End-to-end network slicing orchestration

– A KEY ENABLER FOR INDUSTRY-VERTICAL USE CASES

By automatically matching the particular service requirements of an industry-vertical use case to its specific deployment areas, transport-aware network slicing orchestration makes it possible to ensure end-to-end QoS without over-provisioning.

PAOLA IOVANNA,
MALGORZATA SVENSSON,
ALEXEY SHAPIN,
GIULIO BOTTARI,
FABIO UBALDI,
FILIPPO PONZINI,
MARZIO PULERI

To support use cases with extreme performance requirements in a cost-efficient manner, communication service providers need comprehensive, end-to-end (E2E) network slicing orchestration that covers the full range of network resources – RAN, core, cloud and transport.

■ The acceleration of industrial digitalization and the rise of new applications such as cloud robotics and remote-assisted surgery are leading to a high demand for the new capabilities available in 5G. At the same time, future use cases like massive twinning, immersive telepresence and collaborative robotics are helping to shape the journey toward 6G. There is also a desire to achieve ubiquitous radio access in specific deployment areas, supporting a high density of users and access points to the network. The goal is to achieve all of this without increasing the cost/revenue ratio.

Based on Service Level Agreements (SLAs), new services for industry verticals [1] – including those that demand extreme performance – will require E2E QoS at scale over three main network deployment areas: local, confided wide and general wide. Examples of the local area include indoor locations and campuses, while examples of confined wide areas include ports and railway systems. General wide area refers to larger geographical locations such as cities.

E2E QoS depends on factors such as throughput, latency, availability, reliability and resilience. To ensure service delivery with the required E2E QoS without resorting to over-provisioning of network resources, the corresponding RAN, core and transport resources must be made available in the specific deployment areas. The network must have the ability to support a mix of heterogeneous services in the corresponding deployment areas. Our research shows that by including transport awareness in orchestration and slice operations, it is possible to tailor the network resources to the actual QoS values.

Our end-to-end slice orchestration concept

In a 5G network, there is always at least one default slice and each UE (user equipment) is associated to a slice. The 5G network slicing feature makes it possible to set up independent logical networks on a shared physical and virtual infrastructure. A slice can, for example, ensure ultra-reliable low-latency communication (URLLC) to support a service related to the remote control of robots in a factory. Each slice operates on specific tracking areas (TAs) served by a set of gNodeB base stations along with the Access and Mobility Management Function. This means that each network function can be placed in accordance with both the area and the service conveyed by the related slice.

An industry-vertical service that spans a specific deployment area can be mapped on a slice that includes multiple TAs. The service is characterized by specific E2E QoS requirements that the slice must support within the deployment area. The E2E QoS is defined by the combination of the QoS in the radio layer (RAN/Core Network (CN)) and the QoS in the transport layer. An optimal combination can be achieved by using automatic and dynamic techniques to create smart mapping between the RAN/CN QoS and the transport QoS that takes into account the specific technology of the infrastructure. For example, the 5G QoS Identifier of the RAN could be mapped to the corresponding Differentiated Services Code Point of the transport.

> “E2E QoS DEPENDS ON FACTORS SUCH AS THROUGHPUT, LATENCY, AVAILABILITY, RELIABILITY AND RESILIENCE”

The transport connections are implemented according to the specific transport data-plane technologies such as IP, VPN and VLAN. Because a particular service is available in a specific deployment area, there can be multiple transport connections for it. As a result, it is essential that there is a procedure during service provisioning that automatically maps the E2E QoS parameters (peak rate, guarantee rate, resilience, availability and so on) of the slice on all the available transport connections. A requirement that all the transport connections support the slice peak rate would result in a waste of resources, while splitting the slice peak rate among all the transport connections could adversely affect the service level. An effective solution assigns the

Terms and abbreviations

AC – Admission Control | **AI** – Artificial Intelligence | **AR** – Augmented Reality | **CIR** – Committed Information Rate | **CN** – Core Network | **CNF** – Container Network Function | **E2E** – End-to-End | **eMBB** – Enhanced Mobile Broadband | **HSS** – Home Subscriber Server | **IoT** – Internet of Things | **mMTC** – Massive Machine-Type Communication | **PIR** – Peak Information Rate | **PoP** – Point of Presence | **SLA** – Service Level Agreement | **TA** – Tracking Areas | **URLLC** – Ultra-Reliable Low-Latency Communication | **vEPC** – Virtual Evolved Packet Core | **VNF** – Virtual Network Function

