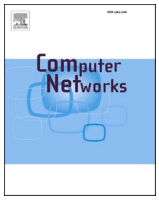
Computer Networks 243 (2024) 110287

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet



Relay-empowered beyond 5G radio access networks with edge computing capabilities

I. Vilà *, O. Sallent, J. Pérez-Romero

Signal Theory and Comunications Department of Universitat Politècnica de Catalunya (UPC), Barcelona, Spain

ARTICLE INFO

Keywords:
Beyond 5G
Edge computing
Relays

ABSTRACT

Relevant services envisaged for beyond 5G (B5G) systems, such as extended reality and holographic communications, have extremely demanding user experience requirements with significant computational and communication demands. While edge computing aims to address the computation requirements by offloading the computational tasks to edge servers near the user, the communication will take advantage of the technologies developed for 5G New Radio jointly with an never-before-seen degree of network densification. This paper proposes the use of relays with edge computing capabilities. The approach's potential for B5G are identified, and a system model is defined to characterize both computational and communications viewpoints. Based on this, results are provided to highlight the gains and limitations of the proposed approach from a system-level perspective. Finally, the main challenges for enabling relays with computing capabilities in B5G deployments are discussed.

1. Introduction

The evolution of communication networks Beyond 5G (B5G) and towards 6G is expected to meet the various and complex requirements of a wide range of vertical services [1,2]. For instance, applications based on eXtended Reality (XR) and holographic representations are expected to become a key asset in a broad range of scenarios, fusing the digital and the real worlds to provide end users with new experiences such as metaverse environments, immersive online gaming, real-time 3D communications, and so on. These emerging applications, which have stringent user experience requirements, are becoming more demanding in terms of both the computational and communication capabilities of B5G communication infrastructures.

To address the computation challenge, the conventional approach of offloading heavy tasks to powerful computing elements residing in the cloud (i.e., cloud computing) is no longer capable of meeting the latency requirements of such applications. Edge computing has rapidly evolved in response to these needs as a revolutionary paradigm that delivers computational power and resources closer to where the data is generated, significantly lowering response times with a much reduced carbon footprint [3].

In terms of communications requirements, 5G NR is comprised of multiple technological components (e.g., multiuser massive MIMO, smart beamforming), resulting in a large increase in the achievable spectral efficiency. However, to realize the promise of dramatically increased data rates (from Mbps to Gbps) and ultra-reliable low latency (from tens of milliseconds down to milliseconds), network densification has long been identified as an essential part of 5G network rollout [4]. The importance of network densification is exacerbated as high 5G frequency bands, which have worse propagation characteristics, are more integrated [5]. Millimeter wave (mmWave) signals at these frequencies exhibit reduced diffraction and more specular propagation than their microwave counterparts, making them far more susceptible to blockages [6]. As a result, huge capital expenditure (CAPEX) on the deployment of 5G infrastructure rollout will be necessary to meet the required capacity and coverage needs. To prevent financial strain, Mobile Network Operators (MNOs) must discover innovative and inventive ways of managing and deploying their 5G and beyond Radio Access Network (RAN) infrastructures.

In light of the foregoing, this paper advocates for B5G RAN deployments that make use of relay nodes with edge computing capabilities. Compared to the traditional edge computing vision, where edge servers are co-located with base stations, relays nodes equipped with computing servers will bring edge computing capabilities deeper into the RAN and thus closer to the user. Therefore, the envisaged solution will take advantage of relay nodes to further exploit the task offloading

* Corresponding author.
E-mail address: irene.vila.munoz@upc.edu (I. Vilà).

https://doi.org/10.1016/j.comnet.2024.110287
Received 9 October 2023; Received in revised form 17 January 2024; Accepted 23 February 2024
Available online 24 February 2024

benefits and as a mechanism to truly meet service requirements, particularly service continuity, which can be jeopardized due to poor coverage footprints. As a result, the synergy between relay-enhanced B5G RAN and edge computing can provide enhanced computation and communication capabilities for applications located at the boundary of MNOs' networks. In particular, this approach can bring several benefits in terms of e.g. reduction of both radio resource occupation and computational load at the base station as well as latency and power reductions.

The remainder of the paper is organized as follows. Section 2 summarizes the related work and highlights the novelties and contributions of this paper. Section 3 discusses the various options for a relay-empowered B5G RAN with edge computing capabilities, while Section 4 describes the corresponding system model. Section 5 contains some performance assessment results, highlighting the gains and limitations of edge computing-enabled relays. Section 6 examines different challenges of enabling relays with computing capabilities taking into account standardized architectures. Finally, Section 7 summarizes the conclusions and the future work.

2. Related work

Edge computing consists in placing computational infrastructure at the network edge. Edge servers can be either general-purpose servers (i.e., the same servers used on cloud environments), new platforms specifically designed for the edge requirements, or other platforms designed for specific use cases (e.g., automotive) [7]. The implementation of edge computing relies on virtualization technologies such as Network Function Virtualization (NFV), Information-Centric Networks (ICN) and Software-Defined Networks (SDN) [8]. There is an increasing number of emerging mobile applications that will benefit from edge computing by offloading their computation-intensive tasks to edge servers [8]. As identified in Hu et al. [9], potential applications include augmented reality, intelligent video acceleration, connected cars and IoT gateway.

Several research efforts have been made in the area of edge computing, as reflected in survey papers such as [7,10]. Moreover, many standardization activities are underway to support the deployment of edge computing with mobile networks [11]. The most relevant standardization activities are carried out by the ETSI Industry Specification Group (ISG) Multi-access Edge Computing (MEC), which has created an open and standardized IT service environment that allows third-party applications to be hosted at the edge, and by the Third Generation Partnership Project (3GPP), where various specification groups are working on the architectures that enable edge computing and its management. Moreover, the work by GSMA and 5G-PPP/6G IA (6G Smart Networks and Services Industry Association) focuses on setting the requirements and implementation agreements for edge computing.

The option of deploying relay stations to extend the coverage and capacity in cellular networks has been well considered in the literature for many years (see e.g., [12]), although practical implementation has been limited to rather specific use cases (e.g., extending coverage in a tunnel). However, the interest in relays has recently revamped, for example, with the Integrated Access and Backhaul (IAB) technology, which provides an alternative to fibre backhaul by extending 5G New Radio (NR) to support wireless backhaul [13,14]. Similarly, vehicle-mounted relays are considered in a recent study item in the 3GPP Release 18 [15] and in some previous works [16,17]. In turn, the capability of User Equipment (UE) to relay the traffic of another UE to/from the network is included by 3GPP as the UE-to-network relaying connectivity model of [18], identifying different scenarios, requirements and key performance indicators. In this respect, Pérez-Romero and Sallent [19] presented a vision of a B5G scenario where the UE actively complements the RAN infrastructure by acting as a relay, and thus empowering the RAN with enhanced flexibility to support different use cases.

The use of relays with computing capabilities is at an incipient stage. The ETSI MEC work item in [20] focuses on extending the edge computing platform to the far edge by incorporating computing capabilities to "constrained devices", which can be small cells, vehicles, UEs, flying objects such as drones, etc. These devices can incorporate relaying capabilities as well to serve others (e.g. through Device-to-Device (D2D) communications between UEs, drones, etc.). Besides, only a few works in the literature have proposed solutions for the use of relays with edge computing capabilities [21–26]. Among these, Liang et al. [21] and Chen et al. [22] consider the problem of task forwarding in cooperative wireless systems, where a task is sent from a source user to a destination user through a relay. In this context, the work in Liang et al. [21] proposes three different relay selection schemes that optimize the maximum transmission rate on the radio channel, the maximum computational capability and the total task computation delay (i.e., the delay encompassing the uplink (UL), the downlink (DL) transmission, and the computation of the task in the relay), respectively. The authors in Chen et al. [22] propose a solution that jointly optimizes the energy consumption and the delay by selecting the percentage of a task to be offloaded to the relay, the power allocation for the UL and DL and the computational resources allocation. In contrast to Liang et al. [21] and Chen et al. [22], which are designed for cooperative wireless systems scenarios, the scenario considered in [23–26] consists of a cellular network with Base Stations (BSs) and relays, where tasks can be offloaded to a relay or the BS as considered also here. The authors of Yao et al. [23] propose an energy optimization algorithm that selects the offloading mode of a user's task by choosing between its computation at the device, at one relay, at one BS, or at one BS connected through a relay. Instead, the work in Cao et al. [24] proposes a partial offloading scheme, where time-constraint tasks are divided into three parts. One part is sent to the relay, the other to the BS and the last one is computed locally in the user device. A protocol to send the three parts according to a time-division scheme is proposed and, then, the authors propose an energy consumption and computation time optimization algorithm that determines the task partition and the joint computation and radio resources allocation. In [25], a new protocol is proposed for the same task as in Cao et al. [24] but for the case where an Orthogonal Frequency-Division Multiple Access (OFDMA) scheme is considered. The proposed protocol allows the split tasks to be sent to the relay and the BS simultaneously for cooperative computation of the task. The associated resource allocation problem is formulated as a Mixed-Integer Programming (MIP) problem and solved by successive convex approximation. The work in Hu et al. [25] is extended in Luo and Huang [26], where a task offloading and computation strategy is designed for scenarios where reconfigurable intelligent surface (RIS)-aided OFDM-Non-Orthogonal Access (NOMA) schemes are used. For this case, a new protocol is proposed for relaying tasks and the associated resource allocation problem is formulated as a MIP and solved by various optimization techniques.

