Aalborg Universitet

AALBORG UNIVERSITY
DENMARK

Network-coded cooperation and multi-connectivity for massive content delivery

Leyva-Mayorga, Israel; Torre, Roberto; Pla, Vicent; Pandi, Sreekrishna; Nguyen, Giang T.; Martinez-Bauset, Jorge; Fitzek, Frank H. P.

Published in:
IEEE Access

DOI (link to publication from Publisher):
10.1109/ACCESS.2020.2967278

Creative Commons License
CC BY 4.0

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

Citation for published version (APA):
Leyva-Mayorga, I., Torre, R., Pla, V., Pandi, S., Nguyen, G. T., Martinez-Bauset, J., & Fitzek, F. H. P. (2020). Network-coded cooperation and multi-connectivity for massive content delivery. *IEEE Access*, *8*, 15656-15672. [8962074]. https://doi.org/10.1109/ACCESS.2020.2967278

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

? Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
? You may not further distribute the material or use it for any profit-making activity or commercial gain
? You may freely distribute the URL identifying the publication in the public portal ?

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from vbn.aau.dk on: November 25, 2020

IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received December 13, 2019, accepted January 7, 2020, date of publication January 17, 2020, date of current version January 27, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2967278

Network-Coded Cooperation and Multi-Connectivity for Massive Content Delivery

ISRAEL LEYVA-MAYORGA[1], ROBERTO TORRE[2], VICENT PLA[3], SREEKRISHNA PANDI[2], GIANG T. NGUYEN[2], JORGE MARTINEZ-BAUSET[3], AND FRANK H. P. FITZEK[2]

[1]Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark
[2]Deutsche Telekom Chair of Communication Networks, Technische Universität Dresden, 01062 Dresden, Germany
[3]ITACA Research Institute, Universitat Politècnica de València, 46022 Valencia, Spain

Corresponding author: Israel Leyva-Mayorga (ilm@es.aau.dk)

This work was supported in part by the European Union's H2020 Research and Innovation Program under Grant H2020-MCSA-ITN-2016-SECRET 722424. The work of Vicent Pla and Jorge Martinez-Bauset was supported under Grant PGC2018-094151-B-I00 and Grant RED2018-102585-T (MCIU/AEI/FEDER,UE).

ABSTRACT Massive content delivery is in the spotlight of the research community as both data traffic and the number of connected mobile devices are increasing at an incredibly fast pace. The enhanced mobile broadband (eMBB) is one of the main use cases for the fifth generation of mobile networks (5G), which focuses on transmitting greater amounts of data at higher data rates than in the previous generations, but also on increasing the area capacity (given in bits per second per square meter) and reliability. However, the broadcast and multicast implementation in 5G and presents several drawbacks such as unexpected disconnections and the lack of device-specific QoS guarantees. As a result, whenever the exact same content is to be delivered to numerous mobile devices simultaneously, this content must be replicated. Hence, the same number of parallel unicast sessions as users are needed. Therefore, novel systems that provide efficient massive content delivery and reduced energy consumption are needed. In this paper, we present a network-coded cooperation (NCC) protocol for efficient massive content delivery and the analytical model that describes its behavior. The NCC protocol combines the benefits of cooperative architectures known as mobile clouds (MCs) with Random Linear Network Coding (RLNC). Our results show the benefits of our NCC protocol when compared to the establishment of numerous parallel unicast sessions are threefold: offload data traffic from the cellular link, reduce the energy consumption at the cooperating users, and provide throughput gains when the cellular bandwidth is insufficient.

INDEX TERMS Cooperation, fifth generation of mobile networks (5G), massive content delivery, random linear network coding (RLNC).

I. INTRODUCTION

Wireless data traffic is increasing dramatically. For instance, the amount of traffic transmitted in 2016 grew 63 percent when compared to 2015 and a monthly data traffic of 49 exabytes is expected by 2021 [1]. This represents an increase of around 700 percent with respect to 2016 and more than three quarters of this traffic will be caused by mobile video. Such a dramatic increase in data traffic poses important challenges to the current 4th generation of mobile networks (4G), but also to the fifth generation (5G), whose deployment has already begun.

The associate editor coordinating the review of this manuscript and approving it for publication was Congduan Li.

A particularly problematic scenario is that of massive content delivery (MCD) applications. In the latter, the exact same content is to be delivered to a large number of user equipments (UEs). Live video streaming (e.g., of popular cultural or sporting events), gaming, and virtual reality are of the most relevant MCD applications in 5G. Distributed ledger technologies (DLTs) (e.g., Blockchain) present another use case for MCD, where every DLT node has to store a copy of a common timestamped and ordered database called *ledger*. The ledgers store the transactions generated in the network and, to ensure consistency in the local copies, the overall ledger update rate is limited by the weakest wireless link.

In the applications described previously, the UEs must receive the content in a synchronized manner.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

Hence, the overall QoS may be greatly affected by UEs with poor wireless conditions. In contrast, there is another set of MCD applications where the data delivery from the source (e.g. BS) to the UEs and its distribution between UEs are not synchronized. Content caching, for example, presents two main challenges: the prediction and placement of the content, and its distribution among the UEs (i.e., MCD). Classic use cases for content caching are traditional video streaming services, location services, and software updates.

The popularity of all the MCD applications described above is growing rapidly due to the great capabilities of modern smartphones and the ever-increasing communication demands of the population. Hence, implementing efficient MCD mechanisms is one of the major communication challenges for 5G.

The research community has been aware of the communication problems that may arise in MCD applications. Consequently, several systems have been deployed in order to provide multicast capabilities to the cellular base stations (BSs). Among these, the Multimedia Broadcast Multicast Service (MBMS) [2] took advantage in the first years after the deployment of 4G. The MBMS is a multicast implementation through small cells defined in the 3rd Generation Partnership Project (3GPP) 4G and 5G specifications [3]. However, several studies found the 4G implementation of MBMS and its evolution, the evolved MBMS (eMBMS), suffer from unexpected disconnections, reduced transmission range, high energy consumption, and poor spectral efficiency [2], [4], [5]. The first phase of standardization of 5G brought no major enhancements to the MBMS. Therefore, these problems are still present in 5G [6].

As a result of the inefficiency of the MBMS system and the lack of other mechanisms for MCD, the state-of-the-art solution in MCD applications is to transmit the exact same content to numerous UEs simultaneously from the cellular BS. In other words, the UEs that request access to a given content (e.g., video streaming) from a cellular BS are connected via a unicast link, regardless of the number of UEs that request the exact same content. Hence, the number of simultaneous unicast sessions is equal to the number of requesting UEs. Needless to say, this approach is highly inefficient in terms of resource utilization.

Cooperative mobile clouds (MCs) are a promising solution to the described MCD scenario [7]. An MC is a cooperative architecture in which a group of UEs share the available wireless resources opportunistically [8]. In an MC, UEs usually have multiple radio access technologies (RATs) and multi-connectivity (MCo) capabilities. That is, the multiple RATs can be used simultaneously. This allows them to use the cellular link to download content directly, and then cooperate through a short-range technology, such as WiFi, which reduces bandwidth utilization the cellular link. In addition, MCs may also reduce the energy consumption of the UEs if less energy is needed for communication in the short-range.

Some cooperative content delivery systems have been proposed in the literature, but these oftentimes consider unicast transmissions in the short-range [9], [10]. Since the UEs within a MC are closely located, the use of multicast short-range links for content delivery is possible and much more efficient than the use of independent unicast sessions. It is in multicast wireless networks where Network Coding (NC) schemes have proven to be highly valuable to ensure a high data rate and a low error ratio [11]. In NC, the transmitter segments the content into batches of packets, known as generations. Then, it combines the packets contained in its coding matrix to create coded packets. As a result, the NC transmission schemes focus on the delivery of sufficient packets to the destination to decode the generation.

One of the most widely used NC schemes is Random Linear Network Coding (RLNC) [12], where each packet is multiplied by a coefficient chosen randomly from a Galois-field of size q, denoted GF(q). Full-vector RLNC is the most common and simple variant of RLNC in which every transmitted packet is coded. On the other hand, the systematic RLNC is a variant in which the packets in a generation are first sent without coding; these are known as source packets. Then, coded packets are transmitted to recover the errors that occurred during the transmission of the source packets. It has been observed that systematic RLNC results in a higher probability of decoding the generation and reduces the decoding complexity at the UEs when compared to full-vector RLNC [13]. Fig. 1 illustrates the difference between traditional data transmission with feedback, full-vector RLNC, and systematic RLNC.

The combination of cooperative approaches such as MCs with RLNC schemes has led to the innovative communication paradigm of Network-Coded Cooperation (NCC) [14], [15]. Research has shown that NCC has the potential to provide increased performance in multicast applications [16].

In this paper, we propose an NCC protocol for massive content delivery in cellular networks. It comprises two phases, namely the cellular and NCC phases. In the cellular phase, the BS segments the requested content in batches of size g packets; hereafter, we refer to the batch size, g, as the generation size. These g packets are transmitted to an MC through multiplexed unicast links. Then, in the NCC phase, the UEs cooperate under the systematic RLNC scheme to distribute these packets through multicast WiFi links.

One of the main drawbacks in existing cooperative systems is the transmission of a large number of feedback messages within the MCs, which are needed to keep track of the state of the UEs [17]. Hence, in this study we propose to eliminate the transmission of feedback messages from the UEs and, instead, use analytical models to calculate the optimal number of coded transmissions. That is, the minimum number of coded transmissions needed to deliver the content to the MC and fulfill the reliability requirements.

As it will be observed throughout this paper, it is complicated to develop an exact model for the proposed NCC protocol. In particular, two challenges must be overcome. The first one is to solve an RLNC multicast problem with

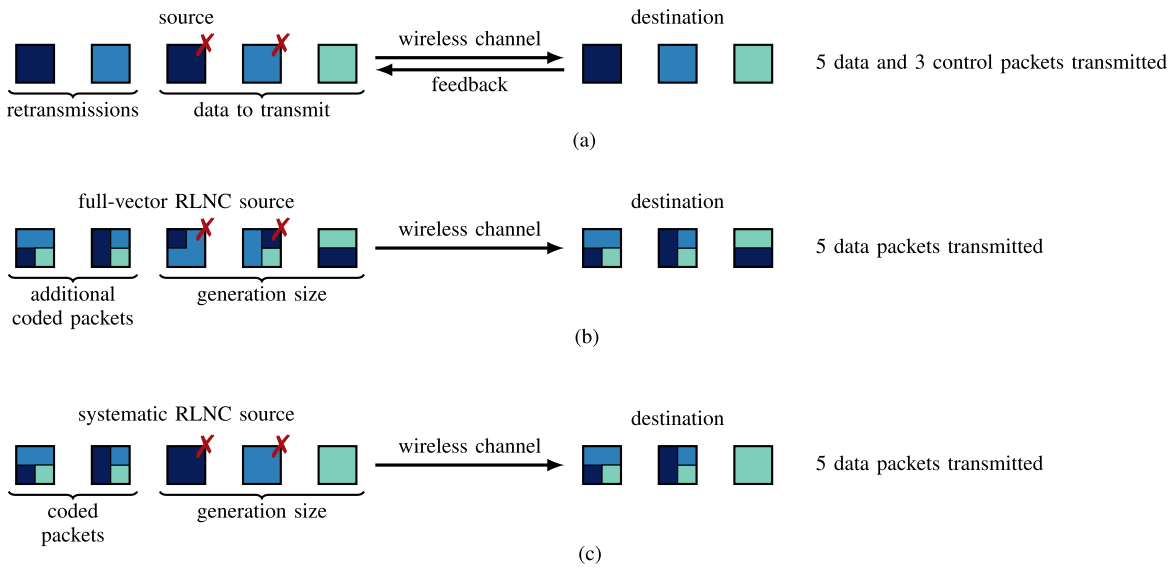

FIGURE 1. Unicast data transmission schemes: (a) traditional with feedback for each successfully received packet; (b) full-vector RLNC; and (c) systematic RLNC. In each case, the second and third packets are lost due to wireless channel errors.

multiple sources. That is, the content is distributed among the UEs in the MC and the packets received at each node are not present at the remaining UEs. As a reference, single-source multicast scenarios under RLNC schemes have been studied in the literature and the formulation of the exact probability of delivering the content to every destination is not trivial [18]. Concretely, exact formulations only exist for the case of one source and two destinations and lower bounds must be used for a higher number of destinations.

