Performance Evaluation of Ultra-Reliable Low-Latency Communication Over Unlicensed Spectrum — Source link

Ayat Zaki Hindi, Salah Eddine Elayoubi, Tijani Chahed

Institutions: CentraleSupélec, Centre national de la recherche scientifique

Published on: 20 May 2019 - International Conference on Communications

Topics: Spectrum management

Related papers:
- Channel Access Enhancements in unlicensed spectrum for NR URLLC transmissions
- Feasibility of URLLC in Unlicensed Spectrum
- Allocation of licensed and unlicensed spectrum in heterogeneous networks
- Discontinuous transmission-enabled licensed-assisted access for energy-efficient hetnets
- Holistic Small Cell Traffic Balancing across Licensed and Unlicensed Bands
Performance evaluation of Ultra-Reliable Low-Latency Communication over unlicensed spectrum
Ayat Hindi, Salah Eddine Elayoubi, Tijani Chahed

To cite this version:

HAL Id: hal-02117089
https://hal.archives-ouvertes.fr/hal-02117089
Submitted on 2 May 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Performance evaluation of Ultra-Reliable Low-Latency Communication over unlicensed spectrum

Ayat Zaki Hindi, Salah-Eddine Elayoubi, Tijani Chahed

1 Orange Labs, 44 Avenue de la Republique, 92320 Chatillon, France
2 Institut Mines-Telecom; Telecom SudParis; UMR CNRS SAMOVAR, 9 Rue Charles Fourier, 91000 Evry, France
3 CentraleSupelec, 3 Rue Joliot Curie, 91190 Gif-sur-Yvette, France

ayat.zaki hindi@orange.com, salaheddine.elayoubi@centralesupelec.fr, tijani.chahed@telecom-sudparis.eu

Abstract—We consider in this paper the transport of Ultra-Reliable Low-Latency Communication (URLLC) uplink traffic over unlicensed spectrum. We specifically consider Licensed Assisted Access (LAA) system and study the feasibility of the strict reliability and delay requirements by the means of an exact formulation which incorporates the use of a timer tracking the lifetime of each packet and identifying delay budget violation. When the delay constraint is too tight, unlicensed spectrum alone is not sufficient when the traffic density is high and we propose the use of 5G licensed spectrum to compensate for the lacking resources. We incorporate this usage of licensed spectrum for delayed packets in the model and show how to calculate the performances for a given amount of additional 5G resources for a grant-free or a grant-based allocation. We validate our analytical models numerically and dimension the joint licensed/unlicensed system resources so as to meet both reliability and delay constraints.

Keywords: URLLC, LAA, LTE, MAC-delay, 5G networks.

I. INTRODUCTION

5G networks are expected to support new applications beyond those currently transported by existing 4G. One of the most important foreseen applications is Ultra-Reliable Low-Latency Communication (URLLC) which is a Machine Type Communication (MTC) involving a potentially huge number of connected objects, with stringent constraints both on reliability and latency, set to 99.999% and 1ms, respectively [1]. These constraints are tightly linked, as retransmissions in case of failure are limited by the delay budget, and a packet that is not received within this budget is considered as lost.

Several methods were proposed to enhance the performance of LTE so as to make it able to transport URLLC [1], but the stringent reliability and delay constraints come at large resource consumption cost. The scarcity of the licensed spectrum calls then for the usage of unlicensed spectrum that is more available, and the de facto system called Multefire [9] which deploys uplink/downlink unlicensed transmission is envisioned as a possible solution for URLLC in confined areas like factories. Multefire is an extension of the LTE License-Assisted Access (LAA) [8], which is a new 3GPP standard proposed to offload part of LTE traffic on the unlicensed spectrum. The objective of this paper is to evaluate the capability of unlicensed spectrum for carrying URLLC traffic and to explore licensed spectrum as a mean to ensure reliability in case of initial failure.

There has been works which study transport of URLLC over LTE, for instance [2], but only a few considered the use of LAA, such as [3], mainly because of the existence of other technologies such as WiFi on the same unlicensed bands, which decreases the reliability of the system. Other studies focused on the impact of LAA on WiFi performance, for instance [4] where LAA is modeled using a Markov chain similar to the well known Bianchi model for WiFi [5]. These works however do not quantify delay which is essential in our case to study URLLC transport performance. For coping with this, [6] proposed a probabilistic formulation to calculate reliability for a given delay budget, and [7] derived probability generating functions to obtain the delay distribution. These methods are computationally intensive and do not allow introducing hard delay constraints as for URLLC.

