IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 29 September 2022, accepted 19 October 2022, date of publication 28 October 2022, date of current version 4 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3218061

RESEARCH ARTICLE

Chained Orchestrator Algorithm for RAN-Slicing Resource Management: A Contribution to Ultra-Reliable 6G Communications

JOSE J. RICO-PALOMO[1], JESUS GALEANO-BRAJONES[1], DAVID CORTES-POLO[2], JUAN F. VALENZUELA-VALDES[3], AND JAVIER CARMONA-MURILLO[1]

[1]Department of Computing and Telematics System Engineering, Universidad de Extremadura, 06006 Badajoz, Spain
[2]Department of Signal Theory and Communications and Telematics Systems and Computing, Rey Juan Carlos University, Móstoles, 28933 Madrid, Spain
[3]Department of Signal Theory, Telematics and Communications, CITIC, Universidad de Granada, 18014 Granada, Spain

Corresponding author: Jose J. Rico-Palomo (jjricopal@unex.es)

This work was supported in part by the Spanish National Program of Research, Development, Innovation, under Grant RTI2018-102002-A-I00; and in part by the Junta de Extremadura under Project IB18003 and Grant GR21097.

ABSTRACT The exponentially growing trend of Internet-connected devices and the development of new applications have led to an increase in demands and data rates flowing over cellular networks. If this continues to have the same tendency, the classification of 5G services must evolve to encompass emerging communications. The advent of the 6G Communications concept takes this into account and raises a new classification of services. In addition, an increase in network specifications was established. To meet these new requirements, enabling technologies are used to augment and manage Radio Access Network (RAN) resources. One of the most important mechanisms is the logical segmentation of the RAN, i.e. RAN-Slicing. In this study, we explored the problem of resource allocation in a RAN-Slicing environment for 6G ecosystems in depth, with a focus on network reliability. We also propose a chained orchestrator algorithm for dynamic resource management that includes estimation techniques, inter-slice resource sharing and intra-slice resource assignment. These mechanisms are applied to new types of services in the future generation of cellular networks to improve the network latency, capacity and reliability. The numerical results show a reduction in blocked connections of 38.46% for eURLLC type services, 21.87% for feMBB services, 12.5% for umMTC, 11.86% for ELDP and 11.76% for LDHMC.

INDEX TERMS 6G, RAN-slicing, reliability, capacity, latency, resource management, channel estimation.

I. INTRODUCTION

The long-awaited all-connected society is rapidly becoming part of our lifestyle. The world is flooded with mobile phones, tablets, laptops, wearables, industrial systems, smart cities, and other devices connected to the Internet, and growth prospects do not seem stagnant. In recent years, the number of smartphones has increased by 93 million, and the number of connected devices now exceeds 5.22 billion worldwide [1]. This increment in the number of devices is reflected in the data traffic, which has already recorded a volume of more than 55 Exabytes per month by 2021. To meet these demands, 5G technology was introduced.

This generation of cellular networks standardises different types of services (eMBB, URLLC and mMTC) to classify applications according to demands, connection requirements and traffic, among others.

However, the development of new applications, such as autonomous vehicles and tele-medicine, requires new latency specifications below 1 ms [2], [3]. High-capacity-demanding applications, such as Virtual and Augmented Reality (VR/AR), add complexity to an already problematic scenario, along with the number of connected, low-power and synchronised devices used by Industry 4.0. These applications require a much larger volume of data,

The associate editor coordinating the review of this manuscript and approving it for publication was Meng-Lin Ku.

in addition to increasing the number of devices connected to the network [4]. These data reflect the huge volume of traffic generated by this type of communication and suggest that conventional technologies may struggle to meet the demands. High mobility, low power consumption of devices and high density are challenges that cellular technology must address. Therefore, the future generation of cellular networks (6G) was devised to address this scenario.

6G Communications offers higher specifications than its previous generation, and a more varied and specialised division of service types to provide dedicated services to new and emerging applications. In addition, given its versatility, it can implement mechanisms to manage its resources. Dynamic resource reconfiguration through RAN-slicing is one of its key point.

Because of the greater specificity of services that 6G Communications can offer, access networks can be provisioned using slices to meet the demands of emerging applications, having the ability to particularise user demands more than in 5G. Specifically, critical communications, such as autonomous vehicles, minimally invasive tele-medicine applications, and industrial Internet, require near-zero latencies and maximum reliability. All of the above demands higher requirements than those offered by previous generations of cellular networks. The particularisation and deployment of slices focused on these types of services make a more efficient use of the resources available on the network, in addition to having the capacity to be deployed in real time. Also, intelligent management mechanisms help 6G networks deploy priority-based RAN-slicing mechanisms that increase the reliability of communication.

The main contribution of this work is the development of an orchestration algorithm for 6G RAN-Slicing, with the objective of ensuring ultra-reliable cellular communication to reduce the number of blocked connections by increasing the average network capacity and latency. This solution is based on the concatenation of resource estimation techniques, dynamic resource management in inter-slice environments, and reallocation of resources between different slices. The performance of the proposed solution was tested by simulations and compared with the standardised baseline link planning and resource allocation for 5G-NR without RAN-Slicing. The numerical results showed an improvement in network reliability depending on the type of service: 38.46% for eURLLC, 21.87% for feMBB, 12.5% for umMTC, 11.86% for ELDP, and 11.76% for LDHMC.

This paper is organised as follows: an introduction to 6G and RAN-Slicing technology is presented in Section II. Section III presents a taxonomy of related works in the research field of this study. Section IV details the modelling used in the simulations. Section V describes the proposed RAN-Slicing algorithm, and Section VI details the experimentation performed to assess the performance of the proposed approach. The last part of the work, in Section VII, comprises the main conclusions of the research and suggests possible approaches to further investigate the matter.

II. RAN-SLICING AS A KEY ENABLER TECHNOLOGY FOR 6G COMMUNICATIONS

New applications developed recently and new requirements make necessary a new classification of service types [5]. 6G Communications establishes a new paradigm that aims to provide full wireless coverage to meet the objective of connectivity anywhere and anytime. It will be able to serve a large number of users with extremely high data rates and exceptionally low latencies, joining satellite, terrestrial, aerial and quantum communications, among others. The 6G ecosystem would also continue the trends of previous generations, which included new services with the addition of new technologies. The services proposed by 5G will evolve to address the latest applications and traffic characteristics. These new services are as follows [6]:

- Further enhance Mobile Broadband (FeMBB): Applications which require a high bandwidth and a lot of capacity to meet the demands. This includes technologies such as Holographic Verticals, Full-Sensory Digital Reality (VR/AR), Tactile/Haptic Internet and UHD/EHD Videos.
- Long Distance and High Mobility Communications (LDHMC): Users who move at high speed while they often are far away from the network access point. Examples include hyper-high-speed railway (HSR), space travel applications and deep-sea sightseeing, among others.
- Extremely Ultra Reliable and Low Latency Communications (eURLLC): This service handles the communications that are critical and has a high priority, where a near-zero error rate must be ensured. Some of these technologies include Fully Automated Driving and Industrial Internet.
- Extremely Low-Power Communications (ELPC): Applications that must have a minimum power consumption but the connection to the grid must be guaranteed. E-Health technologies and nano devices, robots and sensors are examples of such applications.
- Ultra-Massive Machine-Type Communications (umMTC): Ensuring sufficient capacity for the establishment of thousands of connections is one of the most important features. Internet-of-Everything (IoE) and smartcities are two technologies that fall under this type of service.

Early preliminary studies established higher requirements than 5G. To meet these requirements, several enabling technologies are being considered for inclusion in the 6G ecosystem. These include THz-communications, very-large-scale antenna arrays, laser and visible light communications, spatial satellite links, and core/RAN slicing. Artificial Intelligence also plays an important role in this type of networks, as well as cloud/edge/fog computing, blockchain, Software Defined Networking (SDN) and Network Function Virtualisation (NFV).

One of the most promising enabling technologies in 6G is RAN-Slicing [7]. This key enabler technology involves

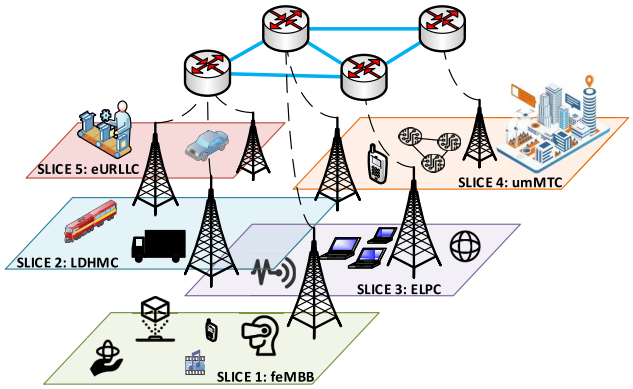


FIGURE 1. Schematic diagram of RAN-Slicing segmentation for each type of service.

the segmentation of the network infrastructure into logically self-contained networks. Each slice, which is designed and deployed for a specific type of service, consists of functions and resources abstracted from underlying communication and network resources. The concept was conceived for the 5G core network; however, to meet user experience expectations, it was necessary to upgrade to the RAN. Figure 1 shows a schematic segmentation of the RAN into slices depending on the service type.

In a conventional RAN, the transmission employs a best-effort strategy without resource reservation, which cannot guarantee Quality of Service (QoS). However, RAN-Slicing implements resource management mechanisms to meet the demands of the users. These mechanisms manage the resources of the different slices to ensure an increase in the Key Performance Indicators (KPIs), to meet the requirements of the next generation of cellular networks. These are developed to supply specific needs and are deployed when necessary. According to the literature, RAN-Slicing techniques can be developed using artificial intelligence, optimisation problems, dynamic resource management, among others, and a combination of these [8], [9], [10], [11].

These resources can be distributed among those available for a slice or ceded by another slice. Therefore, the complete state of the RAN and its slices must be known. For this purpose, orchestrators are used, which are devices that are aware of the state of the RAN, the BSs that compose it and its resources.

This paradigm endows the network with great flexibility and versatility for resource reconfiguration, which is necessary for a new scenario that cellular networks must face. Owing to the variety of services that will be differentiated in 6G, the mechanisms to be deployed on the RAN in operation must be able to reconfigure these resources in real time. They must also be able to differentiate the criticality of each user and reallocate resources proportionally to the priority of their traffic.

