IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 June 2023, accepted 29 June 2023, date of publication 17 July 2023, date of current version 25 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3295838

RESEARCH ARTICLE

Analysis of a Contention-Based Approach Over 5G NR for Federated Learning in an Industrial Internet of Things Scenario

GIAMPAOLO CUOZZO[1], (Member, IEEE), JONAS PETTERSSON[2], AND MASSIMO CONDOLUCI[3]

[1]National Laboratory of Wireless Communications (WiLab), CNIT, 40133 Bologna, Italy
[2]Ericsson Research, 977 53 Luleå, Sweden
[3]Ericsson Research, 164 40 Stockholm, Sweden

Corresponding author: Giampaolo Cuozzo (giampaolo.cuozzo@cnit.it)

ABSTRACT The growing interest in new applications involving co-located heterogeneous requirements, such as the Industrial Internet of Things (IIoT) paradigm, poses unprecedented challenges to the uplink wireless transmissions. Dedicated scheduling has been the fundamental approach used by mobile radio systems for uplink transmissions, where the network assigns contention-free resources to users based on buffer-related information. The usage of contention-based transmissions was discussed by the 3rd Generation Partnership Project (3GPP) as an alternative approach for reducing the uplink latency characterizing dedicated scheduling. Nevertheless, the contention-based approach was not considered for standardization in LTE due to limited performance gains. However, 5G NR introduced a different radio frame which could change the performance achievable with a contention-based framework, although this has not yet been evaluated. This paper aims to fill this gap. We present a contention-based design introduced for uplink transmissions in a 5G NR IIoT scenario. We provide an up-to-date analysis via near-product 3GPP-compliant network simulations of the achievable application-level performance with simultaneous Ultra-Reliable Low Latency Communications (URLLC) and Federated Learning (FL) traffic, where the contention-based scheme is applied to the FL traffic. The investigation also involves two separate mechanisms for handling retransmissions of lost or collided transmissions. Numerical results show that, under some conditions, the proposed contention-based design provides benefits over dedicated scheduling when considering FL upload/download times, and does not significantly degrade the performance of URLLC.

INDEX TERMS 5G, NR, Ultra-Reliable Low-Latency Communication (URLLC), Industrial IoT (IIoT), Federated Learning (FL), contention-based.

I. INTRODUCTION

Next-generation mobile radio networks will support new use cases and, consequently, new traffic types [1], [2], [3], [4], [5]. One exemplary emerging application is the Industrial Internet of Things (IIoT), where wireless technologies ensure the interconnection between industrial assets (e.g., valves, pumps, robotic arms, etc.) and the control rooms of industry plants [6], [7] to realize digital twins of physical industrial entities, promote Extended Reality (XR)-based maintenance operations, or support distributed Machine Learning (ML) frameworks such as Federated Learning (FL) [8], [9], [10], [11], [12], [13].

Notably, the main characteristic of these new data transfers is that they put more effort into the uplink direction, whereas legacy traffics, such as web browsing, are rather downlink-heavy. For instance, uplink performance is as important as downlink for fast convergence of FL algorithm, where devices upload the results of their local training to a central entity (upstream) which performs aggregation and re-distributes the updated model (downstream) until all nodes utilize the same version [14]. In this regard, the literature has been investigating several approaches to optimize the uplink data transmissions that mainly belong to

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

two categories: Contention-Free (CF) and Contention-Based (CB). According to the former, User Equipments (UEs) transmit via dedicated radio resources that can be either time slots (Time Division Multiple Access (TDMA)) [15], [16], frequency channels (Frequency Division Multiple Access (FDMA)) [17], or their combination [18], [19], [20], as well as orthogonal spreading codes in a Code Division Multiple Access (CDMA) approach [21], [22], [23], and spatial beams in a Multiple Input Multiple Output (MIMO) network [24], [25]. As for the CB uplink transmissions, besides the proliferation of well-known studies on ALOHA-based solutions and Carrier Sense Multiple Access (CSMA) protocols [26], [27], [28], [29], [30], a recent hot topic is called Non-Orthogonal Multiple Access (NOMA), where smart receivers are designed to mitigate the interference produced by uplink transmissions that exploit the same radio resource [31], [32].

From a standardization viewpoint, the 3rd Generation Partnership Project (3GPP) has been considering dedicated scheduling as the main approach for uplink data transmission, with the network assigning dedicated radio resources (grants) upon receiving explicit requests from each UE. Radio resources could be either granted in a dynamic way based on the amount of data a UE has in its buffer or could be allocated in a semi-persistent way with an allocation repeating over a certain amount of time. The usage of CB approach has been studied for Long Term Evolution (LTE) to allow UEs to directly transmit data in uplink without having to wait for a dedicated grant [33]. Nevertheless, performance gains of CB over LTE were limited and achievable only in scenarios with low load and small-size uplink data, hence standardization continued to focus on dedicated scheduling as the main approach for uplink data transmission.

However, with the proliferation of new uplink-oriented applications with heterogenous requirements, there is a renewed interest in exploring the potential benefits of CB designs for 3GPP-compliant networks. Additionally, 5th generation (5G) New Radio (NR) foresees substantial differences w.r.t LTE that might really unleash the potential benefits of CB schemes. Therefore, the aim of this paper is to re-visit the work done by 3GPP and to give a first assessment of the achievable performance of CB uplink transmissions applied to 5G NR. We present a CB design for 5G NR Physical Uplink Shared Channel (PUSCH), and we consider different mechanisms for handling retransmissions of lost or collided transmissions. Unlike previous assessments done by 3GPP, we consider extensive network simulations to assess the application-level performance achieved by FL traffic in an IIoT scenario when using the proposed CB design for 5G NR PUSCH, focusing on both upstream and downstream performance. Numerical results show that the considered CB design for 5G NR PUSCH provides benefits over dedicated scheduling under some conditions, and scales well with the number of UEs, by also poorly deteriorating the application-level performance of other higher-priority traffic flows.

The paper is structured as follows. In Sec. II we clarify the original contributions of this paper by reviewing both the academic literature and 3GPP standards. Sec. III describes the considered CB design for NR PUSCH, whereas Secs. IV and V present the system model and the metrics used for the performance evaluation. Finally, in Sec. VI we present the corresponding numerical results, while in Sec. VII we summarize the main achievements and possible future works.

II. RELATED WORKS

A. LITERATURE REVIEW ON UPLINK DATA TRANSMISSIONS

The academic literature analyzes several approaches to shrink the uplink latency provided by dedicated scheduling, where UEs willing to transmit data have to first request radio resources from the network. Some works propose improvements of the semi-persistent allocation mechanisms [34], where the network reserves a given number of dedicated radio resources for a limited amount of time. In this regard, the authors in [35] study predictive algorithms for the radio resource assignments by considering an LTE network, whereas the potential benefits of a traffic-aware semi-persistent scheduler are investigated in [36] for a private 5G NR network tailored to an IIoT environment. Semi-persistent resource allocations reduce the control plane overhead, but fail in managing unpredictable/highly-variable traffic and do not scale well with the offered traffic due to an intrinsic spectral inefficiency.

To overcome the above limitations, the literature is proposing grant-free transmissions [37], [38], that is, a distributed scheme where UEs can autonomously select the radio resources to be used for their uplink transmission without relying on any grant reception, thereby introducing possible collisions. This approach is tailored to aperiodic (or uncertain) traffic but its CB nature undermines communication reliability. Some works [39], [40], [41], [42], [43] try to mitigate the collision impact by studying both, the optimal number of a-priori packet duplications and how to manage the acknowledgments of the duplicates, leading to the consequent trade-off between resource efficiency and reliability. Conversely, others investigate sensing mechanisms and/or interference cancellation techniques [44], [45], [46], as well as considering UEs that leverage ML to learn how to optimally select the radio resources based on their past experience [47]. However, distributed solutions imply a higher complexity at the UE-side which may be unfeasible in some scenarios (e.g., for IIoT applications), and their optimality applies only to particular cases.