In addition to the RLNC multicast problem, the packets generated from each source are incorporated to the coding matrices of the rest of the sources. That is, the second challenge is to model the inclusion of packets received from both, the BS and MC neighbors in the coding matrices of the UEs. This characteristic enhances the throughput when compared to other policies such as only including the packets received directly from the BS in the coding matrix [19]. However, it greatly increases the complexity of the analysis. Therefore, we developed two analytical models that provide tight upper bounds for the probability of successfully delivering the content to the MC; their accuracy is assessed in Section IV.

By using our models, the optimal number of coded transmissions, the achievable throughput, and the energy consumption of the UEs can be calculated with a high precision, given the UEs have accurate channel state information (CSI) on the short-range links. Nevertheless, our models are greatly valuable to, for example, calculate approximate values of the number of coded packets needed even if the CSI is not accurate. In these cases, our model may be used in combination with simple feedback mechanisms. For example, the UEs may send a short ACK through the physical uplink control channel (PUSCH) towards the cellular BS to indicate the generation has been decoded and the next generation can be transmitted.

Our results show that, when compared to parallel unicast content delivery, our protocol leads to energy savings of more than 40 percent while greatly reducing the data transmitted through the cellular link. In addition, up to a three-fold increase in throughput can be achieved with MCo. That is, by using the WiFi and cellular interfaces simultaneously.

The remainder of the paper is organized as follows. The state-of-the-art is presented in Section II. Then, we describe our NCC protocol in Section III and the analytical model in Section IV. We present the parameters that lead to the adequate configuration of the system, along with the achievable gains and overhead in Section V. Section VI concludes the paper.

II. RELATED WORK

The MBMS system is the multicast implementation for 3GPP network. It was developed for 4G (LTE and LTE-A) and no major enhancements were incorporated in 5G. However, it has several drawbacks that make it inappropriate when a reliable content delivery is needed [6]. One of the most important ones is that it suffers from unexpected disconnections and lacks mechanisms to provide the necessary QoS requirements to individual UEs. As a consequence, diverse solutions to the MCD have been developed. For instance, the idea of organizing microcells in cloudlets was first described in [20]. Cooperative relaying was proved to increase network performance in [21] whereas the advantages of Network Coding were first shown in [11]. Moreover, the interplay between subgrouping in cloudlets and network coding was first proposed in [14].

Despite the clear advantage of short-range NC multicast in the cloud, most existing cooperative systems incorporate unicast short-range data transmissions. Some examples are the Microcast [9], and CoopStream [10] systems, whose main focus is to offload data traffic from the BS.

TABLE 1. Comparison between related systems.

		LTE-A	Short-range	FEC
eMBMS	[2]	✓	✗	Raptor Codes
MicroCast	[9]	✓	Unicast	Network Coding
CoopStream	[10]	✓	Unicast	RLNC
NCVCS	[22]	✗	Multicast	Network Coding
NCC system	[15]	✓	Multicast	RLNC

Clearly, the performance of all these previous technologies might increase by using multicast instead of unicast in the short-range. For instance, Wang *et al.* [22] demonstrated the advantages of multicast in their NC-based Video Conference System for mobile devices in multicast network (NCVCS).

The main motivation for this paper is the NCC system first proposed in [23], whose main focus was to offload the LTE-A BS, but also, important throughput and energy gains were observed. Consequently, two demonstrators were developed [15]. Then, initial design of our protocol and a basic performance evaluation was conducted [24]. Our initial results showcased the potential gains provided by NCC, but also the great complexity of its analysis.

Regarding the analytical modeling of multicast with RLNC, a thorough study on the decoding probability in a one-source multicast scenario with both, full-vector and systematic RLNC was conducted by Tsimbalo *et al.* [18]. In the latter study, the importance of the correlation between the packets received at each node is showcased, and the authors concluded that the effect of this correlation is only negligible for the systematic RLNC but not for full-vector RLNC. As it will be seen in Section IV, we deal with a similar but more complex problem because, in our NCC protocol: 1) content delivery within the MCs is performed through multiple multicast sessions, one for each UE; 2) the BS distributes the data packets among the UEs, which creates numerous sub-batches of packets; and 3) coding is performed by combining the packets received from both the BS and from neighboring UEs.

III. PROPOSED NCC PROTOCOL

In this paper, we propose and evaluate the performance of an NCC protocol for efficient MCD. The first step towards cooperation is the creation of groups of UEs called MCs. For this, let n be the number of UEs in an MC, hereafter referred to as the cloud size. The cellular BS is in charge of creating the MC with UEs that: 1) have a direct cellular link to the same BS; 2) request access to the exact same content; and 3) are fully interconnected by a short-range technology, for example, WiFi. It is out of the scope of this paper to develop the rules and the protocol for the formation of the MC. Instead, we focus on the content delivery once the MC has been formed.

Content delivery occurs in two phases: the cellular and NCC phases. In the cellular phase, the BS segments the requested content in batches of g data packets; hence, g is the generation size. These g packets are transmitted to an MC through n unicast sessions. However, the packets transmitted through each unicast session are different. Therefore, at the end of the cellular phase, each packet is only present at one out of the n UEs in the MC. Then, at the NCC phase, the UEs first transmit the packets received from the BS without coding via multicast WiFi links. Afterwards, the UEs generate and transmit coded packets to recover the errors that may have occurred during the previous transmissions. Again, these are transmitted through WiFi multicast links. The cellular and NCC phases are now described in detail.

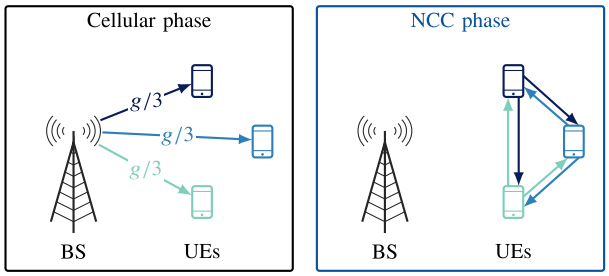

FIGURE 2. The proposed NCC protocol comprises two phases: (a) cellular and (b) NCC. In the former, the BS distributes the packets in the generation among the UEs in the MC. In the latter, the UEs cooperate so that the generation is delivered to every UE in the MC.

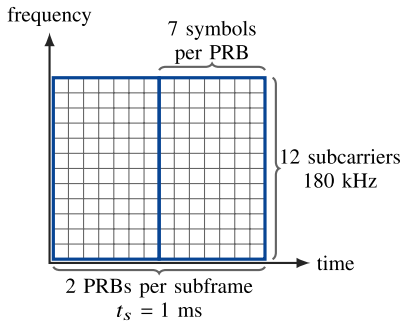

FIGURE 3. Structure of the physical resource block (PRB) in 4G and 5G.

A. CELLULAR PHASE

The BS transmits the g data packets to the n UEs through n unicast sessions. Data transmission in 5G takes place in a time-slotted channel, whose minimum scheduling unit is one subframe, with duration $t_s = 1$ ms [25]. On the other hand, the minimum unit for data transmission (downlink) is the physical resource block (PRB), which is defined as a set of seven consecutive orthogonal frequency-division multiplexing (OFDM) symbols in the time domain and twelve consecutive subcarriers in the frequency domain [25], as shown in Fig. 3. In the time domain, two PRBs fit in one subframe.

We assume the n unicast sessions are multiplexed, either in time or in frequency. For this, each of the n UEs is assigned an index, in the set $\mathcal{N} = \{i \in \mathbb{Z}_+ \mid i \leq n\}$, that defines the order in which they will receive the data packets from the BS. By following this scheme, a total of $g_i < g$ packets are transmitted to the ith UE, and

$$g = \sum_{i=1}^{n} g_i \tag{1}$$

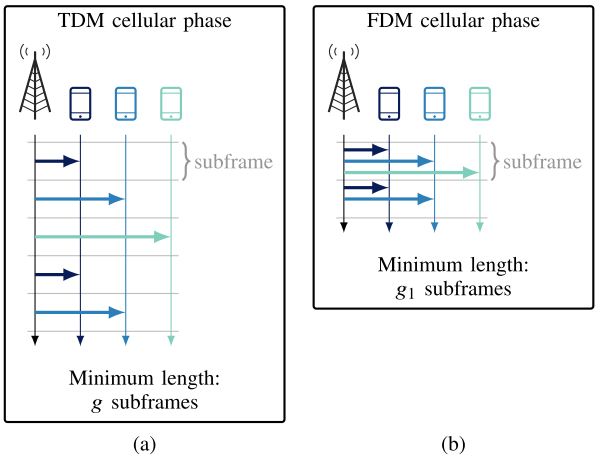

FIGURE 4. **(a) TDM and (b) FDM in the cellular phase given $n = 3$ and $g = 5$.**

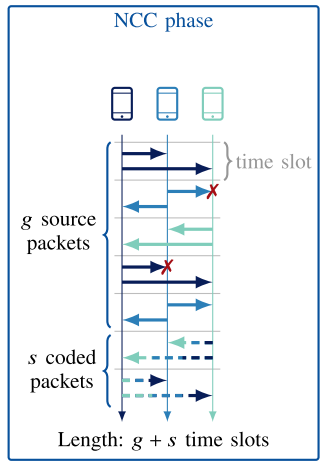

FIGURE 5. **The NCC phase is performed following a TDMA schedule. The errors that occurred in the second and fourth time slots of the NCC phase are recovered with the two coded packet transmissions.**

If time-division multiplexing (TDM) is used at this phase, only one data packet transmission occurs simultaneously towards the MC. On the other hand, if frequency-division multiplexing (FDM) is used, the number of simultaneous data packet transmissions towards the MC is the minimum between n and the maximum number of simultaneous data packet transmissions that can be accommodated in one cellular carrier. The latter is determined by the cell bandwidth and the selected data rate. Please observe FDM transmission may not be possible under certain applications that generate the data on the fly, where the throughput of the BS is limited by the rate at which data is generated and arrives to the BS. Live video streaming applications are clear examples of such applications. For these, TDM unicast must be used.

Fig. 4 illustrates the cellular phase with both TDM and FDM for $n = 3$ and $g = 5$ for a case where no errors occur during the transmissions.

B. NCC PHASE

The UEs are in charge of redistributing the g packets received from the BS in the MC. Since no feedback messages are transmitted, the BS must inform the number of time slots allocated for the content delivery within the MC to the UEs.

The index assigned to each UE in the cellular phase is used to create a TDMA schedule. At each time slot, a UE performs a multicast packet transmission in the short-range link to the remaining $n - 1$ UEs in the MC. Hereafter, we denote $\mathcal{N}_i = \mathcal{N} \setminus i$ as the set of neighbors of the ith UE; $|\mathcal{N}_i| = n - 1$. The transmitting UE changes at each time slot to uniformly distribute energy consumption among the MC members. Please observe that the time slot duration at this phase is not necessarily the same as that of the subframe at the cellular link, hence a higher or lower data rate can be used.

At the end of the cellular phase, g_i packets are present at the ith UE and these are not present in the remaining $n - 1$ UEs. The systematic RLNC scheme is implemented in this phase, hence, g_i packets are transmitted without coding by the ith UE. Therefore, the first g packet transmissions within the MC are not coded; hereafter we refer to these as the source packet transmissions. That is, each UE will forward the packets received at the cellular phase. Then, coded packet transmissions are performed in order to recover the errors that may have affected the g source packets.

Exactly g time slots are needed for the transmission of the g source packets. On the other hand, the BS has to calculate and transmit s, the number of time slots allocated for the transmission of coded packets, to the UEs in the MC. When $g + s$ time slots have elapsed in the NCC phase, the BS continues with the transmission of the next generation if needed, hence a new cellular phase begins. Otherwise, data transmission is terminated.

The timing diagram at the NCC phase for $n = 3$, $g = 5$, and $s = 2$ is illustrated in Fig. 5. In the diagram, an error occurs at the second and one more at the fourth time slot. These errors can be recovered with the two coded packet transmission because packets transmitted by neighboring UEs are included in the coding matrix of every UE.