III. RELATED WORKS

RAN-slicing technology has been extensively studied in 5G, and some challenges are posed in the literature that must be fulfilled. Many of these studies require knowledge of communication channel characteristics to avoid overload and congestion [12]. In addition, a new line is opened focusing on resource reconfiguration algorithms in the slices, considering the QoS; channel variations can affect the QoS of the most critical services [13]. In [14], the need for resource estimation prior to establishing the connection was highlighted. Other challenges were raised in [15], which stated that there is a necessity to develop mechanisms for resource allocation and sharing in a slice-based RAN. Furthermore, in [16], a target for dynamic resource allocation was discussed, showing the need for developing algorithms to reallocate resources between different slices.

Evolving to 6G Communications, a major challenge is to achieve dynamic network orchestration and slice resource management according to real-time network information and service requirements [6]. Some authors also agree on the goal of managing and sharing resources in the slices [8]. In [17], the issue of coexistence between different types of services and resource management in beyond-5G and 6G networks was discussed.

To overcome these challenges, different RAN-Slicing-based solutions and mechanisms have been proposed in the literature and can be divided into four blocks: user-centric solutions, inter-slice-based and intra-slice-based techniques and orchestrator algorithms, which combine the above solutions by means of resource planning algorithms. Table 1 shows a taxonomy of the related works presented in this Section, according to the proposed RAN-Slicing solution.

User-centric mechanisms respond to resource estimation techniques in which the network allocates resources according to the requirements of users at a specific time. In the taxonomy, the different solutions offered to the challenges posed by RAN-Slicing in next-generation cellular networks are presented. Some authors use resource estimation techniques to reduce the number of communication deadlocks, such as [9], which poses an optimisation problem to maximise the resources allocated to UEs by using resource reservation operations, depending on the network demands. In [18], the authors presented a novel latency-sensitive 5G RAN slicing solution based on partitions of radio resources among slices, considering the rate and latency demands of applications. A resource reservation scheme in factory-like environments was proposed in [19] using optimisation techniques.

Numerous studies have been carried out in intra-slice solutions, which are solutions based on resource sharing within a slice. In [20], the authors presented a statistical model that characterises resource sharing in a RAN-Slicing scenario, considering the available resources in layer 3. A solution for resource sharing is also presented in [21], but making use of genetic algorithms to maximise long-term network utility in Slice as a Service (SlaaS). A resource allocation slicing policy for inter-slice isolation across Mobile Virtual Network Operators (MVNOs) was investigated in [22] with a multi-objective optimisation that minimises the inter-slice

TABLE 1. Taxonomy of related works divided by type of proposed solution (Prop.) and compared with the proposal developed in this article.

Solution type	Ref.	Proposed mechanism	Improved metric	Application	Experimental/simulation
User-centric (resource estimation and reservation)	[9]	Optimisation technique	Capacity and power consum.	5G RAN	Simulation
	[18]	Dynamic slice creation	Latency and capacity	5G RAN	Simulation
	[19]	Optimisation technique	Latency and reliability	5G RAN	Simulation
Inter-slice (resource sharing inside a slice)	[20]	Radio resource model	Capacity	5G CN and RAN	Simulation
	[21]	Optimisation technique	Network utility	5G RAN	Simulation
	[22]	Optimisation technique	SINR	5G RAN	Simulation
Intra-slice (resource sharing between different slices)	[23]	Optimisation technique	Capacity	5G RAN	Experimental
	[10]	Optimisation technique	Latency and capacity	5G RAN	Simulation
	[24]	SDN/NFV algorithm	Capacity	5G Virtualized RAN	Simulation
Orchestrator algorithms	[8]	Machine learning	Reliability	6G Communications	Simulation
	[11]	Machine learning	Latency and reliability	6G Communications	Simulation
	[25]	SDN/NFV algorithm	Latency and capacity	5G CN	Experimental
	Prop.	Chained algorithms	Reliability, lat. and capac.	6G Communications	Simulation

interference generated by the simultaneous multiplexing of resource blocks.

Intra-Slice-based resource management solutions propose techniques which share resources between two or more slices. In [23], the authors presented a complete solution for dynamic RAN-Slicing resource allocation, where the optimal slice configuration was computed through a joint evaluation of the slice Service Level Agreements (SLAs) and the real-time evolution of the served traffic of the users. The research carried out in [10] also needs to be highlighted, in which an optimisation problem is proposed to maximise the dynamic allocation of resources in eMBB and URLLC service types. In other works, a customised shape-based heuristic algorithm for users to improve resource utilisation and QoS fulfilment was presented [24].

The solutions based on orchestration algorithms, that is algorithms that manage access to segmented network resources, which are presented in [8] and [11], are focused on 6G Communications and use machine learning techniques to improve the QoS performance for the users and in the whole network, respectively. In [25], an orchestration algorithm-based solution deployed in an experimental architecture was evaluated, achieving high flexibility and scalability by employing SDN and NFV technologies.

Our contribution is encompassed in orchestration-based solutions and provides a chained algorithm for RAN-Slicing resource management, which is applied to 6G Communications to improve network reliability, capacity, and latency, and its performance has been tested by simulations. It combines user-centric channel-estimation techniques and a resource pooling mechanism for dynamic resource allocation in inter-slice domains and intra-slice resource reassignment.

This proposed heuristic solution allows resource reconfiguration at service time when the network is in operation. In addition, in [8] and [21], proposals whose objective is real-time resource reconfiguration were presented, although the proposals use machine learning and genetic algorithms, respectively, contrary to our proposal. In addition, these proposals focus on network metrics, such as the reconfiguration of VNFs to minimise the computation time of network devices and network utility. In our proposal, user-centric resources are studied in comparison to other studies. Furthermore, in [23], the study of mechanisms that can be deployed at service time was also carried out, but by means of a small-scale experimental network. However, our proposal focuses on a dense urban environments and is tested by simulations.

The aforementioned chaining of mechanisms is carried out by a slice orchestrator, who knows the state of all cells and slices. Compared to other works, the considered scenario in which the proposed techniques are deployed is a heterogeneous network, e.g., in [11], another RAN-Slicing strategy was developed, but a mechanism for a single-cell was considered. In contrast to the proposal presented in this paper, in [25] and [20], focus was placed on network slicing solutions at higher layers, without considering the cellular network or considering restrictive admission control at the network layer, respectively.

The solution proposed in this work was developed to increase the reliability of the RAN, minimise latency, and maximise the capacity of the links between the BSs and UEs.

IV. SYSTEM MODEL

This section presents the modelling of the system used for the simulations.

The simulated network is composed of several layers. The first is the RAN, which is composed of a heterogeneous Backhaul Network (BN) consisting of N BSs (macro and small) and their links, distributed over the simulation map. The BN is represented by a set of BSs as $BN_{BS} = \{BS_1, BS_2, \ldots, BS_N\}$. The second layer is the slice orchestrator, which controls the resource management logic of all the BSs. The links

TABLE 2. Summary of metrics used in the simulations.

Metric	Symbol	Units
SINR	$SINR$	dBm
Received power	P_{RX}	dBm or mW
Noise	P_{N_0}	mW
Transmit power	P_{TX}	dBm or mW
Gain (TX and RX)	G_{TX} / G_{RX}	dB
Frequency	f	GHz
Distance	d	meters
Capacity	C	Mbps
Bandwidth	BW	MHz
Spectral efficiency	S_c	bps/Hz
Propagation time (latency)	T_{Prop} and $L_{u,i}$	seconds
Estimated latency	$L'_{u,i}$	seconds
Pilot time	L_{pilot}	seconds
Available resources (link)	$W_{u,i}$	tuple $\{C, L\}$
Traffic demands	D_u	Mbps

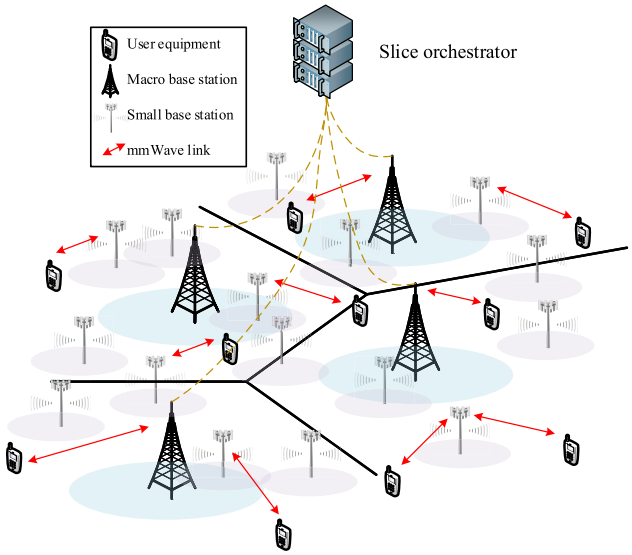


FIGURE 2. Diagram of the deployed network used for the simulations.

are dedicated to each BS and are considered lossless. The last tier of the system model is composed of a group of K User Equipment (UE), defined by $U = \{U_1, U_2, \ldots, U_K\}$, randomly distributed for the scenario and following a Fluid Flow (FF) mobility model. These UEs are modelled by a MIMO array of antennas. Figure 2 shows the layout of the network scenario.

A. COMMUNICATION MODEL

The communication between UE u and BS i is established by means of a millimetre wave (mmWave) link $W_{u,i}$, which is generated by a link planning policy based on Signal-To-Interference-Plus-Noise-Ratio (SINR). This link has a maximum capacity set by antenna technology and bandwidth resources that can be served by the BS. Furthermore, the link has latency, which refers to the time it takes a signal to reach from a sender to a receiver. This definition is explained in Section IV-A4.

1) SINR LINK PLANNING

The link planning algorithm consists of evaluating the SINR level of all $BS \in \{BN_{BS}\}$ and selecting the one that offers the highest value. When a UE $u \in \{U\}$ needs to connect to a BS $i \in \{BN_{BS}\}$, the SINR is calculated as follows:

$$SINR_u = \frac{P_{RX(i,u)}(mW)}{\left[\sum_{j=1\ j\neq i}^{N} P_{RX(j,u)}(mW)\right] + P_{N_0}(mW)} \quad (1)$$

where $P_{RX(i,u)}$ is the power received by the BS i for the UE u in milliwatts, P_{N_0} is the noise power in milliwatts, and $\sum_{j=1\ j\neq i}^{N} P_{RX(j,u)}$ is the interference, i.e. the sum of the power received, by all BS $j, \forall j \in \{BN_{BS}\} - \{i\}$ that works at the same frequency.

