Expanding Edge Computing deeper into Beyond 5G Radio Access Networks

I. Vilà, O. Sallent, J. Pérez-Romero
Signal Theory and Comunications Department of Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain
irene.vila.munoz@upc.edu, sallent@tsc.upc.edu, jordi.perez-romero@upc.edu

***Abstract*— Relevant services envisaged for beyond 5G (B5G) systems, such as extended reality and holographic communications, present stringent user experience requirements with high computational and communication demands. While edge computing aims to address the computation requirements by offloading the computational tasks to edge servers close to the user, the communication will leverage the technologies developed for 5G New Radio together with an unprecedented level of network densification. This paper advocates for deploying relays equipped with edge computing capabilities. The potentials of this approach for B5G are identified and a system model is presented to characterize both computational and communications perspectives. Based on this, results are provided to show the benefits and limitations of the proposed approach from a system-level perspective.**

Keywords—Beyond 5G, Edge Computing, Relays

I. Introduction

The evolution of communication networks Beyond 5G (B5G) and towards 6G is expected to deal with the diverse and challenging requirements of a broad range of vertical services [1][2]. For example, applications based on eXtended Reality (XR) and holographic representations are expected to become a key asset in a wide range of scenarios, fusing the digital and the real world to deliver new experiences to end users, such as metaverse environments, immersive online gaming, real-time 3D communications, etc. These emerging applications with stringent user experience requirements are becoming increasingly demanding in terms of both the required computation and communication capabilities of B5G communication infrastructures.

To address the computation challenge, the traditional approach of offloading heavy tasks to powerful computing elements residing in the cloud (i.e., cloud computing) is no longer capable to meet the latency requirements of such applications. In response to these needs, edge computing has been rapidly evolving as a novel paradigm that brings computational power and resources closer to where the data is generated, thus considerably reducing response times with a much lower carbon footprint [3].

Regarding the communications' requirements, 5G NR has been built out of multiple technology components (e.g., multiuser massive MIMO, smart beamforming) leading to a significant increase in the achievable spectral efficiency. However, to realize the promise of vastly increased data rates (from Mbps to Gbps) and ultra-reliable low latency (from tens of milliseconds down to milliseconds), network densification has been identified for a long time as an integral part of 5G network deployment [4]. The relevance of network densification is exacerbated as high 5G frequency bands with poorer propagation characteristics will become more integrated [5]. Millimeter wave (mmWave) signals at these frequencies exhibit reduced diffraction and more specular propagation than their microwave counterparts, and hence they are much more susceptible to blockages [6]. Consequently, massive capital expenditure (CAPEX) on the deployment of 5G infrastructure will be required to provide the desired capacity and coverage needs. Therefore, Mobile Network Operators (MNOs) need to find new and creative ways of managing and deploying their 5G and beyond Radio Access Network (RAN) infrastructures to avoid seeing their finances stretched.

With all the above, this paper advocates for B5G RAN deployments exploiting relay nodes with edge computing capabilities. The envisaged solution expands the edge computing paradigm deeper into the RAN by leveraging relay nodes as a way to further exploit the task offloading benefits and as a mechanism to truly meet service requirements, particularly concerning the necessary service continuity that can be put at risk due to poor coverage footprints. Hence, the synergy between relay-enhanced B5G RAN and edge computing can provide computing and communication capabilities for applications residing at the boundary of MNOs' networks.

The rest of the paper is structured as follows. Section II describes the related work and outlines the novelties and contributions of this paper. Section III elaborates on the different possibilities for relay-empowered B5G RAN with edge computing capabilities, while this scenario is characterized in Section IV. Section V includes some performance assessment results, highlighting the benefits and limitations of edge computing-enabled relays. Finally, Section IV summarizes the conclusions.

II. 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, and 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 [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 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, [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 edge computing capabilities has been considered only in a few works in the literature [20]-[23]. Among these, [20] and [21] 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 [20] 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 [21] 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 [20] and [21], which are designed for cooperative wireless systems scenarios, the scenario considered in [22] and [23] 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 [22] 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 [23] 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. 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. However, none of the previous works has addressed the joint optimization of communication and computational resources when incorporating relays with computing capabilities in cellular networks, which is the main novelty of this paper. Furthermore, in contrast to previous works, which have not considered any specific radio technology, this paper puts an special focus on B5G RAN deployments, considering 5G NR parameters for the communication model.

III. Relay-empowered B5G RAN with Edge Computing Capabilities

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).

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.

- *Fixed relay*. In this case, the IAB solution can be leveraged [24], 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 to a planning and dimensioning process, both from communication and computing perspectives. From the communication side, the deployment could be motivated to improve 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 likely offers better motivation to also allocate computing resources at the relay. 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.