B. STANDARDIZATION REVIEW ON UPLINK DATA TRANSMISSIONS

Dynamic Scheduling (DS) is the main approach used in 3GPP-compliant networks to support the transmission of uplink data with variable size and no periodic patterns [48]. Fig. 1 shows the timing diagram of DS. First, a UE with no allocated grants (i.e., dedicated radio resources) waits for an

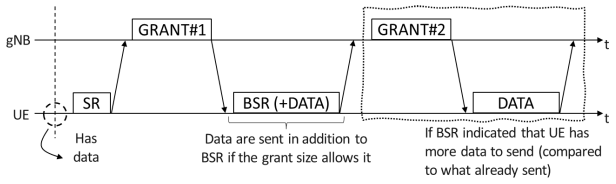


FIGURE 1. High-level time diagram of the basic dynamic scheduling principle.

occasion to send a Scheduling Request (SR) to indicate to the Next Generation Node Base (gNB) that it has new data to be sent, then the gNB replies with a grant ("Grant#1" in Fig. 1) containing the set of radio resources that are allocated for the first uplink transmission. Consequently, the UE will create a Transport Block (TB) (i.e., Medium Access Control (MAC) Protocol Data Unit (PDU)) based on the received grant. The TB will be used to carry (i) the Buffer Status Report (BSR), i.e., a MAC Control Element (CE) indicating the number of bytes left in its transmission buffer, and (ii) any data that may fit into it[1]. The number of resources allocated by the first grant could be enough to allow the UE to transmit all data in its queue, but this cannot be guaranteed as the gNB has not yet information on how much data the UE has to send. Hence, depending on the received BSR, the gNB could send one or more new grants to allow the UE to free up its buffer (transmissions highlighted with a dashed box in Fig. 1).

As a matter of fact, DS is a very flexible approach because it allows tailoring the radio resources allocated to a UE based on its buffer status and cell load, as well as adjusting transmission parameters (e.g., Modulation and Coding Scheme (MCS)) based on its channel quality. Nevertheless, the interval from when new data reaches UE's buffer to when the gNB knows how much data the UE has actually to send is not negligible and this impacts the overall uplink latency performance.

The 3GPP studied possible uplink latency reduction techniques for LTE in Rel. 9 [33], [49] and in Rel. 14. [50], [51], [52]. For DS, it was proposed to increase the frequency of SR occasions to reduce the first component of uplink delay. Other solutions were based on Semi-Persistent Scheduling (SPS), with periodic fixed-size allocation dedicated to a UE which would allow a UE to directly start transmitting its buffered data. However, since the UE could have no data to transmit in a given SPS occasion, it could do padding or skip the transmission opportunity depending on the configuration sent by the gNB. Solutions [33], [51], [52] were instead based on the usage of a CB PUSCH, where a UE could directly transmit its uplink data using a pre-configured PUSCH allocation which is shared among multiple UEs, thereby introducing collisions in the network. Regarding handling of collisions, the proposal [51] considered that the gNB could not distinguish among colliding UEs. A colliding UE will not receive any feedback (acknowledgment of successful reception) by the gNB and this will trigger a retransmission. In this scheme, the UE will perform backoff when selecting the next CB PUSCH occasion for transmission. The proposal [52], instead, considered that the gNB could distinguish colliding UEs through DeModulation Reference Signal (DMRS)-based UE identification [48], [53]. In this way, the gNB can, at least, acquire knowledge of which UEs collided and consequently schedule a dedicated PUSCH resource for their retransmission (thus avoiding further collisions). Moreover, the study in [33] provided an analysis of achievable performance when using CB for traffic upload and download and considering one UE, whereas relationships between uplink load, collision probability, and uplink latency characterizing the aforementioned solutions can be found in [51] and [52]. By considering these works, the calculations in [50] and [54] highlight that the uplink delay for CB PUSCH transmissions is difficult to be kept stable if the collision probability (which depends on how many UEs share the same CB allocation) becomes too high, whereas solutions based on SR frequency increase and on SPS allow a more predictable delay performance at the expense of a reduced uplink capacity. Consequently, [49], [50] concluded that the gains of the CB PUSCH solutions were too limited for LTE compared to DS or SPS to motivate the required extra standardization work.

C. ORIGINAL CONTRIBUTIONS OF THIS PAPER

Besides the degradation of performance caused by collision, the study from 3GPP did not provide an exhaustive analysis of the behavior of CB over LTE. Furthermore, 5G NR brought different changes compared to LTE which could influence the achievable performance of a CB approach. Overall, the contributions of this paper are:

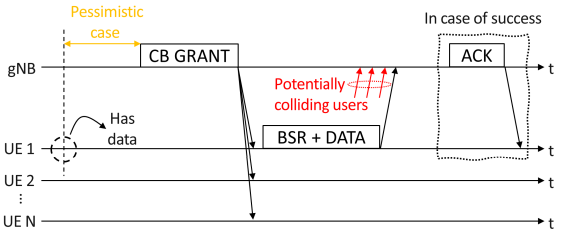
- Introduce a CB design for 5G NR PUSCH.
- Analyse performance when legacy DS and CB for 5G NR PUSCH are simultaneously used, by also assessing the impact of two different retransmission mechanisms which are inspired from previous 3GPP studies [51], [52].
- Analyse performance of CB for 5G NR PUSCH when applied to a FL-based IIoT scenario, where there is a correlation among uplink transmissions of FL UEs, thus creating a more challenging scenario for CB.
- Analyse both downstream and upstream flows.
- Analyse the trade-off and the relationships among different metrics related to CB for 5G NR PUSCH.

III. CB FOR NR PUSCH

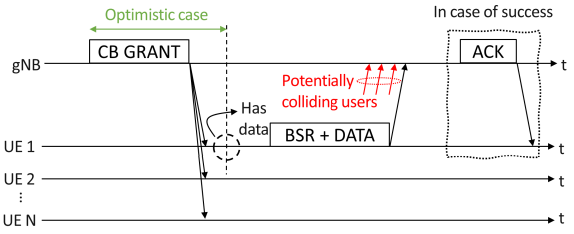
A. GENERAL 5G RADIO FRAMEWORK

The time axis is divided into *slots*, composed of 14 Orthogonal Frequency Division Multiplexing (OFDM) symbols each, whereas the frequency axis is partitioned into Resource Blocks (RBs), that is, sets of 12 subcarriers [48]. We consider an Frequency Division Duplexing (FDD) scheme where the gNB manages the uplink/downlink radio resources, that is, a set of OFDM symbols and RBs (space/power domains are not considered in this paper).

[1] SRs, grants, and the message containing BSRs plus data are mapped to Physical Uplink Control Channels (PUCCHs), Physical Downlink Control Channels (PDCCHs) and PUSCHs, respectively.



(a) Pessimistic case, where the UE has new data to transmit but has not yet received a CB grant.



(b) Optimistic case, where the UE has already received a CB grant before new data has to be transmitted.

FIGURE 2. Timing diagram of the considered CB for NR PUSCH design, showing the relationship between reception of CB grant and availability of uplink data.

B. CONTENTION-BASED DESIGN

The gNB allocates a portion of uplink radio resources to a given set of UEs. We refer to this allocation as *CB grant, CB resource, or CB allocation*, equivalently. In our current implementation, we consider that a CB grant is created at each slot[2]. The CB grant is broadcasted to the UEs associated with that CB resource and contains the following two main pieces of information:

1) The time/frequency location and dimension (in terms of number of OFDM symbols and RBs) of the CB resource[3];
2) The MCS associated to the CB allocation. In our current implementation, we consider that the gNB has Channel State Informations (CSIs)[4] of the UEs, hence the MCS is chosen according to the UE in the worst channel condition. If CSIs are not available, the most conservative MCS is selected.

Only the UEs with non-empty queues will exploit the CB grant to transmit BSR and data, whereas the others will ignore it. In case of successful CB transmission, the gNB replies with a positive Acknowledgment (ACK).

Fig. 2 shows the timing diagram of the aforementioned CB design, focusing on the time relationship between the availability of a CB allocation and the presence of data in the UE's buffer. Fig. 2a depicts the pessimistic case, i.e., the UE has new data available at its buffer but has no CB resources granted for transmission, so it has to wait to receive a CB grant. Indeed, this is possible because the creation of CB grants and data are independent events. Fig. 2b represents the optimistic case, where the UE has already received a CB grant and thus new data which reached its buffer can directly be sent over the CB resources.

However, a CB uplink transmission can fail either due to collisions or link failures. In the former case, we consider that collisions are always harmful, i.e., no capture effect is considered. Nonetheless, regardless of the reason for the missed reception, the gNB will not reply with any ACK, thereby triggering a retransmission [48], [51].

C. RETRANSMISSION MECHANISMS

By recalling that 5G NR relies on Hybrid Automatic Repeat reQuest (HARQ) at the MAC layer, we defined a maximum number of retransmissions for each TB, N_{RX}. In particular, when a UE wants to retransmit a TB after $N_{RX} + 1$ times, the MAC layer declares a HARQ failure, and an Radio Link Control (RLC) retransmission is triggered since we consider Acknowledge Mode (AM) RLC.

Specifically, two retransmission mechanisms are considered, where all UEs implement the same mechanism within one simulation round.

1) RETRANSMISSIONS ON DEDICATED RESOURCES

In this case, we assume that the gNB can retrieve the identity of the colliding UEs, and thereby it can reserve dedicated radio resources for each colliding UE to retransmit the failed TB. In particular, the dedicated grant will indicate which CB resource was used by the failed attempt so that the UE can know what TB to retransmit.

