing resources that are available as a service within a computing continuum.

- Benchmarking, testing and debugging of quantum programs is still at a very early stage and needs to be explored at a fundamental level.

### 5.8.3 Research challenges

| **Research Theme** | **SW architectures** | | |
|---|---|---|---|
| Research Challenge | Timeline | Key outcomes | Contributions/Value |
| 1. Exploring offloading of computationally intensive and delay-sensitive workloads | Short-term (finished in 3y) | SW architectures and mechanism for task offloading | Alignment with computing continuum |
| 2. Exploring telecom specific problems suitable for quantum computers | Mid-term (finished in 5y) | Quantum algorithms for complex problems in the telecom domain | Alignment with computing continuum |
| 3. Quantum software engineering | Long-term (finished in 7y+) | Implementation and integration of quantum algorithms in the telecom domain | Alignment with computing continuum |

### 5.8.4 Recommendations

| **Research Theme** | **SW architectures** | | |
|---|---|---|---|
| Action | Subtheme 1 | Subtheme 2 | Subtheme 3 |
| *International Collaboration* | Investigating common interfaces and standards for offloading | | |
| *Open Data / Open Source* | | | SDKs for implementing quantum algorithms |
| *Large Trials* | PoCs to test task offloading approaches | | PoC of implemented quantum algorithms |
| *Cross-domain research* | | Developing Quantum algorithms requires interdisciplinary research | |

## 5.9 Human centricity and digital trust

### 5.9.1 Data authenticity and trusted digital interactions in dynamically composed service environments

The discussion in[C5-16] points out that the impact of SNS will be limited if risk-averse users refuse to share data fearing it may be misused or reject advanced applications because they feel manipulated. Without trust, services such as health care or applications such as online elections may not be viable.

Therefore, ways to verify data authenticity and truthfulness will be needed, along with trusted digital interactions, especially in dynamically composed service environments. Trusted identities

and authentication services for software and devices as well as humans are essential, along with access control mechanisms that users can understand, to manage their data and protect themselves from manipulation. Telco operators will have the opportunity to play the role of an identity provider and manager and, thus, controlling access to a world of applications.

Smart contracts and distributed ledgers, trusted hardware, and homomorphic encryption may provide links in the chain of trust, but the key is to use measures to support a holistic network of trust between stakeholders. Fact-checking services based on AI may play a role, as may services that govern the AI to ensure that novel technologies and applications remain compatible with societal needs such as the right of individuals to freedom of expression. Authenticity must be demonstrable, not just for data but also for the consequences of using data in AI-enabled decision-making algorithms. Technologies alone will not be enough – they must be deployed in a citizen-centric fashion, giving humans control over their interactions.

To achieve high levels of trust, software engineering methodologies and tools must provide the trust anchors needed by stakeholders: software developers, service operators, business customers and consumers, regulators and certification agents. Certification of products and services will be important, and will play an essential role in regulation to ensure security and safety in sectors such as medical IoT. As certification procedures are expensive, software engineering methods will be needed to implement them for dynamically changing systems in a cost-effective way, e.g. by focusing on critical sub-systems or operational contexts. In some areas, new procedures and standards (similar to ISO 26262) will be needed, e.g. to certify software based on machine learning/AI, for which there are no established methods today.

### 5.9.2 Human-centric software engineering and codes of ethics for software development

Future networks will enable a rich environment for multi-user interaction and will support “assistive” technologies such as tactile gloves and devices offering gesture recognition and haptic feedback. These devices and the related services will deliver information that is relevant to users’ tasks at hand. The interactions will be mediated by algorithms in many cases. The overall system will be a complex constellation of software and content components operated by a multitude of ecosystem participants.

Designing, developing, deploying and maintaining these software-based systems is technically very challenging and many issues remain to be solved [C5-22]. For the requirement engineering, it is essential to understand for example how and who could take control in these spaces, how the information will flow and whether it is necessary to settle protection mechanisms against dominant positions, leading to serious threats to human integrity. The human-centric and ethical issues are often overlooked in the building of such systems.

### 5.9.3 Research challenges

| Research Theme | Human centricity and digital trust | | |
|---|---|---|---|
| Research subtheme | Timeline | Key outcomes | Contributions/Value |
| 1. Collaborative research work required in the context of identity and privacy in open ecosystems | Short-term (finished in 3y) The research activity should progress together with architectural achievements in 6G | Trusted identities and data authentication services offered by SNS | Human-centricity of SNS |
| 2. Privacy and responsible software development | Mid-term (finished in 5y) | 6G networks will enable new generation disruptive services. The activity should produce a set of requirements and tools for helping software developers and service providers to embed human-centric aspects. | Human-centricity of SNS |

### 5.9.4 Recommendations

| Research Theme | Human centricity and digital trust | |
|---|---|---|
| Action | Subtheme 1 | Subtheme 2 |
| *Cross-domain research* | Multi- and interdisciplinary research to address human-centricity | Multi- and interdisciplinary research to address human-centricity |

## 5.10 Digital twins

### 5.10.1 Software engineering of telco digital twins

Digital twins are dynamic virtual representations of entities such as assets, persons, and processes. Digital twins are built by developing models for interpreting data at different speeds to be used for creating a one-to-one association with their real-world twins. Digital twins do not exist independently from an enabling software platform. It is thus of paramount importance to investigate software engineering practices and tools for simplifying the development of telco-world twins.

It is also worth to note that, in a broad view there are two main categories of digital twins, offline and real time. On one hand, for the real-time variation, it mimics the behaviour of existing physical products almost simultaneously. From 6G system perspective, ultra-low latency, high availability of connectivity and computational edge resources, high reliability, trustworthiness, and high security are some key requirements to be considered. On the other hand, digital twins that are more computationally intensive and thus cannot keep up in real-time, can also be run after the sensor data has been captured in an offline manner over a public/private cloud. In this case, there might not be extremely hard requirements on 6G systems. What mentioned so far are from the perspective of 6G systems to support digital twin of vertical sectors, e.g. digital twin of an engine or a road. However, the network itself as a physical entity should have its own digital twin too. Defining and realization of a proper and harmonized model with the current available solutions in the vertical sectors for the network gears, e.g. UE, RAN and Core, is an important research direction to follow.

At the same time, the twin's software development life cycle should be restructured for managing requirements, including sustainability requirements (cf., for example, Twin Transition [C5-23]), models, data & metadata, with special attention to the validation and verification procedures and improvement cycles comprising operational feedback. Moreover, new, and more sophisticated software engineering techniques are aimed at taming the complexities of higher-level twins. Considering the very wide range of competencies needed for developing new twins and for making them a more affordable investment for a larger number of companies, more effort on "composable" digital twins' development tools and techniques is needed. Twins of telco networks may begin at the network design stage, facilitate deployment of network nodes, optimise the deployment of edge nodes, and provide integration of information and decision support to enable highly automated and remote operation while ensuring high efficiency, safety, and environmental awareness.

Such digital twins are complex since they must integrate a wide range of information, algorithms, and models for processing real-time operational data as well as a large amount of business-relevant data. They might include simulation and predictive capabilities to support improved operational decision making, possibly in distributed and decentralised contexts, crossing ownership domains. Smart Network Twins could be engineered not only for optimisation and planning purposes but also, for example, for reacting to cyber-attacks, providing an immediate forecast of the consequences, and allowing efficient mitigation of them.

### 5.10.2 Research challenges

| **Research Theme** | **Digital twins** | | |
|---|---|---|---|
| Research Challenge | Timeline | Key outcomes | Contributions/Value |
| 1. Managing the life cycle of digital twins | Mid-term (finished in 5y) | SW engineering approaches for telecom digital twins | Optimisation of network planning and operation |
| 2. Composition and interworking of digital twins | Mid-term (finished in 5y) | Standard interfaces for digital twins | Optimisation of network planning and operation |

### 5.10.3 Recommendations

| **Research Theme** | **Digital twins** | |
|---|---|---|
| Action | Challenge 1 | Challenge 2 |
| *International Collaboration* | | To develop standardized interfaces for digital twins |
| *Open Data / Open Source* | Open source implementations of digital twin frameworks | |
| *Large Trials* | | PoCs for testing and demonstrating the interworking of digital twins |
| *Cross-domain research* | SW engineering for managing the life cycle of digital twins | |

# 6. Radio Technology and Signal Processing

*Editor: Wen Xu*

This chapter aims to address the enabling technologies for the next generation radio interface, including

1) Spectrum reutilization, interference management, subnetworks and wireless edge caching;
2) Optical wireless communications (OWC);
3) Millimeter-wave and terahertz communication;
4) Massive and ultra-massive MIMO including intelligent reflecting surfaces (IRSs) and cell-free massive MIMO;
5) Waveform, non-orthogonal multiple access and full-duplex;
6) Enhanced modulation and coding;
7) Integrated sensing and communication;
8) Grant-free random access for massive connections;
9) Machine learning empowered physical layer.

## 6.1 Vision and Requirements

Each generation of wireless communication has provided new services. 5G is no exception, and in addition to the enhanced mobile broadband (eMBB), the new services ultra-reliable low latency communication (URLLC) and massive machine-type communication (mMTC) were introduced.

During the coming decade digitalization will evolve even further, since information and communications technologies (ICT) have the potential to address many of the societal challenges ahead. Thus, our vision for 6G is to include societal needs in addition to the traditional technical requirements for advanced digital services to humans and machines. The target year for 6G deployment is well aligned with the target year for UNs seventeen sustainable development goals (SDGs), both being 2030. Thus, the research should from the start focus on meeting requirements on a Sustainable 6G. 6G should also properly capture the societal challenges as defined by the SDGs, targeting 6G for Sustainability services. To this end, the 6G system should be assessed using new key value indicators (KVIs) [C6-1] on e.g., digital inclusion, trustworthiness and sustainability, in addition to the traditional service oriented key performance indicators (KPs), such as data rates, low latency and highly accurate positioning. In addition, the traditional KPIs driving the design of 6G should also be revisited to capture holistic requirements on CO2 footprint and material flows.

6G should go beyond connectivity and become a trusted platform including communication, computation and storage, to provide new capabilities as intelligence at the edge and in the cloud, joint communication and sensing, and highly accurate positioning. This 6G platform will support new applications, e.g., cyber-physical systems (CPS) utilizing the ubiquitous connectivity, sensing and contextual awareness; and extended reality (XR) utilizing high data rates, low latency, precise positioning and sensing.

Technical 6G requirements include Tbps data throughput, sub-ms latency, extremely high reliability, everywhere mMTC, extreme energy efficiency, very high security, cm-level accuracy radio localization, and global coverage though integration of non-terrestrial networks [C6-2][C6-3]. These

new capabilities and requirements will require continued research and development of radio technologies and signal processing and protocols. A natural way forward to deal with these challenges is to consider electromagnetic spectrum at higher frequencies such as the sub-THz or THz spectrum, as well as infrared and visible light spectrum, since these frequencies offer wider bandwidths for higher data rates and higher resolution localization and sensing. Besides, the centimeter and millimeter wave spectrum currently utilized for 5G and other legacy wireless systems need to be re-farmed and more efficiently reused. To that end, interference management and co-existence issues should also be carefully addressed. In addition to further enhancing the widely used technologies (such as waveform, modulation and coding, non-orthogonal multiple access, full-duplex, massive MIMO, etc) to approach the theoretic limits, e.g. in terms of spectral and energy efficiency, further research is needed in several other domains, such as modelling, designing and optimizing the use of intelligent reflecting surfaces (IRSs) for communications, localization and sensing. Further research on integration of high precision localization and sensing in mobile communications networks, as well as exploring the potential for joint sensing and communications using the same hardware and software systems. Additional areas that deserves attention are grant-free random access for massive connections, and wireless edge caching and computing. Moreover, machine learning (ML) and artificial intelligence (AI) as a tool has been successfully applied in many applications. For the application in communication technologies and radio interface design, further research is still needed on all layers. Meanwhile, distributed learning and inference over the wireless links will be a common norm in 6G networks, where the radio transmission technologies would deserve another look to minimize overhead in both spectrum and energy, as well as to optimize the joint communication and AI performance with the consideration of the specific learning and inference features.

Compared with 5G [C6-4], 6G air interface may need to fulfill more stringent KPIs and requirements, such as

- **Energy efficiency (bit/Joule):** It is the capability to minimize the energy consumption in relation to the traffic capacity provided.
- **Spectral efficiency (bit/s/Hz):** This is a metric widely used in Shannon information and coding theory for optimizing a communication system or its building blocks. For example, the 5th percentile user spectral efficiency is the 5% point of the CDF of the normalized user throughput. Note that in the case of very high frequency scenario (e.g. sub-THz), spectral efficiency may not be the most important design metric. User experienced data rate, defined as the 5% point of the CDF of the user throughput, can be used instead.
- **Peak date rate:** Maximum achievable data rate or throughput under ideal conditions per user/devices (in Tbit/s).
- **Area traffic capacity (bit/s/km$^2$):** The total traffic throughput served per geographic area. This becomes more and more important for uplink when sensing and distributed AI become common usage scenarios in 6G.
- **Coverage:** This is usually the 3D global coverage and full connectivity by terrestrial and non-terrestrial coverage.
- **Reliability:** The success probability (e.g. 1-10$^{-7}$) of transmitting packet of different size within the maximum allowed latency at a certain channel quality.

- **Mobility:** Maximum speed at which a defined QoS and seamless transfer between radio nodes can be achieved (in km/h), where the nodes may belong to different layers and/or radio access technologies (multi-layer-RAT).
- **Air interface latency (user plane):** The contribution by the radio network to the time from when the source sends a packet to when the destination receives it (in ms).
- **Connection density:** Total number of connected and/or accessible devices fulfilling a specific quality of service (QoS) per unit area. Note here in 6G, the QoS could have broader sense than the traditional communication QoS, e.g. it could be sensing resolution or AI accuracy and efficiency.
- **Positioning accuracy:** It is the difference between the calculated horizontal/vertical position and the actual horizontal/vertical position of a terminal.
- **Privacy:** Capability of protecting the radio access from eavesdropping.

Besides quantitatively measurable KPIs, such as **spectral efficiency** in terms of bits per second per Hz, **energy efficiency** in terms of bits per Joule, **reliability** in terms of success rate, etc., there are some soft KPIs, such as **intelligent level** of a network in terms of the network intelligence from the perspectives of action implementation, data collection, analysis, decision and demand mapping according to the degree of manual participation in network operation, **coverage** which strongly depends on the number of deployed base stations, **controllable radio environment** in terms of dynamically changing the characteristics of radio propagation environment and creating favorable channel conditions to support higher data rate communication and improving the coverage, **privacy**, etc.

0 shows the connections between different enabling technologies and different 6G requirements and KPIs.

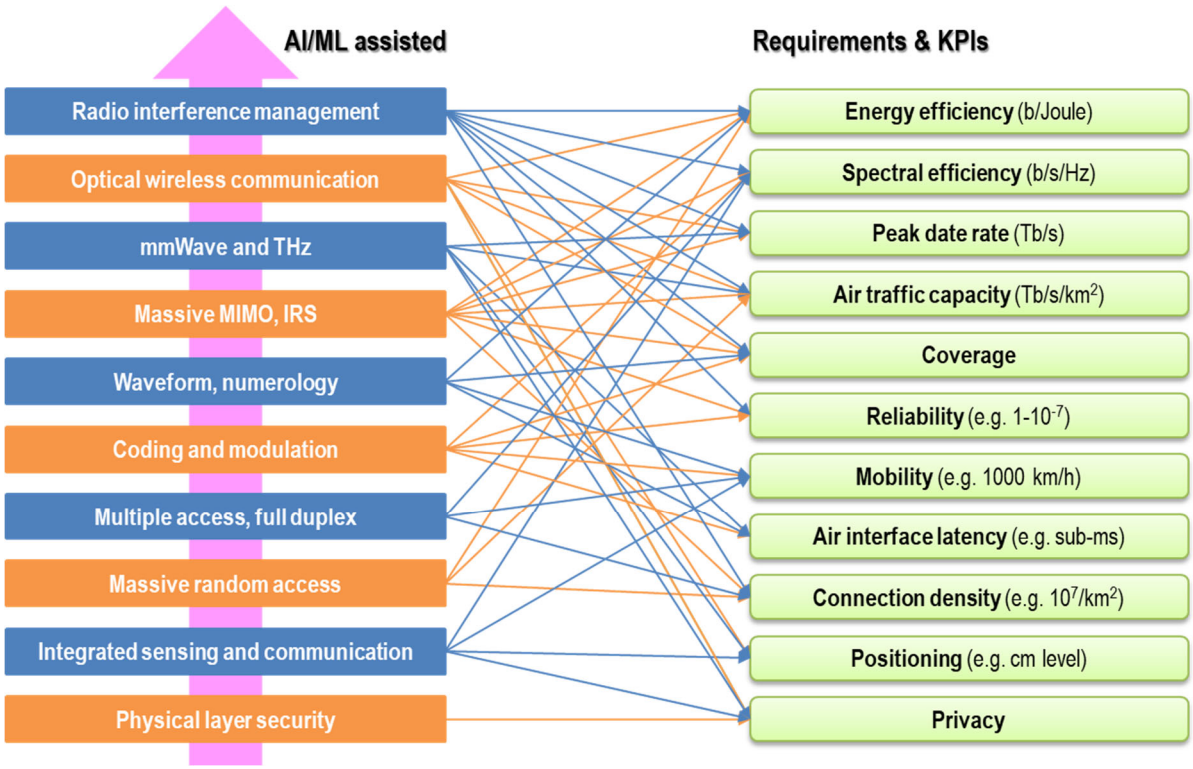



Figure 6-1 Enabling technologies with main contributions to different 6G requirements and KPIs.

## 6.2 Radio Interference Management

### 6.2.1 Spectrum re-farming and sharing

Allocated frequency spectrum is one of the main factors that determines the system capacity. However, radio spectrum is a very scarce resource with the distinctive feature that the lower frequency bands are especially precious and tightly regulated. In order to satisfy the high bandwidth demands of upcoming generations of mobile systems, it is crucial to reutilize the existing spectrum resources and optimize the access to new frequency bands. While the traditional approach allocates a dedicated spectrum to each radio access technology (RAT), spectrum reutilization between RATs and other frequency bands offers a more efficient and flexible utilization of resources, e.g., for load-balancing. Spectrum reutilization, also known as spectrum sharing, can be applied to harmonize the joint utilization of both licensed and unlicensed bands.

A straightforward approach to spectrum reutilization is *spectrum re-farming* which performs, e.g. static assignment of spectrum resources to different RATs. Note that a static nature usually leads to a poor spectrum utilization.

*On a more finely resolved time scale*, efficient spectrum utilization is achieved by dynamic inter-band resource allocation and scheduling with optimized multi-RAT handover and interference coordination. This has traditionally been based on a centralized radio resource management. However, the associated high signaling overhead motivates the exploration of decentralized strategies. Spectrum sharing is supported by multi-RAT connectivity, which allows the UE (user equipment) to choose the best RAT depending on the link qualities. This added diversity does not only increase the performance due to better spectrum utilization, it also makes the network more robust and resilient towards shadowing effects, thus improving reliability and availability. Besides, opportunistic spectrum access opens the door to further improvements by means of multiuser diversity, making it possible for UEs to access spectrum channels, e.g. on a CSMA-like basis.

A key point in the path to 6G networks is the autonomy from human intervention regarding network configuration, implying network's self-organization and management (SOM) mechanisms that smartly consider the characteristics of the environment. To achieve this in the context of joint utilization of licensed and unlicensed spectrum, adaptive and dynamic spectrum sharing strategies are required. In this line, cognitive environment concepts, in which spectrum awareness, e.g. based on a combination of advanced SIGINT (signal intelligence) and AI (artificial intelligence) techniques, can be employed to ensure co-existence with existing (e.g. analogue) in-band services. On the other hand, spectrum awareness and reutilization can help to increase security at radio level, e.g. through detection and countermeasure of threats such as RF jamming or spoofing. While decentralized SOM approaches can alleviate computational load and signaling overhead at the cost of suboptimal solutions when compared to centralized SOM, the trade-off between network load and performance optimality is subject to the characteristics of the multi band environment.

Such considerations and new concepts for spectrum licensing and reutilization are particularly important in the context of new radio technologies such as the millimeter wave, optical wireless, and terahertz communications discussed below, which have a radically different interference footprint compared to the conventional sub-6GHz communications. Their highly directional links

and susceptibility to blockage reduce interference, which significantly increases the potential gains of spectrum sharing and simplifies its use [C6-5]. Dynamic sharing of the different bands will imply a highly complex scenario in which autonomous AI-based mechanisms are envisioned to be crucial.

### 6.2.2 Subnetworks and coexistence

While the current generation of mobile communications (5G) has already made a big step towards supporting use cases with strict requirements in terms of latency, reliability, and throughput, 6G is expected to move this one step further, with latency of loop-cycles down to 100 µs, more than six nines reliability and multi-gigabit data rates (not necessarily concurrently). Different scenarios such as factory floor level communications (e.g., within a robot or a production cell, i.e. in-robot and in-production cell communications), sensor/actuator traffic inside a car (in-car communications) and in-body communications call for these extreme requirements. At the moment, wired solutions are mostly used for these deployments, but strong benefits could be provided in terms of flexibility by using wireless.

One approach to achieve those extreme requirements is the introduction of so-called subnetworks [C6-6]: For many scenarios requiring ultra-low latencies, both the origin and the recipient of a given message are close by. The vision is then to have an access point that a) controls and serves the needs of these devices next to each other and b) is, at the same time, a sort of special device connected to an overlay public or private network. That allows avoiding the delays from forwarding the traffic through the core network. Note that subnetworks may also serve non-critical traffic arising within its coverage area targeted to applications outside of the subnetwork. In this case, the access point is required to be attached to the overlay network.

The scenarios we target are usually covering rather small areas (few meters up to few tens of meters), therefore many characteristics linked to small cells (pico, femto) apply here as well (e.g., low-power transmissions, antenna configurations, etc.).

To be able to meet these extreme requirements, a dedicated air interface (e.g., supporting much shorter symbols), access protocols and diversity mechanisms need to be selected and designed.

In some scenarios it may happen that the subnetwork loses its connection to the overlay network (e.g., a car driving into a tunnel or entering an underground parking lot). As subnetworks are supporting life-critical applications, it is fundamental that mechanisms that allow them running autonomously even when out-of-coverage must be present.

Several subnetworks may be present in close vicinity (e.g., cars on a congested road, production units on a factory floor, people attending a crowded event, etc.). Interference among those subnetworks may then arise and needs to be handled through both respective design choices (e.g., allowing for multiple orthogonal channels) and more sophisticated resource management procedures and interference mitigation techniques (both in a centralized manner exploiting the connection to the overlaying network and decentralized to ensure reliable connections when the subnetwork needs to act autonomously).

A single subnetwork may be required to serve several tens or even hundreds of individual nodes. So, channel access procedures are required. Also, those nodes are usually required to be rather cheap, impacting the degrees of freedom we can make use of for designing them.

While there is very limited or close to no mobility between subnetworks (e.g., a sensor within a car does not switch to another car), individual subnetworks may be mobile. That implies the need for mobility techniques to be applied and also makes the interference very dynamic.

As indicated above there are many open questions to be answered and even more fundamental ones such as: What frequency ranges (FR1, FR2, FR3) are reasonable selections for subnetworks and what are the related implications? How much spectrum do we need? Is licensed or unlicensed access a better choice? Do we use dedicated spectrum, or should we go the ultra-wideband (UWB) underlay route? How can we make the system more pro-active and less re-active? How can we benefit from the fact that subnetworks are attached to an overlay network at least most of the times? How do we integrate the subnetwork into the 5G/6G overlay-network w.r.t. architecture and protocols?

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

Network slicing and edge network function virtualization (NFV) also contemplate multi-RAT operating scenarios, based on highly reconfigurable software defined radio (SDR) hardware featuring heterogeneous processing resources (i.e., general purpose processing elements tightly coupled with hardware accelerators). The functionality of such agile SDR units could be updated at run-time according to traffic context, signal propagation conditions and required performance (e.g., in terms of throughput, latency, and resiliency). An efficient way to achieve field updates of this type is by jointly optimizing the multi-RAT radio and processing resources through suitably selected machine learning (ML) techniques.

To evaluate these complex multi-RAT scenarios, open source simulation models for 4G and 5G technologies from 3GPP releases and different IEEE standard amendments in multiple bands, are needed for an end-to-end and high-fidelity evaluation of smart solutions especially for academia but also complementing private simulation systems from industry. The simulation models need to capture the wide range of spectrum considered for communication services, e.g., from 0.4 up to 71 GHz for 5G NR Rel-17, and consider the multiple heterogeneous spectrum paradigms like licensed, unlicensed, dedicated and shared, which are to be harmoniously used through intelligent frameworks in order to take the best advantage of spectrum resources.

Existing *short-range wireless communication* technologies, including WiFi, Bluetooth and Zigbee, share the same spectrum, e.g. in 2.4GHz. Co-existence of different wireless network technologies in/near such a carrier frequency may cause radio interference, which can lead to relatively high error rate in data transmission. This problem happens especially in unlicensed bands. How to efficiently share the spectrum and improve the co-existence needs careful considerations. Scalability and

power efficiency are critical for the success of a macro, micro, or pico network. Current short-range communication technology provides either high throughput with high power, or low throughput and low power consumption. Whereas IoT devices operate in a very low power mode most of the time, they need to support a short-time high bandwidth transmission. Scalability is needed to support both short-time high throughput transmission and low power transmission. A unified and scalable architecture will be beneficial to support both low data rate (e.g. with Bluetooth, ZigBee, RFID, NFC, etc) and ultrahigh data rate (e.g. up to 100Gbps within 10m coverage). Further requirements to be considered include, e.g. scalable network topology supporting P2P (point-to-point), MP2MP (multipoint-to-multipoint), as well as the smart home and smart building coverage; more power/cost efficient designs, e.g. for zero-power consumption in some dedicated scenarios; and the capability of information and energy simultaneously transporting (IEST).

The wide mmWave spectrum region accounts for different access paradigms, including licensed (e.g., 28 GHz bands), unlicensed (e.g., 60 GHz bands) and shared (e.g., 37 GHz bands) for various applications such as vehicular and cellular. Co-existence of multiple technologies and standards like 5G NR-U (NR in unlicensed), NR V2X (vehicle-to-everything communications) and 802.11ad, 802.11ay, 802.11bd in different spectrum bands should be properly addressed considering various regulatory requirements and access mechanisms. Innovative solutions that increase spectral and energy efficiency need to be considered [C6-7].

### 6.2.3 Wireless edge caching

Wireless communication networks have become an essential utility for citizens and businesses. Wireless data traffic is predicted to increase by 2 to 3 orders of magnitude over the next five years [C6-8] [C6-9]. The implications of these trends are very significant: while continued evolution is to be expected, the maturity of current technology (e.g., LTE-Advanced for cellular and IEEE 802.11ac for WLAN) indicates that the required orders of magnitude throughput increase cannot be achieved by an incremental “more-of-the-same” approach. As far as wireless capacity is concerned, the forthcoming 5th Generation (5G) of standards and systems is focused to a certain extend on the traditional view of “increasing peak rates” [C6-10]. In contrast, it is widely recognized that a major driver of the wireless data traffic increase is on-demand access to multimedia content (Wireless Internet) [C6-8][C6-9]. Peak rates do not necessarily yield an improved user Quality of Experience (QoE). For example, typical video streaming requires rates ranging from ~400 kbps (standard quality) to ~2 Mbps (high quality). What really matters for the end user QoE is the availability and stability of such rates, so that a video can be played anywhere, at any time, and without interruptions. Also, we observe that the users’ content consumption pattern and the operators' data plans are dramatically mismatched. For instance, a standard monthly data plan in the EU includes ~3 Gbytes of LTE traffic at a cost ranging between 15 and 50 EUR, while a single movie requires ~1.5 Gbytes of data, such that the whole plan would be depleted by streaming ~2 movies.

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

rates, which have been already demonstrated by various “5G-ready” experimental platforms [C6-11][C6-12]

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

Recent research in information theory and wireless communication has shown that content distribution over a wireless network (e.g., on-demand video streaming) can be made much more efficient than current state-of-the-art technology by caching content at the wireless edge [C6-13][C6-14][C6-15][C6-16]This means pre-storing segments of the content files at the base stations, at dedicated “helper” nodes, and also in the user terminals.

Traditional caching (e.g., prefix caching) decreases the transmission load by the fraction of data already present (pre-cached) at the destination. With these novel modern techniques, based on extensive use of network coding, it is possible to show that a constant (non-vanishing) per-user throughput can be achieved while the number of users grows to infinity. We refer to this behavior as “full throughput scalability” [C6-17] For the sake of concreteness, consider the analogy with conventional TV broadcasting: in this case, leveraging the broadcast property of the wireless medium, an infinite number of users can be served with a finite transmission resource, i.e., a finite bandwidth and transmit power. For example, this approach is taken in the so-called enhanced multicast-broadcast multimedia service (eMBMS) in 4G networks. Now, the reason for which eMBMS turned out not to be a huge success is that users do not consume wireless multimedia as they used to consume traditional live TV: they wish “on-demand” services, to access what they want at the desired time and location, and not at the time decided by a TV broadcaster. With on-demand delivery, the broadcast nature of the wireless medium cannot be exploited in a direct and trivial manner. In fact, streaming services today treat the on-demand traffic as unicast individual traffic, as if the content was individual independent data. An important consideration here is security. The data can be stored on user’s local cache that depends on the demand of other users in the network. This leads to the possibility of spying and tampering. Authors in [C6-18] formulate a shared-link caching model with ‘private demands’ with the goal to design a two-phase private caching scheme with minimum load while preserving the privacy of the demands of each user with respect to other users.

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

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

- Coding (e.g., combining edge caching with modern multiuser MIMO physical layer schemes);
- Protocol architectures (e.g., combining edge caching with schemes for video quality adaptation such as Dynamic Adaptive Streaming over HTTP (DASH) [C6-19]);
- AI/ML based content popularity estimation and prediction, to efficiently update the cached content [C6-20].

### 6.2.4 Research challenges

| **Research Theme** | **Radio Interference Management** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Spectrum re-farming and sharing | Mid- to long-term<br><br>➔ To be specified in 6G standard. | Advanced methods, and protocols. | Improved KPIs: Energy efficiency, spectral efficiency, capacity, throughput, reliability. |
| Subnetworks and coexistence | Mid- to long-term<br><br>➔ May be specified in 6G standard. | New radio access node setting up a sub-network serving local nodes:<br>- Vertical use cases (e.g. inside a car or robot).<br>- Consumer use cases (smart wearables).<br>- Very low latency, very high reliability, extreme data rates, reduced energy consumption. | New use cases for wireless communications. Replace wired connections through wireless systems (e.g. CAN-Bus).<br>Support of the metaverse. |
| Wireless edge caching | Mid- to long-term<br><br>➔ May be specified in 6G standard. | Advanced methods, and protocols. | Improved KPIs: Energy efficiency, spectral efficiency, capacity, throughput, reliability, quality of experience (QoE). |

## 6.3 Optical Wireless Communication

Despite the tremendous improvements due to the small cell concept and the allocation of new radio frequency (RF) spectrum, the continued exponential growth in mobile traffic [C6-21] means that inevitably the RF part of the electromagnetic spectrum will not be sufficient to be able to drive the cyber-physical continuum which is centered around immersive user experience in an XR environment, digital twins, the convergence of computing, sensing, control and robotics. [C6-22].

It is, therefore, essential to consider the infrared and visible light spectrum, both of which are part of the electromagnetic spectrum for future wireless systems for terrestrial, space and subsea applications. Light based wireless communication systems will not be in competition with RF communications, but instead these systems follow a trend that has been witnessed in cellular communications by inspecting all the generations developed during the last 30 years. Light-based

wireless communications simply add new capacity – the available unregulated spectrum is 2600 times larger than the entire RF spectrum. While most light-based communication systems are based on intensity modulation / direct detection (IM/DD), recently coherent optical wireless systems have been proposed [C6-23].

An important advantage is that off-the-shelf optical devices can be used to harness these unregulated and free transmission resources. By using advanced devices, lab demonstrations showed 8 Gbps from single light emitting diodes (LEDs) and 17.6 Gbps using laser diodes (LEDs) [C6-24]. Moreover, 26 Gbps where demonstrated using a dual wavelength laser device emitting in the visible and infrared spectrum [C6-25]. This work has been extended to a 10x10 WDM system and 105 Gbps were demonstrated at CES 2022. Furthermore, a record of received data rates of 1.1 Gbps by using a single solar cell has been demonstrated. The use of these types of 'data' detectors has the appealing advantage of achieving simultaneous energy harvesting and high-speed data communication – a feature that will become ever more important in order to meet UN [C6-1]. By 2026, it is expected that micro-LED technologies and spatial multiplexing techniques will be mature and cost effective such that white light based on different wavelengths will unlock throughput, leading to potentially 100 Gb/s plus for ultrahigh-data-rate VLC access points [C6-26].

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

Cellular wireless networks which are based on VLC are referred to as LiFi (light fidelity) [C6-27]. LiFi enables bi-directional networked communication including multiuser access and handover. The major research efforts in the last 15 years have been focused on enhancing link data rates of IM/DD optical wireless communication systems. With the advent of LiFi the research focus has begun to shift to challenges related to networking issues using light.

Like in RF networks, there are issues surrounding interference management and interference mitigation in LiFi networks. However, since, for example, there is no multipath fading because the detector sizes are much larger than the wavelength, techniques developed for RF systems may only be sub-optimal. There are also fundamental differences as a result of IM/DD, in that signals can only be positive and real-valued. Consequently, new LiFi-bespoke wireless networking methods must be developed. Moreover, because light can be confined spatially by using very simple and inexpensive optical components, interference can be controlled much easier. This feature also allows step-change improvements of the small cell concept as single cells might cover sub-$m^2$ areas leading to data densities of 88 Gbps/$m^2$ [C6-28].

Furthermore, due to the extremely short wavelength, the active detector sizes are very small, and massive MIMO structures can be implemented at chip-level [C6-29]. This property can be used to

develop unique and LiFi-bespoke MIMO systems, networked MIMO approaches, and new angular diversity techniques in conjunction with low computational complexity cooperative multipoint systems. Diversity techniques in LiFi systems are especially powerful to combat random blockages that naturally occur in a mobile scenario. Moreover, the spatial confinement of signals in LiFi enables the development of radically new physical layer security concepts.

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

**Convergence with 3GPP access:** LiFi communication is bi-directional. Due to the abundance of optical spectrum, typically the visible spectrum is used for the downlink by piggy-backing on lighting systems, while the infrared spectrum is used for uplink transmission. The simplicity of IM/DD in conjunction with advanced modulation techniques [C6-30] enable highly energy- and spectrum-efficient transmission systems suitable for the uplink. These modulation techniques are based on multicarrier approaches. Therefore, it could be argued that a *tight interaction between radio and optical components should be considered at the level of baseband processing*. Since OFDM transmission (e.g. 5G waveforms) is feasible on a free-space IM/DD optical link, it is definitely worth investigating the use of the same basic waveform and protocol stack for radio and LiFi systems. This would allow for a *common baseband processing platform* in both the small-cell transmitters and terminal receivers. Moreover, the 3GPP access-layer protocols are perfectly adapted to the use of downlink-only component carriers.

### 6.3.1 Research challenges

| **Research Theme** | **Optical Wireless Communication** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Advancing transmitter and detector technology including solar cell data detectors acting as simultaneous energy harvesting devices | Mid- to long-term<br><br>➔ May be specified in 6G standard. | Devices that deliver optical-to-electrical (O/E) and electrical-to-optical (E/O) conversion efficiencies, e.g. for electrical bandwidth > 10 GHz at hundreds of mW optical transmit power, and receivers sensitivities less than 40 dBm. | Toward zero carbon footprint of future networks.<br>Tbps wireless multiuser access networks.<br>Extreme MIMO/WDM, > 1000 channels. |
| Developing optimized multiuser access and interference mgmt | Mid-term<br><br>➔ May be specified in 6G standard. | Achieving extreme networks densification indoors towards 100 Gbit/$m^2$ at 100X improved energy efficiency in a multiuser scenario | Extreme network densification, > 100 Gbps/$m^2$.<br>Sub ms latency. |
| Developing supporting optical wireless backhaul | Short- to long-term | Point to point backhaul achieving > 1 Tbps at distances up to 1 km | Backhaul, > 2 Tbps indoors and outdoors. |
| Bespoke RIS technology for OWC | Long-term | RIS to support mobility in indoor and outdoor scenarios | Ultra reliability, mobility. |

## 6.4 Millimeter-Wave and Terahertz Communication

In the last decade, major device, communication and networking features have led to the development and commercialization of millimeter-wave (mmWave) wireless technology. Today, wireless local area networks operating in the Industrial, Scientific and Medical (ISM) 60 GHz frequency band and orchestrated by the IEEE 802.11ad, the IEEE 802.11ay and IEEE 802.11be, are a reality. Similarly, 5G wide area networks operating in the licensed Frequency Range 2 (FR2) between 24 GHz and 52 GHz (and soon to be extended to 71 GHz) are already deployed in several countries. Higher data-rates (approaching 20 Giga-bits-per-second or Gbps) and lower latencies (approaching few milliseconds) are some of the promises of mmWave technologies to enable long-awaited applications including immersive augmented and virtual reality, the tactile internet, and autonomous unmanned networks, among others, all within different contexts, from entertainment to education to remote work telepresence. Moreover, besides communications, the mmWave spectrum has also enabled exciting applications in the field of wireless sensing, from precise localization and radar, to the extraction of body features for security applications.

All the aforementioned commercial technologies and the majority of the research solutions explored to date are for systems operating under 100 GHz. However, this is changing. Today, there are several major academic and industrial research initiatives worldwide aimed at developing wireless solutions in the Terahertz band, broadly defined between 100 GHz and 10 THz [C6-31][C6-32]. The US National Science Foundation Spectrum Innovation Initiative, the Semiconductor Research Corporation (SRC) and DARPA Communication and Sensing at Terahertz frequencies (ComSenTer), the National Natural Science Foundation of China, multiple European projects funded by the Beyond 5G track of the Horizons 2020 program, and several industry-led efforts (e.g., Nokia, Samsung, Huawei) are just a sample set of a growing pool. In addition, there is even a standard above 275 GHz, the IEEE 802.15.3d approved in 2017. When moving to these frequencies, not only there are larger contiguous bands for ultrabroad band communication and networking systems, but electromagnetic radiation interacts with the environment in a more intimate manner, i.e., at the molecular level, giving a whole new meaning to what wireless sensing means.

Of course, such exciting opportunities come with several challenges spanning devices, wave propagation, communication, signal processing and networking. In terms of **THz devices**, major progress has been achieved to close the so-called Terahertz technology gap. The key hardware building blocks of a THz wireless communication and sensing system include the 1) analog front-ends, 2) the antenna systems, and 3) the digital back-ends. There are three main approaches to the development of analogue *THz front-ends*. First, in an electronic approach, frequency multiplying chains can be utilized to up-convert a microwave signal to the terahertz band. By moving from Silicon and Silicon-Germanium-based transistors to III-V semiconductor -based transistors and Schottky diodes, on-chip transceivers able to deliver a few hundreds of milliwatts at 300 GHz and a few milliwatts above 1 THz have been demonstrated [C6-33]. Second, in a photonic approach, difference frequency generation based on laser photomixing is at the basis of several THz transceivers operating at a few hundreds of GHz [C6-34]. While their output power is lower than electronic systems, their phase noise is lower and the potential bandwidth is larger. Third, direct generation and modulation of THz signals with plasmonic devices built with graphene and other two-dimensional materials has been proposed [C6-35][C6-36]. Their high efficiency, combined with their very small footprint (micrometric in size for the entire front-end), is at the basis of future ultra-

massive MIMO systems (see Sec. 6.5). Independently of the approach, high gain directional *antenna systems* in transmission, reception as well as in reflection are needed to overcome the lack of higher transmission power and high propagation losses. Beyond fixed high-gain directional antennas and antenna arrays, lenses and lens arrays as well as metasurfaces can be used to engineer the radiation, propagation and detection of THz signals [C6-37]. Finally, *high-speed data-converters* and digital signal processors (DSPs), able to sustain multi-Gbps and Tbps are needed. Very high-order parallelization, enabled for example by Radio-Frequency Systems on Chip (RFSoc), is one of the clear paths moving forward [C6-38].

Multibeam antennas are also critical components in future wireless communications networks. The quasi-optical techniques are expected to become in the design of THz and mm wave antennas. New challenges for research are the design of lenses with metamaterials and transmitarrays based on metavolumes. Multibeam antenna design also requires the design of low-loss distribution networks, such as groove gap technology, integrated with electronic components. It is necessary to develop very efficient analysis methods for inhomogeneous and anisotropic dielectrics [C6-39].

In parallel to THz technology development, much has been accomplished in terms of understanding **THz** propagation, by following both physics- [C6-40][C6-41] and data-driven approaches [C6-42][C6-43]. There are three main phenomena affecting the propagation of THz waves, namely, spreading loss, molecular absorption loss and blockage. The *spreading loss* accounts for the attenuation due to expansion of the wave as it propagates through the medium and,, because of the small effective area of antennas as we move up in frequency, becomes critical at THz frequencies. This is why high-gain directional antennas are needed. The *molecular absorption loss* accounts for the attenuation that a propagating wave suffers because a fraction of its energy is converted in vibrational kinetic energy in molecules (especially water vapor). Absorption does not occur at all the frequencies, but only at known absorption peaks and, while it is generally perceived as a problem, it can also be at the basis of enhanced physical layer security. Beyond line of sight propagation, high reflection loss, diffused scattering and diffraction by obstacles need to be captured. Ultimately, stochastic multi-path channel models are needed to statistically characterize the channel. In this direction, massive experimental measurement campaigns in different indoor and outdoor scenarios are being performed. These should guide future THz infrastructure deployments and in the development of tailored real-time channel estimators and equalizers.

In light of the capabilities of THz devices and the peculiarities of the THz-band channel, there is a need to develop new communication algorithms and networking protocols, tailored to THz communication systems. At the physical layer, innovative modulations are needed. For short-range communications (below one meter), the use of impulse-radio-like communication based on the transmission of one-hundred-femtosecond-long pulses following an on-off keying modulation spread in time has been proposed [C6-44]. For longer communication distances, enabled in part by ultra-massive MIMO systems (Sec. 5.4), new **dynamic bandwidth modulations** are needed to not only overcome but even leverage the unique distance-dependent bandwidth created by molecular absorption [C6-45][C6-46]. As of today, both single-carrier (e.g., M-QAM, M-PSK, APSK) and multi-carrier (e.g., DFT-spread-OFDM) waveforms for THz systems have been developed, but it is not yet determined which is going to be the waveform(s) for 6G THz systems. Independent of the modulation, and like any wired or wireless Tbps communication system, physical-layer

synchronization (in time, frequency and phase) becomes a major challenge. Additional challenges include new channel coding strategies as well as new physical layer security schemes capture and leverage the impact of molecular absorption.

Moving up in the protocol stack, at the link layer, novel **MAC protocols** are required for THz-band communication networks, since classical solutions do not capture the peculiarities of this band. The very large available bandwidth almost eliminates the need for nodes to contend for the channel. The transmission of very short signals also minimizes the chances for collisions. However, the need for high gain directional antennas simultaneously at the transmitter and the receiver to establish links over realistic distances, makes the beam discovery and tracking a major challenge. In this direction, innovative neighbor discovery strategies that leverage the full antenna radiation diagram as well as new receiver-initiated medium access control protocols have been proposed [C6-47]. At the network layer, multi-hop relaying and routing strategies are needed to support mobile ad-hoc THz networks. Cross-layer solutions that capture the trade-offs between antenna beamwidth, communication distance, available bandwidth, processing overhead and buffer capacity are needed [C6-48], and the use of different forms of machine learning (ML) can help to operate such networks. At the transport layer, as wireless multi-Gbps and Tbps links become a reality, the aggregated traffic flowing through the network will dramatically increase. These will introduce many challenges at the transport layer regarding **congestion control** as well as end-to-end reliable transport. For example, we expect that a revision of the TCP congestion control window mechanism will be necessary.

To support the development of the field, new **experimental platforms** and simulation tools are needed. For the time being, the majority of the experimental platforms developed to date are channel sounders or physical layer testers that rely on non-real-time DSP, and mostly at sub-THz frequencies, but this real-time platforms become a must for testing of anything beyond a point to point link. Finally, in parallel to all the scientific developments, work needs to be done towards regulation and standardisation of the THz-band.

### 6.4.1 Research challenges

| **Research Theme** | **Millimeter-Wave and Terahertz Communication** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Development of THz transceivers | Long-term<br><br>➔ May be specified in 6G standard. | Highly efficient THz transceivers with small footprint | |
| High directional THz antenna systems | Long-term | THz antenna systems (e.g., lens arrays and intelligent reflecting surfaces) that overcome the high path loss of THz bands | Improvement of coverage |
| THz propagation | Mid- to long-term | Develop THz channel models that accurately model propagation in the THz band | Understanding of THz channel models results in improved |

| | | | |
|---|---|---|---|
| | | | coverage, throughput, and reliability |
| Radio methods for THz bands | Mid- to long-term<br><br>➔ To be specified in 6G standard. | Develop modulation schemes, waveforms, synchronization and channel coding that take into account the THz band characteristics | Improvement of spectral efficiency |
| Novel MAC protocols for THz bands | Mid- to long-term<br><br>➔ May be specified in 6G standard. | Development of new protocols that go beyond current collision based schemes | Improvement of device density |

## 6.5 Massive MIMO

### 6.5.1 Ultra-massive MIMO

The grand challenge for mmWave, THz-band and optical communications is posed by the very high and frequency-selective path loss, which easily exceeds 100 dB for distances over just a few meters. As mentioned in Sec. 5.4, high-gain directional antenna and lensing systems are needed to communicate over distances beyond a few meters.

Similarly, as in lower frequency communication systems, antenna arrays can be utilized to implement MIMO systems, which are able to increase either the communication distance by means of beamforming, or the achievable data rates by means of spatial multiplexing. Spatial multiplexing can also be used in a point-to-point line-of-sight link, which is called LoS-MIMO. A special case of LoS-MIMO, which is based on concentric uniform circular arrays is called orbital angular momentum (OAM). In the last decade, the concept of massive MIMO (mMIMO) was introduced and heavily studied in the context of 5G systems [C6-49][C6-50][C6-51]. In such schemes, very large antenna arrays with tens to hundreds of elements are utilized to increase the spectral efficiency to communicate over a large distance. In these arrays, it is important to take mutual coupling between the antenna elements into account in a physically consistent way [C6-52][C6-53]. In addition, MIMO DPD (digital predistortion) can be used to handle the varying nonlinearities in the power amplifier related to a varying impedance mismatch between the power amplifiers and the antenna array due to mutual coupling [C6-54]. Very large antenna arrays have been proved to be very useful for mmWave communication systems [C6-55][C6-56]. These enlarged arrays can be centralized or distributed, giving birth in the latter case to the **cell-free massive MIMO** concept.

When moving to the THz-band, antennas become even smaller and more elements can be embedded in the same footprint. Beyond refining massive MIMO solutions with more antennas, new solutions that can manipulate radiation in space and frequency in unprecedented ways are needed. This is how ultra-massive MIMO enters the game. The concept of **ultra-massive MIMO (umMIMO)** communications, enabled by very dense plasmonic nano-antenna arrays, has been recently introduced in [C6-57] and [C6-58]. Instead of relying on conventional metals, nanomaterials and metamaterials can be utilized to build plasmonic nano-antennas (see Section 6.4) which are much smaller than the wavelength corresponding to the frequency at which they are designed to operate. This property allows them to be integrated in very dense arrays with innovative architectures. For example, even when limiting the array footprint to 1 mm × 1 mm, a total of 1024

plasmonic nano-antennas designed to operate at 1 THz can be packed together, with an inter-element spacing of half a plasmonic wavelength. Such plasmonic nano-antenna arrays can be utilised both at the transmitter and the receiver (1024×1024) as well as in reflection (in the form of IRS) to simultaneously overcome the spreading loss problem (by focusing the transmitted signal in space) and the molecular absorption loss problem (by focusing the spectrum of the transmitted signal in the absorption-free windows).

By properly feeding the antenna array elements [C6-57], different operation modes can be adaptively generated. In *ultra-massive beamforming*, all the nano-antennas are fed with an amplitude- and phase-manipulated version of the same single plasmonic feed. In *ultra-massive spatial multiplexing (um-SM)*, different plasmonic signals are sent through physically or virtually grouped array elements to communicate with different users. Obviously, any combination in between UM Beamforming and UM Spatial Multiplexing is possible. In addition, to maximize the utilization of the mmWave- and THz-channel and enable the targeted Tbps-links, more than one spectral window could be utilized at the same time. In this direction, *multi-band umMIMO* enables the simultaneous utilization of different frequency bands by leveraging the electrically tunable frequency response of graphene-based plasmonic nano-antennas. One of the key advantages is that the multi-band approach allows the information to be processed over a much smaller bandwidth, thereby reducing overall design complexity as well as improving spectral flexibility. In this direction, advanced *space-time-frequency coding and modulation techniques* need to be developed for the umMIMO systems to exploit all of the spatial, temporal and frequency diversities, and hence, promise to yield remarkable performance improvements. In general, there are still considerable challenges in terms of cost, implementation complexity and efficiency. Besides the challenges related to the plasmonic nano-antenna array technology, the realization of any kind of ultra-massive MIMO communication, even at lower frequencies, requires the development of novel **accurate channel models** able to capture the impact of the very large dimensions of the array, where spatial non-stationarities emerge [C6-59]. Similarly, ways to estimate and equalize an extremely large number of parallel broadband channels are needed. In these cases, the use of ML-driven approaches might be the solution. Recently, the term of holographic MIMO [C6-60] has been used to refer to structures with capabilities very similar to that of umMIMO.

### 6.5.2 Intelligent reflecting surfaces

A new and revolutionizing technique able to substantially improve the performance of wireless communication networks is smartly changing the propagation characteristics of the wireless channel through the use of **intelligent reflecting surfaces** (IRSs), or sometimes referred as reconfigurable intelligent surfaces (RISs) or large intelligent surfaces (LIS), which are made of a large number of low-cost reflecting elements able to independently change the amplitude and/or phase of the incident signal so as to achieve specific propagation effects [C6-61][C6-62][C6-63][C6-64]. Since most of the IRSs are nearly-passive, they can be fabricated with light weight, low profile, and even made to be conformal to various objects. As a result, they can be easily deployed in a wide range of scenarios such as walls, ceilings, billboards, lamp-posts, and even on the surface of vehicles to support several applications for smart factories, stadiums, shopping centers, airports, etc. Moreover, IRSs can be deployed as energy-efficient auxiliary devices that are transparent to the

wireless users, without the need for modifying the hardware configuration of the end-user devices. This offers high flexibility and compatibility with legacy wireless systems. Overall, IRS can be used to improve the coverage, reduce interference levels, and increase system capacity in a power efficient manner. Additionally, they can be employed to increase physical layer security, positioning accuracy and even support wireless power transfer.

One approach is to have a large number of low complexity antennas connected to a processing unit. These elements are activated according to the user location and its transmission requirements. This allows unprecedented capacity gains [C6-64], as well as accurate positioning [C6-65]. Although the IRS is made of antenna elements with very low complexity, its implementation may still involve considerable challenges possibly due to the large number of antennas and the associated circuitry. Radio stripes [C6-66] are another interesting variant of IRS, with the antennas placed over a stripe instead of a surface. Similar to other IRS implementations, radio stripes may provide a low-cost implementation to gain capacity and enable cell-free systems [C6-67].

The communication using IRSs and/or radio stripes schemes will require advanced, low complexity techniques for the signal separation, as well as new resource allocation spatial aspects (i.e., which antennas/panels are activated for a given user). To further improve performance, disruptive techniques that take advantage of hardware imperfections such as nonlinear and/or memory effects can be employed [C6-68][C6-69]. In the case of IRSs at higher frequencies, the concept of ultra-massive MIMO in reflection introduced in the previous section can be leveraged.

The large dimensions of IRSs and radio stripes (several tens of meters), together with the relatively short communication ranges (tens of meters or even less), leads to near field communication effects, with its inherent potentials and challenges. The channel estimation can be a considerable challenge due to the large number of parameters to estimate [C6-70]. To overcome these difficulties, parametric channel estimation and tracking techniques [C6-71] can be employed, eventually supported by positioning information. Moreover, the impact of various physical-layer performance limitation parameters has not been investigated yet, including the large scale fading and spatial correlation. An investigation related to these performance deterioration effects is very challenging but is necessary in order to propose appropriate fading mitigation techniques.

### 6.5.3 Distributed and cell-free massive MIMO

Network densification, which increases the number of antennas per site and results in smaller cells, is one of the solutions to achieve the high data rates targeted for 5G and beyond [C6-67]. The antennas of such massive MIMO systems can be deployed either in a collocated fashion where a large array of antennas is mounted in a single location in a compact way or in a distributed fashion with antennas spread over the covered area. The former approach is known as the centralized mMIMO [C6-72] and the latter the distributed mMIMO [C6-73]. Distributed mMIMO can be implemented with a cell-based approach where the access points (APs) are divided into disjoint clusters and APs of each cluster cooperate to serve the user equipments (UEs) within the cell defined by the cluster. This scheme is called coordinated multi-point (CoMP) with joint transmission in 3GPP LTE [C6-74], but unfortunately it did not provide much practical gains [C6-75]. This can be mainly attributed to the considerable amount of backhaul signaling for Channel State Information (CSI) and data sharing resulting from a network-centric approach to coherent transmission [C6-76], whereby the APs in a cluster cooperate to serve the UEs in their joint coverage region. The practical

implementation of JT-CoMP was also hindered by other attributes of LTE, such as frequency division duplex operations and a rigid frame/slot structure, which did not allow for effective channel estimation.

The cell-centric approach can be changed to a user-centric one, where the cluster serving a particular UE is be determined dynamically by choosing the subset of APs closest to the UE. The basic idea of a cell-free system, denoted as resource pooling for frameless network architecture, was proposed and analyzed already in [C6-77], and a UE-centric JT-CoMP scheme denoted "Cover-Shifts" was proposed in [C6-78]. The combination of TDD and mMIMO operations with the dense distributed network topology and the user-centric approach leads to the concept of **cell-free massive MIMO** in which all APs are able to serve UEs cooperatively without any cell restrictions. The cooperation among the APs can be implemented via a fronthaul connection between each AP and CPU and a backhaul connection between CPUs. Compared to its cell-centric counterpart, the cell-free mMIMO is considered as a promising technology [C6-79] due to its improvements in terms of spectral and energy efficiency, especially for indoor and hot-spot coverage scenarios [C6-80]. Nevertheless, some crucial questions remain open for cell-free massive MIMO, such as the relevant initial access, power control, distributed processing considering encoding/decoding, resource allocation, channel modelling, compliance with existing cellular standards and prototype design. Moreover, UAV-enabled communication are expected to play an important role in emerging distributed MIMO scenarios, in which massive number of service requests are expected for a short period of time, e.g., in crowded events. In these cases, low complexity and power efficient UAV-association solutions will be proposed, targeting to improving the spectral efficiency, without inducing signal processing overhead [C6-81].

### 6.5.4 Research challenges

| **Research Theme** | **Massive MIMO** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Ultra-massive MIMO | Mid- to long-term<br><br>➔ May be specified in 6G standard. | - Implementable MIMO DPD for wideband massive arrays.<br>- Utilization of THz-band with the use of plasmonic nano-antenna arrays | MIMO DPD provides enhanced radio equipment sustainability.<br>Increase in spectral efficiency and throughput. |
| Intelligent reflecting surfaces | Mid-term<br><br>➔ May be specified in 6G standard. | Use of reflecting elements to achieve specific propagation effects | Improved energy-efficiency and coverage<br>Increase in capacity<br>Important for dense networks and in security sensitive scenarios |
| Distributed and cell-free massive MIMO | Mid- to long-term<br><br>➔ May be specified in 6G standard. | Distributed implementations of cell-free massive MIMO encompassing a very large number of antennas.<br>Centralized and distributed algorithms for coordinated | Improved performance in very crowded scenarios with high user-perceived throughput and low energy consumption.<br>Increased area spectral |

| | | | |
|---|---|---|---|
| | | transmission/reception involving large numbers of users. | efficiency, and energy efficiency |

## 6.6 Waveform, Multiple Access and Full-Duplex

Cyclic prefix orthogonal frequency division multiplexing (CP-OFDM) has been adopted in several wireline and wireless standards such as ADSL, Wi-Fi, LTE, and recently in 5G NR [C6-82]. CP-OFDM divides the bandwidth into several orthogonal subcarriers. Fine time and frequency synchronization is then required to maintain the subcarrier orthogonality. However, strict synchronization is limiting in certain scenarios. For example, sporadic access in internet of things (IoT) and machine-type communications (MTC) requires relaxed synchronization schemes, in order to limit the length of the signaling overhead [C6-83]. Ideally, the massive number of devices could just transmit their messages asynchronously; being only coarsely synchronized [C6-83]. This could also be advantageous for low-latency communications. However, in multi-user asynchronous access, the CP-OFDM subcarriers are no longer orthogonal, which introduces high inter-carrier interference [C6-84]. Therefore, CP-OFDM is no longer viable in such scenarios. CP-OFDM is also sensitive to phase noise [C6-85], which is larger in state-of-the-art oscillators as we move to higher frequencies. Moreover, the performance of CP-OFDM is challenging in scenarios with very high time-variability, which we find in vehicular applications and high-speed trains.

Several waveforms, e.g. filter bank multi-carrier (FBMC), generalized frequency division multiplexing (GFDM) which is also known as cyclic block filtered multi-tone (CB-FMT) [C6-86], universal filtered multi-carrier (UFMC), and filtered OFDM (f-OFDM) [C6-87] may be more suitable since their subcarriers are better localized in the frequency domain, and therefore limit the inter-carrier interference. A good frequency localization may also be beneficial due to other reasons, e.g. sensitivity to phase noise in mmWave, required accuracy of frequency-synchronization, etc.