The received power P_{RX} is calculated using the well-known link budget formula:

$$P_{RX}(dBm) = P_{TX}(dBm) + G_{TX}(dBm) + G_{RX}(dBm) - PL(dB) \quad (2)$$

where P_{TX} represents the transmit power in dBm, G_{TX} and G_{RX} are the transmitter and receiver gain respectively, and PL is the path losses of the link.

2) PROPAGATION CHANNEL AND PATH LOSSES

The free-space path losses (PL) follow the ABG model standardised by 3GPP in [26]. This model is a large-scale propagation path loss model. It can be parameterised in terms of distance, frequency, and shadow factor. The formula that describes its behaviour is:

$$PL(dB) = PL^{ABG}(f, d)[dB] = 10\alpha \log_{10}\left(\frac{d}{1\text{m}}\right) + \beta + 10\gamma \log_{10}\left(\frac{f}{1\text{GHz}}\right) + X_{\sigma}^{ABG} \quad (3)$$

where $PL^{ABG}(f, d)$ denotes the path loss in dB over frequency f and distance d. α and γ are coefficients showing the dependence of path loss on distance and frequency, respectively. β is an optimised offset value for path loss. X_{σ}^{ABG} is the shadow factor of the ABG model. For each propagation scenario, the α, β, γ, and σ values vary. Table 3 shows the parameters used for the simulations depending on the scenario (*Scen.*) and environment (*Env.*) type, where d is the distance range in meters, f is the frequency range in GHz and β and σ are expressed in dB. *UMa*, *Umi* and *Ind.* correspond to Urban MacroCell, Urban microcell and Indoor scenarios, respectively.

3) LINK CAPACITY MODEL

Link capacity is the product of the spectral efficiency S_c and the bandwidth BW_u assigned to UE u,, and represents the maximum amount of data that can be transmitted over a communication link. Is measured in Mbps. The spectral

TABLE 3. Parameters of the ABG model used in the simulations [27].

Scen.	Env.	d	f	α	β	γ	σ
UMa	**LOS**	60-930	2-38	1.9	35.8	1.9	2.4
	NLOS	61-1238	2-38	3.5	13.6	2.4	5.3
Umi	**LOS**	27-54	28-73	1.1	36.8	2.1	4.3
	NLOS	48-235	2.9-73	2.8	31.4	2.7	6.8
Ind.	**LOS**	4-49	2.9-73	1.6	32.9	1.8	1.8
	NLOS	4-67	2.9-73	3.9	19	2.1	2.1

efficiency, measured in bps/Hz, is defined by:

$$S_c = \log_2\left(det\left[I_{N_{RX}} + \frac{SINR}{N_{TX}} H * H^{T'}\right]\right) \tag{4}$$

where $det[]$ is the determinant of [], $I_{N_{RX}}$ is the identity matrix whose dimensions are the number of receiver MIMO antennas, N_{TX} is the number of transmitter antennas and H is the channel matrix, which is generated randomly using a complex normal distribution. $H^{T'}$ is the conjugate transpose of the matrix [28]. The rows and columns of the channel matrix are defined by the numbers of receiver and transmitter antennas, respectively.

To estimate the channel conditions, it is necessary to randomly generate H matrices using complex normal distribution $\mathcal{N}(\mu,\ \sigma^2)$, according to the parameters detected in the reception of the signal:

$$H' = \sum^{N_H}\sum^{M} \frac{H_{1(N_{tx}\times N_{rx})} \sim \mathcal{N}(\mu,\ \sigma^2)}{\sqrt{2}} + \frac{jH_{2(N_{tx}\times N_{rx})} \sim \mathcal{N}(\mu,\ \sigma^2)}{\sqrt{2}} \tag{5}$$

where H' is the estimated H matrix composed of N_H matrices of size $N_{tx} \times N_{rx}$, summed M times, which corresponds to the number of samples used. The final capacity was the average of all samples obtained. H_1 and H_2 correspond to the real and imaginary parts of the H-matrix, respectively.

The objective of the proposed mechanism is to offer the highest link capacity between between UE and the BS. Therefore, it is necessary to determine the highest BW assigned to the UE and increases the spectral efficiency by finding the highest SINR that can be offered by the BS to which it is connected:

$$max(C_{u,i}) = max(BW_{u,i}, SINR_{u,i}) \tag{6}$$

4) LATENCY MODEL

The latency model use is a composition of three values affecting uplink communication [29]. First, the propagation time T_{prop}, which is the time required for the wave with the information to travel from the transmitter to the receiver. Then, there is the tail time T_{tail}, which is the time required for information to wait in the BS queue. Finally, the handling time T_{hand}, that is the response time of the BS computing devices:

$$T_{total} = \text{link latency} = T_{prop} + T_{tail} + T_{hand} \tag{7}$$

assuming that $T_{hand} = \frac{1}{\mu(1-\beta)}$, and $T_{tail} = \frac{\beta}{\mu(1-\beta)}$ follows a $GI|M|1$ queue model.[1] In our system modeling, the base station queues are considered infinite; therefore, handling and tailoring times are negligible. The propagation time is defined as follows:

$$T_{prop} = \frac{2(t_{slot} - E[T_v])}{1 + f_{err}\left(\frac{\delta(f,d)}{\sqrt{2\sigma}}\right)} \tag{8}$$

where t_{slot} is the slot time between resource blocks defined by the standard, f_{err} is the error function and $\delta(f, d) = P_{TX} + P_{N_0} - PL(f, d)$. $E[T_v]$ refers to the propagation characteristics produced by mobile blockers.[2] This implies that there are moments in time when the link has No-Line-of-Sight (NLoS). These mobile blockers are modelled following an $M|G1|\infty$ queue, where the arrival is interpreted as the crossing between a blocker and the LOS link. This blockage time distribution can be approximated by using the mean waiting times:

$$E[T_v] = \frac{E[T_{LOS}]\,E[T_{NLOS}]}{E[T_{LOS}] - E[T_{NLOS}]} \tag{9}$$

where $E[T_{LOS}]$ is the mean time that the link is not blocked and $E[T_{NLOS}]$ is the mean time that the link is blocked by a mobile blocker.

In order to estimate the link latency, the following equation is used:

$$L'_{u,i} = t_{pilot}(u, i) - T'_{prop}(W_{u,i}), \quad \forall i \in \{BN_{BS}\}, \forall u \in \{U\} \tag{10}$$

where $L'_{u,i}$ is the estimated latency, $t_{pilot}(u, i)$ is the time taken for a pilot signal to travel from UE u to BS i and $T'_{prop}(W_{u,i})$ is the estimated propagation latency of the link.

The objective of the proposed mechanisms is to provide the lowest possible latency in the link between the UE and BS. Therefore, it is necessary to minimise the impact of mobile blockers and determine the highest power received by the BS to which the UE is connected:

$$min(L_{u,i}) = min(E[T_v]), max(\delta(f, d)) \tag{11}$$

B. SERVICES AND TRAFFIC MODELS

Numerous demands generated by the users are modelled according to various traffic models defined by 3GPP [30] and other entities, compiled in [31]. Table 4 presents the definition of the implemented traffic models and their priorities. Each traffic model has been assigned a priority P (a real number between 1 and 10) according to the capacity and latency requirements, depending on the type of services defined in the previous sections. In addition, the type of service assigned to traffic models is indicated.

[1] GI|M|1 is a queue in which inter arrival times follow a general arbitrary distribution (G), service times follow an exponential distribution (M), and 1 denotes that the model has a single server. The services times for handling and processing queues are μ and β respectively.

[2] Mobile blockers are objects that temporally interpose themselves in the Line-of-Sight (LoS) of the link between the transmitter and receiver (e.g. pedestrians, vehicles, etc).

TABLE 4. Implemented traffic models in simulation framework.

Applications	Service	Priority P	Type	Model	Reference
Gaming	feMBB	Medium (5)	Experimental	Real data rates and arrival times	[32], [33]
V2X	eURLLC	Very high (9)	Mixed	Mixes control, data and VOD traffic, and others	[34], [35]
Smartcities	umMTC	High (7)	Stochastic	Poisson process	[31], [36]
EHD VOD	feMBB	Medium (5)	Mixed	Real data rates and Truncated pareto-based arrival times	[37], [38]
AR/VR	LDHMC	Low (3)	Mixed	Real data rates and arrival times	[39], [40]
IoT	ELPC	Medium (5)	Stochastic	Poisson process	[41], [31]

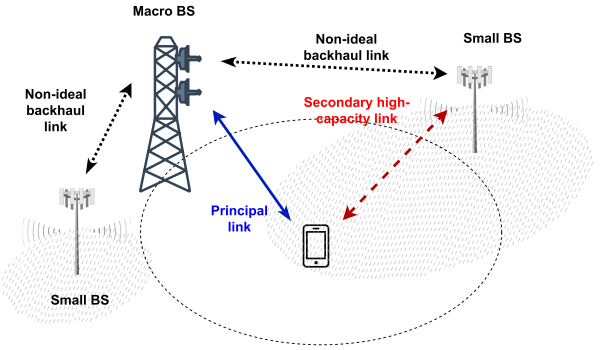


FIGURE 3. Schematic operation diagram of DC technology.

UE demands, defined by D_u, will consume the available resources of the UE-BS link $W_{u,i}$. If the BS or $W_{u,i}$ have lack of available resources to be satisfied, the demand will be blocked and discarded. These blockages are used as a metric of reliability, i.e., the number of errors in the traffic flows [42].

C. MULTI-CONNECTIVITY MODEL

According to [43], multi-connectivity (MC) can be used to enhance user throughput, coverage, and/or reliability. In terms of latency, to reduce it from the communication system. Complementary dual links are used for load balancing in the case of capacity or latency requirements.

Dual Connectivity (DC) technology (standardised in 3GPP Release 12 [44]) has been implemented in the simulation tool. As stated in Release 12, the UE must be configured to "utilise radio resources provided by two distinct schedulers, located in two NodeBs connected via a non-ideal backhaul", i.e., the UE is simultaneously connected to two non-collocated nodes (master and secondary). These links do not have to operate at the same frequency or be of the same cell type. In fact, the UE is trying to be connected to a MacroBS and SmallBS simultaneously. The secondary links were activated and deactivated according to the chosen schedule and according to the needs of the UE at any given time. A schematic of this operation diagram is shown in Figure 3.