Remark 1: The MCS associated to retransmissions on dedicated resources is no longer dependent on the worst channel conditions but is tailored to the CSI of the specific UE (if available).

Remark 2: In real-world implementations, the use of orthogonal signals, such as those obtained with a proper mapping of DMRS symbols, can let the gNB know the identity of the colliding UEs [48]. However, these types of signals are usually in a finite number, and this may limit the number of UEs that can exploit a CB allocation. This aspect is left to future studies, i.e., we have assumed that such a mechanism can distinguish all UEs associated with the CB resource.

2) RETRANSMISSIONS ON CONTENTION-BASED RESOURCES

In this case, when the ACK is not received, each UE will perform backoff before retransmitting again via CB resources. During backoff, each UE will stay silent, i.e., it will not exploit any CB grant, for a number of slots uniformly distributed over the interval $[0, T_{BO}]$. Remarkably, since retransmissions refer to the same TB created for the failed

[2]Of course, other approaches could be considered, e.g., creating the CB grant as a semi-persistent allocation thus avoiding the transmission of CB grants at each slot. Nevertheless, we considered this approach for simplicity of implementation, and because it allowed us to analyze scenarios with a dynamic variation of the resources allocated to the CB grant.

[3]Notice that, in our design, the colliding transmissions are completely overlapped in time and frequency, and this means that the gNB can avoid performing blind decoding, that is, blindly searching for possible transmissions within a given time-frequency resource.

[4]Specifically, the gNB computes the CSI upon receiving the periodical Channel Quality Information (CQI) transmissions made by the UEs.

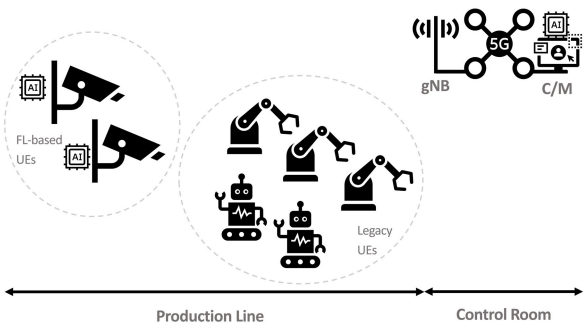


FIGURE 3. The considered 2030-like industrial scenario, where industrial assets are equipped with legacy or FL-based UEs that are served by a private 5G network. These UEs communicate with a C/M that is physically located in the control room of the factory.

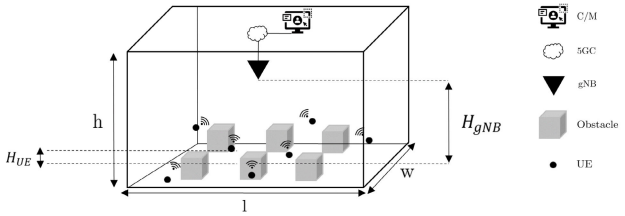


FIGURE 4. The deployment model of the considered 2030-like industrial scenario.

transmission, the MCS of the CB allocation used for the retransmission may not be compatible with that associated with the failed TB, since this depends on how the channel has changed for the worst UE. Therefore, when using this retransmission policy, UEs do not rely on HARQ retransmissions, that is, we set $N_{RX} = 0$. Conversely, an RLC retransmission is triggered directly, thereby generating a new set of TB.

IV. SYSTEM MODEL

A. SCENARIO

This paper considers the 2030-like industrial scenario foreseen in [8], where a heterogeneous set of IIoT UEs coexist in the same factory. In particular, Fig. 3 illustrates the considered network architecture, where N industrial assets can either be legacy (e.g., robotic arms) or FL-based (e.g., cameras performing image recognition through neural networks trained via FL). Both types of entities are deployed in the production line of the factory and they have to communicate with a C/M located in the control room. To this aim, the industrial devices are equipped with 5G UEs (one per industrial asset), and the factory is controlled by a private 5G network which consists of a dedicated Radio Access Network (RAN) and 5G Core (5GC).

B. DEPLOYMENT MODEL

The factory floor has been modeled as a parallelepiped of length l, width w, and height h, as indicated in [7]. Inside the factory, 5G communication between the UEs and the gNB can undergo severe attenuation due to obstructing elements (also referred to as ''obstacles'' in the rest of the paper), such

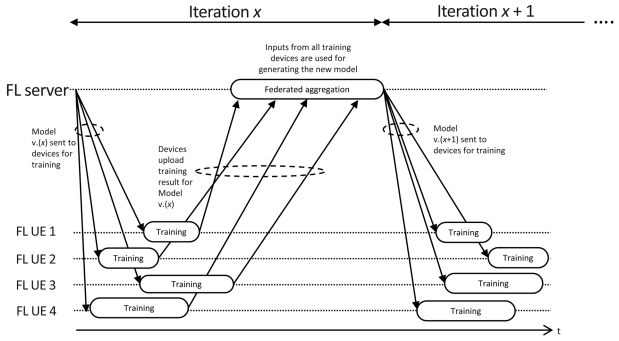


FIGURE 5. Timing diagram of the considered FL traffic, where a FL server performs aggregation during iteration x by using inputs from all FL devices (four UEs are depicted as an example), and it generates a new model version which will be sent during iteration x+1. For the sake of simplicity, the gNB time axis is not shown, but it clearly acts as the forwarding node between FL UEs and the FL server.

as walls or metal slabs. Obstacles are modeled as cubes and they are distributed inside the factory based on a given density B_D, whereas N UEs are randomly and uniformly distributed inside the factory at a given height H_{UE} from the ground floor. The gNB is instead located at height H_{gNB}, as shown in Fig. 4.

C. TRAFFIC MODEL

As anticipated in Sec. IV-A, we consider a factory containing two types of industrial assets. This means that we consider two different cathegories of UEs, i.e., (i) UEs that produce Ultra-Reliable Low-Latency Communication (URLLC) traffic (hereinafter referred to as URLLC UEs) and (ii) UEs that generate FL traffic (hereinafter referred to as FL UEs). Hence, among the N UEs which are randomly and uniformly distributed in the factory, N_{UR} are URLLC UEs and N_{FL} are FL UEs.

1) URLLC TRAFFIC

It is modelled as a periodic bidirectional traffic, where UEs transmit and receive application layer PDUs of P^U_{UR} and P^D_{UR} bytes, respectively, with a fixed periodicity τ. Depending on the transmission direction, that is, uplink or downlink, an application layer PDU is discarded if it has been received with a delay exceeding τ^U_B or τ^D_B, respectively. In particular, the delay is defined as the time elapsing from the instant when new data for transmission is generated at the sender, and the instant when it is entirely received by the recipient.

2) FL TRAFFIC

It is modeled according to the synchronous FL framework described in [14], i.e., an iterative procedure where a FL server, embedded in the C/M, trains a global model (e.g., the parameters of a neural network) by aggregating local models coming from the FL devices. The approach of one, generic, FL iteration is shown in Fig. 5. At the beginning of the iteration, the FL server sends the current version of the global model to the FL UEs, i.e., model v.(x) during iteration x, where $x \in \{1, 2, \ldots, X\}$, being X the total number of FL iterations. Upon reception of the model, the devices

perform local training and then send their updated version to the server. Finally, the server computes the new version of the model, i.e., model v.(x+1), that will be sent in downlink during iteration x+1. Specifically, a unicast download of the model is assumed, that is, the server individually sends the same current version of the model to all UEs[5]. In this regard, τ^M represents the time taken by the C/M application layer to generate the corresponding Physical (PHY) PDUs. From the communication perspective, having a unicast download means that UEs will receive the model at different instants, and this spreads the subsequent uplink traffic over time, even due to potentially different training times of separate UEs. However, in this synchronous FL paradigm, the server has to receive all models before generating the new version. To avoid that the server stops due to errors (e.g., missing data fragments), we leverage Transmission Control Protocol (TCP) at the transport layer, thus introducing retransmissions of lost data at layer 4 of the protocol stack.

D. CHANNEL MODEL

The channel model is taken from [57], where the blockage model B is used to determine the multipath attenuation caused by each of the obstacles using a knife-edge diffraction method, in addition to the path gain matrix and 3D channel data for all possible devices' locations.

E. APPLICATION OF CONTENTION

In such an IIoT scenario, the objective of this paper is to assess whether the CB approach for *uplink* transmissions (i.e., NR PUSCH) described in Sec. III can provide any benefit. In particular, we apply the CB design for NR PUSCH to the FL UEs only, because there exist other scheduling algorithms, such as semi-persistent scheduling [34], [35], [36], which are better tailored to the URLLC traffic characteristics. For example, the stringent availability requirements of the URLLC traffic [58], [59], [60], [61], cannot be easily met by a design where transmissions can also fail due to collisions in addition to channel impairments.