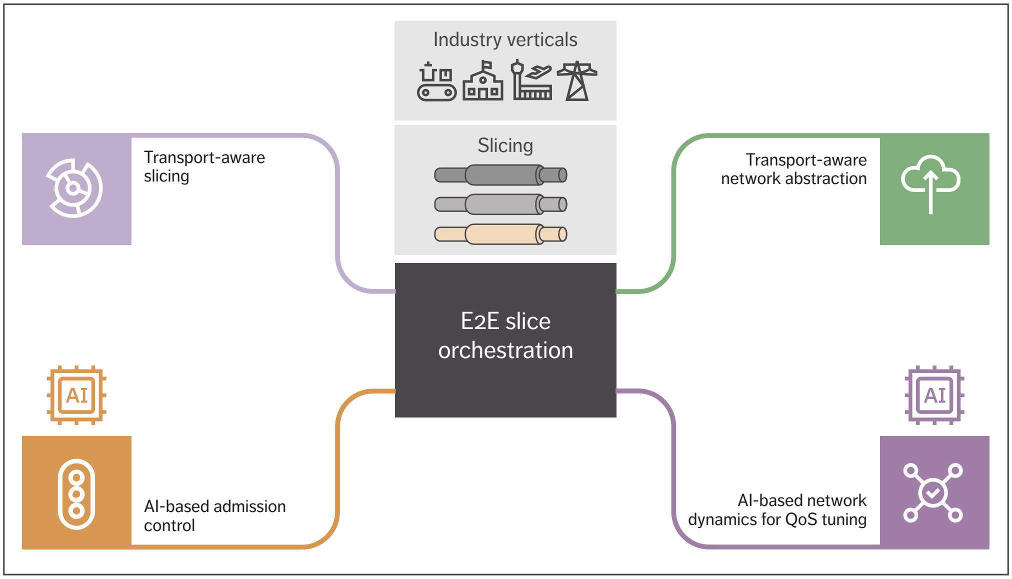
Industry verticals
Slicing
Transport-aware slicing
Transport-aware network abstraction
E2E slice orchestration
AI
AI
AI-based admission control
AI-based network dynamics for QoS tuning

Figure 1 The four core components of Ericsson's E2E network slicing orchestration solution

rate for each connection according to the actual needs.

In fact, the orchestration of all the various infrastructure components should be extended to ensure effective mapping based on actual traffic behavior. The mapping must be fully dynamic and automatic for each specific service, including those that impose a guaranteed and deterministic performance level.

Moreover, each network slice should operate as an isolated E2E network, tailored to fulfill the specific requirements requested by a particular application. The most appropriate type of isolation for a particular slice depends on the transport technology it uses. For example, in the case of optical networks, it is possible to utilize the separation provided by the wavelength channel, while a dedicated scheduler could be used for packet networks. Our E2E orchestration solution bases decisions about isolation on information from the transport domain related to the types of isolation that it can provide.

As shown in *Figure 1*, our transport-aware, E2E network slicing orchestration solution consists of four main components:

- Transport-aware slicing including the interaction between RAN/CN and transport
- Transport-aware network abstraction
- Artificial intelligence (AI) based admission control (AC)
- AI-based network dynamics for QoS tuning.

Transport-aware slicing

One important aspect of network slicing orchestration is to map traffic from a single slice or group of slices to transport resources that match the required E2E QoS for that slice or group of slices. Dedicated transport may also be required when latency is an issue, when there is a need for transport observability per slice, or to guarantee isolation.

Since the transport layer is logically separated from the radio layer and the expected radio needs are known, the standard approach to transport resource allocation in this scenario is to base it on the peak of expected radio needs. The downside of this approach is that it frequently results in the over-provisioning of the transport layer, which may not always be feasible or economically justifiable.

As an alternative to the current approach in which the QoS of the RAN/CN and the QoS of transport are orthogonally associated and independently configured, we have developed an approach that avoids over-provisioning by making the radio layer orchestration aware of the transport resources. The traffic flows for a single slice (or group of slices with heterogeneous SLA needs) are mapped to the most appropriate transport connection in a shared (not dedicated) transport network.

It is important to note that our transport-aware approach requires some changes to current 5G slicing practices. Firstly, it requires that the service and its deployment area are associated to a particular slice. Secondly, it requires that the slice TAs are chosen to cover the deployment areas of the service. Thirdly, the E2E QoS parameters of the service need to be mapped to the corresponding network resources (RAN, CN and transport) associated to the slice by using a RAN/CN and transport abstraction to expose a suitable view of the network resources to the orchestrator.