The schedule of these links considers the needs of the UE connection. If the demand cannot be served because there are no resources available on the link, the scheduler commands a secondary link to be opened. When the demand is served and terminated, it is closed.

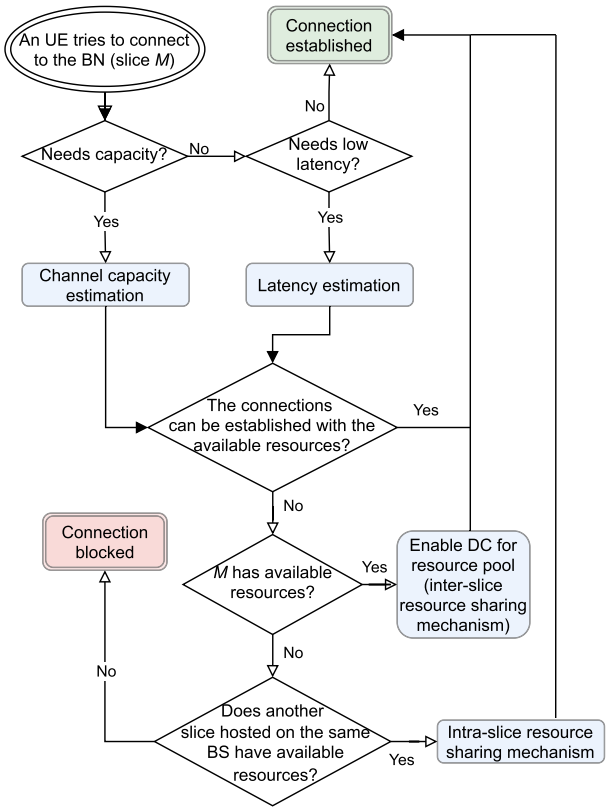


FIGURE 4. Decision diagram of the slice orchestrator algorithm.

V. PROPOSED ALGORITHMS

In this section, a chained orchestration algorithm for 6G RAN-Slicing resource management is described. It is based on dynamic resource management in a network, focusing on the capacity, latency, and reliability. This algorithm acts in a cascading process, that is, by chaining several techniques one after the other to reduce the number of blocked connections of the users. A flow diagram of the orchestrator decisions is shown in Figure 4.

This orchestrator solution is divided into three blocks, depending on the part of the network in which it operates:

- Resource estimation mechanisms: This block covers user-centric techniques based on channel estimation. This estimation can be used to determine the latency or capacity.

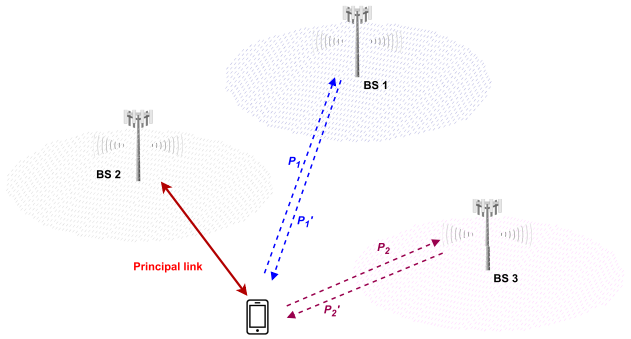


FIGURE 5. Process of sending and receiving pilots. The UE, without dropping the main link or stopping the information flow, continuously sends and receives pilots from other nearby BSs.

- RAN-centric intra-slice algorithms: In this block, a resource pooling technique is proposed for transferring resources between BSs inside the same slice. This mechanism is based on a DC.
- RAN-centric inter-slice algorithms: For the transfer of resources between different slices, different algorithms for efficient RAN bandwidth management are proposed in this block.

The type of service and criticality of the communication will be the determining factors that will enable this set of mechanisms. The slice orchestrator established in the RAN must know the status of the entire network and incoming demands. Knowing the behaviour, the algorithms can be run in cascade, i.e., triggered when the previous mechanism fails to take effect.

A. USER-CENTRIC MECHANISMS FOR RESOURCE ESTIMATION

The proposed resource estimation mechanism is based on the needs of the service and its criticality. If a UE needs lower latency or more capacity than its link can offer, it will connect to another BS that can guarantee these resources. These resources were guaranteed by reducing the number of blocked connections. Therefore, communication errors are reduced, i.e., reliability is increased. These solutions are based on channel estimation, which provides the UE with a complete view of the state of the RAN, in agreement with the two metrics to be evaluated.

The use of pilot signals is necessary to estimate link and channel properties. Pilots are signals that are used for supervisory, control, equalisation, continuity, synchronisation, or reference purposes. This method delegates the computation of the estimation to the UE and not to the RAN. For this purpose, the UE sends a pilot to the BS, which is returned by the BS. The operating scheme is illustrated in Figure 5.

1) LATENCY ESTIMATION

The time taken for the pilot to reach from the UE to the BS is the estimated propagation latency t_{prop}. The UE knows only the total time between sending its pilot and receiving it from the BS. This time is denoted as t_{pilot}. This time is broken down into the sum of: (i) the time it takes for the pilot of the UE to reach the BS (the same as t_{prop}), (ii) the time it takes for the BS to process that pilot and send its own T_{proc} and (iii) the time it takes for the pilot of the BS to reach the UE t_{BS}.

t_{prop} and t_{proc} are propagation times following the Equation 8, and are calculated as a function of transmit power P_{TX} and frequency link f: $t_{pilot} = t_{prop}(P_{TX_{UE}}, f) + t_{proc} + t_{BS}(P_{TX_{BS}}, f)$

Knowing the total time (pilot time), and knowing from the pilots the frequency of the link and the transmit power of the BS, t'_{BS} can be calculated. Because the processing time is negligible, the link latency between the UE and all BSs can be estimated.

2) CAPACITY ESTIMATION

Capacity modelling, as explained before, is a function of the number of transmitter and receiver antennas (N_{tx} and N_{rx}, respectively), bandwidth BW_u, channel properties H and SINR. To estimate the link capacity C', the UE must know the characteristics of receiving system, such as the bandwidth and the number of MIMO antennas at the link peer.

Pilot signals provide this information through the same mechanism, returning some information from the BS. The number of receiving MIMO antennas is available in the link information, and the bandwidth is known to be the bandwidth allocated to the UE if it is connected. This bandwidth is allocated using a simple resource management policy.

B. RESOURCE MANAGEMENT IN RAN-SLICING

Each BS $i \in BN_{BS}$ has a certain amount of resources, grouped in a tuple $R_i = \{C_i, L_{i,u}\}$. Each C_i corresponds to the available capacity per BS, and each $L_{i,u}$ defines the estimated latency that BS i can offer to UE u, depending on its position and propagation channel, based on resources management planning (with or without a slice schedule). $R(u)$ are the resources allocated to UE u. The base case of this resource management planning is defined by Algorithm 1. This algorithm proposed by 3GPP is the baseline for numerical results [45].

The algorithm works as follows: a user u is chosen from the set of active users in the scenario (Line 1). The user is connected to the BN (Line 2). This connection establishment uses a well-known schedule based on SINR, following Eq. 1. The signal level of a base station and the interference received by all the others are evaluated. After evaluation, u is connected to the one that offers the highest SINR level. This candidate base station is referred to as i (Line 3). Once the connection is established, u begins to generate traffic demands D_u (Line 4). These demands will be characterised by the required throughput and the minimum latency it needs to be established (Line 5 and 6 respectively). Both metrics are encompassed in a tuple $R(u)$, which represents the resources needed to satisfy the demand (Line 7). If the available resources by BS i (R_i) are greater than or equal to those needed to satisfy the demand (Line 8), the demand

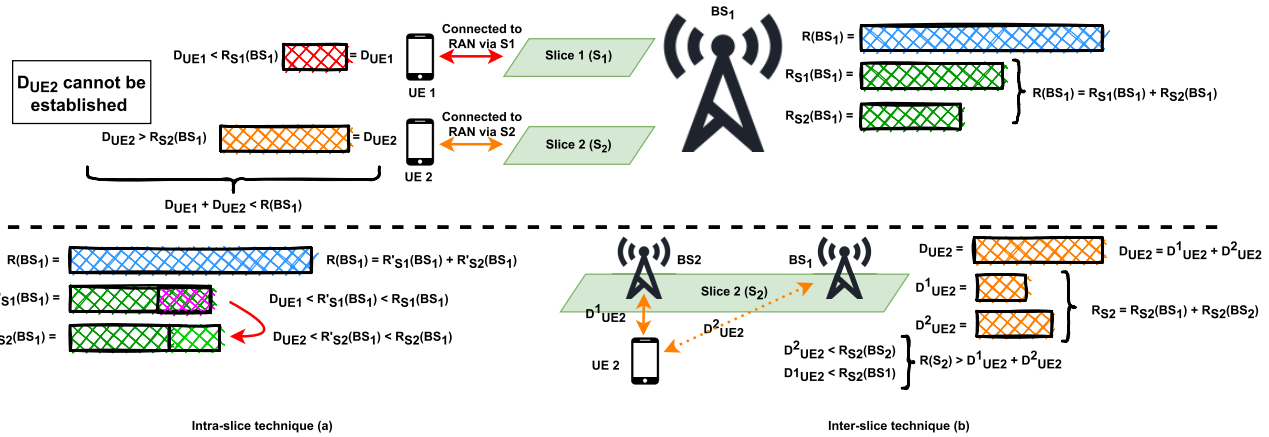


FIGURE 6. Operation of proposed RAN-based resource allocation techniques. (a) intra-slices techniques for reallocation of resources between different slices in the same BS (when there are no resources available within the same slice). (b) inter-slices techniques for sharing resources between different BSs within the same slice (DC).