While the previous works have mainly focused on the algorithmic design of task forwarding solutions in [21,22] and task offloading solutions [23–26], our recent work in Vilà et al. [27] is the first to provide a quantitative assessment of the benefits of incorporating relays with computing capabilities in B5G RAN deployments from a joint communication and computation perspective. To this end, the work in Vilà et al. [27] characterizes the communication model by considering 5G NR parameters, while works [21–26] did not consider any specific standardized radio technology.

This paper builds upon our previous work [27] by providing an upgraded and more complete view of the benefits and challenges of incorporating relays with computational capabilities. This is attained through two main contributions. The first one is the extension of the system model presented in the previous work [27] by including the characterization of the power consumption. The main motivation for this is that sustainability is a relevant aspect for 5G systems, but it is expected to become a priority for B5G deployments aiming at green communications with reduced carbon footprint [28]. Therefore,

quantifying the reduction in power consumption that can be achieved through the use of relays with edge computing capabilities becomes a relevant aspect to justify the need for this type of computational-enabled nodes. This quantification is included in this paper for different configuration parameters of the base station and the relay available in the literature, reflecting different implementations. The second contribution is the discussion of the challenges involved in including edge computing-enabled relays of different types in the 3GPP architecture. The future realization of relays with computing capabilities needs to be aligned with the standardization activities in 3GPP and ETSI MEC. Therefore, this paper intends to provide some insights in the realization of the concept from an architectural and functional perspective.

3. Relays with edge computing capabilities in B5G RAN

The use of relays with computational capabilities brings several benefits over the scenario where computing resources are only available in the BS. First, the computational load of the BS can be reduced since some of the computations of users in the BS area would be performed in the relays. Second, the load on the radio channel between the relay and the BS can also be reduced since all the traffic generated to offload the tasks to the edge server in the BS through the relay will be cut off at the relay. Third, in the case of time-constrained tasks, offloading them in the relay can improve the delay associated with the computation of the task in the edge (i.e., embracing the upload of the task to the edge server through the UL, its computation and the download of its result through the DL) since usually the radio conditions of the channel between the user and the relay will be better than the one with the BS (i.e., due to closer distances of the user with the relay). Fourth, the power consumption in the system can also be reduced as generally relays consume less power due to the better radio conditions of the users towards the relay and the reduced load of the channel between the relay and the BS.

The upper part of Fig. 1 illustrates the benchmark scenario, where a BS is provided with edge computing capabilities and various UEs can connect to the BS via 5G NR air interface. The computing capabilities at the BS enable more responsive service provisioning to the UEs. The lower part of Fig. 1 depicts the envisaged B5G scenario, which includes different types of relays. Providing edge computing capabilities to each type of relay embraces different considerations, as elaborated in the following.

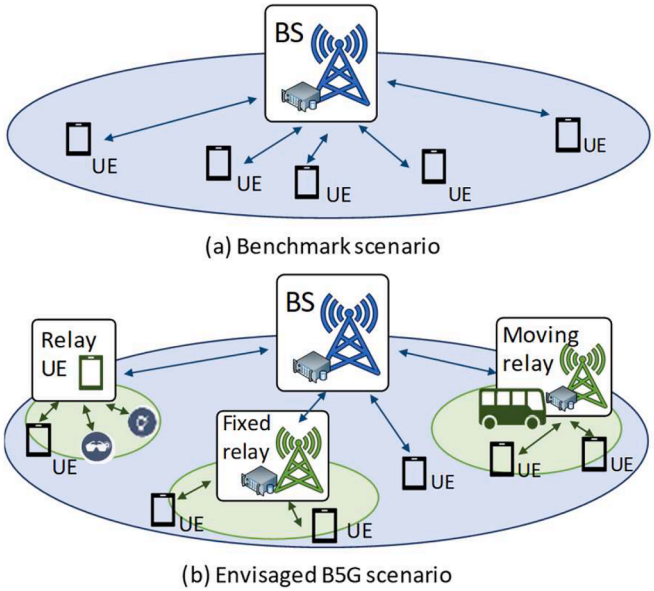


Fig. 1. Benchmark and envisaged B5G scenarios.

- *Fixed relay*. In this case, the IAB solution can be leveraged [29], enabling a fast and flexible deployment of new IAB nodes. The BS, referred to as IAB-donor, serves relay nodes, referred to as IAB-nodes, and other UEs that are directly connected to it, considering 5G NR for all links. Given the fixed nature of the relay node, its deployment needs to be associated with a planning and dimensioning process, both from communication and computing perspectives. From the communication side, the deployment could target the improvement of coverage (i.e., to provide coverage extension within the BS's service area) and/or capacity (i.e., to deploy the relay closer to a traffic hotspot). The latter case can also motivate the allocation of computing resources at the relay, e.g. to properly serve the computational needs of the users in the traffic hotspot. Considering that this embraces some costs to the MNO, edge computing capabilities should be properly dimensioned depending on the expected number of UEs and applications benefiting from task offloading over the relay node.
- *Moving relay*. To satisfy highly demanding user experience requirements in mobile environments, such as trains and buses, the development and deployment of mobile relays are envisaged. Onboard mobile relays at the vehicles enable efficient access for the in-vehicle UEs through wireless backhaul links [30]. Several advantages of mobile relaying have been identified: the reduction of high vehicle penetration loss up to 20–35 dB, the avoidance of the signalling storm problem due to group handover, etc. [31]. The interest in the deployment of moving relays with edge computing capabilities would be closely related to some use cases. For instance, passengers on a tour bus could view immersive content projected onto the front window of the vehicle, superimposed on the landscape or monuments they observe while touring [32]. Another example is an autonomous tram [33]. During its daily service, a tram equipped with an Obstacle Detection and Tracking (ODT) system continuously scans the track area in the front of the vehicle to search for potential collision objects. The onboard computing platform collects raw data from sensors, like radars, laser scanners or cameras, and then synchronizes and associates them to the possible target tracks. A moving relay could provide a reliable tram-to-ground connection to send warnings/alarms to the Operations Control Center (OCC). These use cases again embrace some dimensioning exercise to determine the amount of radio and computing resources to be allocated to the onboard relay node.
- *Relay UE*. This case is sustained in the support of D2D communications, in which two UEs in proximity can directly communicate. For D2D operation, 3GPP defined the PC5 interface between UEs sustained on a new radio link for direct transmissions between devices, denoted as sidelink. The vision of UEs acting as relays, as proposed in Pérez-Romero and Sallent [19], includes the necessary mechanisms and intelligence at the MNO's service management and orchestration (SMO) layer to embrace relay UEs as an integral part of the so-called augmented RAN. Using UEs as relays can be the most appealing use case for MNOs since the communication and computing resources are leveraged by the users themselves. A relay UE would exploit its communication capabilities to transmit/receive traffic from/to another UE to/from the BS and its computing capabilities to offload tasks from another UE and/or the BS. Certainly, this is also the most challenging use case from technical and business perspectives. For the former, the specification and development of certain management functions and corresponding interfaces would be required. For the latter, proper incentive mechanisms should be developed by MNOs to attract customers and motivate through win-win mechanisms their willingness to contribute with their devices to the augmented RAN vision.

4. System model

Let us consider the system depicted in Fig. 2 with a BS and a relay,

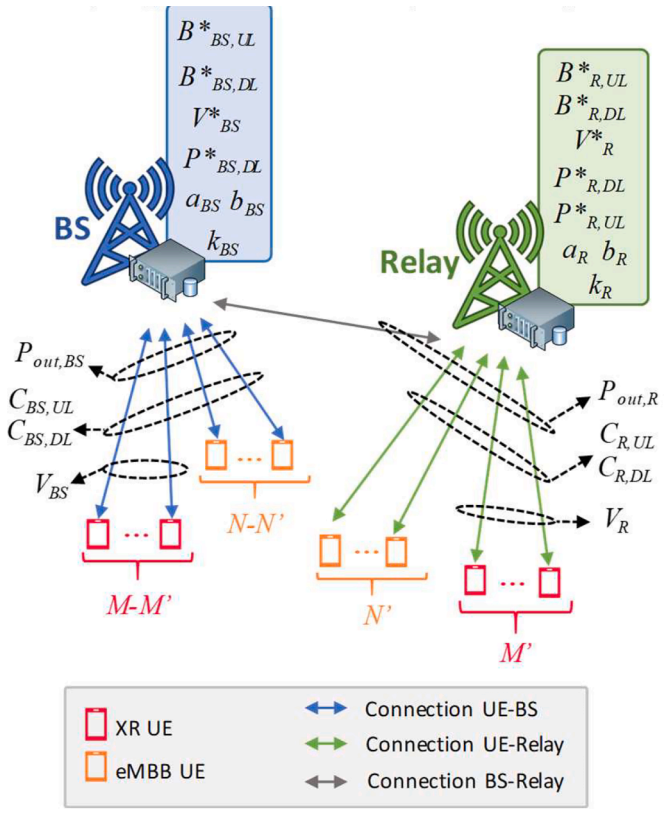


Fig. 2. System model.

which can be a fixed relay, a moving relay or a relay UE. The BS and the relay operate with 5G NR technology and embed edge computing capabilities. In the BS coverage area, there are M UEs that generate computationally intensive tasks subject to be offloaded at any edge computing platform and N UEs that only require 5G connectivity. For example, the first type of UEs could be provided with XR services, which send XR tasks such as rendering tasks that can not be computed locally on the device to the edge. In turn, examples of the second type of UEs could be those having an enhanced Mobile Broadband (eMBB) service for Internet access. Without loss of generality in the context of this paper, UEs generating computationally intensive tasks are referred to as "XR UEs" while UEs with only 5G connectivity requirements are referred to as "eMBB UEs".