Conversely, the study of achievable performance with CB strategies for the FL traffic may be interesting due to the following reasons:

1) It is likely spread over time due to (i) unicast download of the model, (ii) a non-negligible τ^M, and (iii) possibly different training times of the UEs. Indeed, by design, CB solutions work well when UEs do not have to transmit at the same time;
2) It is characterized by an on-off pattern, i.e., the download of a version of the model is followed by an upload of the new version and vice versa, but these two events never occur together. Since the download of a model is also characterized by the uplink transmissions of the corresponding TCP ACKs, the immediate consequence of this property is that the uplink transmissions of TCP ACKs and FL models are not simultaneous and thus they cannot collide;
3) FL UEs can, in principle, exploit their local ML capabilities to also learn *when* to use the CB resources based on their past experience. However, this aspect is not considered in this study but it might be the subject of future works;

[5]This choice avoids considering the technical difficulties of multicast transmissions, such as the selection of the MCS [55], [56].

V. PERFORMANCE METRICS

This section describes the metrics used to assess the performance of the considered CB design when applied to the FL traffic and referring to the IIoT system model presented in Sec. IV. In particular, each metric refers to one traffic type, that is, either URLLC or FL traffic.

A. APPLICATION LAYER AVAILABILITY

Let us introduce a Bernoulli state variable for the i-th URLLC device, $X_i(t)$, that is zero if the last reception (at the application layer) has failed, either due to link failures or exceeding delay bound (see Sec. IV-C). Consequently, the application layer availability for the i-th URLLC UE can be defined as follows:

$$a_i(t) := \begin{cases} 0, & \text{if } \int_{t-T_{SV}}^{t} X_i(\tau)\, d\tau = 0 \\ 1, & \text{otherwise} \end{cases} \tag{1}$$

where T_{SV} is the survival time, i.e., the interval of time during which the application can tolerate failures, i.e., missed reception of data.

Therefore, the application layer availability for the i-th URLLC UE can be written as:

$$a_i := \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} a_i(t)\, dt \tag{2}$$

Finally, the average application layer availability, averaged over the total number of URLLC UEs N_{UR}, can be computed as $\overline{a} = \frac{\sum_{i=1}^{N_{UR}} a_i}{N_{UR}}$. Moreover, depending on the transmission direction (uplink or downlink), two average application layer availabilities can be defined, that is, $\overline{a^U}$ and $\overline{a^D}$.

B. COLLISION PROBABILITY

The collision probability of the n-th FL UE, with $n \in \{0, 1, \ldots, N_{FL}\}$, is defined as follows:

$$p_n^C = \frac{C_n}{T_n} \tag{3}$$

where C_n is the number of CB allocations where the n-th UE has collided, and T_n is the total number of utilized CB resources. It immediately follows that the average collision probability, averaged over the total number of FL UEs, is $\overline{p^C} = \frac{\sum_{n=1}^{N_{FL}} p_n^C}{N_{FL}}$.

C. MODEL DOWNLOAD TIME

Fig. 6 formalizes the different timings characterizing a generic iteration x and referring to the n-th FL UE. As already

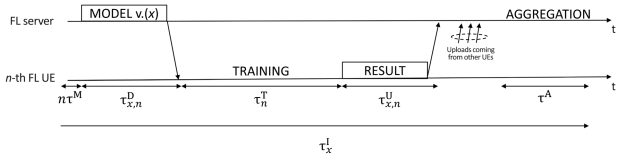


FIGURE 6. Timing diagram of a generic FL iteration x, where the download of model v.(x) for the n-th FL UE starts after $n\tau^M$ w.r.t the beginning of the iteration and lasts $\tau^D_{x,n}$. When the download ends, the n-th FL UE performs training for τ^T_n and then it takes $\tau^U_{x,n}$ to upload the modified version of the model. Upon reception of the training outcomes from all UEs, the FL server ends iteration x by taking τ^A to create model v(x+1).

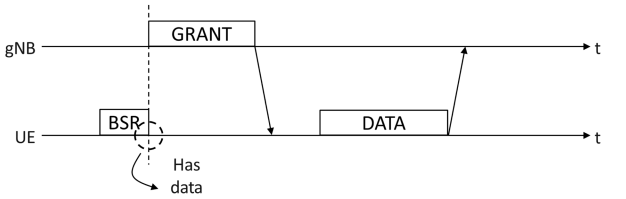


FIGURE 7. Timing diagram of the Instantaneous Buffer Information (IBI) approach, an ideal version of DS used for comparison with CB for NR PUSCH, where the gNB immediately knows the BSR of the UEs as soon as they generate new data and it thus reserves dedicated radio resources for them.

mentioned in Sec. IV-C, the iteration starts when the FL server has a new model ready to be transmitted in unicast to all FL UEs, and the download of the model intended for the n-th FL UE starts after $n\tau^M$ w.r.t the beginning of the iteration.

In this regard, the model download time $\tau^D_{x,n}$ is defined as the time elapsing from the transmission of the first bit of model v.(x) to the reception, by the n-th UE, of its last bit. It immediately follows that the average model download time, averaged over the total number of FL UEs and iterations X, can be computed as $\overline{\tau^D} = \frac{1}{XN_{FL}} \sum_{x=1}^{X} \sum_{n=1}^{N_{FL}} \tau^D_{x,n}$.

D. MODEL UPLOAD TIME

Upon receiving the model v.(x), the n-th FL UE performs local training for a given amount of time τ^T_n. Afterward, it will transmit the result of the training, i.e., the updated version of the model, to the FL server.

Hence, the model upload time $\tau^U_{x,n}$ (see Fig. 6) is defined as the time elapsing from the transmission, by the n-th FL UE, of the first bit of the local updated version of model v.(x), to the reception by the FL server of its last bit. It immediately follows that the average model upload time, averaged over the total number of FL UEs and iterations X, can be computed as $\overline{\tau^U} = \frac{1}{XN_{FL}} \sum_{x=1}^{X} \sum_{n=1}^{N_{FL}} \tau^U_{x,n}$.

E. ITERATION TIME

When the FL server receives the updated versions of model v.(x) from all the FL UEs, it takes τ^A to perform aggregation, i.e., to generate the new version v.(x+1). Since the iteration time x is defined as the time elapsing from the generation of model v.(x) to the creation of model v(x+1) at the server-side, the average iteration time $\overline{\tau}$, averaged over the total number of iterations X, can be computed as $\overline{\tau^I} = \frac{1}{X} \sum_{x=1}^{X} \tau^I_x$.

VI. PERFORMANCE EVALUATION

A. ANALYZED POLICIES

This section briefly summarizes the different policies (and their nomenclature) used in the performance evaluation campaign.

1) DYNAMIC SCHEDULING

It is the basic scheduling mechanism of 5G NR described in Sec. II-B, and it will be labeled as *DS*.

2) INSTANTANEOUS BUFFER INFORMATION

It is an ideal version of DS where the gNB immediately knows the BSR of the UEs as soon as they generate new data, and it will be labeled as *IBI*. More precisely, the IBI approach is summarized in Fig. 7, where, without any reception of SR or BSR, the gNB reserves dedicated radio resources to the UEs that have new data to transmit.

This kind of ideal version of DS is useful for comparison with the considered CB for NR PUSCH design because, in both cases, the SR is not needed, i.e., the impact of the control plane is lower. However, differently from contention, no collisions are present in this case. Indeed, as it will be shown in Sec. VI-B, the considered CB for NR PUSCH outperforms the IBI scheme only in very specific occasions, and overall the performance of IBI is close to the best achievable.

3) CB FOR NR PUSCH WITH RETRANSMISSIONS ON DEDICATED RESOURCES

It refers to the CB design for NR PUSCH described in Sec. III-C, where retransmissions are scheduled via dedicated resources, and it will be labeled as *CB for NR PUSCH with re-tx on dedicated*.

4) CB FOR NR PUSCH WITH RETRANSMISSIONS ON CONTENTION RESOURCES

It refers to the CB design for NR PUSCH described in Sec. III-C, where retransmissions are scheduled via CB resources, and it will be labeled as *CB for NR PUSCH with re-tx on contention*.

B. NUMERICAL RESULTS

Simulation parameters, if not otherwise specified, are reported in Table 1. In particular, some additional information should be provided:

- Based on the chosen numerology (i.e., $\Delta f = 30$ kHz), the 5G slots are 0.5 ms long. Hence, each simulation is formed by 300000 slots since $T_S = 150$ seconds. All results have been obtained by averaging over 10 simulations, that is 3000000 slots, where each simulation was associated with a diverse seed, i.e., a different distribution of UEs and obstacles, as well as an independent channel evolution;

TABLE 1. Simulation parameters.

