Grant-Free Coexistence of Critical and Non-Critical IoT Services in Two-Hop Satellite and Terrestrial Networks

Rahif Kassab*, Andrea Munari†, Federico Clazzer†, and Osvaldo Simeone*

*King's Communications, Learning and Information Processing (KCLIP) Lab, King's College London, UK

†Institute of Communications and Navigation, German Aerospace Center (DLR), 82234 Wessling, Germany

Emails: *{rahif.kassab,osvaldo.simeone}@kcl.ac.uk , †{Andrea.Munari,Federico.Clazzer}@dlr.de

Abstract

Terrestrial and satellite communication networks often rely on two-hop wireless architectures with an access channel followed by backhaul links. Examples include Cloud-Radio Access Networks (C-RAN) and Low-Earth Orbit (LEO) satellite systems. Furthermore, communication services characterized by the coexistence of heterogeneous requirements are emerging as key use cases. This paper studies the performance of critical service (CS) and non-critical service (NCS) for Internet of Things (IoT) systems sharing a grant-free channel consisting of radio access and backhaul segments. On the radio access segment, IoT devices send packets to a set of non-cooperative access points (APs) using slotted ALOHA (SA). The APs then forward correctly received messages to a base station over a shared wireless backhaul segment adopting SA. We study first a simplified erasure channel model, which is well suited for satellite applications. Then, in order to account for terrestrial scenarios, the impact of fading is considered. Among the main conclusions, we show that orthogonal inter-service resource allocation is generally preferred for NCS devices, while non-orthogonal protocols can improve the throughput and packet success rate of CS devices for both terrestrial and satellite scenarios.

Index Terms

Beyond 5G, IoT, Grant-Free, satellite networks, mMTC, URLLC

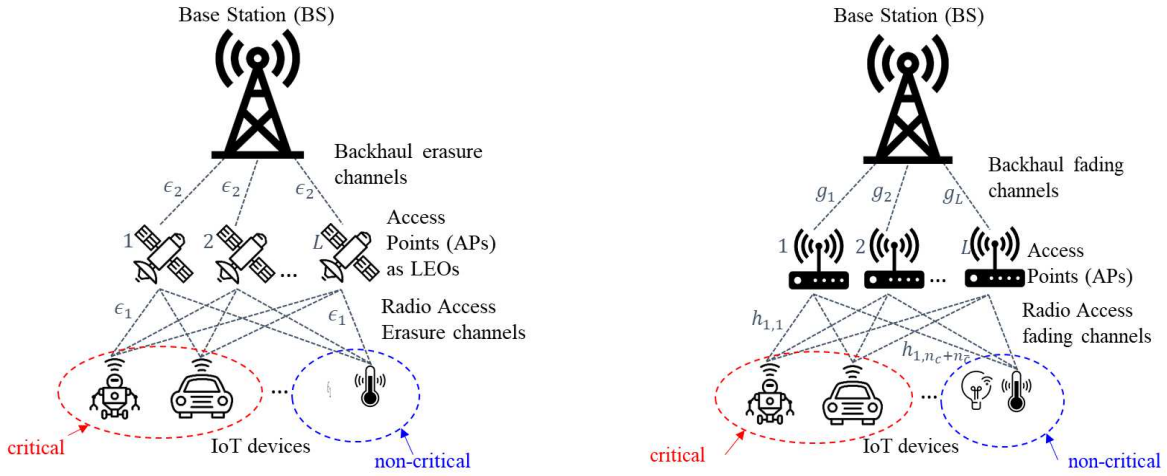
The work of Rahif Kassab and Osvaldo Simeone has received funding from the European Research Council (ERC) under the European Union Horizon 2020 research and innovation program (grant agreement 725731).

I. Introduction

Future generations of cellular and satellite networks will include new services with vastly different performance requirements. In recent 3GPP releases [1], a distinction is made among Ultra-Reliable and Low-Latency Communications (URLLC), with stringent delays and packet success rate requirements; enhanced Mobile Broadband (eMBB) for high throughput; and massive Machine Type Communications (mMTC) for sporadic transmissions with large spatial densities of devices [2], [3]. In this paper, we focus on Internet of Things (IoT) scenarios, which are typically assumed to fall into the mMTC service category [4]. We take a further step compared to the mentioned 3GPP classification by considering a beyond-5G scenario characterized by the coexistence of heterogeneous IoT devices having critical and non-critical service requirements. Devices with critical service (CS) requirements must be provided more stringent throughput and packet success rate performance guarantees than non-critical service (NCS) devices. We note that CS for IoT will be introduced in 3GPP release 17 [5] under the name *enhanced Industrial IoT*, while NCS IoT is a typical use case for *New Radio-light*, which will also be studied in release 17 [5].

As illustrated in Fig. 1, we consider a two-hop wireless architecture with radio access channels followed by backhaul links. In these topologies, space diversity is provided by multiple Access Points (APs) that play the role of relays between the devices and the Base Station (BS). For terrestrial networks, C-RAN can be described as two-hop networks [6] while for satellite networks, the model applies to communications through Low-Earth Orbit (LEO) mega-constellations, e.g., Amazon Kuiper [7] and SpaceX Starlink [8] projects. A new 3GPP work item for IoT over non-terrestrial networks (NTN) was recently introduced [5]. With thousands of LEO satellites, these constellations will offer connectivity to each earth location with multiple satellites at a time. Satellite based IoT deployments can hence extend connectivity to remote areas with low or no cellular coverage, enabling applications in many industries, such as maritime and road transportation, farming and mining.

In the presence of a large number of IoT devices requiring the transmission of small amounts of data, conventional *grant-based* radio access protocols can cause a significant overhead on the access network due to the large number of handshakes required. A potentially more efficient solution is given by *grant-free* radio access protocols where devices transmit whenever they have a packet to deliver without any prior handshake [9]–[11]. This is typically done via some



(a) (b)

Fig. 1: An IoT system with grant-free wireless radio access and shared backhaul with uncoordinated APs, in which IoT devices generate CS or NCS messages. The setup in (a) illustrates a satellite communications scenario with binary erasure channels modeling the presence/absence of a line-of-sight link. The setup in (b) illustrates a terrestrial communications scenario with fading channels.

variations of the classical ALOHA random access scheme [12]. Grant-free access protocols are used by many commercial solutions in the terrestrial domain, e.g., by Sigfox [13] and LoRaWAN [14]; as well as in the satellite domain, using constellations of low-earth orbit satellites (LEO), e.g. Orbcomm [15] and Myriota [16].