The waveforms differ in whether they are orthogonal, whether and how they employ a cyclic prefix, and how the subcarriers are filtered to make them well localized in the frequency domain [C6-88]. FBMC is orthogonal, performs per-sub-carrier filtering and eliminates the cyclic prefix, but care must be taken in the implementation since contrary to OFDM, GFDM and UFMC, it uses offset quadrature amplitude modulation (OQAM). There are also proposals to use FBMC with QAM modulation, but in a non-critically sampled system like filter bank orthogonal frequency division multiplexing (FB-OFDM). FB-OFDM can be orthogonal due to the non-critical sampling and does not need a cyclic prefix either [C6-89]. It is reported in [C6-90] that by applying DFT spreading to FBMC, complex orthogonality can be restored. GFDM also performs per-subcarrier filtering and reduces the overhead of the cyclic prefix by employing it for several symbols, instead of per symbol as in OFDM. GFDM can be orthogonal or non-orthogonal. Non-orthogonality introduces self-interference even if the transmitters are perfectly synchronized. This requires a more complex receiver using e.g. successive interference cancellation. UFMC eliminates the cyclic prefix and applies a filtering for a sub-band consisting of several subcarriers, where the subcarriers within a sub-band are orthogonal to each other but the sub-bands are non-orthogonal, introducing less inter-carrier interference compared to GFDM. Numerous comparisons between those waveforms have been made regarding implementation complexity, spectral efficiency, robustness towards multi-user interference (MUI) and resilience to power amplifier non-linearity etc., see e.g. [C6-91][C6-92].

There are further new waveforms, including orthogonal time frequency space (OTFS) modulation [C6-93] that are proposed to deal with the fast time variability of the channel. OTFS can be considered as a special case of multicarrier code division multiple access (MC-CDMA). It uses long spreading sequences that are well localized in the delay-Doppler domain. This kind of spreading sequences have originally been designed for radar systems [C6-94].

Constant envelope OFDM (CE-OFDM) [C6-95]uses phase modulation to modulate an OFDM signal onto a carrier to reduce the peak to average power ratio (PAPR). A low PAPR is advantageous, as it enables more efficient power amplification, as the lower the PAPR is, the smaller the power backoff can be. It uses Hermitian-symmetric inputs to the IDFT, which leads to a real valued output used to modulate the phase. Low-complexity receivers for CE-OFDM have been studied in [C6-96].

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

Relaxing the orthogonality constraint generally leads to a more efficient and flexible use of the wireless channel. Non-orthogonal multiple access (NOMA) has attracted significant attention in recent years, as it does not only result in larger achievable rates for scheduled uplink and downlink transmissions, but also provide means to cope with packet collisions for MTC scenarios with grant-free access [C6-97][C6-98][C6-99]. Challenges for NOMA research include

- *User pairing:* With a careful design, more than two users can be paired to use the same resource [C6-100]. Yet the challenge to find the optimal one is still broadly open. The main focus is to find a balance between error rate performance, number of paired users, each user's throughput and overall throughput.
- *Power control:* The design of power control in NOMA can affect other performances such as receiver interference level and throughput. E.g. the work in [C6-101], where the power constraint is jointly allocated in full-duplex NOMA, can be further extended to multi-cell scenario.
- *Physical layer security:* In most NOMA cancellation techniques, one user can decode another user's signal in its own device. Such an issue needs further investigations (see e.g. [C6-102]).
- *Code-domain multiplexing:* Different users are allocated different codes and multiplexed over the same time-frequency resources. These schemes include multiuser shared access (MUSA), low-density spreading (LDS), and particularly sparse code multiple access (SCMA), which can be potentially combined with other technologies such as mmWave communications or physical layer security and applied in massive MIMO systems [C6-103]. The main challenge is the design of low complexity SCMA systems, an aspect that still requires research work.

Furthermore, advanced self-interference cancellation techniques can potentially double the spectral efficiency, and enable in-band full-duplex (IBFD) transceivers that offer a wide range of benefits, e.g., for relay, bidirectional communication, cooperative transmission in heterogeneous networks, joint communication and sensing, and cognitive radio applications [C6-104][C6-105]. However, for the full-duplex technique to be successfully employed in next generation wireless systems, there exist challenges at all layers, ranging from antenna and circuit design (e.g. due to hardware imperfection and nonlinearity, non-ideal frequency response of the circuits, phase noise, etc, especially when taking MIMO and massive MIMO into account), to the development of

theoretical foundations for wireless networks with IBFD terminals, and including AI-based algorithms that are capable to perform self-interference cancellation in multiple radio frequency bands. Much work remains to be done, and an inter-disciplinary approach will be essential to meet the numerous challenges ahead [C6-105].

### 6.6.1 Research challenges

| **Research Theme** | **Waveform, Multiple Access and Full-Duplex** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Optimize the waveforms for mmWave, THz, OWC and ISAC applications | Mid- to long-term<br><br>➔ To be specified in 6G standard. | Waveform that take into account the hardware properties and impairments that are important for these applications (phase noise, PAPR, power amplifier non-linearity, ICI even in asynchronous communications, …) | Improved performances in crowded scenarios, high mobility, also contributing to accurate positioning, with reduced energy consumption |
| Enhanced NOMA | Mid- to long-term<br><br>➔ To be specified in 6G standard. | Code design, resource allocation, and receiver algorithms | NOMA can provide higher capacity, energy efficiency, device density |
| Full-duplex transceivers | Mid- to long-term<br><br>➔ May be specified in 6G standard. | Broadband full-duplex RF frontends for massive MIMO | Full-duplex can increase the throughput and enable spectrum sensing or mono-static radar |

## 6.7 Coding and Modulation

Channel coding aims to correct errors to establish reliable communication and can be regarded as one of the most complex parts of the baseband transmission chain [C6-106]. For decades, researchers sought for channel codes with good error correction performance approaching Shannon's capacity limits with manageable complexity. Modern channel coding schemes such as Turbo, LDPC and Polar codes with excellent performance made their way into several communication standards after advancements in semiconductor technology. However, as the decoders for those codes are very complex, there will be implementation bottlenecks (w.r.t. computational complexity, algorithm parallelization, chip area, energy efficiency, etc.) to be addressed for high throughput (e.g. when throughput is over multiple Gigabits per second) and/or low latency applications are targeted by future communication standards.

For 6G system, the peak data rate can attain 100~1000Gbps, and it is difficult for the legacy 5G LDPC decoder to support the ultra-high data rate since the legacy 5G NR LDPC design is mainly applicable for block parallel decoder, which has much lower throughput capability compared to row parallel decoder or full parallel decoder. Therefore, a new LDPC design which adapts to row parallel decoder or full parallel decoder should be investigated to support ultra-high peak data rate of 6G system.

Even with full parallel decoder, the current LDPC design is not likely to achieve the 1000Gbps throughput with an acceptable chip area efficiency. Other coding schemes such as polar codes with SC decoding and new LDPC codes deserve further study. The challenge in designing 6G LDPC codes is how to satisfy the combination requirement of ultra-high throughput, acceptable complexity, comparable performance and flexibility to 5G LDPC codes. For Polar codes, the state of the art CRC aided successive cancellation list (CA-SCL) decoding doesn't scale up well with throughput due to its serial nature of the algorithm. Hence, iterative algorithms like multi-trellis BP (belief propagation) decoding [C6-107][C6-108] may be considered. Furthermore, modified polar code constructions can be adopted to improve the performance of iterative BP algorithms. Approaches like unfolding the iterative decoders using deep neural networks can be used to improve the latency and throughput of the decoders [C6-109][C6-110]. In addition, if polar codes are considered as channel coding candidates for 6G data channel, a new code extension structure should be investigated to support IR-HARQ functionality.

Even though these modern coding schemes show near-capacity error correction performance for many channels (e.g. binary input additive white Gaussian/BI-AWGN channels), their combination with higher order modulation schemes (such as QAM) can lead to a sub-optimal performance. One reason for this degradation is the so-called 'shaping loss' caused by the probability distribution of the transmitted symbols [C6-111]. In order to approach capacity, the transmitted symbols need to have a certain probability distribution (e.g. discrete Gaussian distribution is needed for the transmission over AWGN channels) and using uniformly distributed symbols results in a performance loss, which can be up to 1.53 dB on AWGN channels.

Several solutions for constellation shaping are proposed to compensate this loss. One option is to optimize the locations of the modulated symbols in the constellation diagram to obtain non-uniform constellations (NUC), as adopted in the ATSC3.0 standard [C6-112]. This scheme is also called geometric shaping and shows improvements compared to uniform signaling. Another approach is the so-called probabilistic shaping [C6-113][C6-114][C6-115], where a shaping encoder is employed to encode messages in a way that the transmitted codewords have a non-uniform probability distribution, resulting in a capacity achieving distribution when combined with simple QAM symbols. This approach is shown to perform close to channel capacity. Another feature of probabilistic shaping is that the probabilities of transmitted symbols can be changed to adapt the transmission rate without changing the FEC code. This is of particular importance since a single FEC code design is sufficient for rate-adaption. Considering the diverse requirements of future communications systems, several shaping encoders suitable for both high throughput and ultra-low latency (short blocks) have been proposed in the literature [C6-114][C6-116]. However, a unified and implementable modulation/demodulation algorithm of probabilistic shaping should be studied for different code sizes, different code rates and different modulation orders. In addition, hardware implementation of efficient shaping encoders and decoders needs further investigations.

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

### 6.7.1 Research challenges

| Research Theme | Coding and Modulation | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| New channel coding | Mid-term<br><br>➔ To be specified in 6G standard. | Channel encoder and decoder for 1) extremely high throughput or/and, 2) extremely high reliability or/and, 3) extremely low latency or/and, 4) extremely low power consumption, e.g. to support the ultra-high data rate ranging from 100-1000 Gbps, reliability of 1-$10^{-7}$, etc. | Improved throughput, spectral efficiency, and energy efficiency. |
| New modulation | Mid-term<br><br>➔ To be specified in 6G standard. | Advanced modulation & coding scheme with signal shaping loss removed. | Improved spectral efficiency, and energy efficiency. |

## 6.8 Integrated Sensing and Communication

Especially in the massively connected world of the "Internet of Things" (IoT), it is getting more and more important to be aware of where all these "things" are located, e.g., via positioning and sensing. Future mobile radio systems will play an essential role in providing high accuracy positioning of the "things". In cellular systems like 4G LTE, positioning is performed in a way where several base stations (BSs) send reference signals to the user equipments (UEs) in the downlink, while in the uplink, the UE sends a reference signal to the nearby base stations. This approach is good enough to support the requirements imposed by the FCC for localization of emergency calls (so-called E-911), where accuracy in the order of 50 m is required [C6-117]. However, nowadays, the enormous technological evolution stimulated use cases, e.g., V2X, smart factory, and others, requiring much higher localization accuracy. For instance, for vulnerable road user discovery in vehicular scenarios, a localization error as low as 10 cm is needed (see, e.g. [C6-118]). Currently, 3GPP is considering a positioning accuracy below 20 cm in some scenarios, leveraging the **higher frequencies and large signal bandwidths**, **dense** deployments, and **device-to-device** communications capabilities offered by 5G. However, wireless positioning defined by the current 5G NR standard is only applicable to locate UEs with communication capability. To gain knowledge of the physical conditions of environment or objects, **radar-like abilities**, denoted with sensing can be utilized. In this context, a very promising solution consists of integrating communication with sensing.

While it is by now known that MIMO systems improve spatial diversity and result in spatial multiplexing gains, their power in improving positioning accuracy has not yet been fully exploited. Large antenna arrays at the BS result in very fine angular sampling, which can be leveraged for positioning methods. Further, existing positioning methods only work well in strong LoS environments in general. Many environments, however, experience strong multipath which causes performance degradations and reduces position accuracy. For that reason, the existing methods

need to be revised or new methods need to be developed to accommodate multipath propagation. Such methods can additionally leverage the presence of large antenna arrays at the BS [C6-119]. Clearly, having multiple antennas at the UE can improve positioning. In particular, the ability for a receiver to measure the **time-of-arrival, angle-of-arrival, and angle-of-departure** of distinct multipath components improves not only the ability of the UE to exploit the LoS path (including the possibility to determine the UE's orientation), but also its ability to **map the environment**, in order to determine the location and the extent of dominant reflectors, which can also assist to develop simultaneous localization and mapping (SLAM) schemes [C6-120]. Note that SLAM may lead to high-complexity for the UE due to iterative nature of the algorithm. Environment sensing by network may simplify the process for UE by providing the information about dominant paths which can be used for UE positioning without requiring complicated processing at UE. Such **radar-like** (sensing) **abilities** can occur in either bistatic operation (piggybacking on standard positioning reference signals) [C6-121], or in monostatic operation (requiring full-duplex processing at the BS) [C6-122]. The price to pay is the complexity of the associated simultaneous localization and mapping algorithms [C6-121], which likely need to be solved through mobile edge computing. Moreover, fully harnessing these physical dimensions will require **novel signals tailored to fully exploit temporal, spatial, and frequency domain** [C6-123].

Once such sensing abilities of communication systems are available, a convergence of radar and communications technologies becomes likely. However, passive radar technologies by using communication transmitters that are not under control of the locating entity may not be suitable for critical applications where **service availability and reliability** is crucial. In this context, active radar may be needed. The more active radar systems will be employed, the more interference will be experienced. Radio resource management is one means to cope with the interference. Well-known technologies from cellular communications can be employed using communications links for exchanging such control information. Ultimately, new waveforms can be deployed for combining sensing and communication [C6-124][C6-125]. The full integration of communication and sensing functionalities aims to maximize the efficiency of spectrum usage and minimize resources (hardware, energy) in performing both functionalities. Therefore, high-accuracy sensing without weakening wireless communication will be indispensable for future networks, including short-range communication.

Cooperation can boost the positioning accuracy [C6-126][C6-127], especially in massively connected scenarios. In cooperative positioning, the UEs can send and receive signals and exchange their position-relevant information. If the density of UEs is large, it is likely that there are line of sight (LOS) propagation conditions to each UE from several UEs, which is significantly increasing achievable localization **accuracy and coverage**. There are two different approaches to position calculation, a centralized approach where a central entity calculates the position and a decentralized approach where UEs calculate their position based on the position estimates of the UEs in their vicinity. With side-link communication in 5G, new opportunities for localization and sensing arise, not only in signal design, but also in protocols and algorithms. Important use cases are in the vehicular and unmanned autonomous vehicles (UAVs) contexts, where relative location information from cooperative links can have direct implications for safety and global situational awareness.

Accurate positioning can be leveraged to enable **sensing-assisted communications** [C6-128], e.g., design of narrow beams targeted towards the intended user in traditional cellular systems, facilitate autonomous driving, etc. These effects will become increasingly pronounced as communications systems shift to ever higher carrier frequencies (0.1 THz and beyond). Furthermore, accurate positioning is a prerequisite for emerging industrial and factory applications. Therefore, in contrast to legacy systems, positioning has a big impact on the operation of future communication systems. For these reasons, investigating new positioning paradigms, e.g., for joint communication, positioning and sensing, is essential, as it can further improve spectral efficiency, energy efficiency, and reduce latency. Similar to other applications such as ultra-massive MIMO, data-driven approaches (e.g., based on machine learning) can be used for positioning where complex propagation environments cannot be accurately modelled.

In addition to high-precision positioning and ranging, sensing will also be able to detect object size, shape, material characteristics, motion state, presence or proximity of adjacent objects, etc. Sensing is also expected to play an important role in high-precision imaging and environment reconstruction. These capabilities will be used for industrial automation, Internet of Things, V2X, smart homes, public safety, medical care, smart cities, etc.

Integrated sensing and communication is currently gaining more and more attention by paving the way for a new plethora of services offered by future mobile networks. However, its development poses several challenges, including integrated waveform design, integrated baseband and hardware design, sensing algorithms, multi-band sensing technology cooperation, fusion with other sensing and localization technologies, computational requirements, etc.

### 6.8.1 Research challenges

| **Research Theme** | **Integrated Sensing and Communication** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Integrated waveform design | Mid-term<br><br>➔ To be specified in 6G standard. | New waveform flexible to accommodate communication and sensing.<br>Monostatic, multistatic sensing supported.<br>Resource allocation and optimal transmission parameters for sensing and communications. | Enabler for 6G systems with unprecedented sensing capabilities.<br>Numerous applications, including safer cities and workplaces. |
| Multi-band sensing technology | Mid-term | New technological solutions for sensing at different frequencies including sub-THz and THz bands. | Use of network infrastructure for sensing and localization.<br>E.g. a new network of cross-sector competences between the ICT industry and the radar/sensing industries. |

| | | | |
|---|---|---|---|
| Distributed and cooperative sensing | Mid- to long-term<br><br>➔ May be specified in 6G standard. | Methods and solutions for distributed sensing with data fusion capabilities.<br>Use of AI for data fusion, object recognition, and environment mapping. | Full exploitation of network resources: pervasive deployment, distributed computation, backhaul and core infrastructure.<br>Enabler for the perceptive network paradigm and key element for mapping the physical world into the digital one. |
| Sensing aided communication | Mid- to long-term<br><br>➔ May be specified in 6G standard. | Sensing and communication methods with/without full-duplex.<br>Methods to exploit the radar information for communications (e.g., channel estimation) | Improved spectral/energy efficiency, throughput and positioning accuracy. |

## 6.9 Massive Random Access

The future vision of IoT envisages a very large number of connected devices, generating and transmitting very sporadic data. The challenge here is how to coordinate such a network without consuming much of the network resources and node energy for protocol overhead. Modern information theoretic research has formalised this problem as follows: consider a number of nodes, each of which makes use exactly of the same code, which is hardwired into the device for system simplicity and cost reasons. These nodes access a common transmission resource at random in a very sporadic manner. The receiver (e.g., a base station) must decode the superposition of codewords without knowing a priori who is transmitting [C6-129]. After decoding the messages (payload), the ID of the transmitter can be found as part of the message, if necessary. For example, in some applications it is important to know the transmitter, but there are applications in which it is important to get the data and not the identity of the transmitter. The challenge now is to design such new random-access codes for which the superposition of up to K distinct codewords can still be uniquely decoded. As there is no scheduling of the transmission resource by the base station, the massive random access (MRA) is contention-based. In contention-based MRA, the collisions of multiple packets in the same slot are inevitable. To solve these collisions and support a MRA of high user loading, non-orthogonal multiple-access (NOMA) techniques should be considered. NOMA has been well researched in grant-based schemes. However, in MRA, the transmission is grant-free, the global power control, resource allocation and configuration cannot be used, which poses a challenge to deal with inter-user interference (IUI). The one-dimension discrimination of power domain brought by the near-far effect of MRA is not enough to deal with severe IUI. Therefore, higher-dimension domains like code domain and spatial domain should be introduced. In code domain MRA schemes, the transmitters randomly select their non-orthogonal spread codes [C6-130]. At the receiver side, the codes are detected and used to alleviate IUI. The prior knowledge of the statistic properties of data (e.g., constellation shape), codebook, and CRC result should be fully utilized for advanced blind detection [C6-131].

Spatial domain is an effective way to increase the spectrum efficiency. Although the orthogonality of the spatial domain cannot be guaranteed in MRA transmissions, it is still very efficient as multiple receive antennas increases the degrees of freedom without extra resource consumption. However, using conventional transceiver to acquire spatial degrees of freedom is very challenge in MRA transmission. As there is no coordination of the transmission resource by the base station, active UEs in MRA autonomously select pilot sequences from the predefined pilot sequence set. Inevitably, multiple active UEs may select the same pilot sequence, which is called 'pilot collision'. For a given pilot set, the probability of pilot collision increases rapidly with the increase of the number of active UEs. Pilot collision will lead to miss detection and inaccurate channel estimation of collided UEs, which severely degrades both the suppression of IUI and the compensation of channel distortion experienced by the data symbols. Moreover, considering the extremely simple transmit procedure of MRA, the received signals could experience large time offsets (TO) and frequency offset (FO) : 1) Due to the lack of uplink timing alignment/timing advance (TA) procedure, the transmitted signals of active UEs may arrive at the BS with different time delays, with each UE's delay being determined by its distance to the BS, thus the received signals from the UEs near the edge of the cell would experience large TOs; 2) Due to the lack of tight frequency synchronization, the oscillator misalignment and Doppler effect could cause a large FO. Large TO/FO will further increase the symbol distortion on the basis of distortion induced by wireless multi-path channels, which makes the channel estimation and symbol demodulation more difficult. As a result, it's very challenging for the multiuser detection (MUD) of MRA transmission as it will encounter not only heavy IUI and severe distortion on the received symbols, but also uncontrollable pilot collision. To achieve a better MUD performance for MRA, different transceivers have been proposed. One solution is data-driven method not relying on pilots via blind receive beamforming and blind equalization [C6-131][C6-132] Another way is enhancing the pilot design to reduce pilot collisions, for example, multiple independent pilot scheme [C6-133] and extremely sparse pilot scheme [C6-134] can be used.

In MRA, the design of channel access protocols departs from conventional approaches used for predictable, persistent, and synchronized data sources. This new random-access paradigm is inherently related to **group testing**: A set of statistical procedures for which it is possible to identify the presence of certain individual agents by sampling combinations thereof [C6-135] and [C6-136]. A related setting consists of coded slotted Aloha, where sparse codes with iterative message passing decoding are developed along multiple random transmissions, to effectively eliminate interference by a sort of low-complexity successive interference cancellation [C6-137]. The performance can be further improved using low-rate channel codes in combination with multi-user detection at the physical layer [C6-138]. While traditional access protocols were designed to avoid interference, the key idea of such innovative approaches lies on the ability to harness information from multi-user interference and constructively utilize it for contention resolution, in combination with advanced signal processing techniques at the receiver [C6-139].

A related problem consists of activity detection, e.g. using a receiver with a large antenna array: In this case, users are given unique signature sequences and transmit at random in a completely uncoordinated way. The base station has multiple antenna observations and must identify the “active set” of users that are transmitting. This problem is related to **compressed sensing** where the sparse vector to be estimated is the vector of 0s and 1s, denoting “absence” or “presence” of the transmitters. Modern techniques based on approximated message passing (AMP) can be used for

this purpose [C6-140] and preliminary research results show the exact trade-off between the length of the signature sequences (protocol overhead) and the number of active users, such that the probability of identification error can be made as small as desired [C6-141] and [C6-142]. Compressed sensing-based multi-user detection may also be combined with coded random access schemes [C6-143].

Massive MIMO technology can be efficiently exploited in massive random access to improve the activity detection accuracy by leveraging the high spatial multiplexing gains. The combination of massive MIMO with non-orthogonal multiple-access (NOMA) techniques emerges as a promising area for the design of novel random access protocols. With the aid of multiple-measurement vector compressed sensing techniques [C6-144], the user detection error in grant-free random access can be driven to zero asymptotically in the limit as the number of antennas at the base station goes to infinity. Another approach of jointly addressing the problems of activity detection and collision resolution is the grant-based strongest-user collision resolution protocol, able to resolve collisions in a distributed and scalable manner by exploiting special properties of massive MIMO channels [C6-145].

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

In a second step, this line of research should consider waveforms adapted for low-latency sporadic access for the cyber-physical systems characteristic of the tactile Internet [C6-146]. Here, sub-ms latencies may be required in order to control moving or even flying objects (passenger drones) or other similar scenarios requiring the combination of ultra-reliable communication with centralized control systems. Similar mechanisms will also be required for evolved Industry 4.0 applications [C6-147]. It is envisaged that the physical-layer transport mechanisms will be associated with real-time cloud computing (mobile edge computing) in proximity to the radio network to implement the necessary control loops. This concerns primarily sub-6GHz access for the uplink and massive connectivity of objects to wireless infrastructure. The objective is to provide solutions for the evolution of cellular IoT uplink waveforms and protocols that scale to huge number of connected devices with stringent energy and potentially latency constraints.

Another promising research direction lies on the use of data-driven methods for the design of new generalized random-access protocols, where the receiver exploits certain side information about the (possibly correlated) activation patterns of the devices. In this context, AI/ML techniques have the potential to build on the availability of data and identify features that could enable the interaction with the underlying random access protocols, e.g., reduce connectivity overhead and prevent the under-utilization of the scarce radio resources [C6-148].

### 6.9.1 Research challenges

| Research Theme | Massive Random Access | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Code design and NOMA for random access | Mid-term<br><br>➔ To be specified in 6G standard. | Codes for which a superposition of codewords can be uniquely decoded<br>Advanced receiver algorithms to resolve packet collisions for contention-based access | Improve the area spectral efficiency and reliability by exploiting interference |
| Synchronization, channel estimation, and beamforming for random access | Mid-term<br><br>➔ To be specified in 6G standard. | Pilot sequence design and autonomous pilot selection schemes<br>Channel estimation without tight time/frequency synchronization<br>Blind beamforming and equalization algorithms | Reduce energy consumption and support mobility by relaxing the synchronization requirements, which enables deep sleep modes of inactive devices |
| User activity detection | Mid-term<br><br>➔ To be specified in 6G standard. | Algorithms to determine the set of active users based on group testing or compressed sensing<br>Joint activity and data detection<br>AI/ML techniques to exploit correlated activity patterns | Efficient user activity detection is crucial for grant-free transmission with high device density |

## 6.10 Machine Learning Empowered Physical Layer

Application of artificial intelligence (AI) and machine learning (ML) in communication systems spans a wide range from optimizing a specific function to end-to-end learning of the entire communication system [C6-149]. AI/ML can also be used to simplify network operations and reduce the need for human participation and supervision. With proper use of AI/ML, the system can automatically address the vast majority of network anomalies and only require human intervention in a small number of extremely unusual cases.

For these reasons, ML techniques are expected to redefine the classical approach in communication system design to achieve global optimization or performance improvement. 6G is expected to be a large-scale and self-organized system that integrates terrestrial and non-terrestrial networks to provide seamless wireless connectivity elsewhere, and ML techniques can play a meaningful role to develop this concept with two different viewpoints. A block-structured strategy focuses on the application of a specific ML algorithm for each block of the communication systems to optimize the individual performance of each block. This methodology inherits the advantages of former models and algorithms. For example, channel estimation can be studied as a regression problem and the selection of modulation and coding scheme (MCS) can be learnt from an exhaustive exploration of the environment [C6-150]. A second perspective addresses end-to-end (E2E) communication and targets at the optimization of the complete communication system, from the transmitter to the receiver. An E2E ML approach may consist in the representation of both blocks as black boxes and the characterization of the transmitter as an autoencoder that can infer MCS through data-driven analysis [C6-151]. ML techniques applied to spectrum/environment awareness may assist in the

adoption of both strategies given the scarcity of tools to face 6G challenges such as 3D transmission models, the adoption of new frequency bands (THz), the introduction of new network elements (e.g. IRS), beamforming in ultra-massive MIMO communications, and non-identified interference sources.

In the following, we describe some of the topics where we expect AI/ML research to be important to improve the physical layer design and performance.

- *Overall physical layer:* At the highest level, we can consider the entire physical layer transmitter-receiver chain as an auto-encoder. While promising, it does not make use of the vast knowledge established during a century of communications research. Therefore, a more feasible approach could be to learn only parts of the physical layer, while still training the whole link in an end-to-end manner. Practically, we believe that the greatest benefits can be reaped where there exists a model deficiency or large variations in individual units. The model deficiency can manifest itself in reality being too complex to model with sufficient fidelity at an affordable effort, e.g., some radio propagation channels. Individual variations are expected in e.g., low-cost hardware for IoT or distributed massive MIMO.
- *Channel learning:* The wireless channel can be complex and rapidly changing, in particular as we are progressing to even higher carrier frequencies and into the light spectrum. At higher radio frequencies, communication is often beam-based. To learn the wireless channel and be able to correctly predict time, frequency, and spatial properties allows for performance improvement and overhead reduction. A further example is molecular communications discussed in Chapter 11, where an accurate model describing the molecular channel is yet to be developed.
- *Radio interface design:* AI/ML has the potential to design new signal waveforms or modifying existing signals. The signal constellation is often designed or selected based on the prevailing channel conditions. It has been shown that learning the signal constellation can give improvements both in performance and in control signaling reduction. AI/ML can be used to design schemes beyond QAM, such as fully pilotless waveforms, and find new modulation types for THz carrier frequencies. Other functions in the transceiver chain that have received considerable interest include FEC design and decoding. At the receiver, AI/ML can compensate for the loss in performance when operating under non-ideal conditions, e.g., frequency offset, colored noise, interference, cross-talk, etc.
- *Multi-antenna systems:* MIMO, massive MIMO, and distributed MIMO offer rich opportunities for AI/ML research. Current MIMO systems use precoders and beam pair search procedures. The antenna arrays are assumed to be uniform and the RF chains equal. AI/ML can be used to optimize precoders for non-ideal conditions, where the previously mentioned assumptions do not hold. As the number of antenna elements increases and the deployments go from regular arrays to almost random positions, the need for AI/ML algorithms increases even further.

  Fully digital beamforming requires independent, affordable RF chains. This may lead to increased hardware impairments and variations between individual RF chains. AI/ML solutions would be desirable to compensate for this since manually measuring and calibrating such systems would be prohibitively complex. Moreover, an important aspect in large-scale multi-channel transmission scenarios is the optimization of the energy efficiency. To this aim deep reinforcement learning and/or federated learning approaches are expected to balance the trade-off between the power consumption and achieved throughput [C6-152]
- *Learning over the air:* The wireless medium mixes signals from different sources. This is often a source of nuisance since it causes interference. However, if it can be controlled, it becomes

possible to use this to perform “learning over the air”. It should be further investigated how reflective surfaces can be used to create a virtual ML model.

- *Hardware:* The dominating hardware architectures for AI/ML algorithms are CPUs and GPUs, but these are not optimal for real-time physical layer algorithms. So-called in-memory computing is a more promising hardware solution for real-time physical-layer ML solutions. Moreover, inspired by the structure and energy efficiency of brains, new neuromorphic hardware architectures have been devised and with them, pulse-based – spiking – neural networks. These and other new architectures and methods to realize learning algorithms should be investigated.
- *Hardware impairment modeling:* In low-cost devices and higher operating frequencies, non-linear effects and other impairments are more pronounced. Complementing the hardware models that exist, AI/ML should be used to compensate for the performance loss that would result if these impairments are not compensated for. Here it is relevant to establish a useful trade-off between algorithmic complexity and hardware simplicity. Another scenario is where the optimal solutions are computationally demanding and/or not possible for practical hardware architectures. In this context, ML can be used to approximate those optimal solutions with lower complexity, albeit at a performance loss. Examples include maximum likelihood detection, channel estimation, etc.
- *Performance vs. resource trade-off:* Many network nodes are energy constrained, e.g. when they are battery-powered or the heat dissipation should be limited. Frequent retraining of large AI/ML models may contribute negatively to the sustainability of future systems. Thus, an important high-level topic is the trade-off between performance and resource/energy use or AI/ML algorithms.
- *In radio network AI computing:* Model training and model interference of AI/ML methods can lead to large amount of computations. Computation efficiency is important to apply AI/ML in radio networks. The computation tasks can be performed by multiple nodes in the network, e.g., jointly by BS and UEs, then the data, parameters of AI/ML models, and outputs of AI/ML models need to be exchanged among network nodes. In this case, communication and computation can be jointly designed. Technologies like over-the-air computing (AirComp) [C6-153] or coded computing [C6-154] can be considered to improve the efficiency or reliability of the computing in radio networks.
- *Physical layer security:* Ubiquitous access and a massive number of low-cost devices will be key features of 6G and, at the same time, will substantially increase the number of potential threats and cyberattacks to the reliability and security of both users and networks. In particular, a major concern is that, due to the propagation medium’s broadcast nature, wireless transmissions are exposed to attacks that may require very unsophisticated techniques, e.g. a jamming attack just needs a very powerful transmitter. To guarantee security in wireless networks, two approaches emerge to provide secure wireless communication in addition to network security mechanisms. Physical layer security (PLS) is driven by the exploitation of the physical characteristics of the wireless channels to combat jamming and eavesdropping attacks [C6-153]. While PLS techniques have traditionally leaned on artificial noise generation or diversity, a new wave of ML techniques supported by massive MIMO or full duplex mmWave can face the new security challenges on the radio interface. On the other hand, ML techniques may exploit spectral and signal analysis to detect radio attacks. This paradigm may extend the utilization of signal processing and ML to the detection of jamming, eavesdropping or rogue base stations.
- *Trustworthiness of AI/ML algorithms:* Topics in this area include explainable AI/ML, uncertainty outside the training distribution, and spoofing.

Compared to model-based algorithms, an issue with current-day AI/ML algorithms is that they appear as black-box solutions and their intermediate states cannot (always) be interpreted in a meaningful manner. Explainable AI and how to incorporate existing model-based knowledge in training can bring increased transparency and trust in the models.
Mathematical models can be extrapolated to understand their asymptotic behavior. For AI/ML algorithms we cannot do so when the input is far from their training set distribution. Methods should be developed to not only allow AI/ML models to make accurate predictions but also know when they are operating far from their training distribution.
Many AI/ML algorithms can achieve super-human performance on e.g., image recognition problems. However, it has been shown that AI/ML algorithms can be tricked into misclassifying images when a noise pattern, imperceptible to humans, is added to the original image. Designing AL/ML algorithms robust to unintentional and intentional spoofing attempts is increasingly important as AI/ML algorithms enter 6G systems.

### 6.10.1 Research challenges

| **Research Theme** | **Machine Learning Empowered Physical Layer** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Hardware architecture | Mid-term | New architectures and methods to realize real-time learning algorithms | Improved hardware architectures |
| Hardware impairment modeling | Mid-term | Compensate for the performance loss due to non-linear effects in low-cost devices by complementing the hardware models using AI/ML. | Reduce the complexity |
| Overall physical layer | Mid-term<br><br>➔ May be specified in 6G standard. | Using AI to model scenarios/systems with model deficiencies or mismatches such as communication with non-uniform antenna arrays. | Improved spectral efficiency and throughput |
| Radio building blocks | Mid-term<br><br>➔ May be specified in 6G standard. | Using AI to design modulation schemes beyond QAM, pilotless waveforms, beam search procedures in massive MIMO systems, and enhanced channel prediction mechanisms. | Improved spectral efficiency and throughput |
| Enhance RAN adaptivity and intelligence without increasing complexity at RU end | Mid- to long-term | Transition from RU-centric to cloudified RAN AI-driven PHY building blocks. | Leverage on edge efficient and sustainable power supplies and higher end specialized heterogeneous computing HW to shorten training times by one additional order of magnitude and simplify the RU architecture. |
| Reduce complexity in neural network modeling of | Mid- to long-term | Applying intelligent histogram-based and alternative training data selection mechanisms, | Reduce neural network training times by two orders of magnitude to enable faster |

| | | | |
|---|---|---|---|
| multidimensional nonlinear effects. | | together with feature selection and feature extraction techniques can contribute to reduce the dataset dimensionality. The latter can also be applied to prune the neural network and reduce its complexity for multidimensional nonlinear problems. | adaptivity, reduce complexity by one order of magnitude and power consumption by a factor between 2 and 5. |
| In radio network AI computing | Mid- to long-term<br><br>➔ May be specified in 6G standard. | Apply in radio network computing techniques to improve the efficiency or reliability of the computing tasks. | Improve the computing efficiency of AI/ML algorithms or reduce the latency of deploying AI/ML models. |
| Physical layer security | Mid-term<br><br>➔ May be specified in 6G standard. | Physical layer security as well as ML based techniques to combat the increased number of potential threats and cyberattacks to secure both users and networks. | Increase security risks against cyber attacks. |
| Trustworthiness of AI/ML algorithms | Mid-term | Increase transparency and trust in the models by utilizing 'explainable AI'.<br>Robust against unintentional and intentional spoofing attempts. | Gain human trust in the network by making the algorithms transparent. |

# 7. Optical Networks

*Editor: Raul Munõz*

## 7.1 Introduction

Within the next decade, the world will go digital, improving our quality of life and boosting the industrial productivity. Artificial intelligence will free us up from routine tasks and unleash human creativity and product innovation. We will enter a new era in which billions of things, humans, and connected vehicles, robots and drones will generate Zettabytes of digital information. All this information needs to be transported, stored, and processed in an efficient way.

Smart connectivity will be the foundation of this new digital world: Always available, intrinsically secure, and flexibly scaling. A programmable network infrastructure will be the nervous system that the digital society, industry, and economy will heavily rely upon. Delivering the required performance, resilience, and security levels, while satisfying cost, energy efficiency and technology constraints, presents a formidable research challenge for the next decade.

Overcoming the challenges in scaling electronic interconnect speeds, advanced electro-photonic integration will enable a new generation of optical networking and IT equipment. Combining the advantages of optics and electronics is the way forward to deliver unprecedented functionality, compactness, and cost-effectiveness.

Optical networks have long been the solution of choice for submarine, long-haul, and metro applications, residential/business fixed access, and mobile fronthaul/backhaul networks, thanks to the unparalleled capacity, energy efficiency and reach of optical fibre transmission. In recent years, optical network technologies have conquered inter and intra data center networks and have created tremendous growth in this sector.

From ground-breaking discoveries such as new types of optical fibres and EDFAs over products such as WDM systems and 100 Gb/s transponders to global standards such as SDH and OTN, Europe has been at the forefront of optical communications R&D for many years.

Seven out of the top 20 network operators are headquartered in Europe while five out of the 10 largest optical equipment manufacturers have major R&D centres in Europe. By revenue, they represent more than 50% of the global optical equipment market. Two of the largest component manufacturers have operations in Europe and more than a hundred SMEs and universities provide complementary innovation on network, system, or component levels. Optical technologies leverage a telecommunication infrastructure market of 350 billion EUR and impact more than 700,000 jobs in Europe [C7-1].

Yet, innovation cycles are fast, and competition is fierce. New research challenges require a continued effort to defend and strengthen Europe's leading position.

## 7.2 Vision

Optical communications and networking technologies are essential to provide high-speed, cost-effective, energy-efficient, secure, and reliable connectivity services for 6G, spanning from the fixed access to the transport network, as well as for inter and intra data center communication, as shown in Fig. 7-1. Open and disaggregated packet and optical technologies will be further developed to provide a converged packet-optical network with a more granular and large-scale management of flows with dedicated and deterministic QoS in support of B5G/6G mobile networks, IoT/V2X, free-space optics, non-terrestrial networks, and fixed networks (enterprise, residential). The need for extending the cloud towards the network edge (e.g. Street cabinet, Cellsite, RSU) will require the deployment of edge computing integrated with the packet optical networks, providing a wide ecosystem where packet, optical, edge computing and cloud converge.

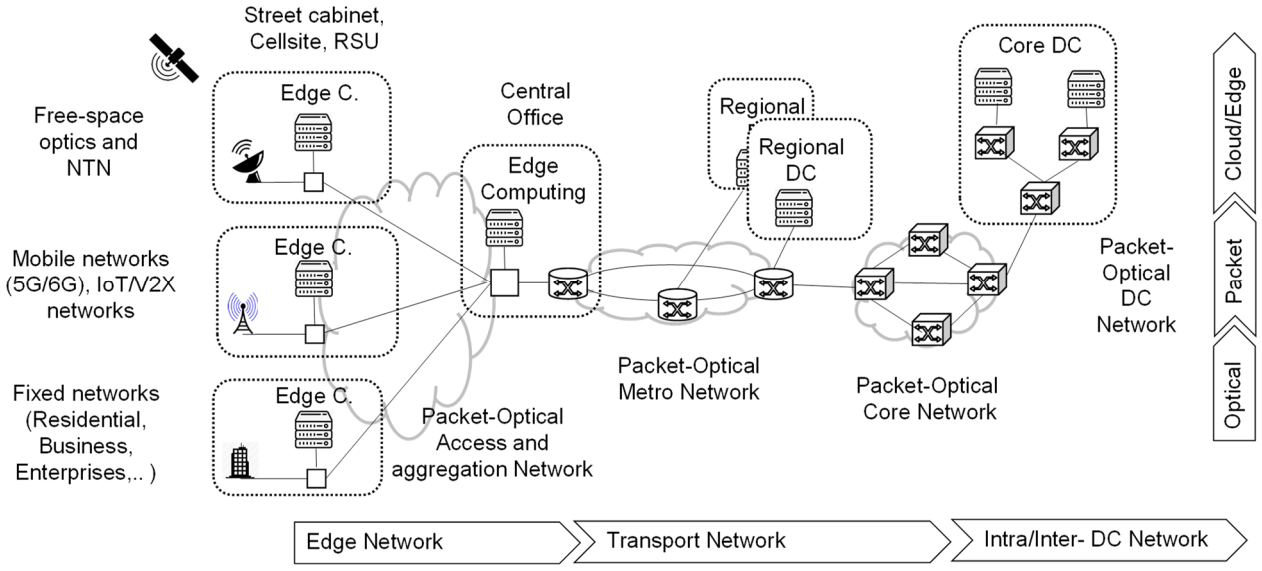



*Figure. 7-1 End-to-end optical network scenario*

The following high-level requirements are identified for the target vision:

- ***High-capacity scaling and reliable connectivity*** through the adoption of spectrally and spatially multiplexed systems with suitable photonic technologies and devices to support network reconfigurability and dynamicity.
- ***Cost-effective and energy-efficient systems*** integrating suitable photonic technologies and devices that meet mixed requirements in terms of cost, reach, throughput, power consumption and footprint.
- ***Advanced electro-photonic integration*** enabling a new generation of optical networking and IT equipment, overcoming the challenges in scaling electronic interconnect speeds. Combining the advantages of optics and electronics is the way forward to deliver unprecedented functionality, compactness and cost- and energy-effectiveness.
- ***Access to everything.*** Current optical access network solutions will evolve further to also fulfil requirements of future applications demanding ultra-high speed and low latency. New architectures, derived from low-cost Fibre-to-the-Home solutions need to be scalable to support dense 6G deployments at a low-cost point.

- ***Deterministic networking*** *in optical/packet networks* to encompass multiple switching capabilities, being able to accommodate traffic flows with a more granular and large-scale management with deterministic QoS.
- ***Edge-cloud continuum*** *to* provide a converged packet, optical, edge computing and cloud ecosystem, requiring the deployment of edge computing integrated with the packet optical network based on the need for extending the cloud towards the network edge.
- ***Full network programmability***, considering new deployment models for open and disaggregated optical networks, in which open devices and sub-systems result in a programmable optical transport network with open and standard interfaces.
- ***Network and service automation*** by deploying closed-loop control and advanced telemetry enabling AI mechanisms, using telemetry data for network continuous optimization in a proactive way with machine learning-assisted analytics, that may anticipate to the problems and events, and propose corrective actions.
- ***Network multi-tenancy*** with intent-based policies and software defined security to deploy smart, secured, and trustworthy end-to-end network and transport slices in open and disaggregated networks.
- ***Efficient Integration of optical technologies for radio access networks*** to provide optical connectivity to each radio antenna and deliver ultra-high speed and low latency at a cost that is compatible with the revenue generated by smaller and smaller cells.
- ***Integration of free space optical technologies*** to complement with next-generation technologies, such as 5G and 6G wireless networks to be widely deployed in various indoor (e.g., data centers), terrestrial (e.g., mobile networks), space (e.g., inter-satellite and deep space communication), and underwater systems (e.g., underwater sensing).
- ***Secure communications*** deploying optical quantum communications and related technologies (such as the adoption of QKD and quantum cryptography in optical networks) in coexistence with the actual deployed network infrastructure.

## 7.3 Sustainable capacity scaling

Global data traffic in optical networks has been growing a high and steady pace of x2 every 2-3 years over the past 15 years and there is no sign that this pace will be slowing down significantly in the upcoming decade. Hence networks need to urgently adapt. Not all segments will be equally hit. For the sake of efficiency and latency, data will be stored closer to the users of these data, hence metropolitan and edge optical networks will grow considerably faster than long-haul fiber networks. At the same time, cloud providers will continue to massively offload the public internet into their private intranets. Projections of future traffic predict required data rates of 10 Tb/s line interfaces and over 1 Pb/s for optical fibre systems by 2025 [C7-2], while optical interconnect capacity are expected to be aligned with the Ethernet roadmap of line interface speeds (~6.4 Tbit/s in 2030). Networks also need to provide headroom for unexpected traffic increases, as observed in several EU member states during the health crisis of 2020-2022.

### 7.3.1 Scaling to Petabit/s capacities in core and metro networks

This evolution stumbles upon the most fundamental limits of physics that are: Moore's law on Silicon integration and Shannon's limit on optical fibre capacity, both of which are considerable barriers to growth. New research efforts to radically improve the dense integration of high-speed electronics and optics (separately, or together) are needed, especially for metropolitan networks,

where capacity growth is more constrained by cost. However, there is a clear danger that very soon, a two-fold increase in the requested capacity could be requiring doubling the amount of optical/electronic hardware. This would increase cost in a linear fashion and threaten future capacity growth. Obviously, disruptive approaches are now needed.

To expand network capacity beyond the Shannon's and Moore's limits, given by current fibre and integration technology, we need to exploit all dimensions in space and frequency, opening new optical wavelength bands and space division multiplexing. The exploitation of new wavelength bands will require advances in a multitude of technologies ranging from optical amplifiers, tailored to these new bands, to a large variety of opto-electronics devices and sub-systems; namely, tuneable lasers, optical multiplexers, couplers, optical mixers, photodiodes, and wavelength selective switches. Advances in fiber technology will facilitate this evolution. In particular, hollow-core fibres promise both larger bandwidth (up to several hundreds of nm) and much lower latency (~30% lower) compared with standard fibres, at the cost of higher attenuation for now. But this attenuation is expected to decrease in the next decade. In parallel, improvement of the attenuation in the C-band is still possible, for instance using pure silica core fibres, showing that innovations are still possible in this field. Besides, system design should be revised and updated taking into account the new physical impairments which will undoubtedly come up in the new bands. Intensive research efforts are necessary along these lines.

In parallel, space division multiplexing must be investigated. This approach can offer significant capacity increase, either by multiplying fibre count in cables, or by introducing multicore or multimode fibres. Here again, new node and system architectures, new digital signal processing, new space division multiplexers, new switches and new optical amplifiers are needed, along with new fibre types.

Finally, capacity can also be gained through margin reduction. Recent publications show that a doubling of network capacity, or even more, is possible through careful margin reduction never hitting the guaranteed limit of resilience, which has been rendered possible by the availability of a wealth of monitoring data in new optical networks, as discussed in paragraph 6.7.

### 7.3.2 Next generation terabit/s transceivers

Recent successful innovations will be exploited far beyond the current status. It can be predicted that optical communications are moving to coherent transmission everywhere. Once viewed as prohibitively expensive, coherent technologies will massively expand from long-haul systems into all fields of optical communications: to support the new generations of wireless systems of 6G, to offer enhanced broadband access, to cope with the growth of inter data center communications, to make edge cloud a reality, and even to allow a new breed of intra-data center networks. Coherent is the most promising technology to bridge the gap which is caused by the Shannon limit, leveraging "shaped" modulation formats, flexible rates, higher than 100 Gbaud symbol rates, and increased density WDM.

Photonic integration and co-packaged optics will be important for efficient scaling, particularly for replacing electrical interconnects with optical interfaces in highly parallel packet fabrics. They will allow for the integration of multiple optical devices into a single platform and the closer integration

of the optical devices with their corresponding electronic circuits. Overall, a change of scale in component count per square millimetre will be required to less that 1 pJ/bit/s in the medium-term future.

### 7.3.3 Research Challenges

The research challenges from the previous subsection are summarized below:

| **Research Theme** | **Sustainable capacity scaling** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| 1 Petabit/s over 1000 km in a long haul cable<br>Overcome next major milestone of system capacity | Long-term (finished in 7y+) | *Provide mix of technologies to meet the goal, and design rules, including compensation of impairments.<br>*Achieve 2Tbit/s per lambda in 5y | High-capacity scaling and reliable connectivity |
| Beyond Shannon and Moore in metro<br>Enable massive parallelism in space and wavelength domains | Mid-term (finished in 5y) | *Multi-band WDM amplification (20THz per fiber in 3y, 200THz line amplifier node in 5y)<br>*subsystems for space-division mux (>10 modes or cores or fibers per device) | High-capacity scaling and reliable connectivity<br>Advanced electro-photonic integration |
| Petabit/s energy-efficient interconnects<br>Cost per bit and power per bit reduction | Mid-term (finished in 5y) | *Make leap in optical integration and co-packaged optics for lower consumption interconnects (<1pJ/bit/s) | High-capacity scaling and reliable connectivity<br>Cost-effective and energy-efficient systems<br>Advanced electro-photonic integration |

### 7.3.4 Recommendations for Actions

| **Research Theme** | **Sustainable capacity scaling** | |
|---|---|---|
| Action | scaling to Petabit/s capacities | Next generation terabit/s transceivers |
| International Calls | X | X |
| International Research | To leverage industry and academic efforts vertically for critical mass. | To federate industry and academic efforts across Europe for design and manufacturing of challenging components and systems |

## 7.4 New switching paradigms

Technologies such as autonomous driving, augmented/virtual reality, and augmented workspace, are about to become reality in a not so far future. This will impose significant challenges for the network. New network architectures with edge clouds close to the end user and centralized clouds with flexible function split are required. In order to enable this flexibility, new switching paradigms are needed to connect real-time programmable optical devices in distributed architectures. The steep learning curve in photonics integration will for example allow optical flow switching

approaches, which were previously considered too costly and/or complex. This can pave the way to a new generation of switches with optimized mix between optical and electronic processing functions. They should be operating over multiple wavelength bands and spatial dimensions and have a smart network fabric relying on software programmability and slicing, addressing multiple protocol layers and network domains. This applies also to intra-DC applications, where new switching concepts mixing optical and electronic switching technologies could lead to higher performance and lower power consumption. In addition to ultra-fast switching speeds, the capability of switching on different levels of granularity and a high overall switching throughput, future switching architectures need to take the energy-efficiency of the switches itself, but also that of the network they are supporting, into account. Another topic of interest would be disaggregated switching platforms in comparison to purpose build solutions.

### 7.4.1 Ultra-fast Multi-granular Switching Nodes

Flexgrid technology on the optical layer, the utilization of new wavelength bands (beyond C+L+S band) and the advent of multi-core/multi-mode transmission will require new multi-granular switching node architectures. This will allow an even more flexible network slicing in the wavelength as well as in the spatial domain. An operation over multiple wavelengths, wavelength bands and spatial dimensions requires new switch and transponder architectures that have not been discussed in great detail yet. Some applications may require network resources only for a very short time. Consequently, approaches enabling a faster reconfiguration (< 1 ms) on the optical layer and taking into account concerns such as amplifier power transients need to be developed.

### 7.4.2 Switching Architectures guided by Energy-Efficiency

Future networks will face the need to reduce their energy consumption. Therefore, switching architectures need to take energy-efficiency into account at a very early stage and on all network levels. On the hardware level, switching architectures with an intelligent mix between optical and electrical switching functions, will be required. In addition to a power reduction in the switching components itself. On the control & management layer, the switching needs to be optimized to perform switching functions in the domain, electrical or optical, with the lowest power consumption. In that respect the switching operations could benefit from the larger degree of freedom in multigranular switching architectures (wavelength, waveband and space).

### 7.4.3 Research Challenges

The research challenges from the previous subsection are summarized below:

| **Research Theme** | **New switching paradigms** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Multi-granular Switch allowing to switch between space, wavebands and wavelengths | Mid-term (finished in 5y) | Operation over S+C+L-band (1460 – 1625 nm)<br>Operation with multicore fiber<br>Switching granularity from 50 GHz (single wavelength), over 5-8 THz (waveband switching) to 21 THz (complete fiber/core switching) | Deterministic networking<br>High-capacity scaling and reliable connectivity |
| Switches with fast reconfiguration times | Mid-term (finished in 5y) | < 1ms | Full network programmability |

| | | | |
|---|---|---|---|
| Switching architectures with optimized mix between optics and electronics for energy-efficient networks | Long-term (finished in 7y+) | | Cost-effective and energy-efficient systems |

### 7.4.4 Recommendations for Actions

| Research Theme | New switching paradigms | |
|---|---|---|
| Action | Ultra-fast Multi-granular Switching Nodes | Switching Architectures guided by Energy-Efficiency |
| *International Calls* | To leverage industry and academic efforts vertically for critical mass | To federate industry and academic efforts across Europe for design and manufacturing of challenging components and systems |
| *International Research* | To leverage industry and academic efforts vertically for critical mass | To federate industry and academic efforts across Europe for design and manufacturing of challenging components and systems |

## 7.5 Deterministic networking

While today's Internet is built on a best-effort traffic paradigm, an increasing number of applications require reliable end-to-end transmission with guaranteed throughput and bounded latency. Examples range from RAN transport over vehicular/robotic/ industrial control to augmented/virtual/extended reality. Stringent network requirements also result from accurate positioning, navigation and timing (PNT) services. Future human-centric use cases such as the "Internet of senses" and holographic communications are expected to increase the demand for deterministic network behaviour even more. Application requirements are diverse and therefore a flexible quality-of-service (QoS) framework is necessary: A control application for instance only consumes low bandwidth but needs high reliability and relies on definite time window for packets arrival. In turn, an extended reality applications may be able to tolerate more timing variations, yet high bandwidth is mandatory for good user experience.

While mechanisms exist to control throughput, latency, jitter and packet loss in packet-optical networks, they often provide statistical QoS only and do not guarantee a deterministic network behavior. Available timing signals rarely offer the necessary accuracy and/or reliability to allow precise time synchronization for mission critical applications. End-to-end services are often delivered over heterogenous network infrastructure in which different QoS and traffic-engineering mechanisms are employed and need to interwork which each other. Fundamental architectural questions such as where to control traffic, which data plane support is necessary, and how to facilitate end-to-end service assurance need to be answered in light of a diverse set of application requirements. Novel solutions are needed which trade-off performance improvements against scalability limitations and implementation complexity. Activities should leverage technologies from standards bodies such as ITU-T (e.g. OTN, PON), IEEE (e.g. TSN, 1588, EPON), and IETF (e.g. DetNet, L4S), OIF (e.g. FlexE), address deficiencies and develop novel solutions as extensions.

### 7.5.1 Resilient solutions for high-precision, network-assisted timing distribution

Precise timing information is not only necessary to operate communication networks, it is also an enabler for the digital transformation of critical infrastructures. Real-time control, positioning and

navigation but also accurate event recording and threat mitigation rely on the availability of precise time information. High-precision time distribution becomes an additional service delivered by a new generation of smart networks.

Cost-effective and scalable time distribution solutions are required which can operate over a heterogeneous network infrastructure and are robust against failures and attacks. For high reliability, a resilient timing network combining information from multiple reference sources and offering sufficiently long local hold-over capabilities is necessary. Hardware assistance is required for high-time resolution and low timing error. Pluggable or embedded time synchronisation modules can provide precise timing capabilities to network or user equipment not possessing such capabilities by default. A control, telemetry and analytics framework is required to deliver timing services, assess their quality and take corrective actions where needed.

### 7.5.2 Reliable data & control plane solutions for deterministic network services

The performance of packet-optical networks is crucially dependent on packet processing and traffic management functions such as shaping and queuing. Deterministic services need to be given preferential treatment without burdening the network with overly complex processing for lower priority services. Flexible data plane and control plane solutions are required which can cope with changing traffic patterns and a variable traffic mix. Architectural trade-offs are needed to avoid network inefficiencies on one hand and insufficient performance on the other hand. Mechanisms to apply back-pressure and to communicate between network and client equipment can help to avoid a QoS deterioration by mitigating network overload conditions.

Some of the most challenging requirements are driven by mobile fronthaul applications and comprise <100µs latency[7], <8ns relative timing error, and several tens of Gb/s throughput. Data plane optimizations are necessary to fulfill the demanding latency and timing requirements. Applications in which packets have to arrive in a certain time window need further research, especially if such services have to be delivered over large networks or a heterogenous infrastructure which is only partly timing-aware. If deterministic network services serve mission critical applications, measures need to be taken to protect these services against outages, rogue device behavior, and attacks by malicious actors.

### 7.5.3 Tools for service assurance in deterministic networks

The performance of deterministic network services can strongly depend on network size, number of nodes, network traffic as well as on the characteristics of the used network equipment and applied control strategies. Planning tools are required to estimate the attainable service performance and determine an optimized network configuration. Methods of network calculus, heuristics and information about the internal structure of network equipment can be used for this step. Information from network planning but also network telemetry and equipment status (e.g. queue fill levels) can then be leveraged to get a current view of the network and service quality, forecast future evolution, and make adjustments where necessary. Network analytics and machine learning can help to accurately predict the network behavior. A digital network twin approach may be used to test changes before deploying them in an operational environment.

[7] including fibre transmission which adds 5µs/km

### 7.5.4 Research Challenges

The research challenges from the previous subsection are summarized below:

| Research Theme | Deterministic networking | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Resilient network-assisted time distribution | Short-term (finished in 3y) | Accurate time-of-day delivery to network endpoints and user devices<br>Resilience against GPS/GNSS jamming and other attacks, indoor performance w/o GPS/GNSS reception<br>KPI values are application specific | Deterministic networking<br><br>Secure communicatons |
| Reliable data & control plane solutions for deterministic network services | Mid-term (finished in 5y) | Reliable transport of deterministic traffic alongside lower priority traffic<br>Guaranteed throughput, bounded latency, high availability<br>KPI values are application specific | Deterministic networking<br>Network multi-tenancy<br>High-capacity scaling and reliable connectivity |
| Tools for service assurance in deterministic networks | Mid-term (finished in 5y) | Performance prediction of deterministic network services in complex networks, assessment of network optimizations on digital twin<br>Forecast accuracy, network utilization, predicted versus measures service quality | Network and service automation |

### 7.5.5 Recommendations for Actions

| Research Theme | Deterministic networking | | |
|---|---|---|---|
| Action | Resilient network-assisted timing distribution | Reliable data & control plane solutions for deterministic services | Tools for service assurance in deterministic networks |
| *International Calls* | | EU-Japan collaboration | |
| *International Research* | EU collaboration, worldwide standards | EU collaboration, worldwide standards | EU collaboration, worldwide standards |
| *Open Data* | | | Network statistics |
| *Large Trials* | | Scalability tests in relevant environment(s) | Scalability tests in relevant environment(s) |
| *Cross-domain research* | Collaboration with vertical industries (manufacturing, multimedia, …) | Collaboration with vertical industries (manufacturing, multimedia, …) | Collaboration with vertical industries (manufacturing, multimedia, …) |

## 7.6 Optical technologies for radio networks and systems

The expected tenfold increase of traffic growth and the tight latency constraints dictated by new 6G services will require a substantial evolution not only of the RAN but also of the architecture and the technology of the underlying mobile transport network. Optics is an enabler for 6G not only as regards the new mobile transport network, but novel optical Interconnect technologies will play a key role in future advanced antenna systems, impacting their architectures. Finally, new advances in photonic integration open new opportunities to apply suitable combinations of optical, radiofrequency, and digital electronics to radio systems. These three application macro-areas lead to a wide set of new challenges for optical technologies in radio access, as illustrated in the next sections.