Throughout this study, we consider that cellular and WiFi transmissions can be performed either one at a time or at the same time. Therefore, UEs may support either single-connectivity (SCo) or MCo. If only SCo is supported, cellular and NCC phases can only occur one after the other. On the other hand, if MCo is supported, these can occur simultaneously, as long as the UEs in the MC have data to transmit in the short-range. That is, source packets in the NCC phase are only transmitted once. Consequently, if some packets are lost in the cellular phase, the implicated UEs must wait until these are retransmitted and received to perform the corresponding source packet transmissions. In the following, we provide two analytical models to calculate adequate values for parameter s and to evaluate the benefits of the proposed NCC protocol.

IV. ANALYTICAL MODELS

This section presents two analytical models to that describe the operation of the proposed NCC protocol and can be used to optimize its operation.

Let s be the number of coded packet transmissions performed in the MC from every $i \in \mathcal{N}$. The value of parameter s is selected by the BS and transmitted to the UEs in the MC. Building on this, we seek to obtain s^*, defined as the minimum value of s needed to deliver the content with the desired reliability τ. Once s^* has been obtained, the maximum throughput and the minimum average energy consumption per UE will be calculated.

The focus of our models is to find the optimal number of transmissions in the NCC phase of the proposed protocol. Therefore, we assume that wireless channel errors in the cellular link are due to interference and small-scale fading, and can be recovered by the implemented hybrid automatic repeat request (HARQ) mechanisms of 5G without impacting the NCC phase.

To find s^*, let S be the RV that defines the total number of coded packet transmissions needed to deliver the content to the n UEs in the MC. Therefore, S has a phase-type (PH) distribution that describes the probability that the UEs in the MC are able to decode the generation.

In order to decode the generation, the coding matrix of a UE must be full rank. This occurs when the matrix has exactly the same number linearly independent rows as columns. The linearly independent rows are commonly known as degrees of freedom (DOFs). The rank of a matrix can be calculated by performing Gaussian elimination so to have the matrix reduced to row echelon form and then counting the number of ones in the diagonal, these ones are known as pivots.

Hereafter, we refer to S as the probability of successful content delivery, whose support is the number of time slots allocated for the transmission of coded packets s. Building on this, s^* is defined as

$$s^* \triangleq \min_s \{s \mid F_S(s; n) \geq \tau\} \tag{2}$$

where F_S is the cumulative distribution function (CDF) of RV S. That is, τ is a threshold for S and its value must be selected depending on the needs of the content delivery application. The process to calculate S is described in the following.

At the end of the cellular phase, g source packets have been distributed among the n UEs in the MC following a round-robin scheduling. We have previously defined g_i as the total number of data packets received by the ith UE in the cellular phase. These g_i packets are exclusively at the BS and the ith UE. Then, it follows that g_i is also the number of source packet transmissions performed by the ith UE in the NCC phase, and is given as

$$g_i = \left\lceil \frac{g - (i-1)}{n} \right\rceil. \tag{3}$$

Please recall $\mathcal{N}_i$ is the set of neighbors of the ith UE. Now, we define the time index $s_i \in \mathbb{Z}_{\geq 0}$ as the number of coded packet transmissions towards the ith UE (i.e., from every $j \in \mathcal{N}_i$). Time index s_i is a function of s, the UE index i, the cloud size n, and the generation size g given as

$$s_i = f(s, i, n, g) = s + g_i - \left\lceil \frac{g + s - (i-1)}{n} \right\rceil. \tag{4}$$

Please observe that time index s_i may be different for each UE at each s. Next, let $\left\{X_{s_i}^{(i)}\right\}_{s_i \in \mathbb{N}}$ be the stochastic process that defines the rank of the coding matrix of the ith UE, whose support is $x = \{0, 1, \ldots, g\}$.

The RV of the stochastic process defined above at $s_i = 0$ is $X_0^{(i)}$ and defines the rank of the coding matrix of the ith UE at the end of the source packet transmissions at the NCC phase. This will later used as the vector of initial probabilities for the DTMC that describe the transmission of the coded packets in the MC.

Like in the cellular link, we assume that wireless channel errors in the short-range link are due to interference and small scale fading. Therefore, the coherence time of the channel is in the order of one subframe and errors in subsequent subframes occur independently. As a consequence, we assume the short-range link between the ith and the jth UEs is fully characterized by ϵ_j, which is the packet error ratio (PER).

To obtain the probability mass function (pmf) of $X_0^{(i)}$, let $X_0^{(ij)}$ be the RV that defines the number of source packets transmitted successfully to the ith UE from the jth UE. Hence, the pmf of $X_0^{(ij)}$ is given as

$$\Pr\left[X_0^{(ij)} = x\right] \triangleq \binom{g_j}{x} (1 - \epsilon_j)^x \, \epsilon_j^{g_j - x}. \tag{5}$$

As a consequence, the pmf of $X_0^{(i)}$ is defined as

$$p_{X_0}(x; i) = \Pr\left[X_0^{(i)} = x\right] \triangleq \sum_{j \in \mathcal{N}_i} X_0^{(ij)}. \tag{6}$$

Given that each pair of RVs $X_0^{(ij)}$ for all $j \in \mathcal{N}_i$ are independent, the pmf of $X_0^{(i)}$ can be calculated as a series of $n - 2$ discrete-time convolutions.

For notation simplicity, hereafter we assume $\epsilon = \epsilon_j$ for all $j \in \mathcal{N}_i$. That is, we assume the same PER between the ith UE and each of his neighbors. This is a valid assumption since the UEs within an MC are closely located and form an interference-limited network. Hence, the levels of interference experienced by two UEs in the same MC are greatly similar. However, a generalization of our simplified model, presented in Section IV-A, can be easily obtained to accommodate different PERs between each UE pair.

Building on this assumption, the pmf of $X_0^{(i)}$ can be calculated as

$$p_{X_0}(x; i) = \binom{g - g_i}{x - g_i} (1 - \epsilon)^{x - g_i} \, \epsilon^{g - x}. \tag{7}$$

Please observe that, since only source packets have been transmitted up to this point, $X_0^{(i)}$ is also the number of non-zero columns in the coding matrix of the ith UE at $s_i = 0$.

Next, coded packet transmissions are performed at every $s_i \in \mathbb{Z}_+$. At this point, we define the RV $S^{(i)}$ as the number of coded transmissions from every $j \in \mathcal{N}_i$ that are needed for the coding matrix of the ith UE to be full rank. Clearly, the RV $S^{(i)}$ also has a PH distribution whose domain is the set of values for time index s_i; the CDF of $S^{(i)}$ is defined as

$$F_S(s_i; i) \triangleq F_{X_{s_i}}(g; i) = \Pr\left[X_{s_i}^{(i)} = g\right]. \tag{8}$$

Naturally, the distribution of $S^{(i)}$ depends on the PER ϵ and on the probability of linear independence of each of the s_ith coded packet transmissions, denoted as p. Nevertheless, the correlation between the packets received at each pair of UEs is needed in order to obtain the exact value for p. Therefore, we define the stochastic process $Z_{s_i}^{(ij)}$ as the number DOFs that are missing from the coding matrices of both, the ith (receiver) and jth (transmitter) UEs at s_i. The joint pmf of $X_0^{(i)}$ and $Z_0^{(ij)}$ is given as

$$\begin{aligned} &p_{X_0,Z_0}(x, z; i, j) \\ &= \epsilon^{g-x+z} \sum_u \left[\binom{g_j}{u} \binom{\gamma}{x - g_i - u} \right. \\ &\quad \left. \times \binom{\gamma - x + g_i + u}{z} (1-\epsilon)^{\gamma+u-z} \right] \end{aligned} \tag{9}$$

where $\gamma = g - g_i - g_j$ and u represents the number of DOFs in the coding matrix of the ith UE that were transmitted by the jth UE. In other words, u is the number of DOFs transmitted from the jth to the ith UE. The summation in (9) is performed in the set of possible values

$$\left\{u \in \mathbb{Z}_{\geq 0} \;\middle|\; \max\{0, x - \gamma - g_i + z\} \leq u \leq \min\{g_j, x - g_i\}\right\}.$$

Then, the exact value of p for a given x and z is defined as

$$\begin{aligned} p(x, z) &\triangleq \Pr\left[X_{s_i+1}^{(i)} = x + 1 \,\middle|\, X_{s_i}^{(i)} = x \cap Z_{s_i}^{(ij)} = z, \epsilon = 0\right] \\ &= 1 - q^{x+z-g}. \end{aligned} \tag{10}$$

That is, the selected Galois-field size q and the generation size g are the parameters of $p(x, z)$. Naturally, different pairs of UEs $\{i, j\}$ may have different joint distributions of $X_0^{(i)}$ and $Z_0^{(ij)}$, as these depend on g_i, g_j, and γ. Furthermore, the joint pmf of $X_{s_i}^{(i)}$ and $Z_{s_i}^{(ij)}$ is different at each s_i. Therefore, the joint pmf of $X_{s_i}^{(i)}$ and $Z_{s_i}^{(ij)}$ must be calculated for every possible s_i and for each receiver-transmitter pair $\{i, j\}$ in order to calculate the exact $p(x, z)$ at each coded transmission. Specifically, the number of stochastic processes that are needed to describe the exact state of the whole MC is $n + \binom{n}{2}$, whereas the minimum number of stochastic processes to obtain the exact pmf of $S^{(i)}$ is $n + 1$. This makes our problem intractable even for relatively small values of n. For instance, a related problem has only been solved for one transmitter and two receivers by Khamfroush *et al.* [26], but no exact formulations exist for a higher number of receivers.

To better illustrate the complications of this problem, let $\mathscr{S} = \left\{(x, z) \in \mathbb{N}^2 \;\middle|\; x + z \leq g\right\}$ be the state space

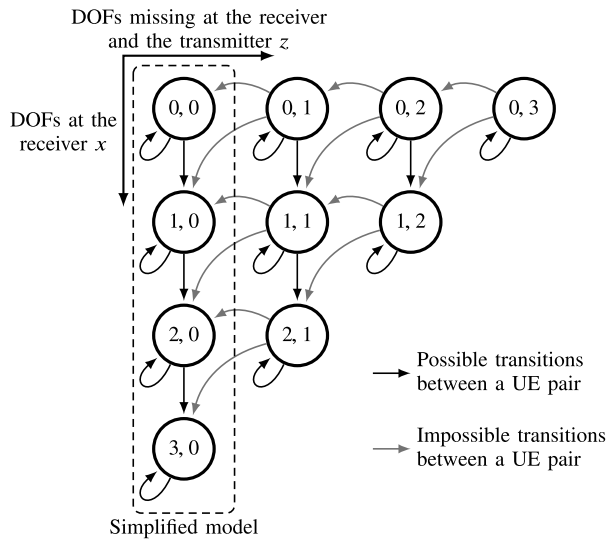

FIGURE 6. Discrete-time Markov chain (DTMC) that describes the transitions at a coded packet transmission for both, the joint and the simplified model. Gray arrows indicate transitions that cannot occur when considering a UE pair $\{i, j\}$. The probability of these transitions is approximated in our joint model.

of the bivariate (two-dimensional) discrete-time Markov chain (DTMC) shown in Fig 6. The latter describes the possible transitions at each coded packet transmission when considering a single UE pair $\{i, j\}$. Naturally, $\Pr\left[X_{s_i}^{(i)} = x + 1 \,\middle|\, X_{s_i-1}^{(i)} = x\right]$ and $\Pr\left[X_{s_i}^{(i)} = x \,\middle|\, X_{s_i-1}^{(i)} = x\right]$ are simply defined by $p(x, z)$, given by (10), and the PER; these transitions correspond to the black arrows in Fig.6. Instead, the state of the system (x, z) does not provide sufficient information to derive the probabilities of decreasing $Z_{s_i}^{(ij)}$ for the next transmissions, where a different $j \in \mathcal{N}_i$ will be the transmitter. Therefore, $\Pr\left[Z_{s_i}^{(ij)} = z - 1 \,\middle|\, Z_{s_i-1}^{(ij)} = z\right]$ and $\Pr\left[Z_{s_i}^{(ij)} = z \,\middle|\, Z_{s_i-1}^{(ij)} = z\right]$ are impossible to derive; these transitions correspond to the gray arrows in Fig. 6.