In classical cellular IoT scenarios, *orthogonal inter-service resource allocation* schemes are typically used [17]. However, due to their static nature, orthogonal schemes may cause an inefficient use of resources under unpredictable traffic patterns in grant-free IoT systems. To obviate this problem, *non-orthogonal resource allocation*, which allows access of multiple devices to the same time-frequency resource, presents a promising alternative solution [18]–[21]. Recent work has proposed to apply non-orthogonal resource allocation to heterogeneous services [22]–[24]. In order to mitigate the impact of interference in non-orthogonal schemes, one can leverage successive interference cancellation (SIC) [25], time diversity [26], and/or space diversity [27] [28].

Main Contributions: In this work, we aim to answer the following questions: *What is the*

impact of two-hop topologies on the grant-free coexistence of CS and NCS in IoT systems? Is there any advantage in using non-orthogonal access techniques across the two services? In order to do this, we study grant-free access for CS and NCS in space diversity-based models for both satellite and terrestrial applications. We analytically derive throughput and packet success rate measures for both CS and NCS as a function of key parameters such as the number of APs and frame size. The analysis, which generalizes conventional models used for slotted ALOHA (SA), accounts for orthogonal and non-orthogonal inter-service access schemes, as well as for binary erasure channels modeling NTN (Fig. 1(a)) and for fading channels modeling terrestrial networks (Fig. 1(b)). Finally, two receiver models are considered, namely, a collision model, where packets transmitted by the same device are assumed to undergo destructive collision, and a superposition model, where packets transmitted from the same device are superposed at the receiver.

Preliminary results for our model were presented in [29] and [30]. The main differences with our two previous works are as follows:

- Reference [29] considers a single-service set-up with the erasure channel model and a single and simpler decoder model. In contrast, this work considers the coexistence of two services with both erasure and fading channel models in addition to a more practical decoder model.
- Paper [30] studies CS and NCS model under erasure channels assuming that the decoder cannot recover a packet if multiple instances of the same packet are received. In contrast, this work considers fading channel model in addition to a more practical decoder model.

The rest of the paper is organized as follows. In Sec. III we describe the system model used and the performance metrics. In Sec. IV we study the system in the presence of a single service while Sec. V and VI tackle the heterogeneous services case under the general collision and superposition model respectively. Finally, the heterogeneous service case is evaluated under the fading channel model in Sec. VII, conclusions and extensions are discussed in Sec. VIII.

Notation: Throughout our discussion, we denote as $X \sim \mathrm{Bin}(n,p)$ a Binomial random variable (RV) with n trials and probability of success p; as $X \sim \mathrm{Poiss}(\lambda)$ a Poisson RV with parameter λ. We also write $(X,Y) \sim f \cdot g$ for two independent RVs X and Y with respective probability density functions f and g.

II. Related Works

Satellite networks, through constellations of LEO satellites, have been increasingly receiving interest from academia and industry as a potential technology to extend connectivity to rural and remote areas and thus enabling new applications in fields like maritime, transportation, farming and mining where internet connectivity is a big challenge. A new 3GPP work item for IoT over non-terrestrial networks (NTN) was recently introduced [5]. Academic research in the wireless community has been focusing on the integration of satellite networks in next generation cellular networks. For example, [31] considers the problem dynamic resource allocations for LEO satellites while taking into consideration the power consumption and mobility management constraints. [32] analyzes IoT traffic in LEO satellites networks while [18], [33], [34] consider different architectures integrated with 5G and beyond cellular networks while highlighting the advantages of such architectures which include space and access diversity. Similarly to previous works, this work considers a two-hop architecture to model a satellite system, however, in contrast to previous works, we derive the throughput expressions with coexisting critical and non-critical IoT services.

Non-orthogonal multiple access is a promising technology to provide massive connectivity by allowing for multiple users to transmit over the same radio resources in the uplink while using superposition coding in the downlink. For the uplink, reference [35] shows that NOMA with Successive Interference Cancellation (SIC) at the base stations can significantly enhance cell-edge users' throughput. As for the downlink, NOMA was demonstrated to achieve superior performance in terms of ergodic sum rate of a cellular network with randomly deployed users in [21]. NOMA can leverage different technologies to differentiate the users, e.g., in the code or power domains. We refer to [36] and references therein for a detailed discussion.

Although NOMA can provide massive connectivity, a related challenge is the access to radio resources. Grant-based access, which is currently used in LTE/LTE-A [37], is not suited for IoT networks characterized by high density of devices and sporadic traffic of small data units. Indeed, grant-based access entails an increased protocol overhead, inducing latency and possibly network overload. By eliminating the handshake procedure, *grant-free random access* was proposed as a potential solution for these problems [38]. Many works have proposed to combine grant-free based access and NOMA. For example, power-based grant-free NOMA was proposed in [25] using slotted ALOHA, and code-based NOMA was presented in [39] using sparse codes. We

refer to [40] for a comprehensive survey on grant-free NOMA techniques. In contrast to previous works, this work aims to study the benefits of grant-free NOMA in the context of satellites and terrestrial two-hop networks with coexisting IoT services.

III. System model and Performance metrics

A. System Model

We first consider the system illustrated in Fig. 1, in which L APs, e.g., LEO satellites, provide connectivity to IoT devices. The APs are in turn connected to a BS, e.g., a ground station, through a shared wireless backhaul channel. The main motivation to use the two-hop architecture in Fig. 1 is to benefit from enhanced coverage and space diversity, owing to the presence of multiple APs in the vicinity of each IoT device. We assume that time over both access and backhaul channels is divided into frames and each frame contains T time slots. At the beginning of each frame, a random number of IoT devices are active. The number of active IoT devices that generate CS and NCS messages at the beginning of the frame follow independent Poisson distributions with average loads $\gamma_c G$ and $(1-\gamma_c)G$ $[\text{packet/frame}]$, respectively, for some parameter $\gamma_c \in [0,1]$ and total load G. Users select a time-slot uniformly at random among the T time-slots in the frame and independently from each other. By the Poisson thinning property [41], the random number $N_c(t)$ of CS messages transmitted in a time-slot t follows a Poisson distribution with average $G_c = \gamma_c G/T$ $[\text{packet/slot}]$, while the random number $N_{\bar{c}}(t)$ of NCS messages transmitted in slot t follows a Poisson distribution with average $G_{\bar{c}} = (1-\gamma_c)G/T$ $[\text{packet/slot}]$.