At a certain point in time, M' and (M–M') XR UEs exploit edge computing capabilities at the relay and the BS, respectively. In turn, there are N' eMBB UEs that get connected through the relay using two hops (UE-Relay and Relay-BS) and the remaining (N–N') UEs are connected via a direct link with the BS.

In the following, the considered task, computation, communication and power consumption models in the system are detailed. The basic acronyms and notations of the characterized system model are summarized in Table 1.

4.1. Task model

XR UEs with computationally intensive tasks generate non-divisible tasks. For the i-th XR UE, tasks are generated according to a certain probability distribution with mean λ_i (tasks/s). A task is characterized by its length L_i (bits) and by the length of the result L_i' (bits). The computation of the task requires a number of floating point operations (FLOP), O_i, to be completed in a maximum delay time $D_{\max,i}$ (s).

Table 1
List of acronyms and notations.

Acronyms	
BS	Base Station.
DL	Downlink.
eMBB	enhanced Mobile Broadband.
FLOP	Floating Point Operations.
FLOPS	Floating-point operations per second.
MCS	Modulation and Coding Scheme.
NR	New Radio.
RAN	Radio Access Network.
RF	Radio Frequency.
RI	Rank Indicator.
SINR	Signal to Interference and Noise Ratio.
UE	User Equipment.
UL	Uplink.
XR	eXtended Reality.
Notations	
$B^*_{BS,DL}$	Total available bandwidth in the BS for the DL in Hz.
$B^*_{BS,UL}$	Total available bandwidth in the BS for the UL in Hz.
$B^*_{R,DL}$	Total required bit rate capacity at the BS for the DL in bps.
$B^*_{R,UL}$	Total available bandwidth in the relay for the UL in Hz.
B_i	Bandwidth assigned to the i-th link in Hz.
$C_{BS,DL}$	Total available bandwidth in the relay for the DL in Hz.
$C_{BS,UL}$	Total required bit rate capacity at the BS for the UL in bps.
CP	Inefficiency factor due to the cyclic prefix.
$C_{R,DL}$	Total required bit rate capacity at the relay for the DL in bps.
$C_{R,UL}$	Total required bit rate capacity at the relay for the UL in bps.
$D_{c,i,BS}$	Required time to compute the task of the i-th UE at the BS in seconds.
$D_{c,i,R}$	Required time to compute the task of the i-th UE at the relay in seconds.
$D_{DL,i}$	DL transmission delay of the task's result of the i-th UE.
$D_{\max,i}$	Maximum time to complete the task of the i-th UE.
$D_{T,i,BS}$	Total delay time for offloading, computing the task, and downloading the task's result of the i-th UE at the BS in seconds.
$D_{T,i,R}$	Total delay time for offloading, computing the task, and downloading the task's result of the i-th UE at the relay in seconds.
$D_{UL,i}$	UL transmission delay of the task of the i-th UE.
L_i	Length of a task offloded by the i-th UE in bits
L_i'	Length of the i-th UE's task result in bits.
M	Total number of UEs in the BS that generate computationally intensive tasks to offload (referred to XR UEs)
M'	Number of XR UEs that are connected to the relay.
N	Total number of UEs in the BS that only require 5G connectivity (referred to as eMBB UEs).
N'	Number of eMBB UEs that are connected to the relay.
OH_i	Overhead inefficiency factor of the i-th link.
O_i	Required FLOP to complete the task of the i-th UE.
P	System power consumption.
$P^*_{BS,DL}$	Maximum RF output power for the DL at the BS in Watts.
$P^*_{R,DL}$	Maximum RF output power for the DL at the relay in Watts.
$P^*_{R,UL}$	Maximum RF output power for the UL at the relay in Watts.
P_{BS}	Consumed power at the BS in Watts.
$P_{out,BS}$	RF output radiated power at the BS in Watts.
$P_{out,R}$	RF output radiated power at the relay in Watts.
P_R	Consumed power at the relay in Watts.
R_i	Transmission data rate of a generic i-th wireless link in bps.
RI_i	Rank Indicator for the i-th link.
V^*_{BS}	Maximum computation speed at the BS in FLOPS.
V^*_R	Maximum computation speed at the relay in FLOPS.
V_{BS}	Required computation speed at the BS in FLOPS.
V_R	Required computation speed at the relay in FLOPS.
$a_{BS,}\ b_{BS}$	Power consumption parameters for the BS.
$a_{R,}\ b_R$	Power consumption parameters for the relay.
k_{BS}	Number of antennas for transmission in the BS.
k_R	Number of antennas for transmission in the relay.
m_i	Number of bits per symbols to be transmitted over the i-th link.
r_i	Code rate for the i-th link.
λ_i	Mean of the probability distribution for task generation of the i-th UE.

4.2. Computation model

The computational resources at a certain node are characterized in terms of the number of floating-point operations per second (FLOPS) that can be supported.

The required computation speed at the relay, V_R (FLOPS), is given by:

$$V_R = \sum_{i=1}^{M'} O_i \cdot \lambda_i, \text{ subject to } V_R \leq V_R^* \quad (1)$$

where V_R^* (FLOPS) is the maximum computation speed at the relay. Similarly, the required computation speed at the BS, V_{BS}, is defined as:

$$V_{BS} = \sum_{i=1}^{(M-M')} O_i \cdot \lambda_i, \text{ subject to } V_{BS} \leq V_{BS}^* \quad (2)$$

where V_{BS}^* (FLOPS) is the maximum computation speed at the BS.

The required time to compute a task of the i-th XR UE at the relay, $D_{c,i,R}$ (s), is given by:

$$D_{c,i,R} = O_i \,/\, V_R^* \quad (3)$$

Similarly, the required time to compute a task of the i-th XR UE at the BS, $D_{c,i,BS}$ (s), is given by:

$$D_{c,i,BS} = O_i \,/\, V_{BS}^* \quad (4)$$

4.3. Communication model

The transmission data rate, R_i, in bps in a generic i-th wireless link between a transmitter and a receiver (e.g., UL/DL UE-Relay, Relay-BS, UE-BS) is given by [34]:

$$R_i = B_i \cdot m_i \cdot r_i \cdot RI_i \cdot CP \cdot (1 - OH_i) \quad (5)$$

where B_i is the bandwidth assigned to this specific link, m_i is the number of bits per symbol to be transmitted and r_i is the code rate (i.e., the ratio between useful bits and total coded bits as a result of the channel coding process). The values of m_i and r_i are determined by the Modulation and Coding Scheme (MCS) according to the Signal to Interference and Noise Ratio (SINR) of the user. The value of RI_i is the Rank Indicator (RI), which specifies the number of layers used in MIMO. In addition, CP in (5) is an inefficiency factor due to the cyclic prefix, computed as the fraction of useful symbols duration in a slot, and OH_i captures the overhead inefficiency due to control channels and reference signals.

Focusing on the UL, the total required data rate capacity (bps) at the relay, $C_{R,UL}$, is given by:

$$C_{R,UL} = \sum_{i=1}^{M'+N'} R_i, \text{ subject to : } \sum_{i=1}^{M'+N'} B_i \leq B_{R,UL}^* \quad (6)$$

where the summation operator in $C_{R,UL}$ refers to the required data rate to support the existing UE-Relay ULs, considering both the XR and the eMBB UEs connected to the relay. Also, (6) considers that the aggregated B_i of all the existing UE-Relay ULs must be lower or equal to the total available bandwidth in the relay for the UL, $B^*_{R,UL}$. Note that the assigned B_i to the i-th user is obtained according to (5) by considering a required bit rate R_i. Moreover, in the case that the resulting bandwidth aggregated for all UEs exceeds the limit $B^*_{R,UL}$, the bandwidths of all $M' + N'$ UEs are reduced proportionally to their required bit rates in order to fit the bandwidth limit.