This paper considers unlicensed spectrum for URLLC and verifies its ability to ensure reliability and latency in confined environment like factories, i.e., without interference from other systems like WiFi. We adopt an analytical approach extending the Bianchi model [5] to a system operating in our case on reliability and delay constraints. Our model includes a timer which expires when a certain delay budget is reached. The joint use of this timer with the one used for the backoff reduces the scalability of the model when it comes to large timers and large number of backoff stages. In order to remedy to this, we further propose an approximation which handles efficiently the presence of these two delay timers. We validate our model using event-driven simulations and show a good match.

Our results show that unlicensed spectrum could handle URLLC traffic for low traffic conditions, while a large traffic leads to excessive delay due to collisions, even for a system with no external interference. We then study the usage of 5G licensed spectrum as a support solution for URLLC packets that could not be served within a certain delay budget. As the URLLC traffic does not allow waiting for scheduling grants, a grant-free system is considered as allowed by the current 5G standard [10]. In this case, a pool of 5G resources are reserved for URLLC and packets are sent there by selecting randomly a resource. Collisions are then possible and a proper dimen-
sioning of resources is needed for ensuring the reliability. We obtain in this case the minimum resources which 5G needs to add in order to meet our performance requirement and show that the presence of unlicensed spectrum reduces drastically the amount of reserved licensed resources compared to a stand-alone 5G network.

The remainder of this paper is organized as follows. Section II presents the system under study. We describe in Section III our models for the MAC access both for unlicensed and licensed spectrums. Section IV contains our numerical results. Section V eventually concludes the paper and indicates some future work perspectives.

II. SYSTEM

Our system consists of \( N \) stations operating on unlicensed spectrum within a range of coverage, without the presence of any interfering sources (such as Wi-Fi), which may represent the communicating machines in the illustrative scenario of an isolated plant in an industry 4.0 use case scenario. In the following, we will use the term LAA when referring to the operation on the unlicensed spectrum as we will consider later the possibility of using the licensed spectrum as a support system, denoted by the term licensed 5G. Note that our usage of the term LAA implies implicitly that LAA is extended to the uplink, which is a natural evolution of LAA standards.

We consider URLLC traffic, generated by potentially huge number of machines, each requiring a high reliability, on the order of 99,999\%, and a strict delay constraint, typically 1ms. LAA may turn to be not sufficient for the transport of such traffic with such constraints, we hence propose the joint use of licensed 5G, whenever LAA alone fails to meet the reliability-delay performance.

A. Operation on unlicensed spectrum

LAA stations access the unlicensed medium using a load-based Listen-Before-Talk (LBT) mechanism, similar to the slot based WiFi’s Distributed Coordination Function (DCF) protocol. When a packet is ready to be transmitted, a random number, called the Backoff Timer (BT), between 0 and \( W_0 - 1 \) is chosen, where \( W_0 \) is the contention window size. If the medium is sensed to be idle, BT is decremented by one, else it is suspended until an idle slot is sensed again. When BT reaches zero, the station transmits the packet, then waits for the BS reply to know whether the transmission was successful or it encountered a collision. If a collision occurs, the station retries to send the packet as previously until the packet reaches the maximum allowed number of retrials \( m \), in which case the packet is dropped.

Every trial to transmit is called a backoff stage, and the size of the contention window in every stage may vary depending on the used category of LBT, for example in cat3 LBT the window size is fixed in all stages, but in cat4 LBT, it is doubled after every stage until it reaches the maximum size \( W_{\text{max}} \) [4].