When all of these conditions are met, a consistent association of the QoS of slices in the RAN/CN and the QOS of transport will be performed automatically, and both layers will be automatically configured.

Transport-aware abstraction

Transport-aware abstraction is a compact description of all network resources (radio, transport and cloud) that expose the corresponding QoS parameters (latency, bandwidth, resilience and so on) to the E2E orchestrator. Abstraction simplifies the resource details (such as quantity, vendors, location of the resource, physical details, real topology and so on) for the E2E orchestrator so that it can consider the essential transport features in a simplified way concurrently with the features of radio and cloud resources.

A network service is constituted by the sequence of virtual network functions (VNFs), physical network functions (PNFs) and container network functions (CNFs). Transport provides the connectivity among them. One of the main challenges is the optimization of resource placement on top of the underlying transport infrastructure. For example, VNFs/CNFs can be connected through a simple point-to-point transport link or, alternatively, through a meshed geographical transport network. These two options imply different latency values or different availability. Knowledge about transport characteristics is particularly relevant in the case of services with critical performance requirements. Transport-aware abstraction provides a flat view of all the network resources, including transport, to facilitate the best resource selection and enable cross-optimization.

By logically separating the service from the infrastructure technologies, the abstraction technique makes independent services from technologies, allowing these two elements to evolve independently. Additionally, the abstraction enables a clear separation of responsibility and roles between the infrastructure provider and the service provider.

Our proposed abstraction models a geographical site as a point of presence (PoP) that is composed of a set of functions. The various PoPs in the network are connected by virtual links that represent the transport connections. In the evolution of 5G toward 6G, some of the RAN functions can be located at geographically distant sites (in the cloud RAN scenario, for example), and some of the CN functions can migrate toward the access site up to the antenna sites. In other words, each site can host a mix of RAN and CN functions. The transport network is abstracted by point-to-point virtual links with associated QoS parameters, including information related to resilience and availability.

> “OUR TRANSPORT-AWARE APPROACH REQUIRES SOME CHANGES TO CURRENT 5G SLICING PRACTICES”

Most relevant industry verticals

- Manufacturing
- Energy and utilities
- Transportation
- Health care
- Media and entertainment
- Smart cities
- Governments

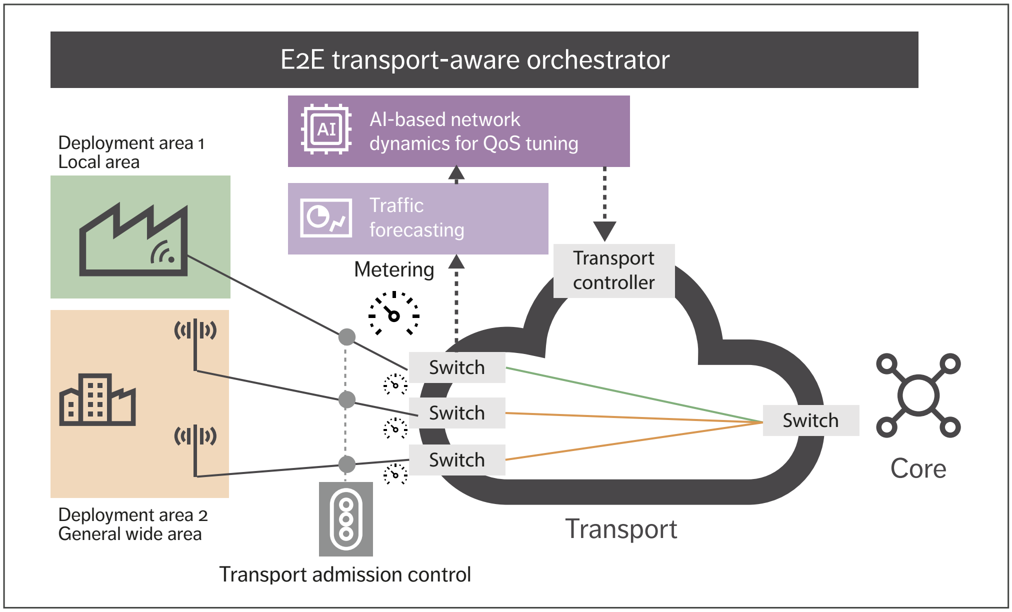

Figure 2 Functional blocks in the reference network

AI-based network dynamics for QoS tuning

Effective QoS tuning is challenging for transport-aware network slicing orchestration. The traffic associated with a specific service will change over the deployment area, both in space and time, which will affect the QoS parameters. Therefore, the assignment of QoS parameters to the transport tunnels supporting a slice should be done dynamically with appropriate mechanisms. In most cases, service traffic is by nature dynamic and at least partially related to predictable situations or historical trends that influence traffic load such as the time of day, weekday versus weekend and the like.