### 7.6.1 Optical technologies for radio access networks

Optical technologies are well known for their high bandwidth, transparency to the format of the carried signal and low latency and play a role as important as packet networking in this evolution of the mobile transport networks. However, they need to evolve in parallel to it to achieve a further level of energy efficiency, miniaturization, cost effectiveness and fast reconfigurability. Among optical technologies, wavelength division multiplexing (WDM) is the most promising one as regards supported bandwidth, distance (a prerequisite for centralization) and compatibility with several network topologies (ring, mesh, point-to-multipoint) but it is also the one where the technology evolution needs to be more radical. The current cost figures of essential WDM components (high speed transceivers, wavelength switches, etc.) is at least one order of magnitude higher than the ideal cost figure (<0.1 $/Gbit/s) for an access network. Moreover, in this network segment, the link attenuation is significant due to the presence of optical add-drop nodes and passive splitters, and optical amplifiers cannot be used in a challenging environment like an outdoor radio unit placed at the top of an antenna pole. Finally, fibre's chromatic dispersion is very significant when moving at high bit rates, even over short distances.

This picture calls for the following research and innovation challenges:

1) A first challenge is the development of high speed WDM interfaces (from 100G over 400G to 1.6T) at a much lower cost than today. This can be achieved in two ways: making cheaper and energy efficient the coherent optical interfaces or extending the domain of direct detection systems to higher capacities. Both the paths could exploit new ML-based DSP for signal equalization, line coding, pulse shaping, polarization recovery, etc. Moving into the optical domain functions usually performed by electronic ICs is today possible, thanks to the advent of integrated photonics, and could improve the energy efficiency.
2) A second challenge is enabling longer distance and higher attenuations, which is essential in CRAN. Small footprint optical amplifiers in a pluggable format or as a function provided in a PIC and able to work in a harsh environment will be an important enabler. Radically reducing cost and getting low power, industrial temperature, zero-touch operation, and high loss budget for moderate capacities are the key features to develop.
3) But high capacity and distance are not all: the 6G network will be a dynamic network, ideally capable to provision every type of service on demand and in real time. So, a third challenge is ensuring dynamicity to the underlying transport network to avoid cost and waste of energy for bandwidth overprovisioning. Dynamic bandwidth allocation is something packet networks are very good at doing but energy consumption and latency may unacceptably grow in presence of congested traffic. Optics may be used for traffic offload, but it is basically

static. Reconfigurable optical components exist (tuneable lasers, tuneable filters, wavelength selective switches, optical switches) but, despite many research efforts never achieved sufficient cost effectiveness or TRL for mobile transport applications. Innovation Actions to accelerate the time to market of fast configurable (nanoseconds) optical subsystem could help in this sense.

### 7.6.2 High speed optical interconnects in radio systems

The traditional fronthaul links carrying digitized samples of the signal on air are very bandwidth inefficient and in 6G would lead to an explosion of capacity. A solution is moving back to the antenna digital processing functions previously performed at the baseband unit (BBU). This reduces the fronthaul bandwidth but requires high interconnection capacity between the digital integrated circuits (ICs) and the radio frequency ICs (RFICs) in the radio unit (RU). In extreme scenarios, the total throughput can be hundreds of Terabit/s.

This leads to the research challenge to provide high bandwidth density (hundreds Gbps/mm2) optical interconnects, in presence of these constraints:

- Energy efficiency of the order of 1 pJ/bit. Power dissipation is critical in radio systems, mostly relying on passive cooling.
- Low latency (a few nanoseconds), including FEC and bits-to-symbols mapping.
- Low Bit Error Rate, <10−15.
- Wide temperature Range. Since an antenna operating outdoor is not actively cooled, the internal operating temperature can be > 100 ◦C.

This challenge calls for developing new or evolving emerging optical interconnect technologies, like

- Co-packaged optical (CPO) transceivers, i.e., optical transceivers mounted on the same substrate of the IC they are connected to)
- Monolithically integrated optical transceivers, where optical front-end and electronics are integrated on the same chip.
- Passive optical routing solutions (e.g., Si on glass interposer) to mitigate the high loss, power consumption, signal degradation and electromagnetic interference issues electrical interconnects suffer from at high bit rates.

### 7.6.3 Optically enabled radio functions

Transport networks and high-speed interconnects are two sweet spots for optical technologies but new advances in photonic integration open new opportunities to apply suitable combinations of optical, radiofrequency, and digital electronics to radio systems.

A first related research challenge is the low phase noise (PN) generation of frequency references for clock and radio frequency. Electronic generations schemes are not accurate enough as the frequency rises up to the THz range but also the traditional optical schemes based on the heterodyning of two independent lasers in a photodiode suffer from the frequency and linewidth instability of the laser, making them unsuitable for radio applications. New solutions will require phase-locking of the beating sources, as in Mode-Locked Lasers (MLL). However, state-of-the-art MLL are not compatible with radio applications due to the high cost, power consumption, and large footprint so that developing new integrated photonic components will be necessary. Photonic integrated optoelectronic oscillator is an alternative to MLLs to generate stable low PN radio

carriers. They rely on modulating the light's intensity from a laser and feeding back the detected RF modulation to the modulator's input port. A third approach modulates a Continuous Wave (CW) laser with a low frequency carrier signal through a non-linear optical modulator, so generating multiple frequencies of the carriers. The main challenge with this method is achieving sufficient conversion efficiency. Independent of the frequency generation method, the possibility of selecting different laser modes shall be provided for the flexible generation of RF carriers with tunable frequency.

A second research challenge is increasing the density of antenna elements per surface area, as required by massive MIMO systems at very high radio frequencies. This requires integrated photonic solutions for the distribution of optical signals inside the antenna system, as optical waveguides embedded in the Printed Circuit Board (PCB) and glass interposers. With optical waveguides in the PCB, the signal is distributed by means an optical layer below the antenna elements and then routed vertically to PD near the element. A glass interposer could be instead placed on top of the RFICs, so that the optical signal is routed down vertically to the r PD. These solutions his requires the co-packaging of RFIC, TIA, PD and optical waveguides, which is a formidable technology challenge since these components are based on very different technologies.

A third radio-related challenge where optical technologies already demonstrated their potential is beamforming. Current solutions enable squint-free, high pointing accuracy beamforming but suffers from high insertion loss, proportional to the number of AEs and photonic phase shifters or true time delay elements. Integrated optical amplification, either based on III-V integration or rare earths doping of silicon waveguides could mitigate the issue but should not negatively affect the phase noise of the system.

### 7.6.4 Research Challenges

| **Research Theme** | **Optical technologies for radio networks and systems** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| **High speed WDM optical transmission in RAN** (from 100G over 400G to 1.6T) 10 times more cost effective than today. Both coherent and direct detection are in scope. Optical processing and ML-based DSP are among the enabling technologies. | Mid-Term | Energy and cost efficient digital and optical processing schemes for coherent optical interfaces | High-capacity scaling and reliable connectivity.<br><br>Main application is the 6G mobile transport network, to avoid capacity bottlenecks |
| **Cost-effective optically amplified networks** having low power consumption, Industrial temperature operation capability and zero-touch operation as the key features. | Short-Term<br>Mid-Term | Cost efficient, small form factor and plug & play optical amplifiers | Access to everything<br><br>Enabler for Cloud Ran deployments over high distances, where opex savings and high coordination is ensured by deeply centralized processing functions. |

| | | | |
|---|---|---|---|
| **Fast reconfigurable optical RAN** based on tuneable lasers, tuneable filters, wavelength selective switches, optical switches) for RAN. | Short-Term | High-TRL fast-reconfigurable integrated optical switches and tuneable lasers | Full network programmability<br><br>Enabler for optimizing the bandwidth resources in packet RANs where with highly variable traffic load |
| **High bandwidth density (hundreds Gbps/mm2) optical interconnection systems** with 1 pJ/bit energy efficiency, ns latency, <10-15 BER, high temperature operation (> 100 ◦C.). | Mid-Term | Highly energy efficient, high-performance solutions for co-packaged optical transceivers, monolithically integrated optical transceivers and passive optical routing solutions. | Advanced electro-photonic integration.<br>High-capacity scaling and reliable connectivity and Cost-effective and energy-efficient systems<br><br>Application area is 6G RAN equipment (remote units, digital units, x-haul switches) where high capacity must be provided in small space and power dissipation is crucial. |
| Tuneable **low phase noise (PN) generation of frequency references** for clock and radio frequency. up to the THz range based on integrated photonics schemes. | Mid-Term<br>Long-Term | Integrated photonic based solution for generation of high radio frequency with extremely high accuracy | Efficient Integration of optical technologies for radio access network<br><br>Enabler for 6G when moving to sub-THz frequencies |
| High-bandwidth density solutions for the distribution of optical signals inside an antenna system, | Long Term | Solutions for the distribution of optical signals in an antenna system based on the co-packaging of heterogeneous technologies (RFIC, TIA, PD, optical waveguides, etc.) in the same device | Advanced electro-photonic integration<br>Efficient Integration of optical technologies for radio access network<br><br>Enabler for 6G when moving to sub-THz frequencies, to guarantee high performance (low noise, high output power, high EMF immunity) |
| MIMO systems based on low noise optical amplification | Long Term | Integrated optical amplification to provide high Tx or Rx in antenna systems based on optical generation and distribution of radio signals | Efficient Integration of optical technologies for radio access network<br><br>Enabler for 6G when moving to sub-THz frequencies, to guarantee high performance (low noise, high output power,) |

### 7.6.5 Recommendations for Actions

| Research Theme | Optical technologies for radio networks and systems | | | | | | |
|---|---|---|---|---|---|---|---|
| Action | High speed WDM optical transmission in RAN | Cost-effective optically amplified networks | Fast reconfigurable optical RAN | High bandwidth density optical interconnection systems | Low phase noise generation of references | Distribution of optical signals inside an antenna system | MIMO systems based on low noise optical amplification |
| *International Calls* | Having a shared program with non-EU based research and market leaders in coherent optics and ICs | Having a shared program with non-EU based research and market leaders in optical amplification | | Involving US market leaders in co-packaged optics | | Need to involve foundries that can integrate PICs and electronic ICs with a mature and reliable process | |
| *International Research* | Having shared program with non-EU based research and market leaders in coherent optics and ICs | Having a shared program with non-EU based research and market leaders in optical amplification | | Involving US market leaders in co-packaged optics | | Need to involve foundries that can integrate PICs and electronic ICs with a mature and reliable process | |
| *Large Trials* | Trials with leading EU mobile operators in their RAN | Trials with leading EU mobile operators in their CRAN | Need to demonstrate production in large scale and high reliability | | | | |
| *Cross-domain research* | | | | | It requires skills in both optics and radio systems, a combination that seldom | It requires to involve different expertise: photonic systems, | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | engineers have | experts, radio designers, PIC and IC technology experts | |

## 7.7 Optical network automation

Optical network automation is key to achieve operators' business goals and in supporting new complex services. Aspects related to automation must be developed in the areas of service deployment, network planning and overall network operation. Initial research should be focused in automating repetitive, error-prone tasks or tasks with very well-established workflows, with applications in single domain scenarios, such as service activation (aiming at OpEx reductions due to more efficient workflows and considerably reduced execution times), increased flexibility in offering services, better service level agreement performance, and faster issue resolution. Automation is critical in optical networks supporting increasing data rates given, for example, the complexity of modelling of physical impairments, or the large number of parameters and their interdependencies. Outcomes related to automation in single domains shall form the basis for more ambitious cross-domain automation (across technology layers or network segments). AI/ML solutions in support of network operations should be further developed beyond policy- / expert-/rule- based systems, and control and orchestration architectures should become increasingly modular, leveraging the flexibility of deployment in hybrid clouds while relying on open and standard data models, protocols, interfaces, and frameworks (including, for example, proven and mature open-source projects and initiatives).

### 7.7.1 Network Telemetry and Optical Network Sensing

This aspect should address activities related to optical monitoring, network streaming telemetry and overall secure and efficient data collection, storage, and subsequent use, with applicability and focus on *large-scale scenarios.* Telemetry systems should allow maximum flexibility, including *at-origin* or *intermediate* filtering and aggregation of data. Research activities range from the definition and subsequent standardization of data models for the telemetry data; the definition of efficient and secure protocols, in terms of latency and encoding to the definition of architectures in support of overall infrastructure monitoring and telemetry. Telemetry should be enabled at a device or system level as well as at the domain or network level. Telemetry systems should unify aspects related to network state synchronization, alarm reporting (incl. threshold crossing alerts), performance monitoring and fault management. This aspect also encompasses research on *Network Sensing* (the use of network and computing/storage infrastructure to design solutions enabling the detection of events of interest and related applications). This includes, in particular, *optical fibre sensing*, for applications like intrusion detection and prevention, repairing network outages towards self-healing networks, and the systematic use of such sensing techniques – coexisting with actual user traffic -- at scale to predict and pinpoint physical layer issues along with software automation, design, and operational tools to mitigate those issues. The aspect should address applications, services, technologies, and challenges/benefits of network sensing.

### 7.7.2 Control and Orchestration architectures for Network Automation

This aspect covers research activities related to the definition of control and orchestration architectures for heterogeneous multi-layer, multi-domain, or multi-technology scenarios, addressing the shortcomings of current (e.g., SDN-based) approaches such as scalability, reliability, or deployment agility, as well as the improved support of emerging cases such as infrastructure sharing. The architectures should exploit and enable multi-tenancy in a more cloudified environment, while empowering users with the capability to manage their own services, while full leveraging network slicing. In a short-term, control and orchestration systems should rely on open data models and interfaces, suitable extended for recent development such as multi-band networking, space division multiplexing or improved support for physical impairment modelling in view of beyond 100G systems. Novel or refined architectures should be investigated, including full support of service lifecycle management, integration, and migration of current operators OSS/BSS systems, encompass telemetry and network sensing and apply to heterogeneous environments, such as integrating wired/wireless access networks (i.e., PON/RAN) and transport segments in over-arching control systems. Enable data export and the use of 1st party and 3rd party systems (for resource allocation, path computation, AI/ML systems), including dynamic algorithms and heuristics enabling almost real-time end-to-end quality of experience with solutions to minimize end to end delay/latency and jitter and performing coordinated resource allocation. Further research is needed in the integration of packet and optical networks, ensuring efficient and low-latency service provisioning. In a medium-term, research should address how to enable or adopt the application of IT and DevOps principles for network control and management, further leveraging tools for process automation, data visualization. Support higher abstractions in terms of service definitions, further exploiting and refining intent-based approaches. In the long-term, support truly autonomous networks by integrating monitoring, telemetry and 1st party/ 3rd party AI/ML assisted network operation closed-loops, with dynamic instantiation of customized control and management systems in hybrid clouds in support or consolidated business models of infrastructure sharing and multi-tenancy. Research on the applicability and role of analog computing for network operation with extremely reduced power consumption.

### 7.7.3 AI/ML in support of Network Operation

This aspect encompasses the use of AI/ML-mechanisms in support of network operation, e.g., service and infrastructure management, both in single domains and cooperatively across different domains. This includes: i) the definition of ML models, the use of training data sets; applicability and reusability of existing/previously used models; ii) development of use cases and scenarios involving general resource allocation, function placement (for example, selection of functional splits based on multi-objective problem formulation and dynamic traffic patterns) and iii) Research on distributed self-management control infrastructures based on multi-agent systems able to autonomously coordinate resources near real-time for end-to-end service assurance

### 7.7.4 Reliability and Security of Control, Orchestration and Management

This research aspect covers activities in support of the reliability and security of the control, orchestration and management functional elements, systems, and devices, considered as critical infrastructure systems. It should be formulated in terms of novel architectures, interfaces, protocols, relevant data security, and overall system integrity. Also related to Research Aspect 4, this includes enablers regarding the usage of the underlying infrastructure (network sensing) for

intrusion detection and prevention, to leverage the safety and reliability of network infrastructures. This aspect also should address research on the analysis of attack methodologies against network automation systems (ML-based or not), identification of their vulnerabilities and the definition of mitigation and/or defence countermeasures, or the use of distributed ledger architectures in support of network control and service management, especially in multi-actor scenarios as well as designs addressing security considerations including privacy, avoiding data exfiltration and leakage.

### 7.7.5 Optical Network Digital Twin

This aspect covers the use of digital twins for optical networks. This includes the definition of new use cases and applicability statements of the digital twin concept (including, but not limited, to soft-failure/anomaly detection, localization, identification; dynamic operation and rollback of state changes, root cause analysis, discrete event emulation); mechanisms for state synchronization between the physical entity (e.g., optical device, link, node, network) and the digital twin; techno-economic studies related to savings associated to the used of digital twins.

### 7.7.6 Research Challenges

In view of the research aspects, the next table summarizes key research challenges:

| **Research Theme** | **Optical network automation** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Large Scale Telemetry and Efficient and Reliable Network Sensing Leverage operational and instrumental data from devices such as transceivers or ROADMs, while supporting: i) an increasingly high number of devices and ii) low-latency closed-loop systems. Open Data Repositories Enabling controlled data exchange and common repositories from data from multiple sources (e.g., network segments) Related to Research aspects 1, 3, 4 | Short-Term<br><br>Mid-Term | Optical Telemetry systems and platforms, with data models, including scalability analysis. Applications and solutions related to the use of Network Sensing **Properties and KPIs: a**ble to manage large scale systems with aggregated Terabit/s and Petabit/s telemetry data. Support hundreds of thousands of optical devices. | **Requirements:** Full network programmability and network and service automation<br><br>Contributions/value: Civil engineering Distributed Sensing Health care applications Network operation Failure localization |
| **System Interoperability** enabling modular systems and avoiding lock-in; Exploit increasingly complex device/service programmability Addressing technologies such as multi-band, SDM and improved physical impairment modelling. new devices and extended capabilities. Fast Service Creation and Modification with efficient Packet/Optical integration and Networking Leveraging new pluggable/co-packaged interfaces in hybrid scenarios with low latency requirements. Optical bypasses and | Mid-Term | Open and Standard Data models (e.g., with YANG) for new devices or extended capabilities for 400G and beyond. Common protocols and frameworks. Promote reuse and adapting industry best practices. Applicability analysis/trade-offs in terms of speed, complexity, efficiency in resource allocation and function placement. Improved hardware designs and system control for reduced | **Requirements:** High-Capacity Scaling, Cost-effective and energy-efficient. Full network programmability and network and service automation Contributions/value: Increased interoperability with reduced vendor lock-in. Increased Efficient use of optical spectrum with more advanced algorithms for resource |

| | | | |
|---|---|---|---|
| cost efficiency. Dynamic service provisioning and reconfiguration based on a set of constraints. Leverage programmability of optical devices with operating modes and parameters for increased efficiency. Related to Research aspects 1, 2, 3 | | optical device reconfiguration latency. **KPIs:** Adoption by SDOs, industry actors and reference implementations. Service Creation O(seconds); Target device reconfiguration operations O(100ms), depending on device complexity and reduce service provisioning time by an order of magnitude compared to manual operation (90% reduction). | allocation and function placement. Reduced CapEx due to a more competitive market. Reduced OpEx due to automation. Increased user satisfaction and reduced service activation times. Agile operations and reduced OpEx. |
| Support new services related to **infrastructure sharing enable Multitenancy** in a cloudified environment, while empowering customers with the capability to manage their own services, consolidate network slicing. Related to Research aspects 1, 2, 4. | Mid-Term | New architectures in support of optical infrastructure sharing | **Requirements:** Cost-effective. Full network programmability and network and service automation. Network Multitenancy |
| **Agile operation** Application of DevOps principles and IT practices for network management. This includes cloudification of OSS/BSS, adoption of open-source projects and frameworks, application of Continuous Delivery/Continuous Integration, Improved workflow automation, network optimization and development of an ecosystem of suitable automation applications. Unified short-term provisioning and long-term network-planning, with closed loops at different timescales Related to Research aspects 1, 2 | Mid-Term, Long-Term | Novel network automation application ecosystem. Common software frameworks for unified short-term provisioning and long-term planning and dimensioning. Open interfaces for 1st party and 3rd party applications (e.g., resource allocation and function placement) | **Requirements:** Cost-effective. Full network programmability and network and service automation. Network Multitenancy<br><br>New markets and business opportunities related to specialized services in support of network automation. |

| | | | |
|---|---|---|---|
| **Network Domain Automation** via i) AI/ML assisted decision-making processes and issuing recommendations and ii) improved resource allocation and function placement algorithms.<br>Network Domain Automation via AI/ML with Direct Control, truly autonomous networks<br>Network Automation via AI/ML in Cross-domain settings addressing challenges related to trust, security and optimality.<br>Predict and/or replicate network behaviour based on potential events and actuations<br>Related to Research aspects<br>1,2, 3,4, 5 | Short-Term<br><br>Mid-Term<br><br>Long-Term | Recommendation based / Direct Control Closed-Loop network automation.<br>Digital Twin Implementations for Optical Networks and Systems with different levels of abstraction, modularity, and reusability.<br>**KPI:** Increased resource efficiency, reduced blocking probability or improved energy efficiency; >25% of OpEx savings compared to manual operation; Reduced time to deploy services in cross-domain scenarios over an order of magnitude shorter than current static practices. Improved prediction of network outages and service impact; Reduced rate of reconfiguration errors; More efficient network planning and capacity upgrades | **Requirements:** Cost-effective. Full network programmability and network and service automation. Network Multitenancy<br><br>**Contributions/value:** Increased resource efficiency, reduced blocking probability or improved energy efficiency. Efficient network planning and optimization. Seamless and optimum capacity upgrades, including migrating scenarios. |
| Secure Control Systems<br>Address security requirements incl: privacy, preventing data exfiltration, leakage, functionality bypass, spoofing, or isolation violations<br>Related to Research aspects<br>1,2,3,4,5 | Short-Term & Mid-Term | New architectures and functional requirements for software systems.<br>Development of new protocols and interfaces addressing security requirements. Applicability statement and assessment of QKD systems for critical applications<br>**KPI:** Reduced rate of security-related incidents and lower implications | **Requirements:** Secure communications, resilience.<br>**Contributions/value:** Important implications in the overall design of control and orchestration systems. More secure systems and increased robustness against attacks.<br>Increased confidence of end users. Large impact on network operation (direct) and end users (indirect) |

### 7.7.7 Recommendations for Actions

| **Research Theme** | **Optical network automation** | | | | |
|---|---|---|---|---|---|
| Action | Telemetry and Sensing | Architectures, Data Models | AI/ML for Netw. Op. | Reliability & Security | Opt. DT |
| *International Calls* | X | X | | X | X |

| | | | | | |
|---|---|---|---|---|---|
| *International Research* | Promote the development of telemetry platforms enabling sharing of data is challenging in a world-wide setting. | Align research program control and management architectures across SDOs<br>Japan has research groups on optical networking.<br>US several projects and initiatives are US based (TIP, ONF) | | Japan has developed programs targeting reliability post 2011 events. | US has a solid experience on Digital Twin. The concept was initially developed at NASA. |
| *Open Data* | Promote the sharing of (anonymized) telemetry data from operators and relevant sources | | Promote the sharing of both ML models and ML datasets, enabling reuse and robust/faster training | | Promote the design of Open Digital Twins, in a comparable way to Open-Source Software or Open-Hardware |
| *Large Trials* | Network/Fiber sensing to be demonstrated in operators' networks in environments close to production with coexisting traffic | | Need to test AI/ML systems at a large scale, and to evaluate the required telemetry systems | | Large Scale DT need to be developed and assessed in terms of e.g. computation requirements. |
| *Cross-domain research* | | | Leverage AI/ML expertise from other domains | Strong inter-dependencies with Security Experts and Cryptography | Leverage the use of Digital Twins in other domains |

## 7.8 Security for mission critical services

The ever-increasing interconnectedness not only of people but also of devices starting from huge power plants down to billions of IoT devices like sensors or appliances does not only increase the dependence on the network infrastructure but also expand the threat surface and therefore the vulnerability of every individual and of the society as a whole. Important threats do not only include hacking and espionage, but also network outages due to natural catastrophes as well as terrorism and sabotage targeting critical infrastructure. Therefore, it is getting more important to better safeguard our network infrastructure against data leakage and unexpected service outages.

The higher flexibility of optical networks, enabled through software-controlled network elements (software defined networking, SDN), also increases the vulnerability of such networks to various

kind of attacks and therefore security and resilience aspects need to be part of the concepts from the beginning (including both the hardware and software layers of the network). More generally, the design of network equipment needs to employ modern security and reliability paradigms (security by design) and apply state-of-the-art software technology to foster efficient and secure implementation of increasingly complex network elements.

### 7.8.1 Quantum-safe cryptography

A signal on an optical fiber can be tapped once the physical access to the fiber is available. At this point, the data of millions of users and billions of applications is exposed to theft and manipulation. Therefore, authenticity, privacy and data integrity are essential and need to be kept at a level playing field with increasing threat scenarios, e.g., by allowing for crypto-agility. Improvements need to consider quantum-safe solutions for authenticating the communication partners, for protecting the data against tampering and for exchanging secret keys by employing post-quantum cryptography or secure quantum communication, e.g., quantum key distribution.

### 7.8.2 Physical layer security

Physical layer security aims at providing alternatives to algorithmic based encryption, key-exchange and authentication. The topic is already an established research area in wireless communications. Examples are private transmission without keys, deriving secure keys from channel properties or physical unclonable functions (PUF). In many cases physical layer security primitives will be used in addition to the established algorithmic protocols since their security is based on different mechanisms and also possible attacks are very different. Hardware-based authentication mechanisms (e.g., PUF) can enable zero-trust communication and improve the security of modular systems and networks. In addition, such security anchors can be used to prevent product piracy.

### 7.8.3 Network resilience

Adding redundancy is the conventional, but also expensive way to improve the reliability and resilience of networks. System concepts for low cost and low power implementation of redundancy solutions (e.g., using high radix optical switches) should be studied. Alternative concepts, that are high on the research agenda today, are increased flexibility, massive monitoring and software control of optical networks. It should be possible to employ this functionality beyond the borders of a single networking domain. Also, the monitoring of optical distribution networks like passive optical networks should be improved, which is especially difficult due to their passive implementation.

### 7.8.4 Intrusion detection and mitigation

Based on the data generated by the monitoring solutions, data processing (e.g., by means of ML methods) can help to detect upcoming or hidden problems early and counteract them in advance with the available flexibility. While machine learning often has a competitive edge over conventional algorithms, in most cases it is not clear how a certain ML method gets its results (ML as a ‘black box’). This might lead to unexpected behaviour (e.g., if the input data is not within the standard range), and could be used to fool an ML implementation (c.f., adversarial ML). Secure usage of ML requires exhaustive testing and ideally some degree of explainability, i.e., some insight into the inner working of the algorithm.

### 7.8.5 Research Challenges

| Research Theme | Security for mission critical services | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Long distance QKD<br>QKD devices for Telco environments with sufficient secure key rates and an attenuation budget considerably beyond current limitations. | Mid-term | KPIs: Attenuation budget, secure key rate | Secure communications. Data privacy and improved security for critical infrastructure |
| HW-based authentication<br>Concepts for secure authentication and device authenticity | Mid-term | Optical component or macro to include into PIC<br>Improved entropy vs. commercial solutions | Secure communications. Data privacy and improved security for critical infrastructure |
| AI/ML based intrusion detection<br>Methods to enable secure and robust intrusion/anomaly detection | Mid-term | Truly reliable and secure AI/ML algorithms and concepts to ensure this | Secure communications. Data privacy and improved security for critical infrastructure |

### 7.8.6 Recommendations for Actions

| Research Theme | Security for mission critical services | | |
|---|---|---|---|
| *Action* | Quantum-safe cryptography | Physical layer security | Intrusion detection |
| *International Calls* | X | **X** | X |
| *International Research* | Leverage knowledge in European but non-EU countries | Cryptography is a truly international endeavour | Benefit from strong AI/ML footprint in the US |
| *Open Data* | | | Open training and test data can be essential to progress the community |
| *Cross-domain research* | Strong interdependencies with classical security | Strong interdependence with classical security | Leverage knowledge from strong AI community |

## 7.9 Ultra-high energy efficiency

Communications networks are a pillar of our daily life. It enables to connect between people and machines, providing an enormous range of services.

The ongoing pandemic has strengthened the need for data connectivity relying on higher capacities and forecasts agree that the traffic growth will further intensify, at an annual rate well above 20% that has to be matched by the capacity of the optical transceivers and line systems. At the same time, requirements for reduced latencies in data processing as well as higher energy-efficiencies, become more stringent. Especially the focus on greener networks and processing systems, becomes a top priority for our society. Currently the ICT sector accounts for 5-9% of electricity use and around

3% of the global greenhouse emissions8. The EC estimates that at this pace, the ICT footprint could increase to 14% of global emissions by 20409,10. Hence, solutions to make sustainable both power consumption and footprint of world-network infrastructures are mandatory. As fibre optics and photonic devices are the key technologies underlying the worldwide telecommunication infrastructure, it will play a highly relevant role in reducing the total power consumption of the ICT. A matter that has been recently highlighted also by the activity of the F5G within the ETSI on green energy. In this context, some key questions arise:

1. How to reduce the overall network energy consumption while increasing capacity?
2. How can photonic integration help to reduce energy consumption? How to replace bulk-optics with PICs for subsystems/functionalities like WSSs/spectral-lanes switching or to replace power consuming electronic processing?
3. How can network controllers be used to reduce energy-critical network functions?
4. Which functions are energy-critical?
5. How to make next-generation networks sustainable and multi-generational so that electronic waste is drastically reduced?
6. Can AI-based network management help to reduce energy consumption?
7. How to optimize service placement in the networks?
8. How to optimize the processing and amount of information, to avoid this being carried out every time and everywhere?
9. What is the impact of distributed compute & storage resources for low latency services on network energy consumption?

The following areas of research might help to answer the above questions

### 7.9.1 Simplified and fully configurable flexible E2E optical networks

Optical networks consist of a variety of domains and segments, which are traditionally being individually optimized using specific solutions. This might be cost-effective locally, but not in an end-to-end (E2E) fashion across the optical domain. However, as many services, e.g., in access, are terminated after ~20 km, E2E is possible only by assuming some form of bypass.

Next-generation optical networks will cope with even more stringent requirements than today. Existing architectures are sub-optimal and introduce too many unnecessary processing stages and overhead or are relying in electronics processing that introduces scalability limitations due to limited bandwidth. In this sense and from a purely “hardware” perspective, all signal processing functions that can be moved from electronics to photonics will contribute to the saving of energy consumption. Consequently, we must simplify it by developing new network architectures and technologies that enable efficient multi-layer/domain IP routing & optical transport integration. This also requires a high level of intelligence, easiness in managing and deploying. Such a solution would help full configurability of E2E connections, better planning, and lower power consumption.

Lines of research in this field might include: (i) development of technologies that consent to remove unnecessary opto-electronic-opto (OEO) conversions and processing to decrease power consumption and footprint; (ii) realization of intelligent and configurable components and that can

[8] Although it is worth mentioning that thanks to ICT the number of travelling is reduced.
[9] Shaping Europe’s digital future.
[10] This is gross amount, which does not consider the reduction of power consumption enabled by the ICT.

be operated via software, and optimized, e.g., for performance or low power consumption; (iii) Simplification and optimization – by minimizing the amount of packet processing of information, i.e., avoiding when possible IP routing – of the way information is processed. Information is analogue in nature and is important to identify how/where it can be more efficiently processed with electronics, or photonics in the digital (electronics) or analogue (electronics/photonics) domains. This could benefit from concepts relying on electronics/photonics co-integration and by the introduction of configurable programmability in photonics; (iv) optical packet/burst switching in wide network enabled by novel components and new concepts (like e.g. deterministic networking), and by a close research collaboration – aiming also at standards – among operators, vendors, and component manufacturers; (v) realize truly cloudified software-based configurability at the component/subsystems level so that each building block is optimized for specific applications. Programmability and softwarization should be also used at the platform level to optimize the implementation of designs and the performance of the resulting components/subsystems while reducing power consumption and footprint, and finally at the network level with the goal to enhance capacity allocation, reduce OPEX as well as increase network reliability (proactively identify potential failure, optimize planning and operations, etc.)

### 7.9.2 Energy efficient transceivers

Although the power consumption of individual transmitters and receivers is negligible – in comparison to the total consumed power by the network – their pervasive use and scaling to higher-rates requires careful engineering to achieve the targets of energy savings. There exist several ways to reduce the energy within the transceivers: (i) modern transceivers allow a multitude of transmission modes. These could be exploited to minimize consumed power and spectral occupancy. For example, the FEC overhead can be tuned based on the current specific margins, or, similarly, the spectral occupancy could be adjusted according to the given traffic demands that need to be served. Autonomous and flexible transceivers, also in terms of adaptive spectral occupancy, need to be flexible and follow the dynamically changing traffic variations; (ii) Specific power-hungry functions could be outsourced to photonic devices such as optical FPGAs, which enable dynamic and energy efficient processing via PICs. For instance, the dispersion compensation within the DSP could be carried out by self-adaptable programmable optics, at least for specific link parameters. Optical FPGA could also offload functions/calculations in data centres and high-performance computing infrastructures; (iii) Parallelism in the optical domain is key for scaling/improving performance and can rely on either the spatial (i.e., SDM) or the spectral (i.e. UWB) domain. Transceivers and their building blocks can be designed such that they can be reconfigured and reduce the consumed energy when they are operated under certain conditions; (iv) Co-packaging can significantly reduce the power needs of next-generation optical transceivers, including pluggable ones. Research could deal with the different approaches to interface photonics to the electronics while considering how close the photonics can be to the electronics. So far two solutions have been proposed: (a) low-data rate and (b) high-data rate but limited by SERDES. Research is needed to overcome the limitation caused by co-packaging; (v) Photonic programmable chips to replace ASIC partially or totally.

### 7.9.3 Energy-aware optical networks and components

Nowadays, optical network architectures are not optimized to minimize the power consumption and achieve energy awareness, although the transition to a greener world is becoming a major sustainability concern for our society. Furthermore, the associated components are also not

optimized to avoid waste of energy. As a result a series of new research directions supporting the greener networks is listed hereafter: (i) optimized design of components via parallelism. Often it is not needed to use the best performance, and part of the components composing an optical transmitter might be switched-off, e.g., a DAC does not always need to deliver the best performance, and if intelligently designed, part of it could be switched off - e.g., by turning off un-used capacity and active components, also as function of the instantaneous traffic; (ii) Usage of network telemetry approaches to monitor the actual consumed power and enhancement of control plane operation to enable energy-aware network via software configuration. Guidelines for system CO2 monitoring and requirements; (iii) optimize the application server locations to minimize the power consumption, and latency; (iv) enable network operators to minimize energy and resource requirements through load-adaptive network control; (v) optimized How to optimize services placement in data-centres across the networks? This cannot be done only by networking. Optimize low latency. Optimize where the resources are available.

### 7.9.4 Zero-electronic waste and scalable optical networks.

Existing optical networks are not scalable. Upgrades to new generations are carried out by generating large electronic waste (e-waste). For example, when moving from 10G to 25G PON, all boxes and transceivers at the end user and within the electrical aggregation stages need to be replaced, regardless it is required or not. Next generation optical networks need to be further developed by relying on improved technologies that enable programmability/reconfigurability, so that the network can be upgraded in a dynamic way only when and where is needed. Some lines of relevant research may include: (i) development of technologies that permit the co-existence of multiple generations of optical devices so that upgrades are local and not network wide. These novel approaches should also enable full interoperability among different domains and vendors; (ii) design of components and network elements e.g., racks, switches, so that the overall power consumption is reduced, and that scalability is optimized; (iii) utilization of new materials and new fibres – e.g., with lower attenuation and nonlinearity – so that the transceivers and network can be simplified; (iv) enable techniques supporting reconfigurability; e.g. to transmit only what it is needed and process only what it is required. E.g., nowadays we transmit 100G / 400G, regardless of the real traffic which is actually transmitted; (v) Consider new cooling techniques by involving collaboration with other fields of research.

### 7.9.5 Research Challenges

| **Research Theme** | **Simplified and efficient high-speed optical networks** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Flexible E2E optical networks<br>Current optical networks are divided into too many domains that require the usage of different OEO conversion stages.<br>Programmable Integrated Photonic Processing hardware<br>Introduce new Optical Node and Transceivers Architectures that are fully dynamic and configurable to support intelligent on demand | Long-term (finished within 5-7y) | An ecosystem of devices and components to create E2E flexible optical networks.<br>Highly scalable and flexible, same E2E technology.<br>**KPIs**: (I) reduction of 50% of the transceivers | Sustainability and sovereign digital infrastructure.<br>Adding software-defined intelligence and dynamically optimized performance (spectrum management, capacity allocation and energy saving) to photonic transceivers and switching elements.<br>Incorporation of a future-proof and scalable matched |

| | | | |
|---|---|---|---|
| processing of traffic in an optimized manner. | | | interface between the fibre and the wireless segments of access networks to support 5/6G and future extensions conveying data and sensing information. |
| Zero-electronic waste and scalable optical networks<br>Current optical networks are not multi-generational. When a new generation is introduced, old and low-speed transceivers cannot talk to newer and higher-speed ones. Need to introduce network architectures and device technologies that are scalable and configurable for that to allow network upgradability.<br>Basic building-blocks of past high-capacity backbone networks/systems may be reutilized in the future in lower performance access/short reach infrastructures (in line with the concept of Cyclic-Economy). | Long-term (finished within 5-7y) | A comprehensive strategy to realize fully scalable optical networks.<br>KPIs: (I) 0.15 W/Gbps, a 90% reduction respect to 100G / wavelength platforms in data centers and their interconnect. | Sustainability |

### 7.9.6 Recommendations for Actions

| **Research Theme** | **Simplified and efficient high-speed optical networks** | |
|---|---|---|
| Action | Flexible E2E networks | Zero-electronic waste |
| *International Calls* | | X |
| *International Research* | X | |
| *Open Data* | X | |
| *Large Trials* | | X |

## 7.10 Optical integration 2.0

The foundation for the development of cost- and energy-efficient systems with high reliability lies in the integration of multiple optical and electrical functionalities, as scaling down the number of high-speed interfaces will reduce the power consumption of the network components. In addition, improving the repeatability of manufacturing will increase the reliability of photonic components and reduce their cost. At the same time, the performance of the components needs to be enhanced to support a wider spectral range for new optical bands and higher speeds for increased data rates per channel. These challenges require the investigation of novel material platforms and, ultimately, the combination of multiple platforms up to the manufacturing level.

To meet all these challenges, the following aspects must be considered in order to enable the photonic layer to support the challenges of the system and networking layers.

### 7.10.1 Multi-band exploitation

Further expansion of overall system capacity requires exploitation of fiber wavelength windows beyond the C-band. To this end, passive and active optical components need to support ultra-wide-band operation with optical bandwidths exceeding 100nm, posing challenges from the point of view of component design and material properties alike.

### 7.10.2 High-capacity interfaces for spectrally and spatially multiplexed systems

Increased data traffic in optical transport networks will require more and more high-speed optical interfaces, when exploiting higher network capacities enabled by wavelength and spatial multiplexing. To avoid scaling of cost and power consumption with the exponentially increasing data traffic, the development of standardized components for future spectral and spatial unit cells are required.

### 7.10.3 New materials

Moving to higher channel bandwidths, further performance gains are necessary and might be achieved by (monolithically) adding organic or ferro-electric (e.g., BaTiO3) materials into the Silicon platform, providing potentially very high electro-optical coefficients and reducing the required driving power. This will be beneficial for the integration of optical functionalities on every size scale and allows to meet the demand to drive up analogue bandwidth and data rate per port by a factor of 10 by 2030. Other new materials, e.g., Lead Zirconium Tantalate (PZT) for phase actuation and thin film lithium niobate (TFLN) for modulation, have a potential to significantly contribute to the exploitation of low-power actuation and switching and a wider optical bandwidth range through high performance modulation capability in combination with negligible parasitic propagation losses. The integration with mature photonic platforms such as InP enables on-chip laser integration, leading to compact, energy-efficient, and low-cost transceiver solutions.

### 7.10.4 Optical chip interconnects

Advances in electronic integration follow Moore’s law and lead to increased throughput requirements of electronic ICs on or between printed circuit boards (PCBs). Packaging and I/O limitations will require a transition from electronic to optical chip interconnects when further scaling up electronic processing capabilities. Silicon-compatible, compact, and low power datacom transceivers are required to facilitate an integration into next-generation multi-chip switch and processor modules. Silicon is known to provide good passive optical properties for routing, modulation, and detection of light. It needs to be mentioned, however, that while Moore’s scaling of electronic memory and processors yields ever smaller structures in Silicon, this miniaturization is not feasible for optical components, where the telecommunication wavelengths on the order of a micrometre pose a limit on the structure sizes. Further integration of photonic and digital processing functions will require scale adaptation.

### 7.10.5 Multi-platform manufacturing

For the optical transport use case, typical volumes are in the range of only ~100,000s/year/design, while quality and performance of the transceivers is paramount. These volumes are subscale for typical silicon semiconductor fabs. Scaling III/V wafer processes to 4” and 6”, hybridly cointegrating III/V actives with low-loss passive waveguides like stoichiometric Si3N4, is essential to scale up optical transport with the ever-increasing long distance internet traffic.

### 7.10.6 Photonic-electronic integration

As the demand for on-chip functionalities continues to grow, it is expected that electronics technology will keep on focusing on the integration with photonic circuits to keep up with the Moore's Law scaling. The types of electronic-photonic integration are monolithic, heterogeneous, or partially monolithic. The unique advantage of Si photonic-based photonic integrated circuits (PICs) is the CMOS compatibility which enables the seamless integration with electronics and thus, the realization of cost-effective, high yield manufacturable solutions.

### 7.10.7 Reliability and repeatability

To foster commercial exploitation, PICs require a high degree of repeatability in manufacturing under somewhat varying processing, material and layout conditions. Establishing mature Process Design Kits (PDKs) requires robust passive and active optical components that perform within narrow margins over multiple dies, wafers and wafer-runs. Reducing statistical deviations in PIC performance as opposed to setting up hero-experiments and focusing on on-wafer testability of PICs is likely to enhance industrial uptake.

### 7.10.8 Research Challenges

The research challenges from the previous subsection are summarized below:

| **Research Theme** | **Optical Integration 2.0** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| To utilize the full capacity of the fiber, new materials and designs for optical wide-band operation must be developed. | Mid-term (finished in 5y) | Optical components for multi-band operation, based on new materials or robust building blocks<br>KPI:<br>Optical operating bandwidth >100 nm | High-capacity scaling and reliable connectivity<br><br>Increased system capacity while using existing fiber infrastructure |
| Parallelized optical transceivers, supporting multiple channels in the spatial and wavelength dimensions | Mid-term (finished in 5y) | Modular transmitter/receiver components to support multiple optical channels<br>Concepts for multi-band operation with unified transceiver components<br>KPI:<br>Transceivers for termination of >8 channels with power consumption of < 2 today's channel equivalents<br>Optical operating bandwidth >100 nm | High-capacity scaling and reliable connectivity; Cost-effective and energy-efficient systems<br><br>Increased system capacity with less space and less energy consumption |
| Integration of optical interconnects between electrical processing modules on a chip | Mid-term (finished in 5y) | Optical interconnections on a chip using multiple material systems including on-chip mode matching / spot size converters<br>Electro-optical interconnection on the same chip allowing monolithic co-integration of RF electronics and photonics<br>KPI:<br>Energy consumption < 10^5 eV / bit<br>Interconnect loss <0.5 dB/facet | Cost-effective and energy-efficient systems; Advanced electro-photonic integration<br><br>Higher chip processing power with less energy consumption for connections between processing modules<br>Increased RF-bandwidth systems at reduced footprint, lower cost and lower power consumption. |

| | | | |
|---|---|---|---|
| Multi-platform manufacturing | Mid-term (finished in 5y) | High-performance, long-distance transceivers using cointegrated best-of-class materials<br>KPI:<br>Full O- to L-band coverage<br>Power consumption 2-4x below SotA | High-capacity scaling and reliable connectivity; Cost-effective and energy-efficient systems<br><br>Increased system capacity with less space and less energy consumption while using existing fiber infrastructure |
| Reliable, repeatable and testable PDKs | Short-term (finished in 3y) | PDKs that perform within narrow margins under varying conditions over chips, reticules, wafers, wafer-runs.<br>Electro-optical on-wafer test vehicles enabling simultaneous testing of optical, DC and RF parameters | Cost-effective and energy-efficient systems; Advanced electro-photonic integration<br><br>Increased reliability of PICs, higher value of foundry services, reduced barriers to enter PIC market |

### 7.10.9 Recommendations for Actions

| **Research Theme** | **Optical Integration 2.0** | | | | |
|---|---|---|---|---|---|
| Action | Multiband exploitation using new materials | Multi-channel transceivers | Integration of chip interconnects | Multi-platform manufacturing | Reliability and repeatability |
| *International Calls* | X | X | X | Establishment of European multi-platform fabs for co-integration of multiple photonic IC technologies | Improvement of fab performance |
| *International Research* | European cooperation required to achieve goals | European cooperation required to achieve goals | European cooperation required to achieve goals | | |
| *Cross-domain research* | Cooperation with Photonics21 PPP | Cooperation with Photonics21 PPP | Cooperation with Photonics21 PPP | Cooperation with Photonics21 PPP | Cooperation with Photonics21 PPP |

## 7.11 Optical access beyond FTTH

The evolution of optical access technologies has so far been driven primarily by Fiber-to-the-Home (FTTH) network architectures and services. The economies of these networks demand ultra-low cost, still highly performant optoelectronic and digital processing capabilities. The optical system and network architectures must be optimized for efficiently aggregating dynamic traffic to and from

many nodes in a given area while respecting differentiated service requirements, and at the same time meeting challenging physical layer specifications.

After two decades of increasing system capacities (bit rates) and optical power budgets on the one hand, and of refining the TC Layer (Transmission Convergence, i.e. a system-wide MAC layer) for supporting differentiated service requirements on the other hand, PON-based access solutions have now reached a level of technical maturity and cost efficiency that makes them attractive also for other many-nodes/short-reach networks beyond traditional FTTH. First new market segments can be addressed already today using existing PON systems without the need for modifications: commercially available Passive Optical LANs to be used in place of conventional office and enterprise LANs, as well as for backhaul links - and fronthaul in near future - in small cells mobile networks.

Building on this solid foundation, the future evolution of optical access technologies and architectures will bring about further increased system capacities, highly flexible system and network reconfiguration, meshed and resilient network topologies, coexistence of best effort and deterministic traffic, secured transmission over complex architectures, and much more as will be pointed out in the subsections below. The objectives shall be to make these solutions suited for diverse applications and network scenarios in public and private area environments, with single- and multi-tenant business models, for vertical markets and industrial applications, for 5G and emerging 6G mobile networks, for small intra-datacenter networks, and more to come.

### 7.11.1 Increased capacities and flexible configuration of access transmission systems

Optical transmission systems for access networks today are worldwide predominantly based on TDM-PON technologies (GPON family (ITU-T) and EPON family (IEEE)). Single channel line rates of commercial systems will soon reach 50 Gb/s, employing NRZ IM/DD modulation and reception schemes, supported by FEC and DSP for mitigating noise and dispersion induced transmission errors. Multilevel signaling (real or complex valued) and field modulation/coherent reception will come into play for further increased capacities. In an alternative approach, multiple wavelength channels can be combined for achieving higher system capacities beyond 100 Gb/s. For addressing end nodes (users) in complex scenarios with diverse loss budgets and link distances, a third deployment option will comprise a combination of different line rates, modulation formats and FEC / DSP levels on single channels or across multiple parallel wavelength channels. This approach is similar to the Modulation Coding Scheme levels in 5G radio transmission, but has not yet been applied in optical access systems. Anticipating complex architectures in future access and similar short reach optical networks, this approach will enable graceful and cost-efficient upgrade options in deployed fiber networks.

The focus of this work shall be on design studies and proof of concept demonstrations of suitable system and network configurations, as well as of enhanced system protocols for managing and operating such multi-dimensional ultra-high capacity transmission schemes in most flexible and efficient ways, making them useful for many application scenarios as indicated in the introduction of this section. Along with TDM-PON based wavelength channels (TWDM) also non-TDM wavelength channels shall be considered for supporting sustained unshared high- capacity links to individual nodes in the network, preferably as a combined flexibly configurable TWDM/WDM-PON solution.

### 7.11.2 Flexible realtime and non-realtime resource assignment

In TDM-PON for residential users, the transmission bandwidth per end node (user) is assigned on a non-realtime basis, taking into account predefined minimum (guaranteed), assured and maximum bandwidth values per service, as well as the observed utilization of previously assigned bandwidth and on status reports from ONUs. This established process shall be further improved for optimizing the time varying bandwidth assignments towards more precisely meeting the actual service requirements and thus increasing the overall bandwidth efficiency of TDM-PON systems. For instance, more precise prediction of near future bandwidth requirements per service are expected to be achievable by adding appropriate AI/ML algorithms that can account for observed traffic patterns.

Network slicing in TDM-PON for enabling multi-tenancy business models is another topic that needs precise modeling, comparison and implementation of different bandwidth assignment strategies to enable meeting different prioritization and fairness KPIs, as agreed upon in the operators' SLAs.

The TDM-PON bandwidth assignment algorithms mentioned above are set to accommodate all services (on average over multiple PON frames (125 µs)) in the best possible way before allocating time slots fro transmission. However, for challenging service requirements such as low latency, ultra-low jitter, low packet loss rate etc., more predictable, i.e. deterministic, assignment strategies are needed. Direct time slot allocation per service and per frame is a promising approach. For client traffic with precisely predictable bandwidth needs over time (e.g. constant or strictly periodic), the slot allocation on TDM-PON can be preconfigured. For dynamically changing bandwidth needs, a low-level logical interface can be used for mutual realtime communication between client and TDM-PON to dynamically predict varying near future client needs and PON capabilities (e.g. CTI (Cooperative Transport Interface) for 5G fronthaul links over PON). The above strategies are rather simple to design for a single node or only a few nodes on the network. For many nodes, however, and especially when designing for an efficient utilization of the available transport bandwidth, appropriate strategies and algorithms are needed.

Meeting low latency and low jitter KPIs needs realtime and even isochronous realtime synchronization between network nodes, depending on the service precision requirements. The implementation of suitable processes for frequency syntonization and ToD (Time-of-Day) synchronization is addressed in the Deterministic Networking section of this document. However, the implementation of highly precise ToD synchronization processes in PON supporting many end nodes needs special consideration, in particular for sub-nanosecond precision that is required e.g. for precise positioning use cases.

In complex access networks with different segments and technologies (application, mobile, optical, computing), the non-realtime orchestration of coordinated individual realtime resource assignments is crucial for meeting stringent service requirements. Flexible reconfiguration of logical channels, involving time and spectral domains in different segments will add another level of complexity that needs to be addressed in this challenge.

### 7.11.3 Redundant, meshed and flexible optical layer network architectures

Optical access networks today are deployed on point-to-point (ptp) or point-to-multipoint (ptmp) fiber optical distribution networks (ODN), employing passive optical power splitters (for TDM-PON

and WDM-PON) or passive wavelength routers (for TWDM-PON and WDM-PON) for distributing optical signals to the end nodes. The ODNs implement a logical (via time slots in TDM-PON) or physical (via wavelength channels in WDM-PON) star topology from the OLT to all ONUs, in the majority of cases without redundant attachment.

Critical services will require redundant system layouts for improved resilience and and ultra-high network availability and service reliability. Redundant ODN layouts have been described for simple tree architectures already in early GPON documents. More sophisticated and resilient architectures will, however, be needed for certain use cases e.g. in the industrial space or in small cells x-haul networks. These architectures shall provide resilient connections in the distribution and drop section of the ODN, and at the same time allow for redundant attachment to the metro or equivalent aggregation networks.

Some latency sensitive use cases will (in addition) benefit from short local interconnects between ONUs in the same ODN or in neighboring ODNs, without going through the central node (OLT). Employing selective optical loop-backs at the passive remote nodes will establish a local meshed topology among the ONUs. Depending on the required interconnection patterns and the group size of interconnected ONUs, as well as on the required local link capacities and allowed latencies, different solutions shall be devised for optimized bandwidth efficiency, cost efficiency, energy efficiency and other KPIs. Additional optical ports on the ONUs may be considered for this overlay network. Solutions are sought for ptmp ODNs (PON), but as well for ptp ODNs.

Deployment related, operational, network migration related considerations are calling for reconfigurable (nominally passive) remote nodes in the ODN. This reconfigurability shall allow for flexible reconfiguration of the interconnection topology on an ODN on time scales well above milliseconds, with low energy consumption as much as possible. Suitable remote node solutions along with the supporting network architectures, including the associated management tools for configuring the remote nodes shall be elaborated and practically demonstrated.

### 7.11.4 Optical layer multi-tenancy in access networks

The cost of rolling out new fiber connections in access has frequently been a major blocking point in providing early FTTH services, and is again considered a blocking point for optical x-haul in future small cells 5G and 6G mobile networks. Sharing the fiber infrastructure among multiple players (either competing with each other or complementing each other in their service portfolio) may become a beneficial business model for such scenarios.

Neutral host models have been implemented in some optical access networks in Europe and in Asia. Those are based on one player leasing a dark fiber infrastructure to one system and service provider, but sharing a dark fiber infrastructure on the optical layer among multiple system and service providers has not been implemented yet. For making this a viable business proposal, appropriate technical additions to current ways of deploying and managing fiber networks as well as operating data services on such networks must be developed. For instance, a neutral host providing the dark fiber must be able to monitor and manage the ODN, however, without having access to telemetry, OAM and management data provided by the service networks using the infrastructure. On the other hand, sharing a fiber link in the wavelength domain e.g. needs definite upper bounds on optical power and other optical transmission system parameters in order to avoid mutual linear and non-

linear signal distortions. Related implementation and operation details must be elaborated, both for the neutral host and for the system and service providers, and suitable business models must be evaluated. As a side remark: one such service can be provided by an additional player using the fiber for various kinds of optical fiber sensing projects.

### 7.11.5 Research Challenges

Feel free to add some text, conclude with the Synthesis table:

| **Research Theme** | **Optical access beyond FTTH** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Challenge 1 :<br>Optical systems capacity increase beyond 100 Gb/s by combining diverse modulation coding schemes on multiple TWDM and WDM channels | Short-term (finished in 3y) | Scalable architecture solutions and TWDM/WDM PoCs for well beyond 100 Gb/s syxstem capacity, applicable to diverse use cases as addressed in the introduction of the section<br><br>SotA : single TDM-PON channel per system with up to 50 Gb/s serial bit rate (ITU-T), or 2 parallel 25 Gb/s channels aggregated by WDM (IEEE) | Addresses the High Capacity Scaling and the Cost-effective and Energy Efficient Systems requirements of the target vision |
| Challenge 2 :<br>Dynamic resource assignment, synchronization and orchestration | Short- to Mid-term (finished in 5y) | Non-realtime assignment (2y)<br>SotA : no AI/ML involved<br>Realtime assignment (3y)<br>SotA : first research proposals<br>Precise synchronization (5y)<br>SotA : N/A<br>Orchestration of assignment and synchronization (5y)<br>SotA : N/A | Addresses the Deterministic Networking, the Efficient Integration of optical technologies for radio access networks, and the Network Multi-tenancy requirements of the target vision |
| Challenge 3 :<br>Redundant, meshed and flexible optical layer network architectures | Short- to Mid-term (finished in 5y) | Quantitative elaboration of redundant network architectures for critical services (2y)<br>Full network architecture studies and PoC for meshed ODNs (3y)<br>Full network architectures and PoC supporting flexible reconfiguration | Addresses the Reliable Connectivity, the Cost Effective and Energy Efficient Systems, and the Deterministic Networking requirements of the target vision |

| | | | |
|---|---|---|---|
| *Challenge 4 :*<br>Optical layer multi-tenancy in access networks | Long-term<br>(finished in 7y) | Elaboration of technical add-ons and design guidelines for enabling optical layer multi-tenancy in fiber access networks, and related business models<br>SotA : N/A | Addresses the Network Multi-Tenancy, the Cost Effective and Energy Efficient Systems, and the High-Capacity Scaling requirements of the target vision |

### 6.1.1 Recommendations for Actions

| **Research Theme** | **Optical access beyond FTTH** | | |
|---|---|---|---|
| Action | Research Aspect 1 | Research Aspect 2 | Research Aspect 3 |
| International Calls | | | |
| International Research | optical components related | | |
| Open Data | | | |
| Large Trials | | | |
| Cross-domain research | optical components related | AI/ML related | |

# 8. Non-Terrestrial Networks and Systems

*Editor: Alessandro Vanelli-Coralli*

## 8.1 The 6G NTN Vision

### 8.1.1 6G as umbrella for NTN

5G has now rolled out in most developed countries in the World and we continue to see improvements and developments in the standards bodies. From 2021 to 2025 is the period in which NTN including satellites will work towards 5G integration with TN's and full commercial operation. Now is the time to start considering what techniques and technologies might feed into standardisation for 6G. The roadmap is shown in Figure 8-1.

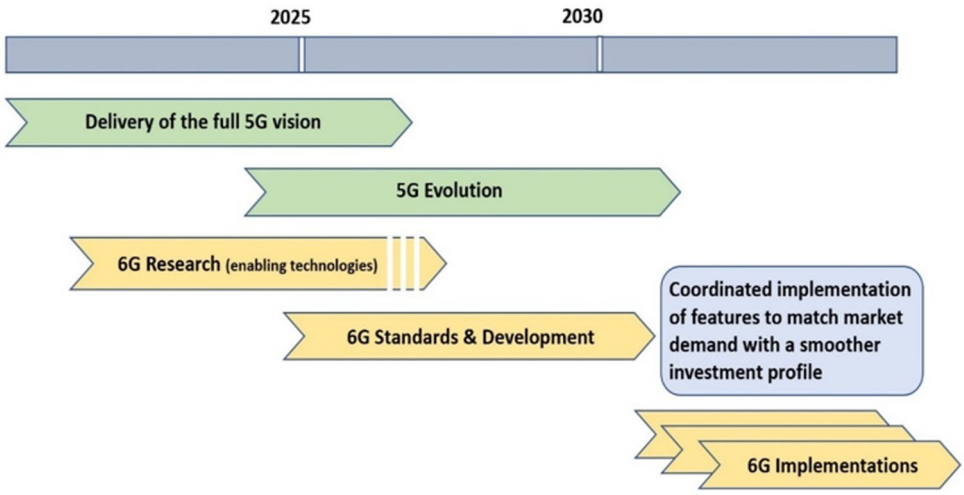



*Figure 8-1 -5G to 6G Roadmap time lines*

6G must be different from previous generations; it must be designed to meet Global Challenges and at the same time to include the technology to support both cost-effective coverage and radically new innovative services. The vision is that it should include converged digital and communications infrastructure in an ecosystem that serves humanities needs but at the same is affordable to the users and economic to the providers. In addition, contributing to environment sustainability and helping towards a net-zero agenda.