Radio Access Model: As in, e.g., [29], [42], [43], we model the access links between any device and an AP as an independent interfering erasure channel with erasure probability ϵ_1. In satellite applications, as represented in Fig. 1a, this captures the presence or absence of a line-of-sight link between the transmitter and the receiver. A packet sent by a user is independently erased at each receiver with probability ϵ_1, causing no interference, or is received with full power with probability $1-\epsilon_1$. The erasure channels are independent and identically distributed (i.i.d.) across all slots and frames. Interference from messages of the same type received at an AP is assumed to cause a destructive collision. Furthermore, CS messages are assumed to be transmitted with a higher power than NCS messages so as to improve their packet success rate, hence creating significant interference on NCS messages. As a result, in each time-slot, an AP can be in three possible states:

- a CS message is retrieved successfully if the AP receives only one (non-erased) CS message and no more than a number K of (non-erased) NCS messages. This implies that, due to their lower transmission power, NCS messages generate a tolerable level of interference on CS messages as long as their number does not exceed the threshold K;
- a NCS message is retrieved successfully if the AP receives only one (non-erased) NCS message;
- no message is retrieved otherwise.

We note that packet re-transmissions are ignored due to latency and low power constraints.

Backhaul model: The APs share a wireless out-of-band backhaul that operates in a full-duplex mode and in an uncoordinated fashion as in [29]. The lack of coordination among APs can be considered as a worst-case scenario in dense low-cost terrestrial cellular deployments [44] [28] and as the standard solution for constellations of LEO satellites that act as relays between ground terminals and a central ground station. In fact, satellite coordination, although feasible through the use of inter-satellite links [45], may be costly in terms of on-board resources. In each time-slot $t+1$, an AP sends a message retrieved on the radio access channel in the corresponding slot t over the backhaul channel to the BS. APs with no message retrieved in slot t remain silent in the corresponding backhaul slot $t+1$. The link between each AP and the BS is modeled as an erasure channel with erasure probability ϵ_2, and destructive collisions occur at the BS if two or more messages of the same type are received. As for the radio access case, erasure channels are i.i.d. across APs, slots and frames. We note that different forwarding strategies and buffering can be modelled by defining a forwarding probability at each AP. This can easily be accounted for in our model by simply multiplying the probability of successfully receiving a packet at each AP by the forwarding probability. We refer to [46] for details. Due to the sporadicity and unpredictability of IoT traffic, grant-based access is problematic as it induces additional latency and under-utilization of radio resources. In this paper, we assume that radio and backhaul access is carried out using grant-free non-orthogonal slotted ALOHA [37]. Furthermore, We assume that, over both the access and backhaul channels, a simple power-based non-orthogonal multiple access scheme is used whereby the packet received with the strongest power can be decoded. For practical considerations, we do not implement successive interference cancellation by exploiting the coherent sum of different message copies.

In order to model interference between APs, we consider two scenarios. The first, referred to as *collision model*, assumes that multiple messages from the same or distinct device cause

destructive collision. Under this model, in each time-slot, the BS's receiver can be in three possible states:

- as for radio access, a CS message is retrieved successfully at the BS if only one CS message is received from any AP, along with no more than K NCS messages;
- a NCS message is retrieved successfully if no other CS or NCS message is received;
- no message is retrieved at the BS otherwise.

In the second model, referred to as *superposition model*, the BS is able to decode from the superposition of multiple instances of the same packet that are relayed by different APs on the same backhaul slot, assuming no collisions from other transmissions. The superposition model is motivated by the fact that, when multiple copies of the same message are received by a receiver over the same radio resource, they undergo superposition over the wireless channel. This results in an effective channel given by the sum of the channels affecting each copy of the message. Therefore, the message can be decoded without the need for interference cancellation. In practice, this can be accomplished by ensuring that the time asynchronism between APs is no larger that the cyclic prefix in a multicarrier modulation implementation. Synchronization can be ensured, for example, by having a central master clock at the BS against which the local time bases of APs are synchronized [47]. Overall, the BS's receiver can be in three possible states:

- a CS message is retrieved successfully at the BS in a given time-slot if no other CS message and no more than K NCS messages are received by the BS;
- a NCS message is retrieved successfully if no CS messages and no other NCS messages are received in the same slot;
- no message is retrieved at the BS otherwise.

Note that the main difference between the collision and superposition models is that, under the collision model, a packet can not be recovered when multiple instances of the same packet are received at the BS, while this same packet can be recovered under the superposition model.

Inter-service TDMA: In addition to non-orthogonal resource allocation whereby devices from both services share the entire frame of T time-slots, we also consider orthogonal resource allocation, namely *inter-service time division multiple access* (TDMA), whereby a fraction αT of the frame's time-slots are reserved to CS devices and the remaining $(1-\alpha)T$ for NCS devices. Inter-service contention in each allocated fraction follows a SA protocol as discussed above. In the following, we derive the performance metrics under the more general non-orthogonal scheme

described above. The performance metrics under TDMA for each service can be directly obtained by replacing T with the corresponding fraction of resources in the performance metrics equations and setting the interference from the other service to zero.

B. Performance Metrics

We are interested in computing the throughput R_c and $R_{\bar{c}}$ $[\mathrm{packet/slot}]$ and the packet success rate Γ_c and $\Gamma_{\bar{c}}$ $[\mathrm{packet/frame}]$ for CS and NCS respectively. The throughput is defined as the average number of packets received correctly in any given time-slot at the BS for each type of service. The packet success rate is defined as the average probability of successful transmission of a given user given that the user is active, i.e., that it transmits a packet in a given frame.

C. Considerations on the System Model Assumptions

Before moving to the analysis, some remarks on the considered system model are in order. First, the binary erasure collision channel model with i.i.d. erasures-also referred to as on-off fading model in the literature [48]-entails some simplifications that need to be discussed. When the aggregate channel traffic is high, in a realistic scenario, the aggregate interference power, even if very small on a single packet level, may become relevant and hinder the correct reception of a data unit when many concurrent packets are transmitted. This particular aspect is not captured by the considered channel model, which only models independent erasures. However, the relevance of this scenario remains limited since high channel load conditions in an SA-based system yield unacceptable packet success rates [49]. From this standpoint, the model we consider captures well the behaviour of practical systems in the more interesting low-to-moderate channel load conditions.