Parameter	Description	Value
f_c	Carrier frequency	2.6 GHz
B	Overall system bandwidth	40 MHz
Δf	Subcarrier spacing	30 kHz
T_S	Simulation time	150 s
A_{gNB}	gNB antenna number	2
A_{UE}	UEs antenna number	1
P_{TX}^{U}	UE transmit power	0.2 W
P_{TX}^{d}	gNB transmit power	0.5 W
B_D	Obstacle's density	0.15 obstacles/m^2
S	Side of the obstacles	9 m
l	Length of the factory floor	15 m [7]
w	Width of the factory floor	15 m
h	Height of the factory floor	11 m
H_{UE}	UEs' height	1.5 m
H_{gNB}	gNB height	10 m
N_{UR}	Number of URLLC UEs	10
P_{UR}^{U}	Uplink application layer PDU for URLLC UEs	64 B
P_{UR}^{D}	Downlink application layer PDU for URLLC UEs	80 B
τ_B^U	URLLC uplink delay bound	10 ms
τ_B^D	URLLC downlink delay bound	3 ms
τ	URLLC uplink/downlink transmission periodicity	5 ms
τ^M	Time taken by the C/M to generate the PHY PDUs of a given FL model	10 ms
τ^T	Training time of FL UEs	10 s[a]
τ^A	Aggregation time of the FL server	10 s
T_{BO}	Backoff interval	10 slots
T_{SV}	Survival time	15 ms

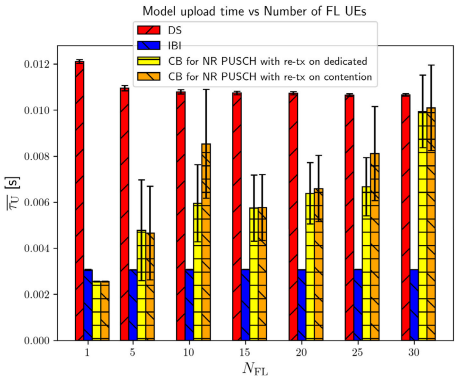
[a]This setting was chosen so as to properly balance the average number of collisions at each FL iteration and considering that the real number depends on many implementation-specific factors (e.g., the hardware used) which are out of the scope of this work.

- Based on the chosen bandwidth and numerology, the overall number of RBs is 112, and the latter constitutes the upper bound for the size of the CB grant in the frequency domain. Conversely, there is no limit in the time domain, i.e., on the available OFDM symbols to be used for CB uplink transmissions;
- The total number of FL iterations, X, is not fixed a priori (and thus it does not appear in Table 1) since it depends on the specific settings of a given simulation run, such as the number of UEs, the model size, the considered retransmission policy, etc.;
- As far as the FL model is concerned, we follow the approach of [14], where both the FL server and UEs are assumed to implement a Deep Neural Network (DNN) following the MobileNets architecture [62], i.e., a class of DNNs models based on a streamlined architecture for mobile and embedded vision applications. However, it is important to underline that we did not implement any DNN because we focus here on the communication part of this traffic. Hence, we set the related FL timings (i.e., training and aggregation time) and model sizes based on the study in [62];
- All results show a confidence interval with a probability of 95%;
- When FL UEs have to rely on DS, SRs periodicity is 1 slot, i.e., 0.5 ms;
- As already described in Sec. III, the maximum number of HARQ uplink/downlink retransmissions, N_{RX}, is set to 0 when FL UEs retransmit via contention, otherwise it is set to 10. On the other hand, N_{RX} is set to 3 or 2 for uplink and downlink URLLC transmissions, respectively;
- Among the same category of UEs, proportional fair is used as the radio resource assignment algorithm [63];
- URLLC UEs have higher priority w.r.t FL UEs. Remarkably, retransmissions have a higher priority w.r.t first transmissions, and this means that retransmissions of FL TB are prioritized w.r.t first transmissions of URLLC TB. In a nutshell, the different cases can be sorted in descending priority order as follows:
  1) Retransmissions of URLLC UEs;
  2) Dedicated retransmissions of FL UEs (when considering CB for NR PUSCH with re-tx on dedicated, as described in Sec. III-C1);
  3) First transmissions of URLLC UEs;
  4) First transmissions or CB retransmissions of FL UEs (when considering CB for NR PUSCH with re-tx on contention, as described in Sec. III-C2).

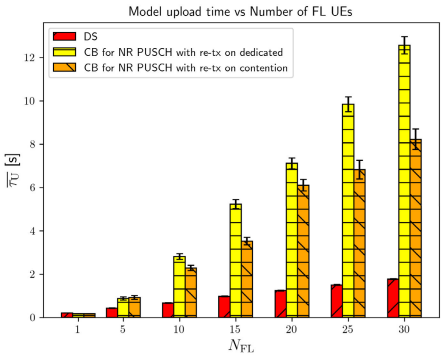
All that being said, Fig. 8 is a collection of four plots showing the average model upload/download time as a function of the number of FL UEs, and by considering FL model sizes, P_{FL}, of 12 and 16 kB, as well as 2 MB (the latter is the minimum size considered in [62]). It is clearly evident that the considered CB approach outperforms DS when the model size is 12 and 16 kB, for all the considered values of N_{FL}, both in upload and download. Notice that the considered CB for NR PUSCH is applied to any uplink communication of the FL UEs; therefore, the gain in the model download time is due to an improvement of the time needed to transmit the TCP ACKs.

However, the gain of CB for NR PUSCH ceases to be true for a larger model size of 2 MB, even for a small number of FL UEs (i.e., $N_{FL} > 1$), and this is more evident for the model upload time. Indeed, in upload, a larger model size increases both the number of transmissions (due to segmentation [48]) and their dimensions, whereas, in download, the number of TCP ACKs transmissions increases but their size remains unaffected.

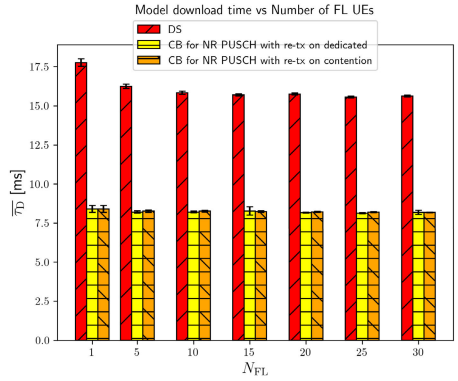
The two retransmission policies perform similarly, especially for low model sizes, because of a sufficiently low amount of new collisions during the retransmissions. Nonetheless, counterintuitively, Fig. 8b shows lower model upload times when retransmitting via CB grants. The reason for that is two-fold. On the one hand, reserving dedicated resources for retransmissions of a non-negligible size significantly shrinks the available resources that can be used for future CB allocations (due to the higher priority given to



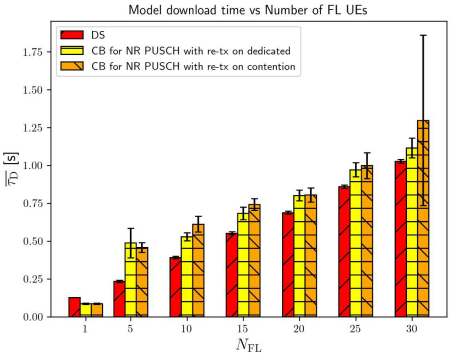
(a) Model upload time as a function of the number of FL UEs when $P_{FL} = 12$ kB.



(b) Model upload time as a function of the number of FL UEs when $P_{FL} = 2$ MB.



(c) Model download time as a function of the number of FL UEs when $P_{FL} = 16$ kB.



(d) Model download time as a function of the number of FL UEs when $P_{FL} = 2$ MB.

FIGURE 8. Model upload/download time as a function of the number of FL UEs, and by considering model sizes of 12/16 kB and 2 MB. The different curves represent four different situations, that is, the FL UEs are scheduled (i) via DS, or (ii) IBI, or (iii) perform CB for NR PUSCH with re-tx on dedicated resources, (iv) or perform CB for NR PUSCH with re-tx on CB resources.

retransmissions w.r.t first transmissions), thereby prolonging the overall transfer time of the case with retransmissions on dedicated resources (this aspect will be better clarified later when discussing Fig. 9). On the other hand, looking at behaviors of single UEs, we noticed that the number of UEs which do *not* complete all iterations within the simulation time is higher (on average) when retransmissions happen via CB resources compared to retransmitting on dedicated resources (the reader can recall from Fig. 5 that a UE finalizes an iteration only when it concludes its model upload). Due to this second reason, the lower model upload time for the case of retransmissions on CB resources when $P_{FL} = 2$ MB is not fully reflecting a better performance from a FL iteration point of view compared to retransmissions on dedicated resources. Indeed, the latter approach, although with higher upload transfer times, allows more UEs to complete the FL iterations also in case of higher load as retransmissions are (i) without collisions, and (ii) can exploit a per-UE link adaptation process for the MCS selection.