In the following, we propose two different analytical models to approximate the distribution of $S^{(i)}$. From there, s^* can be calculated. The main difference between the two models relies on the number of agents in the model and, as a consequence, on the number of stochastic processes that are used to describe the state of the system. We refer to these as the simplified and joint models; these are described in the following.

A. SIMPLIFIED MODEL

The simplified model has two agents: the receiver i and the $n - 1$ neighbors of i. These $n - 1$ neighbors and their coding matrices are aggregated into a single transmitter j. By doing so, it is clear that $\Pr\left[Z_0^{(ij)} = 0\right] = 1$. That is, every missing DOF from the receiver i is present at the aggregated transmitter j for all coded packet transmissions $s_i \in \mathbb{N}$. Hence, only the states enclosed by the dashed line in Fig. 6 are considered, for which $z = 0$. Consequently, the state space of the system is $\mathscr{S}_s = \{x \in \mathbb{N} \mid x \leq g\}$. This allows us to use the pmf of $X_{s_i}^{(i)}$

TABLE 2. MSE between the approximate and exact probability of linear independence of the first coded packet transmission.

	$n=3$		$n=100$	
	$g=10$	$g=100$	$g=10$	$g=100$
$\epsilon=0.02$				
$q=2$	$6.95\cdot10^{-5}$	$3.09\cdot10^{-4}$	$1.77\cdot10^{-4}$	$5.58\cdot10^{-4}$
$q=2^8$	$1.64\cdot10^{-8}$	$5.38\cdot10^{-8}$	$4.12\cdot10^{-8}$	$8.20\cdot10^{-8}$
$\epsilon=0.16$				
$q=2$	$2.54\cdot10^{-3}$	$1.33\cdot10^{-5}$	$4.63\cdot10^{-3}$	$5.92\cdot10^{-7}$
$q=2^8$	$4.86\cdot10^{-7}$	$1.53\cdot10^{-10}$	$7.69\cdot10^{-7}$	$1.44\cdot10^{-12}$

alone instead of the joint pmf of $X_{s_i}^{(i)}$ and $Z_{s_i}^{(ij)}$ to calculate $S^{(i)}$. Hence, transitions only depend on

$$p\left(x \mid z=0\right)=1-q^{x-g} \tag{11}$$

and the PER.

Clearly, the results obtained with this simplified model correspond to a tight upper bound in performance. That is, (11) is exact for $n=2$ since $g=g_i+g_j$ in this latter case, which leads to $\Pr\left[Z_{s_i}^{(ij)}=0\right]=1$ for all s_i. For any other cases, the mean squared error (MSE) of the upper bound in (11) can be calculated as

$$\mathrm{MSE}=\sum_{\forall x,z} p_{X_{s_i},Z_{s_i}}\left(x,z \mid i,j\right)\left(q^{x+z-g}-q^{x-g}\right)^2. \tag{12}$$

Table 2 shows the MSE for the first coded transmission in the MC for characteristic values of n, g, ϵ, and q. The first coded transmission for $n=3$ and for $n=100$ is performed by the second and the first UE, respectively. Therefore, the MSE was obtained with $i=1$ and $j=2$ for $n=3$, and with $i=2$ and $j=1$ for $n=100$.

Next, we proceed to provide the exact formulations for the probability of successful delivery to the ith UE $S^{(i)}$ with this model; instead DTMCs will be used to obtain the pmf of $S^{(i)}$ with the joint model, presented below. The rationale behind this is that exact formulations provide a higher degree of tractability and are more illustrative than results derived from DTMCs. Hence, we advocate for the former approach whenever possible.

First, let $\boldsymbol{C}_{r\times c}$ be a coding matrix of dimension $r\times c$ s.t. $r\in\mathbb{N}$ and $\{c\in\mathbb{Z}_+ \mid c\le g\}$, whose elements are selected uniformly at random from GF(q). The probability that matrix $\boldsymbol{C}_{r\times c}$ is full rank, denoted as $F\left(r,c\right)$, is defined as

$$F\left(r,c\right)\triangleq\begin{cases}0 & \text{for } r<c,\\ \prod_{j=0}^{c-1}\left(1-q^{j-r}\right) & \text{otherwise.}\end{cases} \tag{13}$$

Then, we use (13) and (4) to obtain the conditional CDF of $S^{(i)}|X_0^{(i)}$ as

$$F_{S|X_0}\left(s \mid x;i\right)=\sum_{u=g-x}^{s_i}\binom{s_i}{u}\left(1-\epsilon\right)^u\epsilon^{s_i-u}F\left(u,g-x\right). \tag{14}$$

That is, at time index s, s_i out of the total coded packet transmissions are performed by the UEs in $\mathcal{N}_i$ and the remaining $s-s_i$ transmissions are performed by the ith UE. Please observe that s_i corresponds to the number of coded transmissions performed by the aggregated transmitter j in this simplified model.

Next, we calculate the marginal CDF of $S^{(i)}$ from (7) and (14)

$$\begin{aligned}F_S\left(s;i\right)&=\sum_{x=g_i}^{g}p_{X_0}\left(x;i\right)F_{S|X_0}\left(s_i \mid x;i\right)\\&=\sum_{u=g}^{g+s_i}\left(1-\epsilon\right)^{u-g_i}\epsilon^{g+s_i-u}\\&\quad\times\sum_{x=x_{\min}}^{g}\binom{s_i}{u-x}\binom{g-g_i}{x-g_i}F\left(u-x,g-x\right)\end{aligned} \tag{15}$$

where $x_{\min}=\max\{g_i,u-s_i\}$. This concludes the simplified model.

B. JOINT MODEL

The joint model has three agents: the receiver i, the transmitter j, and the $n-2$ neighbors of i and j. These $n-2$ neighbors and their coding matrices are aggregated into a single auxiliary node k. In this model, the exact joint pmf of $X_0^{(i)}$ and $Z_0^{(ij)}$ is calculated and used to approximate the transition probabilities for all z and $s_i\in\mathbb{Z}_+$.

It is easy to observe that each and every one of the g DOFs is either in the coding matrix of i, j, or k, as all of them were transmitted from the BS. Therefore, $\Pr\left[Z_{s_i}^{(ijk)}=0\right]=1$ for all s_i. Consequently, all the missing DOFs from $\{i,k\}$ are in j.

It is also clear that the state of the system with i, j, and k can be completely described by the joint distribution of $X_s^{(i)}, X_s^{(j)}, X_s^{(k)}, Z_s^{(ij)}, Z_s^{(ik)}$, and $Z_s^{(jk)}$. However, only $X_{s_i}^{(i)}, Z_{s_i}^{(ij)}$, $Z_{s_i}^{(ik)}$, and $Z_{s_i}^{(jk)}$ are needed to obtain $S^{(i)}$. Building on this, let $\mathscr{S}_j=\left\{(x,z,z_k)\in\mathbb{N}^3 \;\middle|\; (x,z,z_k)\le g\right\}$ be the state space of the three-dimensional DTMC, where z_k denotes the number of missing DOFs from the pair $\{i,k\}$. From there, it is clear four possible outcomes exist at the s_ith transmission from the jth UE at a given state (x,z,z_k); these are denoted as follows.

- $(x,z,z_k)\to(x+1,z,z_k-1)$: with probability $p_{xz_k}(x,z)$
- $(x,z,z_k)\to(x+1,z,z_k)$: with probability $p_x(x,z)$
- $(x,z,z_k)\to(x,z,z_k-1)$: with probability $p_{z_k}(x,z)$
- $(x,z,z_k)\to(x,z,z_k)$: with probability $p_0(x,z)$

Naturally,

$$p_{xz_k}(x,z)+p_x(x,z)=p(x,z)(1-\epsilon). \tag{16}$$

In words, the probability of increasing x is simply given by the probability of linear independence $p(x,y)$ for the given x and z (see (10)) and the PER ϵ. Analogously,

$$\begin{aligned}p_{z_k}(x,z)+p_0(x,z)&=1-p_{xz_k}(x,z)+p_x(x,z)\\&=1-p(x,z)(1-\epsilon).\end{aligned} \tag{17}$$

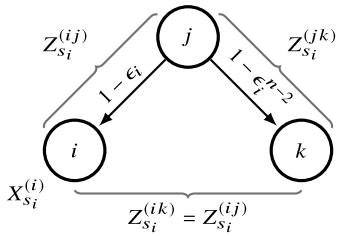

FIGURE 7. Aggregated system described by the joint model.

However, only the joint pmf of $X_0^{(i)}$ and $Z_0^{(ij)}$ is known, and does not provide sufficient information to derive $Z_0^{(ik)}$ nor $Z_0^{(jk)}$. Without this information, the vector of initial states, denoted as α, cannot be calculated.

To solve this problem, we assume all the missing DOFs from node k are at the transmitter j, hence $\Pr\left[Z_0^{(jk)} = 0\right] = 1$ and also that $Z_{s_i}^{(ik)} = Z_{s_i}^{(ij)}$. Please observe that only the states in which $z = z_k$ are possible under this latter assumption. As a consequence, states (x, z, z) form a closed set and the state space of the three-dimensional DTMC $\mathscr{S}_j$ can be easily reduced to that of $\mathscr{S} = \left\{(x, z) \in \mathbb{N}^2 \mid x + z \leq g\right\}$. Analogously, the transition probabilities p_{xz_k} and p_{z_k} are hereafter simply be denoted as p_{xz} and p_z, respectively.

The latter set of assumptions can be interpreted as follows. The UE that will perform a coded transmission after j is contained in auxiliary node k. Then, after each coded transmission, the previous transmitter will be incorporated to k, from where the new transmitter j will be selected. If we assume the same PER between UE pairs, the number of missing DOFs at each UE is highly correlated. Furthermore, it is clear to see that the number of missing DOFs in $\{i, k\}$ decreases only if a coded packet transmission is successfully delivered to the $n-2$ UEs aggregated in k and if it is linearly independent to their aggregated coding matrices. That is, the common information in k increases only if new information arrives to every UE in it. Fig. 7 illustrates the system considered for the joint model at an arbitrary s_i. It includes the three agents, the known pmfs, and the basic assumptions.

Based on this model, the resulting transition probabilities are given as

$$p_{xz}(x, z) = \left(1 - q^{x+z-g}\right)\left(1 - q^{-z}\right) \times (1 - \epsilon)\left(1 - \epsilon^{n-2}\right); \tag{18a}$$

$$p_x(x, z) = \left(1 - q^{x+z-g}\right)(1 - \epsilon) \times \left(\epsilon^{n-2} + q^{-z} - \epsilon^{n-2}q^{-z}\right); \tag{18b}$$

$$p_z(x, z) = \left(1 - q^{-z}\right)\left(1 - \epsilon^{n-2}\right) \times \left(\epsilon + q^{x+z-g} - \epsilon q^{x+z-g}\right); \tag{18c}$$

$$p_0(x, z) = \left(\epsilon^{n-2} + q^{-z} - \epsilon^{n-2}q^{-z}\right) \times \left(\epsilon + q^{x+z-g} - \epsilon q^{x+z-g}\right). \tag{18d}$$

Please observe that the probabilities p_{xz} and p_z listed above correspond to the transitions identified with gray arrows in the two-dimensional DTMC from Fig. 6. With this information, it is now possible to formulate the vector if initial states and the transition matrix that represent the coded transmissions at the NCC phase.