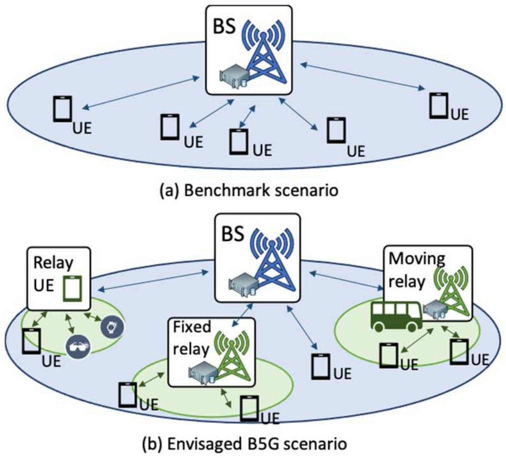

Fig. 1. Benchmark and envisaged B5G scenarios

- *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 [25]. 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. [26]. 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 [27]. Another example is an autonomous tram [28]. 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 device-to-device (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 [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 from 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.

IV. System Model

Let us consider the system depicted in Fig. 2 with a BS and a relay, which can be either fixed, moving or relay UEs. 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 that are subject to be offloaded at any edge computing platform and *N* UEs that only require 5G connectivity. Both types of users are non-Guaranteed Bit Rate (non-GBR). At a certain point in time and following certain decision-making criteria, *M'* and (*M-M'*) UEs exploit edge computing capabilities at the relay and the BS, respectively, while *N'* and (*N-N'*) only-connectivity UEs gain connectivity through the relay or via a direct link with the BS, respectively.

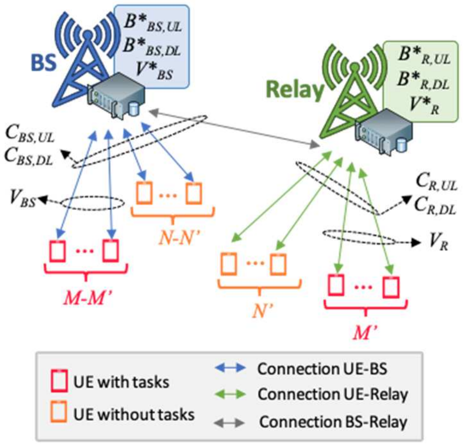

Fig. 2. System model

In the following, the considered task, computation and communication models in the system are detailed.

A. Task model

UEs with computationally intensive tasks generate non-divisible tasks. For the *i*-th 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).

B. 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:

$$V_R = \sum_{i=1}^{M'} O_i \cdot \lambda_i \,, \quad \text{subject to } V_R \le V^*_R \tag{1}$$

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:

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

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

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

$$D_{c,i,R} = O_i \,/\, V^*_R \tag{3}$$

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

$$D_{c,i,BS} = O_i \,/\, V^*_{BS} \tag{4}$$

C. 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 [29]:

$$R_i = B_i \cdot m_i \cdot r_i \cdot RI_i \cdot CP \cdot (1\text{-}\ OH_i) \tag{5}$$

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:

$$C_{R,UL}=\sum_{i=1}^{M'+N'} R_i \;, \qquad \text{subject to: } \sum_{i=1}^{M'+N'} B_i \le B^*_{R,UL} \tag{6}$$

where the summation operator in $C_{R,UL}$ refers to the required data rate to support the existing UE-Relay ULs. Also, (6) considers that the the aggregated B_i of all the existing UE-Relay ULs must be lower or equal than the total available bandwidth in the relay for the UL, $B^*_{R,UL}$.

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

$$C_{BS,UL}=\sum_{i=1}^{N'} R_i+\sum_{i=1}^{(M-M')+(N-N')} R_i \,, \qquad \text{subject to: } \sum_{i=1}^{N'} B_i+\sum_{i=1}^{(M-M')+(N-N')} B_i \le B^*_{BS,UL} \tag{7}$$

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. 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 than 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 UE to the relay or the BS in the UL is:

$$D_{UL,i} = L_i/R_i \tag{8}$$

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 is:

$$D_{DL,i} = L'_i/R_i \tag{9}$$

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:

$$D_{T,i,R} = D_{UL.i} + D_{c,i,R} + D_{DL,i} \tag{10}$$

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

$$D_{T,i,BS} = D_{UL.i} + D_{c,i,BS} + D_{DL,i} \tag{11}$$

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

V. Performance Evaluation

A. Considered scenario

The considered scenario to illustrate the role and potential of embracing relays with edge computing capabilities in B5G deployments consists of two different types of services associated to the UEs in the area of a single BS, namely *eXtended Reality* (*XR*) and *enhanced Mobile Broadband* (*eMBB*) users. While *eMBB* users only require communication capabilities in the scenario, *XR* users generate tasks that are offloaded to the edge site (i.e., BS/Relay). The considered requirements for UEs are specified in Table I.