Dynamically tuning the parameters of slices to support current needs is especially useful in cases where the QoS parameters – peak information rate (PIR) and committed information rate (CIR) , for example – that are assigned to the services are not well known in advance and could potentially be overestimated. This plays an important role in the bandwidth partitioning that is needed to support services for industry verticals, by making it possible to do it dynamically according to actual traffic need.

AI-based admission control

The role of AI-based AC is to check whether network resources are available to support the QoS and traffic parameters of an incoming connection. Three types of AC should be considered when supporting services for industry verticals:

- AC related to the radio domain (managed by the Radio Resource Controller)
- AC for the transport domain
- E2E AC that combines the AC from radio and transport and manages the E2E service accordingly.

Reference network

Figure 2 shows the three main functional blocks in our reference network: transport AC, machine learning based or statistical traffic forecasting, and

AI-based network dynamics for QoS tuning. The combined use of these three blocks enables optimal dimensioning and operation of the transport network and reduces the level of over-provisioning.

Transport AC is invoked in two main phases. The first is when the slice for the E2E service is created, to dynamically verify the availability of the transport resources before their configuration/placement. If resources are not available, the connection is rejected and a notification is sent back to the originator or requester of the service. The second main phase is during service transmission, to ensure that QoS is in accordance with the SLA. Our proposed E2E orchestrator includes a novel function of transport AC that can be combined with the radio AC.

The traffic forecasting engine, which is either AI-based (machine learning) or statistical, utilizes metering functions on the traffic that enters the transport network nodes. This data is integrated with the current network status and with information related to specific circumstances (time of day, special events and so on). The traffic forecasting engine is responsible for determining traffic trends and their time behaviors over the considered deployment area. It also provides insights on actual radio traffic conditions that could not be observed and understood otherwise.

The QoS tuning functionality in the reference network is responsible for allocating and optimizing performance and network resources for all the admitted services, deciding at runtime the best routing based on a transport snapshot and the trends derived by traffic forecasting. Guided by policy, a certain amount of bandwidth is assigned to each transport tunnel (VLAN/VPN). The QoS parameters (effective bandwidth, PIR, CIR and committed burst size) are tuned according to actual needs, based on traffic prediction and measurements.

Case study: a smart factory

In a smart factory, many use cases are realized indoors in parallel. A smart factory pilot hosted at facilities operated by Comau, an industrial automation company, and TIM, a telecom operator, in Turin, Italy [2, 3] is a good illustration of this. Each use case in the smart factory puts specific, and often challenging, performance requirements on the telecommunication network. Failing to meet those requirements would immediately translate into bottlenecks in the manufacturing process.

The experimental area in the Comau factory is covered with a 5G network (RAN/CN) that is connected to TIM's central office (CO). The pilot includes a shared transport infrastructure, based on optical technologies deployed by Ericsson, which conveys radio traffic with appropriate E2E QoS. Cloud platforms, located on site and at TIM's CO, enable the implementation of Network Functions Virtualization and the support of Comau's applications that are running remotely. *Figure 3* visualizes the pilot architecture, where the orchestration functionality runs in a specific server hosted at TIM's CO (in the bottom-right corner).

Three use cases have been deployed in the Comau pilot, as shown on the left side of Figure 3. The first use case captures the motion of a real robot and, through an ultra-low latency radio link, produces a synchronized digital twin. The movement of the mechanical robot and of the respective virtual renderings are perfectly aligned in time.

The second use case is dedicated to demonstrating real-time monitoring of the industrial assets. Data is captured from a massive number of sensors and sent to an application deployed by Comau that runs in TIM's cloud. This application uses the acquired data to plan predictive maintenance, improve the accuracy of its production planning forecast, and improve quality, among other things.

The third use case demonstrates immersive telepresence for an enhanced remote support scenario in which the maintenance staff are assisted by a remote expert to investigate and solve a failure using augmented reality (AR) and step-by-step digital tutorials.