Algorithm 1 Base Case Algorithm for Resource Management (Without RAN-Slicing)

Input:
$U \leftarrow$ Set of UEs
$BN \leftarrow$ Backhaul Network (set of BSs)
$R_i = \{C_i, L_{i,k}\} \leftarrow$ Available resources from BS i
begin
[1]: **foreach** u **in** U **do**
[2]: Connect u to BN
[3]: BS $i \leftarrow$ candidate $BS \in BN$
[4]: $D_u \leftarrow u$ generates a traffic connection
[5]: $C(u) \leftarrow$ required throughput by D_u
[6]: $L(u) \leftarrow$ min. latency required by D_u
[7]: $R(u) \leftarrow \{C(u), L(u)\}$
[8]: **if** $R(u) \leq R_i$ **then**
[10]: Assign $R(u)$ to u
[9]: $R_i \leftarrow R_i - R(u)$
else
[11]: $U' \leftarrow$ subset of U connected to BS i
[12]: $R'(u) \leftarrow \frac{R_i}{len(U')+1}$
[13]: Assign $R'(u)$ to u
[14]: Assign $R'(u)$ to each u' in U'
[15]: $R_i \leftarrow 0$
end if
end foreach
end

can be established. The resources that u needs are allocated to it (Line 9) and the available resources by the BS i are updated, subtracting those it has allocated (Line 10). When $R_i < R(u)$, i.e., BS i have no available resources to satisfy the demand, two events can occur. On the one hand, if D_u cannot support a reduction in QoS and cannot be served with $R'(u)$ because of their priority and the criticality of their type of service, the connection will be blocked. On the other hand, if the traffic can be served with $R'(u) < R(u)$ even if the QoS decreases, the connection will be established with fewer resources (lower QoS). In this case, the resources of all users connected to that base station are reallocated. First, a subset of users who comprehend the users connected to BS i is extracted. This subset is denoted by U' (Line 11). Second, a newly assigned resource is calculated using the formula presented in Line 12. The available resources R_i are divided by the number of connected users and the number that needs a connection. These new resources are assigned to u (Line 13), and consequently reassigned to all users of U', which were previously connected to BS i (Line 14). Finally, the resources available at the base station are updated (Line 15). If any of the demands of users belonging to U' do not support a decrease in their QoS, the demand is blocked.

The total complexity of the algorithm is $O(n^2 + n)$ for each user u in U. The connection of user u to the BN has complexity $O(n^2 + n)$, depending on the number of BSs in the scenario. This algorithm is based on Eq. 1: for each evaluated BS $i \in BN$, it is necessary to iterate all BSs $j \neq i$. In the worst case, after the connection, all users in U' must be iterated to reallocate their resources. The complexity of this operation is $O(n)$.

1) RAN-SLICING IN CELLULAR NETWORKS

Consider $S = \{S_1, S_2, \ldots, S_M\}$ a set of M slices in the RAN. Each slice S_m is distributed over the BN. Depending on the type of service it has been assigned (in this case, eURLLC, LDHMC, ELPC, feMBB and umMTC), S_m will have access to a set of resources R_m, which are represented by a tuple $R_m = \{C_m, L_m\}$, entailing portions of the BSs resources. In turn, each slice hosts a set of UEs $U'_{S_m} \in U$ of dimension $N' \leq dim(U)$. UEs belonging to U'_{S_m} are

physically connected to the BN via mmWave link to BSs and logically connected to S_m, having access to S_m resources, whether or not they are hosted on the BS to which they are connected.

When the RAN-Slicing schedule is enabled in the cellular network, the resource management planning follow the Algorithm 2. For this schedule, intra- and inter-slice algorithms act when the RAN cannot serve the connections owing to a lack of resources.

Algorithm 2 RAN-Slicing Algorithm for Resource Management

Input:
$U \leftarrow$ Set of UEs
$BN \leftarrow$ Backhaul Network (set of BSs)
$R_i = \{C_i, L_{i,k}\} \leftarrow$ BS i available resources
$S \leftarrow$ Set of slices
$R_m = \{C_m, L_m\} \leftarrow S_m$ available resources

begin
[1]: **foreach** u **in** U **do**
[2]: Connect u to BN
[3]: BS $i \leftarrow$ candidate $BS \in BN$
[4]: $S' \leftarrow$ slice to which u belongs
[5]: $D_u \leftarrow u$ generates a traffic connection
[6]: $C(u) \leftarrow$ required throughput by D_u
[7]: $L(u) \leftarrow$ min. latency required by D_u
[8]: $R(u) \leftarrow \{C(u), L(u)\}$
[9]: **if** $R(u) \leq R_i$ **then**
[10]: Assign $R(u)$ to u
[11]: $R_i \leftarrow R_i - R(u)$
[12]: $R_{S'} \leftarrow R_{S'} - R(u)$
else
[13]: **if** $R(u) \leq R_{S'}$ **then**
[14]: RP over DC for S'
else
[15]: Reallocate resources for S
end if
end if
end foreach
end

This algorithm works as follows: one user u is chosen from the set of active users in the scenario (Line 1). User u is connected to the BN following an SINR schedule (Line 2). Consequently, a candidate BS is chosen (Line 3). Depending on the type of service, u is assigned to a slice by the orchestrator. This slice is referred to as S' (Line 4). When u generates traffic demand D_u (Line 5), two metrics are extracted: the necessary throughput $C(u)$ (Line 6) and the minimum latency required to establish the connection $L(u)$ (Line 7). The tuple that collects both metrics is denoted as $R(u)$ (Line 8). If the resources available by BS i (R_i) are greater than or equal to those needed to satisfy the demand (Line 9), the resources that u requests $R(u)$ are assigned to it (Line 10), and the connection is established. The resources available for BS i are updated (Line 11). Consequently, the resources available for slice S' are also updated (Line 12). In both cases, the resources allocated to u are subtracted. If the BS i had no available resources, i.e. $R(u) > R_i$, the available resources are compared by slice S' (Line 13). If the slice S' has resources available to satisfy $R(u)$, an inter-slice mechanism is triggered (Line 14). This mechanism is Resource Pool over Dual Connectivity, explained in Subsection V-C. If it has no resources, an inter-slice mechanism is triggered for all slices hosted in BS i (Line 15). This mechanism is based on resource reallocation using priority-based schedules, explained in Subsection V-D.

The complexity of this algorithm is also $O(n^2+n)$, for each u in U. Similar to the base case, the connection of user u to the BN has a complexity of $O(n^2 + n)$. The worst case from the point of view of complexity would be the RP over DC for all base stations belonging to S'. The operation of that algorithm that adds the highest complexity is the connection of u to BN', i.e., BSs belonging to S'. In the worst case, all the BSs in the scenario belong to S'. Thus, the complexity of RP over DC is $O(n^2 + n)$.

C. RAN-CENTRIC MECHANISMS FOR INTER-SLICE RESOURCE SHARING: RESOURCE POOL OVER DUAL CONNECTIVITY

The BN is logically divided into slices, hosted in portions of the BSs. The capacity and latency available per slice are distributed among the individual capacities and latencies of each BSs. Therefore, the BSs resources are shared within the same slice. These shared resources must be available through the UEs in several ways. This logical storage of all available resources is known as Resource Pool (RP). It is a logical container where all available capacity or latency is logically stored, so that a UE which needs it can use it, depending on the decision of the orchestrator.

This logical use of resources must translate into physical access to the RAN. BSs can communicate with each other via non-ideal backhaul to determine the resources that are available. As if the objective is to make this process transparent to the UE, the slice orchestrator must take care of the decision making, and the BN should take care of the rest of the process.

The method proposed in this study to share resources with the UE is to open a secondary link with the BS that can provide the required resources. The slice orchestrator must know the UE needs, and if the BS to which it is connected does not have sufficient resources, it must find one (by SINR planning) that can serve its demands. The selected BS forces the orchestrator to open a secondary link to the UE. This secondary link will have dedicated resources demanded by the UE. The operation of these mechanisms is shown in Figure 6 (b).

DC has been selected as the enabling physical technology for this resource sharing process. UE demands are split between the two links to minimise the number of blocked

connections. When the UE needs to change, the orchestrator rearranges the link configuration.

D. RAN-CENTRIC MECHANISMS FOR INTRA-SLICE RESOURCES REALLOCATION

The orchestrator must reallocate resources between the different slices to serve high-critically communications. This occurs when the slice with the high-priority UE connected has no available resources. Two algorithms are proposed for the proportional allocation of these resources between different slices. These mechanisms penalise lower priority connected UEs by reallocating resources, to free up resources for the candidate UE. The lower the priority of the UEs (i.e. the less critical the type of communication), the more resources it will give to the higher priority UE.

The solutions mentioned are detailed below and are shown in Figure 6 in part (a).

1) OUTWEIGH PRIORITY ALLOCATION

Suppose a UE $u \in \{U\}$ with high-priority P_u connected to slice S_u needs bandwidth BW_u, and the inter-slice methods cannot ensure their allocation. The BS to which it is connected has an available bandwidth $BW' < BW_u$, reserved for S_u. The other connected UEs, with priorities $p_n \in \{P\}$ are use the remaining bandwidth, assigned to other slices. To compensate for the bandwidth required by the highest priority UE, resources are subtracted proportionally from the other slices, depending on the priorities of the other connected UEs, as shown in Equation 12. A scheme of the algorithm operation can be seen in Figure 7. This transfer is proportional to the weighted inverse averaging method [46].

$$BW_u = BW' + \sum_{n=0}^{N'} BW_n \cdot f(n, P) \tag{12}$$

where N' are the same- or lower-priority UEs and $f(n, P)$ is the priority factor, which is defined as the portion of resources that UEs n contributes according to the total priorities P:

$$f(n, P) = \frac{\sum_{i=0}^{N'} P_i}{P_n \cdot \sum_{i=0}^{N'} \frac{\sum_{j=0}^{N'} P_j}{P_i}} \tag{13}$$

This technique allows the algorithm to balance the resources available in the slice. This is used so that critical communications can be handled without depriving other lower-priority UEs of resources, even if the candidate UE cannot obtain all the necessary resources.

2) DISCARD PRIORITY ALLOCATION

This technique follows the same philosophy as the previous one but ensures that the candidate UE obtains all required resources, at the cost of discarding other lower priority UE by requesting their reconnection to the network. When a UE attempt to connect to a BS, the slice orchestrator recalculates the necessary bandwidth for all the other UEs already connected to the same BS but in different slices to assign the necessary bandwidth to the candidate UE. After this reassignation, the orchestrator disconnects from the BS all UEs who have left with fewer resources than required for their type of service. Instead of decreasing the QoS of these UE, as in the previous algorithm, it discards them. When a UE is discarded, the orchestrator requests other BSs in the slice to reconnect the UE. In this way, the slice ensures that the UEs of the most critical communications (i.e. those with the highest priority) will always obtain the necessary resources. However, very low priority UEs, who are less affected by reconnection owing to their low criticality, are reallocated to other BSs in the slice.