The total required radio capacity at the BS in the UL, $C_{BS,UL}$, is given by:

$$C_{BS,UL} = \sum_{i=1}^{N'} R_i + \sum_{i=1}^{(M-M')+(N-N')} R_i,$$
$$\text{subject to : } \sum_{i=1}^{N'} B_i + \sum_{i=1}^{(M-M')+(N-N')} B_i \leq B_{BS,UL}^* \quad (7)$$

where the first summation in $C_{BS,UL}$ refers to the required data rate to support existing ULs between the relay and BS and the second to the ULs between UEs and the BS to support the 5G connectivity of the N' eMBB UEs in the relay. In line with (6), (7) considers that the overall assigned bandwidth in the BS (i.e., the aggregated B_i of all UE-BS and BS-Relay ULs) must be lower or equal to the total available bandwidth in the BS for the UL, $B^*_{BS,UL}$.

Regarding the DL, the expressions for the total required radio capacity at the relay, $C_{R,DL}$, and at the BS, $C_{BS,DL}$, can be obtained by considering their respective total available bandwidth in the DL, $B^*_{R,DL}$ and $B^*_{BS,DL}$, in (6) and (7), respectively.

Considering the above, the transmission delay of a task from the i-th XR UE to the relay or the BS in the UL, $D_{UL,i}$, is:

$$D_{UL,i} = L_i / R_i \quad (8)$$

Similarly, once the task has been computed, the transmission time to download the task's result from the relay or the BS to the UE, $D_{DL,i}$, is:

$$D_{DL,i} = L_i' / R_i \quad (9)$$

The total delay time for computing a task in the relay, $D_{T,i,R}$, including the transmission of the task to the relay, the computation time in the relay and the transmission of the task's result back to the UE, is given by:

$$D_{T,i,R} = D_{UL,i} + D_{c,i,R} + D_{DL,i} \quad (10)$$

Correspondingly, the total delay time for computing a task in the BS, $D_{T,i,BS}$, is:

$$D_{T,i,BS} = D_{UL,i} + D_{c,i,BS} + D_{DL,i} \quad (11)$$

The values of $D_{T,i,R}$ and $D_{T,i,BS}$ need to be smaller than $D_{\max,i}$, i.e., $D_{T,i,R} \leq D_{\max,i}$ and $D_{T,i,BS} \leq D_{\max,i}$.

4.4. Power model

The Radio Frequency (RF) output radiated power at the BS, $P_{out,BS}$, in Watts is given by:

$$P_{out,BS} = \frac{P^*_{BS,DL}}{B^*_{BS,DL}} \cdot \sum_{i=1}^{(M-M')+N} B_i \quad (12)$$

where $P^*_{BS,DL}$ is the maximum RF output power for the DL at the BS and the summation operator refers to the total allocated bandwidth in the DL at the BS that accounts for the $(M-M')$ XR UEs that are connected directly to the BS and the N eMBB UEs that are connected to the BS either directly or through the relay. Note that in (12) it is assumed that $P_{out,BS}$ depends on the occupied bandwidth, as considered in previous works (e.g., Fantini et al. [35]).

At the relay, the RF output radiated power, $P_{out,R}$, in Watts considers two transmitters, one for the Relay-UE DL and another for the Relay-BS UL, and is given by:

$$P_{out,R} = \frac{P^*_{R,DL}}{B^*_{R,DL}} \cdot \sum_{i=1}^{M'+N'} B_i + \frac{P^*_{R,UL}}{B^*_{BS,UL}} \cdot \sum_{i=1}^{N'} B_i \quad (13)$$

where $P^*_{R,DL}$ and $P^*_{R,UL}$ are the maximum RF output power for the Relay-UE DL and Relay-BS UL, respectively. The first summation operator refers to the total assigned bandwidth for the Relay-UE DLs supporting M' and N' users. Instead, the second summation operator refers to the total assigned bandwidth for the Relay-BS ULs supporting N' users.

Based on several references [35–38], the following model is considered for the consumed power at the BS, denoted as P_{BS} (in Watts):

$$P_{BS} = a_{BS} \cdot P_{out,BS} + b_{BS} \cdot k_{BS} \quad (14)$$

where $P_{out,BS}$ is the RF output power at the BS in (12), a_{BS} captures the linear dependency between the radiated power and the power consumption, and b_{BS} is the power consumption associated with circuits, baseband processing, etc., that is given in Watts and is multiplied by the number of antennas used for transmission at the BS, denoted as k_{BS}. The values of a_{BS} and b_{BS} can be parametrized for different implementations.

Correspondingly, the power consumption at the relay, denoted as P_R, in Watts, is given by:

$$P_R = a_R \cdot P_{out,R} + b_R \cdot k_R \quad (15)$$

where a_R and b_R are the equivalent parameters to a_{BS} and b_{BS} parameters in (14) for the relay, and k_R is the number of antennas for transmission at the relay. It is assumed that the term b_R already captures the consumed power for both the transmitters Relay-UE DL and Relay-BS UL.

The system power consumption, P, is thus the aggregation of consumed power at the BS and the relay, that is:

$$P = P_{BS} + P_R \quad (16)$$

Note that the power consumption model is characterized from the perspective of the network operator and focuses on the communications infrastructure, thus excluding the power consumption of UEs terminals and the consumption at the edge computing servers.

5. Performance evaluation

5.1. Considered scenario

To illustrate the role and potential of embracing relays with edge computing capabilities in B5G deployments a scenario has been considered with the requirements of the XR UEs and eMBB UEs specified in Table 2. The values of the parameters considered for the computation, communication and power models are included in Table 3. Regarding the total available bandwidth and the maximum computation speed at the BS and the relay, different values are considered in each analysis, so they will be detailed in the corresponding subsections.

To illustrate the benefits of the proposed B5G scenario in a clearer though meaningful manner, the UEs are concentrated in a certain region in the BS area, so all of them experience similar radio conditions with respect to the BS and the relay. Three different situations are studied: *Situation A*, where it is considered that the relay has been properly deployed (i.e., the relay is located close to the traffic hot spot and has good visibility towards the BS), *Situation B*, where the relay has been deployed close to the UEs but with not so good visibility towards the BS and, *Situation C*, where the relay has not been properly deployed (i.e., bad channel conditions with the BS and far from the traffic hot spot). Table 4 summarizes the considered Channel Quality Indicator (CQI) and the associated $m_i \cdot r_i$ value for the different links and the abovementioned situations [35], which are valid for both the UL and DL directions of each link. According to these values, results have been obtained by assessing the system model analytically.

5.2. Analysis of computing and communication resources requirements gains

This section compares the bandwidth and computational requirements of the proposed approach with those in two benchmarks, considering that the total number of XR and eMBB users in the scenario are $N = 10$ and $M = 8$, respectively. In *benchmark* #1, there is only the BS (i.e., all the $(M + N)$ users are connected to the BS and the relay is not present). In turn, in *benchmark* #2, both the BS and the relay are present

Table 2
UEs requirements.

Parameter		Value
Required data rate per UE (R_i)	XR UE	UL and DL: 7 Mbps
	eMBB UE	UL and DL: 1 Mbps
Task specification (*XR UE*)	Task size (L_i)	70 kbits
	Task result size (L_i')	70 kbits
	Task required (O_i)	$1 \cdot 10^8$ FLOP
	Av. task generation rate (λ_i)	100 task/s
	Maximum delay ($D_{max,i}$)	30 ms

Table 3
System model parameters.

Parameter		Value
Number of antennas	Relay (k_R)	2
	BS (k_{BS})	2
Rank Indicator (RI_i)		2
Cyclic Prefix inefficiency factor (CP)		14/15
Overhead inefficiency factor (OH_i)		0.08
Maximum RF transmission power	BS DL ($P^*_{BS,DL}$)	40 W
	Relay DL ($P^*_{R,DL}$)	5 W
	Relay-BS UL ($P^*_{R,UL}$)	10 W
Power consumption parameters ([32])	a_R	20.4
	b_R	13.91 W
	a_{BS}	28.4
	b_{BS}	156.38 W

Table 4
Considered CQI values.

Link	Situation A		Situation B		Situation C	
	CQI	$m_i \cdot r_i$ (bps/Hz)	CQI	$m_i \cdot r_i$ (bps/Hz)	CQI	$m_i \cdot r_i$ (bps/Hz)
UE-BS	4	1.47	4	1.47	4	1.47
UE-Relay	11	5.11	11	5.11	6	2.406
Relay-BS	11	5.11	6	2.406	6	2.406

but the relay only offers communication capabilities. In this case, the number of XR and eMBB users that are connected to the relay are $M' = 4$ users and $N' = 5$ users, respectively, while the rest are connected to the BS. Then, we evaluate the proposed approach (i.e., relay with communication and computing capabilities) with the same M' and N' as in *benchmark* #2 and we obtain the reduction in bandwidth and computation requirements with respect to both benchmarks. For all the cases, the available bandwidth in the BS is $B^*_{BS,UL} = B^*_{BS,DL} = 28.8$ MHz and in the relay is $B^*_{R,UL} = B^*_{R,DL} = 14.22$ MHz (corresponding to a nominal bandwidth of 30 MHz and 15 MHz, respectively, in the 5G NR specifications with 15 kHz of subcarrier separation [39]), and the maximum computation speed in the BS is $V^*_{BS} = 100$ GFLOPS and in the relay $V^*_R = 50$ GFLOPS.