Let $\alpha_{x,z}^{(0)} = \Pr\left[X_0^{(i)} = x \cap Z_0^{(ij)} = z\right]$ denote the probability that the DTMC begins at the transient state $(x < g, z)$. From there, the vector of initial states can be structured as

$$\boldsymbol{\alpha}^{(0)} = \left[\alpha_{0,0}^{(0)}\, \alpha_{0,1}^{(0)} \ldots \alpha_{0,g}^{(0)}\, \alpha_{1,0}^{(0)} \ldots \alpha_{1,g-1}^{(0)} \ldots \alpha_{g-1,1}^{(0)}\right]. \tag{19}$$

Next, let $\boldsymbol{A}(x)$ and $\boldsymbol{B}(x)$ be the substochastic matrices that represent the transitions between transient states at level x. The former is a square matrix with dimension $g - x + 1$ and the dimension of the latter is $(g - x + 1) \times (g - x)$. These are given as

$$\boldsymbol{A}(x) = \begin{bmatrix} p_0(x, 0) & & & & \\ p_z(x, 1) & p_0(x, 1) & & & \\ 0 & p_z(x, 2) & p_0(x, 2) & & \\ & & \ddots & \ddots & \\ & & & p_z(x, g-x) & p_0(x, g-x) \end{bmatrix}$$

$$\boldsymbol{B}(x) = \begin{bmatrix} p_x(x, 0) & & & & \\ p_{xz}(x, 1) & p_x(x, 1) & & & \\ 0 & p_{xz}(x, 2) & p_x(x, 2) & & \\ & & \ddots & \ddots & \\ & & & & p_{xz}(x, g-x) \end{bmatrix}$$

Next, let $\boldsymbol{T}$ be the substochastic matrix that represents the transitions within every possible transient state. The latter is given as

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{A}(0) & \boldsymbol{B}(0) & & & \\ & \boldsymbol{A}(1) & \boldsymbol{B}(1) & & \\ & & \ddots & \ddots & \\ & & & \boldsymbol{A}(g-2) & \boldsymbol{B}(g-2) \\ & & & & \boldsymbol{A}(g-1) \end{bmatrix}$$

Having defined $\boldsymbol{\alpha}^{(0)}$ and $\boldsymbol{T}$, it is easy to calculate the joint pmf of $X_{s_i}^{(i)}$ and $Z_{s_i}^{(ij)} = Z_{s_i}^{(ik)}$ after $s_i \in \mathbb{N}$ coded packet transmissions as

$$\boldsymbol{\alpha}^{(s_i)} = \boldsymbol{\alpha}^{(s_i-1)}\boldsymbol{T}; \tag{20}$$

given the vector of initial states $\boldsymbol{\alpha}^{(0)}$ is known. Finally, the CDF of $S^{(i)}$ can be easily obtained as

$$F_S(s; i) = 1 - \boldsymbol{\alpha}^{(s_i)}\mathbf{1}. \tag{21}$$

C. ASSESSING THE PERFORMANCE OF THE PROPOSED NCC PROTOCOL

There exist numerous possible definitions for the probability of successful content delivery S. For example, Tsimbalo *et al.* [18] define the probability of successful content delivery as the probability that each and every UE in the MC decodes the generation; this same definition of S was used in our previous work [24]. While the definition adopted

by Tsimbalo *et al.* is intuitive, it presents some drawbacks in terms of complexity and, as a consequence, precision. In particular, calculating the exact pmf of S under this definition is not possible for $n \geq 3$, so simplifying assumptions must be made for larger cloud sizes. The drawbacks of this latter definition are further discussed in the Appendix. Instead, in this paper we propose the following definition of S.

Definition 1: Probability of successful content delivery: Let S be the RV that defines the number of coded packet transmissions needed so the coding matrix of any of the $n \in \mathcal{N}$ UEs in the MC is full rank. Therefore, the CDF of S is

$$F_S(s) \triangleq \min_i F_S(s; i) = \min_i \Pr\left[X_{s_i}^{(i)} = g\right] \quad \forall s \in \mathbb{N}. \tag{22}$$

That is, S is the probability that the generation is decoded by any of the UEs in the MC. Clearly, S is a PH distribution ans its definition implies $F_S(s_i; i) \geq F_S(s)$. Furthermore, there exists a value of i at each s, denoted as $i_\ell(s)$, for which $F_S(s; i_\ell) = F_S(s)$. In words, $i_\ell(s)$ is the index of the UE with the lowest probability of decoding the generation at time index s. Therefore, (2) can be rewritten simply as

$$s^* \triangleq \min_s \{s \mid F_S(s; i_\ell) \geq \tau\}. \tag{23}$$

Throughout our experiments, we observed that $i_\ell(s) = (g \bmod n) + 1$ for all s if every UE pair has the same PER.

Once s^* has been obtained, we can calculate the maximum achievable throughput per UE R^*, given in bits per second. For this, let T be RV that defines the length of the cellular phase in subframes. That is, the number of subframes needed to deliver the generation from the BS to the MC. In a general scenario, the distribution of T depends on the selected multiplexing scheme, the generation size g, the PER at the cellular link, and the round trip time (RTT) of the implemented HARQ mechanism in 5G.

Let t^* be the minimum length of the cellular phase in subframes. Clearly, t^* depends on the multiplexing scheme used for unicast data transmission and the generation size g.

In this paper we consider both TDM and FDM. In TDM, we simply have $t^* = g$. On the other hand, in FDM, t^* depends on the generation size g, the cloud size n, the selected cellular data rate R, and the maximum achievable throughput in the carrier B. The latter in turn depends on the selected modulation and coding scheme (MCS) and the carrier bandwidth. Hence, we have

$$t^* = \begin{cases} g & \text{for TDM,} \\ \left\lceil \dfrac{g}{\min\left\{n, \left\lfloor \frac{B}{R} \right\rfloor\right\}} \right\rceil & \text{for FDM.} \end{cases} \tag{24}$$

Note that wireless channel errors in the cellular phase do not affect the operation of the NCC phase with SCo. This is because the NCC phase is performed after the cellular phase. A similar case may arise with MCo because the BS possesses a great deal of channel state information (CSI). Hence, the BS can calculate the minimum number of transmissions in the cellular link that are needed prior to the beginning of the NCC phase, so the latter is not affected by the wireless channel errors in the cellular phase. That is, a time shift between the beginning of the cellular phase and its corresponding NCC phase can be calculated, so the latter can be performed normally.

Building on this, hereafter we assume no errors occur in the cellular phase so $\Pr\left[T = t^*\right] = 1$. This assumption provides a clear picture of the operation of our NCC protocol. On the downside, it minimizes the achievable throughput gains and energy savings w.r.t. traditional unicast transmission.

To proceed with the calculation of the achievable throughput, let ρ be the ratio of WiFi to cellular data rate. The achievable throughput is then given as

$$\begin{aligned} R^*(n) &= \frac{\ell}{t_s} \frac{g}{\delta_t + \frac{1}{\rho}(g + s^*)} \\ &= \frac{R}{\frac{\delta_t}{g} + \frac{1}{\rho}\left(1 + \frac{s^*}{g}\right)}, \quad \text{for } n \geq 2; \end{aligned} \tag{25}$$

where

$$\delta_t = \begin{cases} t^* & \text{for SCo,} \\ 1 & \text{for MCo} \end{cases} \tag{26}$$

indicates the minimum number of subframes between the beginning of a cellular phase and its corresponding NCC phase. That is, at least one packet must be successfully present at the MC so the NCC phase can begin with MCo-capable UEs. It is clear that using FDM in the cellular phase reduces δ_t when compared to TDM, so the former is preferred.

To calculate the minimum average energy consumption per UE $\overline{E}^*_{\text{ue}}(n)$, let $s_i^* = f(s^*, i, n, g)$ as defined by (4) and

$$\mathbb{E}\left[S^{(i)} \,\middle|\, s^*\right] = \sum_{s=0}^{s_i^*} s\, p_S(s; i) \tag{27}$$

be the expected number of subframes that the ith UE is in reception mode and in which coded packets are transmitted. Naturally, $p_S(s; i)$ is the pmf of $S^{(i)}$, which can be easily obtained from its CDF calculated either with the simple (15) or the joint model (21).

Next, we calculate the expected number of source packets received at each of the UEs as

$$\mathbb{E}[X_0] = \frac{1}{n} \sum_{i=1}^{n} \sum_{x=0}^{g} x\, p_{X_0}(x; i). \tag{28}$$

Finally, if we assume that the energy consumption during communication is fixed (i.e., the UEs use a unique power level), we can calculate $\overline{E}^*_{\text{ue}}(n)$ as

$$\begin{aligned} \overline{E}^*_{\text{ue}}(n) = \frac{\ell}{n}\Bigg[& g\, E_{\text{cel,rx}} + (g + s^*)\, E_{\text{wifi,tx}} \\ & + \left((n-1)\, g + \sum_{i=1}^{n} \mathbb{E}\left[S^{(i)} \,\middle|\, s^*\right]\right) E_{\text{wifi,rx}} \\ & + s^* E_{\text{e}}(q) + \Big[n\big(g - \mathbb{E}[X_0]\big)\Big] E_{\text{d}}(q) \Bigg] \end{aligned} \tag{29}$$

TABLE 3. Parameter settings.

Parameter	Symbol	Settings
Carrier bandwidth	–	20 MHz
Achievable carrier throughput in the downlink	B	97.896 Mbps
Cloud size	n	$\{2, 3, \ldots, 100\}$ UEs
Desired reliability	τ	$1 - 10^{-3}$
Generation size	g	$\{32, 64\}$ packets
Galois-field size	q	$\{2, 2^8\}$
Packet error ratio (PER)	ϵ	$\{1, 2, 4, 8, 16\} \cdot 10^{-2}$
Subframe duration	t_s	1 ms
Packet length	ℓ	1470 bytes

where $E_{\text{cel,rx}}$, $E_{\text{wifi,rx}}$, and $E_{\text{wifi,tx}}$ define the energy consumed per bit in the LTE-A transmission, and WiFi reception and transmission, respectively. $E_{\text{e}}(q)$ and $E_{\text{d}}(q)$ are the energy consumed per bit to encode and decode a packet for a given Galois-field size GF(q), respectively.

V. RESULTS

This section presents relevant results regarding the accuracy of our analytical models with respect to simulations and the achievable gains with our NCC protocol. Hence, we first compare the results obtained by our model with those obtained by Monte Carlo simulations. Afterwards, we present the optimal number of coded transmissions s^* as a function of the cloud size n, the generation size g, and the PER ϵ. Finally, we discuss the achievable throughput and energy gains that can be achieved with our NCC protocol.

For our analyses, we selected $g \in \{32, 64\}$ for the generation size and $q \in \{2, 2^8\}$ for the Galois-field size. These are two of the values that provide the highest benefits in RLNC for each of these two configuration parameters; hence, these are widely used in the literature [27], [28]. We also selected a typical cellular bandwidth of 20 MHz, which gives an achievable throughput of 97.896 Mbps when 256 quadrature amplitude modulation (QAM) is used [29, Table 7.1.7.2.1-1].

In the cellular phase, up to one data packet is transmitted from the BS to each UE in the MC per subframe and the packet length is $\ell = 1470$ bytes. The latter is a typical packet length in user datagram protocol (UDP). Given the subframe duration is $t_s = 1$ ms, this gives a cellular data rate per UE of $R = 11.76$ Mbps. Throughout this section, we assume the same PER for each of the WiFi links in the MC ϵ. These and other relevant configuration parameters are listed in Table 3.

Energy consumption parameters for communication in the cellular and WiFi interfaces were obtained from [30] and [31], respectively. The same energy per bit is consumed during transmission and reception over WiFi, and the ratio of WiFi to cellular data rate is denoted by ρ. The energy consumption during encoding and decoding is obtained from the work of Sørensen *et al.* [28], where a Samsung Galaxy S5 was examined. Please observe that, while a different amount of energy can be consumed during encoding than during coding, the difference between these two was observed to be negligible [28]. Hence it is safe to assume $E_{\text{e/d}}(q) = E_{\text{e}}(q) = E_{\text{d}}(q)$.

TABLE 4. Energy consumption parameters.

Model	Symbol	Parameter	Energy (nJ/bit)
Cellular [30]	$E_{\text{cel,rx}}$	$R = 11.76$ Mbps	78.76
WiFi [31]	$E_{\text{wifi,tx}} = E_{\text{wifi,rx}}$	$\rho = 0.5$	55.27
		$\rho = 1$	37.64
		$\rho = 2$	28.82
		$\rho = 4$	24.41
RLNC [28]	$E_{\text{e}} = E_{\text{d}}$	$q = 2$	1.6
		$q = 2^8$	3.5

TABLE 5. JSD between the pmfs of S obtained with the simplified model with respect to simulations and the relative decrease in the JSD obtained when using the joint model instead for $\epsilon = 0.01$.

	$n = 3$		$n = 50$	
	$g = 32$	$g = 64$	$g = 32$	$g = 64$
JSD simplified model $(\cdot 10^{-6})$				
$q = 2$	8.60	96.22	406.18	15.87
$q = 2^8$	8.63	30.54	250.51	151.35
Relative decrease $G_{\text{JSD}} \cdot 10^{-2}$				
$q = 2$	2.28	1.93	1.17	25.32
$q = 2^8$	76.88	13.80	6.42	10.80

The energy consumption parameters used throughout this section are listed in Table 4.