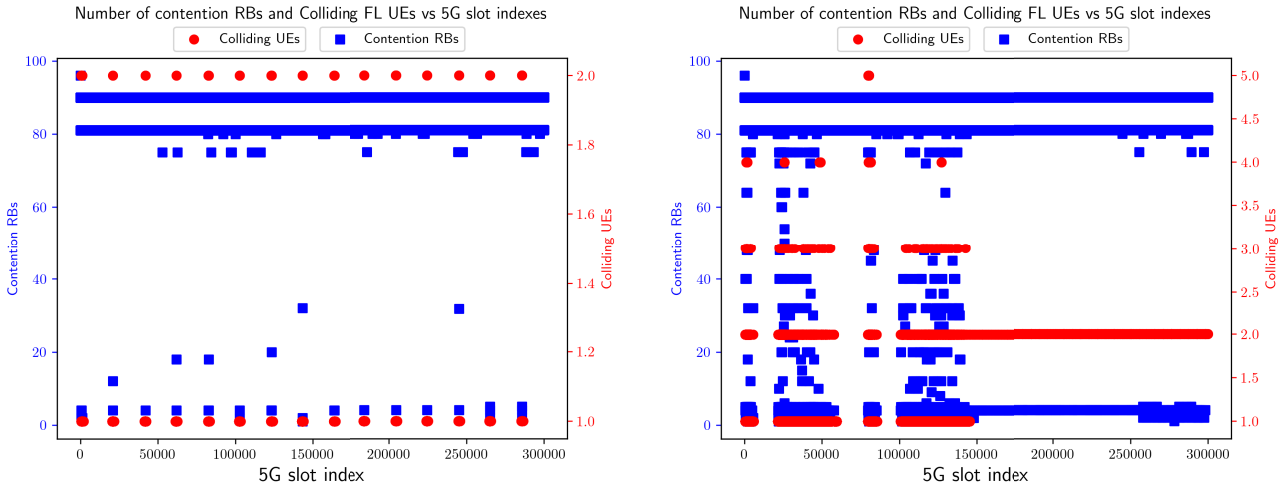
As expected, CB for NR PUSCH outperforms IBI only in a very specific case, i.e., when considering a single FL UE for small model sizes. This is because, the absence of collisions highlights the gain provided by CB, i.e., the single UE is likely to immediately perform the few transmissions needed to upload the small model because it already received the CB grant when the data is generated (see Fig. 2b). When the number of UEs increases, IBI remains the best-performing

TABLE 2. Conditions under which CB for NR PUSCH with retransmissions on dedicated resources provides lower model upload times w.r.t DS as a function of the model size P_{FL} and maximum bandwidth allowed for CB transmissions B_{CB} when considering $N_{FL} = 30$.

		P_{FL} [kB]			
		1	**8**	**12**	**16**
B_{CB} [MHz]	**5**	YES	NO	NO	NO
	10	YES	YES	YES (up to 5 UEs)	NO
	20	YES	YES	YES (up to 20 UEs)	NO
	40	YES	YES	YES	NO

TABLE 3. Average collision probability and average iteration time as a function of the number of FL UEs, when considering the model size of 2 MB, and the CB approach for NR PUSCH with retransmissions on dedicated resources.

N_{FL}	$\overline{p^C}$	$\overline{\tau^I}$ [s]
1	0	20.2498
5	0.0206	22.2051
10	0.0426	25.1477
15	0.0649	29.6163
20	0.0826	32.1696
25	0.1017	37.2457
30	0.1238	42.4982



(a) Case with $P_{FL} = 12$ kB. (b) Case with $P_{FL} = 2$ MB.

FIGURE 9. Number of RBs used for CB allocations (blue dots) and number of colliding FL UEs (red dots), as a function of the 5G slot indexes contained in one simulation run, when considering 30 FL UEs transmitting and receiving models made of 12 kB (on the left) or 2 MB (on the right) by means of CB for NR PUSCH with retransmissions on dedicated resources.

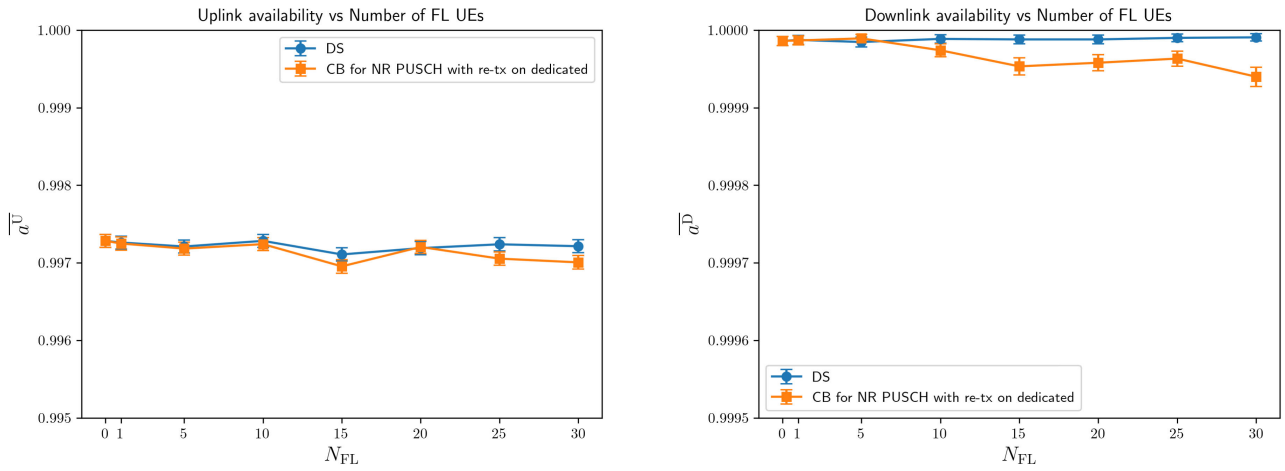
policy due to its ideality. For this reason, we will not further consider the performance of IBI in the analysis.

To explain the reasons behind the choice of 12 and 16 kB as the FL model sizes for Figs. 8a and 8c, Table 2 shows the conditions under which CB for NR PUSCH with retransmissions on dedicated resources provides lower model upload times w.r.t DS as a function of the model size, P_{FL}, and maximum bandwidth allowed for CB transmissions B_{CB} (out of the overall system bandwidth B) when considering $N_{FL} = 30$. As can be seen, independently of the considered B_{CB} value, a model size of 12 kB is the maximum value for which the considered CB design provides benefits over DS, thereby motivating the model size choice of Fig. 8a. Of course, the same holds for the model download times, i.e., 16 kB is the maximum model size for which there are gains, as well as when considering the design employing retransmissions on CB resources.

Next, in Table 3 we show the average collision probability, $\overline{p^C}$, and average iteration time, $\overline{\tau^I}$, when considering the model size of 2 MB, and the considered CB for NR PUSCH with retransmissions on dedicated resources. It can be seen that the average iteration times can be tens of seconds even for low loads, thus indicating that a FL training can be quite large if it involved higher loads (i.e., a non-negligible training phase has to be performed before having FL-based cameras ready to perform image recognition).

It is interesting to notice that the average collision probabilities are relatively small (at most $\sim 12\%$). This consideration suggests that the introduction of a collision framework in a 5G NR IIoT network produces an additional phenomenon that cannot be merely controlled by looking only at the collision probability. This thought is confirmed through Fig. 9, where it illustrates both, the number of RBs used for CB allocations (blue dots) and the number of colliding FL UEs (red dots), as a function of the 5G slot indexes contained in one simulation run. Two model sizes are compared, i.e., 12 kB (on the left) and 2 MB (on the right). The plot refers to one cell, and a total amount of 30 FL UEs which are scheduled via CB for NR PUSCH with retransmissions on dedicated resources. Among the total amount of 112 RBs, the gNB never allocates more than 96 RBs for the CB allocation due to the presence of the higher priority always-on URLLC traffic. With a fixed periodicity, the CB allocation shrinks to 81 RBs due to periodic control plane signals, such as CQIs. However, it can be clearly noted that, when considering model sizes of 12 kB, the number of collisions is sporadic and they never involve more than 2 UEs. Consequently, the number of RBs used for the CB allocation (blue dots) remains high for most of the time. This ceases to be true when considering model sizes of 2 MB, because, for the vast majority of the simulation, the CB allocation is shrinked to a few RBs (close to 5), thus resulting in very long download/upload transfer times (as previously shown in Fig. 8b and 8d). Indeed, the size of the model is such that, even the retransmissions of a few number of colliding UEs (no more than 5) need most of the dedicated radio resources. Consequently, the higher priority given to such retransmissions dramatically reduces the amount of resources that can be used for future first transmissions via CB allocations. This phenomenon explains the performance degradation of the considered CB scheme when dealing with larger model size despite a relatively low collision probability, and, at the same time, it motivates the benefit in retransmitting on CB resources for this specific case (as explained when describing Fig. 8b).