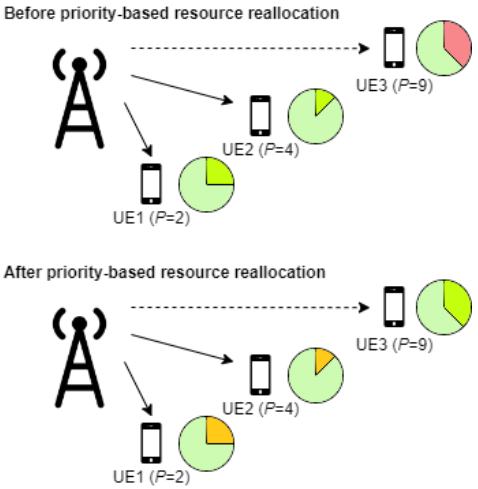


FIGURE 7. Operation of Outweigh priority allocation algorithm. The resources that cannot be assigned (red) to a higher priority user are reallocated by taking away resources from other lower priority users (orange). Resources in green are those that remain allocated to UEs.

VI. SIMULATION RESULTS AND DISCUSSION

To test the performance of the developed algorithms, a dense urban environment has been assumed in the design of the simulation setup. To select the scenario parameters, the 3GPP recommendation for simulation scenarios based on Dense Urban environments [47] is assumed:

The BSs are distributed in the scenario following a Poisson Point Process and two levels of cells are considered. The backhaul links are mmWave technology that connects each small BS to the nearest Macro BSs, and there are no connections between small BSs. The UEs are random-distributed over the simulation scenario and move following the FF movement model. These UEs are divided into five groups, according to the type of service and the overall traffic they carry. This division is as follows:

- 80 UEs using feMBB services, as EHD VOD and gaming.
- 80 UEs using eURLLC services, as VOD, remote control, and FTP (V2X).
- 80 UEs using LDHMC services, as augmented and virtual reality.
- 80 UEs using ELPC services, as IoT devices.

TABLE 5. Cellular network parameters used in the simulations.

Parameter	Macro BS	Small BS	UE
Nº	4	16	480
Height	40 m	15 m	1.7 m
Nº antennas	8	36	4
Gain	12 dBi	8 dBi	5 dBi
TX power	28 dBm	12 dBm	-
Band	n41	n257	-
Carrier frequency	2.6 GHz	28 GHz	-
Available BW	200 MHz	3 GHz	-
Simulation time			60 minutes
Scenario size			10 Km^2
Propagation scenario	UMa and Umi random distributed		
Backhaul links mean capacity			10 Gbps

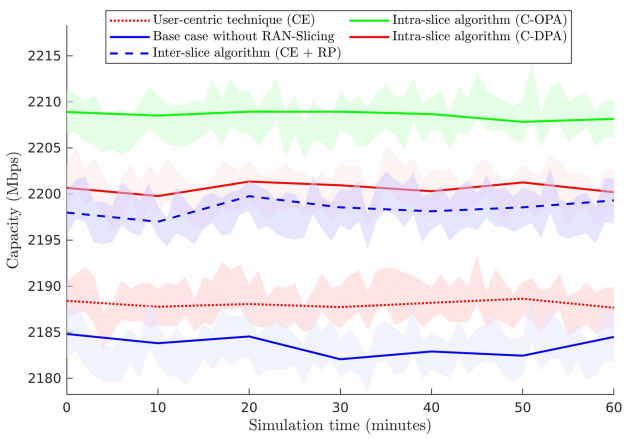


FIGURE 8. Average capacity offered by the network according to the proposed resource management algorithms, compared to the base case.

- 160 UEs using umMTC services, as smarthome and smartcities.

UEs with ELPC and umMTC services are fixed nodes disposed of in the scenario following a uniform distribution. The IoT and smart devices are assumed to be static devices in smart cities and environments.

The frequency planning used is obtained from [48]. The transmission and reception parameters have been chosen according to the frequencies used and cell type [49], considering new generation antennas [50]. The UEs parameterisations have been obtained from [51]. A summary of the selected parameters is provided in Table 5.

Latency- and capacity-focused slices are defined. Each slice hosts users that generate the traffic corresponding to each type of service. In the base case, bandwidth allocation per user is 10MHz without any resource management algorithm deployed.

To verify the performance of the proposed solutions, a set of simulations was performed to obtain the results and their subsequent processing. To minimise the randomness of some models, 35 simulations have been run with the same configuration for each test. Stochastic distributions for antenna positioning and user movement changes in each simulation. The results presented are the average values of all the executions with a confidence interval of 95%.

Network reliability analysis is performed by calculating the number of blocked connections for each type of service. Therefore, it is necessary to perform capacity and latency analyses for the entire network. These latency and capacity results were translated into a number of blocked demands. As capacity and latency-focused slices have been defined, specific results have also been obtained for the types of service that mainly require these metrics (feMBB and eURLLC, respectively).

A. CAPACITY ANALYSIS

To calculate the variability of the capacity offered by the network using the proposed mechanisms, a capacity analysis was performed. In this analysis, the average measurements of the capacities offered by the network to the UE have been made, without discriminating the types of service.

Figure 8 shows the average capacity offered by the network to the UEs according to the algorithm used, compared to the base case without RAN-Slicing. It shows the measurements obtained with: (i) user-centric techniques, which in this case is capacity estimation (CE), (ii) chaining with inter-slice techniques based on Resource Pool (CE + RP), and chained algorithms with inter-slice techniques (iii) based on Chained Outweigh Priority Allocation (C-OPA) and (iv) based on Chained Discard Priority Allocation (C-DPA).

The tests performed show that when the resource estimation technique is used, the capacity increases by approximately 7 Mbps. When this technique is chained with the Resource Pool over DC (inter-slice), the increase is approximately 15 Mbps. If the proposed orchestration algorithm works in full chaining (with intra-slice algorithms), the results are 30 Mbps in the best case with C-OPA and 20 Mbps when using C-DPA (both approximate results).

B. LATENCY ANALYSIS

Under the same motivation as that of the capacity analysis, an average latency analysis was performed on the network. This latency is the delay time that base stations can offer to UEs, on average, regardless of the type of service. The general consideration of this analysis, as in the previous one, is that is does not discriminate between the type of service or dedicated slices. This considers the full performance of the network.

Figure 9 shows the average latency analysis performed with the different algorithms in the chain compared with the base case. The solutions obtained with latency estimation (LE), their chaining with inter-slice based techniques with Resource Pool (LE + RP) and the complete chaining of the proposed algorithm with C-OPA and C-DPA are presented.

A decrease in the average latency offered by BSs to the UEs was observed. This decrease as 6.6% when using latency-based resource estimation techniques (user-centric

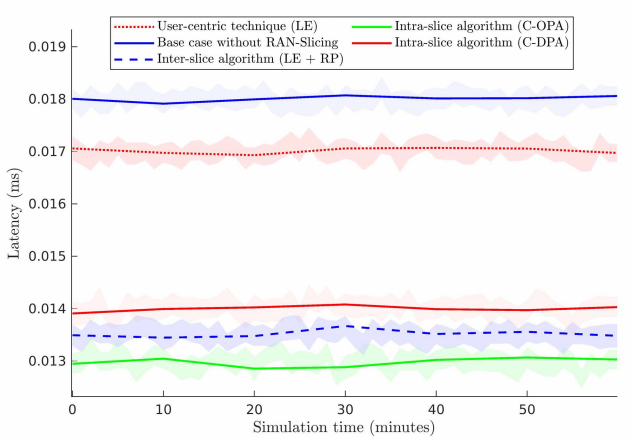


FIGURE 9. Average latency offered by the network according to the proposed resource management algorithms, compared to the base case.

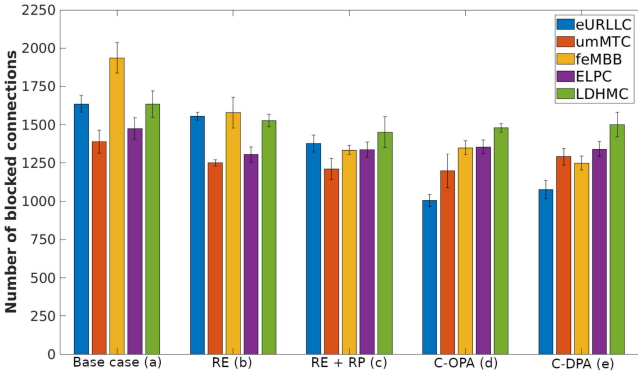


FIGURE 10. Number of blocked connections for each type of service depending on the resource management algorithm used. (a) Base case without RAN-Slicing, (b) resource estimation, (c) resource estimation and resource pool, (d) chained algorithm with Outweigh Priority Allocation, (e) chained algorithm with Discard Priority Allocation.

solutions) and 22.3% when these algorithms are chained with inter-slice-based ones (RP over DC), compared to the base case. If the proposed orchestration algorithm acts in full chaining, the decrease is higher, on the order of 25% with C-DPA, and up to 35.5% with C-OPA.

These results translate into a decrease in blocked connections when user demands are considered, which in turn means an increase in network reliability in average terms.

C. RELIABILITY ANALYSIS

Reliability is defined as the number of errors in communication. The more errors there are, the less reliable the system is. To assess the reliability of the network with the proposed solution deployed in the RAN, a measurement of blocked connections has been performed in order to obtain the number of error in the traffic flows, i.e., blockages in connections. A connection suffers from blockage when the links it passes through do not have sufficient available resources to meet demands. These link resources are allocated by the BS (or slice) at the time of connection establishment, depending on the available resources.

The same simulations have been carried out to measure the same metrics, but to modifying the resource management mechanism. The results, shown in Figure 10, are compared to those of the resource management algorithm in the base case without RAN-Slicing.

The results show the number of connections blocked by the network for each user type. As traffic is generated according to stochastic distributions, an average of the number of blocked connections is considered. At best, the simulations show a reduction of 38.46% for eURLLC, 12.5% for umMTC, 21.87% for feMBB, 11.86% for ELDP and 11.76% for LDHMC communications. Therefore, the reliability of the network increase.