The Comau pilot features two of the four key components of our E2E transport-aware slice orchestration solution: the transport-aware slicing

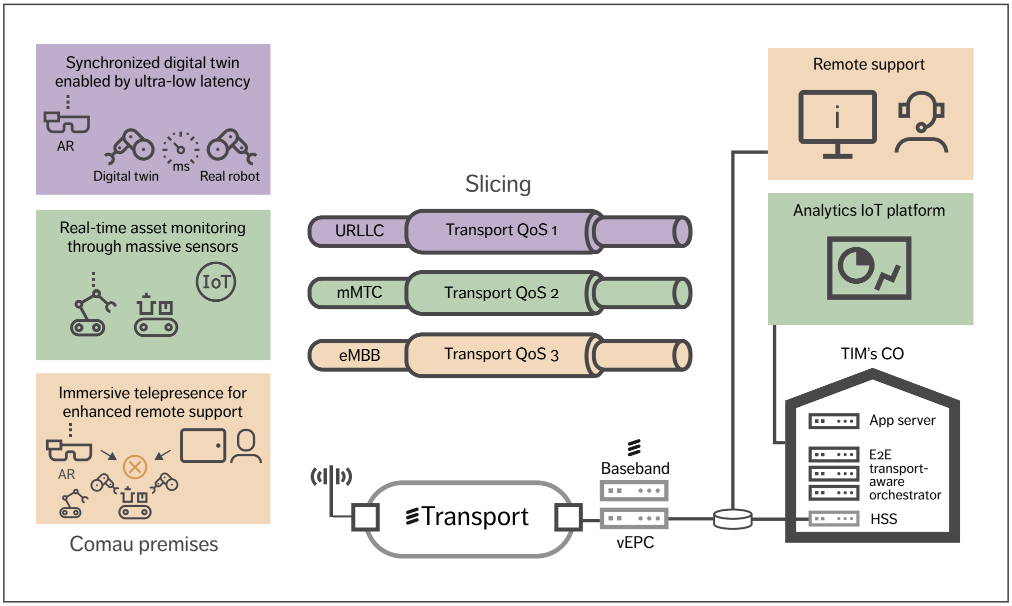

Figure 3 Architecture of the Comau pilot

and the transport-aware network abstraction functionalities shown in Figure 1. As a result, the solution ensures the alignment of all the resources (radio, transport and cloud) involved in the provisioning of the services with the related E2E QoS characteristics. Based on the QoS defined on the radio network, the corresponding requirements on transport are identified by the orchestrator, which then sends the requests to the transport domain. E2E QoS is managed automatically and dynamically through a unique infrastructure composed of radio and transport.

AI-based network dynamics for QoS tuning and AI-based AC – the other two key components of the solution presented in Figure 1 – have been defined and demonstrated as software modules in Ericsson laboratories.

The Comau pilot clearly demonstrates that the transport domain can maintain the properties of the network slice(s) it transports without the need to dimension it to the peak of the radio traffic, which will be a critical enabler for a wide variety of uses cases. Our concept of introducing transport-awareness into the slicing mechanism, and the other concepts demonstrated in the Comau pilot, have already gained significant recognition in our industry [4, 5].

Conclusion

The digital services of the future will demand new capabilities in 5G and beyond, including appropriate end-to-end (E2E) QoS at scale, in specific deployment areas (local, confined wide, general wide), where services vary dynamically in time and space. To meet these requirements, the network infrastructure will need to support a mix of heterogeneous services, including those demanding extreme performance.

Transport-aware network slicing orchestration will serve as a key enabler of services for industry verticals because its ability to manage network resources according to actual needs ensures cost-efficient service support. It does this by mapping the

E2E QoS parameters of the service associated to the slice on the corresponding network resources, including the transport connections in the deployment area. The infrastructure domains can automatically maintain the properties of the network slice(s), from the service provisioning phase, without the need to dimension the transport network resources to the peak of the radio traffic. Our solution also features an artificial intelligence method to perform traffic forecasts and dynamically tune the QoS of the transport connections in the service deployment area, with the scope to optimize the usage of resources while guaranteeing the required performance level.

In essence, to support services for industry verticals, our research indicates that 5G slice orchestration should be extended in several aspects to increase automation and awareness of transport, and to maximize the amount of traffic served by reducing over-provisioning.