Fig. 3 shows the reduction in the bandwidth requirement in the BS in the UL and the DL. The reductions obtained for the BS with respect to *benchmark* #1 (i.e., only BS is present) take values between 40% and 50 %. These are due to the lower bandwidth requirement in the BS when using relays as the radio conditions in the Relay-BS link are better than in the UE-BS link in all cases. The differences observed between *Situation A* (i.e., good conditions in all links) and *Situation B-C* (i.e., worser conditions in the Relay-BS link and worse conditions in all links, respectively) are because much better conditions in the Relay-BS link are experienced in *Situation A*. Therefore, the bandwidth requirement in the

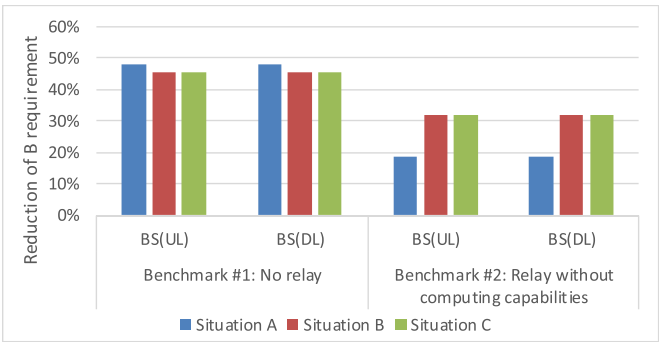


Fig. 3. Percentage of reduction of the required bandwidth at the BS in the UL and DL with respect to the benchmarks.

BS for *Situation A* is smaller than the one for the other two situations, which leads to a higher bandwidth reduction.

As for the reductions compared to *benchmark* #2 (i.e., relay without computing capabilities), smaller values are obtained than for *benchmark* #1 in the BS because only the reductions due to the incorporation of computing capabilities at the relay are captured. The reason for these reductions is that the bandwidth of XR UEs no longer needs to be allocated at the BS. Higher reductions are obtained in the BS for situations B-C than for *Situation A* due to its better radio conditions in the Relay-BS link. For the relay, no reductions are obtained for any of the situations since the same number of UL and DL connections will be established in the proposed approach and in *benchmark* #2. Overall, the results of Fig. 3 show that the introduction of relays with computing capabilities offers promising reductions of the required bandwidth in the BS. From the computational perspective, a reduction of 50 % of the required computational speed in the BS is obtained when including computing capabilities in the relay with respect to benchmarks #1 and #2 (i.e., only computing capabilities at the BS) since half of the operations are conducted in the relay according to the values of M and M'.

5.3. Analysis of capacity gains

This section considers that a certain amount of bandwidth and computing resources are available in the system and analyses the impact of the distribution of these resources between the BS and the relay on the maximum number of supported users in the system. Specifically, three cases are evaluated: *Distribution* #1, where 100 % of the resources are provided in the BS because there is no relay, *Distribution* #2, where 50 % of the resources are allocated to the BS and 50 % to the relay, and *Distribution* #3, where the communication resources are distributed as in *Distribution* #2 but 30 % of the computational resources are allocated to the BS and 70 % to the relay. For all the distributions, the maximum number of *XR* users is obtained by deriving the total computing delay at the BS, $D_{T,i,BS}$, and at the relay, $D_{T,i,R}$, for different values of M' and selecting the maximum value that fulfils $D_{\max,i}$. Note that the computation of the delays considers that the actual computational speeds and the data rate in the links are adjusted to the requirements of computational/communication resources and the total available resources in the BS/relay (i.e., if the required resources are higher than the available ones, the provided computational speeds and data rates are reduced according to the excess).

Fig. 4 shows the maximum number of *XR* users for distributions #1-#3 under the assumption that there are no *eMBB* users (i.e., $N = 0$). The total channel bandwidth in the scenario is 28.8 MHz for the DL and 28.8 MHz for the UL and the total computational capability is 100 GFLOPS. Results in Fig. 4 show that more XR users can be supported in the scenario for *Distribution* #2 (i.e., balanced resources in the BS and the relay) than for *Distribution* #1 (i.e., all the resources in the BS) for all situations with increasing factors of 78 % for *Situation* A (i.e., good conditions in all links) and *Situation* B (i.e., poorer conditions in the relay-BS link), and 14 % for *Situation* C (i.e., poor conditions in all links). Indeed, in *Distribution* #2 with the same resources in the BS and the relay, the relay can support 216 % more users than the BS for situations A and B and 67 % more in *Situation* C. These differences are consistent with the increase in the $(m_i \cdot r_i)$ values of the UE-Relay link with respect to the UE-BS link.

Another fact observed in the results is that *Distribution* #3 (i.e., more computing resources at the relay) allows increasing the number of supported XR users by 8 % with respect to *Distribution* #2 (i.e., balanced computing and radio resources between BS and relay) for situations A-B. The reason for this improvement is that in *Distribution* #2, as more users are supported due to good channel conditions, the computing resources are the limiting factor of the maximum number of *XR* users. This is addressed in *Distribution* #3 by providing more computing resources in the relay and reducing those in the BS. This results in an increase in the supported users in the relay, which is much higher than the reduction of the users in the BS. In the case of *Situation* C, no benefits for *Distribution* #3 are observed since the channel conditions in the Relay-UE channel are worse and are the limiting factor.

Fig. 5 shows the maximum number of XR users that can be supported in *Situation* A (i.e., good conditions in all links) when increasing the number of eMBB users, *N,* and for distributions #1–#3. Note that the distribution of the connected eMBB UEs to the BS-Relay is 100–0 % for *Distribution* #1, 50–50 % for *Distribution* #2 and 30–70 % for *Distribution* #3. Results show that the maximum number of supported XR users decreases when increasing *N* for all the distributions, as eMBB users consume bandwidth in the BS and the relay. The decrease in *Distribution* #1 is at a higher slope than in the other two distributions due to the worse channel conditions in the BS. Because of this, the gain of *Distribution* #2 over *Distribution* #1 increases with *N*, taking values of 78 % for $N = 0$, and 183 % for $N = 60$. However, the gains of *Distribution* #3 over *Distribution* #2 remain similar.

Fig. 6 depicts the maximum number of XR users when varying the value of the task size L_i and the task's result size L_i' for distributions #1-#3 and $N = 0$. A first observation is that the comparison between distributions #1-#3 in terms of supported XR users follows a similar trend for all the task sizes like in the results discussed before. Then, when assessing the impact of the task size, Fig. 6 reflects that the maximum number of XR users supported decreases as the task size increases for all the distributions. To get more insights in this effect, Table 5 summarizes the percentage of reduction at system level (i.e. aggregate XR users supported by BS and relay), BS level and relay level when increasing $L_i = L_i$' from 50 kbits to 70 kbits (i.e. 28.5 % increase) and from 70 kbits to 90 kbits (i.e. 22.5 % increase). It is shown that the percentage of reduction in the maximum number of users in the system follows approximately the percentage of increase of the task size in most cases. This effect is also observed at BS and relay levels. Nevertheless, in some other cases more deviating values are obtained. For example, this occurs in *Distribution* #2 when increasing the task size from 70 kbits to 90 kbits. The reason for this is that the task size only affects the transmission delay (see (8) and (9)). Hence, in cases where the delay due to computational

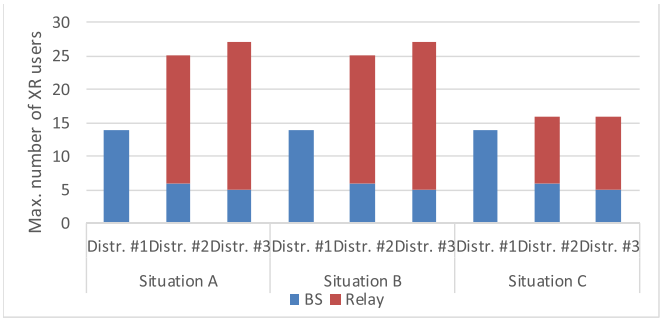


Fig. 4. Maximum number of XR users in situations A-C for different distributions of the resources in the scenario for $N = 0$.

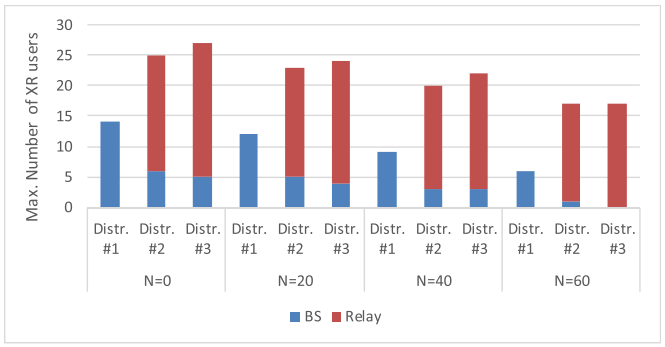


Fig. 5. Maximum number of XR users in *Situation* A when increasing *N*.