As distinct from the start of research for previous generations, in 6G we have a radically different world order to address. In particular, the following are now key challenges:

- Changed working practices in a post Pandemic era
- Climate change and the reduced reliance on fossil fuels
- The need to radically reduce energy usage and improve its security

Innovation in technology marches on at an increased pace with major innovations blurring the boundaries between the physical and virtual worlds and enabling natural interactivities between them. Increased and massive softwarisation, Artificial Intelligence (AI) and Machine Learning (ML), sensing of the environment are all key new innovations. The task for 6G is to embody the new

technology advances to address the global challenges in a system that will provide services that are affordable and usable by the population at costs that are sustainable by the operators.

Scanning the many 6G vision documents from around the world and looking into the contents of previous chapters, we abstract the following common themes:

- New Human centric services—AR/VR/MR—Teleportation
- KPI's that exceed those possible in 5G in latency and reliability
- Seensing at the user terminal merged with communications
- AI based network and massive virtualisation.
- A 3D space network including UAV's-HAPS-Satellites (NTN)
- New Frequency bands
- Increased security across integrated networks
- Address space customization and semantic forwarding
- Massive IoT
- Timing and positioning accuracy
- Computing as a network service

The vision of 6G incorporates a range of human multi-sensory experiences enabled by digital solutions and hyperfine geolocation with context awareness provided by massive localized sensors. In addition to human and local information sensing, system-level sensing will be essential for efficient and intelligent system operation. This implies fine time and frequency synchronization to microseconds and guaranteed ultra-low latency (ULL) not provided by 5G. This will enable the provision of a tranche of new services for verticals across telecom networks. The vision is of a hybrid **network of networks** from short-range and ultra-high capacity to the widest coverage via a new **space network** dimension (Figure 8-3). NTN will become one of the key enablers in providing coverage, security, and resilience in this 6G vision.

Within NTN, satellites remain key. Today's satellite communication systems have expanded from GEOs to include MEOs for regional coverage, which operate with digital payloads including multiple beam antennas to provide frequency reuse and very high throughput, approaching a Terabit/s. A significant innovation has been the emergence of massive constellations of LEO satellites or the adaptation of older ones (see below), offering very high throughput and low latency matching the demands of 5G and potentially 6G services.

As of October 2022, several efforts are considering the development of an Internet service delivered from space and principally from satellites orbiting the Earth. The most known are Starlink and OneWeb; their service is offered from very dense constellations of satellites in LEO orbits. Starlink is reaching several tens of thousands of satellites at 550km altitude, whereas OneWeb is a little less, at 1200km.

Considering older constellations, Iridium is being upgraded to Iridium Next and is running about 70 satellites at a 780 km altitude. Molniya, an older Russian service, runs from a Highly Elliptical Orbit (HEO) which ranges from LEO to MEO. Newer efforts, less mature at this time, are Kuiper (backed by Amazon) at between 590 and 630 km, GuoWang and GalaxySpace (China), Astra Space, Lynk (a 5G base station on satellite) at 500 km, Telesat at probably 100 km (lower than LEO), Boeing V-Band

constellation, Viasat XVI and Link-16 (military) and Aalyria (backed by Google). In Europe, an initiative started in 2022 towards an "EU space-based secure connectivity system".

Successive generations of these constellations will use optical inter-satellite links (ISLs) and higher frequency bands above Ka-band;—Q/V and up to optical for feeder links, as well as ISLs. In addition, the next generation of GEO and LEO satellites will include regenerative payloads to provide improved connectivity. Most applications are currently for backhaul or direct to fixed and mobile-specific terminals (often serving multi-users). However, some companies are experimenting with direct to HH -UT's operating in the mobile bands (ASTmobile and Lynk). The latter requires major technological advances and are likely to be a longer-term realization.

The role of satellites has traditionally been to provide coverage into regions not economic for terrestrial infrastructure and to provide resilient backup to terrestrial services. We see these features remaining as key drivers in 6G. Nonetheless, NTN provides also **flexibility, efficiency, service continuity, and fast and low-cost global coverage** (e.g., for IoT application). As terrestrial networks pursue lower latency service offerings, satellite constellations at very low altitude (vLEO) with Inter Satellite Links (ISL's) offer comparable and even lower latency for longer links. Thus, these systems are of interest for 5G and now inclusion in 6G. Due to restricted spectrum and satellite power, capacities have in the past been limited and hence more expensive than terrestrial. However today using frequency reuse, dynamic resource allocation and onboard processing both GEO and LEO satellites have increased to circa 1 Terabit/s and the costs of the space system have drastically reduced.

NTN should be integrated from the start in 6G rather than being bolted on, as with earlier generations. For satellites to play an integrated role in 5 and 6G, some commonality of standards is required. Until recently, satellites remained outside mainstream standards and had developed their own air interface standards — DVB-S2X (and its predecessors), which was initially based on video broadcast. More recently, and seeing the advantages of integration, satellites have joined the 3GPP standards groups responsible for 5G and now 6G standards. The 3GPP Rel. 16 (2017), on which the current rollout of 5G terrestrial networks is based, does not include satellites. However, starting from rel. 17, Non-Terrestrial Networks (NTN) are part of the 5G ecosystems and their development continues into Rel. 18 &19 towards the goal of integrated standards. There is thus a pathway to full integration as shown in Figure 8-2 with the period to 2025 used to getting 5G and satellite established and the period up until 2030 having satellite established as a unified part of 6G.

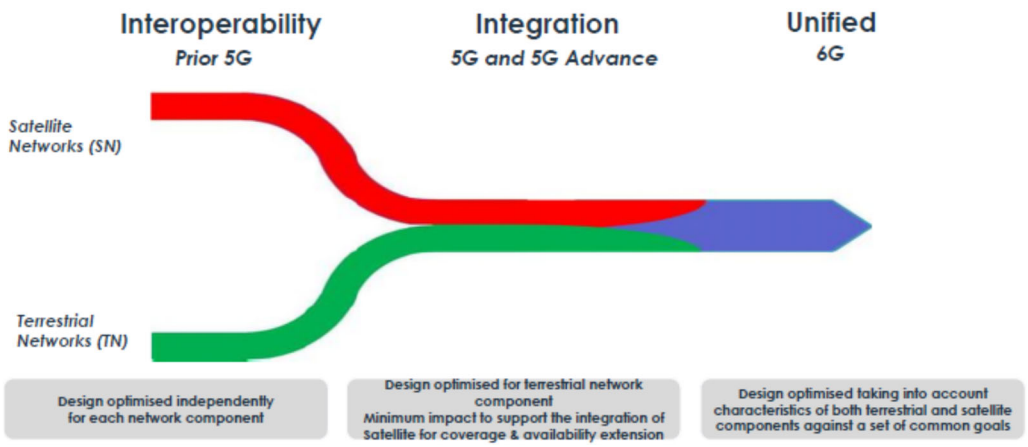



*Figure 8-2 - Pathway to full integration (source Thales Alenia Space).*

### 8.1.2 Satellites as key components in 6G

The key role of satellites for communications is in coverage and resilience, however, they are essential in enabling other critical services such as earth resources, positioning-navigation, and timing (PNT) and for continuous control and connectivity to aerial devices (UAVs) and maritime vehicles. The ultrahigh capacities needed for some 6G services will only be available on short-range terrestrial connections using Terahertz links in urban areas. A range of 6G services will be required by users traveling out of these areas and thus the pathway to 100% coverage can only be economically provided by satellite. The satellite will also be used to backhaul mobile cells in rural and remote areas or/and to provide backup resilience. Of course, connection to ships and aircraft will necessitate satellite backhaul. Connection to premises or events can be provided quickly by satellite via small antennas.

In addition to the above it will be seen that the addition of the extra dimension of space to create a 3D network is implicit in the 6G vision and this is where satellites fit into a broader picture. As shown in Figure 7-3 this leads to the concept of a multi-layer network which adds satellites in GEO, MEO and LEO to lower altitude HAPS and even lower aerial devices such as drones; orbits such as HEO (highly-elliptical orbit, see Molniya) and VLEO (Very-Low Orbits, also mentioned above as vLEO) might be considered too. The network architecture connecting these components will be service dependent as some architectures will better suit the requirements of specific services. The network functions can also be distributed amongst the entities to optimise performance. In all cases we will have a highly integrated E2E cross-network system.

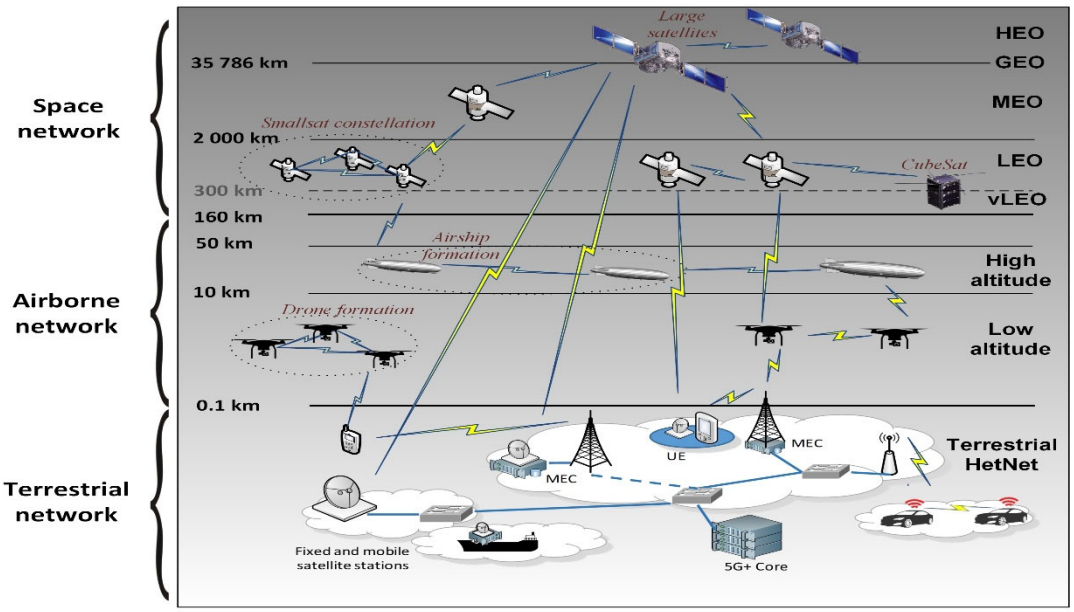



*Figure 8-3- 3D space network [C8-01]*

Today it is too early to list detailed technologies and use cases for 6G. The time to do this would be at the start of the 6G standardisation process – possibly in 2025. However, there are emerging themes such as the increased use of open systems and the integration across various networks as shown in the 3D layered model. We see 6G as a fabric and an ecosystem bringing together a number of new technologies providing ultra-high resolution, communications at the edge, supported by massive intelligence in the core.

The 3GPP standards group and NTN are continuing their work on 5G+ looking at applications into vehicle to everything (V2X) transport and 5G based IoT [C8-2]. Also considerations regarding energy and spectrum efficiency, location sensing, carrier aggregation and advanced antenna arrays are under way. In other areas advances in software defined radio (SDR) and digital processing will contribute to more flexibility in radio access and in core networks. Security across the whole system will be critical and will be embedded in the design. This will require the use of intelligent firewalls, context-aware domain level protection, and advanced cryptography supported by cloud quantum computing. These and other innovations will feed into the base definitions for 6G.

For 6G we will need new and advanced techniques that enable deeper integration between satellite and terrestrial networks which are seamless from a user perspective, moving from satellite to terrestrial coverage. With the introduction of large LEO satellite constellations with high mobility this introduces new challenges ranging from intelligent and dynamic spectrum sharing to seamless handover and maintenance of QoS.

### 8.1.3 Key Challenges for satellites in 6G;

- Unified T/NTN architecture
- Full network integration of all layers in the 3D SDN Network,
- Direct connectivity to smart phones, outdoor and light indoor and in vehicle,
- Ultra Low Latency support for vertical sectors
- Merging networking and computing -not just at the edge

- Integrated and flexible Air Interface for multi services.
- Ultra-accuracy of positioning and timing
- Embedding AI in network and RAN
- Providing Security across the network- elements
- A new IP for space networks
- New spectrum and sharing across the network components
- Supporting massive IoT
- Solving the problem of massive antennas in space

5G is now commercially launched and a programme of integrating satellites is in place; it is now time to set in motion a programme of research and development towards **satellites in 6G** which we outline herein as an input to full 6G standardisation starting in 2025.

## 8.2 Research theme: Architecture and System-Level Aspects

The architecture of an NTN system consists of a set of elements belonging to the space segment and/or the ground/user segment which interact as an integrated system. System integration allows to better differentiate the overall set of design solutions potentially increasing system flexibility; it can be addressed as integration of different technologies (e.g. radio and optical), integration of services (e.g. localization, Earth observation and communication services), integration of space and non-space systems (e.g. GEO/MEO/LEO/VLEO satellites, aeronautical components, stratospheric systems, aerial vehicles and terrestrial stations).

### 8.2.1 Multilayer Architecture

Multilayer 3D design aspects should start from the definition of the roles and tasks of each component guaranteeing flexibility of functions, together with their method of interfacing with other components. The definition of roles and functions of each node (e.g. network access, routing/bridging/switching, proxy, control/learning) should take into account both interconnection options but also the geometry of the overall network. However, the geometric design of a multilayer 3D network is the first step of the design process requiring an interdisciplinary approach. Satellite constellation design even for a single orbit height is a complex task that varies vastly based on the mission objectives. One needs to take into account mission and service requirements including coverage areas, throughput for end users and latency requirements. For some IoT or Earth observation designs the coverage can be non-continuous whereas broadband satellite constellations have to provide continuous connectivity to the coverage area. Various constellation creation methods, such as Walker constellation, which is a globally symmetrical configuration, or streets-of-coverage are used for LEO constellation designs. The latter refers to the swath on ground with continuous coverage. In addition, analytical methods such as stochastic geometry analysis that abstracts the generic networks into uniform binomial point processes have been developed to support constellation design and analysis. The overall design also has to include use of inter-satellite links, location of ground stations, data links and scheduling algorithms. Since 6G satellite communications architecture is thought to be three-dimensional, the constellation design needs to be updated to include satellites in GSO and NGSO orbits. This brings both new challenges and possibilities to the design task. How to optimize the number of satellites for the required coverage and service taking into account different types of satellites in different orbit heights? There is a need to develop new design methodologies and simulation tools for flexible operations and capacity

estimations across layers. Future designs might require the use of machine learning frameworks in order to cope with the vast dynamicity and complexity of the multilayer system consisting of a very high number of moving nodes.

### 8.2.2 Satellite-as-a-Service and Ground-Segment-as-a-Service

Traditionally, the significant costs of satellite missions have prevented many business customers from embarking on space technology. New economic trends based on sharing an infrastructure have the potential to attract many stakeholders that cannot invest in a full project. With a satellite-as-a-service (SaaS) or a ground-station-as-a-service (GaaS), customers would be renting out the platform and payloads of satellites or ground stations, respectively, for their own purposes, either acquiring data, doing edge computing, or using communication capabilities. Distributed Ledger Technologies (DLT) can enable the necessary transparency, interoperability and accountability, by recording every transaction and event for a seamless service. One main challenge for the implementation of DLT in space is the high demand of computing resources, and this requires further research on the architectural solutions and practical constraints. A step beyond SaaS is software-as-a-service (SWaaS) from satellites, such that stakeholders can develop and run decentralized applications in space based on open-source and real-time operating system and supported by DLT for immutability and trustworthiness.

### 8.2.3 Autonomous Networking

Driven by the vision of sustainable communication systems, there is the need to support a flexible, scalable and low-cost operation of networks encompassing flying autonomous vehicles, satellite constellations or a combination of both. Hence there is the need to develop a networking architecture able to support the exchange of large amounts of data over a set of intermittent available links as devices, while supporting the creation of autonomous networking systems, for instance by combining computing and networking functionalities. Such autonomous systems should be able to adapt their behavior to changing contexts, as well as to predict future constraints and requests in order to optimise their performance.

### 8.2.4 Mobility Management

Mobility Management in NTN is challenging due to the movement of satellites nodes, like LEO satellites moving approximately at 7.5 km/s. In rel. 17 2 different deployment scenarios were designed:

- Earth moving cells, where the cells move with the movement of the satellite across the surface of Earth.
- Earth fixed cells, where the satellites continuously adjust their beams to form fixed cells on Earth.

The latter can lower the number of handovers at the cost of a lower link budget and performance. As we are targeting high performance in the next releases the number of handovers will be significant compared to terrestrial cells. Handover solutions exist to ensure virtually zero failures, but the number of handovers is still large. The cost of every handover to the UE is that new information needs to be read from the new cell in both idle and active mode. This increases the power consumption of the UE significantly and should be minimized in future releases by avoiding having to read new cell information. At the same time the network needs to know where the UE is such it can be paged, which puts, which puts a constraint on how large one can design tracking areas

and registration areas of UE. The trade-off between the size of those vs the paging cost and UE power consumption should be optimized and the full design should be reconsidered in order to accommodate fast moving transmission nodes.

### 8.2.5 Autonomous Positioning

Most of the mobility management techniques developed for mobile satellite communications assume that the user terminal is provided with a GNSS receiver sharing user terminal location information with the satellite network control system. Also 3GPP Rel. 17 and Rel. 18 rely on GNSS. Consequently, the power consumption and the processing capability of the user terminal are increased as required by the functioning of a GNSS receiver. However, this is in contrast with the limitations of Internet of Remote Things (IoRT) terminals in terms of power generation, usually provided by small solar panels. Furthermore, in urban canyons where the minimum GNSS signals visibility is not guaranteed, the NTN service may not be available even if, in contrast, NTN signals are received.

A straightforward solution to overcome these effects is to integrate a positioning service with the communication service, possibly using the same signals designed for communication purposes or proposing modifications and/or additional features. This additional option allows to offer to NTN terminals an autonomous positioning service which does not exploit any GNSS signal. Two different approaches may be considered: 1) a network-based positioning approach, where NTN entities computes the user terminal location exploiting measurements extracted from the signals transmitted by the terminal; 2) a user-based positioning approach, where the user terminal computes its position using the signals received by the NTN nodes. It is worth noting that autonomous positioning is achieved using any of the two approaches, but only the first approach allows to decrease the power consumption and the processing capability requirements of the user terminal.

### 8.2.6 Expected Impact

The design of a NTN component through a solid integration of 3D sub-components opens the door to a more effective use of radio resources, enhanced service flexibility, and improved user engagement even under challenging conditions.

The implementation of computation and storage capabilities in 3D NTN nodes allow to pre-process and store several different types of information also including auxiliary information thus helping to reduce user terminal requirements in terms of computational complexity, power, mass, storage and increase both spectral and power efficiency of each link.

In summary, the key point to achieve the previous results is to find the optimum trade-off in terms of functions and capabilities of each different node of the 3D NTN system.

The following list of KVI and KPI may be used to measure how well the expected impact of the proposed system-level design approaches is achieved.

#### *8.2.6.1 Key Value Indicators (KVI):*

- Autonomous positioning capability
- Service flexibility
- Support of IoT services to remote locations

*8.2.6.2 Key Performance Indicators (KPI):*

- Communication latency as a function of the geographical distance between source and destination in comparison with the latency of a terrestrial-only network.
- Positioning error for the autonomous positioning service.
- QoS levels to maritime and aeronautical users.

## 8.3 Research theme: Air Interface

While most of the research development of 5G NTN were focused on adapting the conceived air-interface for its application on NTN scenarios, 6G aims to address NTN peculiarities from its conception phase. This will involve addressing terrestrial and non-terrestrial trade-offs of the different entities that constitute the air-interface.

In the following we introduce the main challenges of next generation air-interface 6G satellite communication, ranging from the waveform design to novel conceptions of the shared access.

### 8.3.1 Waveform Design

*8.3.1.1 Evolution Radio Technologies*

**Orthogonal frequency division multiplexing (OFDM)** is the modulation technique adopted in 5G-NR. However, at frequencies foreseen for B5G networks or at frequencies typically used in satellite communications, OFDM is challenged by several effects, which suggest considering alternative modulation formats. In fact, in LEO and B5G scenarios OFDM requires frequent adaptation due to mobility and fading, which would lead to prohibitive overheads at high carriers due to the short coherence times. In addition, OFDM suffers from a high peak-to-average power ratio (PAPR), leading to reduced power efficiency, which becomes a limiting factor at high-frequency carriers.

These drawbacks of OFDM have stimulated the interest in advanced modulation formats. An example is represented by **Orthogonal Time Frequency Space (OTFS) modulation** which is a promising candidate, as it has lower PAPR, is much more robust to time-varying channels, has a much lower overhead, and is more resilient to Doppler. In contrast to OFDM modulation in the time–frequency domain, OTFS relies on the delay–Doppler domain. This implies that a time-varying channel with constant Doppler will appear time-invariant to OTFS.

Yet a relevant approach is to consider end-to-end data-based physical layer designs. In contrast to classical source channel coding approach, the new wave of AI-based techniques is promoting novel developments of communication sub-systems (e.g. synchronization, modulation,…). This will allow new schemes to optimally adapt to the peculiarities of NTN, namely non-linear space segment operations, Doppler effects, etc.

*8.3.1.2 Optical Wireless*

While most of satellite communications are currently taking place via radio components, optical connections are being increased, specially in professional links and inter-satellite (inter- and intra-orbit) communications.

In terrestrial in single-mode fiber-optics transmissions, the state-of-the-art waveform design is represented by **coherent systems**, since they present many advantages:

- the possibility to adopt high order modulations, thus reducing the speed of the required electronic processing;

- the absence of nonlinear transformations at the receiver that degrade the information content of the received signal;
- the availability of advanced signal processing techniques (such as, for example, single- and multi-carrier predistortion, equalization, advanced detection algorithms, etc.) to compensate for the possible impairments;
- the possibility to adopt sophisticated state-of-the-art techniques such as, for example, time-frequency packing, probabilistic constellation shaping, polarization multiplexing, orthogonal frequency-division multiplexing, etc.

On the other hand, the disadvantages are represented by the need of a more sophisticated hardware required for the opto-electronic transformation, the need to recover the transmit carrier, in phase and frequency, and a much higher sensitivity to phase noise than **intensity-modulation/direct-detection (IM/DD) systems**, thus also calling for a more sophisticated digital processing at the receiver.

The evolution experienced by fiber-optics systems and networks and the maturity of optical coherent technologies have a significant impact optical wireless satellite communications, currently based on IM/DD systems. Although the propagation impairment and architecture of the satellite scenarios can be very different than those in fiber-optics systems, and thus an adaptation is required, the benefits can be significant.

The stringent 6G requirements on low energy consumption, data rate, and users to be served may imply that optical satellite communications play a key role. However, given the large variety of the scenarios to provide service, the waveforms may suffer many different impairments. To name a few, they have to support the turbidity in underwater, turbulence in space and multipath in ground segments.

In order to improve their performance, advanced signal processing tools such as MIMO, modulation, coding may be investigated.. Toward this regard, the properties of the optical channel shall be considered to obtain additional levels of diversity. In particular, systems operating with hybrid radio-optical components (e.g. optical feeder links in transparent satellite payloads) shall be modelled in an end-to-end basis. As a matter of fact, whenever both radio and optical links are simultaneously present, diversity techniques take a very relevant role in order to increase the communication availability.

### 8.3.2 Multi-antenna solutions

#### *8.3.2.1 Satellite Antenna Evolution*

The industry in the space sector is working to develop efficient antenna solutions for the satellites, able to reduce the associate size, complexity and cost. As opposed to conventional reflectors, which in most cases generate fixed multibeam footprints with high efficiency, phased array antennas offer highly flexible coverage by using advanced beam-forming technology, with lower efficiency when they operate as direct radiating arrays. While there are some common functional elements which are identical for GEO and NGSO payload architectures, the different orbital geometry and required EIRP and G/T pose different challenges; thus, antennas can be smaller for NGSO satellites, while subject to tight integration constraints. Reflect array technology encompasses the virtues of both

reflectors and phased-arrays, and appears as a promising candidate for new satellites in need for flexible coverage provided by efficient antennas which occupy limited real state on the satellite. By properly modifying the phase-shifts at the discrete elements, they can provide high-gain focused beams with dynamic steering. This is the same physical principle as that used by Reconfigurable Intelligent Surfaces (RIS), which are emerging as a supporting technology for smart wireless communications, so that fruitful synergies can be expected between terrestrial and satellite communications when it comes to the design of reconfigurable reflecting surfaces. Under different configurations, that one based on array-fed reflectors is perhaps the best positioned for further R&D efforts, in order to provide an increase of the antenna efficiency in broadband applications with wide-aperture antennas.

As a matter of fact, the continuous 'softwarization' of the space segment will guide the evolution of the radio control technology as extremely user radio resource granularity will be accomplished yet with that a very short time-to-react. Ideally, the space segment could autonomously decide beam-pattern coverage reconfigurations.

At the user terminal, it is evident that next generation satellite systems will eventually use high frequency ranges at the user segment. This will entail the development of new aperture developments devoted to specific verticals and attending to SWaP restrictions. At the same time, terminals at V and W bands will provoke revisiting classical beamforming satellite on-the-move (i.e. a combination of time reference and spatial reference) solutions. In order to address certain verticals (e.g. automotive), the use of metamaterial antennas will be required together with holographic beamforming.

#### *8.3.2.2 Satellite Beamforming in Satellite Swarms*

The emergence of broadband services provided by *satellite communications* (SatCom) has been drastically pushing the requirements for more efficient communication payloads. In addition, the possibility of integrating the satellite component into *terrestrial networks* (TNs) opens up a myriad of opportunities to be explored. Indeed, the coverage of both future cellular and *Internet of things* (IoT) TNs would greatly benefit from a complementary non-TN. These benefits, however, come at the cost of dealing with minimal SatCom capabilities of smartphones and low-cost IoT terminals with very small antennas. A possible solution to address this issue would be deploying spacecraft with very large antenna apertures. In order to meet the stringent requirements to close the link with *user equipment* (UE) with minimal SatCom capabilities, the current spacecraft's antenna apertures should be increased by at least a factor of 10 to reach apertures of tens of squared meters. The likelihood of having even more stringent requirements is high, given the dependency of the beamforming gain on the link budget. However, this is entirely impractical since the associated costs do not scale well with the antenna aperture increase for a monolithic satellite. Indeed, deploying such satellites is rather costly, considering the design/build costs and the launch costs of these bulky pieces of equipment.

On the other hand, the current trend in the space economy, commonly termed *NewSpace*, relies on miniaturization, mass-production, and increased capabilities of spacecraft. Instead of launching one large, monolithic spacecraft, many small, sometimes kg-sized, spacecraft can be launched, benefiting from advances in software capabilities and electronics miniaturization. These novel space systems change the paradigm of space-mission design by adding redundancy (due to multiple

spacecraft) and lower capital expenditures (due to cost spread over time). In terms of comparable beamforming capabilities, a state-of-the-art *low-Earth-orbit* (LEO) satellite weighing about 1000 kg can be replaced by small satellites collectively weighing 100 kg or less. In this context, a promising solution is to let a swarm of small spacecraft act as if it were one (non-monolithic) satellite with a huge *virtual antenna array* (VAA) that can establish high data-rate links with UE with minimal SatCom capabilities via *distributed beamforming* (DBF). The main challenges related to the spacecraft's physical limitations and flight dynamics affecting the swarm include: the accuracy and miniaturization of position-keeping systems, which cannot maintain unchanged the geometry of the swarm under realistic flight dynamics; the requirements of very accurate clock/carrier-phase synchronization; the existence of grating lobes and long propagation time-offsets due to large inter-satellite distances; and the power/computational requirements for the DBF operations.

In addition, apart from DBF for antenna gain increase, for transmission to multiple UEs on the ground the possibility for *distributed spatial multiplexing* (DSM) from LEO satellite swarms would also be of interest. Multi-stream transmission would not be possible from the antennas of the same satellite since their wavelength order of magnitude distance and the strong LoS characteristics of the satellite-to-ground channels would result to a low rank of the corresponding multiple-input and multiple-output (MIMO) matrix. A way to counteract this and increase the rank of the MIMO channel would be to leverage the very large distances among LEO satellites and enable them to send independent streams in a coordinated fashion, hence operating as a virtual array for DSM, similar to DBF operations.

#### *8.3.2.3 Beam Management*

Beam management in 5G NR refers to the set of procedures, and in particular of beam sweeps, that permit identifying a suitable gNB-UE beam pair for communication. Although they have been designed for terrestrial applications, they offer a good base line support to NTN as well. However, enhanced beam management solutions will be required for B5G and 6G NTN systems.

First, they need to support multi-connectivity schemes across the 3D network. Therefore, current solutions need to evolve to multi-beam solutions capable of identifying multiple beam pairs in different layers of the 3D network. For fast and efficient operation, the capabilities of active antenna arrays need to be fully exploited enabling hierarchical, multi-beam and-or multi-band beam sweeps. Besides, geolocation information must be efficiently exploited to reduce the angular sectors that need to be scanned thus reducing the time to acquire the satellite position/signal. Second, in 6G a more flexible use of the spectrum could be envisaged, permitting frequency reuse across the different vertical layers of the network or between TN and NTN resources. This framework will require an interference-aware beam management, which should not only identify a suitable beam pair, but also discover the presence of potential interferers and mitigate them through null-steering solutions. Third, beam management will not only be needed in terrestrial communications or in space-to-Earth/ Earth-to-space links, but also for inter-satellite links, which may be realized within the same layer or across different network layers.

On the standardization side, ISL represent a new framework since the gNB-UE roles cannot be assigned, thus new beam management procedures are needed. On the technical side, ISL may leverage on the same beam management solutions, but the requirements in terms of mobility will be different from those including a terrestrial terminal and should be properly considered. Finally,

beam management needs to be further enhanced to allow joint communication and sensing solutions to perform environment sensing predicting link blockage for example or for increasing Space Situational Awareness.

### 8.3.3 Integrated Communications and Sensing

Driven by the need to reduce costs and to exploit at best the available payload resources depending on the overflown area, there has been lately a growing interest for multi-purpose low Earth orbit satellite systems. This new paradigm fosters the development of systems able to support communication services as well as Earth observation, sensing and/or navigation/localization applications. Different approaches can be considered to realize such multi-service networks. A first solution consists in a flexible allocation of on-board digital processing resources to support digital communication functionalities when flying over densely populated areas while enabling otherwise post-processing of Earth observation/sensing signals. The advent of highly reconfigurable and powerful system-on-chip for space systems is key for the further development of such systems. The agile distribution of the data processing load within the network also requires a tight coordination of the payloads in the constellation via intersatellite links. The inherent distributed architecture of constellations and swarms can also be used to apply Frequency Difference of Arrival (FDOA) and Time Difference of Arrival (TDOA) techniques on communication signals to locate emission sources without the need to transmit position data. Another approach driven by the multi-mission paradigm is the introduction of novel waveforms able to fulfil the requirements for both communication (e.g., IoT) and observation/sensing/localization application scenarios. An opportunistic exploitation of existing communication signals for non-communication services can also be envisioned. Sharing waveforms for multiple applications is of high relevance to address spectrum congestion issues and enable a more aggressive reuse of frequency resources.

The success of multi-purpose systems can be evaluated via the following key performance indicators:

- Overall utilization of on-board payload resources in the distributed satellite network
- Number of services supported in a given frequency band

### 8.3.4 Next Generation Multiple Access and Resource Management

The full integration of NTN into 6G will not be possible with incremental technology evolutions. Among others, breakthroughs in the massive reuse of resources –time, frequency, beams, and antennas, among others- and very large antennas to provide adequate link budgets to handheld terminals are called for a competitive and complementary role of satellite services. Adaptive Resource Allocation becomes a many-dimensions problem, with the need for real-time optimization of beam placement and shaping, beam-gateway routing, frequency assignment and power allocation, all to serve time-varying non-uniform user distributions. All these subproblems are entangled, so that system level AI-based approaches will have to be explored to reap the benefits of this holistic approach. The full potential of beam-free approaches, with several satellites in sight from a given location, and each with high reconfigurability capabilities at the radiated waveforms, is also to be disclosed, following the emerging enthusiasm with promising terrestrial cell-free solutions.

#### *8.3.4.1 Multi-satellite and multi-RAT connectivity*

The coexistence of satellite systems at different orbits is foreseen to form a very dense space network. Accordingly, it is highly likely that users could establish links with multiple satellites in the field of view. In non-geostationary (NGSO) constellations, as soon as the visible satellites approach the local horizon, new satellites take their place. Hence, it is important to remark that some links are only available for a short period of time. The presence of multiple satellites is traditionally exploited in NGSO constellations to ensure service continuity by means of inter-satellite hand-over (HO) mechanisms. Having in mind the objective of enhancing the spectral efficiency as well as the transmission reliability, a different approach can be followed to operate the satellites in a coordinated fashion. It is worth highlighting that the challenges and the potential solutions are specific to the layer at which the signals from the different involved satellites are combined. In alignment with the scope of this section, we focus the attention on the combining at PHY layer.

One option to benefit from multi-satellite transmission is to separate satellite signals at the receiver, e.g., in time, frequency, space or code domain, and employ a dedicated synchronization circuit for each signal. This implementation allows compensating the time and the frequency misalignments individually on each link, which are especially significant in NGSO satellite systems. Then, the combination can be performed either at the symbol or at the bit level. The former option imposes the same modulation and coding scheme (MCS) in all the links that belong to the cooperative set. By contrast, the combination at the bit level provides more flexibility on the mapping and the adaptive coding and modulation (ACM). This is an opportunity for future research activities to cover enhanced schemes that optimally allocate the information and the redundancy bits among the resources of the cooperative satellites and decide the MCS of each link. The goal is to satisfy the QoS requirements with the highest spectral efficiency. Another aspect that deserves research efforts is the buffer requirements to successfully apply message combining, mainly when satellites could be located at different orbits and may be subject to different round trip time (RTT) delays. In this multi-connectivity scenario, the HO and the hybrid automatic repeat request (HARQ) need to be revisited in order to take into account the particularities of the satellite channel as well as the availability of several satellites' links per connection. For example, the HO triggering event has to be properly configured to ensure service continuity on the different links. Additionally, HARQ should increase the number of processes as well as the transmission/reception buffers along with efficient mechanism of retransmission request over the link with the best quality for instance.

There is another approach to exploit joint transmission of multiple satellites, which departs from signal separation. The technique consists of combining the signals coherently, by exclusively using a single synchronization circuit. This entails achieving a synchronized reception in time, frequency, and phase. Unfortunately, the framework specified by the latest 3GPP releases to serve users coherently by multiple transmitters cannot be applied to satellite communications, as the working principles are valid for static transmitters that exhibit similar RTT delays. This is not the case with most satellite constellations, where usually inter-satellite distances are larger than inter-site distances in terrestrial deployments. The immediate consequence is that perfect synchronization at the transmitter does not always imply perfect synchronization at the receiver, due to the large propagation delay between satellites. This impairment together with the continuous slang range variation between ground users and the target satellite are the main obstacles to directly benefiting from the coordinated multi-point (CoMP) schemes specified in 3GPP. It is noteworthy to mention

that recent works have proposed joint transmission mechanisms, where multiple satellites serve multiple users by means of a distributed precoder. Different solutions have been investigated to compute the precoding matrix using localization or channel state information in a centralized or decentralized manner, by exploiting inter satellite links (ISLs). Typically, the synchronization requirements of distributed precoding are very strict, and thus their fulfilment will introduce additional requirements on the inter-satellite distance. To overcome possible constraints imposed by the synchronization, it is interesting to investigate coherent schemes that are tolerant to the differential Doppler frequency shifts and the propagation delays that arise in the multi-satellite transmission context. Remarkably, the adoption of the orthogonal time frequency space (OTFS) modulation can be more efficient than single carrier or orthogonal frequency division multiplexing (OFDM) schemes to deal with multi-satellite connectivity scenarios. In this respect, OTFS can deal - up to some extent- with the satellite channel dynamics by inserting one cyclic prefix (CP) per frame, which encompasses several OTFS symbols. Hence, the use of advanced waveform designs that facilitate the application of distributed precoding in multi-satellite transmission schemes could be a promising starting point.

#### *8.3.4.2 Next Generation Multiple Access*

The envisioned network of 6G supports a multi-layer architecture as sketched in Figure 8-4, with a massive number of participants and moving components, each with different radio environment characteristics and impacts. Thus, techniques for advanced spectrum management are necessary in order to fulfil the IMT2030 requirements by the support of dynamic RAN components in hybrid 6G Networks.

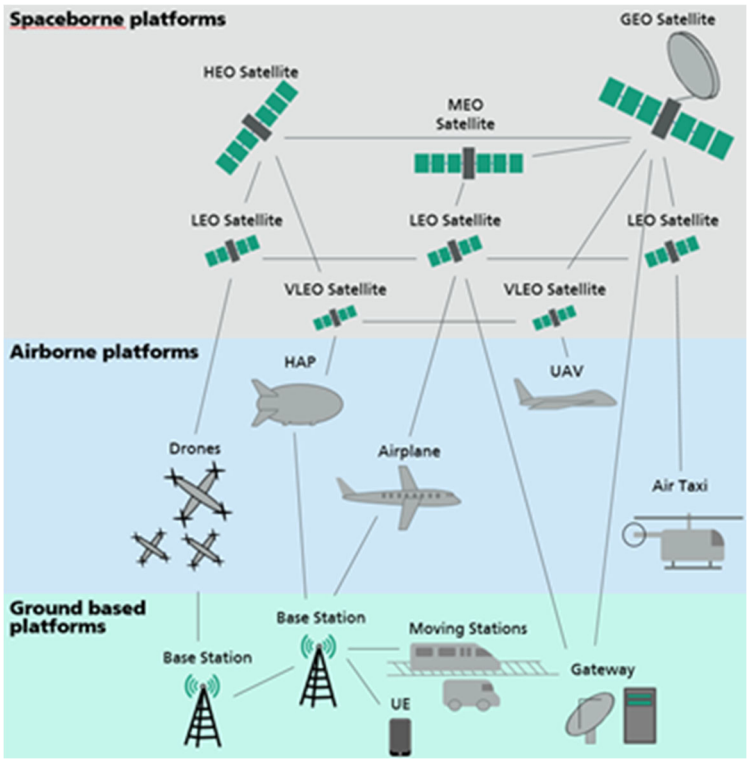



*Figure 8-4 Heterogeneous and dynamic radio access network for 6G*

For the next generation RAN of 6G, we investigate physical layer techniques, which may facilitate the handling dynamic massive connectivity and radio access. The current multiple access techniques

are divided into two categories, namely orthogonal transmission strategies and non-orthogonal transmission strategies. Due to low complexity and interference avoidance, orthogonal transmission strategies have been extensively employed in practical wireless communication systems e.g., orthogonal frequency division multiple access (OFDMA) in the 4G, where users are allocated with orthogonal frequency resource blocks. However, given the ever-increasing number of wireless devices and the limited amount of available spectrum, conventional orthogonal transmission strategies have become inefficient due to their low spectral efficiency (SE) and the limited number of users/devices that can be supported due to the inflexible resource allocation. Thus, concepts based on non-orthogonal radio access (NORA) concepts in downlink and uplink are interesting candidates for radio access in next generation networks. Specifically, concepts based on non-orthogonal multiple access (NOMA), space division multiple access and multiplexing (SDMA) and Rate-Splitting Multiple Access (RSMA) are promising. In principle, these techniques are already well investigated. However, putting these techniques into the context the 6G RAN arises new unknown challenges, which are related to the key requirements of

- Enhanced QoS
- Energy efficiency
- Dynamic hybrid systems

In summary, this leads to new system design parameters

- Drivers system scenarios
- Heterogeneous traffic
- Massive access and connectivity

These new aspects necessitate the investigation of new physical layer techniques for SISO as well as MIMO transmission e.g., advanced precoding and receiver algorithms, in consideration of developments in other domains. Advances in distributed/federated satellite on-board signal processing may open up new directions for joint communications and sensing (JCS). AI may be helpful to encounter the highly complex 6G systems, where developments in hardware domain enables the implementation of these techniques. NORA concepts can be applied for multiplexing users in the downlink as well as for the uplink multiple access. However, the 6G key requirements imply unknown challenges to the uplink, as the uplink received only little interest from academia in general so far (in contrast to the downlink). Hence, a special focus shall be given to the uplink by the investigation of the next generation multiple access (NGMA). The ultimate goal of NGMA is to enable a tremendous number of users/devices to be efficiently and flexibly connected with the network over the limited wireless radio resources. The key idea of non-orthogonal transmission strategies is to allow different users to share the same resource blocks. To handle the interference, extra techniques like superposition coding (SC), rate splitting (RS), successive interference cancellation (SIC), and message passing (MP) are used. Although the non-orthogonal transmission strategies increase the transmitter and receiver complexity, but significant benefits can be achieved, such as supporting massive connectivity, achieving high SE, energy efficiency (EE) and guaranteeing user fairness. These benefits make non-orthogonal transmission strategies promising candidates for NGMA. In next generation wireless networks, NGMA has to facilitate diverse application scenarios and be compatible with other promising physical layer techniques along with achieving high bandwidth efficiency and supporting high connectivity. Finally, it shall be emphasized that the

investigation of NORA/NGMA does not exclude the use of conventional orthogonal schemes for 6G systems. In contrary, a suitable alternation/combination of orthogonal and non-orthogonal radio access is most likely to be the solution to address the challenges of dynamic massive connectivity and radio access for next generation hybrid networks with satellite-aided and terrestrial communication.

#### *8.3.4.3 Spectrum Sharing*

Dynamic spectrum management needs to be updated to the 6G era, taking into account the multi-layer architecture, application requirements, and ever higher frequency bands to be utilised. There are many topics to be addressed to make this successfully, including regulatory domain. First, defining the most suitable frequency bands for systems and links, as well as use cases where dynamic spectrum sharing is possible. Second, developing spectrum sharing mechanisms to manage the complexity of a dynamic and mobile 3D network. There can be satellites in different orbits as well as terrestrial and airborne systems sharing the same frequency bands. In order to coordinate the spectrum use and enable predictable QoS for all users, information must be shared between the entities controlling the involved layers or towards a third-party entity controlling the spectrum use. This entity could be a database-assisted system that limits the number of secondary users accessing the primary user spectrum as well as provides means to control their resource use.

Key Value Indicators:

- Ultra-flexible satellite offered capacity
- Unified chipset and telecom infrastructure
- Seamless interoperability
- Smartphone satellite connectivity for voice and video.

Key Performance Indicators:

- User-level service level agreement accomplishment.
- Spectral efficiency
- Terrestrial and non-terrestrial handover delay.
- Radio terminal power consumption.

## 8.4 Research theme: Network of Networks

### 8.4.1 Network Architecture Evolution Perspective

#### *8.4.1.1 Dynamic NG-RAN functional splitting*

5G functional split is one of the main concepts introduced to push forward the idea of optimised network operations, relying on deep network softwarisation principles (e.g., cloud-RAN, SDN/NFV, data/control planes programmability). In this respect different split options have been proposed and discussed, whose applicability pretty much depends on the performance implications they bring and eventually on the requirements exposed by the verticals operating on the considered 5G network. More advanced investigations have also considered the possibility of dynamically selecting the most appropriate splitting model on the basis of the requested QoS guarantees by means of deep network softwarisation concepts combined to AI-based solutions. It is straightforward to see

that this challenge becomes even more evident in the case of 3D-NTN networks comprising multiple space assets, so that the identification of specific network functions and their related positioning across the space segment is a very complex problem from both network performance optimisation and overall network implementation standpoint. The functional splitting of NG-RAN across data and control planes has been already partly addressed in the context of "traditional" NTN networks, where the split is operated directly on ground or in a more advanced configuration in space assuming a satellite regenerative payload. The latter case is however limited to standalone NTN assets, so that the problem of functional split distributed in space is not so critical from a complexity standpoint, although a deeper understanding of all related performance implications needs still additional studies and related validations.

#### *8.4.1.2 Cognitive-based Intent-Based Networking for 3D-NTN.*

A unified 6G-NTN large scale network will be able to adapt the network operation to external stimuli while respecting a set of operational intents. Such cognitive network will rely on a set of distributed agents embedded in the network with computing, communication, caching and cognitive capability. Each of these agents will be controlled by intents that specify expectations including requirements, goals and constraints. Each agent will encompass networking functions as well as cognitive functions that observe the environment under control and draw conclusions from the acquired data, in order to take actions to adapt the operation of deployed services based on current or predicted network behavior while fulfilling the expectations of the intents (expressed by Humans). In order to support such self-organized behavior, there is the need to include mechanisms able to support the automatic translation of operation intents into end-to-end services defined as a set of network functions. The cognitive functions use Artificial Intelligence (AI) features to draw conclusions from raw data. Examples of such features include machine-learning models that can produce insights and self-organized capabilities (e.g., particle swarm optimization) to execute on intents. Besides the usage of AI and self-organized optimization to furnish the needed automation and prediction, the envisioned management plan needs to be able to interact with a variety of network technologies, such as service function chaining, software defined network and network function virtualization. This means the networking functions implemented in network agents are software modules installed in a virtual infrastructure and used to create service chains that have an end-to-end meaning. In this context, of importance is the development of methods for virtual network function orchestration, namely to decide about the best place to install network functions to fulfill operational intents. Due to the large volumes of data that may be transported in a unified 6G-NTN network and the spatial distribution of computing and resources (e.g., on the 6G RAN, ground station or even inside the space segment), the self-organization of such network may pass by using distributed artificial intelligence, which is an approach suitable to solve complex learning, planning, and decision-making problems, as the ones that may show up in a large scale 6G-NTN network. The usage of distributed, and trustful AI aims to facilitate learning and perception to solve problems that require large datasets, through distributing the problem to autonomous processing agents.

#### *8.4.1.3 Programmable Data Plane*

A future unified 6G-NTN network will address a large variety of end-to-end services, each one with a specific set of networking and computing requirements. The current development of Edge Computing solutions is the initial set to allow the network to host more flexible services. However, there is the need to analyze different networking solutions to move beyond packet interception as

the basis of network computation. While existing solutions employ rudimentary languages for programming network elements, a richer programmability framework is required to support emerging workloads, such as edge network analytics, and distributed machine learning. There is the need to investigate new solutions to address the operation of large-scale network as is the case of unified 6G-NTN, including automation, self-management, orchestration across components and federation across network nodes to enable emerging services and applications. Strong consistency in distributed networks, such as 6G-NTN, are important. However, maintaining consistency requires multiple communications rounds in order to reach agreement, leading to a potential creation of bottlenecks in the network. In this context, consensus mechanisms for ensuring consistency are of extreme importance in managing large amounts of data related to the operation of large networks. In this scenario, one initial challenge if to find a good tradeoff between reducing the coordination overhead, while keeping the number of inconsistencies low.

#### *8.4.1.4 Effective network slicing driven by AI-based network orchestration*

Network Slicing is considered an important enabling technology for the next generation 5G network architecture for multiple logical networks over a common physical infrastructure. Several VNFs can be embedded together to form a network slice associated with a particular networking service. VNF embedding and placement over several virtualized networking platforms is an important problem to be solved. Different services can have heterogeneous requirements and VNF placement needs to consider such requirements before performing VNF placement. Each service request comes with a specific demand. Required service latency and the data rate are two main characteristics of demanded services. As an example, for avoiding the excess transmission delay between VNFs, to satisfy the critical latency requirement of service, it is preferred to place VNFs associated with URLLC services in proximity. On the other hand, for satisfying the user data rate requirements of eMBB, several VNFs of eMBB slices can be placed on platforms with large computation capabilities. Such an approach can have many benefits for serving many users with efficient utilization of EC resources. An ML-based intelligent strategy can be applied for designing a function for associating a proper weight for each VNFs placement over different platform layers. By analyzing the individual service request data and the available resources of EC platforms, an intelligent approach can be defined, for solving both slice function placement and resource allocation problem over a multi-EC enabled VN.

### 8.4.2 Network Orchestration/Management

#### *8.4.2.1 Orchestration for converged NTN -TN Infrastructures*

The orchestration and management of network functions over terrestrial infrastructure has been largely investigated in recent years. The achievements have enabled the standardization of a Network Function Virtualization (NFV) Management and Orchestration (MANO) by the European Telecommunications Standard Institute (ETSI). The operation features and performance achieved with this system to manage the network has contributed to the usage of these mechanisms over satellite systems. These systems are characterized by being resource constrained systems with the capability to serve a wide and heterogeneous number of users in a single covered area. The ability to manage services or network functions in a satellite system enables them to optimize their resources, as well as differentiate services among the users in the satellite footprint. The benefits of this approach have been shown in prior work, demonstrating this flexibility to manage satellite resources and services. Additionally, the location of NFV and the network controllers (of an SDN

architecture) have also been previously studied. There are benefits of deploying these network controllers on ground and in-space, depending on the constellation architecture. Specifically, the location of the controllers in the TN enables achievement of a low latency service for small constellations. On the other hand, these controllers provide a better system performance if they are located in space when managing large constellations. The technological challenge still remains in how to integrate current operational MANO technologies—designed for TN—with NTN infrastructure. The development of a unified orchestrator would support service ubiquity, flexibility, scalability, and cost-efficiency towards the realization of NTN- and TN-as-a-service.

The envisioned goal of an autonomous orchestration of services in NTN infrastructure can be achieved by modifying current orchestration technologies (i.e., MANO). Specifically, a MANO technology capable of deploying services in NTN and TN infrastructures would simplify their management by satellite-agnostic network operators, and facilitate their integration. Future research will develop a framework capable of managing satellite and UAV mobility, and their intermittent connectivity with ground systems. This temporal visibility will impact in the monitoring of the services, as well as anticipate service execution over specific areas, thus achieving a continuous service strategy.

In this respect, the complexity and peculiarity of NTN might be best managed by conceiving a specific orchestration level, interacting with the MANO Network Functions Virtualization Orchestrator (NFVO) of the TN through well-defined APIs, in order to present an abstract view of the space network, with Network Control Centers acting as “mediation points”. A similar role is currently played by the Telco Operations Support System (OSS) in architectural solutions based on the separation of concerns between Vertical Applications and Network Functions. In this type of frameworks, a Vertical Applications Orchestrator (VAO) takes care of the orchestration management of the chains of micro-services that make up an application graph, and conveys the communications requirements toward the NFVO by means of a Slice Intent descriptor through the NorthBound interface of the Telco OSS; the same interface is used by the NFVO to dynamically interact with the VAO in the backward direction for slice adjustment and adaptation. In the interaction/integration between TN and NTN, the orchestration management of the latter might be delegated to a specific TNT-NFVO, in order to relieve some of the burden from the TN-NFVO, with a clear definition of the respective roles.

#### *8.4.2.2 Orchestration Management in 3D Networks*

The orchestration of services and resource management of NTN infrastructures is directly related to the network architecture. Specifically, each node (satellite or aerial) in the NTN architecture is characterized by different features depending on their design (e.g., available resources and capabilities) and their location in the multi-layered structure. Node mobility is one of the technical challenges that needs to be addressed to achieve the envisioned service orchestration. For instance, GEO satellites are characterized by having a wide coverage area in a fixed region in which large numbers of users can be served, although the propagation time is not negligible. On the other hand, LEO satellites may have global coverage with low propagation time, but their visibility in a region is temporal. The service orchestrator in this scenario must leverage on this heterogeneity and cope with node mobility. Developments for satellite constellations have leveraged their predictive movement. Nevertheless, their approach is mainly centered on a predefined and custom

constellation (i.e., Iridium, Starlink). Another relevant feature of NTN architectures is the interconnection of their nodes, composing the entire network. Depending on these interconnections and the mobility effects, different network approaches can be deployed: snapshot Networks, LEO Satellite Networks, Multi-Layered Satellite Networks (MLSN), and Delay/Disruption Tolerant Network (DTN) are examples of these heterogeneous architectures in the satellite domain. Each of them is characterized by managing incoming traffic differently (e.g., store-and-forward, hierarchical routing, predictive approach, etc.), which impacts on the service that can be provided (e.g., delay-tolerant, low-latency, etc.) The service orchestration can take consideration of these features, and manage services depending on their characteristics. This heterogeneity is also represented in the resource capabilities of each node. Different satellite platforms may coexist in the same NTN, as well as HAPS systems and UAV's. All present different resources and capabilities that need to be optimized depending on the generated traffic. For all these aspects, the coordination and orchestration of services in this heterogeneous and dynamic network requires optimization mechanisms that considers all these aspects, and handle infrastructure resources (e.g., power, bandwidth, time, space dimensions, nodes, coverage, and topology) for a more flexible and dynamic system with overall better performance, efficiency, and sustainability.

Research will contribute to the optimization of resource orchestration applying AI-based techniques. Specifically, an efficient and adaptive orchestration can be developed according to traffic and NTN status. Based on Open Source MANO (OSM), these novel orchestration systems will manage services and resources taking in consideration the dynamism of the network (due to satellite movement), topology hierarchy, incoming traffic, and satellite/aircraft resource availability. This orchestration will be supported by AI/ML techniques which will take the decisions to allocate the resources, such as RL-based radio resource allocation for multi-beam satellites with regenerative payload. The orchestration will ensure the transfer of the service context over satellites/AUV's in order to ensure the continuity of this one over a target area, making it transparent to the end-user. The aim is to develop MANO, based on OSM, capable of interacting with NTN and to orchestrate services in this dynamic and hierarchical infrastructure.

In this management and orchestration framework, a hierarchical control architecture of some sort should be foreseen, where the parameters characterizing control strategies acting at fast control loops can be automatically adjusted by means of online adaptation, possibly performed by functional approximators applying AI/ML techniques. In the space segment, this entails the presence of ground-based coordinating units, capable of constantly monitoring and dynamically reconfiguring the interconnection, routing, scheduling, caching, and off-loading operations performed by TNT nodes operating at faster time scales. In such hierarchical orchestration framework, techniques to ensure that QoS requirements are preserved when traversing multiple cascaded queues at different orchestration levels, possibly performing aggregation, different protocol encapsulation and framing, should be devised. Their design can range from the application of model-based control (classical queueing models or continuous flow models) to AI/ML algorithms (with the latter not excluding the adoption of models to describe the systems' dynamics, when available).

### 8.4.3 IP-Forwarding Payload

#### *8.4.3.1 Context*

Traditionally most communications satellites have acted as bent-pipes for a variety of very valid reasons including (a) the desire to maximize the throughput by ensuring as much of the satellite power is used for transmission as possible; and (b) maximizing the satellites' in-orbit operational usefulness by keeping all service defining aspects that might change from time-to-time on the ground. Three factors have changed significantly in recent times. Firstly, the costs of all aspects have dropped significantly from the launch cost through re-usable launch vehicles to the cost of space-worthy electronics driven by the resulting "newspace" marketplace. Finally, the massive increase in software-defined systems for almost every possible communication component means that space-born assets can be massively reconfigured post-launch. This means that perhaps the spacecraft economical operational life can be reduced and that systems on the spacecraft can be considered for data management such as the IP-Forwarding discussed in this section and the routing in space discussed in the next.

#### *8.4.3.2 Networking in the presence of intermittent connectivity [C8-3]*

Independently of the topology (star or meshed), a satellite network may use either a non-regenerative or a regenerative satellite platform. The former is commonly called" bent-pipe", since the satellite does not terminate any networking layers, simply transferring signals from the terminal links to the feeder link. On the other hand, regenerative payloads, also called On-Board Processing (OBP), provide additional functionality in the satellite, based on the capability of terminating one or more layers of the protocol stack. This capability allows the internetworking of the satellite system with Internet Protocol (IP) networks at different levels, namely internetworking below the IP layer (Bridge internetworking function), at the IP layer (IP internetworking function) and above the IP layer (Gateway function).

The operation of large-scale satellite networks requires OBP platforms able to implement at least the layer 3 of the IP protocol stack. The ability to route traffic in space helps to reduce latency especially in the case of multi-layer satellite system so to avoid multiple hops over ground and space segments (GEO satellites already experience 250 ms summing up down- and uplink latencies), as well as to reduce satellite transponder costs. Routing IP traffic on board satellites can increase throughput and enable flexible bandwidth-on-demand real-time applications between users in different geographic regions without static configuration. However, forwarding data in space faces the challenge of dealing with the intermittent connectivity of the satellite system. One of the solutions devised to tackle such challenges is the usage of a Delay Tolerant Networking (DTN) platform. In this context, the DTN bundle protocol allows the chaining of different TCP sessions between custodian satellites, in order to achieve end-to-end connectivity over a set of intermittently connected satellites. From the routing viewpoint, data exchange is rather challenging in such networks since paths between any pair of nodes may never exist or delay may be too long to be accepted by current data transport protocols. There are already a significant number of proposals to opportunistically route data based on time-variant graphs used by DTN, each with a different goal and based on different evaluation criteria.

#### *8.4.3.3 Content-oriented networking in space*

Since the creation of the Internet, the volume of exchanged traffic has grown considerably, from less than 100 GB per month in the late 1980s to an expected volume of nearly 400 billion GB this

year. Due to the fact that the Internet is still growing, in terms of traffic and coverage, with the envisioned integration of NTN and terrestrial networks, there is the need to investigate suitable architectural alternatives to the existing one. One of the most prominent future Internet architectures is Information-Centric Networking (ICN), which addresses data using data identifiers and forwards packets based upon such identifiers instead of host identifier. This shifts the current host-centric Internet paradigms towards a new data-centric approach. ICN enables a consumer to request a given data object in the network without any knowledge about the location of the requested data. The paradigm shifting from a host-centric to a data-centric approach brings several benefits to the operation of large-scale satellite networks, namely the adaptation to intermittent connected networks based on a pull communication model and in-network caching, as well as extra flexibility to handle different types of traffic, based on an extended set of forwarding strategies. While some analysis about the development of ICN based satellite systems have been made, some new ICN based architectures have been proposed to support a universal networking system able to encompass also space borne and airborne platforms.

#### *8.4.3.4 Traffic engineering and flexible forwarding*

In order to support a scalable management of large-scale networks, as is the case of a unified 6G and NTN network, there is the need to develop a flexible networking system able to manage large amounts of traffic over a mobile mesh network encompassing heterogeneous nodes (e.g., different types of satellites) connected by high-speed wireless links (e.g., free space optical links). Technical solutions are needed to devise a suitable management/control plane able to sustain large scale 6G/NTN networks taking into account the capability of: i) exploiting a set of free space optical links as well as omnidirectional radio links; ii) forward packets at high data rates; iii) isolate and serve traffic from different types; iv) control congestion, over intermittently available links and devices v) provide differentiated forwarding paths for specific types of traffic over a large scale network while respecting sets of Service Level Agreements (SLA). Although large-scale satellite networks will exploit high-throughput FSO links, the probability of congestion is high due to resource constraints (e.g., energy constraints on satellites) and network dynamics (e.g. intermittent connectivity). Hence there is the need to control the traffic load in the network, taking into account differentiated QoS treatments. Due to the particular properties of ICN it is assumed that the direct application of IP QoS models (e.g., IntServ, DiffServ) may not be the best option. Hence there is the need to develop a QoS framework able to exploit the particular properties of ICN, such as in-network caching and different forwarding strategies, in order to provide controlled load of differentiated QoS services. Another challenge is related to the possible locations for traffic engineering functionality, such as load balancing: distributed in routers or implemented in a controller. In the former case, load balancing is achieved by a forwarding strategy, allowing routers to select the most-fit forwarding face for every packet. Although this method has the advantage of not having a single point of failure, it can be demanding in terms of resources. Solutions based on a controller rely on periodic probes to collect metrics from service replicas. For instance, based on collected information about link utilization and switch/router load, a reinforcement learning agent can adapt the link weights aiming to balance load in queues, or can compute new optimal paths upon congestion, being the forwarding process implemented in P4.