Thirdly, it is important to note that our model accounts for various types of collisions among packets, namely, inter-CS and inter-NCS collisions, CS to NCA collisions and finally NCS to CS collisions.

Secondly, assuming independent erasures across all links provides an upper bound on the achievable performance in more realistic configurations in which fading events can render the link towards multiple relays correlated [50]. In general, this bound is expected to provide a good approximation for LEO constellation links towards different satellites due to the spatial separation of the channels. In contrast, this assumption may be an optimistic estimate of the performance for terrestrial systems with denser deployments.

Finally, the use of uncoordinated access in the APs-to-BS links should be considered against the use of coordinated-access mechanisms. While the latter can ideally provide higher throughputs, it may also become inefficient when the traffic activity seen at the APs varies heavily, is difficult to predict, or when APs move as in a LEO constellation. In such cases, the overhead required to coordinate the access among relays overcomes the increase in data delivery and thus uncoordinated access may become preferable.

IV. Single Service Under Collision Model

A. Performance Analysis

We start by considering the baseline case of a single service under the collision model. While this can account for either CS or NCS, we consider here without loss of generality only the CS by setting $\gamma_c = 1$. We note that this model is similar to a multiple-relay SA often referred to as modern random access protocol [29]. An AP successfully retrieves a packet when *only one* of the $N_c = n_c$ transmitted packets arrives unerased, i.e. with probability

$$p_{n_c} = n_c \, (1 - \epsilon_1) \, \epsilon_1^{n_c - 1}, \tag{1}$$

where the term $\epsilon_1^{n_c-1}$ is the probability that the remaining $n_c - 1$ packets are erased. Removing the conditioning on N_c, one can obtain the average radio access throughput as

$$\mathbb{E}_{N_c}[p_{n_c}] = \sum_{n_c=0}^{\infty} \frac{G_c^{n_c} \, e^{-G_c}}{n_c!} \cdot p_{n_c} = G_c(1 - \epsilon_1) \, e^{-G_c(1-\epsilon_1)}, \tag{2}$$

which corresponds to the throughput of a SA link with erasures.

The overall throughput R_c depends also on the backhaul channel. In particular, for a successful packet transmission, an AP must successfully decode one packet, which should then reach the BS unerased over the backhaul channel. This occurs with probability

$$q_{n_c} = p_{n_c} \, (1 - \epsilon_2). \tag{3}$$

In addition, the packet should not collide with other packets. By virtue of the independence of erasure events, the number of incoming packets on the backhaul during a slot follows the binomial distribution $\mathrm{Bin}(L, q_{n_c})$. Recalling that collisions are regarded as destructive under the collision model, a packet is retrieved only when a single packet reaches the BS, i.e. with probability $q_c = L q_{n_c} (1 - q_{n_c})^{L-1}$. The CS throughput can then be derived as

$$R_c = \mathbb{E}_{N_c}[q_c] = \sum_{n_c=0}^{\infty} \frac{G_c^{n_c} e^{-G_c}}{n_c!} \cdot L \, q_{n_c} (1 - q_{n_c})^{L-1}. \tag{4}$$

This can be computed in closed form as stated in the following proposition.

Proposition 1: *Under the collision model, assuming $\gamma_c = 1$ (single service), the throughput R_c is given as function of the number of APs L, channel erasure probabilities ϵ_1 and ϵ_2, and CS packet load G_c as*

$$R_c = \sum_{\ell=0}^{L-1} (-1)^{\ell} \, L \binom{L-1}{\ell} \left[\frac{(1-\epsilon_1)(1-\epsilon_2)}{\epsilon_1} \right]^{\ell+1} \cdot e^{-G_c} \cdot \mathcal{H}_{\ell+1}\big(G_c \, \epsilon_1^{\ell+1}\big) \,, \tag{5}$$

where the auxiliary function $\mathcal{H}_m(x)$ is defined recursively as

$$\begin{aligned} &\mathcal{H}_0(x) = e^x \\ \text{and } &\mathcal{H}_m(x) = x \sum_{\ell=0}^{m-1} \binom{m-1}{\ell} \mathcal{H}_\ell(x) \quad m \geq 1. \end{aligned} \tag{6}$$

Proof: Denoting $\beta = (1-\epsilon_1)(1-\epsilon_2)$, and recalling the definitions of probabilities p_{n_c} and q_{n_c} in equations (1) and (3), the throughput (4) can be written as

$$\begin{aligned} R_c &= \sum_{n_c=0}^{\infty} \frac{G_c^{n_c} e^{-G_c}}{n_c!} \cdot L \, \beta \, n_c \epsilon_1^{n_c-1} \left(1 - \beta \, n_c \epsilon_1^{n_c-1}\right)^{L-1} \\ &\overset{(a)}{=} \sum_{i=0}^{L-1} (-1)^i \, L \binom{L-1}{i} \frac{\beta^{i+1} \, e^{-G_c}}{\epsilon_1^{i+1}} \\ &\qquad \cdot \sum_{n_c=0}^{\infty} \frac{\left(G_c \, \epsilon_1^{i+1}\right)^{n_c}}{n_c!} \cdot {n_c}^{i+1}, \end{aligned} \tag{7}$$

where (a) follows by applying Newton's binomial expansion and after some simple yet tedious rearrangements. Let us now introduce the auxiliary function

$$\mathcal{H}_m(x) = \sum_{n_c=0}^{\infty} \frac{x^{n_c} \, {n_c}^m}{n_c!}. \tag{8}$$