B. Operation on licensed 5G

In licensed 5G, operation is done at the Transmission Time Interval (TTI) level which is defined as the smallest time unit during which a transmission can happen. All 5G stations are required to be synchronized. The existing methods to access the licensed medium in the uplink are Grant-based (GB) scheduling and Grant-free (GF) on a common pool. In GB scheduling, the Base station (BS) is responsible for resource allocation. When a station wants to send a packet, it sends a scheduling request to the BS at the beginning of the TTI. After the BS receives the request, it processes it and transmits a scheduling grant to that station which indicates the position of the allocated resources in time and frequency, and accordingly, the station sends its packet. The time spent in each step (scheduling request, request processing, scheduling grant and data transmission) is considered to be one TTI, thus this approach is time consuming and does not meet URLLC requirements of latency although it is collision free and can achieve a high reliability.

The substitute solution is GF transmission, where resources are accessible without prescheduling. Collisions are present here because resources are chosen randomly by the stations, but they can be handled if a multiple access scheme is deployed, for example, using Code Division Multiple Access (CDMA) or Non-Orthogonal Multiple Access (NOMA).

C. Joint transmission on both systems

5G systems profit from their capability of using both licensed and unlicensed spectrum, but it is not obvious what the optimum method to utilize each of them is. We propose in this work two methods to jointly utilize them:

- Duplication: The idea is to send a copy of the packet over both links; it is enough to receive one copy correctly to consider it as a successful transmission. In this case, the delay budget on each system is equal to the overall delay constraint. The principle drawback of this method is duplicating the load of the system which may saturate the licensed link in case of limited resources.
- Series sharing: With this method, we assume that all packets are first transmitted over the unlicensed link, but if a packet is not served within some delay budget, it stops attempting to be sent over the unlicensed spectrum and switches to the licensed link. The allowed delay budget on the unlicensed link must be smaller than the total delay constraint, and more precisely the time needed to send the packet over the licensed link is deducted from the unlicensed delay budget.

Both methods will be evaluated in the numerical application section.

III. MODELING MAC ACCESS

1) Classical LAA model: Classically, the MAC access procedure of WiFi-like systems has been modeled using discrete time Markov chain, as in [5], and is illustrated in Figure 1. We denote by \( p_c \) the probability of collision seen by one station in one slot, in other words \( p_c \) is the probability that at least one
of the other \( N - 1 \) stations is transmitting during the current time slot. LAA stations are considered to be unsaturated because of the nature of URLLC traffic; we denote by \( p_0 \) the probability that the station does not have a packet to send. As the generation time between packets is usually larger than the delay constraint, we allow ourselves to neglect the queuing delay and consider that the packet delay starts once it enters the backoff procedure.

Figure 1: Classical LAA model

Each state of the 2-dimensional Markov chain is a composed of two stochastic processes \( \{s(t), b(t)\} \), representing the backoff stage and BT at time \( t \), respectively. The above model is useful for calculating several performance metrics, but does not quantify the time spent by the packet in the chain until it reaches one of the absorbing states: Success or Failure. For URLLC applications, the most important measure is the delay of the packet; therefore, we propose next some modifications on the previous model that allow us to calculate the packet delay in the chain and hence the system.

2) LAA model with delay constraints: Since URLLC packets carry critical information, delivering them after a certain time will be pointless, thus the packet must be dropped if it exceeds a given delay budget, denoted by \( T \). We assume that all packets are of same length, that no collision avoidance mechanism is deployed (such as RTS/CTS in WiFi that add delay to the system), and that transmission and collision durations are identical and constant \([7]\), denoted by \( x \). \( x \) includes the time until the reception of Ack/Nack and since our system is slotted, both \( T \) and \( x \) are expressed in time slot unit.

In this model, we propose to add a third stochastic process, \( d(t) \), representing the packet delay at time \( t \), obtaining thus a 3-dimensional discrete-time Markov chain, whose states are represented by the triplet \( \{s(t), b(t), d(t)\} \). It is quite complex to illustrate the corresponding graph of the model even for a small number of states, so we illustrate in Figure 2 the first backoff stage of the system.

Figure 2: Example of LAA model with delay constraints for \( T = 3x \) and \( W_0 = 4 \)

The process advances in the horizontal direction similar to classical model, but instead of looping when the medium is sensed busy, the process advances in the new vertical direction, increasing only the value of delay by transmission duration \( x \). The process continues advancing until either the packet is dropped because the delay budget is reached or for transmitting the packet when \( BT = 0 \). If the packet transmission fails while still having delay budget, the process can move to a new backoff stage (represented by "Next stage" in figure 2) following the same process as before.