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 II summarizes the considered Channel Quality Indicator (CQI) and the associated $m_i{\cdot}r_i$ value for the different links and the abovementioned situations [30]. In addition, RI_i=2, CP=14/15 and OH_i=0.08 are considered for all the situations. According to these values, results have been obtained by assessing the system model analytically through a Python development.

TABLE I. 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 FLOP (O_i)	$1{\cdot}10^8$
	Av. task generation rate (λ_i)	100 task/s
	Maximum delay ($D_{max,i}$)	30 ms

TABLE II. 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

B. Analysis of potential 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 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 channelization of 30 MHz and 15 MHz, respectively, in

5G NR with 15 kHz of subcarrier separation), 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 take values between 40%-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* and *Situation B-C* are because much better conditions in the Relay-BS link are experienced in *Situation A*. Therefore, the bandwidth requirement in the 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, 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 since half of the operations are conducted in the relay according to the values of *M* and *M*'.

C. *Capacity assessment*

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 not *eMBB* users (i.e., *N*=0). The total channel bandwidth in the scenario is 30 MHz for the DL and 30 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 than for *Distribution* #1 for all situations with increasing factors of 78% for situations A-B and 14% for *Situation* C. 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·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 allows increasing the number of supported *XR* users by 8% with respect to *Distribution* #2 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 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. Overall, the presented results in this section have highlighted the gains of including relays in 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 only-connectivity users is remarked.

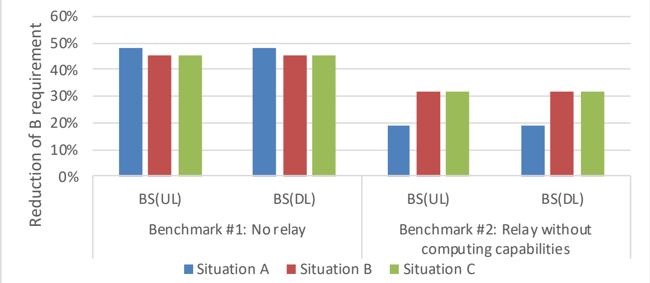

Fig. 3 Percentage of reduction of the required bandwidth at the BS in the UL and DL with respect to the benchmarks

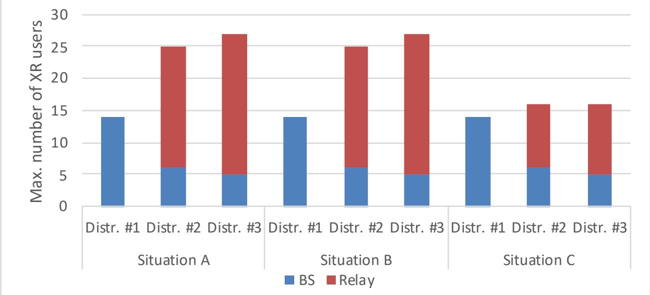

Fig. 4. Maximum number of XR users in situations A-C for different distributions of the resources in the scenario for *N*=0.

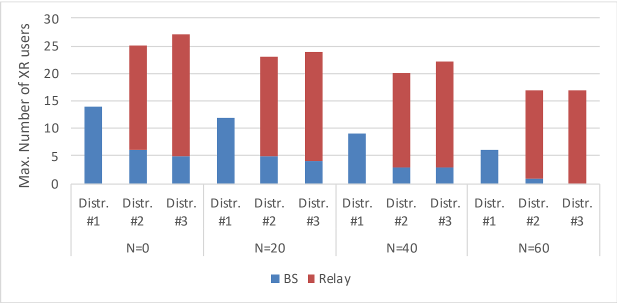

Fig. 5. Maximum number of *XR* users in *Situation* A when increasing *N*.

VI. Conclusions and Future Work

This paper has elaborated on the use of relays with computing capabilities in beyond 5G (B5G) 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, considering 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 on 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. Future work is devised to study the implications of different types of relays and different extended reality applications with multiple computing-enabled relays.

Acknowledgment

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 European Union-NextGenerationEU, 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 and T. Taleb, “The Road beyond 5G: A Vision and Insight of the Key Technologies,” *IEEE Network*, vol. 34, no. 2, pp. 135-141, March/April 2020.

[2] W. Saad, M. Bennis, M. Chen, “A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems”, *IEEE Network*, vol.34, 2020.

[3] AIOTI, “High Priority Edge Computing Standardisation Gaps and Relevant SDOs”, April, 2022.