#### *8.4.3.5 Content distribution*

State of the art: The use of GEO satellites for the distribution of content to consumers, known as direct-to-home (DTH), is a well-documented success story for the sector. Whilst some erosion if this market is foreseen it is holding up quite well in many countries despite the growth of over-the-top TV distribution such as Netflix. Various studies have considered how GEO satellite capacity may be used to distribute both live streams along with cache content to TN RAN nodes for onwards distribution towards the end users however this analysis is incomplete as is the best way to serve nodes connected using LEO for their unicast traffic.

Beyond state of the art: Research into the changing landscape for content will be needed – most likely through review of more detailed research in to the various technologies. Demand can be anticipated from applications such as the extended, augmented and virtual realities spectrum, new consumer TV standards such as 8k TV, developments in gaming, edge computation, and so on. Each application will have its own set of unique requirements in terms of reliability, latency, end user bit rates and standards followed. Techno-economic analysis will be needed to determine where the major issues are that a satcom solution can be assist. Technical analyses of the role for satcoms will be needed to determine how best to serve this evolving landscape. Sites served by non-GSO NTN connections have transient and moving coverage not best suited to multicast/broadcast content. Options for GEO to complement these locations include using a terminal capable of receiving direct from GEO whilst communicating via the other NTN connection and the GEO transmitting to the NTN nodes for onwards relay. Developments in underlying technologies and their integration into the 6G fabric such as content-centric networking, MEC based caching, fog and cloud-based systems; all in a 3D network of multiple NTN and TN connections.

### 8.4.4 Routing in space

#### *8.4.4.1 Context*

Global- and low-latency connectivity in space requires a routing algorithm for deciding, either at the terrestrial or space source, or in every intermediate node, the directions to be used to reach the terrestrial or space destination. Due to the predictability of the network topology, the conventional approach to routing in NGSO constellations is to centrally compute all the paths in a terrestrial location register, or a GSO satellite if it is a hybrid NGSO/GSO network, and then broadcast the information to all the satellites. In any case, this approach does not scale well, especially in mega constellations, and it creates a dependency on the limited contact times with the terrestrial or the high latencies of the GSO segment. Other challenges of space routing are related to the dynamic imbalanced load and the stringent energy and computing constraints.

#### *8.4.4.2 Semantic Routing*

Large scale satellite networks aim to transfer large volumes of data between tens of thousands of satellites that move continuously at high speeds in different orbits. To support this aim, there is the need to develop new protocols for routing traffic from different services. Such protocols can leverage existing solutions to route data over a large set of mobile devices, based on specific algorithms such as Contact Graph Routing as well as new paradigms such as information-centric networking. Future satellite networks should be designed to carry traffic belonging to different services with distinct specifications in terms of traffic performance, reliability and robustness. However, the continuous motion of satellites poses significant difficulties to traditional routing

protocols. For instance, as satellite constellations become very large, the routing may never fully converge, resulting in a sub-optimal network. Moreover, there is the need to develop routing strategies able to consider different service semantics described by a combination of fields in the packet header as well as a transported set of instructions. Such routing strategies require a data plan able to support programmable network functions (e.g., forwarding) and services. Semantic routing is the process of achieving enhanced decisions based on semantics added to IP headers aiming to provide differentiated paths for different services. The additional information or ”semantics” may be placed in existing header fields (e.g., the IPv6 Traffic Class field), may be added to new header fields, or it may be encoded in the payload or on additional headers, such as the IPv6 Extension Header. The application of semantic routing allows packets from different services to be marked for different treatment in the network. The packets may then be routed onto different paths according to the capabilities and states of the network links and nodes, in order to meet the performance requirements. For example, one service may need low latency, while another may require ultra-low jitter, and a third may demand very high bandwidth. Examples of existing semantic routing usage in IP-based networks include: i) using addresses to identify different device types so that their traffic may be handled differently; ii) expressing how a packet should be handled as it is forwarded through the network; iii) enable Service Function Chaining (SFC); iv) forwarding packets based on carried data rather than the destination addresses; v) or formatting geographic location information within addresses.

### 8.4.5 Advanced networking trends for 3D-NTN

#### *8.4.5.1 Service-centric Networking*

In 6G networks, an integrated terrestrial - NTN network can address gaps that currently exist in terrestrial networks. Services carried over such an integrated network are expected to cover broadband services, mobile backhaul, Internet-of-Things and Vehicle-to-Everything.

To support the envisioned set of future services [C8.4], besides being able to route traffic within the space network, there is the need to devise intelligent and flexible networking solutions able to support not only packet switching, but also data storage and processing, while being able to react in real time to the requirements of new operational intents. Such future satellite network aims to reduce data rate requirements, increase energy efficiency, and guarantee end-to-end network connectivity. Therefore, a new network architecture for large-scale multi-orbit satellite systems should be able to abstract a set of networking, storage and computing resources in the form of end-to-end services, deployed to fulfill a set of operational intents that may change over time. To sustain a large set of services, the envisioned network architecture should rely on a programmable data plane, flexible enough to support different services based on a chain of virtual network functions, such as distinct forwarding mechanisms.

From an end-to-end perspective there is the need to develop routing strategies able to consider different service semantics (e.g., load balancing), as well as to exploit a mesh of free space optic links. This goal may require converging optical transport with routing functionality, increasing power efficiency and scalability.

As mentioned before the new information-centric networking paradigm brings several benefits to the operation of large-scale satellite networks, namely the adaptation to intermittent connected links based on a pull communication model and in-network caching. Although a data-centric

approach reflects better the current operation of the Internet, services are an essential component of its operation: users request services rather than data. Hence, there is the need to extend the information-centric networking paradigm to support a network operation based on services and not on data objects. Such service-centric architecture should not alter the information-centric networking primitives used to create its pull communication model, but should instead extend it to support services that are seen by the network as software functions that can be requested by service consumers.

Service function chains is an essential concept for the development of a service-centric architecture, in the same sense that" network slicing" was applied to the specific case of 5G networks. Service function chains are applied at the network layer to steer packet flows through network functions, such as security, load balancing or even DTN custodians: service chains are sequence of service instances (network functions). Packets should be tunneled between service instances using encapsulation, by using techniques such as the Network Service Header or Segment Routing, which may be more widely applicable based on programmable data planes.

#### *8.4.5.2 Non-IP networking*

One of the main results of the studies carried out in the framework of the ITU FG NET 2030 and the ETSI NGP (Next Generation Protocol, now discontinues) ISG and NIN (non-IP networking, continuing and extending the work done within the ETSI NGP ISG) relates to the fact that IP networking as currently deployed may not efficiently fulfill all requirements posed by new services or by the new radio technologies capabilities envisioned to be offered by 6G. In this context, NTN (referred often to as SAtellite-Terrestrial Integrated Networks, SATINs) may suffer from the current limitations of IP to effectively support mobility in large mobile systems, as would be the case for LEO constellations or more notably for complex multi-dimensional space network, comprising diverse space assets all interconnected in a dynamic manner. Similar motivation for developing new protocols has been the baseline taken by Huawei to conceive the so-called "new IP" in the attempt of re-engineering a new protocol able to cope with the current IP deficiencies, although the actual value of such a solution has not been fully accepted by the community and standardisation bodies, mostly because the continuous IP modernization and evolution process is supposed to address most of its present limitations. Nonetheless, the action of designing new networking paradigms is of paramount importance to let diverse technology networks to converge and effectively interwork. From this standpoint, the need of novel technologies is certainly well received in the NTN context because of the floating nature of links and the possible variations of network topologies, along with the possible exploitation of multiple space links at the same time. Likewise, the opportunity to exploit self-organised networking principles and to implement software-as-a-service functionalities call for more advanced networking technologies, able to decouple location and identity of a given service, so as to allow a more effective addressing and routing concepts as well as providing more evolved means towards content-oriented network programmability. All these building blocks can pave the way towards the exploitation of information-centric networking principles applied together th wiin-network computation concepts. Moreover, the objective of achieving a network model more flexible and agile towards online programmability is also in line with the possible exploitaiton of active networking. All these networking directions need however proper investigation for their scalable implementation in space in light of the hardware and resources constraints imposed by space nodes as well as the possible physical limitations exhibited by the space links themselves. Last

but not the least, the application of these new concepts in an isolated manner over NTN systems introduces the necessity of protocol convergence between the different parts of the NTN-6G integrated network, whereby effective data format translation and service mapping functionalities have to be implemented and supported by a multi-layer network orchestration architecture.

### 8.4.6 Expected Impact

The evolution of 5G/6G network so as to embrace NTN segments too will increase the overall complexity of the resulting integrated ecosystem, whereby novel and more agile network orchestration and routing concepts will be of key importance to support existing use cases and enable new ones. On the one hand, coordinated multi-layer multi-domain network orchestration concepts will be necessary to pave the way towards effective network operation convergence across TN and NTN sections of a 6G system. On the other hand, the demanded high flexibility of TN-NTN network architectures will call for more effective routing solutions tailored not only to the characteristics of the links composing an end-to-end path but also to the specific contents being transported, hence introducing the semantics as key attribute in the overall routing decision process. Last but not the least, the deployment of consistent network softwarisation concepts throughout integrated NTN-TN networks will have to cautiously consider the current programmability capabilities as well as the physical constraints exhibited by the existing NTN nodes. As such evolution of microelectronics components and related capabilites for use in space systems will play a prominent role to achieve fully performant and operational NTN-6G networks.

## 8.5 Research theme: Edge Computing

### 8.5.1 Scenario

Computing and communication technologies are primary examples of emerging technologies that reflect how society has been evolving and integrating such tools into its social arrangements and which effects such technologies have been producing toward the development of institutions and of the social progress. Asking how computing platforms will affect equality of opportunity in our society leads us to acknowledge that certain realities fall short of our ideals of life, culture, and gender equality. Therefore, providing access and computing equality is a mission of utmost importance for research.

Cloud computing platforms require huge investments to set up and entail considerable consumption of land as well as the energy resources needed to run them. Edge platforms, Content Delivery Networks, and other distributed services also rely on the computing infrastructures of Telco providers or small operators in the sector, which highlight the need to already have a market operating in the territory. This represents one of the main obstacles to the global deployment of these technologies and the inability to bridge the digital divide in less digitised countries and will only increase this gap.

Recent war scenarios in Europe have brought to the forefront long-forgotten realities such as the lack of bombed-out communication infrastructure, relief services, and information for the population. One of the responses that was somewhat quick to address these shortcomings was the Starlink satellite infrastructure that was generously offered to reactivate information channels. However, these scenarios are well-known realities in other regions of the world and due to the

digital isolation in which they live, they do not come to the global attention of the media. These tragic examples help to understand the strategic importance of NTN platforms and how their technological evolution has so far only been limited to improved performance in terms of bandwidth, massive access, lifetime, and equipment miniaturisation. Differently, what is needed is a step and paradigm shift by trying to bring into orbit the resources to implement emerging computing services such as artificial intelligence, high-performance computing, storage, content delivery, as well as communication and networking.

In fact, the technologies mentioned above have been mostly developed for a terrestrial and fixed infrastructure and are unlikely to be reapplied in a highly dynamic and totally wireless NTN context: the resilience of the hardware installed in orbit is very different from that on earth, as are the problems of energy supply and heat dissipation.

Terrestrial Cloud and NFV systems are designed to guarantee 99.999% system reliability by implementing redundancy mechanisms to compensate for possible service interruptions in essentially static systems and to perform periodic migrations of services and resources accordingly. A system based on orbital computing infrastructures, on the other hand, constitutes a completely different and much more complex and dynamic scenario. One only must think of the visibility time of a low-orbit satellite and the need to guarantee its reachability through inter/intra-satellite links even after the visibility period, ensuring the reachability of all the services allocated to it within the maximum service time imposed by requirements.

This incredible technological challenge, however, represents a concrete response to the need to reduce the digital divide in order to provide all nations with the opportunity to be informed and to react with the right tools to the challenges we are called upon to respond to, such as the recent Covid pandemic. This could definitely implement a digital democracy.

### 8.5.2 Motivations

Modern users and applications require an increased amount of data processing. However, user-side computation capacity is falling short, especially when new latency-critical and data-intensive applications are considered. The Cloud computing facilities can reduce the computation burden of new services; therefore, users can transmit the portion/complete task to the cloud servers having enormous computation and communication power. In general, the cloud facilities are located far from the users, in the core network, and introduce some drawbacks, e.g., huge transmission costs, backhaul link congestions, data security threats due to long-distance communications. Such issues can be addressed by integrating the Edge Computing (EC) facilities into the Network, bringing cloud computing resources in the proximity of end-users.

EC facilities can be integrated through the deployment of several EC servers in the Terrestrial Network (TN). Such an approach has achieved lots of success by enabling new latency-critical services especially for smart city and mobility scenarios. However, the limited capacity and coverage of EC servers is the main bottleneck while exploiting EC advantages. TNs are becoming more and more used, with many new users requesting services with specific demands. Limited coverage into rural and remote areas, unreliable service during natural disasters like tsunamis and earthquakes, new security challenges, poor link budget with additional interferences are some of the main challenges that need to be considered while utilizing TN-based EC platforms. With the presence of

different users, the dynamically changing resources of EC servers adds additional challenges while integrating the TN-based EC services. With this limitation, integrating TN-based EC platforms into VN alone cannot be a sufficient solution for the new futuristic services and applications, which will have more stringent requirements in terms of latency and computation resources.

Encouraged by the new technological developments and the additional interest shown by several tech giants (i.e., Facebook, Google, etc.), the Non-Terrestrial Networks (NTN), including space and air networks are growing these days mainly for providing global connectivity. Several new platforms such as new satellite constellations, unmanned aerial vehicles (UAVs) swarms, small-fuelled aircraft, balloons have been deployed at different heights from the ground users to achieve the global connectivity challenge. Better connectivity, scalability, and reliability are some of the advantages of NTN based communication platforms. With the addition of modern communication technologies, such as multi-beam antennas, the NTN platforms can also provide EC-based services with an onboard computing server. Such NTN-based EC platforms can complement the TNs for solving several problems, including limited capacity and coverage. However, the higher transmission delay, introduced by space networking platforms (i.e., satellite constellations), is a major challenge while considering NTN platforms for serving latency critical applications. Compared with space networks, new aerial platforms such as Low and High-Altitude Platforms (LAPs and HAPs) have a considerable advantage with reduced transmission distances, low deployment time and costs, and reduced communication channel losses.

In what follows the technical challenges have been organized into three main pillars, namely, Edge Architecture/System, Management, and Applications.

### 8.5.3 Architecture/System

Both TN and NTN can enable EC-based services through placements of edge computing servers along with their distributed infrastructures. Therefore, a Multilayered joint T-NTN constituted by different EC platforms over TN and NTN can be utilized for providing heterogeneous services requested by users with demanded quality. A typical service area is composed by several users demanding latency-critical and data-intensive services. The EC-enabled network architecture is constituted by several elements: users on the ground, small-cell BS, Macro BS, LAPs, HAPs, LEO, MEO and GEO. Each of the layer has specific characteristics, in terms of coverage, availability, processing power, communication rate, leading to a heterogeneous system where the multiple layers can complement each other.

Each EC platform layer of the network architecture can adapt different computation and communication strategies. Several virtualization techniques (i.e., Virtual Machines (VMs), containers) can be used for the efficient utilization of EC resources. Furthermore, an SDN-based centralized control approach can be applied for managing the computation and storage resources of individual EC platforms. Multiple operator-based communication technologies can be adapted for enabling the communication between EC nodes of the same and distinct layers.

The main problems to be addressed are listed here.

#### *8.5.3.1 Multi-layer architectures for Edge Computing service deployment*

In this scenario, we consider the Network Function Placement (NFP) problem applied to multiple slices with heterogeneous requirements over the multi-EC enabled NTN. Based upon each slice

requirement, VNF placement is performed over different EC platforms by considering their limited resources.

Resource Allocation problem: Users often request multiple slices with stringent requirements. Each network slice can be embedded as multiple VNFs embedded together. Each VNF of a network slice can have specific computation demands. Allocating enough resources for each slice, considering their requirements is a challenging problem over Muli-EC-enabled NTN and needs to be addressed carefully for reducing service failures.

VNF Placement Problem: Each service request can have specific demands in terms of latency, data rate, etc. To satisfy such requests performing a proper VNF placement over multi-EC enabled NTN is beneficial. For example, to fulfill the stringent latency requirements of Ultra-Reliable Low Latency Communications (URLLC) based slices, placing several VNFs into proximity of an end-user (i.e., over RSU-EC or LAP-EC) can be beneficial with reduced transmission delays. However, in the case of enhanced Mobile Broadband (eMBB) slice, requiring high data rates, it is needed to explore higher layers of EC platforms, (i.e., HAP-EC, LEO-EC, etc.) for having sufficient resources.

#### *8.5.3.2 In-Network Computing/Edge-to-Cloud Continuum approaches for multi-layer satellite Networks*

The presence of multiple layers between the users/data generators and the cloud has introduced in the past the possibility of extending the concept of computation offloading towards a more flexible and fluid approach where all the layers are active part of the computing task tailored to specific applications. This approach has been often referred in the literature as in-network computing or Edge-to-Cloud continuum, where the first terms is more related to the possibility that the networking aspects are merged with processing aspect to create a system where different tasks are processed within the network. The second term is instead more related to the concept where the traditional Cloud XaaS models are extended to wards the edge to create a continuum where each layer can also work with a specific service model depending of the users QoE/QoS requirements. Despite this approach has been already proposed in the literature to a generic network architecture, the introduction of a spatial multilayer architecture like that considered in a NTN can introduce several new challenges and opportunity. The integration of a physically spatial multilayer architecture where several layers from LEO to GEO with a logical multilayer infrastructure from edge to cloud could bring several possibilities. Among others we can point out several options for distributing the processing effort as well mapping different AI schemes at different layers.

#### *8.5.3.3 Protocol Architecture implications for edge computing in 5G-enabled satellite system*

Application of edge computing in NTN-based systems introduces important communication challenges from a service identification viewpoint and then for the actual implementation of edge computing onboard satellites. On the one hand, the mobility of satellite systems with respect to users on ground makes the overall service itself happening in a mobile manner, which is quite uncommon in comparison to usual edge computing services that are deployed in datacenters or in any case fixed locations. This introduces in particular some challenges from a service discovery and migration viewpoint in that the mobility of a given edge computing service producer implies that service migration to a neighbor instance of the service has to be carried out and that in any case the service consumer (i.e., a user device) has to discover the new service point of attachment and accordingly to register to the new entity in a way that the service continuity is preserved or at least

that any possible short interruptions do not affect the users' quality of experience. On the one hand, the MEC protocol architecture specified within the ETSI MEG ISG offer a solid framework to handle interaction between user and MEC agents. On the other hand, such a protocol framework was not designed for mobile systems and in particularly can suffer from the satellite mobility pattern especially if NGSO satellite systems are envisioned. Even more critical in this respect is the case of interconnected multi-orbit systems, so that in principle edge computing capabilities could be implemented across different space assets, hence meaning inter-edge communication and therefore raising the need for more advanced service discovery and migration. These networking aspects are obviously tightly coupled with the corresponding routing issues, which have been already addressed in the previous section.

Another key aspect of edge computing implementation onboard satellite is the protocol processing than can afforded from a satellite payload from a computation standpoint. On the one hand, the implementation of a fully compliant MEC protocol framework implies a full protocol stack implemented onboard, which certainly introduces important power consumption and possible additional latencies due to the protocol layer processing prior to the actual MEC operations. Further to this, the integration of 5G and NTN in a 3GPP-oriented context introduces the necessity of having UPF onboard satellite whereby a full gNB implementation is expected to run in space, with all due implications from power consumption perspective especially because of the resource-hangry tasks performed by the control plane functions. These considerations are particularly relevant in the case of 5G-direct access across NTN elements. In the case of satellite components not 3GPP aligned, gNB implementation can remain on ground but the challenge mentioned above to implement a full protocol stack still plays an important role, whereby proper adaptation of protocol functions, smarter implementation, and more effective satellite payload design from a power-consumption standpoint are amongst the key tasks to be carried out to enable effective edge computing capabilities from space.

#### *8.5.3.4 ICN/NFN networking models for enabling edge computing in space*

Evolution of communication networks towards semantic-oriented networking models has resulted in the conception of information-centric networking architectures, wherein networking operations act at content level rather than on "blind" packets by means of content names. Following extension has been to apply this concept to the so-called named-function networking so that calling of specific functions is happening by specifically invoking computing services by means of their names which are directly mapping to service resources and locations. This obviously bring the important advantage of decoupling edge computing and, overall, in-network computation from the host-centric model reinforced by IP-networking, hence making the deployment of a content-based service-oriented architecture more straightforward. Though still belonging to the research and experimentation domains, these concepts are certainly attractive to achieve a more scalable and dynamic deployment of edge computing services also in scenarios like those satellite-based, where network changes and mobility of nodes make the application of classical edge computing implementation complicated. This is particularly the case for distributed computing scenarios so that data and related computation tasks can be available across the entire network and not simply centralized in given locations. More interestingly, the case of orbital edge computing opens the door to allocating tasks and storing results of these tasks onboard satellites and eventually making those available to users upon request. In such a case, a named-function networking approach can simplify

the service and data discovery operations also in the case of complex multi-tier non-terrestrial networks, where data are distributed across multiple space assets and or could be migrated from one NTN node to a neighbor one in order to boost the data access and eventually a task computation. Moreover, exploitation of semantic routing principles (as outlined in the previous section) would allow for a more agile and dynamic distribution of data and related access independently of the specific underlying network topology that can be highly variable in 3D-NTN systems. Though certainly appealing, the power of information centric networks supporting edge computing may come at non-zero cost because of the additional protocol overhead and the change of networking paradigm for which current space systems may not be fully ready for. Deploying a content/function-oriented networking paradigm would still require coexisting with traditional IP-based networking models and therefore call for adapted protocol interfaces or mapping between service locations and IP addresses. All these operations necessitate then an adequate processing power and capability onboard satellite, which must be traded off against the traditional satellite power/mass dimensioning targeting traditional data communication mechanisms.

### 8.5.4 Management

*8.5.4.1 Orchestration of Edge Computing resource and allocation (AIaaS/SaaS/DaaS, etc.)*

In 6G networks Edge and Ubiquitous Computing will play a fundamental role due to the transition of the telecom architecture towards distributed (micro)service-based platforms that will have to provide computing resources at “zero delay”. This will impose local or proximity processing that cannot be guaranteed without fostering on NTN solutions, as well as service composition models like SaaS (Software as a Service), FaaS (Function as a Service), AIaaS, etc., and novel computing technologies of virtualization like containers, Unikernels, and relative supervisors. Such a novel environment calls for peculiar technologies for service allocation and resource management. In fact, In the NTN edge computing scenario, satellite resources allocation and tasks schedule are fundamental research topics. There are specific aspects to consider when bringing the edge paradigm to the constellation of satellites, such as satellite dynamics: conversely to typical infrastructural edge architectures, satellites are not located in a fixed position; this leads to frequent service handovers. However, it is a dynamic behavior that differs from the one typically affecting highly dynamic distributed architectures (e.g., unpredictable disconnections due to mobility or faults) that, in turn, calls for the conception and development of ad-hoc mechanisms. Another key aspect relates to the possible interplay of low orbit constellations with medium or geostationary constellations. As a matter of fact, due to the low or no shift of positions that characterize these constellations, with respect to low-orbit ones, it could be considered the possibility of reflecting the hierarchical structure of edge computing environments (cloud, near and far edge) on satellites located on very different orbits. This would lead to the opportunity to assess the validity of those techniques and optimizations that have been adopted in hierarchical edge environments, on orbital edge computing ones.

*8.5.4.2 Zero-touch NTNs management*

Zero-touch networks (ZTNs) is a term used to indicate those networks that can heal and tune themselves, based on the signals in the data they collect and analyze across all network activity. Zero-touch networks are so highly automated that monitoring networks and services are able to act on faults with minimal or no human intervention, including in the early detection of emerging problems, autonomous learning, autonomous remediation, decision making, and support of various

optimization objectives. This technology is able to support service, telecommunications, and infrastructure providers that are now looking into more flexible ways of having fully automated and E2E service management spanning under the umbrella of distinct domains, both administrative and technological. To address such a problem, the Zero Touch Network and Service Management (ZSM) specification group from ETSI was formed to discuss relevant use cases, and requirements, and specify an end-to-end management reference architecture that allows such E2E service deployments. Bringing ZSM concepts and approaches to NTN could complement the work conducted by ETSI under this perspective, also opening to novel approaches and a twofold advantage: on the one hand to challenge the ZSM specification in an attempt to improve its flexibility and adaptability with respect to new architectures. On the other hand, it could be a chance for NTN to take advantage of technologies allowing a modular, flexible, scalable, extensible and service-based architecture for edge computing.

#### *8.5.4.3 Context-aware NTN overlays for data sharing*

A key advantage of low-orbit NTN is the limited amount of energy requested to ground transmitters and receivers. Such a feature opens to interesting scenarios allowing to provide connectivity even to small devices in those areas otherwise not possible to cover using ground infrastructures. Typically, once packets reach the constellation, satellites are committed to send the data to the ground, by implementing the necessary routing algorithms. However, as the ambition of novel constellations is to borrow concepts from edge computing environments, bringing computing and storage capabilities onboard to satellites, it does make sense to provide the methodological and algorithmic instruments to properly leverage the features characterizing these computational environments, that proved to be able to reduce the usage of network bandwidth, overall. In fact, this goes beyond the mere virtualization capabilities but also encompasses the approaches enabling distributed data management and sharing as well as the collaboration that such a facility would enable and require. Specifically, it could be envisioned the need and opportunity for building aggregation of data spanning across different satellites, depending on peculiar features of the satellites, the users of such data, or the data itself. A constellation could provide specific mechanisms and solutions to NTN application providers to automatically and dynamically manage data sharing across satellites in a context-aware (pertaining to data itself or also metadata) fashion, e.g., ensuring the sharing of a given piece of data only among the satellites associated to a geographic area, or those satellites that received data whose contents are characterized by a given distribution, etc.

### 8.5.5 Application

In multi-service environments, several users randomly distributed over large coverage areas generates different kinds of services requests. With limited onboard resources, users alone cannot handle such requests and require computation offloading services from different EC platforms over joint T-NTN, allowing them to offload a complete/certain portion of their workload towards EC nodes. This allows users to share their workload with EC nodes for halving the overall task processing costs (e.g., task processing latency energy). However, in the case of a multi-EC multi-service environment, for handling many service requests of different kinds, the selection of a proper EC platform and its individual EC node/nodes for offloading can be a challenging problem to be solved. This is due mainly due to the different natures of services, different coverage areas, and available resources of EC platform layers. Moreover, selecting the wrong EC platform can add larger latency

and energy overheads or might even create service failures. During computation offloading, VUs might need to transfer sensitive data towards EC-node. Therefore, a proper assessment of EC-nodes security is required to be done in advance. Furthermore, individual nodes of EC platforms can only have a limited number of services available with them mainly because of resource limits. Therefore, selecting a proper EC node of an EC platform is required to avoid offloading service failures. Finding a proper EC platform and selecting a proper EC-node along with the amount to be offloaded based upon VUs service request constraints and mobility limitation is an important problem to be solved.

#### *8.5.5.1 Distributed Intelligence and task offloading in hierarchical/multi-layer satellite networks (Orbital Edge Computing - OEC)*

In the distributed learning process, even if an increasing number of clients can generate a more accurate shared global model, the amount of data generated to update the global model can lead to overflows and instability of the data queues at the edge of the network infrastructure. This comes out more critical in network topologies like the Mega LEO ones, where the “snapshot” in which a set of satellites and the relative computing resources are visible and available to a multitude of ground-federated clients is very limited in time. One way to prevent this overflow and keep the data queues stable is to select just a subset of clients to send their models to the Edge according to a selection algorithm.

Ground users are equipped with multiple sensor devices that can generate tons of data in real-time. These data, through proper analysis, can be utilized for providing a better QoS to the end-users. Several ML tools are available for analyzing data and inferring important information from them. Furthermore, there are multiple ways for performing the ML model training over dynamic wireless environments. A centralized ML model training approach is not much suitable technique to perform, mainly because of high communication and computation overheads. If a centralized training approach is adapted, users need to transmit their complete data towards the centralized server which performs training operations. Such an approach adds significant cost in terms of data transmission and training latency at the server. Thus, other distributed and collaborative training approaches can be a better approach for ML model training over user data.

#### *8.5.5.2 Air ground energy-aware computation offloading strategies for on-board and on-ground processing*

The size and energy restrictions often limit the capability of the EC facilities on the different NTN layers. Due to these limitations, each layer can provide only a limited set of services that can be exploited by users. Therefore, selecting a particular EC node for offloading VUs service data is an important problem to be solved. In addition, due to the limited computing capabilities of each layer, offloading a proper amount towards the selected Edge Node (EN) can improve the performance and service quality. Since the processing operations involve various communication and computation steps latency and energy should be properly considered, aiming at minimizing the overall latency and the energy costs by selecting the proper ENs and the offloading amounts in a multiple EC scenario.

### 8.5.6 Expected Impact

Integrating Edge Computing in future NTN is a requirement for enabling innovative low-latency and heavy processing services. The integration of edge computing in NTN enables also the extension toward different services that as of today cannot be fully enabled through actual technologies. With

this in mind, several challenges could be however considered when EC is extended to 3D NTN. As a first set of challenges from the system point of view, it should be considered how in a 3D spatially dispersed network different functions should be placed when fulfilling different service requirements. To this aim both network and application-level functions should be considered in for service deployment. This leads to the integration of a distributed processing architecture where spatially and logically separated entities are considered to create a in-network processing continuum for data processing. The requirement of specific protocol is also envisaged through the extension of terrestrial protocols already considered as well by integrating new paradigms, such as the ICN approach. Such a complex scenario should be effectively managed. To this aim proper orchestration mechanisms should be considered for properly manage the different resources composing the system. Despite traditional orchestration mechanisms are still challenging when extended to NTN, new paradigm has emerged and should be considered. Two of them are Zero-touch management and context-aware overlay definition, allowing to create a content-driven automatic solution for resource management. As a third set of challenges, form the application point of view, such a complex, physically and logically distributed network, proper computational sharing mechanisms should be considered. To this aim proper energy aware solutions needs to be included for avoiding service disruption due to energy failures as well for reducing the carbon net footprint.

KPI:

- Latency reduction: In order to support new services latency requirements should be considered when multi-layer NTN are deployed. Even if NTN have some inherit limitations due to the spatial distributions of nodes, the possibility of integrating multiple layers at different altitude could reduce the latency
- Autonomous service deployment: Services should be automatically (proactively/reactively) deployed upon user demand
- Global Coverage: Global coverage with Edge Computing solutions
- Distributed Processing: an edge computing infrastructure paves the way toward an integrated Terrestrial/Non-Terrestrail distributed processing infrastructure
- Space In-network computing: Space and NTN nodes are used for in-network computing tasks
- Distributed Learning: Edge computing facilities can be used for implementing distributed learning algorithms

## 8.6 Security on the 6G 3D Networks

### 8.6.1 Motivation

Recently two strategies have emerged for securing communication networks: *Quantum*-based technology and *Blockchain* techniques. The first one takes advantage of the non-cloning theorem from *quantum mechanics* to protect optical point-to-point links. Moreover, quantum computers have already been able to break powerful cryptographic algorithms such as RSA in a reasonable amount of time [C8-5], [C8-6]. This is critical in NT networks like the satellite one, in which large areas of coverage are served wireless which make them inherently vulnerable to cyber and physical attacks. Then, quantum and post-quantum research may help to improve the security of cryptographic-based networks against conventional and quantum computer attacks. This is also critical not only in peace moments but in the war ones. In this latter case, NT networks are key to provide communications since the terrestrial ones are damaged. For all these facts it is expected

that quantum networks will be key for the next generation of communication networks. At network level, technologies for developing quantum repeaters will be required. Thus, cloud QKD networks will be possible.

The second emerging area for securing communication networks is blockchain [C8-9]. Their main motivation consists of providing security to a large network of devices without resorting to centralized network architectures. Thus, it is avoided the paralyzation of the network if the centralized server fails by a Denial of Service (DOS) attack [C8-10]. By resorting to blockchain it is not necessary to trust in a third party which may risk the stored data to deletion or tampering attacks [C8-11]. So decentralized architectures such as the blockchain one may help to increase the privacy of the data and the security at network level. In this latter case federated learning technologies may increase the learning rate of the miners and reduce the latency of the network. However, blockchain is based on cryptographic algorithms and so the combination of quantum technology and blockchain solutions will be expected in the future to increase the security of blockchain networks to quantum computers attacks.

In order to further increase the security of the quantum links and blockchain networks, physical layer security and machine learning strategies will help to detect the eavesdroppers, determine if a data can be offloaded or not according to its secrecy rate, and detect anomalies in the training data of the machine learning algorithms. In addition, the security in all types of networks have to be evaluated from an end-to-end perspective and considering physical and cyber security aspects to encompass a larger diversity of attacks. Toward this regard, the softwarization of the satellite payloads, interdomain orchestration and assistance of AI will permit the development of security by design from end-to-end T/NT networks.

Finally, a section on the expected impacts of these technologies in the security of NT 6G networks in terms of KPI and KVI is provided.

### 8.6.2 QKD on Free-Space NT Networks

*8.6.2.1 QKD in Feeder and Access links.*

Currently initial research has already been conducted to develop secure communications in the feeder of satellite and airbones [C8-12], [C8-13] to name a few of them. These cases consider the communications from satellite and airplane to a ground station respectively. In these scenarios robust QKD to the channel impairments as well as the need of defining attack models have to be developed and tested e.g., position of the potential eavesdroppers in the NT network (i.e., space, ground, sea) and evaluate the information that they are able to capture. Measures for developing channel impairments such as the turbulence, beam wandering, scattering, turbidity (sea use case), and polarization losses have to be investigated. Specially, the loss of the polarization in terms of the position of the network node (e.g., satellite, airplane, maritime node). Thus, it will be possible to increase the secrecy rate against the channel impairments and the presence of eavesdroppers, and attacks from third parties. Extension to the access links will be also of interest. So, a fully secure optical NTN network (sea, ground and space) will be possible. Toward this regard QKD relaying will be of interest.

#### *8.6.2.2 QKD Relaying*

While there have been several efforts being made to investigate Quantum Key Distribution (QKD) links, several research challenges remain to overcome the limitations of the length of such links. To tackle this problem there is the need to develop technology to create and manage a QKD network, allowing two distant QKD nodes to establish secure communication while no direct point-to-point quantum link exists between them. In this scenario there are several research challenges, related to secure key relaying, guaranteed key rate services end-to-end and resource management.

In what concerns key relaying, after a path is established, the nodes should use a Key Repeater strategy to establish key material. This means that a source node may need to send reservation requests to each node in the path and to the destination node. If reservations are successful, the source requests a key from the destination and all the intermediate nodes. However, by defining an end-to-end key relaying architecture, there is the need to define strategies for a quality assurance service in what concerns the end-to-end management of the secret key material. In this aspect, a basic Best Effort service type may be provided, defining only an average key rate and traffic burst, while a Guaranteed Key Rate service type is needed for improved versions of QKD networks. For that we need a key management layer that should be in charge of managing the key storage resources, routing protocols, quality of service (QoS), and so on. Because of scarce resources (generation key rate), communication in a network may be reduced to a minimum, since each additional packet means spending an additional amount of previously established key material. Since communication is usually performed on a hop-by-hop basis that requires the trustworthiness of all nodes in the path, selecting the shortest routing path may be necessary to minimize the number of nodes that can potentially be abducted or attacked by an eavesdropper. Finding alternative routes that can provide a service with a higher degree of quality and quick rerouting in the case of interrupt detection or using multipath communications are key features of the overlay network approach. The use of multipath connections is an often-suggested solution for improving network workloads through protecting against network failures, network load balancing, large bandwidth implementation, low-delay time selection, and more.

#### *8.6.2.3 QKD in intersatellite links*

Inter-satellite links (ISLs) are a promising tool for the exchange of control information as well as recently proposed edge computing and offloading services. In addition, ISLs can be used to further increase the communication range by forwarding the data from one coverage area to another. In a more generalized vision, these links are essential for establishing the so-called Internet of Space Things (IoST), which has been proposed recently [C8-7], [C8-8] and may include various interconnected space objects, such as small satellites (especially cubesats), LEO/MEO/GEO satellites and even extra-terrestrial stations. While the ultimate objective of the IoST is to provide ubiquitous connectivity and services on a global scale including deep space missions, some of the IoST nodes may pursue other goals. In fact, malicious satellites can eavesdrop the information on the optical ISLs or even jam and manipulate it. Accordingly, the use of QKD in future space systems is of special interest. In this context, the impact of

i. pointing errors, especially due to motion of the satellites following either same or different orbits,
ii. resource management for the coexistence of quantum keys with classical data transmissions,
iii. multi-satellite ISLs,

iv. inter-satellite routing of the key flow, etc.

on the network security and resilience need to be investigated. In order to further expand the IoST concept to aerial vehicles, the links between satellites and airplanes, helicopters, high altitude platform systems (HAPSs) and unmanned aerial vehicles (UAVs) may need to be secured via QKD, too. However, these links are subject to the atmospheric impairments, e.g., turbulence, absorption, scattering, etc. Hence, the impact of these impairments on the fidelity of the quantum channels, security statistics of these links and even the choice of the links deemed worth securing need to be determined.

#### *8.6.2.4 Physical Layer Security (PLS) in QKD systems*

Physical Layer Security (PLS) bases its security on sending information from a legitimate transmitter (Alice), to an authorized receiver (Bob) at a rate from the outage capacity of the eavesdropper (Eve) [C8-14]. Thus, Eve cannot decode the message that Alice sends to Bob. This is the so-called secrecy rate. This strategy achieves information-theoretic security, does not require secret-key management and can be applied either to RF and optical wireless communications. However, the optical wireless links are much narrower than the RF ones which provide a higher security level. In practical terms it means that Eve will have more difficulties in intercepting the link between Alice and Bob. As a result, Eve will have a worse channel than Bob which will permit to attain larger secrecy-rates and secret-key-rates (SKRs). However, currently QKD achieves low SKR rates, does not have any study on the use case of eavesdropping in NT networks, and the effect of channel variations in the SKR rates. These facts are critical in long range transmissions such as satellite communications. There the channel impairments due to turbulence, beam wandering and depolarization effects may degrade the quality of the quantum links. Nevertheless, it is possible to take advantage of the channel variations, the particularities of the satellite networks, and the reduced beam size of the optical beams to increase the security of the NTN communications. Moreover, given the information-theoretic nature of PLS it can help to reduce the positions in which the potential/s eavesdroppers may be and to increase the complexity of the satellite-eavesdropper. This is very important in satellite networks since the satellites cannot be in whatever orbit and have strict energy constraints, especially in Low Earth Orbits (LEO).

#### *8.6.2.5 Machine Learning (ML) in QKD systems*

To attain larger data rates and security levels, the knowledge, modeling and characterization of the quantum channel is a must. Toward this regard, artificial intelligence and in particular machine learning strategies may help to obtain accurate predictions and modeling of the free-space quantum channel [C8-15]. However, machine learning can also be useful for QKD to optimize the parameters of the QKD transmissions (e.g., the intensity of the optical transmitters or optical/electrical gains of the photodetectors and transimpedance amplifiers). In long range transmissions, such as satellite networks, it would be necessary to increase the fidelity of the entangled quantum states and the quantum memories [C8-16]. Given the no-cloning theorem of quantum mechanics it is not possible to perfectly copy the quantum states. Nonetheless, it is possible to reproduce them with a large fidelity. In this area, machine learning may help to remove the residual noise of the protocols/procedures used for increasing the fidelity of the replicated entangled quantum states. Another area of interest of ML in QKD is in the management of its keys. That is the key formatting, key storage management and the suspicious behavior detection in the key management layer [C8-

17]. In the key formatting, service information is used to train the machine learning system to predict service characteristics and so select the most appropriate key of the stored ones. Regarding key storage management, machine learning helps to update the stored keys by removing the keys with a long key storage time or excessive jitter. In the suspicious behavior detection machine learning can be used to remove collected training data with unexpected noise and redundancy. By doing so, it is expected that the secret key rate of QKD increases due to a better knowledge of the quantum channel, replication of the quantum states, better parameter control of the quantum transmitter and receiver, and enhanced management of the quantum keys and training datasets. Finally, the machine learning system can also be used to detect the presence of eavesdroppers in case that there were errors in the quantum key exchange and/or message transmission that were not produced by the variations of the quantum channel.

### 8.6.3 Federated Network of Blockchain

#### *8.6.3.1 Blockchain over NT networks*

Given the decentralized nature of blockchain for providing security fits very well with NTN architectures such as the satellite one. Despite the academic research on this field is scarce, there are many applications of Blockchain into the IoT segment [C8-19]. At the same time, there are also many works studying the integration of satellite and IoT [C8-20], especially in LEO constellations [C8-21]. In order to communicate the elements of the blockchain network multicast communications are used. Currently, multicast communications are achieved by resorting to L3. However, the use of L1 strategies such as precoding, beamforming, etc. would be helpful to reduce the computational complexity of the network. For instance, a precoder dedicated to miners would reduce the burden of the clients, as they would not receive the messages intended to miners and reciprocally. Intersatellite links need to be used to enable fully decentralized networks. Besides they can cover the entire planet, they also optimize the global performance, since the latency is reduced. In this approach, no gateways with data planes are needed, as no internet connection is necessary. A set of Ground Network Controls can be used to operate the satellite without deploying a user data plane connectivity. However, the following challenges, from the communications point of view, are identified when blockchain is implemented in a satellite network:

- The satellite must be able to interconnect all clients between them.
- Users can be grouped by miners and clients or both. Multicast can be used instead of broadcast to increase the network efficiency.
- In order to preserve the decentralized nature of the satellite network it is desirable to use intersatellite links to cover all Earth faces without going to a central gateway.
- The latency is a crucial parameter, as it may affect the speed of adding new blocks to the blockchain and the appearance of forks. Despite there being mechanisms to mitigate forks, high latencies may affect them.

#### *8.6.3.2 Federated Learning*

Federated Learning (FL) is an approach to train machine learning (ML) algorithms in a way that the data remains private. Specifically, FL techniques aim to train machine learning algorithms on multiple distributed servers or devices, each with its own local and private data. Some simple and intuitive FL methods have proven to be surprisingly efficient solutions. For instance, averaging at regular intervals the weights of the Neural Networks (NN) trained by different FL participants, called workers, over their local data subsets to update a global model. In turn, the local neural networks

are updated with this new global model for further training. The lessons learned from each local data set are progressively shared among all workers as the global model is updated. Thus, the attackers have to corrupt multiple nodes instead of a single one since the decision of others compensates for the corrupted node. However, this approach presents some challenges:

- *Noise and unbalancing*: clients may return noisy results and a priori there is no knowledge on which clients may have more deviated results (unbalancing). For instance, depending on the data, some clients may have unexpected results that may deviate the optimal average. This issue is difficult to detect by the server since there is no exchange on the client's data.
- *Heterogeneity*: one of the key aspects of AI is the heterogeneity of the data used by the models. If the data is highly uniform, the results will not fit in for other cases and the information cannot be inferred.
- *Bottlenecks*: part of the FL relies on the computational power of clients. If a client has an obsolete device or most of the resources are not dedicated to run the AI model, it may slow the overall performance of the system.
- *Poisoning and corruption*: both are related to the security of the network in front of malicious attacks. One of the important aspects is that it is difficult to detect which data is manipulated and which not. According to this emerge two types of poisoning: Data poisoning and model poisoning. Both may drive the blockchain network to a failure.

#### *8.6.3.3 Physical Layer Security (PLS) for federated Blockchain Networks*

The decentralized nature of blockchain permits it to provide security by replicating information in multiple nodes of the network. However, in order to have such information, multicast information across the blockchain network. As a result, it makes this network sensitive to potential eavesdroppers that work coordinately. In order to overcome this issue, PLS security techniques can be used to avoid transmitting information in links with low secrecy rates. Thus, it is more difficult for the eavesdroppers to capture the messages and/or poison the data in which the nodes of the blockchain work. So, use PLS as a secure technique for offloading data in blockchain networks. Initial studies for vehicular networks and consensus blockchain networks have been conducted [C8-18]. However, they can be extended to NTN such as the satellite one, taking into account the particularities of the satellite network such as the large number of nodes that form it to attain global coverage, latencies, and channel impairments.

### 8.6.4 End-to-end security for integrated TN/NTN networks

Extending 5G+ features, such as multi-access edge computing, holistic service-based architectures and RAN softwarisation, to the NTN domain is an essential aspect for achieving a truly integrated TN/NTN network. Such a radical evolution of satellite networks towards 6G convergence comes with unprecedented benefits, laid out in the previous sections, yet also brings several security challenges. Some of these challenges, along with potential responses, are identified below.

- *Softwarisation of satellite payloads:* supporting aspects as orbital edge computing and software-defined routing in space, requires payloads with a significant degree of programmability and reconfigurability. This introduces novel vulnerabilities, compared to traditional "hard-wired" payloads. Integrity check mechanisms need to be applied onboard, allowing remote attestation of the satellite OS and the whole software stack, preventing any malicious interventions from the ground. A "security-by-design" approach for payloads needs to be followed, to make sure that no unwanted modification in the reconfigurable part of the satellite stack can cause

irreversible damage on the platform itself. A key enabler in this context are Trusted Execution Environments (TEEs) to secure data-sensitive computations. The use of TEEs on board provides (1) protected memory for storing sensitive data and (2) isolated code execution. Data and code are protected with respect to confidentiality and integrity. In addition, exhaustive security auditing processes must be defined for every software module which is to be dynamically deployed onboard.

- *Inter-domain management and orchestration:* in an integrated TN/NTN configuration, it is quite probable that the TN and NTN segments are managed by different operators. Establishing and maintaining a truly end-to-end network service mandates that the management and orchestration (MANO) platform extends its control across both TN and NTN segments, resulting in a satellite network which is no longer "closed" to a single administrative entity. In this context, distributed trust-enabling mechanisms need to be established to prevent attacks to the control plane propagating from one segment to another, which could result in unwanted modification of the network service and deployment of malicious network services on the NTN domain. Strict access control mechanisms need to be engaged, to avoid privilege escalation attacks. Proof-of-transit mechanisms are also relevant, making sure that the intended network path is not violated and no unwanted traffic diversion has taken place. Finally, it is very important that, in multi-domain operations, similar security policies are applied to both the TN and NTN realms.
- *AI-based TN/NTN network management:* The use of AI in network management is very promising, yet comes with threats related to AI-specific vulnerabilities. Attacks on both privacy and security are feasible for AI systems in terms of inference attacks leaking private information and poisoning attacks causing misclassification. Especially For collaborative AI, an adversary may control one or more nodes in the network and may thus additionally poison the model directly during training through sharing the (poisoned) gradient. For this purpose, the development of any AI-based mechanism for NTN network management needs to be accompanied with the appropriate robustness evaluation and application of relevant anti-adversarial countermeasures.

### 8.6.5 Expected Impact

After developing the security section of the NTN chapter, the following KPI and KVI arise.

Key Value Indicators:

- Supporting critical data transmissions from IoT or not: 3D-NTN networks play a key role for providing: i) global coverage, ii) service when terrestrial infrastructure has been damaged or does not exist, iii) emergency services. In this regard, new verticals such as the ones from critical sectors (e.g., infrastructures, governmental security, defense, civil service, etc.) will benefit from the 6G networks. Thus, society will be more protected against possible malfunctions (intentional or not) of these services
- Traffic load in secure networks: In the next year security issues will be introduced in the 6G 3D-NTN networks to include new verticals such as the ones related with critical services. Moreover, the mega-satellite constellations will have a larger number of satellites deployed. In practical terms it means that the traffic load of the 6G 3D-NTN networks will rise.
- Revenue of companies that provide secure TN/NT networks: In parallel to the two previous KVI, if new verticals are included in 6G, the satellite operators have more operative satellites, multi-orbit satellite architectures start to be operative, and new security paradigms are developed, the revenue of the companies that provide secure communications will increase.

Especially the ones that manage critical information and NTN since they must be more resilient to attacks.

Key Performance Indicators:

- Secrecy rate of the NTN communications: In 3D-NTN architectures, the eavesdroppers can be positioned in multiple locations since all satellites relay information. As a result, it is necessary to increase the secrecy of the communications to limit the geographical area from which the eavesdroppers may recollect the information of legitimate transmissions. A metric to measure that is the secrecy rate of the communications.
- Latency of NT network after introducing security constraints: It is well-known that quantum computing may reduce the robustness of current cryptographic schemes based on asymmetric encoding. As a result, new paradigms must be developed. The so-called quantum and post-quantum encoding. However, these encodings cannot introduce large delays. In addition, given the multi-orbit nature of the 3D-NTN networks the number of handovers among the multi-orbit nodes is constrained to the latency that demands 6G. So, a metric to measure this effect is the end-end latency.
- Recovery time after suffering an attack in an NT network: Similarly, to terrestrial networks, it is expected that NTN ones will also suffer attacks. In case that they suffer attacks, a metric that may measure the robustness of the communication networks to them is their recovery time. That means, the time that a NT requires to recover its full operativity. The lower the recovery time, the larger the resilience that 6G networks provide.
- False alarm/Detection probability of being attacked: In order to avoid attacks, 6G networks must identify when they are being attacked. Thus, they can take the corresponding preventive actions to avoid and/or limit the effect of the attack. Toward this regard, a metric that may evaluate this effect is the probability of false alarm/detection of being attacked.

.

# 9. Opportunities for Devices and Components

*Editor: André Bourdoux*

Progresses in all aspects of the wireless and wireline network are highly dependent on electronic technologies, components and devices that are used for implementation. This chapter gives an overview of the expectations towards the component and device researchers, designers and manufacturers to support the requirements of wireless/wireline networks up to the end of this decade. This includes the whole range of components such as processors, memories, analog, RF, converters, antennas, packaging and optical components.

## 9.1 Vision and Requirements

Wired and wireless networks are in constant evolution with the goal of addressing all relevant societal challenges, support the digitization of the industry, improve communications devices, support new applications (see Chapter 1), support the "more AI" trend, .... To reach these goals, we expect future networks to support very low to very high throughputs, increase area coverage, reduce latencies, improve reliabilities, integrate artificial intelligence, support an ever larger number of verticals.

The requirements on the components (e.g. chips, antennas, ...) and devices (in this chapter, "device" is used with the meaning of "user device" in the broad sense but not "transistor device") are very broad: they cover all aspects of infrastructure and human and non-human user device hardware and software. The eleven sub-sections of this chapter provide each their set of requirements for different parts of the hardware including radio transceiver hardware, computing and storage, optical hardware, security and IoT.

## 9.2 Sub-10GHz RF

The market of sub-10GHz has been dominated by cellular networks based on 3GPP standardized radio access and Wi-Fi local area access of 802.11 family. Mobile connectivity has utilized even larger portions of the spectrum at frequencies up to 6GHz. LTE (4G) and advanced Wi-Fi features will be complemented by ISM band applications from Bluetooth to home automation, NFC and IoT with narrowed spectrum allocations as well as satellite-based positioning systems.

The trend has been and will be for a more efficient use of spectrum at the range that is the most suitable for compact and highly integrated electronics, i.e. RFICs with efficient DSP in terms of form factor, cost and power consumption in battery operated devices. Although technologies for RFICs and other components may sound mature and speed of any transistor is not a bottleneck, complexity of the designs has become enormous and on the other hand new data intensive applications require enhanced broadband operation. In addition, limited and scattered spectrum availability will lead to increasing parallelism of signal paths from antennas through RF to DSP. Both carrier aggregation to enhance data rate over several bands and massive MIMO to use spectrum more efficiently at those bands multiply the number of parallel active RF signal paths.

The challenges are obvious but hard to tackle including increased power consumption and interference between simultaneously operating radio paths (co-existence). For those, the solutions cannot be overcome solely by bulky filter banks at the front-ends but require increasing signal purity

at transmitters and improving linearity and internal filtering capability in the receivers. This cannot be simply solved by increasing digital content due to bottlenecks in digital conversion. Even if higher resolution ADCs and DACs can shift part of the processing to digital that is in many ways highly beneficial, the new requirements mandate similar or even faster development of RF circuitry including antennas, external filters and switches as well as RFICs to achieve the goals.

Densifying networks is a must also at lower frequencies and not only starting from mmW region due to better range and frequency utilization including coming cell-free MIMO approaches. Co-optimization from antennas to digital over different technologies and techniques is a core competence in this field in addition to squeeze out the best possible performance from the technology. As RF integration is always balancing between speed of the transistor for digital and optimal performance of RF for power amplifiers, highly linear receivers and the best possible RF filtering contradictory requirements determine case-by-case the system partitioning of the functionality to components and IC's. Large SoC's and multi-chip modules and modem combos have their specific purposes also in the future.

To name the key opportunities in components, programmability and flexibility even beyond well-established topologies is a must. That is not anymore only cleverly placing digital switches inside RFICs, but also techniques that can better jointly optimize antenna-filter-RFIC combination in terms of performance and flexible spectrum use. Holistic view on the system performance gives still many opportunities to boost system performance and minimize cost. Digital content approaches are also needed in classical RF signal processing blocks including digital PLL's, transmitters and, to a certain extent, also receivers with minimal RF content keeping in mind that dynamic range is the key to solve any near-far problem especially in cellular transceivers and co-existence scenarios between systems. In addition, solutions for simultaneous transmission and reception in the same channel i.e. in-band full duplex are still far from maturity even for a single band. Multi-band and MIMO pose huge challenges to in-band full duplex. Similar challenges apply both for RF transceivers in mobile solutions as well as in infrastructure with different trade-off in required performance vs battery lifetime.

Finally, opportunities below 10GHz are not only limited to more efficient use of spectrum but serving different kind of applications from narrow-band IoT to radar. These two are among examples that set very specific requirements for the circuits. In some cases, they can be seen as individual problems for specific devices like temperature sensors or hart beat monitoring of elderly people. However, the opportunity to utilize the same, extremely programmable circuitry to achieve multiple goals could enable a new set of new devices. The search for optimal combinations or to design more optimal circuits to serve different combinations is an emerging challenge and opportunity for various wireless systems effectively utilizing spectrum and hardware at frequencies below 10GHz.

### 9.2.1 Research Challenges

The research challenges from the previous subsection are summarized below:

| Research Theme | Sub-10GHz RF | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Better and more efficient radio hardware | Long-term (finished in 7y+) | Mix of solutions: architectures, circuits, digital RF, silicon technology node | Cost effective, performant, green radio |
| Faster and higher resolution converters (DACs and ADCs) at low power | Long-term (finished in 7y+) | Mix of solutions: architectures, circuits, silicon technology node | Useful for all radios |

### 9.2.2 Recommendations for Actions

| Research Theme | Sub-10GHz RF | |
|---|---|---|
| Action | Better and more efficient radio hardware | Faster and higher resolution converters (DACs and ADCs) at low power |
| International Calls | To leverage industry and academic efforts vertically for critical mass | To federate industry and academic efforts across Europe for design and manufacturing of challenging components and systems |
| International Research | To leverage industry and academic efforts vertically for critical mass | To federate industry and academic efforts across Europe for design and manufacturing of challenging components and systems |

## 9.3 Millimeter-wave and TeraHertz

### 9.3.1 THz Communications:

With an massive amount of unused spectrum, the sub-TeraHertz (sub-THz) frequencies between 90 and 300 GHz are candidates to achieve high-data rate wireless and wireline communications and hence to fulfil the requirements of the next-generation of wireless networks and wireline (e.g. data center) networks [C9-1], [C9-2]. THz signals may also be carried over low-cost waveguide structures such as polymer microwave fibre (PMF) as an alternative to high performance copper and optical interconnect use cases over distances from a few cm's up to a few m's.

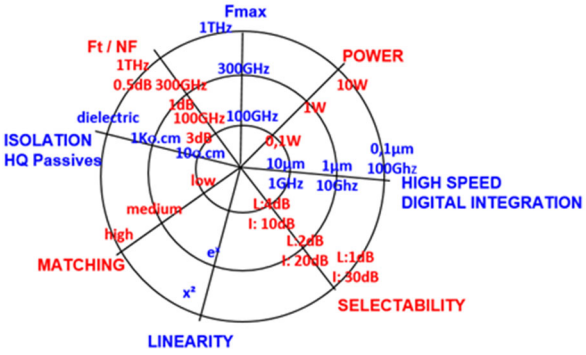
However, before making the deployment of system in the sub-THz band, many challenges need to be addressed. First, the free-space propagation losses increase with the square of the frequency. These losses must be compensated by using high-gain antennas, which entails severe constraints on antenna directivity and alignment. Sub-THz has, in similarity with mm-wave frequencies, a vulnerability to precipitation that limits the practical reach of beyond 100GHz links to less than 1-2 km with telecom grade availability (>99.99%). However, combining lower and higher frequency carriers makes it possible to have the lower frequencies backing up the performance during, rare, high precipitation events while still taking advantage of very high peak rates over long hops (>2-5 km). The use of high frequencies also makes it possible to build small and compact radios beyond 100GHz that will work well in urban environments where there is a need for non-intrusive installations. The ability to create high power, wide bandwidth amplifiers on these bands will make it possible to both increase reach and avoid the need for larger antennas. Larger antennas and simplified deployments have driven a need to investigate electronically controlled steerable beams

that today remains today an open issue. In addition, sub-THz systems could suffer from strong phase impairments due to the poor performance of high-frequency oscillators [C9-3]. Therefore, the study of new digital transmission schemes optimized to mitigate the impact of RF impairments such as phase noise (PN), or strong group delay distortion and polarization rotation in PMF, are essential to guarantee good performance [C9-4].