We distinguish in particular between the states: "Start", "Success" and "Failure", to compute the different probabilities for each, but in fact the sum of states "Success" and "Failure" is equivalent to the state "Start" and the macroscopic view of the states is shown in Figure 3, where MC refers to the set of states with indices \( \{i, j, k\} \) in the Markov chain.

Figure 3: Macroscopic view of LAA model

3) Approximate model: The existence of two possible increments of the delay value: 1 when the medium is idle and \( x \) when the medium is busy, complicates the problem at hand because this generates a huge number of states, making the solution prohibitive. If we neglect the fact that the delay increases by 1 if the medium is sensed idle, then we obtain a much easier system to solve, with smaller number of states. We will propose in the next subsection a simple method to
compensate this approximation.

As the delay increments are now homogeneous (of the same order \( x \)), we can transpose the delay values into multiples of \( x \), and obtain new delay values which are incremented by one every time the medium is sensed busy. The new approximate model is shown in Figure 4.

We denote by \( A_{i,j,k} \) the probability of arriving to state \( \{i, j, k\} \), the equations that solve the system are given in a recursive way as follows:

\[
A_{i,W_0-1,0} = \frac{1}{W_0} A_{prev1,0} \leq i \leq n \\
A_{i,j,0} = (1 - p_c)A_{i,j+1,0} + \frac{1}{W_0} A_{prev1,0} \leq j \leq W_0 - 2 \\
A_{i,W_0-1,k} = p_c A_{i,W_0-1,k-1} + \frac{1}{W_0} A_{prev2,1} \leq k \leq n \\
A_{i,j,k} = (1 - p_c)A_{i,j+1,k} + p_c \frac{1}{W_0} A_{prev2,1} \leq j \leq W_0 - 2, 1 \leq k \leq n \\
A_{i,0,k} = (1 - p_c)A_{i,1,k} + p_c \frac{1}{W_0} A_{prev2,1} \leq k \leq n \\
A_{Success} = (1 - p_c) \sum_{i=0}^{n} \sum_{k=0}^{n} A_{i,0,k} \\
A_{Failure} = p_c \sum_{i=0}^{n} \sum_{j=0}^{W_0-1} \sum_{k=0}^{n} A_{i,j,n}
\]

where \( A_{prev1} = A_{Start} \) for \( i = 0 \), else \( A_{prev1} = p_c A_{i-1,0,0} \), and \( A_{prev2} = 0 \) for \( i = 0 \), or else \( A_{prev2} = A_{i-1,0,k-1} \).

One can verify that \( A_{Failure} + A_{Success} = A_{Start} \) as mentioned above.

For the calculation of \( p_c \), we notice that all of the above equations can be written in terms of \( A_{Start} \). \( A_{Start} \) is calculated by applying the normalization condition:

\[
A_{Start} + \sum_{i=0}^{n} \sum_{j=0}^{w_i-1} \sum_{k=0}^{n} b_{i,j,k} = 1
\]

Then it is possible to use the fixed point analysis where the following equations are numerically solved:

\[
\tau = \sum_{i=0}^{n} \sum_{k=0}^{n-1} A_{i,0,k} \\
p_c = 1 - (1 - (1 - p_0)\tau)^{n-1}
\]

After finding the value of \( p_c \), the arrival probability to each state is then easily calculated by substituting \( A_{Start} = 1 \), since the probabilities are computed for a packet that already entered the process.

Eventually, we obtain \( A_{Success} \) which quantifies the reliability of the system.

4) Approximation Error Compensation: A simple method to compensate the neglected time slots in the approximate model above (for the case when the medium is idle) is by adding a certain value to the delay after every backoff stage, the additional value depends on \( W_0 \), and will be discussed later in section IV.

This compensation will not change the analysis of our approximate model, we just have to calculate \( n \) for every stage, and the total number of stages will be determined dynamically depending on the remaining time budget after every stage.