From the definition of Taylor's series for the exponential function, we have $\mathcal{H}_0(x) = e^x$. Moreover, for $m \geq 1$, we have

$$\mathcal{H}_m(x) = x \sum_{n_c=0}^{\infty} \frac{x^{n_c-1} \, {n_c}^{m-1}}{(n_c-1)!} \overset{(b)}{=} x \sum_{t=0}^{\infty} \frac{x^t \, (t+1)^{m-1}}{t!} \tag{9}$$

$$\overset{(c)}{=} x \sum_{\ell=0}^{m-1} \binom{m-1}{\ell} \sum_{t=0}^{\infty} \frac{x^t \, t^\ell}{t!} \tag{10}$$

$$= x \sum_{\ell=0}^{m-1} \binom{m-1}{\ell} \mathcal{H}_\ell(x), \tag{11}$$

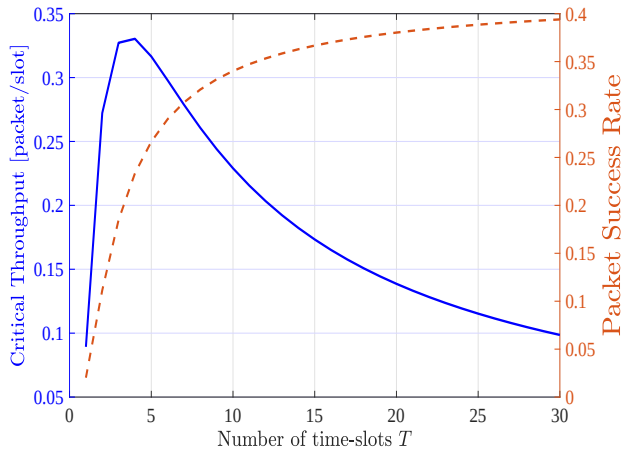


Fig. 2: Single service (here CS) throughput (solid line) and packet success rate (dashed line) as a function of the number of time-slots T ($\epsilon = 0.5$, $G = 16$ [packet/frame], $\gamma_c = 1$ and $L = 3$ APs).

where equality (b) applies the change of variable $t = n_c - 1$ and equality (c) results from applying once more Newton's binomial expansion to $(t+1)^{m-1}$. Plugging this result into the innermost summation within (7) leads to the closed form expression of the CS throughput reported in (5). □

We now turn to the packet success rate. Define the RV $N'_c \geq 1$ to count the number of transmitted messages given that at least one message is transmitted. This RV has the distribution

$$P(N'_c = n'_c | N'_c \geq 1) = (1 - e^{-G_c})^{-1} \cdot \frac{e^{-G_c} G_c^{n'_c}}{n'_c!} \tag{12}$$

which corresponds to a normalized Poisson distribution over the set $\{1, \ldots, \infty\}$. For a given value $N'_c = n'_c$, the probability that the packet of a given user u reaches an AP given that u is active is given by $p_u = (1-\epsilon_1)\epsilon_1^{n'_c - 1}$. Furthermore, the probability that the user's packet reaches the BS is given as

$$q_u = \underbrace{L p_u (1-\epsilon_2)}_{(a)} \underbrace{(1 - q_{n'_c})^{L-1}}_{(b)}, \tag{13}$$

where $p_{n'_c}$ and $q_{n'_c}$ are defined in (1) and (3). In (13), term (a) is the probability that the user's packet is received at the BS from any of the L APs, while (b) denotes the probability that the BS does not receive any CS message from the remaining $L-1$ APs. The packet success rate

Γ_c, can be obtained by averaging (13) over N_c' as $\Gamma_c = \mathbb{E}_{N_c'}[q_u]$. This can be obtained in closed form as stated in the following proposition.

Proposition 2: *Under the collision model, assuming $\gamma_c = 1$ (single service), the packet success rate Γ_c is given as function of the number of APs L, channel erasure probabilities ϵ_1 and ϵ_2, and CS packet load G_c as*

$$\Gamma_c = L\beta(1-e^{-G_c})^{-1}e^{-G_c}\sum_{l=0}^{L-1}\frac{(-\beta)^l}{\epsilon_1^{l+1}}\binom{L-1}{l} \cdot \left[\mathbb{1}_{l=0}(e^{\epsilon_1 G_c}-1)+\mathbb{1}_{l>0}\mathcal{H}_l(G_c\epsilon_1^{l+1})\right], \tag{14}$$

where the function $\mathcal{H}_l(\cdot)$ is defined in (6) *and we have $\beta = (1-\epsilon_1)(1-\epsilon_2)$.*

Proof: The proof follows using the same steps as for *Proposition 1*. □

We note that in the regime of high number of APs, i.e., $L \to \infty$, the throughput and packet success rate derived in *Proposition 1* and *Proposition 2* are equal to zero as detailed in Appendix D. This shows that the number of APs should be carefully selected. This will be further investigated in Sec. VI for heterogeneous services.

B. Examples

Using the expressions derived in *Proposition* 1 and 2, we plot in Fig. 2 the throughput and packet success rate for a single service as function of the number of time-slots T. Increasing T is seen to improve the packet success rate: an active user has a larger chance of successful transmission when more time-slots are available for random access. In contrast, there exist an optimal value of T for the throughput, as hinted by the analysis of the standard ALOHA protocol. Increasing T beyond this optimal value reduces the throughput owing to the larger number of idle time-slots. The asymptotic behaviors of packet success rate and throughput can be easily verified theoretically using the expressions in *Proposition 1* and *Proposition 2* by taking their limit when G_c tends to zero.

V. Heterogeneous Services Under Collision Model

In this section, we extend the analysis in the previous section to derive the throughput and packet success rate of both CS and NCS under the collision model described in Sec. III.

A. Heterogeneous Services with Ideal NCS-to-CS Interference Tolerance

We start by considering the case in which decoding of CS messages is not affected by NCS traffic, i.e., we set $K \to \infty$. Under this assumption, the CS throughput and packet success rate expressions equals the expressions in Propositions 1 and 2. We hence focus here on the performance of NCS, as summarized in the following proposition.

Proposition 3: *Under the collision model with ideal NCS-to-CS interference tolerance, i.e. $K \to \infty$, the NCS throughput $R_{\bar{c}}$ and packet success rate $\Gamma_{\bar{c}}$ can be respectively written as a function of the number of APs L, channel erasure probabilities ϵ_1 and ϵ_2, and CS and NCS packet loads G_c and $G_{\bar{c}}$ as*

$$R_{\bar{c}} = L \sum_{i=0}^{L-1} \sum_{k=0}^{i} (-1)^i \binom{L-1}{i} \binom{i}{k} \left(\frac{\beta}{\epsilon_1}\right)^{i+1} e^{-G} \cdot \mathcal{H}_{i-k}(G_c \epsilon_1^{i+1}) \mathcal{H}_{k+1}(G_{\bar{c}} \epsilon_1^{k+1}) \tag{15}$$

$$\text{and } \Gamma_{\bar{c}} = \sum_{l=0}^{L-1} L \beta^{l+1} \epsilon_1^{-(l+1)} (-1)^{-l} (1 - e^{-G_{\bar{c}}})^{-1} \cdot \binom{L-1}{l} \Bigg\{ \sum_{m=0}^{l} \binom{l}{m} A \cdot B + e^{-G} [\mathbb{1}_{l=0}(e^{\epsilon_1 G_{\bar{c}}} - 1) + \mathbb{1}_{l>0} \mathcal{H}_l(G_{\bar{c}} \epsilon_1^{l+1})] \Bigg\}, \tag{16}$$