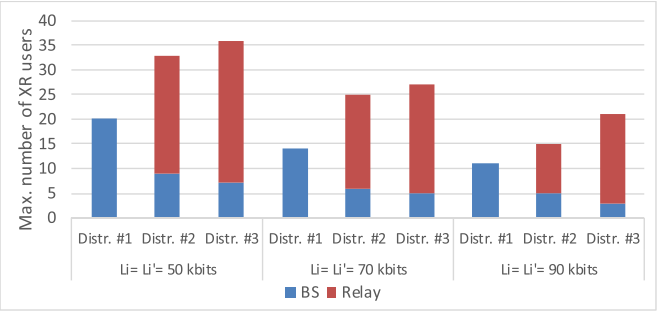


Fig. 6. Maximum number of XR users in Situation A for different values of task size (L_i and L_i') and $N = 0$.

Table 5
Reduction of the maximum number of users in relation to the increase of the task size.

Increase of L_i and L_i'	Distr.	Reduction of the maximum number of users		
		System	BS	Relay
50 kbits to 70 kits (28.5 %)	#1	30.00 %	30.00 %	–
	#2	24.24 %	33.33 %	20.83 %
	#3	25.00 %	28.57 %	24.14 %
70 kbits to 90 kbits (22.2 %)	#1	21.43 %	21.43 %	–
	#2	40.00 %	16.67 %	47.37 %
	#3	22.22 %	40.00 %	18.18 %

issues has a greater impact than that due to transmission, the task size has a lower impact on the maximum number of supported users.

Overall, the presented results in this section have highlighted the gains of including relays in terms of the maximum number of XR users supported in the scenario, showing that the distribution of the available resources in the system can be optimized to maximize the number of supported users. Also, the fact that these benefits remain even with the presence of eMBB users and for different task sizes is remarked.

5.4. Analysis of power consumption savings

This section assesses the impact of relays with computing capabilities from the perspective of power consumption. The assessment is conducted in terms of the power consumption reduction of the proposed approach with respect to the case without relay (i.e. all users connected to the BS). The parameter values for the power consumption assessment correspond to those in Table 3. Besides, the considered available bandwidth in the BS is $B^*_{BS,UL} = B^*_{BS,DL} = 28.8$ MHz and in the relay is $B^*_{R,UL} = B^*_{R,DL} = 14.22$ MHz (same values as the considered in Section 5.2).

Fig. 7 shows the percentage of power reduction when increasing the

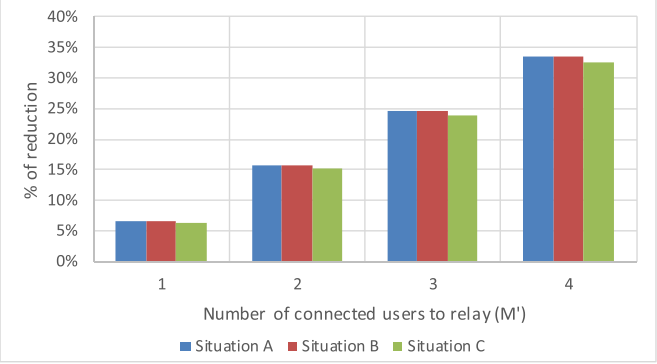


Fig. 7. Power consumption reduction (%) when increasing M' in situations A-C.

number of XR users connected to the relay, M', for situations A–C, considering that the total number of XR users in the system is $M = 8$ and there are no eMBB users (i.e., $N = 0$). Results show that the reduction in power consumption increases with M' for all situations indicating that the introduction of relays also has benefits from the energy consumption perspective. Specifically, significant reductions of more than 30 % are observed for $M' = 4$. The reductions obtained for *Situation* A (i.e., good conditions in all links) and *Situation* B (i.e., poorer conditions in the relay-BS link) are the same since both of them have the same spectral efficiencies in the UE-BS and UE-Relay links, and as there are no eMBB users ($N' = 0$), no data transmission is required over the BS-Relay link, which is the only link that has different spectral efficiencies between the two situations. Therefore, the bandwidth allocation and, hence, the power consumption are the same for both situations. However, the achieved power consumption reduction in *Situation* C (i.e., poorer conditions in all links) is slightly smaller than in *Situations* A-B because the conditions of the UE-Relay link in that situation are worse, resulting in higher bandwidth and therefore a higher power consumption at the relay.

To show the effect of the number of eMBB users, Table 6 includes the power consumption reduction values for $N = 0$ and $N = 10$ for $M = 8$ in situations A–C, considering that 50 % of the users are connected to the relay and the rest to the BS. The results show that the power consumption reductions are slightly smaller with the introduction of eMBB users. The reason is that eMBB users connected to the relay use the radio resources of the BS-relay and UE-relay links, so they consume power in both the BS and the relay. In any case, Table 6 shows that power reductions close to 30 % are still obtained with the introduction of the eMBB users. Moreover, the results in Table 6 show that the power consumption reduction for $N = 10$ is not the same for situations A-B as observed in the previous results. In fact, the reduction is greater in *Situation A* since the spectral efficiency in the BS-relay is larger in that situation.

The power consumption highly depends on the implementation of the BS and relay in terms of the antenna interface (i.e., including the feeder, antenna bandpass filters, duplexers), power amplifiers, circuits, signal processing, cooling system, etc., which is reflected in the parameters of the power consumption model. To assess this impact, different configurations of the parameters a_R and b_R of the relay have been considered, as indicated in Table 7, based on different references in the literature. Configuration #1 corresponds to the one of Table 3, which has been considered for the previous results in this section.

Fig. 8 shows the percentage of power consumption reduction in the proposed approach with respect to the case without relay for configurations #1–#4 when the number of XR users is increased. Results are obtained for *Situation A*, considering that the total number of XR users is $M = 8$ and that there are no eMBB users (i.e. $N = 0$). The obtained power consumption reduction in Fig. 8 increases with M' for all configurations although there are significant differences among them. The maximum reductions are obtained by configurations #3 and #4 with some small differences. For $M' \leq 2$, configuration #4 has a slightly higher reduction than configuration #3, since the value of b_R is the smallest among the considered configurations and the power consumption due to data transmission is low due to the low number of users M'. However, for $M' \geq 3$ the reduction in power consumption for configuration #3 slightly exceeds that for configuration #4. The reason is the lower value of a_R, which limits the impact of increasing the RF output radiated power ($P_{out,R}$) when increasing the assigned bandwidth with the number of users. The lowest reductions are obtained for configuration #2 because the

Table 6
Power consumption reduction (%) in the presence of eMBB users.

N	*Situation A*	*Situation B*	*Situation C*
0	33.54 %	33.54 %	32.52 %
10	30.61 %	29.88 %	28.78 %

Table 7
Power consumption configuration parameters.

Configuration	a_R	b_R [W]	Ref.
Config. #1	20.4	13.91	[35]
Config. #2	2.6	56	[36]
Config. #3	4.0	6.8	[36]
Config. #4	19.2	0.9	[37]

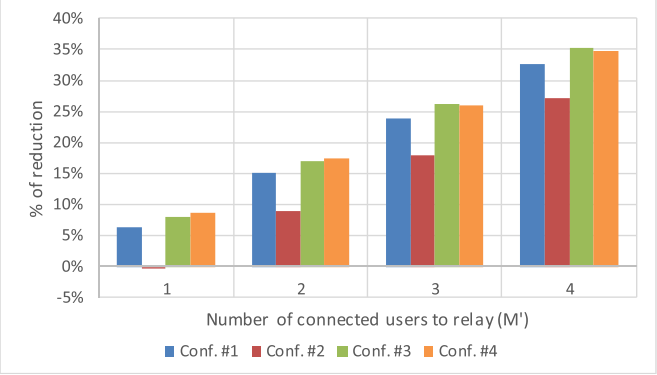


Fig. 8. Power consumption reduction (%) when increasing M' for config. #1–#4 in Situation A.

value of b_R for this configuration is very large. The fact that larger values of b_R lead to smaller power consumption reductions can be also observed when comparing configurations #1 and #4, which have similar values of a_R, but configuration #4 has a significantly smaller value of b_R. Overall, these results illustrate that the benefits of including a relay with computing capabilities in terms of power consumption are highly dependent on the implementation of the relay. In any case, significant power reductions are still observed since, for example, for $M' = 4$ the power consumption reduction ranges between 26 % and 35 % depending on the considered configuration.

6. Challenges

This section discusses different challenges in relation to the introduction of relays with computing capabilities in future B5G systems.

As introduced in Section 3, the support of the different relay types (i.e., fixed, moving, UE relays) can leverage different technologies. These can be characterized by different architectures that involve different entities and functional splits between e.g. the Radio Unit (RU), Distributed Unit (DU) and Central Unit (CU). In this regard, the support of fixed relays can be currently supported by the IAB technology standardized in 3GPP Release 16. In the IAB architecture, illustrated in Fig. 9, the relay, referred to as IAB-node, contains the Mobile Terminal (MT) function that keeps the wireless backhaul with the IAB-donor (i.e., the BS or the gNB in the 3GPP terminology) [40]. Moreover, the relay includes the DU that hosts the Radio Link Control (RLC), Medium Access Control (MAC) and physical (PHY) layers of the radio interface protocol stack to provide connection to UEs through the NR Uu interface between the UE and the IAB-node. In addition, the IAB DU connects to the CU of the IAB donor through the F1 interface. The transmission of this interface is performed on top of the NR Uu interface between the MT and the gNB-DU of the BS using the Backhaul Adaptation Protocol (BAP) sublayer.