Figure 4: Approximate model of LAA with delay constraints
A. MAC access at licensed 5G

We consider a grant-free system that deploys a simple retransmission method, where \( \delta \) replicas of the packet are transmitted over \( \delta \) consecutive TTIs. This system is convenient for URLLC because it does not wait for the acknowledgment after each packet; on the other hand, reliability is enhanced by retransmissions. The system is characterized by a fixed TTI of length \( \tau \), and a delay constraint \( T \), hence the number of possible retransmissions is \( \delta = T/\tau \).

In general, if a packet arrives during a TTI, then its transmission is deferred until the beginning of next TTI. This adds at most one TTI to the delay. We define the transmission unit (TU) as the combination of frequency resources of bandwidth \( w \) and time resources of one TTI that fit one URLLC packet. Assuming that the available licensed bandwidth is \( BW \), then the number of available TUs in one TTI is \( K = \lfloor BW/w \rfloor \).

The transmission procedure is that when a packet arrives to the system, its TU is randomly chosen among \( K \) available ones following a uniform distribution, as well as the TUs of the replicas. A collision happens if two or more packets are transmitted on the same TU, and the packet is lost if and only if all its replicas are in collision. To calculate the probability of packet loss, we designate a random packet being transmitted as the packet of interest (PoI). Denoting \( N \) as the number of possible retransmissions is \( \delta = T/\tau \).

Assuming that the available licensed bandwidth is \( BW \), then the number of available TUs in one TTI is \( K = \lfloor BW/w \rfloor \).

The transmission procedure is that when a packet arrives to the system, its TU is randomly chosen among \( K \) available ones following a uniform distribution, as well as the TUs of the replicas. A collision happens if two or more packets are transmitted on the same TU, and the packet is lost if and only if all its replicas are in collision. To calculate the probability of packet loss, we designate a random packet being transmitted as the packet of interest (PoI). Denoting \( A_n \) the event that \( n \) packets (other than PoI) arrive during \( \delta \) TTIs and \( C \) the event that there is another transmission(s) on the same TU as PoI, the probability that PoI transmits in a given TU during one TTI is \( 1/K \), and the probability that another packet is transmitted in a different TU is \( (1−1/K) \). Hence the probability of collision during \( \delta \) TTIs knowing the number of transmissions is:

\[
P(C|A_n) = (1 − (1 − \frac{1}{K})^n)^\delta
\]

We have that:

\[
P(A_n) = C_{N-1}^n P_a^n (1 − P_a)^{N−1−n}
\]

where \( P_a \) is the probability of having a packet during \( \delta \) TTIs.

The probability of loss of PoI is then:

\[
P_{\text{loss}} = P(C) = \sum_{n=1}^{N−1} P(C|A_n)P(A_n)
\]

\[
P_{\text{loss}} = \sum_{n=1}^{N−1} C_{N-1}^n P_a^n (1 − P_a)^{N−1−n} (1 − (1 − \frac{1}{K})^n)^\delta
\]

It is important to notice the difference of time units between licensed and unlicensed systems, the aforementioned packet arrival probability for one user in the unlicensed system \( P_{aU} = (1−p_0) \) is indicated in one time slot \( T_s \) unit, while the packet arrival probability for one user in licensed system \( P_{aL} \) must be given for one TTI. Usually, \( T_s \) is smaller than TTI, we suggest that \( TTI = zT_s \) and \( z \) is a natural number.

According to section II, packets arrive either directly after they are generated (duplication case), then the probability of packet arrival to the licensed system during one \( T_s \) is \( P_{aL} = P_{aU} \), or after going through the unlicensed system \( P_{aL} = P_{aU} P_{\text{Failure}} \). Supposing that the station generates one packet at most during one TTI, then packet arrival probability during \( \delta \) TTIs is: \( P_a = 1 − (1 − P_a)^{\delta z} \).

IV. Numerical applications

We now evaluate the system performance for users with latency requirement of 1ms and transmission reliability of 99.999%, which can be expressed equivalently with probability of loss of \( 10^{-5} \).

We assume that all users are identical and each generates a packet with a fixed probability \( 1−p_0 \) per \( T_s \), where \( p_0 = 0.999 \) is the probability of not having a packet per \( T_s \), which is equivalent to packet generation rate of 100 packets per second per station.