where

$$A = e^{-G_c}(e^{G_c \epsilon_1^{l+1}} - 1)\mathbb{1}_{m=0} + e^{-G_c} \mathcal{H}_m(G_c \epsilon_1^{l+1}) \mathbb{1}_{m>0} \tag{17}$$

$$B = e^{-G_{\bar{c}}} \mathcal{H}_{l-m}(G_{\bar{c}} \epsilon_1^{l-m+1}) \mathbb{1}_{l-m \neq 0} + e^{-G_{\bar{c}}}(e^{\epsilon_1 G_{\bar{c}}} - 1)\mathbb{1}_{l-m=0}. \tag{18}$$

Proof: The proof is detailed in Appendix A. □

B. Examples

In order to study the performance trade-offs between the two services, we start by investigating the impact of γ_c by plotting in Fig. 3 the CS and NCS throughputs versus γ_c with $\epsilon_1 = \epsilon_2 = \epsilon$, $G = 30$ [packet/frame], $T = 4$ [time-slot/frame], and $L = 3$ APs. For CS, there is an optimal value of γ_c that ensures an optimized CS load as in the standard analysis of the ALOHA protocol, discussed also in the context of Fig. 2. In contrast, the NCS throughput decreases as function of γ_c due to the increasing interference from CS transmissions. The NCS throughput is also seen

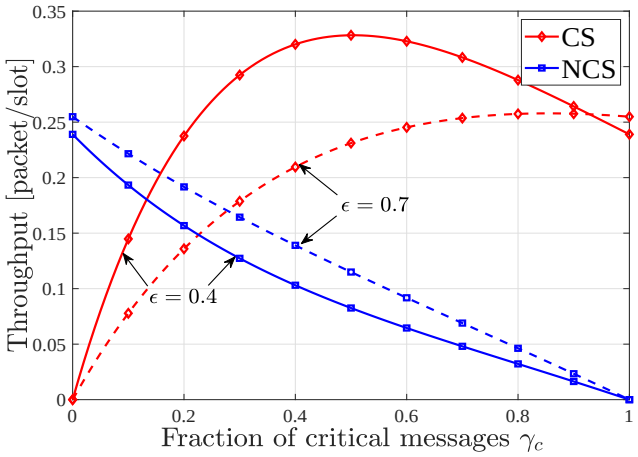


Fig. 3: CS and NCS throughput as function of the fraction of CS messages γ_c for the collision model and using non-orthogonal resource allocation ($\epsilon_1 = \epsilon_2 = \epsilon = 0.4$ or 0.7, $G = 8$ [packet/frame], $T = 4$ [time-slot/frame], and $L = 3$ APs).

to increase as a function of the channel erasure ϵ when ϵ is not too large. This is because a larger ϵ can reduce the interference from CS transmissions.

C. Heterogeneous Services with Limited NCS-to-CS Interference Tolerance

We now alleviate the assumption that CS transmissions can withstand any level of NCS interference by assuming that interference from at most K NCS transmissions can be tolerated without causing a collision from CS traffic. We derive both CS and NCS performance metrics. We note that, perhaps counter-intuitively, both CS and NCS performance metrics are affected by the CS interference tolerance parameter K. In fact, with a lower value of K, a smaller number of CS packets tends to reach the BS, reducing interference to NCS transmissions. We start by detailing the NCS performance metrics.

Proposition 4: *Under the collision model with limited NCS-to-CS interference tolerance, i.e. finite K, the NCS throughput and packet success rate are given as a function of the number of APs L, channel erasure probabilities ϵ_1 and ϵ_2, and CS and NCS packet loads G_c and $G_{\bar{c}}$ as*

$$\begin{aligned} R_{\bar{c}} =& L \sum_{i=0}^{L-1} \sum_{k=0}^{i} (-1)^i \binom{L-1}{i} \binom{i}{k} \left(\frac{\beta}{\epsilon_1}\right)^{i+1} \\ & \cdot e^{-G} \mathcal{H}_{i-k}(G_c \epsilon_1^{i+1}) \Big[\xi_1(K, L, G_{\bar{c}}, \epsilon_1) + \xi_2(K, L, G_{\bar{c}}, \epsilon_1) \Big] \end{aligned} \tag{19}$$

$$\text{and } \Gamma_{\bar{c}} = \mathbb{E}_{N_c,N'_{\bar{c}}}\Big[L(1-\epsilon_1)\epsilon_1^{N'_{\bar{c}}-1}\epsilon_1^{N_c}(1-\epsilon_2)(1-q)^{L-1}\Big], \tag{20}$$

where $q = (1-\epsilon_2)(p_{N_c}\gamma_{K-1}(N'_{\bar{c}},\epsilon_1) + N'_{\bar{c}}(1-\epsilon_1)\epsilon_1^{N'_{\bar{c}}-1}\epsilon_1^{N_c})$,

$$\xi_1(K,L,G_{\bar{c}},\epsilon_1) = \sum_{n_{\bar{c}}=0}^{K} \frac{(G_{\bar{c}}\epsilon_1^{k+1})^{n_{\bar{c}}}}{n_{\bar{c}}!} n_{\bar{c}}^{k+1} \tag{21a}$$

$$\xi_2(K,L,G_{\bar{c}},\epsilon_1) = \sum_{n_{\bar{c}}=K+1}^{+\infty} \frac{(G_{\bar{c}}\epsilon_1^{k+1})^{n_{\bar{c}}}}{n_{\bar{c}}!} \tag{21b}$$

$$\cdot\, n_{\bar{c}}^{k+1}\left[\sum_{l=0}^{K}\binom{n_{\bar{c}}}{l}(1-\epsilon_1)^l\epsilon_1^{n_{\bar{c}}-l}\right]^{i-k}, \tag{21c}$$

and the expectation in (20) *is taken with respect to independent RVs* N_c *and* $N'_{\bar{c}}$*, with the latter distributed as in* (12) *with the index* $\bar{c}$ *in lieu of* c*.*