### 9.3.2 Solid-state technologies for THz applications:

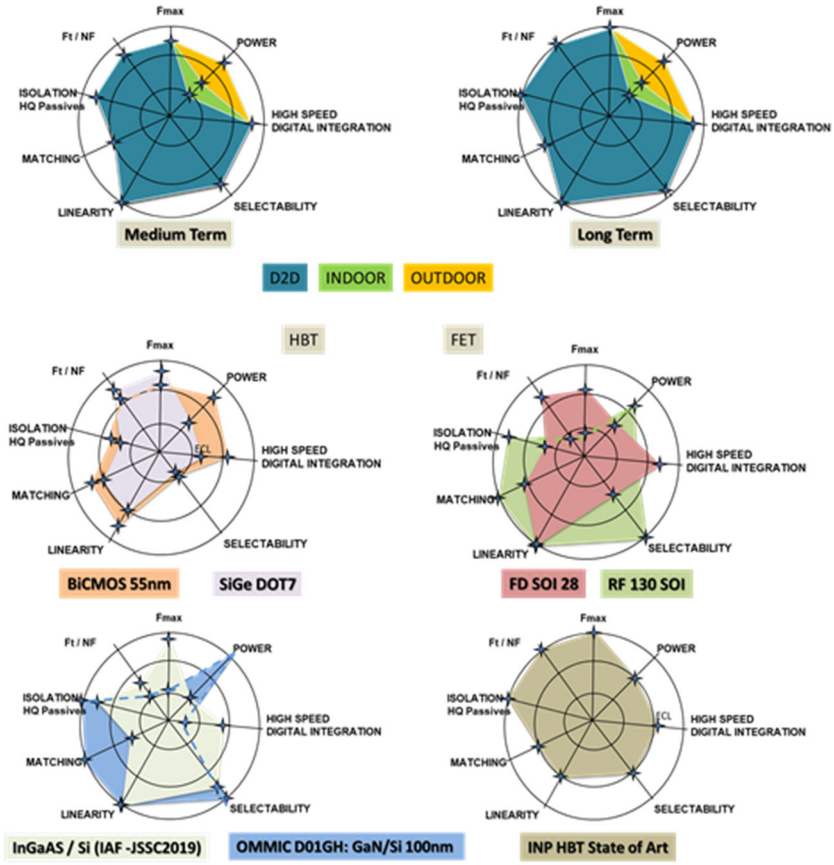
Nowadays, silicon-based technologies offer low-cost solutions for RF and millimetre-wave applications combined with a high complexity in the digital domain. CMOS, however, has its limits in speed and power generation, which become apparent at operation above 100 GHz. This is evidenced in the on-line survey of power amplifiers, maintained by the Georgia Tech University [C9-5]. Hence, the very high-speed part of a THz communication transceiver will need a different technology. Application of this "non-CMOS technology" will be limited to the transceiver functions that cannot be implemented in CMOS: transceiver architectures will be designed such that most functionality can be done in CMOS, keeping the non-CMOS part as limited as possible.

In the H2020 NEREID CSA project [C6-6], IC technologies are compared based on wireless transceivers and their most critical/representative RF building blocks: a power amplifier (PA), a low noise amplifier (LNA), and a voltage-controlled oscillator (VCO). Based on the performance of these circuits a spider diagram can be constructed (see figure below), which uses eight criteria:




The **Power** capability depends on the maximum bias voltages that can be applied over a transistor and the maximum current capability. The **High-Speed Digital Integration** capability is a function of inverter size and efficiency (transit time / current). Here, again, a disclaimer needs to be made that digital functionality will be adapted to enable CMOS integration, even if other technologies could offer better digital circuit performance. Next, the **selectability** is the ability to switch RF signals with high isolation (low loss, high isolation). **The linearity** of transistors is given by their input-output relationship (e.g. $x^2$ for a quadratic device or $e^x$ for a bipolar transistor). Further, **matching** between 2 minimum-size transistors is an important criterion in differential circuits which are widely used, from low frequencies up to mm-wave frequencies.

**Isolation and HQ Passives** depend on the substrate resistivity and on the availability of a thick metal. This gives the ability to limit the effect of pulling of a VCO by a PA, and to have high-Q passives in dielectric conditions. The couple **$f_T$/NF** is the property given by the cutoff frequency $f_T$ and the minimum noise figure $NF_{min}$ of the technology. The cutoff frequency indicates the potential for high-speed digital applications and $NF_{min}$ for the lowest achievable receiver noise. Further, **$f_{max}$** is the frequency where power gain has dropped to 0 dB. Typically, the maximum frequency of operation for RF amplifiers (which need to able to make a minimum amount of power gain) is limited to about $f_{max}/3$.




A connectivity roadmap has been defined in [C9-6] for three main application domains, namely D2D (device to device communication), indoor and outdoor applications. An outlook is given for the medium term (5 years outlook) and for the long term (10 years). Based on this analysis, the connectivity requirements for each process technology can be extracted. This is illustrated in the next figure. The main difference between the technologies is the output power. At low mm-wave frequencies, GaN is the champion here. However, in outdoor applications and also in most indoor ones, beamforming will be used. Then the required transmit power per PA is drastically lowered compared to one single antenna path, making many technologies suitable for the entire front-end including the transmit part. Further, for some D2D applications the linearity constraint can be relaxed. Whatever the application, increasing the operating frequency will impose strong specifications on $f_{max}$ and $NF_{min}$, while increasing the data rate will require a higher $f_T$. NEREID's forecast for these 3 device parameters is 1THz for both $f_{max}$ and $f_T$ and 0.5dB for $NF_{min}$.

In the spider diagrams we consider different (non-CMOS) technologies that are available at this time of writing: silicon-based ones (RF-SOI, FD-SOI, SiGe BiCMOS,,) III-V on silicon substrates (GaAs/Si, GaN/Si) and III-V on native substrates (InP). The benefit of BiCMOS and FD-SOI is the ability to combine mm-wave circuits with complex digital circuits, although to a lesser extent than ultra-downscaled CMOS beyond the 10nm generation. RF-SOI and new GaN/Si bring RF power and selectability. For generation of power in the D-band and in higher frequency bands, the survey of [C9-07] indicates that the best performance is obtained nowadays with InP. A deployment of InP on a very large scale is hindered today by the small number of metal levels that are typically available in commercial foundries and by the small wafer sizes. Integration of high-mobility III-V devices on 300mm silicon wafers as in [C9-08] and going further to co-integration with CMOS is currently investigated. Another route is a further evolution of silicon bipolar transistors, for which cutoff frequencies above 1 THz are predicted [C9-09]. BiCMOS has the advantage of being a silicon technology, with a larger ecosystem than e.g. InP but still, the product of mobility and breakdown voltage of III-V devices is higher. Finally, GaN HEMT devices are also subject to downscaling and might become a strong candidate for D-band operation when gate lengths well below 100 nm can be realized. Operation at 100 GHz with 100nm GaN on SiC devices with gold metallization has already been demonstrated but still with moderate efficiencies for a PA [C9-10]. Also here, a wide

deployment of GaN devices might benefit from integration on 300 mm wafers and with many Cu metal levels as in [C9-10].

Finally, THz communication will use very wide bandwidths to accommodate high data rates. As a result, the bandwidth that needs to be handled by baseband (both analog and digital) circuitry is huge, compared to the early days of wireless communication. This is a challenge for the design of analog-to-digital converters. There will be a need for ADCs with clock rates beyond 10 GHz. The ADC typically resides on the same chip as the digital functionality of a transceiver, which, for mass-market applications, is expected to further follow the CMOS downscaling trend. It still remains an open question how the performance of extremely high-speed ADCs will evolve when logic devices will transition from a finFET device architecture to a gate-all-around structure or even forksheet devices [C9-11].

### 9.3.3 Passive THz Imaging:

THZ Imaging state of the art (SoA) reports two main competing categories of 2D-array image sensors:

1. The above-IC bolometer-based THz image sensors based on a classical infrared (IR) sensor that offer a high sensitivity and currently a good maturity [C9-12]. but, using two different circuits, it is an expensive solution.
2. The monolithic CMOS-based THz imagers have recently emerged as low-cost competitors [C9-5, C9-7]. Even with their current poor sensitivity (1000 times less than bolometer-based sensor), these CMOS-based THz image sensors have proven to be a viable cost-effective alternative to bolometer-based imagers.

Passive THz Imaging has applications in digital health technologies, passive, continuous, home-based monitoring of biochemical markers in biofluids, such as sweat, tears, saliva, peripheral blood and interstitial fluid.

### 9.3.4 Active mm-wave and THz radar imaging:

Active radar imaging makes it possible to add the range and even Doppler dimensions to the image (3D or 4D imaging). On the lower edge of the spectrum, in the mm-wave and low THz bands, radar imaging is evolving fast to satisfy the requirements of ADAS and autonomous driving. The trend there is to resort to MIMO techniques whereby a virtual antenna array is created with a size equal to the product of the number of TX and RX antennas. 79GHz radar imaging with wide field-of-view, resolutions of 1 degree by 1 degree and cm-scale range resolution is experimentally feasible today and radars with wide field-of-view and LiDAR-like resolutions is an active field of applied research. Using higher carrier frequencies such as 140 or 300 GHz is a longer-term trend, resulting in smaller form factor or better angular resolution as well as better range resolution, thanks to the wider bandwidths. Some experimental radar chips show already the potential of CMOS in the low-THz regime (140 GHz) [C9-5]. These highly miniaturized radars will enable new applications, such as intruder detection, gesture and activity recognition, and heart rate and respiration rate monitoring, among many others.

### 9.3.5 Research Challenges

The research challenges from the previous subsection are summarized below:

| Research Theme | Millimeter-wave and TeraHertz | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Better and more efficient radio hardware at mm-wave, sub-THz and THz | Long-term (finished in 7y+) | Mix of solutions: architectures, circuits, digital RF, silicon technology node | Cost effective, performant, green radio |
| Faster (>>1Gsps) and higher resolution converters (DACs and ADCs) at low power | Long-term (finished in 7y+) | Mix of solutions: architectures, circuits, silicon technology node | Useful for all radios with extreme bandwidths |
| Semiconductor technologies CMOS, BiCMOS, III-V | Long-term (finished in 7y+) | Affordable semiconductor technologies for different market volumes | Enable mm-wave to THz radios |

### 9.3.6 Recommendations for Actions

| Research Theme | Millimeter-wave and TeraHertz | | |
|---|---|---|---|
| Action | Better and more efficient radio hardware at mm-wave, sub-THz and THz | Faster and higher resolution converters (DACs and ADCs) at low power | Semiconductor technologies CMOS, BiCMOS, III-V |
| *International Calls* | To leverage industry and academic efforts vertically for critical mass | To federate industry and academic efforts across Europe for design and manufacturing of challenging components and systems | To federate industry and academic efforts across Europe for design and manufacturing of challenging components and systems |
| *International Research* | To leverage industry and academic efforts vertically for critical mass | To federate industry and academic efforts across Europe for design and manufacturing of challenging components and systems | To federate industry and academic efforts across Europe for design and manufacturing of challenging components and systems |

## 9.4 Ultra-low Power Wireless

It is expected the number of IoT nodes will continue to grow to 100 billion by 2030, and ultra-low power wireless connectivity will be the key enabler. However, the existing wireless connectivity has limitation to support such large number of nodes. For example, battery replacement of billions of sensor nodes will not be feasible. Additionally, the trend is to spatial awareness to the IoT end-nodes using the front-end already used for radio communication. Using the sensing capabilities, channel state information can be collected. This allows for new types of applications like presence detection or localization. These functions will pose additional requirements on the radio front-end and thus the design choices. To further scale up the number of IoT nodes, several important challenges need to be overcome.

### 9.4.1 Battery-free operation

Batteries will be the primary limitation of IoT nodes. Manually battery replacement of 100's of billions IoT devices will be too expensive, and the disposed batteries will be a serious environmental

issue. To support a sustainable growth of the IoT devices, battery-free operation will be a key solution.

Most of the existing battery-free wireless communications adopt simple modulations (e.g., backscattering) and protocols. However, they will not be able to scale up to network with large number of nodes. One potential solution as demonstrated in [C9-13] is to adopt a "back-channel compatible" wake-up receiver which monitors the energy profile of the signals sending from the central hub. This wake-up receiver consumes very low power consumption, so it is compatible with battery-free operation. It only activates the main transceiver for sending sensor data only if certain energy profile is detected.

Energy harvesting is another interesting approach for devices requiring extremely low energy. Candidate energy harvesting technologies include thermoelectric, photovoltaic, piezoelectric, RF or wireless, wind and vibration energy harvesting.

### 9.4.2 Spatial Awareness

For spatial radios, we can differentiate between active and device-free localization. In active localization, two (or more) IoT nodes measure the distance between them, using channel state information. For device free localization, time variation of the channel state information is used to detect changes in the propagation environment, e.g. due to human movement [C9-14].

Currently, channel state information is often based on received signal strength (RSS), mostly because it is easy to implement. However, for multipath fading environments and increasing distance, the accuracy is rather poor. To improve robustness against multipath fading, it is well-known that a large radio signal bandwidth is required. This will increase spatial resolution, beneficial to both active and device-free localization.

Using the concept of phase-based ranging [C9-15], a wideband view on the radio channel can be obtained. By sounding the radio channel in a sequential manner over individual narrowband channels using half-duplex bi-direction signals, a wideband measurement of the radio channel is realized. For each individual measurement, only narrowband signals are used, making it suitable for radio front-end used for e.g. Zigbee of Bluetooth [C9-16], but also Wi-Fi [C9-17].

Aside a modification of the radio protocol to incorporate such measurements, also the frond-ends will be impacted, most considerably the Local Oscillator (LO) Generation/Phase Locked Loop. For accurate distance measurements, also the phase of the radio channel should be measured. This means that the generated LO should be continuously, when switching from TX to RX and vice-versa and from channel to channel. This leads to a whole new set of requirements and challenges for PLL design [C9-18].

### 9.4.3 Degradable Devices

One alarming trend in IoT, is the increasing number of disposable devices that are designed to fail and become e-waste once they run out of battery [C9-19]. To solve this problem, we need energy autonomous devices that uses ultra-low power (ULP) radios and harvest the energy they need. However, in order to eliminate the e-waste problem, research is also needed to develop ULP radios that could be manufactured by printing using biodegradable substrates and renewable materials.

This starts to emerge in the RFID domain (e.g. [C9-20]) but would have to be adopted more widely to the Internet of Everything applications in order to avoid environmental problems.

### 9.4.4 Research Challenges

The research challenges from the previous subsection are summarized below:

| **Research Theme** | **Ultra-low Power Wireless** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Battery-free operation and disposable devices | Long-term (finished in 7y+) | Mix of solutions: system concept, protocols, architectures, circuits, energy harvesting, eco-friendly electronics | Green disposable radios |

### 9.4.5 Recommendations for Actions

| **Research Theme** | **Ultra-low Power Wireless** |
|---|---|
| Action | Battery-free operation and disposable devices |
| *International Calls* | To leverage industry and academic efforts vertically for critical mass |
| *International Research* | To leverage industry and academic efforts vertically for critical mass |

## 9.5 Antenna and Packages

### 9.5.1 On-chip antennas, lens-integrated antennas, antenna MIMO arrays

Packaging of mmWave/THz chips for low-cost consumer electronic applications requires low-cost packaging solutions. Conventional low-cost silicon packaging technologies, however, exhibit a typical 1nH/mm lead and wirebond inductances, which are prohibitively high at mmWave/THz frequencies and plastic packaging materials and encapsulants are quite lossy. In addition, even expensive high-performance coaxial cables and connectors have significant losses at mmWave/THz frequencies. As a result of this, future THz packaging technologies must avoid interconnects as much as possible and antennas need to be integrated into the chip package or even on chip. Fortunately, the radiator size scales down at higher frequency and this makes compact and integrated antenna solutions feasible. However, the free-space propagation loss at THz frequencies becomes very high (80 dB for 1 m at 240 GHz) and this loss needs to be compensated with an appropriately high directivity of the antenna system in order to provide sufficient link budget. Due to their large silicon area, however, high directivity antenna arrays are costly on chip. Future solutions, therefore, include alternative lens on-chip assemblies which exhibit a better cost performance ratio.

**On-chip antennas:** For efficient and low-cost THz signal escape from the chip level, appropriate on-chip antenna systems need to be developed. On-chip antennas embedded in the BEOL stack of a lossy silicon chip [C9-21] are very challenging because of potential multi-mode propagation issues (e.g. surface waves) within the volume of an electrically large and thick silicon die leading to very inefficient radiation with very poor control of radiation patterns. Because of very high carrier frequencies with large fractional RF bandwidth, standard design techniques relying on narrowband matching become less efficient and will result in limited circuit performance. Depending on the application, antennas should support very wide operation bandwidth with minimum group delay distortion for high-speed modulation and stable phase characteristics for precise location of the focal point position in an imaging system across the bandwidth of interest. Furthermore, for

sufficiently high frequencies, classical buffering circuits become unfeasible and true antenna-circuit co-design at multiple harmonics simultaneously is necessary for high-fidelity system operation.

**Lens-integrated antennas (chip-on-lens assembly):** Further research is required on new ultra-wideband silicon lens-coupled antenna system allowing efficient coupling of THz radiation into the intrinsic device without classical matching structures. Antenna may further provide dual-polarization functionality with two transmitter/receiver paths connected to each orthogonal polarization.

**MIMO arrays:** Highly directive terahertz antennas can minimize interference between adjacent channels, and frequencies can be reused more frequently, thereby improving spectral efficiency and signal quality. This enables higher channel isolation in a MIMO (Multiple Input Multiple Output) network topology. At THz future MIMO networks could reach data rates of up to one Tbit/s easily. Future MIMO arrays need to support not only faster links, but also real-time operation by rapid channel switching and/or beam-steering/tracking at a very low latency.

### 9.5.2 Metamaterials and metasurfaces

The development of metamaterials (MM) is another promising technology for beyond-5G networks and services scenarios: one remarkable use case, for instance, is the exploitation of smart radio environments (both indoor and outdoor) with ultra-massive MIMO and Artificial Intelligence (AI).

MM are materials which contain inclusions (e.g., metallic or dielectric of various shapes and dimensions) designed and engineered to manipulate electromagnetic (EM) waves. Examples of inclusions embedded into a host metamaterial include EM scattering element and nano-resonators. These properties, for instance, could be used for developing smart antenna and EM processing functions, including methods for AI.

Metasurfaces (MS) are 2D metamaterials (MM). By modifying the structure and spatial distribution of those sub-wavelength reconfigurable passive elements in the metasurface, the electromagnetic characteristics of the elements can be changed, and independent phase shifts are added by different MS elements to incident signals without using any power amplifier, complicated coding and RF processing. In this way, passive reflection, passive absorption, passive scattering can be realized which may even not exist in Nature (e.g., zero or negative refraction) [C9-22], allowing a wide range of EM processing functions and pushing the physical environment to change towards intelligent and interactive.

MS can be seen as arrays of nano-antennas: by shifting the resonant frequency, through the nanoantenna designs, it is possible to effectively control the amount of the phase shifted in the scattered signal. [C9-23] describes a prototype of an information metasurface controlled by a field-programmable gate array, which implements the harmonic beam steering via an optimized space-time coding sequence.

The main advantages of MS include: completely passive and low power consumption, supporting free-duplex and full-band transmission, requiring no high-cost components, being able to be deployed densely and expandable and reconstructing electromagnetic waves at any point on its continuous surface.

It is expected that MS will have several possible applications, such as:

i) radio coverage in areas not well covered by installation of base stations, and face NLoS limitation [C9-24]
ii) smart radio environments (indoor and outdoor): being combined with AI, IoT and edge computing to enhance performance in smart cities, smart homes, health monitoring, and safety inspection,
iii) to serve cell edge users, relief multi-cell co-channel interference, expand coverage, reduce electromagnetic pollution, and implement dynamic mobile user tracking,
iv) automotive applications, vehicle and air networks and high-speed railway scenarios
v) running quantum algorithms directly with EM waves or in optics (e.g., in transformation optics [C9-25], quantum radio-optics devices, ultra-fast switching devices, detecting and recognizing images, holographic applications, etc.)

Furthermore, the possibility of coating surfaces in building or kiosks with intelligent (AI-based) MS will allow creating smart radio environments capable of radio waves propagations by introducing, in a software-controlled way, localized and location-dependent gradient phase shifts onto the signals impinging upon the MS. As a brand-new material, MS can be combined with antenna technology, massive MIMO, millimeter wave, terahertz communication, D2D and other technologies to form a controllable smart radio environment.

### 9.5.3 Research Challenges

The research challenges from the previous subsection are summarized below:

| **Research Theme** | **Antenna and Packages** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| On-chip antennas, lens-integrated antennas, antenna MIMO arrays | Long-term (finished in 7y+) | Mix of solutions: packaging, interconnects, lenses, on-chip vs off-chip antennas, beamforming/MIMO trade-offs | Small form factor, high efficiency antennas for all frequency ranges |
| Metamaterials and metasurfaces | Long-term (finished in 7y+) | Breakthrough antennas/antenna arrays and reflective surfaces | Improved coverage, low losses, passive operation |

### 9.5.4 Recommendations for Actions

| **Research Theme** | **Antenna and Packages** | |
|---|---|---|
| Action | On-chip antennas, lens-integrated antennas, antenna MIMO arrays | Metamaterials and metasurfaces |
| *International Calls* | To leverage industry and academic efforts vertically for critical mass | To federate industry and academic efforts across Europe for design and manufacturing of challenging components and systems |
| *International Research* | To leverage industry and academic efforts vertically for critical mass | To federate industry and academic efforts across Europe for design and manufacturing of challenging components and systems |

## 9.6 Optical wireless convergence

### 9.6.1 Radio-over-fibre communication, sub-systems and components for B5G and 6G networks

In the near future, we can expect significant overhaul of the mobile network, targeting the use of mmWave frequency bands to deliver much higher capacities over the air. Mobile networks will use advanced radio transmission concepts such as coordinated beamforming, coordinated multipoint and massive MIMO (multiple input, multiple out) as well as pico-, femto- and even attocells. It has been long recognized that it is better to centralize (C-RAN, centralized radio access network) the digital signal processing (DSP) required for modulation and demodulation of the RF carriers. Advanced signal processing is now centralized in the baseband units. It is expected that for future mmWave networks, this fronthauling (bringing the data from the antennas to the centralized or cloudified baseband units) will be done through optical fibre given the high capacity and/or frequency of the signals that need to be transported.

While today this fronthauling is built upon standards such as CPRI (common public radio interface) or OBSAI (open base station architecture initiative) in which the digitized IQ samples themselves are transported, for future mobile networks the amount of traffic that will need to be transported will explode. For example, assuming 2GHz modulation bandwidth, 4 carriers, 3 sectors each with 32 antennas, digitization at 8bits, 8B/10B encoding and 10% overhead for control messages, then a total sustained throughput of 25Tb/s will be required in the fronthaul link. To overcome this problem alternative fronthauling techniques will be required:

- Analog radio-over-fibre, in which the RF signals are directly modulated onto optical carriers. This will require the development of highly linear optical modulators, which today form the biggest hurdle in the deployment of analog radio-over-fibre systems for mobile network applications.
- More efficient digitization of the RF signals as opposed to directly transporting IQ samples: one example is sigma-delta modulation in which the RF carrier is oversampled, and the resulting digital signal is transported over the fibre using conventional low-cost optics (as opposed to likely more expensive analog radio-over-fibre) [C9-26].
- High speed PON to facilitate fixed and wireless convergence

### 9.6.2 Optically assisted wireless subsystems

As explained before, new generations of B5G and 6G mobile wireless transmission systems will rely extensively on advanced radio transmission concepts such as beamforming (requiring true time delaying of RF signals), or operate at very high carrier frequencies (100s of GHz, which can be generated by beating lasers on photodetectors spaced apart by the required carrier frequency). Such microwave photonic techniques can play an increasingly important role at these high frequencies.

### 9.6.3 Research Challenges

The research challenges from the previous subsection are summarized below:

| Research Theme | Optical wireless convergence | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Radio-over-fibre and other advanced fronthauling technologies | Long-term (finished in 7y+) | Higher efficiency and higher throughput fronthaul | Key technology for O-RAN, distributed/cell-free MIMO |

| | | | |
|---|---|---|---|
| Optically assisted wireless subsystems | Long-term (finished in 7y+) | Better, higher efficiency circuits and building blocks for sub-THz/THz | Improved functionality, higher efficiency at Sub-THz/THz |

### 9.6.4 Recommendations for Actions

| **Research Theme** | **Optical wireless convergence** | |
|---|---|---|
| Action | Radio-over-fibre and other advanced fronthauling technologies | Optically assisted wireless subsystems |
| *International Calls* | To leverage industry and academic efforts vertically for critical mass | To federate industry and academic efforts across Europe for design and manufacturing of challenging components and systems |
| *International Research* | To leverage industry and academic efforts vertically for critical mass | To federate industry and academic efforts across Europe for design and manufacturing of challenging components and systems |

## 9.7 Baseband Modems

Figure 9-1 shows the range of the processing options that have been explored over the years for base stations baseband modems [C9-27]. Traditionally, most of the heavy lifting was carried out by various application-specific integrated circuits (ASICs) that had moderate programmability. ASICs were necessary because processor performance was limited by transistor count. More recently, flexible solutions that use a reconfigurable processing element, such as field-programmable gate arrays (FPGAs) instead of ASICs, have been studied. Unlike fixed-function ASICs, FPGAs can be reprogrammed dynamically, although the development effort is still high and significantly higher than writing new software. Therefore, there has also been significant interest in truly programmable baseband processors.

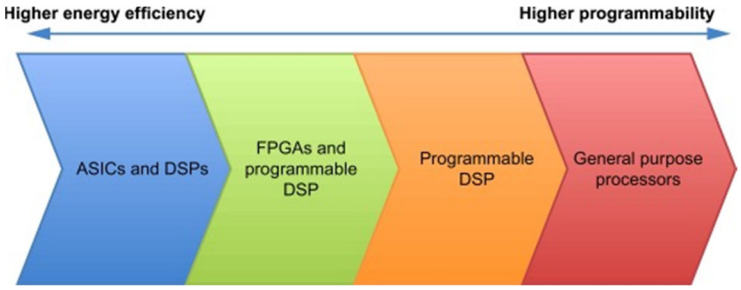



*Figure 9-1. Baseband processing options [C9-27].*

One style of programmable processors integrates the functionality of FPGAs or ASICs into enhanced digital signal processors (DSPs). These designs not only exploit the data level parallelism inherent to baseband processing workloads but also include domain-specific features that are tuned for baseband processing, such as specialized shuffle networks and arithmetic units. A more radical departure from specialized processors and the adoption of very general fully programmable hardware is an attempt to use off-the-shelf CPUs to process all the tasks of the physical layer processing. Such solutions potentially enable wireless operators to further reduce the cost to build and upgrade RAN infrastructure with commodity off-the-shelf CPUs. Current CPUs have a large, and growing, number of cores and integrate single-instruction multiple-data (SIMD) units within each

core. With this high level of parallelism, commodity CPUs can now meet the performance demands of even advanced physical-layer processing. GPUs can also be considered for highly parallelizable tasks that do not require frequent and irregular memory accesses.

While there is a large spectrum for possible processing options, the fundamental trade-offs remain the same. Programmable and reconfigurable processing elements are more flexible in that they can work with different signal frequencies, modulations, and coding schemes, and even completely different channel access methods and processing pipelines. This allows wireless operators to reuse hardware resources when migrating to new wireless technologies. Consolidating the functionality of ASICs into fewer processing elements also greatly reduces the cost of both hardware and software development. Finally, flexible equipment can enable even better resource utilization through a more sophisticated resource scheduling strategy such as dynamic resource allocation between different wireless communication technologies. However, these benefits come at the price of energy efficiency and performance because fewer opportunities for low-level specialization and hardware tuning are available with commodity parts than with specialized fixed-function accelerators. Previous work suggests that the performance and efficiency gap can be 10× to 100× between ASICs and general-purpose processors. With the expected slowdown of device scaling and the benefits it provides for performance and energy-efficiency, the trade-off between energy efficiency and programmability in baseband processing hardware is becoming more important than ever.

For UE modems (e.g. smartphones and IoT devices), ASIC implementations with embedded accelerators, DSP and CPU cores is dominating for power, cost, performance, size perspective. Recently, also dedicated machine learning acceleration is considered within the modem for certain categories of devices. Depending on the advancement regarding machine learning / AI in the PHY, such processing might be more commonplace and demand additional emphasis from an acceleration perspective.

A recent, and potentially disruptive development is the application of deep learning for the physical layer. By interpreting a communication system as an autoencoder, several groups are developing a fundamental new way to think about communication system’s baseband design as an end-to-end reconstruction task that seeks to jointly optimize transmitter and receiver components in a single process [C9-28], [C9-29]. Compared to traditional baseband architectures with a multiple-block structure, the DL based AE provides a new PHY paradigm with a pure data-driven and end-to-end learning-based solution.

Advances in CMOS scaling (following Moore’s law) is crucial to enable progresses in digital implementations. With the current technologies (FinFETs), one approaches several “walls” such as the power wall, performance wall, scaling wall and cost wall. Disruptive approaches are needed to break down these walls to enable evolution towards 1nm (10A) and beyond by the turn of the decade. Promising approaches include gate-all-around (GAA) silicon nanosheet, gate-all-around forksheet, complementary FET (CFET) and exploiting the 3$^{rd}$ dimension. It becomes also essential to have a holistic approach of system-technology co-optimization whereby the application (system), the algorithms, the architecture, the design and the fundamental technologies are considered and optimized jointly.

### 9.7.1 Research Challenges

The research challenges from the previous subsection are summarized below:

| **Research Theme** | **Baseband Modems** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Architecture and processor trade-offs (TPU, GPU, CPU, DSP, ASIC, FPGAs, ASIPs,...) | Long-term (finished in 7y+) | Powerful and Efficient DSP implementation solutions adapted to a broad range of use cases, from simple IoT device to complex base station | High efficiency and computing power |
| Semiconductor technologies (CMOS scaling towards 1nm and beyond) | Long-term (finished in 7y+) | New technologies such as GAA nanosheet, GAA forksheet and complementary FET to extend Moore's law beyond 1nm | Power, performance, scaling and cost for advanced digital implementations |

### 9.7.2 Recommendations for Actions

| **Research Theme** | **Baseband Modems** | |
|---|---|---|
| Action | Architecture and processor tarde-offs (TPU, GPU, CPU, DSP, ASIC, FPGAs, ASIPs,...) | Semiconductor technologies (CMOS scaling towards 1nm and beyond) |
| *International Calls* | To leverage industry and academic efforts vertically for critical mass | To federate industry and academic efforts across Europe for design and manufacturing of challenging components and systems |
| *International Research* | To leverage industry and academic efforts vertically for critical mass | To federate industry and academic efforts across Europe for design and manufacturing of challenging components and systems |

## 9.8 Processors for Cloud-AI, Edge-AI and on-device-AI

The requirements of AI applications are driving the development of dedicated hardware architectures at a rapid pace. CPUs and GPUs are being refined with the purpose of increasing the energy efficiency and reducing the latency. New technological solutions are being leveraged for enabling in-memory compute (i.e. using non-volatile memory technology), multiple chip integration (i.e. chiplets, interposers …), sensor integration.

The rapid pace of adoption of new technologies and ASICs opens up new application segments, since the more processing power is available, the harder the problem addressed. The application space is therefore very broad today but can be split into two main categories: applications that rely on cloud-based solutions and application that run at the edge.

Requirements for cloud-based processors are very specific. First, the cloud is still the workhorse for the learning phase, handling ever larger databases and complex learning algorithms. The compute load must be balanced over many processing units. The first challenge is thus to ensure scalability up to large scales: the associated research areas deal with the interconnect and the memory hierarchy (RDMA over Converged Ethernet being today the solution). Secondly, the cloud must ensure low latency to inference tasks, which are too computation- or memory-intensive to be

handled at the edge. The second challenge is thus to provide accelerators optimized for being efficient when handling low batch sizes (typically a size of 1): the research area is the one of data flow and systolic architectures. Finally, there is also a need for energy efficiency, since the datacentres are a large and growing contributor to greenhouse gases emissions. For that, apart from the classical Moore's law pursuit, work is for example being done on data encoding: this has led to the development of the BF16 (Brain Float 16b) representation, which helps save energy and die area compared to the FP32 representation, at almost no accuracy penalty. The research work must be pursued on dynamic encoding.

AI techniques and methods are necessary for IoT in an on-device or edge computing environment to provide advanced analytics and autonomous decision making, impose additional computation requirements on the hardware architectures.

In particular, for applications that run at the edge or on-device the first and foremost key parameters of interest are the energy dissipation and the memory footprint. Both can be addressed thanks to extreme weight quantization, down to binary synapses. This eases analogue in-memory compute, using non-volatile memory technology. The challenge, in this case, is one of learning algorithm: several tricks must be employed to keep the impact on classification accuracy low. It remains to be seen whether extreme weight quantization is the solution for future applications needs. Indeed, the trend is to have edge platforms or endpoints exhibiting unsupervised or lifelong learning abilities, for applications such as predictive maintenance or adaptation to the environment. The weights accuracy must therefore be higher for the learning algorithm to converge and the on-chip memory larger for storing all the intermediate results. The challenge is to design very dense, local, memory with a low energy access cost. Furthermore, edge or endpoints devices will require sensors for interfacing with the physical world. The difficulty will lie in sensor integration and fusion, with algorithms enabling the use of multimodality (i.e. different input types such as image, sound, vibration). Moreover, research might be needed on flexible on-device operating systems able to cope with open device management ecosystems and AI-based dedicated hardware architectures.

Spiking neural networks is a promising approach to enable bio-inspired learning with extreme efficiency. This event-driven architecture reduces drastically the amount of computing yet achieves excellent performances. Both digital and analog implementations are possible, with analog implementations being the most energy efficient.

### 9.8.1 Research Challenges

The research challenges from the previous subsection are summarized below:

| **Research Theme** | **AI processors** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Processors for Cloud-AI, Edge-AI and on-device-AI | Long-term (finished in 7y+) | Powerful and Efficient AI processors and memories fitting the energy budget at cloud, edge or device. Innovative architectures such as in-memory compute and spiking neural networks | AI enablement |

### 9.8.2 Recommendations for Actions

| Research Theme | AI processors |
|---|---|
| Action | Processors for Cloud-AI, Edge-AI and on-device-AI |
| *International Calls* | To leverage industry and academic efforts vertically for critical mass |
| *International Research* | To leverage industry and academic efforts vertically for critical mass |

## 9.9 Memories

### 9.9.1 Memory technologies towards 2030

#### *9.9.1.1 Entering the zettabyte and yottabyte eras*

The amount of data produced in the world will soon exceed 100 zettabyte, with an annual growth rate of 1.2 to 1.4x. The IP traffic is expected to be about 0.25 zettabyte in 2020 and has a similar growth rate. The yottabyte is the order of magnitude for 2030. This huge amount of data and traffic are partly generated through well-known applications such as Amazon, YouTube, Facebook or Netflix. But emerging IoT applications will make a significant contribution as well, such as autonomous cars, smart buildings, smart city, e-health, etc. Huge amounts of bandwidth are required to transport all this data – from the application to an edge node, then to a base station, and then to a data center – a challenge that will be tackled by 6G and optical networks. Throughout this data flow, stringent requirements will be imposed on memory and storage – in terms of density, bandwidth, cost and energy.

#### *9.9.1.2 Clever data mining, and reduced energy consumption*

At some point in the flow of data transport, the generated data will need to be analyzed and converted into knowledge and wisdom by means of machine learning techniques. The exact point at which this will happen, will significantly impact the requirements on memory and storage. For example, if machine learning can be applied just after data generation, it can help relax the requirements down the data flow. If, on the other hand, data is turned into wisdom later in the process, more raw data will need to be stored throughout the whole process.

The zettabyte and yottabyte eras will also challenge the power that is consumed by the growing amount of data centers, for processing, transporting and storing all the data. Without energy consumption optimization, the energy consumption for these operations, data centers worldwide may use almost 8000 terawatt-hours by 2030. (source: https://www.labs.hpe.com/next-next/energy).

#### *9.9.1.3 The slowdown of today's memory roadmap*

Let us have a closer look at the memories used in a typical laptop. Close to the central processing unit (CPU), fast, volatile embedded static random-access memories (SRAMs) are the dominant memories. Also, on-chip are the higher-level cache memories, mostly made in SRAM or embedded dynamic random-access memory (DRAM) technologies. Off-chip, further away from the CPU, mainly DRAM chips are used for the working memory and non-volatile NAND Flash memory chips for storage. In general, memories located further away from the CPU are cheaper (less expensive per byte), slower, denser and less volatile.

For half a century, Moore's Law has driven the continuous increase of memory densities, and this has translated into cost improvement of memory technologies. However, despite large improvements in memory density, only storage density (NAND Flash devices) has truly kept pace with the data growth rate. With the transition from planar NAND to 3D-NAND devices, density improvement for this storage class is however expected to slow down as well and go below the data growth rate soon.

To meet the memory requirements of the zettabyte and yottabyte eras (i.e., improved density and speed, and reduced energy consumption), multiple emerging memory options must be explored for standalone as well as for embedded applications. Options range from MRAM technologies for cache level applications, new ways for improving DRAM devices, emerging storage class memories to fill the gap between DRAM and NAND technologies, solutions for improving 3D-NAND storage devices, and a revolutionary solution for archival type of applications.

#### *9.9.1.4 MRAM technologies for embedded cache level applications*

Spin transfer torque MRAM (STT-MRAM) technology [C9-39], [C9-31] has emerged as a candidate technology for replacing L3 cache embedded SRAM memories. It offers non-volatility, high density, high speed and low switching current. The core element of an STT-MRAM device is a magnetic tunnel junction in which a thin dielectric layer is sandwiched between a magnetic fixed layer and a magnetic free layer. Writing of the memory cell is performed by switching the magnetization for the free magnetic layer, by means of a current that is injected perpendicular into the magnetic tunnel junction. Because of speed limitations, STT-MRAM are limited to L3 cache.

An MRAM variant, the spin orbit torque MRAM (SOT-MRAM) [C9-32], can potentially replace the faster L1 and L2 cache. In these devices, switching the free magnetic layer is done by injecting an in-plane current in an adjacent SOT layer, as such de-coupling the read and write path and improving the device endurance and stability.

VCMA-based (Voltage Control of Magnetic Anisotropy) MRAM [C9-33] is another interesting emerging option offering low power, high performance and high-density non-volatile memory solution.

#### *9.9.1.5 DRAM scaling*

DRAM is structurally a very simple type of memory. A DRAM memory cell consists of one transistor and one capacitor, that can be either charged or discharged. To downscale the structure, the aspect ratio of the structure must be increased. Another concept could be to place the peripherical logic directly under the array of capacitors and transistors. This logic circuitry controls how data is moved to and from the memory chip, and typically consumes considerable area. Today, the transistor of the DRAM memory cell is however built on silicon. To be able to move the peripherical logic underneath the DRAM array, we need to replace this transistor with a non-Si transistor that is back-end compatible. 3D DRAM integration is yet another improvement path.

#### *9.9.1.6 Storage class memory*

Storage class memory has been introduced to fill the gap between DRAM and NAND Flash memories in terms of latency, density, cost and performance. This new memory class should allow massive amounts of data to be accessed in a very short time. Most probably, more than one novel memory technology will be required to span the entire gap. Candidate technologies include various cross-

point-based architectures for the memory array, such as phase-change-RAM (PC-RAM), vacancy-modulated conductive oxide (VMCO), conductive bridging RAM (CB-RAM) and oxide RAM (OxRAM).

#### *9.9.1.7 3D NAND... and beyond?*

Since its introduction several years ago, 3D NAND [C9-34] has become a mainstream storage technology because of its ability to significantly increase bit density without sacrificing endurance and performance. This is enabled by transitioning from 3 bits per cell to 4 bits per cell. And, instead of traditional x-y scaling in a horizontal plane, 3D NAND scales in the z direction by stacking multiple layers of NAND gates vertically. Today, stacking over 100 layers has become possible, but the density improvement of 3D NAND is expected to slow down and will soon not be able to follow the data growth rate [C9-35].

#### *9.9.1.8 DNA storage: the holy grail of archival storage?*

DNA storage promises storage densities orders of magnitude higher than semiconductor memories. DNA can be kept stable for millions of years. DNA as a medium for storage is also extremely dense and compact. Writing can be performed by encoding binary data onto strands of DNA through the process of DNA synthesis. The DNA strand can be built up with the base pairs representing a specific letter sequence, through a series of deprotection and protection reactions. As from the read side, there is an enormous technology push to sequence DNA faster and faster and at lower cost. Progress in DNA sequencing has been amazing, even outpacing Moore's law. But researchers still have a long way to go before reasonable targets (1Gb/s) can be reached. To realize this, faster fluidics, faster chemical reactions and much higher parallelism are needed than what is possible today.

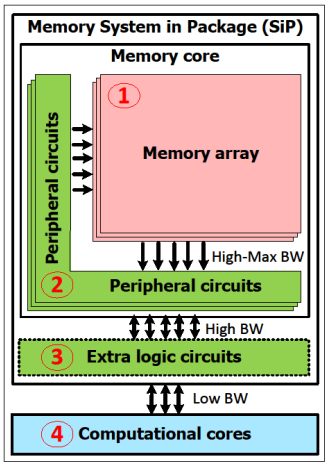
#### *9.9.1.9 Conclusion*

It is clear that the classical memory roadmap cannot handle the zettabyte and yottabyte world in terms of energy, density, speed and cost. It will be crucial to improve and develop new memory and storage technologies.

And, lastly, sustainability brings another aspect of the zettabyte and yottabyte eras forward: recycling. To be able to process and store all the data, massive amounts of devices will be produced. The advent of emerging technologies will also bring in new materials, which today are hardly recycled. The semiconductor industry should therefore also find ways to improve the recyclability of all these materials.

### 9.9.2 Compute-in-Memory

The discussion so far applied to “conventional” von Neuman architectures. This section discusses non-von Neuman approaches.



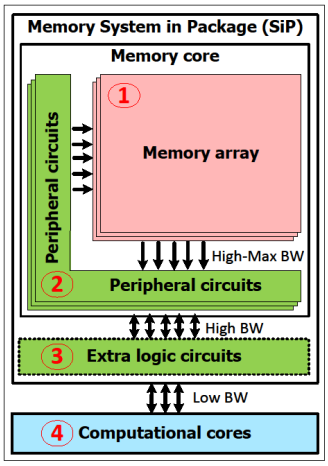




*Figure 9-2 Memories in computer architecture*

Figure 9- provides a classification of computer architectures and highlights the differences [C9-36]. Depending on where the result of the computation is produced, four possibilities can be identified; they are indicated with four circled numbers and can be grouped into two classes: Computation-outside-Memory (COM) and Computation-In-Memory (CIM). In COM the computing takes places outside the memory core, hence the need of data movement; it has two flavours. COM-Far refers to the traditional architectures such as CPU (circle 4 in Figure 9-2) and CIM-Near refers to architectures that include computation units with the memory core(s) to form an SiP such as Hybrid Memory Cubes (circle 3). In CIM (based on memristive OR devices) the computing result is produced *within* the memory core and consists also of two flavours; CIM-A in which the result is produced within the *array* such as IMPLY [C9-37] (circle 1), and CIM-P where the result is produced in the memory *peripheral* circuits such as Scouting Logic [C9-38] (circle 2). Note that CIM architectures have relatively low amount of data movement outside the memory core and may exploit the maximum bandwidth (as operations happen inside the memory array). However, CIM requires more design effort to make the computing feasible (e.g., complex read-out circuits); this may result in large complexity which could limit the scalability. Moreover, as CIM performs computations directly on the data residing inside the memory, the robustness and performance are heavily impacted by data misalignment.

If successful, CIM will be able to significantly reduce the power consumption and enable massive parallelism; hence, increase computing energy efficiency and area efficiency by orders of magnitudes. This may enable new power-constrained computing paradigms at the edge such as Neuromorphic computing, Artificial neural networks, Bio-inspired neural networks, etc. Hence, a lot of application domains can strongly benefit from this computation; examples are IoT devices, wearable devices, wireless sensors, automotive, etc. However, research on CIM (based on memsristive devices) is still in its infancy stage, and the challenges are substantial at all levels, including material/technology, circuit and architecture, and tools and compilers.

- *Materials/Technology*: there are still many open questions and aspects related to the technology which help in making memristive device-based computing a reality. Examples are device endurance, high resistance ratio between the off and on state of the devices, multi-level storage, precision of analog weight representation, resistance drift, inherent device-to-device and cycle-to-cycle variations, yield issues, exploring 3D chip integration, etc.
- *Circuit/Architecture/communications*: Analog CIM comes with new challenges to realize (ultra) low power and simple designs of the array structure, peripheral circuits and the communication infrastructure within the CIM and to the I/O interface. Examples are high precision programming of memory elements, relatively stochastic process of analog programming, complexity of signal conversion circuit (digital to analog and analog-to-digital converters), accuracy of measuring (e.g., the current as a metric of the output), scalability of the analog crossbar arrays and their impact on the accuracy of computing, the partitioning across crossbars and the corresponding intra- and inter-communication under various constrains such as latency, bandwidth and power, etc.
- *Tools/Compilers*: Profiling, simulation and design tools can help the user not only to identify the kernels that can be best accelerated on CIM and estimate the benefit, but also perform design exploration to better guide optimal designs and automatic integration techniques for CMOS and emerging memristive devices (e.g., monolithic stacking).

As of today, most of the work in the public domain is based on simulations and/or small circuit designs. It is not clear yet when the technology will be mature enough to start commercialization for the first killing applications. Nevertheless, some start-ups on memristor technologies and their applications are already emerging; examples are Crossbar, KNOWM, BioInspired, and GrAI One.

### 9.9.3 Research Challenges

The research challenges from the previous subsection are summarized below:

| **Research Theme** | **Baseband Modems** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Memory technologies towards the yottabyte area | Long-term (finished in 7y+) | Technologies with increasing densities for all levels of the memory hierarchy (registers, L1 to L4 cache, DRAM, NAND, storage and cold storage | Enablers for devices, infrastructure, cloud, ... |
| Technologies for In-memory computing | Long-term (finished in 7y+) | More efficient AI | Non von Neuman architecture to better fit AI paradigm |

### 9.9.4 8.9.4 Recommendations for Actions

| **Research Theme** | **Baseband Modems** | |
|---|---|---|
| Action | Memory technologies | Technologies for In-memory computing |
| International Calls | To leverage industry and academic efforts vertically for critical mass | To federate industry and academic efforts across Europe for design and manufacturing of challenging components and systems |
| International Research | To leverage industry and academic efforts vertically for critical mass | To federate industry and academic efforts across Europe for design and manufacturing of challenging components and systems |

## 9.10 Hardware for Security

Due to their cost efficiency and promised performance improvements, decentralized deployments are experiencing an important raise of interest in the industrial community, especially in high-risk environments like production sites. Consequently, two main features related to them, namely security and privacy, are getting more and more implemented as hardware (HW) features. In parallel, one can easily witness how all aspect of nowadays life are increasingly supported by HW with extended lifetimes, which forms the core of the so-called Internet of Everything. Today, all vendors compete towards rapid and widescale deployments of billions of devices in the most diverse fields like autonomous vehicles, smart cities, smart homes, and industrial automation [C9-39]. Unfortunately, while a long lifetime of the HW is required, today's state of the art is unfit to ensure such needed-for long-term security feature. *Sustainable security* is therefore a major concern in the industrial ecosystem, as without it billions of vulnerable but active devices will pose a substantial and increasing security risk to the broader society. Therefore, there is the need for research actions into sustainable security and privacy, which will shape trustworthy devices that can maintain well-defined guarantees (security, privacy, safety) of critical services over extended life-times (e.g. 20+ years) at affordable cost [C9-40].

Today, a constant stream of risks from many sources (SW, HW, Crypto, Infrastructure …) renders devices vulnerable and enables mass-scale attacks such as the Mirai Botnet [C9-41]. Devices can only remain secure under active and costly maintenance (vulnerability management, patching, update), which requires a dedicated development team per vendor supporting legacy devices. In practice, three approaches to long-term security are predominant – none of them satisfactory:

- *No or Time-Limited Maintenance*: The most common approach is to only provide limited-time maintenance and accept the fact that devices remain in operation while security rapidly degrades. This creates a substantial risk to society and to users of critical services.
- *Limit the Device Lifetime*: Vendors sell devices with a limited lifetime (e.g. limited by warranty). Some vendors use remote update to render devices unusable afterwards. This is not satisfying for users and not environmentally sustainable.
- *Continuous Maintenance and Service Contracts*: For some segments, vendors can offer "devices as a service", by which vendors are paid for continuous maintenance including security. While this costly approach works for some industrial settings, it will not be realistic for the majority of the existing scenarios.

To solve the problem of sustainable security and privacy, we believe that multiple research areas must be pursued in parallel to mitigate risks to long-term security:

1. *Long-term Security Maintenance*. Smart systems are increasingly deployed in systems that have a long lifetime. Examples include smart cities, smart infrastructure, industrial, and vehicles. Today, each individual system requires costly maintenance (vulnerability scanning, bug fixing, patching, …). This will create a maintenance nightmare for systems that live 20+ years. *We suggest pursuing research on how to build systems that self-maintain their security for 20+ years with minimal maintenance cost.*
2. *Fail-Security + Survivability under Major Attacks.* Even though everyone would agree that designing secure systems is an indispensable feature, in reality the currently deployed systems are far from perfect: if a system was successfully attacked, security can no longer be guaranteed at all, and systems need most of the time to be manually restored, cleaned, and patched. *We therefore suggest exploring new HW mechanisms that allow graceful*

*degradation under attacks while supporting automated recovery of security while the system maintains its critical services*.

3. HW Security Roadmap towards Post-Quantum Secure Systems: We believe that quantum computing can break today's HW implemented security mechanisms. Since there is no one-size fits all for post quantum security, it is important to analyze a wide range of usages and make appropriate recommendations how to mitigate this risk.

### 9.10.1 Research Challenges

The research challenges from the previous subsection are summarized below:

| **Research Theme** | **HW for security** | | |
|---|---|---|---|
| Research Challenges | Timeline | Key outcomes | Contributions/Value |
| Long-term security | Long-term (finished in 7y+) | Long-term Security Maintenance, Fail-Security + Survivability under Major Attacks, HW Security Roadmap towards Post-Quantum Secure Systems | sustainable security and privacy |

### 9.10.2 Recommendations for Actions

| **Research Theme** | **HW for security** |
|---|---|
| Action | Long-term security |
| International Calls | To leverage industry and academic efforts vertically for critical mass |
| International Research | To leverage industry and academic efforts vertically for critical mass |

## 9.11 Opportunities for IoT Components and Devices

Deploying and managing a large set of distributed devices with constrained capabilities is a complex task. Moreover, updating and maintaining devices deployed in the field is critical to keep the functionality and the security of the IoT systems. To achieve the full functionality expected of an IoT system, research should be done in advanced network reorganization and dynamic function reassignment. Research is needed for providing new IoT device management techniques that are adapted to the evolving distributed architectures for IoT systems based on an open device management ecosystem.

Components (micro-electronic components) and devices mainly for IoT and vertical sector applications are essential elements of future secure and trusted networks and to support the digital autonomy of Europe. With respect to the increasing demand and expectation of secure and trusted networks, especially for critical infrastructures, there should be European providers for such devices as an additional source to latest technologies to complement the European value chain and mitigate the existing gaps.

### 9.11.1 Approach for components

European semiconductor players are stronger in IoT and secured solutions, while volume- oriented market are dominated by US or Asian players. For European industry to capture new business opportunities associated with our connected world, it is crucial to support European technological leadership in connectivity supporting digitisation based on IoT and Systems of Systems technologies.

Increasingly, software applications will run as services on distributed systems of systems involving networks with a diversity of resource restrictions.

It is important to create the conditions to enable the ecosystem required to develop an innovative connectivity system leveraging both heterogeneous integration schemes (such as servers, edge device) and derivative semiconductor processes already available in Europe.

Smart services, enabled by smart devices themselves enabled by components introducing an increasing level of “smartness”, will be used in a variety of application fields, being more user-friendly, interacting with each other as well as with the outside world and being reliable, robust and secure, miniaturised, networked, predictive, able to learn and often autonomous. They will be integrated with existing equipment and infrastructure - often by retrofit.

Enabling factors will be: Interoperability with existing systems, self- and re-configurability, scalability, ease of deployment, sustainability, and reliability, ability to be customised to the application scenario.

All technology and component considerations in the previous sections of this chapter apply also to IoT components.

### 9.11.2 Approach for devices

Devices and especially end devices for IoT and vertical applications including critical infrastructures are an essential part of future networks. In addition to components, they must also fulfil a high security level. System on chip activities can be leveraged for such industrial device activities. The close cooperation between vertical sectors and the ICT industry in Europe will support the development of entire communication and networking solutions in Europe. These activities offer opportunities for start-ups to design communication modems and other components or building blocks devised for many vertical applications.

### 9.11.3 Requirements for IoT devices

Devices with IoT gateway capabilities in support of different IoT connectivity modes, both at local and public network level. In particular, for each supported vertical industrial domain and as well cross vertical industry domains:

- requirements will be derived on which software and hardware capabilities and characteristics these multi-modal IoT devices and network elements should support, when integrated and used into the 5G and beyond 5G network infrastructures. Considering that these IoT devices support e.g., wireless technologies that are non-5G and beyond 5G radio technologies, such as Bluetooth, Wi-Fi, ZigBee, LoRa, Sigfox
- integration and evaluation activities of these multi-modal IoT devices and network elements in the 5G and beyond 5G network infrastructures will be planned and executed.
- Hardware requirements for IoT Devices:
    - Requirements applied for each supported vertical industry domain and as well cross vertical industry domains when integrated and used into the 5G and beyond 5G network infrastructures.
    - At least three different frequency bands for sub-1 GHz (700 MHz), 1 - 6 GHz (3.4 - 3.8 GHz), and millimetre-wave (above 24 GHz) and integrate multiple protocols in addition to cellular ones.

- Functional and performance requirements, such as high data capacity, highest levels of reliability (connectivity), fast response times (low latency), sensing/actuating, processing and storage capabilities; low power consumption.

### 9.11.4 IoT Swarm Systems in the context of 6G:

The concept of swarm applications and/or systems has been introduced some time ago, please see e.g., [C9-42], [C9-43]. In the context of IoT, the Swarm is considered to be an approach in which independent and heterogeneous IoT devices can cooperate with each other to execute tasks synergistically, see e.g., [C9-44]. Concepts for IoT intelligence clustering can be applied as well in 6G enabled devices to promote collaboration and share of resources and functions for performing specific tasks.

Swarm systems are characterized by their intelligence clustering capabilities. The key research challenges related with the application of the swarm and IoT intelligence clustering concept in the context of 6G enabled devices are:

- to dynamically allocate resources such as sensors, communication networks, computation, and information from the edge and cloud in order to execute tasks synergistically,
- to aggregate and use that information to make or aid making decisions
- to dynamically allocate and use actuation resources, while controlling their response by policy, security, and privacy concerns

In addition, standardisation challenges are imposed in the required architecture, such as interfaces, data models and ontologies, security and privacy models.

# 10.Future Emerging Technologies

*Editor: Anastasius Gavras*

This chapter is mainly a point of aggregation of technologies that present such transformational potential that its potential is nor properly represented, or not represented at all, on the previous chapters. It is intended to collect information on transformational technologies beyond the current scope of available lists, such as the ETSI technology radar or the Gartner hype cycle, or that are currently at not receiving due attention. In this aspect, the chapter takes a very different structure of the previous chapters, taking a more speculative view.

Future Emerging Technologies (FETs) will be presented in a storytelling fashion of user scenarios, including aspects as:

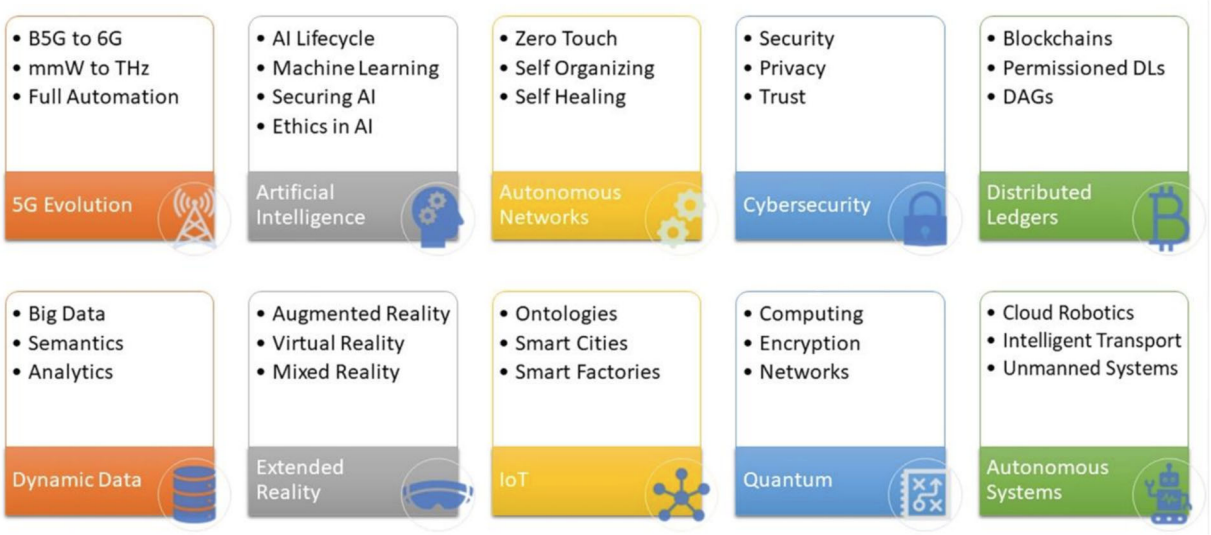
- A description of the user scenario
- A description of the technology (or technologies)
- A view on the potential impact on the UN Sustainable Development Goals (SDGs)

In addition to the storytelling fashion that should cover the above points, several identified technology (or set of technologies) will be accompanied by:

- An estimation of the TRL. We should expect that FETs are at most at TRL2 at the point of publishing this SRIA, so this point is often not needed.
- Information about the context in which the technology has been shown feasible and potentially link it to European inventors and European innovators.
- List of unresolved issues (tech and non-tech, social acceptance, change of human behaviour, ethics...)

## 10.1 ETSI technology radar

ETSI is maintaining a technology radar that aims to capture technology trends as illustrated in the next figure. The first and current edition of the ETSI technology radar was published in April 2021 and is available online[11].