A C-based simulator was developed to assess the accuracy of the analytical models; it incorporates the encoding, transmission, and decoding stages that occur in the proposed NCC protocol. The number of simulation runs is set to ensure the relative margin of error for each point of the pmf of successful content delivery, denoted as $p_{S_{\text{sim}}}(s)$, is less than 0.5 percent at a 95 percent confidence interval.

The accuracy of our model is assessed by means of the Jensen-Shannon Divergence (JSD), which measures the increase in the Shannon's entropy when an approximated pmf is assumed to be the real pmf of an RV. To calculate the JSD, we denote the pmfs of S obtained by our models as $p_S(s)$ and by simulation as $p_{S_{\text{sim}}}(s)$. We assume the latter to be the real pmf of S. Hence, the JSD is calculated as

$$\text{JSD} \triangleq H\left(\frac{p_{S_{\text{sim}}}(s) + p_S(s)}{2}\right) - \frac{H\left(p_{S_{\text{sim}}}(s)\right) + H\left(p_S(s)\right)}{2} \tag{30}$$

where $H(\cdot)$ is the base-e Shannon's entropy. As such, the JSD is upper bounded by $\log 2$ and a JSD of zero indicates both pmfs are identical. Hence, $0 \leq \text{JSD}(\cdot) \leq \log 2$.

Table 5 shows the JSD between $p_{S_{\text{sim}}}(s)$ and $p_S(s)$ obtained with the simplified model JSD_s for typical values of $g \in \{32, 64\}$, $q \in \{2, 2^8\}$, and for widely distinct values of $n \in \{3, 50\}$ given $\epsilon = 0.01$. In addition, Table 5 also shows the relative difference between the latter and the JSD obtained with our joint model JSD_{j}. This relative difference was calculated as

$$G_{\text{JSD}} = 1 - \frac{\text{JSD}_j}{\text{JSD}_s}. \tag{31}$$

Hence, it can be seen as the gain in the Shannon's entropy when using the joint model instead of the simplified model.

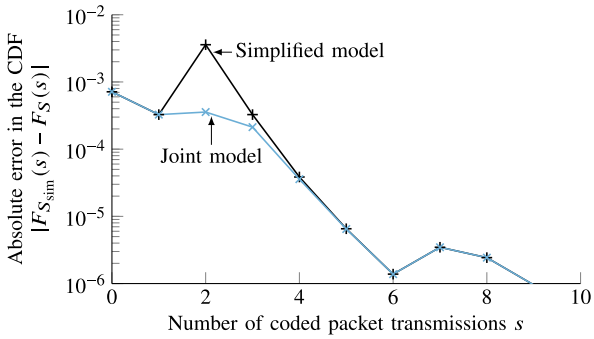

FIGURE 8. Absolute error in the CDF of successful content delivery $|F_{S_{\text{sim}}}(s) - F_S(s)|$ for $g = 64$, $q = 2^8$ and $\epsilon = 0.01$.

The value $\epsilon = 0.01$ was used to assess the impact of the simplifications introduced to our models with respect to the number of missing DOFs from the UE pairs; these in turn affect the probability of linear dependence of the coded packets. As it can be seen, the JSD obtained with the simplified model relatively low regardless of the cloud and generation sizes. As a reference, the minimum JSD for $n = 2$, where our formulations are exact, in combination with $g \in \{32, 64\}$, $q \in \{2, 2^8\}$, and $\epsilon = 0.01$ is $28.37 \cdot 10^{-6}$. The latter was obtained for $g = 32$ and $q = 2^8$. More than ten million simulations were performed for each these cases, hence, we consider all cases that lead to a comparable or lower JSD to be exact.

Table 5 also shows that the joint model is more accurate than the simplified model. In particular, the relative decrease in the JSD goes from 0.01 to up to 0.77. The latter is a sharp decrease that reduces the maximum absolute error in the CDF by more than one order of magnitude with respect to simulations.

We illustrate the importance of a decrease in the JSD in Fig. 8, where the absolute error between the CDF of S obtained by simulation and by the analytical models are calculated for $g = 64$, $q = 2^8$ and $\epsilon = 0.01$. In this latter case, the decrease in the JSD is $G_{\text{JSD}} = 0.11$ and it can be seen that the absolute error for $s = 2$ is reduced by more than one order of magnitude by using the joint model instead of the simplified model. Such a sharp decrease in the absolute error is important because the target reliability τ will oftentimes be extremely close to 1 in practical applications (see Table 3 and Fig. 11 below). However, the Fig. 8 also shows that the benefits of the joint model will only be observed in some specific cases. For instance, the absolute error shown in Fig. 8 decreases rapidly as s increases and is similar for both models. Therefore, both models oftentimes lead to the exact same value of s^*.

For instance, the simplified model is sufficiently accurate in most cases where the target reliability $\tau < 1 - \max\{|F_{S_{\text{sim}}(s)} - F_S(s)|$. For the cases shown in Table 5, this occurs for all JSD $< 10^{-4}$ given $\tau = 1 - 10^{-3}$. Furthermore, the simplified model can be easily extended to incorporate different vales of ϵ for each pair of UEs. This is not straightforward with the joint model. Throughout the remainder of the paper, we present the results obtained with the simplified model. These have been confirmed with the joint model.

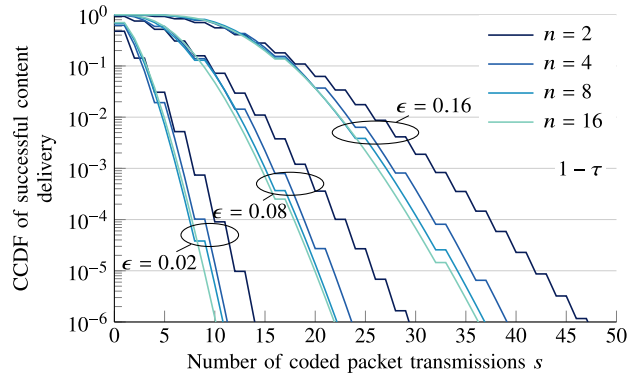

FIGURE 9. CCDF of successful content delivery $S_S(s) = 1 - F_S(s)$ for $q = 2^8$, $\epsilon = \{0.02, 0.08, 0.16\}$, and $n = \{2, 4, 8, 16\}$; y-axis in logarithmic scale.

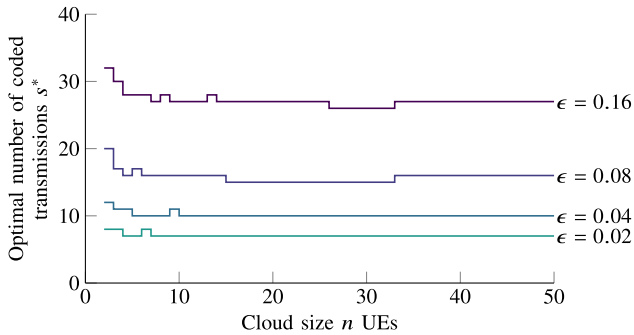

FIGURE 10. Optimal number of coded transmissions s^* given $g = 64$, $q = 2^8$, and $\tau = 1 - 10^{-3}$.

We begin the performance analysis of our NCC protocol by comparing the complementary CDF (CCDF) of successful content delivery $S_S(s) = 1 - F_S(s)$ for $n = \{2, 4, 8, 16\}$ and $\epsilon = \{0.02, 0.16\}$ in Fig. 9. That is, Fig. 9 shows the probability that the worst UE is not able to decode the generation. Hence, lower values indicate a better performance. As it can be seen, large cloud sizes oftentimes reduce the $S_S(s)$ when compared to small cloud sizes. That is, high values of n, for example, $n = 16$, reduce the number of coded transmissions needed to achieve the desired reliability τ when compared to, for example, $n \in \{2, 4, 8\}$. This effect is clearly observable when $S_S(s) \leq 10^{-2}$ in Fig. 9.

The reason for the effect described above is that the ratio of transmissions from the UEs in N_i to total transmissions in the NCC phase s_i/s increases with n. In other words, the frequency of the packets transmitted in the MC towards each of the n UEs increases with n. For example, the curves for $n = 2$ present a step-like shape because each of the two UEs receive up to one packet every two time slots. Therefore, these UEs can only receive up to 50 percent of the total transmissions in the NCC phase. On the other hand, UEs in an MC of size $n = 16$ receive up to 15 packets every 16 time slots. Therefore, these can receive up to 93.75 percent of the total transmissions in the NCC phase.

The effect of cloud size on performance is further illustrated in Fig 10, where we show s^* as a function of n and ϵ for

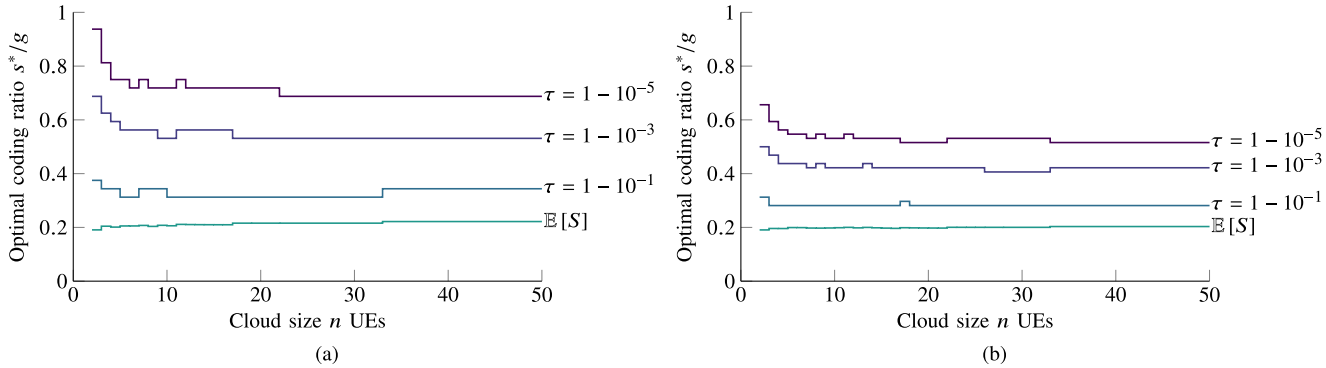

FIGURE 11. Optimal coding ratio for (a) $g = 32$ and (b) $g = 64$ with $q = 2^8$ and $\epsilon = 0.16$.

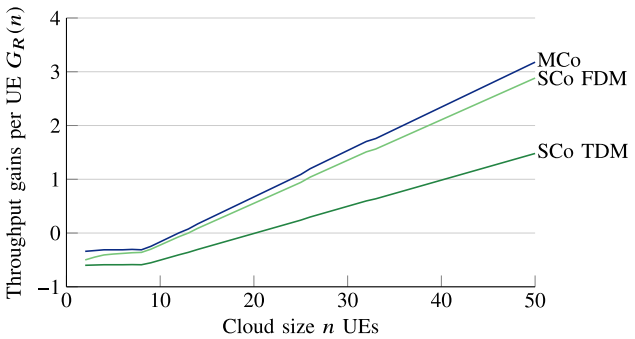

FIGURE 12. Achievable throughput gains with our NCC protocol given: (1) SCo with TDM; (2) SCo with FDM; and (3) MCo are used, for $\epsilon = 0.16$.

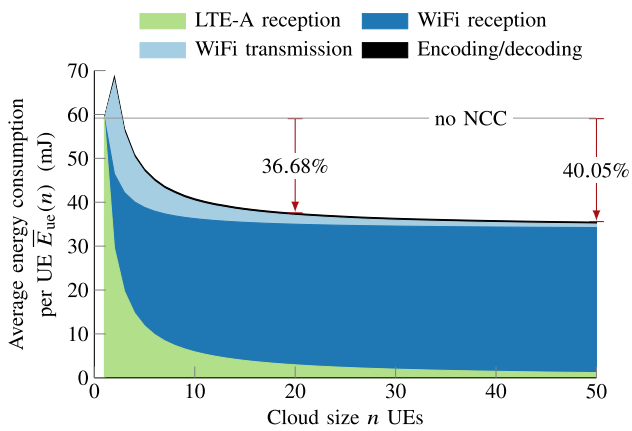

FIGURE 13. Average energy consumption per UE given $\epsilon = 0.16$, $\rho = 1$, $g = 64$, and $q = 2^8$.