These results are a consequence of the improvements observed in the analyses described in the previous subsections. The more the capacity increases or the more the latency decreases, the more the number of blocked connections decreases, and therefore the reliability of the network increases. Service types that require metrics for which slices have been dedicated (feMBB and eURLLC) show further improvements over those already shown for other service types that have not been assigned dedicated slices. Higher priority communications benefit the most from intra-slice algorithms. These algorithms prioritise the establishment of the highest priority connections over others, and reserve more resources for them.

D. PERFORMANCE FOR feMBB AND eURLLC

The performance tests to assess the main metrics of this study have been carried out with a focus on feMBB and eURLLC type users because slices focusing on latency and capacity have been defined. Figure 11 shows the results of the capacity, latency and reliability metrics according to the algorithms used and their concatenation, compared to the base case. The metrics chosen were those used by the IMT for the comparison of different generations of cellular networks [6].

Latency penalisation is observed when capacity is increased and, in contrast, capacity penalisation is observed when latency decreases. This is because the estimation algorithm finds the BS that provides the best metric in each case, without considering the other. This trade-off limits the performance of the proposed method. Another performance bound is the impossibility of inter- and intra-slice algorithms to be launched simultaneously. The capacity and latency resources available by a BS are accessed by inter-slice algorithms when they are triggered. If these resources are being reallocated to another slice, the inter-slice algorithms cannot make use of them. To avoid collisions, the chaining operation of the algorithms must be limited.

The proposed orchestration algorithm can achieve a compromise between latency and capacity when inter-slice (using RP over DC) and intra-slice techniques (using OPA) are

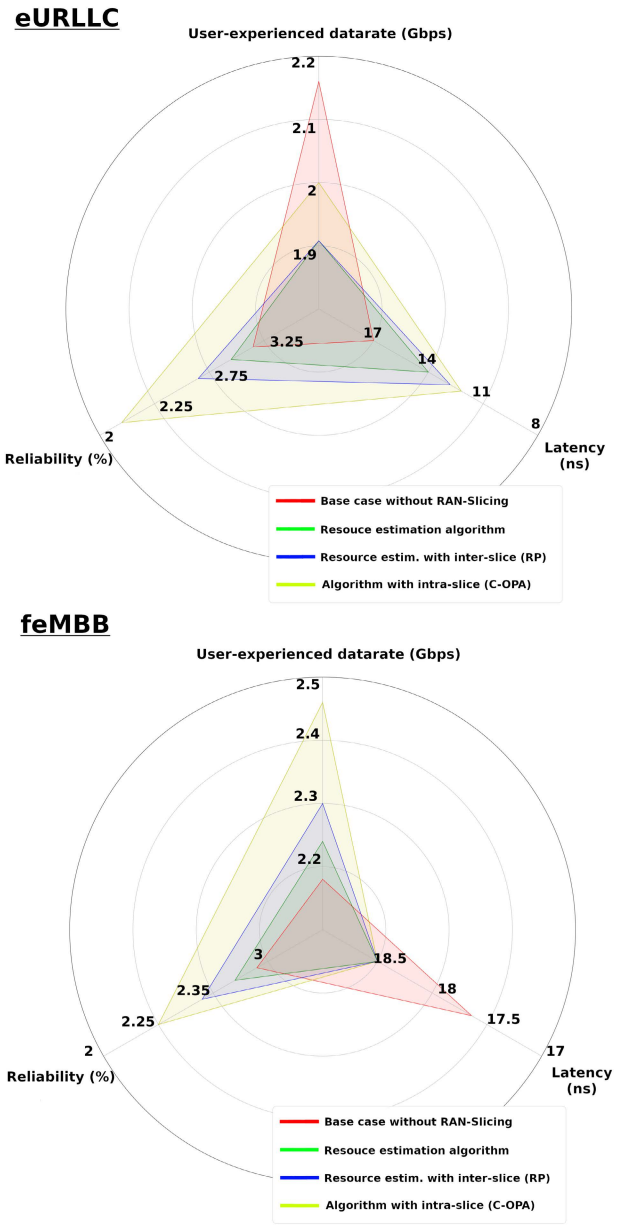


FIGURE 11. Metrics evaluated and their quantification according to the algorithm used for feMBB and eURLLC type users.

chained, which results in an increase in network reliability. Between the base case and the proposed full solution, improvements of 11.2% and 35.3% were observed for capacity and latency, respectively, and blocked connections and reliability showed a reduction of more than a third for eURLLC. For feMBB services, the improvements are 15.8% for capacity, 4% for latency, and a reduction of more than a quarter in reliability.

VII. CONCLUSION

In this paper, we propose a chained orchestration algorithm for resource management in RAN-Slicing applied to 6G cellular networks, focusing on ultra-reliable communications. This orchestration algorithm is based on the concatenation of resource estimation techniques, inter-slice techniques and intra-slice resource reallocation mechanisms. Resource estimation techniques use channel information to calculate the latency and capacity available for users demands. The proposed inter-slice technique makes employ Resource Pool over Dual Connectivity. The proposed intra-slice mechanisms reallocate the available capacity and latency across all slices to meet the demands of a specific slice.

Several studies have been conducted to test the performance of this algorithm from a capacity and latency point of view, resulting in a reliability study, which will be highlighted in this article. Our numerical results were compared with standardised baseline link planning and resource allocation for 5G-NR without RAN-Slicing. These results translate to a decrease in blocked connections. This leads to improved network performance in ultra-reliable 6G communications.

Focusing on the analysis of communications that have a dedicated slice to satisfy their demands (eURLLC and feMBB), latency penalisation is observed when capacity is increased and vice versa because estimation algorithms look for the BS that provides the best metric in each case, without taking into account the other one. The proposed orchestration algorithm can achieve a compromise between latency and capacity when inter-slice and intra-slice techniques are used, resulting in an increase in network reliability. Depending on the traffic requirements, i.e. the type of service, blocked connections decreased by 38.46% for eURLLC and 21.87% for feMBB.

Future work will focus on offering slices to different types of services and on different metrics (power consumption, mobility management, connection density, etc). The impact of the number of handovers on different slices, as well as the dynamic deployment of slices on demand, will also be studied.

The use of genetic algorithms to find a suboptimal solution to the problem will also be a task to be developed in the future. Additionally, linear optimisation is performed to search for the optimal solution. These proposals would be complemented with the search for the optimality gap between the current proposal to maximise the performance of these solutions. These problems could add a recombination of the chaining of the proposed algorithms and different types of constraints in addition to the priority of the type of service.

REFERENCES

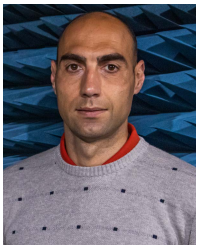
[1] *Digital 2021: Global Overview Report*, Statista, 2021.
[2] A. Schmidt, T. Lio, J. Reers, and A. Regalia, "The electric vehicle—More than a new powertrain," Industry X Accenture, Santiago, Chile, Tech. Rep., 2021.
[3] T. Avot, "5G on the highway to V2X," Altran Technol., Paris, France, Tech. Rep., 2020.
[4] *Forecast Shipments of Virtual and Augmented Reality Headsets Worldwide From 2019 to 2024*, Data Reportal, 2021.