Proof: The proof is detailed in Appendix D. □

We now address the CS analysis. With finite K, a CS message is correctly received at any AP if it is the only non-erased CS message and no more than K NCS messages are received erasure-free at the AP. Conditioned on the number of messages $N_c = n_c$ and $N_{\bar{c}} = n_{\bar{c}}$, the probability of the first event is given by p_{n_c} defined in (1), while the probability of the second event is given by $\gamma_K(n_{\bar{c}},\epsilon_1)$ with

$$\gamma_K(x,\epsilon) = \begin{cases} 1 & \text{if} \quad x \le K \\ \sum_{i=0}^{K}\binom{x}{i}(1-\epsilon)^i\epsilon^{x-i} & \text{otherwise.} \end{cases} \tag{22}$$

Removing the conditioning on $N_{\bar{c}}$, the probability of the second event can be written as

$$\begin{aligned} &\sum_{n_{\bar{c}}=0}^{K}\frac{G_{\bar{c}}^{n_{\bar{c}}}\,e^{-G_{\bar{c}}}}{n_{\bar{c}}!} + \sum_{n_{\bar{c}}=K+1}^{\infty}\frac{G_{\bar{c}}^{n_{\bar{c}}}\,e^{-G_{\bar{c}}}}{n_{\bar{c}}!}\left[\sum_{i=0}^{K}\binom{n_{\bar{c}}}{i}(1-\epsilon_1)^i\,\epsilon_1^{n_{\bar{c}}-i}\right] \\ &= Q(K+1,G_{\bar{c}}) + \xi(K,G_{\bar{c}}), \end{aligned} \tag{23}$$

where the first term in (23) is the regularized gamma function and $\xi(K,G_{\bar{c}})$ represents the second term. For the CS performance metrics we distinguish the following two cases depending on the number L of APs.

1) Small number of APs ($L \le K+1$): In this case, the effect of finite interference tolerance K affects only the radio access transmission phase. In fact, in the backhaul transmission phase, if $L \le K+1$, the number of interfering NCS transmissions on a CS packet at the BS cannot exceed K.

Proposition 5: *Under the collision model, the CS throughput and packet success rate given as a function of the number of APs $L \leq K+1$, channel erasure probabilities ϵ_1 and ϵ_2 and CS and NCS packet loads G_c and $G_{\bar{c}}$ as*

$$R_c = \sum_{\ell=0}^{L-1} (-1)^{\ell} L \binom{L-1}{\ell} \left[\frac{(1-\epsilon_1)(1-\epsilon_2)}{\epsilon_1}\right]^{\ell+1} \cdot e^{-G_c}\, \mathcal{H}_{\ell+1}\left(G_c\, \epsilon_1^{\ell+1}\right) \cdot [\mathsf{Q}(K+1, G_{\bar{c}}) + \xi(K, G_{\bar{c}}, \ell)] \tag{24}$$

$$\text{and} \quad \Gamma_c = \mathbb{E}_{N'_c, N_{\bar{c}}}[L p_u (1-\epsilon_2)(1-q_{N'_c})^{L-1}], \tag{25}$$

where

$$\xi(K, G_{\bar{c}}, \ell) = \sum_{n_{\bar{c}}=K+1}^{\infty} \frac{G_{\bar{c}}^{n_{\bar{c}}}\, e^{-G_{\bar{c}}}}{n_{\bar{c}}!} \left[\sum_{i=0}^{K} \binom{n_{\bar{c}}}{i} (1-\epsilon_1)^i \epsilon_1^{n_{\bar{c}}-i}\right]^{\ell+1} \tag{26}$$

and $q_{n'_c} = n'_c(1-\epsilon_1)\epsilon_1^{n'_c-1}\gamma_K(n_{\bar{c}}, \epsilon_1)(1-\epsilon_2)$ is the probability of receiving any CS packet at the BS. and the expectation in (25) *is taken with respect to independent RVs $N_{\bar{c}}$ and N'_c, with the latter distributed as in* (12).

Proof: The proof is provided in Appendix E. □

Comparing the CS throughput in (24) with the expression (5) for $K \to \infty$ we observe that the effect of a finite interference tolerance is measured by the multiplicative term $[Q(K+1, G_{\bar{c}}) + \xi(K, G_{\bar{c}}, \ell)]$. It can be shown that this term is always smaller than one, which is in line with the fact that a lower CS throughput is expected when K is finite.

2) Large Number of APs ($L > K+1$): In this case, a successful CS transmission occurs in all events where a single CS packet and only up to $K < L$ NCS packets reach the BS. The total probability of these events given $N_c = n_c$ and $N_{\bar{c}} = n_{\bar{c}}$ can be computed as

$$q_c = \sum_{\ell=0}^{K} \binom{L}{1, \ell, L-\ell-1} q'_{n_c} (q_{n_{\bar{c}}})^{\ell} (1 - q'_{n_c} - q_{n_{\bar{c}}})^{L-\ell-1}, \tag{27}$$

where $q'_{n_c} = q_{n_{\bar{c}}} \gamma_K(n_{\bar{c}}, \epsilon_1)\,(1-\epsilon_2)$ is the probability that a CS packet reaches the BS and $q_{n_{\bar{c}}}$ is the probability that a NCS packet reaches the BS (may also be not correctly received due to a collision). Removing the conditioning on $N_c = n_c$ and $N_{\bar{c}} = n_{\bar{c}}$, we get the CS throughput as

$$R_c = \mathbb{E}_{N_c, N_{\bar{c}}}\left[q_c\right], \tag{28}$$

where the expectation is taken with respect to independent RVs $N_{\bar{c}}$ and N_c.

Moving to the CS packet success rate, conditioned on $N_c = n_c$ and $N_{\bar{c}} = n_{\bar{c}}$, the probability p_u of receiving a packet at an AP from a given user u is given as in $p_u = (1-\epsilon_1)\epsilon_1^{n'_c-1}\gamma_K(n_{\bar{c}}, \epsilon_1)$. The probability of receiving successfully a CS packet at the BS is then given as

$$q_c = Lp_u(1-\epsilon_2)(1-q_{n'_c})^{L-1}\underbrace{\sum_{i=0}^{K}\binom{L-1}{i}p_{\bar{c}}^i(1-p_{\bar{c}})^{L-1-i}}_{(a)} \tag{29}$$

where $q_{n'_c} = p_{n'_c}\gamma_K(n_{\bar{c}}, \epsilon_1)(1-\epsilon_2)$ and $p_{\bar{c}} = n_{\bar{c}}(1-\epsilon_1)\epsilon_1^{n'_c}\epsilon_1^{n_{\bar{c}}-1}(1-\epsilon_2)$. The main difference between (29) and the probability inside the expectation in (25) is the multiplication by the term (a) in (29) which corresponds to the probability that a number of NCS packets lower or equal to K should be received in order to be able to recover a CS packet. Removing the conditioning on $N_c = n_c$ and $N_{\bar{c}} = n_{\bar{c}}$, we obtain the packet success rate as

$$\Gamma_c = \mathbb{E}_{N'_c, N_{\bar{c}}}\left[q_c\right], \tag{30}$$

where the expectation is taken with respect to independent RVs $N'_{\bar{c}}$ and N'_c with the latter distributed as in (12).