$g = 64$, $q = 2^8$, and $\tau = 1 - 10^{-3}$. Specifically, the selection of $n = 2$ results in the largest s^* and would lead to the lowest throughput. Conversely, we denote the optimal cloud size n^* as

$$n^* = \min_n \left\{ n \mid s^*(n) \leq s^*(m) \right\} \quad \forall m \in \mathbb{Z}^+ \tag{32}$$

where $s^*(n)$ is denotes the value of s^* for a particular n. It can be observed that n^* increases with ϵ. For instance, Fig. 10 shows $n^* = \{4, 5, 15, 25\}$ for $\epsilon = \{2, 4, 8, 16\} \cdot 10^{-2}$, respectively. Therefore, small cloud sizes should be avoided when the PER at the WiFi link is high.

Next, we evaluate the efficiency of the proposed NCC protocol with usual values of $g \in \{32, 64\}$. For this, we define the optimal coding ratio as s^*/g. The latter can be interpreted as the relative amount of redundancy that must be added to deliver the content successfully.

Fig. 11 shows s^*/g for $g \in \{32, 64\}$ and $\tau \in \{1 - 10^{-1}, 1 - 10^{-3}, 1 - 10^{-5}\}$. Furthermore, Fig. 11 shows s^*/g for the case in which an idealized ACK mechanism is implemented. That is, we assume that an ACK message is immediately transmitted without errors when the content has been delivered to the worst UE in the MC (i.e., the one with the least favorable conditions to receive the generation). The latter case is denoted as $\mathbb{E}[S]$ in Fig. 11, which shows that, on average, at least a redundancy of $s^* \approx 0.2g$ must be added so the worst UE decodes the generation.

It can also be seen from Fig. 11 that selecting $g = 64$ is much more efficient than selecting $g = 32$ as s^*/g with the former is much lower. Therefore, we select $g = 64$ throughout the rest of the paper.

One of the main benefits of our NCC protocol is the offloading of the cellular link. Specifically, the g packets in the generation only have to be transmitted once per MC. As a result, the amount of utilized resources in the cellular link with our NCC protocol is always g, and is independent of n. Hence, the resource utilization with NCC is in the order of $1/n$ when compared to the case when the whole content is transmitted to each UE. This translates to resource savings in the order of $1 - 1/n$. That is, with only $n = 2$, the utilization of resources is cut by half. This in turn ensures the cellular bandwidth is sufficient to serve an MC of any size.

From the parameters listed in Table 3 it is easy to observe that, if parallel unicast is to be used, the maximum throughput per UE is given as $R_u(n) = \min\{R, B/n\}$. Building on this, we calculate the achievable throughput gains as

$$G_R(n) = \frac{R^*(n)}{R_u(n)} - 1. \tag{33}$$

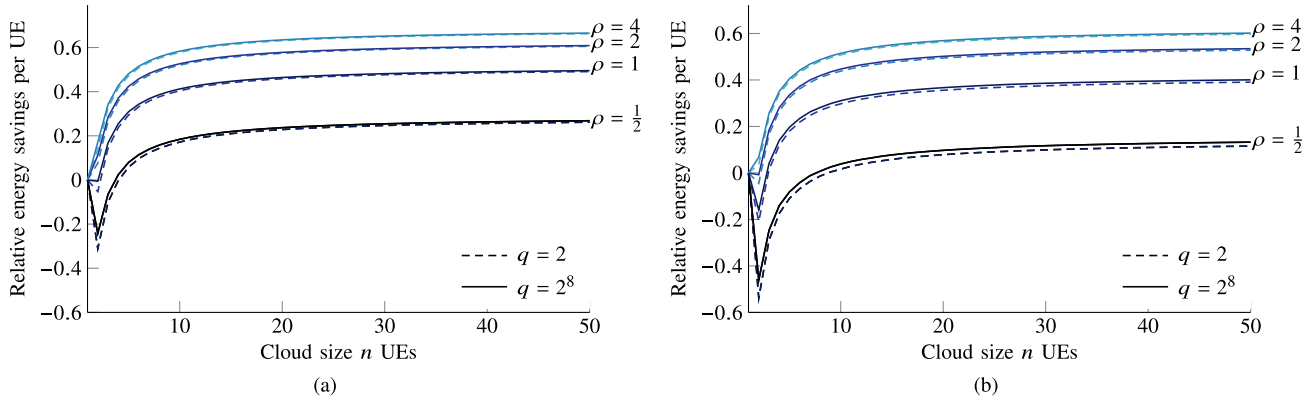

FIGURE 14. Relative average energy savings $G_E(n) = 1 - \overline{E}_{\text{ue}}(n)/\overline{E}_{\text{u}}$ for $\rho = \{0.5, 1, 2, 4\}$, $q = 2$, and $q = 2^8$, given (a) $\epsilon = 0.01$ and (b) $\epsilon = 0.16$.

Three cases are considered to calculate $G_{\text{th}}(n)$: 1) SCo with TDM in the cellular phase; 2) SCo with FDM in the cellular phase; and 3) MCo.

We observed that, for all cases, the achievable throughput per UE with NCC is lower than that of a single unicast session if the same MCS is used, which is $R = 11.76$ Mbps. This slight decrease in throughput is inherent to our NCC protocol, and, as described by (25), occurs because g packet transmissions are performed in the cellular phase, followed by g source packet transmissions and s^* coded transmissions in the NCC phase, for which the PER $\epsilon > 0$. Nevertheless, our NCC protocol results in an n-fold decrease in the number of transmissions from the cellular BS when compared to direct cellular unicast. Furthermore, throughput gains can be achieved for the cases in which the achievable carrier throughput B is not sufficient to maintain $R_{\text{u}}(n) = R$. This is shown in Fig. 12 for $\epsilon = 0.16$ and $\rho = 1$. In particular, Fig. 12 shows that throughput gains can be achieved with $n \geq 20$ with SCo and TDM. But, only $n \geq 13$ and $n \geq 11$ are needed to obtain gains with SCo and FDM and with MCo, respectively. Furthermore, even though a high PER $\epsilon = 0.16$ was selected, more than a three-fold increase in throughput can be achieved with MCo and $n = 50$. The latter value of ϵ can be seen as an upper bound for this parameter as higher values typically lead to the loss of connectivity between devices in real setups. Therefore, $\epsilon = 0.16$ may be seen as a worst case scenario for WiFi transmissions.

Finally, we present the the impact of using our NCC protocol in the energy consumption of UEs. For this, Fig. 13 shows an area plot of the average energy consumption per UE as a function of n for $\epsilon = 0.16$. Different colors indicate the energy consumption at each interface and process, namely, cellular reception, WiFi reception and transmission, and encoding/decoding. For this case, the energy consumption for the direct transmission of the $g = 64$ packets to each UE through the cellular link is $\overline{E}_u = 59.17$ mJ. On the other hand, the energy consumption per UE for $n = 20$ is $\overline{E}_{\text{ue}}(20) = 37.47$ mJ and is further reduced as n increases. At this point it is convenient to define the energy saving provided by our NCC protocol when compared to unicast delivery as

$$G_E(n) = 1 - \frac{\overline{E}_{\text{ue}}(n)}{\overline{E}_{\text{u}}}. \tag{34}$$

Therefore, energy savings of more than 0.37 can be achieved with our NCC protocol, even with relatively small cloud sizes, the same data rate at the cellular and WiFi links, and a high PER in the latter.

Fig. 13 also shows that the main contributing factor to the overall energy savings is that the number of packets transmitted from the BS to each UE decreases as n increases. Conversely, the number of packets transmitted through WiFi to each UE increases with n, but the energy consumption during reception in WiFi is much lower than that in the cellular link. Lastly, the energy consumed for WiFi transmissions is the least contributing factor to overall energy consumption and becomes particularly small for large values for n.

Nevertheless, higher energy savings can be achieved with either a lower PER and with a higher data rate in the WiFi link. This can be observed in Fig. 14, where $G_E(n)$ is shown for $\epsilon \in \{0.01, 0.16\}$, $g \in \{2, 2^8\}$, and $\rho \in \{0.5, 1, 2, 4\}$. As it can be seen, the value of the Galois-field size q has a minor impact on energy consumption, but so does ϵ for high values of ρ, for example for $\rho \in \{2, 4\}$. Furthermore, energy savings can be achieved even with $\rho = 0.5$.

VI. CONCLUSION

In this paper, we presented an NCC protocol and two simple but accurate analytical models that were used to fine-tune its parameters. We assessed the accuracy of our models with respect to simulations by means of the Jensen-Shannon divergence. The main conclusion of this analysis is that the joint model can be up to 0.77 more accurate than the simplified model. This increase in accuracy is essential to meet the reliability requirements of MCD applications. However, the accuracy of the simplified model is adequate for most of the values of the parameters of interest given a relatively low reliability is needed.

Results from the performance analysis of our NCC protocol revealed that important energy savings of more than 37 percent can be achieve when compared to the traditional method for content delivery: replication through parallel unicast links. This is true even with a relatively low cloud sizes of around $n = 4$ UEs and a high PER in the WiFi link. These energy savings increase with the data rate at the WiFi link as communication circuits become more energy efficient. Besides, cellular data consumption is reduced by our NCC protocol by a factor of n w.r.t. traditional parallel transmission. Finally, throughput gains may be achieved if the number of requesting UEs is large and the bandwidth is insufficient to maintain unicast links at maximum throughput. Naturally, these are accentuated by the use of MCo.

Some relevant characteristics that were not considered in our study is that, in practical implementations, the PER between some pairs of UEs will increase with the cloud size and that packet errors may appear in bursts (e.g., due to large-scale fading such as shadowing). As a result, important differences in the PER between pairs of UEs are expected if large clusters are formed and the wireless links can be in outage for extended periods. Hence, future work includes incorporating these aspects to the analysis. On the one hand, to incorporate the impact of varying cloud sizes along with the spatial distribution of the UEs, a specific scenario must be selected. On the other hand, simple models that consider correlation between errors, such as the Gilbert-Elliott model may be considered. However, the novel techniques must be studied to maintain the tractability of the simplified model after incorporating these aspects.

APPENDIX

Here we illustrate the inconvenience of the definition of successful content delivery adopted in our previous work [24] and by Tsimbalo *et al.* [18]. For this, let S_n be the RV that defines the number of coded transmissions needed so that each and every of the n UEs in the MC decodes the generation. The latter is referred to as the probability of successful content delivery in the above-mentioned works. Building on this, the exact CDF of S_n is defined as

$$F_{S_n}(s; n) \triangleq \Pr\left[\bigcap_{i=1}^{n} X_{s_i}^{(i)} = g\right]. \tag{35}$$

But obtaining the exact $F_{S_n}(s; n)$ is complicated because the RVs of the number of DOFs at each UE $X_{s_i}^{(i)}$ for all $i \in \mathcal{N}$ are highly correlated. Specifically, $n + \binom{n}{2}$ stochastic processes are needed to describe the exact state of an MC with n UEs, as described in Section IV.

Instead, to make this problem tractable, it is commonly assumed that $X_{s_i}^{(i)} \perp\!\!\!\perp X_{s_j}^{(j)}$ at each s and for all $j \in \mathcal{N}_i$. Hence, the lower bound

$$F'_{S_n}(s; n) \triangleq \prod_{i=1}^{n} \Pr\left[X_{s_i}^{(i)} = g\right] = \prod_{i=1}^{n} F_{X_{s_i}}(g; i) \tag{36}$$

is commonly used. In particular, Tsimbalo *et al.* found (36) to be a tight lower bound for $F_{S_n}(s; n)$ for a wide range of values of q and g only if the systematic RLNC is used. Despite this fact, it is clear that adopting this previous definition introduces an approximation error.

Besides affecting the precision, we consider using S_n to optimize our NCC protocol to be unfair and provide the following simple proof to support our claim.