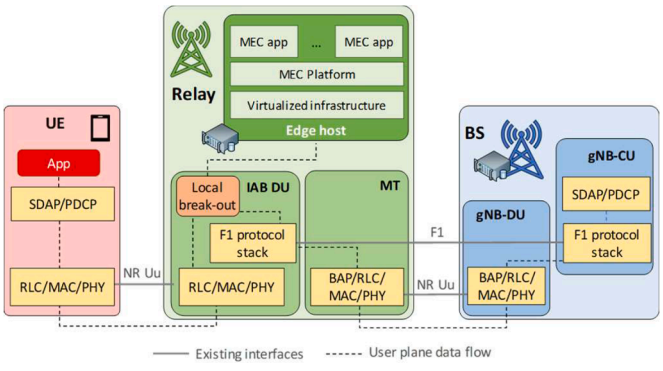


Fig. 9. Architecture of fixed and moving relays with computing capabilities based on IAB technology.

The IAB technology is currently standardized for fixed relays, so it requires to be adapted to support moving relays, denoted as Mobile IAB-nodes. Current works in this direction are conducted by the 3GPP in [15] and [41]. RF and Radio Resource Management (RRM) requirements to support moving IAB-nodes need to be established, including the definition of procedures for the topology adaptation to enable IAB-node mobility, enhancements for mobility of IAB-nodes together with its served UEs, which is related to group mobility, mitigation of interferences due to IAB-node mobility, etc. These aspects need to be addressed considering that Mobile IAB-nodes need to be able to serve legacy UEs and that solutions need to support UE handover and dual connectivity mechanisms.

Regarding relay UEs, the supporting architecture is based on the D2D technology, introduced by 3GPP for the so-called Proximity Services (ProSe). D2D communication relies on the PC5 interface between UEs that is defined on top of a new radio link for direct transmissions between devices, denoted as sidelink. Normative specifications for ProSe with the support of UE-to-Network (U2N) relay were defined in [42]. The fact that relay UEs are battery-constrained devices and typically involve mobility poses challenges that need to be addressed by incorporating new management functionalities. These include, for example, the monitoring of the conditions of relay UEs (e.g., location, coverage conditions, etc.), the determination of the relay UEs that are available in the coverage area of each BS or the activation/deactivation of their relay functionality only when it becomes necessary. Another challenge is the low number of commercial mobile terminals that nowadays support the PC5 interface, which represents a limitation in enabling the communication between a UE and a relay UE. In any case, this limitation is expected to change in the near future thanks to the appearance of new use cases for the PC5/sidelink as defined in [43].

The support of edge computing capabilities requires that relays implement the so-called local break-out to select the traffic that has to be processed locally at the computing resources of the relay. This functionality is devised in the 3GPP vision of edge computing in [44], which considers that the traffic forwarding to the edge platform is performed to a local User Plane Function (UPF) at the local site, which sends the IP-based traffic to be computed by the edge platform. For the case of IAB-based relays, the support of this local break-out is architecturally challenging since the IAB-DU only covers up to the RLC layer of the radio interface. Hence, modifications of the protocol stack (e.g. by incorporating the Packet Data Convergence Protocol (PDCP) and Service Data Adaptation Protocol (SDAP) layers) would be required to allow extracting the IP packets that have to be locally processed. In the case of D2D-based relays (i.e., relay UEs), the implementation of the local break-out depends on the implemented protocol layers supported over the PC5 interface of the relay UE, where two possibilities are devised in the U2N relay normative. The first option, denoted as Layer-3 U2N Relay, considers that the U2N relay has the full protocol stack (from SDAP down to PHY), which would allow to easily implement the local break-out since it enables the exchange of IP packets. The second option, denoted as Layer 2 U2N relay, considers that the U2N relay only includes the RLC/MAC/PHY layers, so similar limitations to the case of IAB-based relays would be encountered.

Another relevant aspect of enabling relays with computing capabilities is to ensure that they support standardized edge computing

platforms. The current standardization conducted by ETSI GR MEC has mainly focused on the development of an architecture composed of different MEC entities (i.e., MEC platform, MEC apps, Virtualization Infrastructure Manager, etc.) that is designed for high computing hosts co-located with the BS [45]. The support of this architecture needs to take into account that relays might have more limited computing capabilities (e.g. in the case of relay UEs), so the deployment of the MEC architecture entities needs to be optimized to minimize the occupation of computing resources (e.g., deploying only selected MEC entities and with reduced functionality). Moreover, to avoid the interruption of the processing of tasks when UEs move in the area, mechanisms are needed to ensure that the MEC platform at a certain relay can interact with the MEC platform at the neighbouring relays and the BS to transfer the tasks. Some of these aspects have been captured in the ongoing ETSI GR MEC work item in [20] on MEC constrained devices.

7. Conclusions and future work

This paper has elaborated on the use of relays with computing capabilities in beyond 5G deployments. Different types of relays envisaged for B5G are identified and considerations on including computing capabilities on them are discussed. Then, the system model including relays with computing capabilities is characterized from computational, communication and power consumption perspectives. The communication model has considered 5G NR parameters. The system is assessed by providing results on a beyond 5G deployment with extended reality users, which is evaluated under different radio channel conditions. Results have shown that: (i) High reductions in the required bandwidth and computational speed in the base station are achieved with respect to benchmark scenarios without relays and relays without computing capabilities; (ii) Deploying the relay in a location with good radio conditions with the BS is relevant to achieve higher bandwidth savings; (iii) The distribution of the available computing and communication resources between the relay and the base station can be optimized to maximize the capacity. (iv) Relevant reductions of power consumption in the order of 30 % are also achieved with the introduction of relays with computing capabilities, although these reductions are highly dependent on the hardware implementation aspects. Finally, the paper has discussed some challenges associated to the future incorporation of different types of relays in B5G, taking as reference standardized architectures for relays and edge computing.

Future work includes the assessment of the benefits of incorporating relays with computing capabilities by using system-level simulations that allow capturing the peculiarities of the different types of relays (i.e., fixed, moving or relay UE) and the effect of moving users. Moreover, extending the power consumption model with the parameters for the computation of the tasks is also envisaged. Finally, the study of the optimization of the distribution of the resources between the BS and the relays, which will be a relevant problem in the planning and deployment of scenarios with relays and BSs with computing capabilities, is also devised.

CRediT authorship contribution statement

I. Vilà: Conceptualization, Methodology, Validation, Software, Formal analysis, Investigation, Resources, Writing – original draft, Writing – review & editing, Visualization. **O. Sallent:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition. **J. Pérez-Romero:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgement

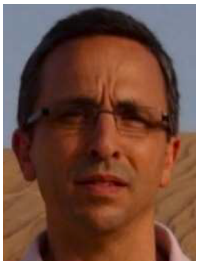
This work is part of VERGE project, which has received funding from the Smart Networks and Services Joint Undertaking under the European Union's Horizon Europe research and innovation programme under Grant Agreement 101096034. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. This paper has also been partly funded by the Spanish Ministry of Science and Innovation MCIN/AEI/10.13039/501100011033 under ARTIST project (ref. PID2020-115104RB-I00). The work of Irene Vilà has been funded by the European Union-NextGenerationEU, the Spanish Ministry of Universities and the Plan for Recovery, Transformation and Resilience, through the call for Margarita Salas Grants of the Universitat Politècnica de Catalunya (ref. 2022UPC-MSC- 94079).