A. LAA model validation

We assume that LAA system is operating on the 5GHz unlicensed band, with a bit rate of 100Mbps, similar to WiFi systems operating on the same band. The time slot duration \( T_s = 9\mu s \). The size of URLLC packet is equal to 100bits, but similar to 802.11 systems, we should calculate the duration of packet transmission until the reception of ACK/NACK, considering all the headers, backoff guard durations and acknowledgment reception. Converting all quantities into \( T_s \) unit, we get \( x = 7T_s \). No RTS/CTS mechanism is considered. The system deploys cat3-LBT with a fixed contention window size of 16.

We compare the simulation results with the analytical ones obtained from the approximate LAA model when no approximation compensation is made and when considering \( W_0/2 \) and \( W_0 \) compensation values. The results are shown in figure 5.

Figure 5: Probability of loss vs number of stations for LAA system

We observe that the model is most accurate when a whole contention window size \( W_0 \) is added after every stage; and not \( W_0/2 \) as one could predict as this is an average value. In fact,
this is because we are not computing average quantities, but exact probabilities. Hence, the approximate LAA model must always be considered with approximation compensation of $W_0$. We can also see from figure 5 that LAA system capacity is able to handle about 75 stations (reliability equal to $1 - 10^{-5}$).

B. Licensed 5G model validation

Due to the various types of traffic to be transported in 5G, and most importantly their delay constraints, 5G system can choose the value of TTI from a number of possible values according to the requirements, contrary to LTE. Clearly, choosing a smaller TTI necessitates larger bandwidth for the same transmission.

We assume that a URLLC packet fits in one conventional LTE RB of 0.5ms and 180KHz bandwidth (12 subcarriers with carrier spacing of 15 KHz), having that TTI length $\tau = 0.125ms$, the bandwidth of one TU is then $w = 720KHz$. When combining the unlicensed/licensed systems, we would like to calculate the minimum bandwidth that satisfies the required reliability and delay, and what number of retransmissions is most appropriate to use. To validate the proposed analytical model of 5G licensed, we implement a discrete-time simulator where $N$ stations generate packets at random at each slot and follow the MAC mechanism explained above (grant-free with replicas). Simulation and analytical results are compared in figure 6 which shows a good match between these two curves.

![Figure 6: Probability of loss vs number of stations for 5G licensed system](image)

C. Joint transmission through LAA and 5G licensed

To explore the methods of unlicensed/licensed systems combination, we choose a number of stations that exceeds the unlicensed system capacity, say $N = 100$, to evaluate the system performance. Our aim is to identify the optimum combination of number of retransmissions $\delta$ and number of time units $K$ that minimizes $K$ and satisfies the system requirements. Figure 7 shows the resource requirements $K$ for the different combinations.

![Figure 7: Probability of loss vs number of TUs for joint unlicensed/licensed system](image)

From figure 7, it is obvious that the duplication method where each packet is sent in parallel on the two systems (corresponding to the "Licensed" curve) is not the best way to combine the two systems as it does not reach a $1 - 10^{-5}$ reliability. We notice that the reliability-delay trade-off is obtained for values of $K$ starting from 10 and $\delta$ equal to 3, 4 and 5; 4 being the optimal in our case.

$K = 10$ TUs with $\delta = 4$ correspond actually to a bandwidth of 7.2MHz. Adding this frequency resource to the unlicensed system, the total system capacity increases to 100 stations (versus 75 stations in case of unlicensed only, as seen above).

V. CONCLUSION

We studied in this paper the transport of URLLC traffic in the uplink using unlicensed spectrum. This type of traffic imposes strict reliability and delay constraints which may not be met by the unlicensed spectrum on its own. In this case, we propose the joint use of 5G licensed spectrum so as to transport this type of traffic.

We modeled the MAC access of the resulting system incorporating a timer corresponding to the delay budget for URLLC traffic and proposed an approximate method which solves the system equations efficiently in the presence of the two delay timers: that of URLLC traffic and that of the system backoff procedure.

We validated our models numerically and showed the performance of the joint licensed/unlicensed system and quantified the quantity of 5G resources to be added to meet the reliability and delay requirements.

Our next step is on the modeling of more advanced resource allocation schemes on 5G, including the transmission of several replicas of the same packet in order to increase reliability. We will also investigate the impact of interference.
in less confined environments, for example originating from WiFi on the performance.

REFERENCES