Further reading

- **Ericsson blog, Unlocking network transport in 5G and 6G networks, available at:** *https://www.ericsson.com/en/blog/2021/12/network-transport-5g-6g*
- **Ericsson blog, Highlights of key end-to-end network slicing capabilities, available at:** *https://www.ericsson.com/en/blog/2019/5/highlights-of-key-end-to-end-network-slicing-capabilities*
- **Ericsson blog, Network slicing orchestration, available at:** *https://www.ericsson.com/en/blog/2018/5/network-slicing-orchestration*
- **Ericsson, service orchestration, available at:** *https://www.ericsson.com/en/service-orchestration*

References

1. **Ericsson, Top 10 network slicing use cases to target, available at:** *https://foryou.ericsson.com/eso-network-slicing-use-cases-report.html*
2. **5GROWTH, 5G-enabled Growth in Vertical Industries, available at:** *https://5growth.eu/*
3. **YouTube, 5G for Industry 4.0: COMAU pilot, available at:** *https://youtu.be/tlyQBmRbNf0*
4. **5GPPP, 5G Infrastructure PPP – Trials and Pilots, December 2020, available at:** *https://5g-ppp.eu/wp-content/uploads/2020/12/5GInfraPPP_10TPs_Brochure2.pdf*
5. **5GPPP, 5G Infrastructure PPP – Trials and Pilots, August 2021, available at:** *https://5g-ppp.eu/wp-content/uploads/2021/10/5GInfraPPP_10TPs_Brochure2021_v1.0.pdf*

THE AUTHORS

Paola Iovanna

◆ joined Ericsson in 2000. She currently serves as a principal researcher driving research activities on transport network and orchestration solutions for next-generation mobile networks (beyond 5G). She has previously led a variety of activities within several EU projects, with responsibility for both pilots and field trials within industry verticals. The author of 70 patents and more than 80 publications, Iovanna holds an M.Sc. in telecommunications engineering from the University of Rome Tor Vergata, Italy.

Malgorzata Svensson

◆ is an expert in operations support systems. She joined Ericsson in 1996 and has worked in various areas within research and development. Svensson has broad experience in business process, function and information modeling, information and cloud technologies, analytics, DevOps processes and tool chains. She holds an M.Sc. in technology from the Silesian University of Technology in Gliwice, Poland.

Alexey Shapin

◆ joined Ericsson in 2017. In his role as senior researcher, he contributes to the radio architecture and protocol design of LTE and 5G New Radio, working closely with 3GPP standardization teams. His research focus is on time-critical communication and ultra-reliable low-latency communication. Shapin holds a Ph.D. in telecommunication from the Siberian State University of Telecommunications and Information Science, Novosibirsk, Russia.

Giulio Bottari

◆ is a master researcher at Ericsson Research in Pisa, Italy. Since joining the company in 2006, his research interests have focused on transport for radio, optical networks and ICT applications for industries. He also served as the innovation manager of the H2020 5G transformer project. Bottari is the author of 80 patents and several articles in publications by the IEEE (Institute of Electrical and Electronics Engineers). He holds an M.Sc. in telecommunications engineering from the University of Pisa.

Fabio Ubaldi

◆ joined Ericsson in 2011. A senior researcher in control plane methods for optical and radio systems, his current research focuses on transport network architecture and orchestration solutions for next-generation radio systems (beyond 5G). He is also active within the frameworks of several EU projects. Ubaldi is the author of 16 filed patent applications and more than 10 publications in the IEEE journal. He holds an M.Sc. in telecommunications engineering from the University of Perugia, Italy.

Filippo Ponzini

◆ is a senior researcher who joined Ericsson in 2007. His expertise includes 5G radio systems, optical transport and integrated photonics, and at present his work focuses on the definition of next-generation radio systems (beyond 5G). Ponzini holds an M.Sc. in telecommunications engineering from the University of Parma, Italy, and an MBA from the Sant'Anna School of Advanced Studies in Pisa, where he also serves as a researcher. Ponzini is the author of more than 30 publications and 40 patents.

Marzio Puleri

◆ is a master researcher whose interests include packet networks, intelligent traffic management, robotics, the Internet of Things, artificial intelligence and the support of industrial and logistics services through mobile networks. Puleri also has extensive knowledge of microelectronics and microwave systems. He holds an M.Sc. in electronic engineering from the Sapienza University of Rome and has worked at Ericsson since 1993.

The authors would like to thank to Massimiliano Maggiari and Gunnar Mildh for their contributions to this article.