References

[1] K. Samdanis, T. Taleb, The road beyond 5G: a vision and insight of the key technologies, IEEE Netw. 34 (2) (2020) 135–141.
[2] W. Saad, M. Bennis, M. Chen, A vision of 6G wireless systems: applications, trends, technologies, and open research problems, IEEE Netw. 34 (3) (2020) 134–142.
[3] AIOTI, "High Priority Edge Computing Standardisation Gaps and Relevant SDOs", 2022.
[4] N. Bhushan, et al., Network densification: the dominant theme for wireless evolution into 5G, IEEE Commun. Mag. 52 (2) (2014) 82–89.
[5] S. Rangan, T.S. Rappaport, E. Erkip, Millimeter-wave cellular wireless networks: potentials and challenges, Proc. IEEE 102 (3) (2014) 366–385.
[6] J.G. Andrews, et al., What will 5G Be? IEEE J. Sel. Areas Commun. 32 (6) (2014) 1065–1082.
[7] M. Caprolu, et al., "Edge computing perspectives: architectures, technologies, and open security issues," 2019 *IEEE International Conference On Edge Computing (EDGE)*, Milan, Italy, 2019, pp. 116–123.
[8] Y. Mao, et al., A survey on mobile edge computing: the communication perspective, IEEE Commun. Surv. Tutor. 19 (4) (2017) 2322–2358. Fourthquarter.
[9] Y.C. Hu, M. Patel, D. Sabella, N. Sprecher, V. Young, Mobile edge computing a key technology towards 5G, in: *ETSI, White Paper* #11, France, 2015.
[10] N. Abbas, et al., Mobile edge computing: a survey, IEEE Internet Things J. 5 (1) (2018) 450–465.
[11] N. Sprecher, et al., Harmonizing standards for edge computing, a synergized architecture leveraging ETSI ISG MEC and 3GPP specifications, in: ETSI White paper #36, France, 2020.
[12] J. Sydir, R. Taori, An evolved cellular system architecture incorporating relay stations, IEEE Commun. Mag. 47 (6) (2009) 115–121.
[13] O. Teyeb, et al., Integrated access backhauled networks, in: *2019 IEEE 90th Vehicular Technology Conference* (VTC-2019 Fall), Honolulu, HI, USA, 2019.
[14] M. Polese, et al., Integrated access and backhaul in 5G mmWave networks: potential and challenges, IEEE Commun. Mag. 58 (3) (2020) 62–68.
[15] 3GPP TR 22.839 v18.1.0, "Study on Vehicle-Mounted Relays; Stage 1 (Release 18)", 2021.
[16] R. Balakrishnan, et al., Mobile relay and group mobility for 4G WiMAX networks, in: 2011 IEEE Wireless Communications and Networking Conference, Cancun, Mexico, 2011.
[17] S. Andreev, et al., Future of ultra-dense networks beyond 5G: harnessing heterogeneous moving cells, IEEE Commun. Mag. 57 (6) (2019) 86–92.
[18] 3GPP TS 22.261 v18.5.0, "Service requirements for 5G system; Stage 1 (Release 18)", 2021.
[19] J. Pérez-Romero, O. Sallent, Leveraging user equipment for radio access network augmentation, in: IEEE Conference on Standards for Communications and Networking (CSCN), Thessaloniki, Greece, 2021.
[20] ETSI GR MEC 036 work item, "Multi-access Edge Computing (MEC); MEC in resource constrained terminals, fixed or mobile", Available at: https://portal.etsi.org/webapp/WorkProgram/Report_WorkItem.asp?WKI_ID=59467.

[21] J. Liang, Z. Chen, C. Li, B. Xia, Delay outage probability of multi-relay selection for mobile relay edge computing system, in: 2019 *IEEE/CIC International Conference on Communications in China* (ICCC), Changchun, China, 2019, pp. 898–902.
[22] X. Chen, et al., Joint cooperative computation and interactive communication for relay-assisted mobile edge computing, in: *2018 IEEE 88th Vehicular Technology Conference* (VTC-Fall), Chicago, USA, 2018, pp. 1–5.
[23] M. Yao, et al., Energy efficient cooperative edge computing with multi-source multi-relay devices, in: 2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), Zhangjiajie, China, 2019, pp. 865–870.
[24] X. Cao, et al., Joint computation and communication cooperation for energy-efficient mobile edge computing, IEEE Internet Things J. 6 (3) (2019) 4188–4200.
[25] D. Hu, et al., Joint task offloading and computation in cooperative multicarrier relaying-based mobile-edge computing systems, IEEE Internet Things J. 8 (14) (2021) 11487–11502.
[26] Z. Luo, G. Huang, Energy-efficient mobile edge computing in RIS-aided OFDM-NOMA relay networks, IEEE Trans. Veh. Technol. 72 (4) (2023) 4654–4669.
[27] I. Vilà, O. Sallent, J. Pérez-Romero, Expanding edge computing deeper into beyond 5G radio access networks, in: *2023 IEEE 9th International Conference on Network Softwarization (NetSoft)*, Madrid, Spain, 2023, pp. 432–437.
[28] C.D. Alwis, et al., Survey on 6G frontiers: trends, applications, requirements, technologies and future research, IEEE Open J. Commun. Soc. 2 (2021) 836–886, https://doi.org/10.1109/OJCOMS.2021.3071496.
[29] 3GPP TS 38.300 V17.3.0, "NR and NG-RAN Overall Description; Stage 2 (Release 17)", December 2022.
[30] G. Noh, H. Chung, I. Kim, Mobile relay technology for 5G, IEEE Wirel. Commun. 27 (3) (2020) 6–7.
[31] G. Noh, et al., Realizing multi-Gb/s vehicular communication: design, implementation, and validation, IEEE Access (2019) 19435–19446.
[32] "Mediapro, Telefónica and TMB to develop the first augmented reality project over 5G on tourist buses". Telefónica. https://www.telefonica.com/en/communication-room/mediapro-telefonica-and-tmb-to-develop-the-first-augmented-reality-project-over-5g-on-tourist-buses/(Accessed March, 27, 2023).
[33] T. Stone. "Trams in Florence, Italy, equipped with traffic data sensors with AV functionality". Traffic Technology Today (TTT). https://www.traffictechnologytoday.com/news/autonomous-vehicles/trams-in-florence-italy-equipped-with-traffic-data-sensors-with-av-functionality.html (Accessed March, 27, 2023).
[34] 3GPP TS 38.306 v17.2.0, "NR User Equipment (UE) radio access capabilities (Release 17)," 2022.
[35] R. Fantini, D. Sabella, M. Caretti, An E3F based assessment of energy efficiency of relay nodes in LTE-advanced networks, in: 2011 IEEE 22nd International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Toronto, Canada, 2011.
[36] G. Auer, et al., How much energy is needed to run a wireless network? IEEE Wirel. Commun. 18 (5) (2011) 40–49.
[37] E. Björnson, M. Kountouris, M. Debbah, Massive MIMO and small cells: improving energy efficiency by optimal soft-cell coordination, in: International Conference on Telecommunication (ICT) 2013, Casablanca, Morocco, 2013, pp. 1–5.
[38] A. Israr, Q. Yang, A. Israr, Power consumption analysis of access network in 5G mobile communication infrastructures—An analytical quantification model, Pervasive Mob. Comput. 80 (2022) 101544.
[39] 3GPP TS 38.214 v15.5.0, "NR; Physical layer procedures for data (Release 15)," 2019.
[40] 3GPP TS 38.401 v17.1.1, "NG-RAN; Architecture description (Release 17)", 2022.
[41] RP-222671 "Mobile IAB (Integrated Access and Backhaul) for NR", 3GPP TSG RAN Meeting #97, 2022.
[42] 3GPP TS 23.304 v17.3.0, "Proximity based Services (ProSe) in the 5G System (5GS) (Release 17)", 2022.
[43] D. McQueen, "The emergence of 5G and 5G-Advanced new features to drive adoption of system-level radio design (Analyst Angle)", 2022. https://www.rcrwireless.com/20221129/5g/the-emergence-of-5g-and-5g-advanced-new-features-to-drive-adoption-of-system-level-radio-design-analyst-angle (accessed: February 2023).
[44] 3GPP TS 23.548 v18.1.1, "5G System Enhancements for Edge Computing; Stage 2 (Release 19)", Technical Specification, 2023-04.
[45] ETSI GS MEC 003 v3.1.1, "Multi-access Edge Computing (MEC); Framework and Reference architecture", Group Specification, 2022-03.

Irene Vilà received a B.E. degree in Telecommunication Systems Engineering in 2015, an M.S. degree in Telecommunication Engineering in 2017, and a Ph.D. degree in Signal Theory and Communications in 2022, all from Universitat Politècnica de Catalunya (UPC), Barcelona, Spain. She is currently a postdoctoral researcher between the CONNECT center in Trinity College Dublin (TCD) and the Department of Signal Theory and Communications (TSC) at UPC. She has been involved in different national and European research projects founded by both public and private organizations. She has coauthored more than 20 papers in international journals and renewed conferences, as well as organized a conference tutorial, and contributed to a book chapter. Her research interests focus on the field of mobile communications, particularly on radio resource management and network optimization, the application of artificial intelligence to radio access network management, and edge computing.



Oriol Sallent is a Professor at the Universitat Politècnica de Catalunya, Barcelona, Spain. He has participated in a wide range of European and national projects, with diverse responsibilities as a principal investigator, coordinator, and work package leader. He regularly serves as a consultant for a number of private companies. He has contributed to standardization bodies such as 3GPP, IEEE, and ETSI. He is the coauthor of 13 books and has authored or coauthored 300+ papers, mostly in high-impact IEEE journals and renowned international conferences. His research interests include 5G RAN (Radio Access Network) planning and management, artificial intelligence-based radio resource management, virtualisation of wireless networks, cognitive management cognitive radio networks and dynamic spectrum access and management, among others.

Jordi Pérez-Romero (Member, IEEE) received a degree in telecommunications engineering and a Ph.D. degree from the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, in 1997 and 2001, respectively.

He is currently a Professor with the Department of Signal Theory and Communications of UPC. He is working in the field of wireless communication systems, with a particular focus on 5G and beyond cellular systems, including radio resource management and network optimization. He has been involved in different European projects with different responsibilities, such as researcher, work package leader, and project responsible, has participated in different projects for private companies and has contributed to the 3GPP and ETSI standardization bodies. He has coauthored more than 300 papers in international journals and conferences. He has also coauthored three books and has contributed to seven book chapters. He is an Associate Editor for the IEEE Vehicular Technology Magazine.