Proof: Please recall our main goal is to find s^*, defined as the minimum number of coded packet transmissions needed to achieve the desired reliability τ. If we substitute F_S with F'_{S_n} in (2), we have

$$s^* = \min_s \left\{ s \;\middle|\; \prod_{i=1}^{n} F_{X_{s_i}}(g; i) \geq \tau \right\}. \tag{37}$$

From (4) we have $s_i^* = f(s^*, i, n, g)$; therefore,

$$\prod_{i=1}^{n} F_{X_{s_i^*}}(g; i) \geq \tau \quad \text{for } n \geq 2 \tag{38}$$

must be fulfilled to achieve the desired reliability.

If we follow the same approach as described above, and aim to obtain the same reliability with two different cloud sizes n_1 and n_2 we have

$$F_{X_{s_i^*}}(g \mid n_1; i) = F_{X_{s_i^*}}(g \mid n_2; i) \frac{\prod_{j=1, j\neq i}^{n_2} F_{X_{s_j^*}}(g \mid n_2; j)}{\prod_{j=1, j\neq i}^{n_1} F_{X_{s_j^*}}(g \mid n_1; j)}. \tag{39}$$

Naturally, this implies $\exists i \in \mathbb{Z}^+ \mid i \leq \min\{n_1, n_2\}$ s.t.

$$F_{X_{s_i^*}}(g \mid \min\{n_1, n_2\}; i) < F_{X_{s_i^*}}(g \mid \max\{n_1, n_2\}; i) \tag{40}$$

□

In words, to achieve a predefined reliability τ with a given n and using S_n, the probability of decoding the generation at each individual UEs must increase with n. That is, if S_n is adopted, each individual UE needs a lower probability to decode the generation in a small MC than in a large MC.

ACKNOWLEDGMENT

The authors alone are responsible for the content of the paper.

REFERENCES

[1] Cisco. (Feb. 2017) *Cisco Visual Networking Index (VNI): Global Mobile Data Traffic Forecast Update, 2016–2021*. Accessed: Oct. 15, 2018. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[2] *LTE Multimedia Broadcast Multicast Services (MBMS)*, Viavi Solutions, San Jose, CA, USA, 2015.

[3] *Radio Resource Control (RRC); Protocol Specification*, document 3GPP TS 36.331 V13.0.0, Jan. 2016.

[4] T.-Y. Chan, Y. Ren, Y.-C. Tseng, and J.-C. Chen, "How to reduce unexpected eMBMS session disconnection: Design and performance analysis," *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 126–129, Feb. 2018.

[5] *Delivery of Broadcast Content over LTE Networks*, EBU TR, Jul. 2014.

[6] G. Araniti, M. Condoluci, P. Scopelliti, A. Molinaro, and A. Iera, "Multicasting over emerging 5g networks: Challenges and perspectives," *IEEE Netw.*, vol. 31, no. 2, pp. 80–89, Mar. 2017.

[7] M. V. Pedersen and F. H. P. Fitzek, "Mobile clouds: The new content distribution platform," *Proc. IEEE*, vol. 100, pp. 1400–1403, May 2012.

[8] F. H. P. Fitzek and M. D. Katz, *Mobile Clouds. Exploiting Distributed Resources in Wireless, Mobile and Social Networks*. London, U.K.: Wiley, 2014.

[9] L. Keller, A. Le, B. Cici, H. Seferoglu, C. Fragouli, and A. Markopoulou, "MicroCast: Cooperative video streaming on smartphones," in *Proc. Int. Conf. Mobile Syst., Appl., Services (MobiSys)*, 2012, pp. 57–70.

[10] L. Aymen, B. Ye, and T. M. T. Nguyen, "Offloading performance evaluation for network coding-based cooperative mobile video streaming," in *Proc. Int. Conf. Netw. Future (NOF)*, 2016, pp. 1–5.

[11] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, "Network information flow," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1204–1216, Jul. 2000.

[12] T. Ho, M. Médard, J. Shi, M. Effros, and D. R. Karger, "On randomized network coding," in *Proc. Annu. Allerton Conf. Commun. Control Comput.*, 2003, vol. 41, no. 1, pp. 11–20.

[13] A. L. Jones, I. Chatzigeorgiou, and A. Tassi, "Binary systematic network coding for progressive packet decoding," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 4499–4504.

[14] M. Di Renzo, M. Iezzi, and F. Graziosi, "On diversity order and coding gain of multisource multirelay cooperative wireless networks with binary network coding," *IEEE Trans. Veh. Technol.*, vol. 62, no. 3, pp. 1138–1157, Mar. 2013.

[15] S. Pandi, R. T. Arranz, G. T. Nguyen, and F. H. P. Fitzek, "Massive video multicasting in cellular networks using network coded cooperative communication," in *Proc. IEEE Annu. Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2018, pp. 1–2.

[16] T. Ho, M. Medard, R. Koetter, D. Karger, M. Effros, J. Shi, and B. Leong, "A random linear network coding approach to multicast," *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4413–4430, Oct. 2006.

[17] Z. Chang, S. Zhou, T. Ristaniemi, and Z. Niu, "Collaborative mobile clouds: An energy efficient paradigm for content sharing," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 186–192, Apr. 2018.

[18] E. Tsimbalo, A. Tassi, and R. J. Piechocki, "Reliability of multicast under random linear network coding," *IEEE Trans. Commun.*, vol. 66, no. 6, pp. 2547–2559, Jun. 2018.

[19] M. Tömösközi, F. H. P. Fitzek, D. E. Lucani, M. V. Pedersen, P. Seeling, and P. Ekler, "On the packet delay characteristics for serially-connected links using random linear network coding with and without recoding," in *Proc. Eur. Wireless Conf.*, 2015, pp. 1–6.

[20] F. H. P. Fitzek, M. Katz, and Q. Zhang, "Cellular controlled short-range communication for cooperative P2P networking," *Wireless Pers. Commun.*, vol. 48, no. 1, pp. 141–155, Jan. 2009.

[21] J. Laneman, D. Tse, and G. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.

[22] L. Wang, Z. Yang, L. Xu, and Y. Yang, "NCVCS: Network-coding-based video conference system for mobile devices in multicast networks," *Ad Hoc Netw.*, vol. 45, pp. 13–21, Jul. 2016.

[23] R. Torre, "Offloading traffic from cellular networks using network coding and cooperation," M.S. thesis, Dresden Univ. Technol., Dresden, Germany, May 2017.

[24] I. Leyva-Mayorga, R. Torre, S. Pandi, G. T. Nguyen, V. Pla, J. Martinez-Bauset, and F. H. Fitzek, "A network-coded cooperation protocol for efficient massive content distribution," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018.

[25] *Physical Channels Modulation*, document 3GPP TS 36.211 V15.4.0, Apr. 2018.

[26] H. Khamfroush, D. E. Lucani, P. Pahlevani, and J. Barros, "On optimal policies for network-coded cooperation: Theory and implementation," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 2, pp. 199–212, Feb. 2015.

[27] A. Paramanathan, M. V. Pedersen, D. E. Lucani, F. H. Fitzek, and M. Katz, "Lean and mean: Network coding for commercial devices," *IEEE Wireless Commun.*, vol. 20, no. 5, pp. 54–61, Oct. 2013.

[28] C. W. Sørensen, A. Paramanathan, J. A. Cabrera, M. V. Pedersen, D. E. Lucani, and F. H. P. Fitzek, "Leaner and meaner: Network coding in SIMD enabled commercial devices," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2016, pp. 1–6.

[29] *Physical Layer Procedures*, document 3GPP TS 36.213 V15.4.0, Jan. 2019.

[30] M. Lauridsen, L. Noël, T. B. Sørensen, and P. Mogensen, "An empirical LTE smartphone power model with a view to energy efficiency evolution," *Intel Technol. J.*, vol. 18, no. 1, pp. 172–193, 2014.

[31] L. Sun, H. Deng, R. K. Sheshadri, W. Zheng, and D. Koutsonikolas, "Experimental evaluation of WiFi active power/energy consumption models for smartphones," *IEEE Trans. Mobile Comput.*, vol. 16, no. 1, pp. 115–129, Jan. 2017.

ISRAEL LEYVA-MAYORGA received the B.Sc. degree in telematics engineering and the M.Sc. degree (Hons.) in mobile computing systems from the Instituto Politécnico Nacional (IPN), Mexico, in 2012 and 2014, respectively, and the Ph.D. degree *(cum laude)* in telecommunications from the Universitat Politècnica de València (UPV), Spain, in 2018. He was a Visiting Researcher at the Department of Communications, UPV, in 2014, and at the Deutsche Telekom Chair of Communication Networks, Technische Universität Dresden, Germany, in 2018. He is currently a Postdoctoral Researcher at the Department of Electronic Systems, Aalborg University, Denmark. His research interests include 5G and beyond, satellite networks, random access protocols for mMTC, URLLC, wireless mesh networks, and network coding schemes.

ROBERTO TORRE received the bachelor's degree in telecommunications technology engineering and the master's degree in telecommunications technology engineering from the University of Valladolid, Spain, in 2015 and 2017, respectively. He is currently pursuing the Ph.D. degree with the Deutsche Telekom Chair of Communication Networks, Technische Universität Dresden (TU Dresden), since August 2017.

VICENT PLA received the B.E. and M.E. degrees in telecommunication engineering and the Ph.D. degrees from the Universitat Politècnica de València (UPV), Spain, in 1997 and 2005, respectively, and the B.Sc. degree in mathematics from the Universidad Nacional de Educación a Distancia (UNED), Spain, in 2015. In 1999, he joined the Department of Communications, UPV, where he is currently a Professor. His research interest includes primarily in the area of modeling and performance analysis of communication networks. During the past few years, most of his research activity has focused on traffic and resource management in wireless networks. In these areas, he has published numerous papers in refereed journals and conference proceedings, and has been an active participant in several research projects.

SREEKRISHNA PANDI was born in Chennai, India. He received the B.E. degree in electronics and instrumentation engineering from Anna University, in 2013, and the master's degree in nanoelectronic systems from Technical University (TU) Dresden, Dresden, Germany, in 2015. Since 2015, he has been a Ph.D. Researcher with the Deutsche Telekom Chair of Communication Networks, TU Dresden.

GIANG T. NGUYEN received the M.Eng. degree in telecommunications from the Asian Institute of Technology (AIT), Thailand, in 2007, and the Ph.D. (Dr. Ing.) degrees in computer science from the Technical University of Dresden, Germany, in 2016. He is currently a Senior Researcher with the Deutsche Telekom Chair of Communication Networks, Technical University of Dresden. His research interests include 5G, and highly adaptive and energy-efficient computing (HAEC). His research focuses on network coding and low-latency aspects of network function virtualization and software-defined networking.

JORGE MARTINEZ-BAUSET received the Ph.D. degree from the Universitat Politècnica de València (UPV), Valencia, Spain, in 1997. He also received the 1997 Alcatel Spain Best Ph.D. Thesis Award in Access Networks. He is currently a Professor with UPV. From 1987 to 1991, he was with QPSX Communications, Perth, Australia, working with the team that designed the first IEEE 802.6 MAN. Since 1991, he has been with the Department of Communications, UPV. His research interests are in the area of performance evaluation and traffic control for multiservice networks.

FRANK H. P. FITZEK received the Dipl.Ing. degree in electrical engineering from the University of Technology and Rheinisch-Westfälische Technische Hochschule, Aachen, Germany, in 1997, the Ph.D. (Dr. Ing.) degree in electrical engineering from Technical University, Berlin, Germany, in 2002, and the Doctor Honoris Causa degree from the Budapest University of Technology and Economy. He became an Adjunct Professor with the University of Ferrara, Italy, in 2002. In 2003, he joined Aalborg University as an Associate Professor and later became Professor. He has co-founded several start-up companies starting with acticom GmbH, Berlin, in 1999. He is currently Professor and the Head of the Deutsche Telekom Chair of Communication Networks, Technical University Dresden, Germany, coordinating the 5G Lab Germany. His current research interests are in the areas of wireless and mobile 5G communication networks, mobile phone programming, network coding, cross layer as well as energy-efficient protocol design, and cooperative networking. He was selected to receive the NOKIA Champion Award several times in a row, from 2007 to 2011. He received the Nokia Achievement Award for his work on cooperative networks, in 2008, the SAPERE AUDE Research Grant from the Danish Government, in 2011, and the Vodafone Innovation prize, in 2012.