[4] N. Bhushan, et al., ‘Network Densification: The Dominant Theme for Wireless Evolution into 5G’. *IEEE Comm. Magazine*, February 2014

[5] S. Rangan, T. S. Rappaport, E. Erkip, ‘Millimeter-Wave Cellular Wireless Networks: Potentials and Challenges’, *Proc. of the IEEE,* March 2014

[6] J. G. Andrews, et al., ‘What Will 5G Be?’, *IEEE Journal on Selected Areas in Communications*, June 2014

[7] M. Caprolu, et al., "Edge Computing Perspectives: Architectures, Technologies, and Open Security Issues," 2019 *IEEE Int. Conf. on Edge Computing (EDGE)*, Milan, Italy, 2019, pp. 116-123.

[8] Y. Mao, et. al, "A Survey on Mobile Edge Computing: The Communication Perspective," in *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2322-2358, Fourthquarter 2017.

[9] Yun Chao Hu, M. Patel, D. Sabella, N. Sprecher, V. Young, “Mobile Edge Computing A key technology towards 5G”, *ETSI, White Paper* #11, France, September 2015.

[10] N. Abbas, et. al, "Mobile Edge Computing: A Survey," in *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 450-465, Feb. 2018.

[11] N. Sprecher, et al, “Harmonizing standards for edge computing, a synergized architecture leveraging ETSI ISG MEC and 3GPP specifications”, *ETSI White paper #36,* France, 2020.

[12] J. Sydir, R. Taori, “An Evolved Cellular System Architecture Incorporating Relay Stations”, *IEEE Comm. Magazine*, June 2009

[13] O. Teyeb, et al., "Integrated Access Backhauled Networks", *IEEE 90th Vehicular Technology Conference* (VTC-2019 Fall), 2019.

[14] M. Polese, et al. “Integrated Access and Backhaul in 5G mmWave Networks: Potential and Challenges”, *IEEE Comm. Magazine*, 2020.

[15] 3GPP TR 22.839 v18.1.0, “Study on Vehicle-Mounted Relays; Stage 1 (Release 18)”, December, 2021.

[16] R. Balakrishnan, et. al, “Mobile Relay and Group Mobility for 4G WiMAX Networks”, *IEEE Wireless Comm. and Net, Conf.*, 2011.

[17] S. Andreev, et. al., “Future of UltraDense Networks Beyond 5G: Harnessing Heterogeneous Moving Cells”, *IEEE Comm. Mag.zine*, June 2019.

[18] 3GPP TS 22.261 v18.5.0, “Service requirements for 5G system; Stage 1 (Release 18)”, December, 2021.

[19] J. Pérez-Romero, O. Sallent, "Leveraging User Equipment for Radio Access Network Augmentation", *IEEE Conference on Standards for Communications and Networking (CSCN)*, December, 2021.

[20] J. Liang, Z. Chen, C. Li and B. Xia, "Delay Outage Probability of Multi-relay Selection for Mobile Relay Edge Computing System," 2019 *IEEE/CIC International Conference on Communications* in China (ICCC), Changchun, China, 2019, pp. 898-902.

[21] X. Cao, et. al, "Joint Computation and Communication Cooperation for Energy-Efficient Mobile Edge Computing," in *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4188-4200, June 2019.

[22] M. Yao, et. al, "Energy Efficient Cooperative Edge Computing with Multi-Source Multi-Relay Devices," 2019 *IEEE 21st Int. Conf. on High Performance Computing and Comm.*; *IEEE 17th Int.. on Smart City; IEEE 5th Int. Conf. on Data Science and Systems (HPCC/SmartCity/DSS)*, Zhangjiajie, China, 2019, pp. 865-870.

[23] X. Chen, et. al, "Joint Cooperative Computation and Interactive Communication for Relay-Assisted Mobile Edge Computing," 2018 *IEEE 88th Veh. Tech. Conf.* (VTC-Fall), Chicago, USA, 2018, pp. 1-5.

[24] 3GPP TS 38.300 V17.3.0, “NR and NG-RAN Overall Description; Stage 2 (Release 17)”, December, 2022.

[25] G. Noh, H. Chung, I. Kim, "Mobile Relay Technology for 5G", *IEEE Wireless Communication*s, June 2020.

[26] G. Noh et al., “Realizing Multi-Gb/s Vehicular Communication: Design, Implementation, and Validation,” *IEEE Access*, vol. 7, Jan. 2019, pp. 19,435–19,446

[27] “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).

[28] 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).

[29] 3GPP TS 38.306 v17.2.0, “NR User Equipment (UE) radio access capabilities (Release 17),” September 2022.

[30] 3GPP TS 38.214 v15.5.0, “NR; Physical layer procedures for data (Release 15),” March 2019.