## 10.2 Quantum technologies

### 10.2.1 Quantum networking

Quantum computing harnesses the collective properties of quantum states of atoms and subatomic particles, such as superposition, interference, and entanglement, to perform calculations quantum processors, which are able to perform quantum logic gates on a certain number of quantum bits (qubits). Quantum communication seeks to utilise quantum mechanics principles for transmission of information. Finally quantum networks facilitate the transmission of information in the form of qubits, between physically separated quantum processors. Quantum networks work in a similar way to classical networks, but are better at solving certain problems.

The most prominent application of quantum properties in communications is Quantum Key Distribution (QKD), which allows two communication parties to produce a shared secret key that can be used to encrypt and decrypt communication among them. More applications have been described, yet the challenge remains to construct large scale quantum networks [C10-1]. A further review of the challenges, with a focus on experimentation towards large scale networks is provided in [C10-2]].

### 10.2.2 Quantum Machine Learning

The advance of quantum computing applied to machine learning promise significant value add in various areas. In the following we use the example of Quantum Machine Learning for Remote Sensing Imagery Classification.

[11] ETSI Technology Radar, First edition - April 2021, ISBN No. 979-10-92620-39-1 https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp45_ETSI_technology_radar.pdf

Currently, different quantum algorithms that could act as building blocks of ML programs have been developed, sometimes related to hardware and software challenges that are not yet completely solved [C10-3]. Given that ML and AI can play fundamental roles in the quantum domain [C10-4], the main benefits of QML, as already summarized in [C10-5], are the following: 1) improvements in run-time, 2) learning capacity improvements, 3) learning efficiency improvements.

However, there is not a shared consensus on how and when QML can be advantageous with respect to its classical counterpart on general classes of problems. For instance, in [C10-6], it is shown how the quality and the amount of data can sensibly affect the performance of classical and QML models in such a way that the quantum advantage is not always guaranteed. With this regard, this paper adds an important element of discussion with respect to the state of the art, by demonstrating how QML could help when dealing with real remote sensing images for a classification problem where multiple classes are used.

**Quantum Machine Learning applications.** Currently, there are several general methods for implementing quantum circuits into ML models, as it can be found in the literature. For instance, in [C10-7] image classification is performed via a QML, while in [C10-8] a quantum support vector machine is used for Big Data classification. In [C10-9] quanvolutional neural networks are employed to carry out image recognition, and instead variational quantum circuits for inductive Grover oracularization are presented in [C10-10]. Lithology interpretation from well logs is discussed in [C10-11], and quantum variational autoencoder presented in [C10-12]. Quantum Neural Networks (QNNs) are often presented as hybrid algorithms that leverage quantum nodes throughout the networks [C10-13][C10-14][C10-15]. QNNs develop a network of both quantum and classical nodes with some given activation functions, convolutional connections, and weighted edges. Here, the quantum nodes can be represented by single qubits or clusters of qubits. QNNs can also present a more complexly integrated circuit with entanglement, where correlations between quantum nodes can be exploited to speed up computation.

**Quantum Machine Learning challenges.** Trying to create complex quantum networks which link together layers of quantum nodes still represents a research challenge. Despite the many possible theoretical applications of quantum computers, there is still significant progress that must be made towards more reliable computation. The QC industry currently finds itself in the Noisy Intermediate-Scale Quantum (NISQ) era, where there is a limit to the number of operations that can be performed on a quantum computer before the information stored becomes useless [C10-16]. Currently, these limitations contribute to the difficulties in scaling up quantum computers. However, all the work in progress is not useless since as soon as scaling quantum computers become viable, they will be able to represent exponentially more information than the classical ones. Fortunately, recent events show promising evidence for moving ahead and away from the NISQ era. In particular, by using QCNN models, researchers have been able to create an optimal QEC scheme for a given error mode [C10-11], and moreover, many QC companies are also projecting similar timelines for developing their architecture. Commercial companies are planning to release error corrected and fault tolerant commercial quantum computers by 2025.

## 10.3 Security

### 10.3.1 Scalable homomorphic encryption

Full homomorphic encryption is a form of encryption that allows computations to be carried out directly on ciphertext. The result of the computation is left in encrypted form which, when decrypted, results in an identical output to that produced had the operations been performed on the unencrypted data. Homomorphic encryption can be used for privacy-preserving outsourced storage and computation, and considering the GDPR rules and various very high requirements for privacy preservation has a large number of applications. This allows data to be encrypted and out-sourced to cloud environments for processing, all while encrypted. While various, usually trivial examples, have been demonstrated since decades, scalable solutions suffer from serious degradation of computing performance. Alternative paths to solve the performance degradation problem have been proposed, and which usually depend on hardware support by Trusted Execution Environments (TEEs) to assist homomorphic encryption by moving time and computationally intensive steps into secured software guard extension (SGX) enclaves . SGX is a feature available in modern Intel CPUs. However, TEE-based techniques are vulnerable to side channel attacks 297.

**Research challenges:** The research challenges pertain to identifying alternative approaches to remove the scalability barriers for full homomorphic encryption allowing its wider application in every application are high very high privacy-preserving requirements.

### 10.3.2 System inherent trustworthiness

The 6G era will introduce potentially new security technology enablers that will try to complement existing approaches to security and trust. However, the emerging challenge is the trustworthiness at the system level, and which must be assured across a very heterogeneous landscape of hybrid public and private clouds and networks, end user devices, sub-networks, IoT devices and applications. The 6G threat vector will be defined by 6G architectural disaggregation, open interfaces and an environment with multiple stakeholders, which is already the case in 5G. In this context a system inherent trustworthiness could possibly be achieved through a holistic consideration of privacy preserving technologies, hardware and cloud embedded anchors of trust, quantum-safe security, jamming protection and physical layer security as well as distributed ledger technologies as well as trusted automated software creation and automated closed-loop security operation [C10-17].

Artificial intelligence and machine learning (AI/ML) mechanisms have found their way into beyond 5G network architecture and their importance as a key technology enabler is augmented by their omnipresence in the future system. Although this pervasiveness will be central for achieving trustworthiness across the full security technology stack and architecture, the downside of AI/ML is the introduction of new, yet to be understood, threat vectors caused by wrong or incomplete models. Such models can emerge through insufficient large data sets used for training the algorithms or even through adversarial machine learning.

## 10.4 Human centric multimodal communication

**Description of user scenario:** Human centric multimodal communications and services (eXtended Reality and holographic telepresence) will become the norm for both social interaction and entertainments as well as professional applications such as tele-operation. In addition to communicating audio/visual data modes, including haptic/tactile and user’s emotion modalities in future communication services unlock some exciting use cases. Firstly, new feats like remote surgery

will become possible, and even early trial surgeries in China[12] have been successful. There are also several industrial/robotic control applications[13], interactive multiplayer gaming[14], as well as education/edutainment[15], automotive[16], athletics training[17], and more. Outside of those more vertical applications, haptics has proven to be very effective at delivering alerts, because the neurons associated with touch respond more quickly than to visual or auditory stimuli. This sort of application has a number of more horizontal uses such as in guiding/alerting visually[18]- or hearing-impaired[19] people, or in applications where sights or sounds are not desirable, such as in certain military operations[20].

**Technology requirements:** Multimodal traffic requirements particularly for teleoperation demands high-rate for XR traffic and low-latency robust link for the haptic feedback control. These place new demands on 6G system to deliver very high spectral efficiency, while ensuring reliable and timely video frame a haptic delivery. A mix of low, mid and high frequency bands from existing 5G-bands as well as the new sub-THz bands provide appropriate coverage for the use case scenario. For multi-modal traffics such as XR and interactive gaming demanding low latency, high throughput and low loss, existing adaptive rate congestion control methods can be applied to enable RAN rate recommendation.

**SDGs overview:** These future emerging technologies enhance the operation of the network infrastructure as well as the user experience. These contribute to lower $CO_2$ footprint (Planet), direct relevance to European inventors and European innovators (Prosperity) and better user experience (People).

**People:** The focal point is to support different verticals including human centric networking considering both the service requirements of local communities and our industry. Technology generation based on different user profiles (including user communication capabilities or impairments) are covered by human centric networking help addressing service inclusion and diversity.

**Planet:** Human centric solutions including terminal design and interface(s) to the network can lead on scaling the number of terminals associated with the device distribution for multimodal communication leading on terminal industry growth can have consequent impacts on environment

**Prosperity:** The human centric networking areas of research innovation (including European R&D centres) impact on 3GPP systems, supporting core service provider business model for technologies impacting sustainability.

---

[12] https://www.pcmag.com/news/china-performs-first-5g-remote-surgery
[13] https://www.landmobile.co.uk/news/glasgow-university-demonstrates-5g-robotic-arm-teleoperation/
[14] https://www.nme.com/news/gaming-news/call-of-duty-black-ops-cold-war-lets-pc-players-enable-trigger-haptics-2995925
[15] https://hal.inria.fr/hal-02479022/file/Howard-PUMAH-demo.mp4
[16] https://www.electronicdesign.com/markets/automotive/article/21145025/surface-haptics-a-safer-way-for-drivers-to-operate-smoothsurface-controls
[17] https://www.abc.net.au/news/science/2021-04-01/vr-teslasuit-simulates-virtual-reality-touch-haptic-feedback/100030320
[18] https://www.freethink.com/series/ramen-profitable/gps-for-the-blind
[19] https://www.auganix.org/this-haptic-suit-lets-you-hear-music-through-your-skin/
[20] https://www.army.mil/article/125405/researchers_develop_hands_free_eyes_free_navigation_for_soldiers

### 10.4.1 Holographic sense

**Use case:** Touch and smell at distance

Future scenarios will seamlessly blend virtual and real environments and holographic objects will become core digital actors. Current AR/VR approaches are limited to audio-visual experience and not suitable for touching, interacting and manipulating environments. Holographic datasets can comprise very high amounts of uncompressed data and computation time for codes can be restrictively high [C10-18]. Coding/decoding latency must be added to the transmission latency of a network.

Similar to AR/VR approaches, current holographic work is largely restricted to audio-visual information. By adding multi-sensory information in the exchanged information [C10-19], one can speak about teleportation. Touch information could require the transmission of about 1 Gbps for an average hand size. The taste and smell are inter-related and could be transposed into chemical reactions. Even though no estimates exist yet, and assuming bit rate and latency would not to be a problem, the replication of chemical reactions at the peer ends of communication depends on the presence of chemical elements involved in the chemical reaction.

### 10.4.2 Augmented cognition through implants or non-invasive

**Use case:** Brain Computer Interface: with the more immediate example of an artificial retina that can directly inject electromagnetic signals to the nervous system creating visual impressions. Effectively this could be extrapolated as far as eliminating the need for displays that use optical modality to transmit information to the brain.

**Description:** Advances in all areas of medicine have turned cyborgs from fiction to reality. Generally the term cyborg refers to humans with bionic, or robotic implants. Here we focus on eye retina implants as a form of cyborgization in medicine. An implant that electrically stimulates the retina by exciting nerve endings can transmit images directly to the optical centre of the brain, bypassing the optical path of the transmission from a display to the eye and to the brain respectively, making the need for displays obsolete.

**What is changing?** Advances in medicine have provided humans with many restorative technologies that restore lost functions, organs and limbs. The key aspect is restoration or repair, with no enhancement of the original capabilities in mind. However, there is only a small step to engaging in activities, which enhance capabilities, e.g. optimising or maximizing performance. Evidence of performance optimisation can be found in the Paralympics, where sprinters with artificial legs are as fast as normal high-performance athletes.

Retina implants[21] are researched for the purpose of restoring useful vision to people suffering vision loss due to degenerative eye conditions or even people that are blind since birth. A retinal implant is a biomedical implant technology currently being developed by a number of private companies and research institutes worldwide. People that lost their vision have learned to interpret the signals of the retina and build images in their brains, however blind since birth people have no concept of

[21] https://en.wikipedia.org/wiki/Retinal_implant

an image like non blinds since birth have. However experiments have shown that even blind since birth people can ”perceive patterns” when the respective nerve endings are electrically stimulated.

The core technology consists of an array of electrodes implanted on the back of the retina and a transmitter that beams electrical signals that correspond to images to the electrode array in the eye. Today the technology, while still rudimentary, allows the user to see a scoreboard type image made up of bright points of light viewed from about arm’s length.

Medicine will progress the ambition of vision restoration in the next years, possibly reaching a similar level of perfection as artificial limbs have reached. This means that blind or almost blind people will be able to restore their vision and get a similarly perfect image transmission of their outer world as before their impairment.

**What is the vision?** Use of the retina implants technology to overlay and transmit images into the brain, bypassing the optical transmission path. Ultimately this means the disappearance of the displays in all forms that we know today. It is a form of augmented reality without head-up displays or eyewear mounted modules.

**What are the challenges, the gaps?** The challenges are located mainly in the medical area, e.g. the safety of surgery and operation of the retina implant. Further challenges are the precise structure of the stimuli signals that should be transmitted to the nerve endings, so that the brain can translate them into images.

Perhaps the brain can learn to interpret any type of signals as long as they are somehow consistently structured and coded. Context switching will need increased attention, since the brain has to communicate somehow to the retina implant and to the transmitting engine that other information is needed and should overlay the vision. Solutions emerge for this purpose as well, such as the Brain Computing Interface (BCI).

Humans would possibly perceive overlay images as augmented reality or synthetic vision, however studies on US air fighter pilots using head-up displays have shown that these are not without side effects. The literature documents that head-up displays can contribute to loss of attention or cognitive capture.

The service and user interface design principles that today focus on device displays have to be redesigned. The safety and security of operation of the devices have to reach degrees that are not available today. For example, how to assure protection against attacks from malware? How do assure only legitimate information is transmitted. The notion of spam may need to be adapted.

The technological challenges are marginal in front of the ethical and societal challenges. What is ethically justifiable out of what is technologically possible? Should a person with a healthy vision undergo a possibly painful surgery and possibly long period of training to learn perceiving overlay transmitted images?

**What are the potential issues?**

- Do we want a solution? – Yes, but for different reasons, mainly for medical.
- Do we want to abandon displays? – Yes, they cost us energy, they draw the batteries empty (on mobiles), they are not very flexible, and they always have too low resolution. No – I

cannot watch a football game with friends in front of a large TV screen. Or maybe I am old fashioned and we can meet in cyber-world and enjoy the game in a very different modality.
- Do we want to attach a whatever-wireless enabled retina implant into our brains? – Each of us should probably answer for him-/herself.
- Does it bring benefits? – Possibly lots of benefits. Most applications of immersive technologies would apply here as well. Perhaps another dimension of everywhere, anytime –We could watch a movie through the eyes of an actor. Can/should I switch off my overlay vision before going to bed?
- Do we need regulatory and policy frameworks that constrain cyborg technology in general?
- Can we afford to not address the technological development in the area of cyborgs in general, even if some of us cannot accept it?

### 10.4.3 Entangled personality

Facebook's metaverse[22], as well as the previously existing concept of avatars, leads to the entanglement of physical objects and humans with their virtual representation that can interact. This could be perceived as an ultimate scenario when the availability of the previously described holographic sense and augmented cognition are realised.

A dystopic interpretation of the concept was shown in the 2009 movie "Surrogates" in which an FBI agent ventures out into the real world to investigate the murder of humanoid remote-controlled robots. These surrogates allow people to interact with society and ultimately assume their life roles, enabling them to experience life in their imagination from the comfort and safety of their homes.

### 10.4.4 The disappearance of the smartphone

**Use case:** Ambient voice recognition in private and public space, or in-ear headsets. Global interconnection of all human-computer interfaces available in a space can provide an intelligent ambient, where the physical smartphone can become obsolete. The user smartphone remains as a concept inside the cloud, as the service communication point for the individual [C10-20]. But the human interface of such a virtual smartphone is build by all the multiple interfaces surrounding the user: the smart TV, the voice-operated house system, the voice controlled services (e.g. Alexa), the car infrastructure, even the devices belonging to other users.

**What are the challenges, the gaps?** The challenges are located mainly in scaling, trust, confidentiality and economical viability of this approach. Transporting the interface of the (global) computing system into "the air" (*talking aloud as "Alexa", or with specific hand gestures*) is possible, but relying on such availability of interfaces everywhere, for everyone, and trusting that this system retains the same security levels of your personal device is a step too large for the current ecosystem, and hardly realizable without profound regulatory and infrastructure changes, which will only be realistic with a complete trusted software redesign for such a large system.

## 10.5 Digital Twinning

### 10.5.1 Digital Twin applied in 6G

Digital Twin technology is considered to be an emerging concept and a multi-disciplinary integration technology, which has already become the centre of attention for industry and as well academia. Digital Twin has several definitions, see for example [C10-17]], and can be seen as the real time

[22] The term originates in the 1992 science fiction novel "Snow Crash" by Neal Stephenson

integration and communication of data between a physical and virtual machine in either direction. In particular, a Digital Twin environment allows for rapid analysis and real-time decisions made through accurate analytics. Several enabling technologies are used to support this enabling environment, such as IoT, AI and underlying infrastructures, such as 5G and 6G; In order to support the rapid analysis and real-time decisions, several requirements are imposed on these enabling technologies, including the underlying infrastructure, such as 6G and as well on the accuracy of the virtual model (representation of the physical model).

**Research Challenges:** Research on addressing the following challenges for the situation the Digital Twin concept is applied in 6G:

- Virtual representation of loT devices mirroring the relevant dynamics, characteristics, critical components and important properties of an original physical abject throughout its life cycle. Real-time update based on reliable multi-sense wireless sensing, cyber-physical interaction and reliable wireless control over interaction points where wireless devices are embedded
- Technique for mesh and multi point over the air (OTA) updates/upgrades
- Simulation and modelling tools for large scale of real-time, robust and seamless interactions among, loT digital twins, humans, machines and environments

### 10.5.2 Accelerating 6G innovation and experimentation via Digital Twinning

A path that has been followed in the recent years in experimentation, is to accurately and reliably create virtual replicas of complex existing systems and technologies, which is reflecting the evolution of digital twins. Here, in order to emulate a large-scale network experimentation platform, a reference representation of all involved technologies in a network ecosystem, that captures their complexity, interfaces, and nature of all involved digital entities is developed. By developing a virtual copy which is fed with real life data and emulates existing technologies and subsystems, various Digital Twin-based experimentation platforms have been developed, where scaling and evolution is not an issue as this relies essentially on the extension of the replicas [C10-21] [C10-22]. Digital twinning relies on the ability to continuously monitor and deliver reliable repeatable results, and brings simplicity and cost-effectiveness, to testing. Emulators are used to test both the performance of a real network as well as those network functions and services that are too remote and complex to configure and access. Data sets stemming from this emulation process are also available for further scenario testing (e.g., reliability etc.)

**Research Challenges:** The digital twin models should provide an infrastructure emulation platform as next-generation evaluation platform, and enable development at scale. The digital twin models should be able to seamlessly interconnect with available testbeds and platforms and play a pivotal role in the creation, deployment, and evaluation of future scenarios in a much faster and agile way. This approach promises faster deployment and operation than using physical testbeds, which may suffer limitations during scale-up operations. The digital twin models can be used for continuous optimization of the physical testbeds, providing assurance in terms of deployment choices, which represents a breakthrough as compared to current testbed approaches. The ambition is to drive innovation in the area of future networks, beyond the state of the art through physical and virtual experimentation in domains such as service orchestration, private network control, organic core network architecture evolution, and others.

## 10.6 Nano, bio-/molecular technologies and communications

This addresses the problem of fow do interface to the nano, bio-/molecular world, and there are multiple use cases that can be considered, as:

**Use case 1:** abnormality detection inside blood vessels with mobile nano-machines

**Use case 2:** “swallow your surgent scenario”

Richard Feynmann in 1959 [C10-22] suggested to shrink computer devices and wires to the 10-100 atom scale structures. This would allow the construction of nanobots that are small enough to travel inside blood vessels and being control via magnetic fields.

**Description** The primary application areas of nano- and molecular-scale technologies are in the biomedical, environmental, military and other industry fields. The basic functions that these technologies are capable of performing are very simple tasks in computing, data storing, sensing and actuation. Networks at this level are relevant in terms of expanding the capabilities of single nano-machines or molecular building blocks in order for them to perform more complex tasks by allowing them to coordinate, share and fuse information. The nano and molecular interfaces and gateways need to be understood, and properly developed, potentially connecting these worlds to the network in some way.

**What is changing?** Nano-technologies emerge as a means for constructing components at the sub-microscopic scale of a nanometer and allowing the fabrication of simple devices ranging in size from 0.1–10 μm. Although largely in the research phase, practical applications have been experimentally demonstrated. Useful applications of nano-machines could be in medicine e.g. to identify and destroy cancer cells, or in the environment for detecting chemicals and their concentration.

Recent progress in nanotechnology and nano-science has facilitated the study of molecular electronics. At the experimental level the advances have facilitated the manipulation of single-molecule electronics. While these artefacts are mostly operating in the quantum realm of less than 100 nm (a scale where quantum mechanic effects become relevant) their collective behaviour can manifest in the macro scale.

**What is the vision?** Research the interfaces from the macro world to the nano and molecular world, in order to usefully interact, observe, control, organize, and exploit the behaviour of nano-machines and molecular building blocks, as well as retrieve useful information from the sub-microscopic world. The research can extend to the programmability of their properties and behaviour.

**What are the challenges, the gaps?** Generally the problem of interfacing is challenging research for the next years, since the known means of communication at this scale differs from the communication means in the macro scale. Important fields of research are securing the macro/nano interfaces in particular in applications which have a direct impact on species and the environment in general. The possible programmability of their properties is enhancing this requirement.

**What are the potential solutions?** The starting point is that it has been demonstrated that electromagnetic waves generated by electro-mechanically resonating nano-materials can be produced and processed at this scale.

## 10.7 Energy

The topic of energy is becoming increasingly important in telecommunications, and shorter term views of this section have already been discussed in Chapter 2 (and Chapter 3), with some topics on implementation also in later chapters. Nevertheless, a more global view with some deeper impacts can be mentioned here.

### 10.7.1 Energy harvesting devices

**Use case:** Energy harvesting from the environment or energy induced from outside on demand to wake up a device for a purpose/task. Applications include environmental monitoring in production environments (temperature, humidity, shock exposure, ...) or monitoring the environment, flora and fauna, pollution etc. In such cases low-cost, zero-energy devices are needed.

Devices - typically sensors - that are deployed in inaccessible places (oceans, woods, sewage) could typically be considered economically not viable to recover for replacing power supply (battery). Therefore energy harvesting from the environment is a potential solution. Energy could be harvested from vibrations, from light, from temperature gradients, or even from the radio-frequency waves themselves [C10-24]].

The power possible to harvest from miniature sources is typically very low. In case of radio-frequency (RF) energy harvesting, the harvested power is often as low as a few microwatts (µW). In comparison, the output power of the RF transceiver circuitry could be in the milliwatt range, which is substantially higher. It becomes necessary for zero-energy devices to store energy. Furthermore, the electronic circuits in a zero-energy device require a minimum input voltage to operate, a voltage that is typically many orders of magnitude larger than the voltage at the output of the antenna. How to efficiently up-convert the input voltage to values that the electronics can run on is another key challenge.

The extremely limited energy supply for such zero-energy devices limits the amount of data possible to transmit, in many cases as low as a couple of bytes, although this is highly dependent on the distance and radio conditions. An option to conserve energy consists in operating the devices in a duty cycled manner. This means that the devices would wake up or be waken-up by an external trigger when there is something to transmit. Mobility handling as it is performed today in cellular networks will be practically impossible. The levels of energy would by far not be enough to constantly measure network conditions in support of hand-over decisions. More challenges are faced in the security area, for example the encryption of the IMSI device identity costs several orders of magnitude more energy than could be harvested. The above constrains call for new physical-layer designs, as the traditional transmission schemes may not be feasible.

### 10.7.2 Energy efficiency with impact on standardisation and policy

Future standardisation should quantify the impact of standards on energy efficiency and energy consumption. Each published standard should consider energy efficiency measures and make statements about $CO_2$ and Green House Gas (GHG) impacts. This topic is now reaching the mainstream, as identified in Chapter 2.

In the future telecommunication system, it will be necessary to concretely address the energy efficiency question with concrete actionable interactions with the system. A customer should be able to query the system about the expected energy consumption of a service that is provisioned

for him (energy expense per hour/day/year). Similarly, a customer should be able to provide an upper limit for the energy use for a service or even more sophisticated to provide indications about how much QoS degradation he is prepared to accept for a give upper bound of energy consumption.

Policy is in place that requires products to specify energy consumption, e.g. in litres of fossil fuel or KWh per passenger-distance travelled (cars, trains, airplanes) or energy and water consumption for white goods (dishwasher, washing machine...). ICT service are currently not subject not subject to similar policy requirements, yet in the future policy may require knowledge about the energy consumption for a 3 minutes phone call.

In order to achieve this, it is necessary to define and agree on a concept on how to sustainably measure the energy requirement for a telecommunications system in a technology neutral way. Hence it is necessary to:

a) Derive models, mechanisms and potential interventions to increase energy efficiency.
b) Specify metrics to capture energy consumption of resources in highly distributed, virtualised environments.
c) Define how the aforementioned model can be instrumented with standard interfaces that allow:
    a. the query and collection of energy consumption metrics
    b. The introduction of target costing in terms of energy requirement per task
d) Specify the relationship of energy consumption with service key performance indicators and related system key value indicators

### 10.7.3 Sustainable ICT

The discussion on green ICT mostly concentrates on the $CO_2$ and other GHG emissions, related directly or indirectly to the use of ICT. The overall environmental sustainability of ICT is rarely in focus. The effects of ICT are commonly ordered in first, second and third order effects. The first order effects are directly related to the mere physical existence of ICT and include production, use and end-of-life treatment. The second order effects are related to the application of ICT and include effects leading to optimisation of processes in other sectors (e.g. traffic optimisation), substitution effects (e.g., e-processes that replace traffic) and induction effects (when ICT creates more demand in other sectors). The third order effects are related to the societal changes that ICT brings along. This includes the deep structural change towards a de-materialisation and de-carbonisation of economy and society, the rebound effects, and the increased dependency on a critical infrastructure. The rebound effects include the stimulation of increased demand due to time-saving optimisation (e.g. increased leisure time traffic), the software-induced hardware obsolescence and the miniaturisation paradox, which indicates that hardware is getting cheaper faster than it is getting smaller.

Considering the first-order effects, we must keep in mind the environmental impacts of ICT caused by the material used in the production (e.g. fossil fuels, water, and chemicals), the possible long-term health effects due to chemical exposure during manufacturing, and exposure to toxic materials in ICT arising from recycling. The manufacturing process of semiconductor chips consumes large amounts of ultra-pure water. Major units of ICT equipment are composed of various materials, which, in turn, consist of a wide range of chemicals, elements and heavy metals. Some of these materials, such as platinum, have a high recovery and recycling efficiency (95%), while others cannot

be recycled at all (e.g. mercury, arsenic and barium). It is essential to make the shift from simply calculating $CO_2$ emissions of ICT production to evaluating the net impact of the technology life-cycle, including operations and use considerations, as well as end-of-life management.

Recycling of e-waste pays off in environmental terms due to the materials recovered, saving energy otherwise used for their primary production. However there is a much more profound reason to recover certain materials from e-waste, namely their sparse occurrence on earth. One example is indium (In), a rare chemical element with soft, malleable and easily fusible properties. Its current primary application is in alloyed form of indium tin oxide to form the transparent, conductive coating of liquid crystal displays (LCD). The amount of indium consumed is largely a function of worldwide LCD production, accounting for more than 50% of its worldwide consumption. Based on the current world-wide reserve base of economically-viable indium and the low recycling rate, it has been estimated that there is about 20-30 years' supply of indium left.

It is necessary to develop models and approached to estimate the global total cost of ownership of ICT in economy and society and include in the model, parameters beyond energy and GHG emissions, such as the use of primary resources, ultra-pure water and the induced second and third order rebound effects.

An adjacent topic for attention is a phenomenon that can be explained with the Jevons Paradox, which states that technological progress that increases the efficiency with which a resource is used, tends to increase (rather than decrease) the rate of consumption of that resource. Although the English economist William Stanley Jevons postulated this in 1865 in the context of consumption of coal, it can be easily transposed to computers, networks and data. So, the more efficiently we capture, store and process data, the more data we capture, transmit and store. This phenomenon can easily jeopardise the efforts towards energy efficiency we undertake, if we consider the global total cost of ownership.

### 10.7.4 Sustainable mobile networks beyond 5G

The connected society we live in will generate a vast amount of data. Mobile networks already have a considerable carbon footprint, and their worldwide energy consumption is expected to rise to 1,700 TWh of electric energy by 2030 (a figure equal to 60% of the total EU electricity consumption in 2019). Emerging AI-driven applications, such as extended reality, smart health, smart factories, and autonomous driving, to name a few, will further push the energy consumption due to the massive amounts of computation they require. As a by-product of the above in the current climate change context, communication networks need to become sustainable, designed and operated with an energy viewpoint to address the environmental dimension in an integral manner.

Another important angle for pursuing energy-sustainable future mobile networks is the non-availability of reliable power grids in providing connectivity in rural areas of developing and underdeveloped countries. For example, in Africa, only 10% of individuals have access to the electrical grid, and cellular coverage is only 15%, as the ICT development cannot keep up with the fast market growth using conventional electricity-hungry BSs. Hence, the design of the mobile networks should go one step further from the traditional energy-efficient design to an energy-neutral paradigm. Self-powered BSs are an essential technological step in the making the above a reality. They will rely on renewable energy such as wind, solar, kinetic, and radiated power, as well

as high-efficiency, high-capacity batteries. Renewable-Energy powered BSs (REBs) could also be incorporated into conventional networks to reduce energy bills and hence the cost per MB seen by the users. The energy-neutrality refers to the zero-sum balance between energy harvested, stored, and consumed during operation, which is a game-changer when a connection to the electricity grid is not available/feasible. Such energy-neutral operation can be achieved through a combination of cost-effective recharging of the batteries, e.g., by using excess self-generated power or recharging during the low-traffic time, as well as by using green energy as an alternative/complement to the electrical grid.

While the energy topic in wireless/mobile networks has been investigated up to a point, we are still far from an energy-neutral operational regime. This points to a need for a new kind of design, as so far, energy networks only use some limited data networ''s knowledge (e.g., as done in smart grids), and on the other hand, the work on data networks only makes some patchy considerations about energy efficiency. Operators need to be aware of energy consumption/provision, and that means being able to answer questions such as who generates energy and through what resources (e.g., renewables) or accounting for the type and quantity of energy spent. Ultimately, we should aim for a network operation that is energy-sustainable, and to make sure we achieve that, we also need to track its operation by having a policy component/roadmap in place which enforces operator behaviors that are energy-responsible (for example, this functionality could sit on top of spectrum sharing/regulation as of today and enhance it). Thus, a highly efficient integrated data-energy network technology for beyond 5G systems that will tackle the enormous energy consumption problem of current and future applications by the interplay with the energy distribution grid is paramount. To tame the growing carbon footprint of beyond 5G networks, it is crucial to devise highly energy-efficient communication and computing techniques, with a holistic look at the underlying computing/learning applications together with intelligent network management at all layers.

### 10.7.5 Energy efficient computing for large scale MIMO

Realization of the next generation large scale massive MIMO system poses the following challenges for implementation of signal processing:

- extremely high processing throughput with low latency to support wide bandwidth and complex algorithms
- ultra-low power consumption to meet power and small footprint requirements
- high scalability and flexibility to accommodate various use cases and deployment scenarios, e.g., on drone, land station, station on ship.

We estimate that a 10–- 20x improvement over current state-of-the-art technology is needed. Existing solutions fall short in providing such sizeable improvements. For example,

- CMOS scaling in the next major technology node offers only 15% to 30% improvement on speed and power over previous node, while manufacturing cost increases significantly.
- New multi-core CPU and DSP design based on von Neumann architecture suffers from "memory wall" problem which limits the achievable power efficiency [C10-25]
- Conventional digital ASIC design flow based on standard logic cell is optimized only for general purpose logic circuits. It does not take full advantage of the unique characteristics of signal processing algorithms.

New approaches are required for the system-on-chip that implements receiver signal processing at the architecture level, signal processing techniques, and packaging. A flow-based massively parallel processor to take full advantage of the data concurrency in digital signal processing is required. The array of processing elements (PE) could be based on a flow-based multi-instruction, multi-data stream (MIMD) processing paradigm in which continuous data stream (signal samples) is flowing through the Input elements and processed by a layer of PEs. Partial results are then passed to another layer of pEs for further processing until the final results are obtained. For the low-power use case, a fraction of the array is needed, while the rest of the array can be powered down.

Near-memory and in-memory processing element should be considered. In-memory PE goes deeper into the integration of memory cell and processing circuitry. In addition, analog circuits for analog multiplication and additions should be considered for low power computing in mixed signal circuits.

## 10.8 Semantic Communications

Shannon in his mathematical theory of communication addressed how one can efficiently transmit bits or symbols from the transmitter to the receiver, which he called the ‘technical problem’. He also described the ‘semantic problem’, which pertains to the question of how precisely the transmitted symbols convey meaning and the ‘effectiveness problem’ which deals with how effectively the communication achieves conduct in the desired way. However, he explicitly chose to focus only on the technical problem.

With significant advances in machine learning, and significant amount communication and information processing happening between machines, it is time to consider goal-oriented communications. In these emerging applications, massive amounts of data from many devices with strict latency requirements is required and hence it is necessary to be very efficient with bandwidth and energy. Developing a theory of semantic and effectiveness communications can lead to major increases in communication efficiency. Notions of entropy and mutual information need to be expanded to include the semantics. This topic has been referred in different chapters before, but its full realization in all layers of the communication system may have deep transformational changes on the operation of any smart infrastructure.

# 11.References

[C1-01] Networld Europe, “Strategic Research and Innovation Agenda” 2020.

[C2-01] H. Tataria, M. Shafi, A. F. Molisch, M. Dohler, H. Sjöland, and F. Tufvesson, “6G Wireless Systems: Vision, Requirements, Challenges, Insights, and Opportunities,” Proc. IEEE, vol. 109, no. 7, pp. 1166–1199, Jul. 2021.

[C2-02] C. Pan et al., “Reconfigurable Intelligent Surfaces for 6G Systems: Principles, Applications, and Research Directions,” IEEE Commun. Mag., vol. 59, no. 6, pp. 14–20, Jun. 2021.

[C2-03] H. Jiang, M. Mukherjee, J. Zhou, and J. Lloret, “Channel Modeling and Characteristics for 6G Wireless Communications,” IEEE Netw., vol. 35, no. 1, pp. 296–303, Jan. 2021.

[C2-04] W. Saad, M. Bennis, and M. Chen, “A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems,” IEEE Netw., vol. 34, no. 3, pp. 134–142, May 2020.

[C2-05] M. Matthaiou, O. Yurduseven, H. Q. Ngo, D. Morales-Jimenez, S. L. Cotton, and V. F. Fusco, “The Road to 6G: Ten Physical Layer Challenges for Communications Engineers,” IEEE Commun. Mag., vol. 59, no. 1, pp. 64–69, Jan. 2021.

[C2-06] X. Qiao, Y. Huang, S. Dustdar, and J. Chen, “6G Vision: An AI-Driven Decentralized Network and Service Architecture,” IEEE Internet Comput., vol. 24, no. 4, pp. 33–40, Jul. 2020.

[C2-07] M. Polese, J. M. Jornet, T. Melodia, and M. Zorzi, “Toward End-to-End, Full-Stack 6G Terahertz Networks,” IEEE Commun. Mag., vol. 58, no. 11, pp. 48–54, Nov. 2020.

[C2-08] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, and M. Zorzi, “Toward 6G Networks: Use Cases and Technologies,” IEEE Commun. Mag., vol. 58, no. 3, pp. 55–61, Mar. 2020.

[C2-09] N. Niebert, A. Schieder, and H. Abramowicz, “Ambient networks: an architecture for communication networks beyond 3G,” IEEE Wirel. Commun., 2004 [Online].

[C2-10] J. Touch et al., “A Dynamic Recursive Unified Internet Design (DRUID),” Comput. Netw., vol. 55, no. 4, pp. 919–935, Mar. 2011.

[C2-11] Dressler, F., Chiasserini, C.F., Fitzek, F.H.P., Karl, H., Lo Cigno, R., Capone, A., Casetti, C., Malandrino, F., Mancuso, V., Klingler, F., Rizzo, G.: V-Edge: Virtual Edge Computing as an Enabler for Novel Microservices and Cooperative Computing. IEEE Network Magazine. (2022).

[C2-12] A. Avizienis, J.-C. Laprie, B. Randell, and C. Landwehr, “Basic concepts and taxonomy of dependable and secure computing,” IEEE Trans. Dependable Secure Comput., vol. 1, no. 1, pp. 11–33, Jan. 2004.

[C2-13] A. Sapio et al., “Scaling Distributed Machine Learning with {In-Network} Aggregation,” in 18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21), 2021, pp. 785–808.

[C2-14] S. Dustdar, V. Casamajor Pujol, and P. K. Donta, “On distributed computing continuum systems,” IEEE Trans. Knowl. Data Eng., pp. 1–1, 2022.

[C3-01] S. Ren et al., “Routing and Addressing with Length Variable IP Address”, Proc. of ACM SIGCOMM NEAT workshop, 2019.

[C3-02] R. Li et al., “A Framework for Qualitative Communications using Big Packet Protocol”, Proc. of ACM SIGCOMM NEAT workshop, 2019.

[C3-03] R. Li, “Network 2030 and New IP”, 15th International Conference on Network and Service Management, IEEE NOMS 2020. Online: http://www.cnsm-conf.org/2019/files/slides-Richard.pdf. Last accessed May 20, 2020.

[C3-04] Pouzin Society, “RINA: Building a Better Network”, available online: http://pouzinsociety.org/education/highlights, Last accessed May 20, 2020.

[C3-05] D. Trossen et al., “Service-based Routing at the Edge”, available online: https://arxiv.org/abs/1907.01293v1. Last accessed May 20, 2020.

[C3-06] SCION Internet Architecture, online: https://www.scion-architecture.net/. Last accessed May 20, 2020.

[C3-07] “62% of all Internet traffic in Asia Pacific will cross Content Delivery Networks (CDNs) by 2021”, online: https://www.prnewswire.com/news-releases/62-of-all-internet-traffic-in-asia-pacific-will-cross-content-delivery-networks-cdns-by-20211-300528121.html. Last accessed May 20, 2020.

[C3-08] “Netflix Eats Up 15% of All Internet Downstream Traffic Worldwide”, Available online: https://variety.com/2018/digital/news/netflix-15-percent-internet-bandwidth-worldwide-study-1202963207/. Last accessed May 20, 2020.

[C3-09] IETF SFC WG. Available Online: https://datatracker.ietf.org/wg/sfc/about/. Last accessed May 20, 2020.

[C3-10] IETF ANIMA WG. Available Online: https://datatracker.ietf.org/wg/anima/about/. Last accessed May 20, 2020.

[C3-11] IETF COIN RG. Available Online: https://irtf.org/coinrg. Last accessed May 20, 2020.

[C3-12] IETF FORCES WG. Available Online: https://datatracker.ietf.org/wg/forces/about/. Last accessed May 20, 2020.

[C3-13] IETF QUIC WG. Online: https://datatracker.ietf.org/wg/quic/about/. Last accessed May 20, 2020.

[C3-14] L. Popa, A. Ghodsi, I. Stoica, ‘HTTP as the narrow waist of the future internet’, Hotnets-IX: Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks October 2010.

[C3-15] Roland Bless, Zoran Despotovic, Artur Hecker, Martina Zitterbart, “KIRA: Distributed Scalable ID-Based Routing With Fast Forwarding”, in proc. IFIP NETWORKING 2022, June 13-16 2022, Catania, Italy

[C3-16] J.H. Saltzer, D.P. Reed and D.D. Clark, “End-to-End Arguments in System Design”, ACM Transactions in Computer Systems, pp. 277-288, vol. 2, ACM, 1984.

[C3-17] B. Carpenter and B. Liu. 2020. Limited Domains and Internet Protocols. RFC 8799. IETF. http://tools.ietf.org/rfc/rfc8799.txt.

[C3-18] B. Carpenter, J. Crowcroft, D. Trossen, “Limited domains considered useful”, ACM Computer Communication Review, pp. 22-28, volume 51, issue 3, ACM, 2021.

[C3-19] D. Clark, “The Design Philosophy of the DARPA Internet Protocols”, Proc. ACM SIGCOMM ‘88, Computer Communication Review Vol. 18, No. 4, August 1988, pp. 106–114.

[C3-20] A. S. Asrese, S. J. Eravuchira, V. Bajpai, P. Sarolahti, J. Ott, “Measuring Web Latency and Rendering Performance: Method, Tools \& Longitudinal Dataset”, IEEE Transactions on Network and Service Management, vol. 16, num. 2, pp. 535-549, IEEE, 2019.

[C3-21] M. Bloecher, R, Khalili, L. Wang, P. Eugster, “Letting off STEAM: Distributed Runtime Traffic Scheduling for Service Function Chaining”, in proc. IEEE INFOCOM 2020.

[C3-22] K. Khandaker, D. Trossen, R. Khalili, Z. Despotovic, A. Hecker, J. Carle, “CArDS: Dealing a New Hand in Reducing Service Request Completion Times”, IFIP NETWORKING 2022.

[C3-23] European Commission, “The European Green Deal”, Available online: https://ec.europa.eu/info/sites/info/files/european-green-deal-communication_en.pdf. Last accessed May 20, 2020.

[C3-24] “Between 10 and 20% of electricity consumption from the ICT* sector in 2030?”, online: https://www.enerdata.net/publications/executive-briefing/expected-world-energy-consumption-increase-from-digitalization.html. Last accessed May 20, 2020.

[C3-25] M. Caesar et al., “ROFL: routing on flat labels”, ACM SIGCOMM Computer Communication Review, August 2006, https://doi.org/10.1145/1151659.1159955.

[C3-26] G. Xylomenos et al, “A Survey of Information-Centric Networking Research”, IEEE Communications Surveys & Tutorials (Volume: 16, Issue: 2, Second Quarter 2014).

[C3-27] Trossen et al., “Name-Based Service Function Forwarder (nSFF) Component within a Service Function Chaining (SFC) Framework”, Internet Society, RFC 8677, online: https://datatracker.ietf.org/doc/rfc8677/. Last accessed May 20, 2020.

[C3-28] D. Clark, J. Wroclawski, K. Sollins, R. Braden, “Tussle in Cyberspace: Defining Tomorrow’s Internet”, Proc. ACM SIGCOMM 2002. Available Online: http://conferences.sigcomm.org/sigcomm/2002/papers/tussle.pdf. Last accessed May 20, 2020.

[C3-29] Maja Curić, Georg Carle, Zoran Despotovic, Ramin Khalili, and Artur Hecker, “SDN on ACIDs”, in Cloud-Assisted Networking workshop, in proc. of ACM CONEXT 2017, Incheon, South Korea, December 2017.

[C3-30] M. Curic, Z. Despotovic, A. Hecker, G. Carle, “FitSDN: Flexible Integrated Transactional SDN”, IEEE LCN 2019.

[C3-31] Tim Höfer, Sebastian Bierwirth und Reinhard Madlener, „C15 - Energie Mehrverbrauch in Rechenzentren bei Einführung des 5G Standards“, available online: https://www.eon.com/en/about-us/green-internet.html. Last accessed May 20, 2020.

[C3-32] E. Masanet, A. Shehabi, N. Lei, S. Smith, J. Koomey, “Recalibrating global data center energy-use estimates,” Science, vol. 367, no. 6481, pp. 984-986, Feb. 2020.

[C3-33] E. Masanet, J. Koomey, “Does not compute: Avoiding pitfalls assessing the Internet’s energy and carbon impacts,” Joule, vol. 5, pp. 1-4, July 2021; https://doi.org/10.1016/j.joule.2021.05.007.

[C3-34] Advanced Configuration and Power Interface Specification. Available: https://uefi.org/sites/default/files/resources/ACPI_6_2.pdf.

[C3-35] ETSI ES 203 682 V1.1.0, “Environmental Engineering (EE); Green Abstraction Layer (GAL); Power management capabilities of the future energy telecommunication fixed network nodes; Enhanced Interface for power management in Network Function Virtualisation (NFV) environments,” 2019.

[C3-36] ITU-T L.1362, “Interface for power management in network function virtualization environments – Green abstraction layer version 2,” 2019.

[C3-37] R. Zoppoli, M. Sanguineti, G. Gnecco, T. Parisini, Neural Approximations for Optimal Control and Decision, Springer Nature, Cham, Switzerland, 2019.

[C3-38] Martin Rapp, Ramin Khalili, Kilian Pfeiffer, Jörg Henkel, “DISTREAL: Distributed Resource-Aware Learning in Heterogeneous Systems”, 36th AAAI 2022, Feb 2022, Vancouver, Canada.

[C3-39] Jing Tan, Ramin Khalili, Holger Karl, Artur Hecker, “Multi-Agent Distributed Reinforcement Learning for Making Decentralized Offloading Decisions”, in proc. IEEE INFOCOM 2022.

[C3-40] NGMN’s report “Green Future Networks – Network Energy Efficiency“, v1.0, Nov. 2021, available online https://www.ngmn.org/wp-content/uploads/211009-GFN-Network-Energy-Efficiency-1.0.pdf.

[C3-41] E. Hossein and F. Fredj, “Energy Efficiency of Machine-Learning-Based Designs for Future Wireless Systems and Networks”, Editorial, *IEEE Transactions on Green Communications and Networking*, vol. 5, no. 3, pp. 1005-1010, Sept. 2021.

[C3-42] R. Bolla, R. Bruschi, F. Davoli, J. F. Pajo, “A Model-Based Approach Towards Real-Time Analytics in NFV Infrastructures,” IEEE Transactions on Green Communications and Networking, published online 20 Dec. 2019; DOI: 10.1109/TGCN.2019.2961192.

[C3-43] K Yeow, A Gani, RW Ahmad, J.J.P.C. Rodrigues, Kwangman Ko, "Decentralized consensus for edge-centric internet of things: A review, taxonomy, and research issues", Vol 6, pp. 1513 - 1524, IEEE Access, 2017.

[C3-44] Cisco Global Cloud Index: Forecast and Methodology, 2016–2021 White Paper

[C3-45] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo and J. Zhang, "Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing," in Proceedings of the IEEE, vol. 107, no. 8, pp. 1738-1762, Aug. 2019.

[C3-46] H. Kim, J. Park, M. Bennis, and S.-L. Kim, “On-device federated learning via blockchain and its latency analysis,” arXiv:1808.03949, 2018.

[C3-47] S. Han, J. Pool, J. Tran, and W. Dally, “Learning both weights and connections for efficient neural network,” in Advances in neural information processing systems, 2015, pp. 1135–1143.

[C3-48] E. Strubell, A. Ganesh and A. McCallum, “Energy and Policy Considerations for Deep Learning in NLP”, Annual Meeting of the Association for Computational Linguistics (ACL short). Florence, Italy. July 2019.

[C3-49] T.-J. Yang, Y.-H. Chen, and V. Sze, “Designing energy-efficient convolutional neural networks using energy-aware pruning,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2017.

[C3-50] R. Zoppoli, M. Sanguineti, G. Gnecco and T. Parisini, “Neural Approximations for Optimal Control and Decision”, Springer, 2019.

[C3-51] A. Lacoste, A. Luccioni, V. Schmidt, Victor and T. Dandres, "Quantifying the Carbon Emissions of Machine Learning", arXiv preprint arXiv:1910.09700, 2019

[C3-52] http://www.green-algorithms.org/

[C4-1] https://undocs.org/en/A/RES/217(III)

[C4-2] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN

[C4-3] https://www.enisa.europa.eu/publications/telecom-security-incidents-2021
[C4-4] https://www.enisa.europa.eu/publications/enisa-threat-landscape-for-5g-networks
[C4-5] https://ec.europa.eu/digital-single-market/en/network-and-information-security-nis-directive
[C4-6] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52017PC0477R(02)
[C4-7] https://ec.europa.eu/digital-single-market/en/news/cybersecurity-5g-networks-eu-toolbox-risk-mitigating-measures
[C4-8] https://www.nextgalliance.org/white_papers/trust-security-and-resilience-for-6g-systems/
[C4-9] https://www.gsma.com/security/resources/gsma-security-landscape-report-2022/6g-ia
[C4-10] https://bscw.5g-ppp.eu/pub/bscw.cgi/d381923/INSPIRE-5Gplus_White_Paper_HLA_FV_GA_Abbrev.pdf
[C4-11] Scott-Hayward, S., Natarajan, S., & Sezer, S.: (2016). A Survey of Security in Software Defined Networks, IEEE Communications Surveys and Tutorials.
[C4-12] Olli Mämmelä, Jouni Hiltunen, Jani Suomalainen, Kimmo Ahola, Petteri Mannersalo, Janne Vehkaperä: ”Towards Micro-Segmentation in 5G Network Security”. EuCNC 2016.
[C4-13] P.Porambage and al. The roadmap to 6G security and Privacy, IEEEE Open Journal of the Communication Society, May 2021
[C5-1] Open Source Software for RAN: https://www.o-ran.org/software
[C5-2] 2021 Accelerate State of DevOps Report, available at: https://cloud.google.com/devops/state-of-devops?utm_source=google&utm_medium=blog&utm_campaign=FY19-Q3-global-demandgen-website-wd-gcp_gtm_stateofdevops
[C5-3] Proposal for a directive of the European Parliament and of the Council on energy efficiency, available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0558
[C5-4] CIS benchmarks, available at: https://www.cisecurity.org/cis-benchmarks/
[C5-5] CNCF Cloud Native Definition v1.0, available at: https://github.com/cncf/toc/blob/main/DEFINITION.md
[C5-6] DS. Dustdar, V. Casamajor Pujol, P. Kumar Donta, "On distributed computing continuum systems", IEEE Transactions on Knowledge and Data Engineering, IEEE, pages: 14, 13 January 2022
[C5-7] Kosar, T., Bohra, S. and Mernik, M., 2016. Domain-specific languages: A systematic mapping study. Information and Software Technology, 71, pp.77-91.
[C5-8] Kleppe, A., 2008. Software language engineering: creating domain-specific languages using metamodels. Pearson Education
[C5-9] Marshall, C. and Rossman, G.B., 2014. Designing qualitative research. Sage publications.
[C5-10] Bass, L., Weber, I. and Zhu, L., 2015. DevOps: A software architect's perspective. Addison-Wesley Professional.
[C5-11] Gartner: “Future of IT Infrastructure Is Always On, Always Available, Everywhere" https://www.gartner.com/en/newsroom/press-releases/2018-12-03-gartner-says-the-future-of-it-infrastructure-is-always-on-always-available-everywhere
[C5-12] 5G-ACIA White Paper, 2021. Integration of 5G with time-sensitive networking for industrial communications. 5G Alliance for Connected Industries and Automation.
[C5-13] ONF Reference Designs: https://opennetworking.org/reference-designs/
[C5-14] one6G Association: https://opennetworking.org/reference-designs/
[C5-15] SND Zoo repository: https://sndzoo.github.io/
[C5-16] Orgalim. Environment: Orgalim position on the Sustainable Products Initiative. https://orgalim.eu/position-papers/environment-orgalim-position-sustainable-products-initiative-0
[C5-17] NESSI, “Software and Smart Networks and Services,” 2020, available at: https://nessi.eu/wp-content/uploads/2020/09/NESSI-Software-and-SNS-issue1.pdf
[C5-18] Coutinho, R.W.L., Boukerche,A., 2022. Design of Edge Computing for 5G-Enabled Tactile Internet-Based Industrial Applications. IEEE Communications Magazine.
[C5-19] Shih-Chun Lin, Kwang-Cheng Chen, Ali Karimoddini, 2021. SDVEC: Software-Defined Vehicular Edge Computing with Ultra-Low Latency. IEEE Communications Magazine.

[C5-20] Shiwei Lai, Rui Zhao, Shunpu Tang, Junjuan Xia, Fasheng Zhou, Liseng Fan, 2021. Intelligent secure mobile edge computing for beyond 5G wireless networks. Elsevier, physical communication, vol. 45.

[C5-21] NESSI, "Software and Quantum Computing", 2022, available at: https://nessi.eu/wp-content/uploads/2022/05/NESSI-Quantum-Computing-issue-1.pdf

[C5-22] NESSI, "Software and Human Centricity", 2022, available at: https://nessi.eu/wp-content/uploads/2022/04/NESSI-Human-Centricity-issue-1.pdf

[C5-23] COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT AND THE COUNCIL. Twinning the green and digital transitions in the new geopolitical context, available at: com_2022_289_1_en.pdf (europa.eu)

[C6-1] M. A. Uusitalo et al., "6G vision, value, use cases and technologies from European 6G flagship project Hexa-X," IEEE Access, vol. 9, pp. 160004-160020, 2021.

[C6-2] N. Rajatheva et al, "White paper on broadband connectivity in 6G," arXiv preprint: https://arxiv.org/abs/2004.14247, 2020.

[C6-3] Next G Alliance Report, "6G applications and use cases," https://nextgalliance.org/white_papers/6g-applications-and-use-cases/, June 2022.

[C6-4] Report ITU-R M.2410-0, "Minimum requirements related to technical performance for IMT-2020 radio interface(s)," 11/2017, https://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-M.2410-2017-PDF-E.pdf.

[C6-5] H. Shokri-Ghadikolaei, F. Boccardi, C. Fischione, G. Fodor, and M. Zorzi, "Spectrum sharing in mmWave cellular networks via cell association, coordination, and beamforming," IEEE J. Sel. Areas Commun., Nov. 2016.

[C6-6] G. Berardinelli et al., "Extreme communication in 6G: Vision and challenges for 'in-X' subnetworks," IEEE Open Journal of the Communications Society, vol. 2, pp. 2516-2535, 2021.

[C6-7] S. Lagen, L. Giupponi, S. Goyal, N. Patriciello, B. Bojovic, A. Demir, and M. Beluri, "New radio beam-based access to unlicensed spectrum: Design challenges and solutions," IEEE Communications Surveys & Tutorials, vol. 22, no. 1, pp. 8-37, March 2020.

[C6-8] Cisco: Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016 - 2021 (White Paper), [Online]. Available at: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html.

[C6-9] Ericsson Expert Analytics. [Online]. Available: https://www.ericsson.com/ourportfolio/products/expert-analytics.

[C6-10] M. Agiwal, A. Roy and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," IEEE Communications Surveys & Tutorials, vol. 18, no. 3, pp. 1617-1655, 2016.

[C6-11] C. Shepard, H. Yu, N. Anand, E. Li, T. Marzetta, R. Yang and L. Zhong, "Argos: Practical many-antenna base stations," in Proc. of the 18th ACM annual international conference on Mobile computing and networking, pp. 53-64, 2012.

[C6-12] J. Vieira, S. Malkowsky, K. Nieman, Z. Miers, N. Kundargi, L. Liu, I. Wong, V. Owall, O. Edfors and F. Tufvesson, "A exible 100-antenna testbed for massive MIMO," in Proc. IEEE Globecom Workshops, pp. 287-293, 2014.

[C6-13] S.-W. Jeon, S.-N. Hong, M. Ji, G. Caire and A. F. Molisch, "Wireless multihop device-to-device caching networks," IEEE Transactions on Information Theory, vol. 63, no. 3, pp. 1662-1676, 2017.

[C6-14] M. Ji, G. Caire and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," IEEE Transactions on Information Theory, vol. 62, no. 2, pp. 849-869, 2016.

[C6-15] M. Maddah-Ali and U. Niesen, "Fundamental limits of caching," IEEE Transactions on Information Theory, vol. 60, no. 5, pp. 2856-2867, May 2014.

[C6-16] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," IEEE Transactions on Information Theory, vol. 59, no. 12, pp. 8402-8413, 2013.

[C6-17] M. Bayat, R. K. Mungara and G. Caire, "Achieving spatial scalability for coded caching over wireless networks," arXiv preprint: arXiv:1803.05702, 2018.

[C6-18] K. Wan and G. Caire, "On coded caching with private demands", submitted to TIT, available online: arXiv:1908.10821 [cs.IT], 2019.

[C6-19] D. Bethanabhotla, G. Caire and M. J. Neely, “Wiflix: Adaptive video streaming in massive MU-MIMO wireless networks,” IEEE Transactions on Wireless Communications, vol. 15, no. 6, pp. 4088–103, 2016

[C6-20] A. Narayanan, S. Verma, E. Ramadan, P. Babaie, and Z. Zhang, “DeepCache: A deep learning based framework for content caching,” in Proceedings of Workshop on Network Meets AI & ML, pp. 48–53, Aug. 2018.

[C6-21] M.S. Islim, R.X. Ferreira, X. He, E. Xie, S. Videv, S. Viola, S. Watson, N. Bamiedakis, R. V. Penty, I.H. White, A.E. Kelly, E. Gu, H. Haas and M. D. Dawson, “Towards 10 Gb/s OFDM-based visible light communication using a GaN violet micro-LED”, Photonics Research, vol. 2, no. 5, pp. A35-A48, 2017.

[C6-22] T. Cogalan and H. Haas, “Why would 5G need optical wireless communications?” in 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), pp. 1-6, 2017.

[C6-23] Y. Tan and H. Haas, “Coherent LiFi system with spatial multiplexing,” IEEE Transactions on Communications, vol. 69, no. 7, pp. 4632-4643, July 2021.

[C6-24] Y. F. Huang *et al*, “17.6-Gbps universal filtered multi-carrier encoding of GaN blue LD for visible light communication,” in Proc. of the *Conference on Lasers and Electro-Optics (CLEO)*, pp. 1-2, 2017.

[C6-25] C. Lee et al., “26 Gbit/s LiFi system with laser-based white light transmitter,” Journal of Lightwave Technology, vol. 40, no. 5, pp. 1432-1439, March, 2022.

[C6-26] E. Calvanese Strinati et al, “6G: The next frontier: From holographic messaging to artificial intelligence using subterahertz and visible light communication,” IEEE Vehicular Technology Magazine, vol. 14, no. 3, pp. 42-50, Sept. 2019.

[C6-27] H. Haas, L. Yin, Y. Wang and C. Chen, “What is LiFi?” *Journal of Lightwave Technology*, vol. 34, no. 6, pp. 1533-1544, March, 2016.

[C6-28] E. Sarbazi, H. Kazemi, M. Dehghani Soltani, M. Safari and H. Haas, “A Tb/s indoor optical wireless access system using VCSEL arrays,” IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications, 2020.

[C6-29] H. Kazemi et al., “A Tb/s indoor MIMO optical wireless backhaul system using VCSEL arrays,” *IEEE Transactions on Communications*, vol. 70, no. 6, pp. 3995-4012, June 2022.