[5] S. J. Nawaz, S. K. Sharma, S. Wyne, M. N. Patwary, and M. Asaduzzaman, "Quantum machine learning for 6G communication networks: State-of-the-art and vision for the future," *IEEE Access*, vol. 7, pp. 46317–46350, 2019.
[6] Z. Zhang, Y. Xiao, Z. Ma, M. Xiao, Z. Ding, X. Lei, G. K. Karagiannidis, and P. Fan, "6G wireless networks: Vision, requirements, architecture, and key technologies," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 28–41, Sep. 2019.
[7] P. Yang, Y. Xiao, M. Xiao, and S. Li, "6G Wireless communications: Vision and potential techniques," *IEEE Netw.*, vol. 33, no. 4, pp. 70–75, Jul./Aug. 2019.
[8] W. Guan, H. Zhang, and V. C. M. Leung, "Customized slicing for 6G: Enforcing artificial intelligence on resource management," 2021, *arXiv:2102.10498*.
[9] X. Fu, Q. Shen, W. Wang, H. Hou, and X. Gao, "Slice merging/spliting operations and tenant profit optimization across 5G base stations," *IEEE Access*, vol. 9, pp. 9706–9718, 2021.
[10] P. Korrai, E. Lagunas, S. K. Sharma, S. Chatzinotas, A. Bandi, and B. Ottersten, "A RAN resource slicing mechanism for multiplexing of eMBB and URLLC services in OFDMA based 5G wireless networks," *IEEE Access*, vol. 8, pp. 45674–45688, 2020.
[11] J. Mei, X. Wang, K. Zheng, G. Boudreau, A. B. Sediq, and H. Abou-Zeid, "Intelligent radio access network slicing for service provisioning in 6G: A hierarchical deep reinforcement learning approach," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 6063–6078, Sep. 2021.
[12] L. Nadeem, M. A. Azam, Y. Amin, M. A. Al-Ghamdi, K. K. Chai, M. F. N. Khan, and M. A. Khan, "Integration of D2D, network slicing, and MEC in 5G cellular networks: Survey and challenges," *IEEE Access*, vol. 9, pp. 37590–37612, 2021.
[13] C. Sexton, N. Marchetti, and L. A. DaSilva, "Customization and trade-offs in 5G RAN slicing," *IEEE Commun. Mag.*, vol. 57, no. 4, pp. 116–122, Apr. 2019.
[14] R. Su, D. Zhang, R. Venkatesan, Z. Gong, C. Li, F. Ding, F. Jiang, and Z. Zhu, "Resource allocation for network slicing in 5G telecommunication networks: A survey of principles and models," *IEEE Netw.*, vol. 33, no. 6, pp. 172–179, 2019.
[15] S. E. Elayoubi, S. B. Jemaa, Z. Altman, and A. Galindo-Serrano, "5G RAN slicing for verticals: Enablers and challenges," *IEEE Commun. Mag.*, vol. 57, no. 1, pp. 28–34, Jan. 2019.
[16] T. C. Chuah and Y. L. Lee, "Intelligent RAN slicing for broadband access in the 5G and big data era," *IEEE Commun. Mag.*, vol. 58, no. 8, pp. 69–75, Aug. 2020.
[17] A. Dogra, R. K. Jha, and S. Jain, "A survey on beyond 5G network with the advent of 6G: Architecture and emerging technologies," *IEEE Access*, vol. 9, pp. 67512–67547, 2021.
[18] J. García-Morales, M. C. Lucas-Estañ, and J. Gozalvez, "Latency-sensitive 5G RAN slicing for industry 4.0," *IEEE Access*, vol. 7, pp. 143139–143159, 2019.
[19] Y. Han, S. E. Elayoubi, A. Galindo-Serrano, V. S. Varma, and M. Messai, "Periodic radio resource allocation to meet latency and reliability requirements in 5G networks," in *Proc. IEEE 87th Veh. Technol. Conf. (VTC Spring)*, Jun. 2018, pp. 1–6.
[20] I. Vila, J. Perez-Romero, O. Sallent, and A. Umbert, "Characterization of radio access network slicing scenarios with 5G QoS provisioning," *IEEE Access*, vol. 8, pp. 51414–51430, 2020.
[21] B. Han, J. Lianghai, and H. D. Schotten, "Slice as an evolutionary service: Genetic optimization for inter-slice resource management in 5G networks," *IEEE Access*, vol. 6, pp. 33137–33147, 2018.
[22] M. Zambianco and G. Verticale, "Interference minimization in 5G physical-layer network slicing," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4554–4564, Jul. 2020.
[23] M. Maule, J. Vardakas, and C. Verikoukis, "5G RAN slicing: Dynamic single tenant radio resource orchestration for eMBB traffic within a multi-slice scenario," *IEEE Commun. Mag.*, vol. 59, no. 3, pp. 110–116, Mar. 2021.
[24] K. Xiong, S. Samuel Rene Adolphe, G. O. Boateng, G. Liu, and G. Sun, "Dynamic resource provisioning and resource customization for mixed traffics in virtualized radio access network," *IEEE Access*, vol. 7, pp. 115440–115453, 2019.
[25] Y. Tsukamoto, R. K. Saha, S. Nanba, and K. Nishimura, "Experimental evaluation of RAN slicing architecture with flexibly located functional components of base station according to diverse 5G services," *IEEE Access*, vol. 7, pp. 76470–76479, 2019.
[26] *Study on 3D Channel Model for LTE (Release 12), v12.1.0*, document 3GPP TR 36.873, Mar. 2015.
[27] S. Sun, T. S. Rappaport, T. A. Thomas, A. Ghosh, H. C. Nguyen, and I. Z. Kovács, "Investigation of prediction accuracy, sensitivity, and parameter stability of large-scale propagation path loss models for 5G wireless communications," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 2843–2860, May 2016.
[28] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Pers. Commun.*, vol. 6, no. 3, pp. 311–335, Mar. 1998.
[29] D. A. Chekired, M. A. Togou, L. Khoukhi, and A. Ksentini, "5G-slicing-enabled scalable SDN core network: Toward an ultra-low latency of autonomous driving service," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 8, pp. 1769–1782, Aug. 2019.
[30] *LTE Physical Layer Framework for Performance Verification*, document 3GPP TSG-RAN1 #48 R1-070674, Feb. 2007.
[31] J. Navarro-Ortiz, P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "A survey on 5G usage scenarios and traffic models," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 905–929, 2nd Quart., 2020.
[32] M. Carrascosa and B. Bellalta, "Cloud-gaming: Analysis of Google stadia traffic," 2020, *arXiv:2009.09786*.
[33] X. Wang, T. Kwon, Y. Choi, M. Chen, and Y. Zhang, "Characterizing the gaming traffic of world of warcraft: From game scenarios to network access technologies," *IEEE Netw.*, vol. 26, no. 1, pp. 27–34, Jan. 2012.
[34] J. Wang, Y. Shao, Y. Ge, and R. Yu, "A survey of vehicle to everything (V2X) testing," *Sensors*, vol. 19, p. 334, Jan. 2019.
[35] C. Storck and F. Duarte-Figueiredo, "A 5G V2X ecosystem providing internet of vehicles," *Sensors*, vol. 19, no. 3, p. 550, Jan. 2019.
[36] A. S. Ibrahim, K. Y. Youssef, H. Kamel, and M. Abouelatta, "Traffic modelling of smart city Internet of Things architecture," *IET Commun.*, vol. 14, no. 8, pp. 1275–1284, May 2020.
[37] *LTE Physical Layer Framework for Performance Verification*, document 3GPP TSG-RAN1#48 R1-070674, Feb. 2007.
[38] Y. Al Mtawa, A. Haque, and B. Bitar, "The mammoth internet: Are we ready?" *IEEE Access*, vol. 7, pp. 132894–132908, 2019.
[39] F. Hu, Y. Deng, W. Saad, M. Bennis, and A. H. Aghvami, "Cellular-connected wireless virtual reality: Requirements, challenges, and solutions," *IEEE Commun. Mag.*, vol. 58, no. 5, pp. 105–111, May 2020.
[40] S. Friston, E. Griffith, D. Swapp, C. Lrondi, F. Jjunju, R. Ward, A. Marshall, and A. Steed, "Quality of service impact on edge physics simulations for VR," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 5, pp. 2691–2701, May 2021.
[41] *3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on RAN Improvements for Machine-Type Communications (Release 11)*, document 3GPP TR 37.868 V11.0.0, Sep. 2011.
[42] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Aug. 2016.
[43] M. Suer, C. Thein, H. Tchouankem, and L. Wolf, "Multi-connectivity as an enabler for reliable low latency communications—An overview," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 156–169, 1st Quart., 2020.
[44] *Technical Specification Group Radio Access Network: Evolved Universal Terrestrial Radio Access (E-Utra) and Evolved Universal Terrestrial Radio Access Network (E-Utran): Overall Description; Stage 2 (Release 12)*, document 3GPP TS 36.300, Jun. 2016.
[45] *5G-Nr (Release 15)*, document 3GPP TR 21.915, Mar. 2015.
[46] J. Hartung, G. Knapp, and B. K. Sinha, *Statistical Meta-Analysis With Applications*. Hoboken, NJ, USA: Wiley, 2008.
[47] *3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on Self-Evaluation Towards IMT-2020 Submission (Release 16) v16.0.0*, document 3GPP TR 37.910, Jun. 2019.
[48] *NR; User Equipment (UE) Radio Transmission and Reception; Part 1: Range 1 Standalone (Release 17), v17.4.0*, document 3GPP TS 38.101-1, Jan. 2022.
[49] N. Lassoued, N. Boujnah, and R. Bouallegue, "Reducing power consumption in C-RAN using switch on/off of MC-RRH sectors and small cells," *IEEE Access*, vol. 9, pp. 75668–75682, 2021.
[50] B. Aqlan, M. Himdi, H. Vettikalladi, and L. Le-Coq, "A circularly polarized sub-terahertz antenna with low-profile and high-gain for 6G wireless communication systems," *IEEE Access*, vol. 9, pp. 122607–122617, 2021.
[51] J. Lee, H. Kim, and J. Oh, "Large-aperture metamaterial lens antenna for multi-layer MIMO transmission for 6G," *IEEE Access*, vol. 10, pp. 20486–20495, 2022.

JOSE J. RICO-PALOMO was born in Badajoz, Spain, in 1996. He received the B.Sc. degree in telecommunications engineering, specializing in telematics engineering, from the Centro Universitario de Mérida of University of Extremadura, Spain, in 2017, and the M.Sc. degree in telecommunications engineering from the School of Technology of University of Extremadura, in 2018. He is currently pursuing the Ph.D. degree. In 2018, he completed a Research Fellowship at the CénitS Supercomputer Center. He is with the GITACA Research Group, Department of Computing and Telematics System Engineering, University of Extremadura. His current research interests include 5G and 6G communications, and propagation channels and models for next generation cellular networks and mobility management.

JESUS GALEANO-BRAJONES received the B.Sc. degree in telecommunication engineering, specialising in telematics, the B.Sc. degree in computer science engineering from the Universidad de Extremadura, Spain, in 2019, the M.Sc. degree (Hons.) in research in 2020. He is currently pursuing the Ph.D. degree with the Department of Computing and Telematics System Engineering, Universidad de Extremadura. He is a member of the GITACA Research Group. His research interests include optimisation algorithms, next-generation networks, artificial intelligence, and network security.

DAVID CORTES-POLO received the degree in computer science and the Ph.D. degree in telematics from the University of Extremadura, Spain, in 2015. He was a Research and Teaching Assistant at University of Extremadura, from 2011 to 2014. Since 2011, he has been the Network Manager with COMPUTAEX Foundation and the CénitS Center. His main research interests include IP-based mobility management protocols, performance evaluation, and quality of service support in future mobile networks.



JUAN F. VALENZUELA-VALDES was born in Marbella, Spain. He received the degree in telecommunications engineering from the Universidad de Málaga, Spain, in 2003, and the Ph.D. degree from the Universidad Politécnica de Cartagena, Spain, in May 2008. In 2004, he joined the Department of Information Technologies and Communications, Universidad Politécnica de Cartagena. In 2007, he joined the Head of Research with EMITE Ing. In 2011, he joined the Universidad de Extremadura, and in 2015, he joined the Universidad de Granada, where he is currently an Associate Professor. He was a Co-Founder of EMITE Ing.—a spin-off company. He also holds several national and international patents. His publication record is composed of more than 80 publications, including 40 JCR indexed articles, more than 30 contributions in international conferences and seven book chapter. His current research interests include wireless communications and efficiency in wireless sensor networks. He has also been awarded several prizes, including the National Prize to the Best Ph.D. in mobile communications by Vodafone and the I-Patents Award by the Spanish Autonomous Region of Murcia for innovation and technology transfer excellence.

JAVIER CARMONA-MURILLO received the Ph.D. degree in computer science and communications from the University of Extremadura, Spain, in 2015. From 2005 to 2009, he was a Research and Teaching Assistant. He has spent research periods with the Centre for Telecommunications Research, King's College London, U.K., and Aarhus University, Denmark.Since 2009, he has been an Associate Professor with the Department of Computing and Telematics System Engineering, Universidad de Extremadura. His current research interests include 5G networks, mobility management protocols, performance evaluation, and the quality of service support in future mobile networks.