Finally, Fig. 10 shows the URLLC uplink/downlink availability (see Sec. V-A) as a function of N_{FL}, by comparing the two cases where FL UEs transmit/receive a model of 2 MB (i.e., the worst case) and are scheduled (i) via DS or (ii) by means of the considered CB design with retransmissions on dedicated resources. As expected,



(a) Average uplink availability of 10 URLLC UEs.

(b) Average downlink availability of 10 URLLC UEs.

FIGURE 10. Average uplink/downlink availability as a function of N_{FL}, by comparing the two cases where FL UEs transmit/receive a model of 2 MB (i.e., the worst case) and are scheduled (i) via DS or (ii) by means of the considered CB for NR PUSCH with retransmissions on dedicated resources.

the uplink/downlink availability remain quite stable when increasing the number of FL UEs due to the higher priority of the URLLC traffic. However, the considered CB design slightly (0.05%) decreases the uplink/downlink availability because FL retransmissions have higher priority w.r.t first transmissions of URLLC UEs (thus the uplink/downlink delay bound is less easily met).

VII. CONCLUSION AND FUTURE WORK

In this paper, we presented a CB 5G NR framework for the PUSCH transmissions of FL UEs in an IIoT scenario which also includes URLLC UEs. By means of near-product 3GPP-compliant network simulations, we showed that the considered CB for NR PUSCH design provides benefits over DS for FL upload/download times in case of sufficiently small model sizes (up to 12/16 kB, as in [62]). Additionally, such a CB design scales well with an increasing number of UEs and does not meaningfully degrade the performance of the URLLC traffic (uplink/downlink availability impacted at most by 0.05%). However, for larger model sizes, DS shows much better robustness of performance and scalability, and gains with CB for NR PUSCH are not present because the size of the CB allocations shrinks very quickly even for relatively low collision probabilities (close to 12% at maximum), thereby leading to longer transfer times.

The study also opens other interesting research trends. For example, additional analyses could add the comparison with other SPS mechanisms (e.g., configured grant), identify the proper metrics that the network could monitor to optimize the CB allocation (e.g., network load), consider more complex CB design (e.g., reserving multiple CB allocations per different sets of UEs), offload CB by only mapping specific traffic types (e.g., information with low-reliability requirements), or assess the impact on the performance of wider areas (e.g., outdoor environments).

REFERENCES

[1] S. E. Elayoubi, M. Fallgren, P. Spapis, G. Zimmermann, D. Martín-Sacristán, C. Yang, S. Jeux, P. Agyapong, L. Campoy, Y. Qi, and S. Singh, "5G service requirements and operational use cases: Analysis and METIS II vision," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Jun. 2016, pp. 158–162.

[2] J. Navarro-Ortiz, P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "A survey on 5G usage scenarios and traffic models," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 905–929, 2nd Quart., 2020.

[3] L. Chettri and R. Bera, "A comprehensive survey on Internet of Things (IoT) toward 5G wireless systems," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 16–32, Jan. 2020.

[4] O. O. Erunkulu, A. M. Zungeru, C. K. Lebekwe, M. Mosalaosi, and J. M. Chuma, "5G mobile communication applications: A survey and comparison of use cases," *IEEE Access*, vol. 9, pp. 97251–97295, 2021.

[5] M. Elsayed and M. Erol-Kantarci, "AI-enabled future wireless networks: Challenges, opportunities, and open issues," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 70–77, Sep. 2019.

[6] M. Gundall, J. Schneider, H. D. Schotten, M. Aleksy, D. Schulz, N. Franchi, N. Schwarzenberg, C. Markwart, R. Halfmann, P. Rost, D. Wübben, A. Neumann, M. Düngen, T. Neugebauer, R. Blunk, M. Kus, and J. Grießbach, "5G as enabler for industrie 4.0 use cases: Challenges and concepts," in *Proc. IEEE 23rd Int. Conf. Emerg. Technol. Factory Autom. (ETFA)*, vol. 1, Sep. 2018, pp. 1401–1408.

[7] *Integration of Industrial Ethernet Networks With 5G Networks*, 5G-ACIA, ZVEI, Berlin, Germany, Nov. 2019.

[8] *Expanded 6G Vision, Use Cases and Societal Values*, Hexa-X, Finland, 2021.

[9] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, and M. Zorzi, "Toward 6G networks: Use cases and technologies," *IEEE Commun. Mag.*, vol. 58, no. 3, pp. 55–61, Mar. 2020.

[10] W. Jiang, B. Han, M. A. Habibi, and H. D. Schotten, "The road towards 6G: A comprehensive survey," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 334–366, 2021.

[11] H. Tataria, M. Shafi, A. F. Molisch, M. Dohler, H. Sjöland, and F. Tufvesson, "6G wireless systems: Vision, requirements, challenges, insights, and opportunities," *Proc. IEEE*, vol. 109, no. 7, pp. 1166–1199, Jul. 2021.

[12] F. Tao, H. Zhang, A. Liu, and A. Y. C. Nee, "Digital twin in industry: State-of-the-art," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2405–2415, Apr. 2019.

[13] M. Ganjalizadeh, H. S. Ghadikolaei, J. Haraldson, and M. Petrova, "Interplay between distributed AI workflow and URLLC," 2022, *arXiv:2208.01352*.