[C6-30] D. Tsonev, S. Videv and H. Haas, “Unlocking spectral efficiency in intensity modulation and direct detection systems,” IEEE Journal on Selected Areas in Communications, vol. 33, no. 9, pp. 1758-1770, Sept. 2015

[C6-31] I. F. Akyildiz, J. M. Jornet and C. Han, “TeraNets: Ultra-broadband communication networks in the Terahertz band,” IEEE Wireless Communications Magazine, vol. 21, no. 4, pp. 130-135, August 2014.

[C6-32] I. F. Akyildiz, C. Han, Z. Hu, S. Nie, J. M. Jornet, “Terahertz band communication: An old problem revisited and research directions for the next decade,” to appear in IEEE Transactions on Communications, 2022

[C6-33] J. V. Siles, K. B. Cooper, C. Lee, R. H. Lin, G. Chattopadhyay, and I. Mehdi, “A new generation of room-temperature frequency-multiplied sources with up to 10× higher output power in the 160-GHz–1.6-THz range,” IEEE Transactions on Terahertz Science and Technology, vol. 8, no. 6, 596-604, 2018.

[C6-34] J. Shi, M. C. Lo, L. Zhang, O. Ozolins, ... and L. K. Oxenløwe, “Integrated dual-laser photonic chip for high-purity carrier generation enabling ultrafast terahertz wireless communications,” Nature communications, 13(1), 1388, March 2022.

[C6-35] J. M. Jornet and I. F. Akyildiz, “Graphene-based plasmonic nano-transceiver for terahertz band communication,” in Proc. of the 8th European Conference on Antennas and Propagation (EuCAP), The Hague, The Netherlands, April 2014.

[C6-36] Crabb, Justin, Xavier Cantos-Roman, Josep M. Jornet, and Gregory R. Aizin, “Hydrodynamic theory of the Dyakonov-Shur instability in graphene transistors,” Physical Review B 104, no. 15, 2021.

[C6-37] Liaskos, Christos, Shuai Nie, Ageliki Tsioliaridou, Andreas Pitsillides, Sotiris Ioannidis, and Ian Akyildiz. “A new wireless communication paradigm through software-controlled metasurfaces,” IEEE Communications Magazine, vol. 56, no. 9, 162-169, 2018.

[C6-38] H. Abdellatif , V. Ariyarathna, S. Petrushkevich , A. Madanayake , J. M. Jornet, “A real-time ultra-broadband software-defined radio platform for Terahertz communications,” in Proc. of the IEEE Conference on Computer Communications Workshops (INFOCOM) - Demo Track, May 2022.

[C6-39] Y. J. Guo, M. Ansari, R. W. Ziolkowski and N. J. G. Fonseca, “Quasi-optical multi-beam antenna technologies for B5G and 6G mmWave and THz networks: A review," IEEE Open Journal of Antennas and Propagation, vol. 2, pp. 807-830, 2021.

[C6-40] J. M. Jornet and I. F. Akyildiz, “Channel modeling and capacity analysis for electromagnetic wireless nanonetworks in the terahertz band,” IEEE Transactions on Wireless Communications, vol. 10, no. 10, pp. 3211-3221, October 2011.

[C6-41] C. Han, A.O. Bicen and I. Akyildiz, “Multi-ray channel modeling and wideband characterization for wireless communications in the terahertz band,” IEEE Transactions on Wireless Communications, vol. 14, no. 5, pp. 2402–2412, 2015.

[C6-42] Y. Chen, Y. Li, C. Han, Z. Yu, and G. Wang, “Channel measurement and ray-tracing-statistical hybrid modeling for low-terahertz indoor communications,” IEEE Transactions on Wireless Communications, vol. 20, no. 12, 8163-8176, 2021.

[C6-43] Y. Xing, T.S. Rappaport, and A. Ghosh, “Millimeter wave and sub-THz indoor radio propagation channel measurements, models, and comparisons in an office environment,” IEEE Communications Letters 25, no. 10, 3151-3155, 2021.

[C6-44] J. M. Jornet and I. F. Akyildiz, “Femtosecond-long pulse-based modulation for terahertz band communication in nanonetworks,” IEEE Transactions on Communications, vol. 62, no. 5, pp. 1742-1754, May 2014.

[C6-45] Z. Hossain and J. M. Jornet, “Hierarchical bandwidth modulation for ultra-broadband Terahertz communications,” in Proc. of the IEEE International Conference in Communications (ICC), Shanghai, China, May 2019.

[C6-46] Han, Chong, and Ian F. Akyildiz. "Distance-aware bandwidth-adaptive resource allocation for wireless systems in the terahertz band." IEEE Transactions on Terahertz Science and Technology, vol. 6, no. 4, pp. 541-553, 2016.

[C6-47] Q. Xia, Z. Hossain, M. Medley, and J. M. Jornet, “Synchronization and medium access control protocol for Terahertz-band communication networks,” IEEE Transactions on Mobile Computing, vol. 20, no. 1., pp. 2-18, January 2021.

[C6-48] Q. Xia and J. M. Jornet, “Multi-hop relaying distribution strategies for Terahertz-band communication networks: A cross-layer analysis,” to appear in IEEE Transactions on Wireless Communications, 2022.

[C6-49] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi and R. W. Heath, “Spatially sparse precoding in millimeter wave MIMO systems,” IEEE Transactions on Wireless Communications, vol. 13, no. 3, pp. 1499–1513, 2014.

[C6-50] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors and F. Tufvesson, “Scaling up MIMO: Opportunities and challenges with very large arrays,” IEEE Signal Processing Magazine, vol. 30, no. 1, pp. 40–60, 2013.

[C6-51] E. Torkildson, U. Madhow and M. Rodwell, “Indoor millimeter wave MIMO: Feasibility and performance,” IEEE Transactions on Wireless Communications, vol. 10, no. 12, pp. 4150–4160, 2011.

[C6-52] M. T. Ivrlač and J. A. Nossek, “Toward a circuit theory of communication,” IEEE Transactions on Circuits and Systems I, vol. 57, no. 7, pp. 1663-1683, July 2010.

[C6-53] T. Laas, J. A. Nossek, S. Bazzi and W. Xu, “On reciprocity in physically consistent TDD systems with coupled antennas,” arXiv:1907.10562v1, 2019.

[C6-54] C. Fager, T. Eriksson, F. Barradas, K. Hausmair, T. Cunha and J. C. Pedro, “Linearity and efficiency in 5G transmitters: New techniques for analyzing efficiency, linearity, and linearization in a 5G active antenna transmitter context,” IEEE Microwave Magazine, vol. 20, no. 5, pp. 35-49, May 2019

[C6-55] E. Bjornson, E. G. Larsson and T. L. Marzetta, “Massive MIMO: Ten myths and one critical question,” IEEE Communications Magazine, vol. 54, no. 2, pp. 114–123, Feb. 2016.

[C6-56] A. L. Swindlehurst, E. Ayanoglu, P. Heydari and F. Capolino, “Millimeter-wave massive MIMO: the next wireless revolution?” IEEE Communications Magazine, vol. 52, no. 9, pp. 56–62, Sep. 2014.

[C6-57] A. Singh, M. Andrello, N. Thawdar and J. M. Jornet, “Design and operation of a graphene-based plasmonic nano-antenna array for communication in the Terahertz band,” IEEE Journal of Selected Areas in Communications, vol. 38, no. 9, pp. 2104-2117, Sep. 2020

[C6-58] I. F. Akyildiz and J. M. Jornet, “Realizing ultra-massive MIMO communication in the (0.06-10) terahertz band,” Nano Communication Networks (Elsevier) Journal, vol. 8, pp. 46-54, March 2016.

[C6-59] E. De Carvalho et al., “Non-Stationarities in extra-large-scale massive MIMO,” IEEE Wireless Commun., vol. 27, no.4, pp. 74–80, Aug. 2020.

[C6-60] C. Huang et al., “Holographic MIMO surfaces for 6G wireless networks: Opportunities, challenges, and trends,” IEEE Wireless Commun., vol. 27, no.5, pp. 118–125, Oct. 2020.

[C6-61] Q. Wu and R. Zhang, “Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming,” *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5394-5409, Nov. 2019.

[C6-62] C. Huang, A. Zappone, M. Debbah and C. Yuen, “Achievable rate maximization by passive intelligent mirrors,” in Proc. IEEE ICASSP, 2018.

[C6-63] R. Liu, Q. Wu, M. Di Renzo and Y. Yuan, “A path to smart radio environments: An industrial viewpoint on reconfigurable intelligent surfaces,” IEEE Wireless Communications, Jan. 2022.

[C6-64] S. Hu, F. Rusek and O. Edfors, “Beyond massive MIMO: The potential of data transmission with large intelligent surfaces,” IEEE Transactions on Signal Processing, vol. 66, no. 10, pp. 2746–2758, May 2018.

[C6-65] S. Hu, F. Rusek and O. Edfors, “Beyond massive MIMO: The potential of positioning with large intelligent surfaces,” IEEE Transactions on Signal Processing, vol. 66, no. 7, pp. 1761–1774, April 2018.

[C6-66] P. Frenger, J. Hederen, M. Hessler and G. Interdonato, “Improved antenna arrangement for distributed massive MIMO,” Patent application WO2018103897, 2017.

[C6-67] G. Interdonato, E. Björnson, H. Ngo, P. Frenger and E. Larsson, “Ubiquitous cell-free massive MIMO communications,” EURASIP J Wireless Com Network, Aug. 2019.

[C6-68] J. Guerreiro, R. Dinis and P. Carvalho, “On the optimum multicarrier performance with memoryless nonlinearities,” IEEE Transactions on Communications, Vol. 63, No. 2, pp. 498 - 509, February 2015.

[C6-69] J. Guerreiro, R. Dinis, P. Carvalho and M. Silva, “On the achievable performance of nonlinear MIMO systems,” IEEE Communications Letters, Vol. 23, No. 10, pp. 1725 - 1729, October 2019.

[C6-70] M. Jian, G. C. Alexandropoulos, E. Basar, C. Huang, R. Liu, Y. Liu, and C. Yuen, “Reconfigurable intelligent surfaces for wireless communications: Overview of hardware designs, channel models, and estimation techniques,” arXiv: 2203.03176, 2022.

[C6-71] P. Pedrosa, R. Dinis, D. Castanheira, A. Silva and A. Gameiro, “Joint channel equalization and trackingfor V2X communications using SC-FDE schemes,” IEEE GLOBECOM, 2019

[C6-72] T. L. Marzetta, “Noncooperative cellular wireless with unlimited numbers of base station antennas,” IEEE Transactions on Wireless Communications, vol. 9, no. 11, pp. 3590-3600, Nov. 2010.

[C6-73] E. G. Larsson, O. Edfors, F. Tufvesson and T. L. Marzetta, “Massive MIMO for next generation wireless systems,” IEEE Communications Magazine, vol. 52, no. 2, pp. 186-195, Feb. 2014.

[C6-74] P. Marsch, S. Brück, A. Garavaglia, M. Schulist, R. Weber, and A. Dekorsy, “Clustering,” in Coordinated Multi-point in Mobile Communications: From theory to Practice, Ed. by P. Marsch, G. Fettweis, Cambridge University Press, New York, pp. 139–159, 2011.

[C6-75] R. Fantini, W. Zirwas, L. Thiele, D. Aziz, P. Baracca, “Coordinated multi-point transmission in 5G,” in 5G Mobile and Wireless Communications Technology, Ed. by A. Osseiran, J. Monserrat, and P.Marsch, Cambridge University Press, Cambridge, pp. 248–276, 2016.

[C6-76] G. Interdonato, E. Björnson, H. Q. Ngo, P. Frenger, and E. G. Larsson, “Ubiquitous cell-free massive MIMO Communications,” EURASIP J. Wireless Commun. and Networking, 2019.

[C6-77] X. Xu, D. Wang, X. Tao, T. Svensson, “Resource pooling for frameless network architecture with adaptive resource allocation”, SCI. CHINA Inf. Sci., vol. 56, 2013.

[C6-78] V. Jungnickel et al., “The role of small cells, coordinated multipoint, and massive MIMO in 5G,” IEEE Commun.Mag., vol. 52, no. 5, pp. 44–51, 2014.

[C6-79] 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, 2017.

[C6-80] 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, 2019.

[C6-81] P. S. Bithas, V. Nikolaidis, A. G. Kanatas and G. K. Karagiannidis, "UAV-to-Ground Communications: Channel Modeling and UAV Selection," IEEE Trans. Commun., vol. 68, no. 8, pp. 5135-5144, Aug. 2020.

[C6-82] 3GPP TR 38.801, Study on New Radio Access Technology (Release 14), 2017.

[C6-83] G. Wunder, P. Jung, M. Kasparick, T. Wild, F. Schaich, Y. Chen, S. Ten Brink, I. Gaspar, N. Michailow, A. Festag, L. Mendes, N. Cassiau, D. Ktenas, M. Dryjanski, S. Pietrzyk, B. Eged, P. Vago and F. Wiedmann, "5GNOW: Non-orthogonal, asynchronous waveforms for future mobile applications," IEEE Communications Magazine, vol. 52, no. 2, pp. 97-105, Feb. 2014.

[C6-84] Y. Medjahdi, et al., "On the road to 5G: Comparative study of physical layer in MTC context," IEEE Access, vol. 5, pp. 26556-26581, 2017.

[C6-85] A. G. Armada, "Understanding the effects of phase noise in orthogonal frequency division multiplexing (OFDM)," IEEE Trans. Broadcast., vol. 47, no. 153–159, Jun. 2001.

[C6-86] M. Girotto and A. M. Tonello, "Orthogonal design of cyclic block filtered multitone modulation," IEEE Trans. Commun., vol. 64, no. 11, pp. 4667–4679, Nov. 2016.

[C6-87] J. Abdoli, M. Jia and J. Ma, "Filtered OFDM: A new waveform for future wireless systems," in *IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 66-70, 2015.

[C6-88] Y. Tao, L. Liu, S. Liu and Z. Zhang, "A survey: Several technologies of non-orthogonal transmission for 5G," China Communications, vol. 12, no. 10, pp. 1-15, Oct. 2015.

[C6-89] X. Yu, Y. Guanghui, Y. Xiao, Y. Zhen, X. Jun and G. Bo, "FB-OFDM: A novel multicarrier scheme for 5G," in Proc. Eur. Conf. Netw. Commun. (EuCNC), pp. 271-276, Jun. 2016.

[C6-90] R. Nissel and M. Rupp, "Pruned DFT-Spread FBMC: Low PAPR, low latency, high spectral efficiency," IEEE Trans. Commun., vol. 66, no. 10, pp. 4811-4825, Oct. 2018.

[C6-91] R. Gerzaguet, N. Bartzoudis, L. G. Baltar, V. Berg, J.-B. Doré, D. Kténas, O. Font-Bach, X. Mestre, M. Payaró, M. Färber and K. Roth, "The 5G candidate waveform race: A comparison of complexity and performance," EURASIP Journal on Wireless Communications and Networking, no. 13, 2017.

[C6-92] M. van Eeckhaute, A. Bourdoux, P. de Doncker and F. Horlin, "Performance of emerging multi-carrier waveforms for 5G asynchronous communications," EURASIP Journal on Wireless Communications and Networking, 2017.

[C6-93] 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 Commun. Netw. Conf. (WCNC), 2017.

[C6-94] A. Fish, S. Gurevich, R. Hadani, A. M. Sayeed, and O. Schwartz, "Delay-Doppler channel estimation in almost linear complexity," IEEE Trans. Inf. Theory, vol. 59, no. 11, pp. 7632–7644, Nov. 2013.

[C6-95] S. C. Thompson, A. U. Ahmed, J. G. Proakis, J. R. Zeidler and M. J. Geile, "Constant envelope OFDM," IEEE Trans. Commun., vol. 56, no. 8, pp. 1300-1312, Aug. 2008.

[C6-96] A. U. Ahmed and J. R. Zeidler, "Novel low-complexity receivers for constant envelope OFDM," IEEE Trans. Signal Process., vol. 63, no. 17, pp. 4572–4582, Sep. 2015.

[C6-97] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," IEEE Communications Surveys & Tutorials, 20(3), 2294-2323, third quarter 2018.

[C6-98] Y. Yuan, Z. Yuan and L. Tian, "5G non-orthogonal multiple access study in 3GPP," IEEE Communications Magazine, vol. 58, no. 7, pp. 90-96, July 2020.

[C6-99] Y. Yuan et al., "Non-orthogonal transmission technology in LTE evolution," IEEE Communications Magazine, vol. 54, no. 7, pp. 68-74, July 2016.

[C6-100] L. Zhu, J. Zhang, Z. Xiao, X. Cao and D. O. Wu, "Optimal user pairing for downlink non-orthogonal multiple access (NOMA)," IEEE Wireless Communications Letters, vol. 8, no. 2, pp. 328-331, April 2019.

[C6-101] H. V. Nguyen, V. Nguyen, O. A. Dobre, D. N. Nguyen, E. Dutkiewicz and O. Shin, "Joint power control and user association for NOMA-based full-duplex systems," IEEE Transactions on Communications, vol. 67, no. 11, pp. 8037-8055, Nov. 2019.

[C6-102] W. Zhang, J. Chen, Y. Kuo and Y. Zhou, “Artificial-noise-aided optimal beamforming in layered physical layer security,” IEEE Communications Letters, vol. 23, no. 1, pp. 72-75, Jan. 2019.

[C6-103] M. Rebhi, K. Hassan, K. Raoof and P. Chargé, “Sparse code multiple access: Potentials and challenges,” IEEE Open Journal of the Communications Society, vol. 2, pp. 1205-1238, 2021

[C6-104] D. Kim, H. Lee and D. Hong, “A survey of in-band full-duplex transmission: From the perspective of PHY and MAC layers,” IEEE Communications Surveys & Tutorials, vol. 17, no. 4, 2017-2046, fourth quarter 2015.

[C6-105] Z. Yuan, Y. Ma, Y. Hu, and W. Li, High-efficiency full-duplex V2V communication, in Proc. 2nd 6G Wireless Summit, Levi, Finland, Mar. 2020.

[C6-106] J. Berkmann, C. Carbonelli, F. Dietrich, C. Drewes and W. Xu, “On 3G LTE terminal implementation – Standard, algorithms, complexities and challenges (Invited Paper),” IEEE International Wireless Communications and Mobile Computing Conference, pp. 970-975, Crete, Greece, Aug. 2008.

[C6-107] A. Elkelesh, M. Ebada, S. Cammerer and S. ten Brink, “Belief propagation list decoding of polar codes,” IEEE Communications Letters, vol. 22, no. 8, pp. 1536-1539, Aug. 2018.

[C6-108] V. Ranasinghe, N. Rajatheva and M. Latva-aho, “Partially permuted multi-trellis belief propagation for polar codes,” IEEE ICC 2020, arXiv preprint arXiv:1911.08868.

[C6-109] A. Balatsoukas-Stimming and C. Studer, “Deep unfolding for communications systems: A survey and some new directions,” arXiv preprint arXiv:1906.05774, 2019.

[C6-110] E. Nachmani, E. Marciano, L. Lugosch, W. J. Gross, D. Burshtein and Y. Be’ery, “Deep learning methods for improved decoding of linear codes,” IEEE Journal of Selected Topics in Signal Processing, vol. 12, no. 1, pp. 119-131, Feb. 2018.

[C6-111] F. R. Kschischang and S. Pasupathy, “Optimal nonuniform signaling for Gaussian channels,” IEEE Transactions on Information Theory, 39(3), pp. 913-929, 1993.

[C6-112] N. S. Loghin, J. Zöllner, B. Mouhouche, D. Ansorregui, J. Kim and S.I. Park, “Non-uniform constellations for ATSC 3.0,” IEEE Transactions on Broadcasting, 62(1), 197-203, 2016.

[C6-113] 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.

[C6-114] M. Pikus and W. Xu, “Bit-level probabilistically shaped coded modulation,” IEEE Communications Letters, vol. 21, no. 9, pp. 1929–1932, Sept. 2017.

[C6-115] T. Prinz, P. Yuan, G. Boecherer, F. Steiner, O. Iscan, R. Boehnke and W. Xu, “Polar coded probabilistic amplitude shaping for short packets,” in Proc. IEEE SPAWC, 2017.

[C6-116] P. Schulte and F. Steiner, “Shell mapping for distribution matching,” arXiv:1803.03614, 2018.

[C6-117] W. Xu, M. Huang, C. Zhu and A. Dammann, “Maximum likelihood TOA and OTDOA estimation with first arriving path detection for 3GPP LTE system,” Transactions on Emerging Telecommunications Technologies (ETT), vol. 27, no.3, pp. 339-356, 2016.

[C6-118] H. Wymeersch, G. Seco-Granados, G. Destino, et al, “5G mmWave positioning for vehicular networks,” IEEE Wireless Communications, pp. 80–86, Dec. 2017.

[C6-119] N. Garcia, H. Wymeersch, E. G. Larsson, A. M. Haimovich and M. Coulon, “Direct localization for massive MIMO,” IEEE Transactions on Signal Processing, vol. 65, no. 10, pp. 2475-2487, May 2017.

[C6-120] J. Yang, C.-K. Wen, and S. Jin, “Hybrid active and passive sensing for SLAM in wireless communication systems,” IEEE Journal on Selected Areas in Communications, vol. 40, no. 7, pp. 2146-2163, 2022.

[C6-121] H. Kim et al, “5G mmWave cooperative positioning and mapping using multi-model PHD filter and map fusion,” arXiv preprint arXiv:1908.09806, 2019.

[C6-122] M. Kiviranta, I. Moilanen, and J. Roivainen, “5G radar: Scenarios, numerology and simulations,” in IEEE International Conference on Military Communications and Information Systems (ICMCIS), 2019

[C6-123] R. Koirala et al, “Localization and throughput trade-off in a multi-user multi-carrier mm-wave system,” IEEE Access, vol. 7, pp. 167099-167112, 2019

[C6-124] C. Aydogdu, G. K. Carvajal, O. Eriksson, H. Hellsten, H. Herbertsson, M. F. Keskin, E. Nilsson, M. Rydström, K. Vanäs and H. Wymeersch, “Radar interference mitigation for automated driving,” submitted to IEEE Signal Processing Magazine, 2019.

[C6-125] J. Khoury, R. Ramanathan, D. McCloskey, R. Smith and T. Campbell, "RadarMAC: Mitigating radar interference in self-driving cars," 13th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), London, pp. 1-9, 2016.

[C6-126] A. Dammann, R. Raulefs, S. Zhang, "On prospects of positioning in 5G," in Proc. IEEE International Conference on Communication Workshop (ICCW), 2015.

[C6-127] W. Xu, A. Dammann and T. Laas, "Where are the things of the internet? Precise time of arrival estimation for IoT positioning," in 'The fifth Generation (5G) of Wireless Communication', A. Kishk, Ed., Intechopen, 2018.

[C6-128] R. di Taranto, S. Muppirisetty, R. Raulefs, D. Slock, T. Svensson and H. Wymeersch, "Location-aware communications for 5G networks: How location information can improve scalability, latency, and robustness of 5G," IEEE Signal Processing Magazine, vol. 31, no. 6, pp. 102-112, Nov. 2014.

[C6-129] Y. Polyanskiy, "A perspective on massive random-access," in Proc. IEEE International Symposium on Information Theory (ISIT), pp. 2523-2527, 2017.

[C6-130] Z. Yuan, G. Yu, W. Li, Y. Yuan, X. Wang and J. Xu, "Multi-user shared access for internet of things," IEEE 83rd Vehicular Technology Conference (VTC Spring), Nanjing, 2016.

[C6-131] Z. Yuan, Y. Hu, W. Li and J. Dai, "Blind multi-user detection for autonomous grant-free high-overloading multiple-access without reference signal," IEEE 87th Vehicular Technology Conference (VTC Spring), Porto, 2018.

[C6-132] Z. Yuan, W. Li, Y. Hu, H. Tang, J. Dai and Y. Ma, "Blind multi-user detection based on receive beamforming for autonomous grant-free high-overloading multiple access," IEEE 2nd 5G World Forum (5GWF), Dresden, Germany, pp. 520-523, 2019.

[C6-133] Z. Yuan, W. Li, Z. Li, Y. Ma and Y. Hu, "Contention-based grant-free transmission with independent multi-pilot scheme," IEEE 92nd Vehicular Technology Conference (VTC2020-Fall), Victoria, BC, Canada, 2020.

[C6-134] Z. Yuan, Z. Li, W. Li, Y. Ma and C. Liang, "Contention-based grant-free transmission with extremely sparse orthogonal pilot scheme," IEEE 94th Vehicular Technology Conference (VTC2021-Fall), 2021.

[C6-135] H.A. Inan, P. Kairouz and A. Ozgur, "Sparse combinatorial group testing for low-energy massive random access," arXiv preprint arXiv:1711.05403, 2017.

[C6-136] J. Luo and D. Guo, "Neighbor discovery in wireless ad hoc networks based on group testing," In 46th Annual Allerton Conference on Communication, Control, and Computing, pp. 791-797, Sep. 2008.

[C6-137] E. Paolini, C. Stefanovic, G. Liva and P. Popovski, "Coded random access: applying codes on graphs to design random access protocols," IEEE Communications Magazine 53, no. 6, pp. 144-150, 2015.

[C6-138] N. H. Mahmood, R. Abreu, R. Böhnke, M. Schubert, G. Berardinelli and T. H. Jacobsen, "Uplink grant-free access solutions for URLLC services in 5G New Radio," 16th International Symposium on Wireless Communication Systems (ISWCS), Oulu, Finland, pp. 607-612, 2019.

[C6-139] F. Clazzer, A. Munari, G. Liva, F. Lazaro, C. Stefanovic and P. Popovski, "From 5G to 6G: Has the time for modern random access come?" arXiv: 1903.03063, 2019.

[C6-140] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in Proc. IEEE International Symposium on Information Theory (ISIT), pp. 2168-2172, 2011.

[C6-141] Z. Chen, F. Sohrabi and W. Yu, "Sparse activity detection for massive connectivity," arXiv preprint arXiv:1801.05873, 2018.

[C6-142] S. Haghighatshoar, P. Jung and G. Caire, "Improved scaling law for activity detection in massive MIMO systems," arXiv preprint arXiv:1803.02288, 2018.

[C6-143] C. Bockelmann et al, "Massive machine-type communications in 5G: Physical and MAC-layer solutions," IEEE Communications Magazine, vol. 54, no. 9, pp. 59-65, September 2016.

[C6-144] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic and E. de Carvalho, "Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the internet of things," IEEE Signal Processing Magazine, vol. 35, no. 5, pp. 88-99, Sept. 2018.

[C6-145] E. Bjornson, E. de Carvalho, J. H. Sørensen, E. G. Larsson and P. Popovski, "A random access protocol for pilot allocation in crowded massive MIMO systems," IEEE Transactions on Wireless Communications, vol. 16, no. 4, pp. 2220–2234, April 2017.

[C6-146] G. Fettweis, H. Boche, et al, “The tactile internet,” ITU-T Technology Watch Report, August 2014.

[C6-147] H. Lasi, P. Fettke, H.-G. Kemper, et al, “Industry 4.0”, Bus Inf Syst Eng 6, pp. 239–242, 2014.

[C6-148] C. Kalalas and J. Alonso-Zarate, “Massive connectivity in 5G and beyond: Technical enablers for the energy and automotive verticals,” in Proc. of 6G Wireless Summit 2020, Levi, Finland, March 2020.

[C6-149] S. Ali, W. Saad & D. Steinbach (Eds.), “White paper on machine learning in 6G wireless communication networks,” University of Oulu, http://urn.fi/urn:isbn:9789526226736, 2020.

[C6-150] H. Ye, G. Y. Li, and B. Juang, “Power of deep learning for channel estimation and signal detection in OFDM systems,” IEEE Wireless Communications Letters, vol. 7, no. 1, pp. 114–17, 2018.

[C6-151] T. OShea and J. Hoydis, “An introduction to deep learning for the physical layer,” IEEE Transactions on Cognitive Communication Networks, vol. 3, no. 4, pp. 563–75, 2017.

[C6-152] A. Giannopoulos, S. Spantideas, N. Kapsalis, P. Karkazis and P. Trakadas, “Deep reinforcement learning for energy-efficient multi-channel transmissions in 5G cognitive hetnets: Centralized, decentralized and transfer learning based solutions,” IEEE Access, vol. 9, pp. 129358-129374, 2021.

[C6-153] K. Yang, T. Jiang, Y. Shi and Z. Ding, “Federated learning via over-the-air computation,” IEEE Transactions on Wireless Communications, vol. 19, no. 3, pp. 2022-2035, March 2020.

[C6-154] S. Li and S. Avestimehr, “Coded computing: Mitigating fundamental bottlenecks in large-scale distributed computing and machine learning,” Foundations and Trends in Communications and Information Theory, vol. 17, no. 1, pp 1-148, 2020.

[C6-155] J. D.V. Sánchez, L. Urquiza-Aguiar, M.C.P. Paredes and D.P.M. Osorio, “Survey on physical layer security for 5G wireless networks,” Annals of Telecommunications, 76 (3), 155-174, 2021.

[C7-01] Photonics21, Vision Paper “Europe’s age of light! How photonics will power growth and innovation”, Nov 2017, online

[C7-02] P. J. Winzer and D. T. Neilson, “From Scaling Disparities to Integrated Parallelism: A Decathlon for a Decade,” Journal of Lightwave Technol., vol. 35, 5, pp. 1099 – 1115, 2017.

[C8-01] M. Höyhtyä, M. Corici, S. Covaci and M. Guta, "5G and beyond for new space: Vision and research challenges," ICSSC-2019, pp. 1-16, doi: 10.1049/cp.2019.1236.

[C8-2] European Vision for the 6G Network Ecosystem, 5GIA; https://5g-ppp.eu/wp-content/uploads/2021/06/WhitePaper-6G-Europe.pdf

[C8-3] 6G Networks for Next Generation of Digital TV Beyond 2030, Paulo Sergio Rufino Henrique & Ramjee Prasad https://link.springer.com/article/10.1007/s11277-021-09070-2

[C8-4] A review on 6G for space-air-ground integrated network: Key enablers, open challenges, and future direction, Partha Pratim Ray; https://www.sciencedirect.com/science/article/pii/S1319157821002172

[C8-5] C. Gidney and M. Ekera, “How to factor 2048-bit RSA integers in 8 hours using 20 million noisy qubits”, Quantum 5, 433 (2021), 1905.09749.

[C8-6] E. Gouzien and N. Sangouard, “Factoring 2048-bit RSA Integers in 177 Days with 13436 Qubits and a Multimode Memory”, Phys. Rev. Lett. 127, 140503, 28 September 2021, doi=10.1103/PhysRevLett.127.140503

[C8-7] Akyildiz, Ian F., and Ahan Kak. "The internet of space things/cubesats." IEEE Network 33.5 (2019): 212-218.

[C8-8] Al-Hraishawi, Hayder, et al. "Multi-Layer Space Information Networks: Access Design and Softwarization." *IEEE Access* 9 (2021): 158587-158598.

[C8-9] S. R. Pokhrel, "Blockchain Brings Trust to Collaborative Drones and LEO Satellites: An Intelligent Decentralized Learning in the Space," in *IEEE Sensors Journal*, vol. 21, no. 22, pp. 25331-25339, 15 Nov.15, 2021, doi: 10.1109/JSEN.2021.3060185.

[C8-10] Distributed Denial of Service (DDoS), ENISA threat Landscape, 2020.

[C8-11] S. Zeba; M. Amjad; D. R. Rizvi, "Sustainable Paradigm for Computing the Security of Wireless Internet of Things: Blockchain Technology," in *Smart and Sustainable Approaches for Optimizing Performance of Wireless Networks: Real-time Applications*, Wiley, 2022, pp.51-66, doi: 10.1002/9781119682554.ch3.

[C8-12] J. et al, “Satellite-based entanglement distribution over 1200 kilometers”, In Proc. of Science, vol.356,nº6343, pp1140-1144, June 2017, doi: 10.1126/science.aan3211.

[C8-13] C. J. Pugh *et al*., "Airborne demonstration of a quantum key distribution receiver payload," *2017 Conference on Lasers and Electro-Optics Europe & European Quantum Electronics Conference (CLEO/Europe-EQEC)*, 2017, pp. 1-1, doi: 10.1109/CLEOE-EQEC.2017.8087396.

[C8-14] A. D. Wyner, "The wire-tap channel," in *The Bell System Technical Journal*, vol. 54, no. 8, pp. 1355-1387, Oct. 1975, doi: 10.1002/j.1538-7305.1975.tb02040.x.

[C8-15] Y. Ismail, I. Sinayskiy, and F. Petruccione,"Integrating machine learning techniques in quantum communication to characterize the quantum channel", In Journal of the Optical Society of America B, vol.36,nº3, March 2019.

[C8-16] J. Wallnöfer, A.A. Melnikov, W. Dür, and H.J. Briegel, "Machine learning for long-distance quantum communications", In proc. PRX Quantum, vol.1, no1, Sept.2020, doi: 10.1103/PRXQuantum.1.010301.

[C8-17] ITU-T Y.3800-series-Quantum Key Distribution networks- Applications of machine learning, 7/2021.

[C8-18] N.Y.Ahn, D. H. Lee, "Physical Layer Security in Autonomous Driving Based Non-Competitive Distributed Consensus Scheme", In Proc. Research Square, May 2021, https://doi.org/10.21203/rs.3.rs-515048/v1

[C8-19] Ali, M. S., Vecchio, M., Pincheira, M., Dolui, K., Antonelli, F., & Rehmani, M. H. (2018). Applications of Blockchains in the Internet of Things: A Comprehensive Survey. IEEE Communications Surveys & Tutorials, 1–1.

[C8-20] M. De Sanctis, E. Cianca, G. Araniti, I. Bisio and R. Prasad, "Satellite Communications Supporting Internet of Remote Things," in IEEE Internet of Things Journal, vol. 3, no. 1, pp. 113-123, Feb. 2016.

[C8-21] Z. Qu, G. Zhang, H. Cao and J. Xie, "LEO Satellite Constellation for Internet of Things," in IEEE Access, vol. 5, pp. 18391-18401, 2017.

[C9-01] T. S. Rappaport, Y. Xing, O. Kanhere, S. Ju, A. Madanayake, S. Mandal, A. Alkhateeb, and G. C. Trichopoulos, "Wireless Communications and Applications Above 100 GHz: Opportunities and Challenges for 6G and Beyond," IEEE Access, vol. 7, pp. 78729–78757, 2019.

[C9-02] J.-B. Dore, Y. Corre, S. Bicais, J. Palicot, E. Faussurier, D. Kt´enas,´ and F. Bader, "Above-90GHz Spectrum and Single-Carrier Waveform as Enablers for Efficient Tbit/s Wireless Communications," in 25th International Conference on Telecommunications (ICT'2018), SaintMalo, France, Jun. 2018.

[C9-03] S. Bicaïs and J.-B. Dore, "Phase Noise Model Selection for Sub-THz Communications," in 2019 IEEE Global Communication Conference (GLOBECOM), December 2019.

[C9-04] S. Bicaïs, J.-B. Dore, G. Gougeon, and Y. Corre, "Optimized Single Carrier Transceiver for Future Sub-TeraHertz Applications", in International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2020.

[C9-05] A. Visweswaran, K. Vaesen, S. Sinha, I. Ocket, M. Glassee, C. Desset, A. Bourdoux, and P. Wambacq, "A 145GHz FMCW Radar Transceiver in 28nm CMOS", in 2019 IEEE International Solid- State Circuits Conference - (ISSCC).

[C9-06] NEREID Roadmap: https://www.nereid-h2020.eu/roadmap

[C9-07] PA survey: https://gems.ece.gatech.edu/PA_survey.html

[C9-08] A. Vais et al., "First demonstration of III-V HBTs on 300 mm Si substrates using nano-ridge engineering », in 2019 IEEE International Electron Devices Meeting (IEDM).

[C9-09] M. Schroter et al., "Physical and Electrical Performance Limits of High-Speed SiGeC HBTs—Part I: Vertical Scaling", IEEE Trans. Electron Devices, Vol. 58, No. 11, Nov. 2011.

[C9-10] U. Peralagu, "CMOS-compatible GaN-based devices on 200mm-Si for RF applications: Integration and Performance", in 2019 IEEE International Electron Devices Meeting (IEDM).

[C9-11] P. Weckx et al., "Novel forksheet device architecture as ultimate logic scaling device towards 2nm", in 2019 IEEE International Electron Devices Meeting (IEDM).

[C9-12] Simoens F., Meilhan J., Nicolas J.-A., "Terahertz Real-Time Imaging Uncooled Arrays Based on Antenna-Coupled Bolometers or FET Developed at CEA-LETI", J. Infrared, Millimeter, Terahertz Waves. 2015; 36:961–985. doi: 10.1007/s10762-015-0197-x

[C9-13] P. Huang, P. Mercier, “A 220µW -85dBm Sensitivity BLE-Compliant Wake-up Receiver Achieving -60dB SIR via Single-Die Multi- Channel FBAR-Based Filtering and a 4-Dimentional Wake-Up Signature,” IEEE ISSCC, 2019.

[C9-14] S. Denis, R.Berkvens, M.Weyn, “A Survey on Detection, Tracking and Identification in Radio Frequency-Based Device-Free Localization”, Special Issue Surveys of Sensor Networks and Sensor Systems Deployments, December 2019

[C9-15] P. Zand, J. Romme, J. Govers, F. Pasveer, and G. Dolmans, “A High-Accuracy Phase-Based Ranging Solution with Bluetooth Low Energy (BLE),” in WCNC, 2019

[C9-16] Boer, J. Romme, J. Govers, and G. Dolmans, “Performance of High-Accuracy Phase-Based Ranging in Multipath Environments,” in 91th IEEE Vehicular Technology Conference, VTC Spring 2020, Antwerp, Belgium, May 25-28, 2020.

[C9-17] D. Vasisht, S.Kumar, D.Katabi, ”Decimeter-level localization with a single Wi-Fi access point”, NSDI'16: Proceedings of the 13th Usenix Conference on Networked Systems Design and Implementation, March 2016, Pages 165–178

[C9-18] E. Bechthum, J. Dijkhuis, M. Ding, Y. He, J. Van den Heuvel, P. Mateman, G-J. van Schaik, K. Shibata, M. Song, E. Tiurin, S. Traferro, Y-H. Liu, C. Bachmann,”A Low-Power BLE Transceiver with Support for Phase-Based Ranging, Featuring 5µs PLL Locking Time and 5.3ms Ranging Time, Enabled by Staircase-Chirp PLL with Sticky-Lock Channel-Switching” IEEE INTERNATIONAL SOLID-STATE CIRCUITS CONFERENCE, ISSCC 2020, San Francisco

[C9-19] Higginbotham, Stacey. "The internet of trash [Internet of Everything]." IEEE Spectrum 55.6 (2018): 17-17.

[C9-20] https://www.storaenso.com/en/newsroom/regulatory-and-investor-releases/2018/11/stora-enso-introduces-sustainable-rfid-tag-technology-eco-for-intelligent-packaging[C11-1] D. D. Clark, C. Partridge, J. C. Ramming, J. T. Wroclawski: “A knowledge plane for the internet”. ACM SIGCOMM 2003 conference.

[C9-21] R. Al Hadi, H. Sherry, J. Grzyb, Y. Zhao, W. Forster, H. M. Keller, A. Cathelin, A. Kaiser, and U. R. Pfeiffer. “A 1 k-pixel video camera for 0.7-1.1 terahertz imaging applications in 65-nm CMOS”. In: IEEE J. Solid-State Circuits 47.12 (Dec. 2012), pp. 2999–3012.

[C9-22] Zheludev, N. I.; Kivshar, Y. S. From metamaterials to metadevices. Nat. Mater. 2012, 11, 917–924.

[C9-23] Zhang, L.; Chen, X. Q.; Liu, S.; Zhang, Q.; Zhao, J.; Dai, J. Y.; Galdi, V. Space-time-coding digital metasurfaces. Nature communications 2018, 9(1), 4334.

[C9-24] M. Jian, Y. Zhao, “A modified off-grid SBL channel estimation and transmission strategy for RIS-assisted wireless communication systems,” IWCMC, 2020

[C9-25] Silva, A.; Monticone, F.; Castaldi, G.; Galdi, V.; Alù, A. & Engheta, N. Performing mathematical operations with metamaterials. Science 2014, 343(6167), 160-163.

[C9-26] C.Y. Wu et.al., “Distributed antenna system using sigma-delta intermediate frequency over fibre for frequency bands above 24GHz”, Journal of Lightwave Technology, Feb. 2020.

[C9-27] A. Gatherer et al, Academic Press Library in Mobile and Wireless Communications, 2016

[C9-28] T. J. O’Shea, J. Hoydis An Introduction to Deep Learning for the Physical Layer, IEEE Transactions on Cognitive Communications and Networking, 2017, pp 563-575.

[C9-29] D. Wu, M. Nekovee, Y. Wang, An Adaptive Deep Learning Algorithm Based Autoencoder for Interference Channel, Proc Second IFIP International Conference on Machine Learning for Networks, MLN2019, Paris December 2019.

[C9-30] E. Grimaldi, V. Krizakova, G. Sala, F. Yasin, S. Couet, G. Kar, K. Garello, P. Gambardella, “Single shot dynamics of spin-orbit torque and spin transfer torque switching in 3-terminal magnetic tunnel junctions”, Nature Nanotechnology 15, 111 (2020)

[C9-31] S. Van Beek, B. O’Sullivan, P.J. Roussel, R. Degraeve, E. Bury, J. Swerts, S. Couet, L. Souriau, S. Kundu, S. Rao, W. Kim, F. Yasin, D. Crotti, D. Linten, G. Kar, “Impact of self-heating on reliability predictions in STT-MRAM”, IEDM 2018

[C9-32] K. Garello, F. Yasin, S. Couet, L. Souriau, J. Swerts, S. Rao, S. Van Beek, W. Kim, E. Liu, S. Kundu, D. Tsvetanova, K. Croes, N. Jossart, E. Grimaldi, M. Baumgartner, D. Crotti, A. Furnémont, P. Gambardella, G.S. Kar, "SOT-MRAM 300mm integration for lo power and ultrafast embedded memories", VLSI 2018

[C9-33] Y. C. Wu, W. Kim, K. Garello, F. Yasin, G. Jayakumar, S. Couet, R. Carpenter, S. Kundu, S. Rao, D. Crotti, J. Van Houdt, G. Groeseneken, G. S. Kar, "Deterministic and field-free voltage-controlled MRAM", VLSI 2020

[C9-34] R. Delhougne, A. Arreghini, E. Rosseel, A. Hikavyy, E. Vecchio, L. Zhang, M. Pak, L. Nyns, T. Raymaekers, N. Jossart, L. Breuil, S. S. V-Palayam, C.-L. Tan, G. Van den bosch, A. Furnémont, "First demonstration of monocrystalline silicon macaroni channel for 3-D NAND memory devices", Proc. VLSI Technology Symposium, p. 203 (2018)

[C9-35] D. Verreck, A. Arreghini, J.P. Bastos, F. Schanovsky, F. Mitterbauer, C. Kernstock, M. Karner, R. Degraeve, G. Van den bosch and A. Furnémont, "Quantitative 3-D Model to Explain Large Single Trap Charge Variability in Vertical NAND Memory", 2019 International Electron Device Meeting (IEDM) Tech. Dig., p. 755 (2019)

[C9-36] A. H. Du Nguyen, Y. Yu, M. Abu Lebdeh, M. Taouil, S. Hamdioui and F. Catthoor, "Classification of Memory-Centric Computing," ACM Emerging Technologies in Computing, vol. 16, no. 2, pp. 1-26, 2020.

[C9-37] K. Kim, S. Shin and S. Kang, "Stateful logic pipeline architecture," IEEE International Symposium of Circuits and Systems (ISCAS), pp. 2497-2500, 2011.

[C9-38] L. Xie, A. H. Du Nguyen, J. Yu, A. Kaichouhi, M. Taouil, M. Al-Failakawi and S. Hamdiou, "Scouting logic: A novel memristor-based logic design for resistive computing," IEEE Computer Society Annual Symposium on VLSI (ISVLSI), pp. 335-340, 2017.

[C9-39] NetWorld2020 ETP, "5G: Challenges, research priorities, and recommendations", White Paper, September 2014. Available online at: https://www.networld2020.eu/.

[C9-40] A. Paverd, M. Völp, F. Brasser, M. Schunter, A.-R. Sadeghi, Asokan, P. Verissimo, A. Steiniger and T. Holz, "Sustainable Security & Safety: Challenges and Opportunities.," in CERTS 2019; pp. 4:1-4:13, 2019.

[C9-41] Mirai botnet code. Available online at: https://github.com/jgamblin/Mirai-Source-Code.

[C9-42] J. M. Rabaey "The swarm at the edge of the cloud - A new perspective on wireless", Symposium on VLSI Circuits - Digest of Technical Papers, IEEE, pp. 6-8, 2011.

[C9-43] E. A. Lee, J. Rabaey, B. Hartmann, J. Kubiatowicz, K. Pister, A. Sangiovanni-Vincentelli, S. A. Seshia, J. Wawrzynek, D. Wessel , "The swarm at the edge of the cloud" IEEE Des. Test., vol. 31, no. 3, pp. 8–20, Jun. 2014.

[C9-44] G. Fedrecheski, L. Caroline, C. De Biase, P. C. Calcina-Ccori, M. Knorich Zuffo, "Attribute-Based Access Control for the Swarm With Distributed Policy Management", Volume: 65, Issue: 1, IEEE Transactions on Consumer Electronics, IEEE, 2019.

[C10-1] Kozlowski, Wojciech and Stephanie Wehner. "Towards Large-Scale Quantum Networks." Proceedings of the Sixth Annual ACM International Conference on Nanoscale Computing and Communication (2019).

[C10-2] Wei, Shihai et al. "Towards Real-World Quantum Networks: A Review." Laser & Photonics Reviews 16 (2022).

[C10-3] J. Biamonte, P. Wittek et al., "Quantum machine learning," Nature, vol. 549, no. 7671, p. 195–202, 2017, doi: 10.1038/nature23474.

[C10-4] V. Dunjko and H. J. Briegel, "Machine learning & Artificial Intelligence in the quantum domain: a review of recent progress," Reports on Progress in Physics, vol. 81, no. 7, p. 074001, jun 2018, doi: 10.1088/1361-6633/AAB406.

[C10-5] F. Phillipson, "Quantum Machine Learning: Benefits and Practical Examples," in QANSWER, 2020, pp. 51–56.

[C10-6] H.-Y. Huang, M. Broughton, M. Mohseni, R. Babbush, S. Boixo, H. Neven, and J. R. McClean, "Power of data in quantum machine learning," Nature Communications 2021 12:1, vol. 12, no. 1, pp. 1–9, may 2021, doi: 10.1038/s41467-021-22539-9.

[C10-7] H. I. G. Hern´andez, R. T. Ruiz, and G.-H. Sun, “Image Classification via Quantum Machine Learning,” arXiv:2011.02831, 2020.

[C10-8] P. Rebentrost, M. Mohseni, and S. Lloyd, “Quantum Support Vector Machine for Big Data Classification,” Physical Review Letters, vol. 113, no. 13, Sep 2014, doi: 10.1103/physrevlett.113.130503.

[C10-9] M. Henderson, S. Shakya, S. Pradhan, and T. Cook, “Quanvolutional Neural Networks: Powering Image Recognition with Quantum Circuits,” arXiv:1904.04767, 2019.

[C10-10] A. I. Hasan, “IGO-QNN: Quantum Neural Network Architecture for Inductive Grover Oracularization,” arXiv:2105.11603, 2021.

[C10-11] I. Cong, S. Choi et al., “Quantum convolutional neural networks,” Nature Physics, vol. 15, no. 12, p. 1273–1278, 2019, doi: 10.1038/s41567-019-0648-8.

[C10-12] A. Khoshaman, W. Vinci, B. Denis, E. Andriyash, H. Sadeghi, and M. H. Amin, “Quantum variational autoencoder,” Quantum Science and Technology, vol. 4, no. 1, p. 014001, sep 2018, doi: 10.1088/2058-9565/AADA1F.

[C10-13] N. Liu, T. Huang, J. Gao, Z. Xu, D. Wang, and F. Li, “Quantum-Enhanced Deep Learning-Based Lithology Interpretation From Well Logs,” IEEE Transactions on Geoscience and Remote Sensing, 2021.

[C10-14] J. Liu, K. H. Lim, K. L. Wood, W. Huang, C. Guo, and H.-L. Huang, “Hybrid quantum-classical convolutional neural networks,” Science China Physics, Mechanics and Astronomy 2021 64:9, vol. 64, no. 9, pp. 1–8, aug 2021, doi: 10.1007/S11433-021-1734-3.

[C10-15] S. Oh, J. Choi, and J. Kim, “A Tutorial on Quantum Convolutional Neural Networks (QCNN),” arXiv:2009.09423, vol. 2020-October, pp. 236–239, sep 2020.

[C10-16] K. Bharti, A. Cervera-Lierta, T. H. Kyaw, T. Haug, S. Alperin-Lea, A. Anand, M. Degroote, H. Heimonen, J. S. Kottmann, T. Menke, W.-K. Mok, S. Sim, L.-C. Kwek, and A. Aspuru-Guzik, “Noisy intermediate-scale quantum (NISQ) algorithms,” arXiv:2101.08448, 2021.

[C10-17] V. Ziegler, P. Schneider, H. Viswanathan, M. Montag, S. Kanugovi and A. Rezaki, “Security and Trust in the 6G Era,” in IEEE Access, vol. 9, pp. 142314-142327, 2021, doi: 10.1109/ACCESS.2021.3120143.

[C10-18] Jon Karafin, Light Field Lab Inc. online https://mpeg.chiariglione.org/sites/default/files/events/08_KARAFIN_LightFieldLab_MPEGWorkshopLB_v01.pdf

[C10-19] Rahim Tafazolli, online https://www.itu.int/en/ITU-T/Workshops-and-Seminars/20190218/Documents/Rahim_Tafazolli_Presentation.pd

[C10-20] Flavio Meneses, Daniel Corujo, Rui Aguiar, “Virtualization of Customer Equipment: Challenges and Opportunities”, ITL internet technology letters Vol. 3, Nº 6, May, 2020f

[C10-21] https://www.spirent.com/assets/wp_simplifying-5g-with-the-network-digital-twin

[C10-22] https://5g-acia.org/whitepapers/using-digital-twins-to-integrate-5g-into-production-networks/

[C10-23] Richard Feynmann, “There is plenty of room at the bottom”, Pasadena, Dec. 1959, https://web.archive.org/web/20170105015142/http://www.its.caltech.edu/~feynman/plenty.html

[C10-24] Ericsson blog - Zero-energy devices – a new opportunity in 6G, retrieved June 2022, https://www.ericsson.com/en/blog/2021/9/zero-energy-devices-opportunity-6g

[C10-25] Brian Rogers, et. Al. “Scaling the Bandwidth Wall: Challenges in and Avenues for Chip Multi-Processor Scaling,” ISCA’09, June 20–24, 2009.

# 12.Contributors

## Editors

**Overall edition** –
Ari Pouttu (University of Oulu),
Jyrky Huusko (VTT),
Rui L. Aguiar (Instituto de Telecomunicações)

**System chapter -** Holger Karl, Hasso-Plattner Institute
**Architecture chapter -** Artur Hecker, Huawei
**Security chapter -** Emmanuel Dotaro, Thales
**Software chapter -** Josef Urban, Nokia
**Radio chapter -** Wen Xu, Huawei Germany
**Optical chapter -** Raul Munõz, CTTC
**NTN chapter -** Alessandro Vanelli-Coralli, University of Bologne
**Devices chapter -** André Bourdoux, IMEC
**FET chapter -** Anastasius Gavras, Eurescom

**Contributors**

Aarno Pärssinen, Oulu Univ.
Agapi Mesodiakaki, Aristotle Unviersity of Thessaloniki
Akis Kourtis, Demokritos
Albert Rafel, BT
Alberto Gotta, CNR
Alessandro Carrega, CNIT
Alessandro Guidotti, University of Bologne
Alessandro Sebastianelli, Università degli Studi del Sannio
Alessandro Ugolini, Università degli studi di Parma
Alexander Hofmann, IIS - Fraunhofer-Institut für Integrierte Schaltungen
Alexandre Petrescu, CEA
Alexandre Valentian, CEA-Leti
Alexandros Stavdas, University of Peloponnese.
Amina Piemontese, Università degli studi di Parma
Ana G. Armada,,Universidad Carlos III de Madrid
Andre Bourdoux,,IMEC
Andrea Giorgetti,,University of Bologna
Andrea Passarella, CNR
Andreas Knopp, Universität der Bundeswehr München
Andrés Meseguer, ITI
Andrew Lord, BT
Antonio de la Oliva, IMDEA networks
Antonio Manzalini, Telecom Italia
Antonio Napoli, Infinera
Antonio Skarmeta, Universidad de Murcia
Artur Hecker, Huawei
Avi Gal, Gilat
Barry Evans, University of Surrey
Beatriz Soret, Aalborg Universitet
Bengt Holter,,SINTEF
Benjamin Wohlfeil, ADVA Optical Networking
Bjørn Skjellaug, Sintef
Carla Amatetti, University of Bologne
Carlos Mosquera, Universidad de Vigo
Christian Hofmann, Universität der Bundeswehr München
Claudio Cicconetti, CNR
Colja Schubert, Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institute.
Constantinos Papadias, The American College of Greece Research Center
Damir Filipovic, AIOTI
Daniel Gaetano Riviello, University of Bologne
Daniele Tarchi, University of Bologne
David Hillerkuss, Huawei
Despoina Petousi, ADVA Optical Networking
Didier Belot, CEA-Leti
Didier Bourse, Nokia
Dimitra Simeonidou, University of Bristol,
Dirk Trossen, Huawei
Elisa Rojas, Universidad de Alcalá
Emmanuel Dotaro, Thales

Enrico Del Re, UNIFI
Ernestina Cianca, University of Rome
Fabio Cavaliere, Ericsson
Fabio Patrone, University of Genoa
Filippo Cugini, CNIT
Francky Catthoor, IMEC
Franco Davoli, University of Genoa
Frank Schaich,,Nokia
Fredrik Dahlgren, Ericsson
George Kalfas, Aristotle University of Thessaloniki
Georgios Gardikis, Sapce Hellas
Georgios Karagiannis, Huawei
Gerald Karam, NOKIA Bell Labs
Gerhard Fettweis,TU Dresden
Giovanni Frattini, Engineering
Giulio Colavolpe, Università degli studi di Parma
Giuseppe Caire,TU Berlin
Gouri Kar, IMEC
Gunnar Mildh, Ericsson
Harald Haas,University of Strathclyde
Helmut Griesser, ADVA Network Security
Henk Wymeersch,,Chalmers University of Technology
Hervé Debar, Telecom Sud Paris
Holger Karl, Hasso-Plattner Institute
Hugo Tullberg, Ericsson
Hui Song, Sintef
Ian F. Akyldiz,Georgia Institute of Technology
Ijaz Ahmad, VTT
Ioannis Tomkos, University of Patras
Jan Craninckx, IMEC
Jari Arkko, Ericsson
Jeroen Wigard, Nokia Bell Labs
Joan Bas, CTTC
Joerg Widmer,IMDEA Networks
Johannes Fischer, Fraunhofer, HHI
Jörg-Peter Elbers, ADVA Optical Networking
Jose Capmany, iPronics
Jose F. Monserrat,,UPV
José Vera, ITI
Josef Urban, Nokia
Josep M. Jornet,Northeastern University
Jyrki Huusko, VTT
Konstantinos Ntontin, University of Luxemburg
Lorenzo Favalli, University of Pavia
Luis Blanco, CTTC
Luis Perez-Freire, Gradiant
Luis Velasco, UPC
Marco Ruffini, TCD
Marios Gkatzianas, Aristotle University of Thessaloniki
Marius Caus, CCTC
Marko Höyhtyä, VTT
Martin Schell, Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institute
Mats Eriksson, Arctos Labs
Mauro De Sanctis, University of Rome
Maziar Nekovee, Univ. of Sussex
Michael Eiselt, ADVA Network Security
Michael Montag, Nokia
Michele Luglio, University of Rome
Miguel Ángel Vázquez, CTTC
Mohand Achouche, Nokia
Mona Ghassemian, King' College London
Musbah Shaat, CTTC
Nandana Rajatheva,,University of Oulu
Natalie Samovich, Enercoutim
Nicolas Chuberre, Thales Alenia Aerospace
Nikolaos Bartzoudis,,CTTC
Nikos Pleros, Aristotle University of Thessaloniki
Oriol Vidal, Airbus
Orlane Bergogne, Airbus
Ovidiu Vermesan, Sintef
Pascal Bisson, Thales
Paul van Dijk, Lionix International
Paulo Jorge Mendes, Airbus
Peter Ossieur, IMEC
Petros S. Bithas,,National and Kapodistrian University of Athens
Philippe Boutry, Airbus
Piet Wambacq, IMEC
Pietro Savazzi,,University of Pavia
Pol Henarejos, CTTC
Pouria Khodashenas, Huawei
Raffaele Bolla, University of Genoa
Raffaele Bruno, CNR
Rainer Wansch, IIS - Fraunhofer-Institut für Integrierte Schaltungen
Ramon Casellas, CTTC
Ravi Kuchibhotla, Motorola
Raymond Knopp,,EURECOM
Razvan Andrei Stoica, Lenovo
Riccardo Campana, University of Bologne
Roberto Bruschi, University of Genoa
Roberto Cascella, ECSO
Rui Dinis,Instituto de Telecomunicações (IT) / Nova University of Lisbon
Rui L. Aguiar, Instituto de Telecomunicações/DETI, Universidade de Aveiro
Rute Sofia, Fortiss
Said Hamdioui, TU Delft
Sebastian, Robitzsch, InterDigital
Sébastien Bigo, Nokia Bell Labs

Silvia Liberata Ullo, Università degli Studi del Sannio
Simon Watts, Avanti
Steven Kisseleff, University of Luxemburg
Thomas Delamotte, Universität der Bundeswehr München
Thomas Heyn, IIS - Fraunhofer-Institut für Integrierte Schaltungen
Thomas Pfeiffer, Nokia Bell Labs Germany
Tim Hentschel,Barkhausen Institut
Tolga Tekin, Fraunhofer, IZM
Tomaso Decola, DRL
Tommaso Foggi, Università degli studi di Parma
Tommy Svensson,,Chalmers University of Technology
Ullrich Pfeiffer, Wuppertal Univ.
Valerio Frascolla, Intel
Vermesan Ovidiu, SINTEF
Vittorio Curri, Politecnico di Torino
Wallace Alves Martins, University of Luxemburg
Xavier Artiga, CTTC
Yao-Hong Liu, IMEC
Yvan Pointurier, Huawei
Yvette Koza,ZTE

https://www.networldeurope.eu/