[14] *5G System (5GS); Study on Traffic Characteristics and Performance Requirements for AI/ML Model Transfer (Release 18*, 3GPP, document TR 22.874, 2021.
[15] P. Huang, L. Xiao, S. Soltani, M. W. Mutka, and N. Xi, "The evolution of MAC protocols in wireless sensor networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 101–120, 1st Quart., 2013.
[16] M. Hadded, P. Muhlethaler, A. Laouiti, R. Zagrouba, and L. A. Saidane, "TDMA-based MAC protocols for vehicular ad hoc networks: A survey, qualitative analysis, and open research issues," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2461–2492, 4th Quart., 2015.
[17] J. Zhang, L.-L. Yang, L. Hanzo, and H. Gharavi, "Advances in cooperative single-carrier FDMA communications: Beyond LTE-advanced," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 730–756, 2nd Quart., 2015.
[18] N. Abu-Ali, A. M. Taha, M. Salah, and H. Hassanein, "Uplink scheduling in LTE and LTE-advanced: Tutorial, survey and evaluation framework," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 3, pp. 1239–1265, 3rd Quart., 2014.
[19] M. Salem, A. Adinoyi, M. Rahman, H. Yanikomeroglu, D. Falconer, Y.-D. Kim, E. Kim, and Y.-C. Cheong, "An overview of radio resource management in relay-enhanced OFDMA-based networks," *IEEE Commun. Surveys Tuts.*, vol. 12, no. 3, pp. 422–438, 3rd Quart., 2010.
[20] E. Yaacoub and Z. Dawy, "A survey on uplink resource allocation in OFDMA wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 2, pp. 322–337, 2nd Quart., 2012.
[21] K. S. Gilhousen, I. M. Jacobs, R. Padovani, A. J. Viterbi, L. A. Weaver Jr., and C. E. Wheatley, "On the capacity of a cellular CDMA system," *IEEE Trans. Veh. Technol.*, vol. 40, no. 2, pp. 303–312, May 1991.
[22] K. Kusume, G. Bauch, and W. Utschick, "IDMA vs. CDMA: Analysis and comparison of two multiple access schemes," *IEEE Trans. Wireless Commun.*, vol. 11, no. 1, pp. 78–87, Jan. 2012.
[23] C. Buratti, G. Cuozzo, and R. Verdone, "OCDMA: A MAC protocol for industrial intra-machine terahertz network," *J. Infr., Millim., THz Waves*, vol. 2022, pp. 1–26, Jan. 2022.
[24] I. Ahmed, H. Khammari, A. Shahid, A. Musa, K. S. Kim, E. De Poorter, and I. Moerman, "A survey on hybrid beamforming techniques in 5G: Architecture and system model perspectives," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3060–3097, 4th Quart., 2018.
[25] W. Choi, A. Forenza, J. G. Andrews, and R. W. Heath Jr., "Opportunistic space-division multiple access with beam selection," *IEEE Trans. Commun.*, vol. 55, no. 12, pp. 2371–2380, Dec. 2007.
[26] K. Akkarajitsakul, E. Hossain, D. Niyato, and D. I. Kim, "Game theoretic approaches for multiple access in wireless networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 13, no. 3, pp. 372–395, 3rd Quart., 2011.
[27] D. Zheng, Y. Zhao, and Y.-D. Yao, "An investigation of cooperative slotted ALOHA with the capture effect," *IEEE Commun. Lett.*, vol. 18, no. 4, pp. 572–575, Apr. 2014.
[28] R. Y. W. Lam, V. C. M. Leung, and H. C. B. Chan, "Polling-based protocols for packet voice transport over IEEE 802.11 wireless local area networks," *IEEE Wireless Commun.*, vol. 13, no. 1, pp. 22–29, Feb. 2006.
[29] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 3, pp. 535–547, Mar. 2000.
[30] M. Ghazvini, N. Movahedinia, K. Jamshidi, and N. Moghim, "Game theory applications in CSMA methods," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 3, pp. 1062–1087, 3rd Quart., 2013.
[31] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2294–2323, 3rd Quart., 2018.
[32] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Oct. 2017.
[33] *Contention Based Uplink Transmission*, 3GPP, document R2-093812, RAN2#66bis, 2009.
[34] Z. Arnjad, A. Sikora, B. Hilt, and J.-P. Lauffenburger, "Latency reduction for narrowband LTE with semi-persistent scheduling," in *Proc. IEEE 4th Int. Symp. Wireless Syst. Int. Conf. Intell. Data Acquisition Adv. Comput. Syst. (IDAACS-SWS)*, Sep. 2018, pp. 196–198.
[35] Y. Feng, A. Nirmalathas, and E. Wong, "A predictive semi-persistent scheduling scheme for low-latency applications in LTE and NR networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.
[36] G. Cuozzo, S. Cavallero, F. Pase, M. Giordani, J. Eichinger, C. Buratti, R. Verdone, and M. Zorzi, "Enabling URLLC in 5G NR IIoT networks: A full-stack end-to-end analysis," in *Proc. Joint Eur. Conf. Netw. Commun. 6G Summit*, Jun. 2022, pp. 333–338.
[37] J. Ding, M. Nemati, S. R. Pokhrel, O.-S. Park, J. Choi, and F. Adachi, "Enabling grant-free URLLC: An overview of principle and enhancements by massive MIMO," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 384–400, Jan. 2022.
[38] M. B. Shahab, R. Abbas, M. Shirvanimoghaddam, and S. J. Johnson, "Grant-free non-orthogonal multiple access for IoT: A survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1805–1838, 3rd Quart., 2020.
[39] H. Grama Srinath, M. Rana, and N. M. Balasubramanya, "Grant-free access for mMTC: A performance analysis based on number of preambles, repetitions, and retransmissions," *IEEE Internet Things J.*, vol. 9, no. 16, pp. 15169–15183, Aug. 2022.
[40] G. Berardinelli, N. H. Mahmood, R. Abreu, T. Jacobsen, K. Pedersen, I. Z. Kovács, and P. Mogensen, "Reliability analysis of uplink grant-free transmission over shared resources," *IEEE Access*, vol. 6, pp. 23602–23611, 2018.
[41] Y. Liu, Y. Deng, M. Elkashlan, A. Nallanathan, and G. K. Karagiannidis, "Analyzing grant-free access for URLLC service," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 741–755, Mar. 2021.
[42] M. C. Lucas-Estañ, J. Gozalvez, and M. Sepulcre, "On the capacity of 5G NR grant-free scheduling with shared radio resources to support ultra-reliable and low-latency communications," *Sensors*, vol. 19, no. 16, p. 3575, Aug. 2019.
[43] T. Jacobsen, R. Abreu, G. Berardinelli, K. Pedersen, P. Mogensen, I. Z. Kovacs, and T. K. Madsen, "System level analysis of uplink grant-free transmission for URLLC," in *Proc. IEEE Globecom Workshops*, Dec. 2017, pp. 1–6.
[44] M. C. Lucas-Estañ and J. Gozalvez, "Sensing-based grant-free scheduling for ultra reliable low latency and deterministic beyond 5G networks," *IEEE Trans. Veh. Technol.*, vol. 71, no. 4, pp. 4171–4183, Apr. 2022.
[45] N. H. Mahmood, R. Abreu, R. Böhnke, M. Schubert, G. Berardinelli, and T. H. Jacobsen, "Uplink grant-free access solutions for URLLC services in 5G new radio," in *Proc. 16th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2019, pp. 607–612.
[46] K. Au, L. Zhang, H. Nikopour, E. Yi, A. Bayesteh, U. Vilaipornsawai, J. Ma, and P. Zhu, "Uplink contention based SCMA for 5G radio access," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2014, pp. 900–905.
[47] F. Pase, M. Giordani, G. Cuozzo, S. Cavallero, J. Eichinger, R. Verdone, and M. Zorzi, "Distributed resource allocation for URLLC in IIoT scenarios: A multi-armed bandit approach," 2022, *arXiv:2211.12201*.
[48] S. Parkvall, E. Dahlman, and J. Skold, *5G NR: The Next Generation Wireless Access Technology*. New York, NY, USA: Academic, 2018.
[49] *Feasibility Study for Further Advancements for E-UTRA (LTE-Advanced) (Release 9)*, 3GPP, document TR 36.912, 2010.
[50] *Study on Latency Reduction Techniques for LTE (Release 14)*, 3GPP, document TR 36.881, 2016.
[51] *Analysis on Resource Efficiency of Uplink Access Solutions*, 3GPP, document document R2-154122, 2015.
[52] *Contention Based Uplink Transmission*, 3GPP, document R2-154191, 2015.
[53] Y. Huang, L. Hu, H. Tong, F. Wang, J. Jin, G. Liu, and Q. Wang, "Reference signal design for demodulation of higher-order MU-MIMO in 3D-MIMO systems," in *Proc. IEEE/CIC Int. Conf. Commun. China-Workshops (CIC/ICCC)*, Nov. 2015, pp. 44–47.
[54] *Performance Evaluation of CB-PUSCH*, 3GPP, document R2-156402, 2015.
[55] 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.
[56] M. Condoluci, G. Araniti, T. Mahmoodi, and M. Dohler, "Enabling the IoT machine age with 5G: Machine-type multicast services for innovative real-time applications," *IEEE Access*, vol. 4, pp. 5555–5569, 2016.
[57] *Study on Channel Model for Frequencies From 0.5 to 100 GHz*, 3GPP, document TR 38.901, 2022.
[58] *Study on Communication for Automation in Vertical Domains (Release 16)*, 3GPP, document TR 22.804, 2020.
[59] *Service Requirements for Cyber-Physical Control Applications in Vertical Domains*, document TS 122 104, ETSI, 2020.

[60] *New Services & Applications With 5G URLLC*, 5G Americas, Bellevue, WAS, USA, 2018.
[61] *5G E2E Technology to Support Vertical URLLC Requirements*, NGMN, Frankfurt am Main, Germany, 2020.
[62] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
[63] A. Goldsmith, *Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

GIAMPAOLO CUOZZO (Member, IEEE) received the B.Sc. degree (Hons.) in electronics and telecommunications engineering, the M.Sc. degree (Hons.) in telecommunications engineering, and the Ph.D. degree in electronics, telecommunications, and information technologies engineering (ET-IT) from the University of Bologna, in 2017, 2019, and 2022, respectively. He is currently a Research Area Leader at the National Laboratory of Wireless Communications (WiLab) of CNIT (the National, Inter-University Consortium for Telecommunications). He is the author of one book, one invention, and six articles. His research activity is focused on the study, development, and validation of wireless networks for the Industrial Internet of Things, with a particular focus on signal processing techniques and MAC protocol design for THz-based systems, as well as scheduling optimization algorithms for 5G NR networks. His interests also include experimental activities that exploit current wireless technologies, like 5G, LoRa, Zigbee and NB-IoT.

JONAS PETTERSSON received the M.Sc. degree in engineering physics from Umeå University, in 1997. He joined Ericsson Research, Luleå, in 1997. He is currently a Master's Researcher with the Protocol and End-to-End Performance Group. He has mainly worked in the areas of quality of experience, radio resource management, and scheduling for 3G, 4G, 5G, and now 6G. His research interests include time-critical communication in 5G and medium access control for 6G.

MASSIMO CONDOLUCI received the B.Sc. and M.Sc. degrees in telecommunications engineering from the Mediterranea University of Reggio Calabria, Italy, in 2008 and 2011, respectively, and the Ph.D. degree in information technology from the Mediterranea University of Reggio Calabria, in 2016, with a focus on access optimization for the IoT traffic and resource allocation for multicasting. From 2015 to 2017, he was a Research Associate with the Centre for Telecommunications Research (CTR), King's College London, U.K., working on fixed-mobile convergence and radio-access optimization for haptic communications. He joined Ericsson Research, in 2018. He has worked on network exposure in 5G systems for automotive use cases, contributing to 3GPP standardization and the 5G Automotive Association (5GAA). He is currently a Senior Researcher with Ericsson Research. He also focuses on beyond 5G mobile network architecture and protocols.