D. Examples

In order to capture the effect of the number of NCS messages K on the CS, in Fig. 4 we plot the throughput region for $K = 2$ and $K \to \infty$, with the latter case corresponding to the analysis in Sec. V-A. The region includes all throughput pairs that are achievable for some value of the fraction γ_c of CS messages, as well as all throughput pairs that are dominated by an achievable throughput pair (i.e., for which both CS and NCS throughputs are smaller than for an achievable pair). For reference, we also plot the throughput region for a conventional inter-service TDMA protocol, whereby a fraction αT for $\alpha \in [0, 1]$ of the T time-slots is allocated for CS messages and the remaining time-slots to NCS messages. For TDMA, the throughput region includes all throughput pairs that are achievable for some value of α, as well as of γ_c.

A first observation from the figure is that non-orthogonal resource allocation can accommodate a significant NCS throughput without affecting the CS throughput, while TDMA causes a reduction in the CS throughput for any increase in the NCS throughput. This is due to the need in TDMA to allocate orthogonal time resources to NCS messages in order to increase the corresponding throughput. However, with non-orthogonal resource allocation, the maximum NCS throughput is generally penalized by the interference caused by the collisions from CS messages,

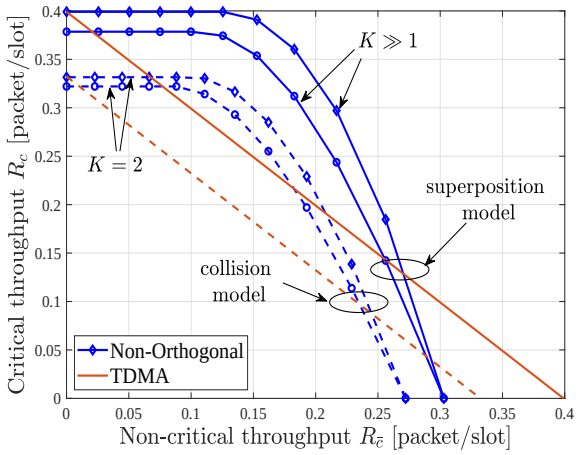


Fig. 4: Achievable throughput region for CS and NCS under superposition and collision models for $K = 2$ and $K \gg 1$ ($\epsilon = 0.5$, $G = 8$ [packet/frame], $T = 2$ [time-slot/frame], and $L = 3$ APs).

while this is not the case for TDMA. In summary, TDMA is preferable when one wishes to guarantee a large NCS throughput and the CS throughput requirements are loose; otherwise, non-orthogonal resource allocation outperforms TDMA in terms of throughput. Furthermore, the throughput region is generally decreased by lower value of K. Experiments concerning packet success rate and performance as function of the number of APs will be presented in the superposition model in the following section.

VI. Heterogeneous Services Under Superposition Model

In this section, we consider the superposition model described in Sec. III.

A. Performance Analysis

Unlike the collision model, in order to analyze the throughput and packet success rate under the superposition model, one needs to keep track of the index of the messages decoded by the APs. This is necessary to detect when multiple versions of the same message (i.e., sent by the same device) are received at the BS. Accordingly, we start by defining the RVs B_i to denote the

index of the message received at AP i and RV B for the BS at any time-slot. Accordingly, for given values $N_c = n_c$ and $N_{\bar{c}} = n_{\bar{c}}$ of transmitted messages, RVs $\{B_i\}$ can take values

$$B_i = \begin{cases} 0 & \text{if no message is retrieved} \\ & \text{due to erasures or collisions} \\ 1 \leq m \leq n_c & \text{if the } m\text{-th CS message} \\ & \text{is retrieved} \\ n_c + 1 \leq m \leq n_c + n_{\bar{c}} & \text{if the } (m - n_c)\text{-th NCS} \\ & \text{message is retrieved.} \end{cases} \tag{31}$$

Note that we have indexed CS messages from 1 to n_c and NCS messages from n_c+1 to $n_c+n_{\bar{c}}$. As for the RV B at the BS, it is defined as

$$B = \begin{cases} c & \text{if a CS message is retrieved} \\ \bar{c} & \text{if a NCS message is retrieved} \\ 0 & \text{if no message is retrieved due to erasures} \\ & \text{or collisions.} \end{cases} \tag{32}$$

Furthermore, we define as $M_m = \sum_{i=1}^{L} \mathbb{1}_{\{B_i=m\}}$ the RVs denoting the number of APs that have message of index $m \in \{0, 1, \ldots, n_c, n_c + 1, \ldots, n_c + n_{\bar{c}}\}$. The joint distribution of RVs $\{M_m\}_{m=0}^{n_c+n_{\bar{c}}}$ given N_c and $N_{\bar{c}}$ is multinomial and can be written as follows

$$\begin{aligned} &\{M_m\}_{m=0}^{n_c+n_{\bar{c}}} | N_c, N_{\bar{c}} \sim \\ &\text{Multinomial}\Big(L, \overbrace{1-p_{n_c}-p_{n_{\bar{c}}}}^{0}, \overbrace{\frac{p_{n_c}}{n_c}, \ldots, \frac{p_{n_c}}{n_c}}^{n_c}, \overbrace{\frac{p_{n_{\bar{c}}}}{n_{\bar{c}}}, \ldots, \frac{p_{n_{\bar{c}}}}{n_{\bar{c}}}}^{n_{\bar{c}}}\Big), \end{aligned} \tag{33}$$

where we used the the probabilities in (1) and (49) that one of the CS or NCS message is received at an AP respectively in a given time-slot. The probability of retrieving a CS message in a given time-slot at the BS conditioned on N_c, $N_{\bar{c}}$ and $\{M_{m\prime}\}_{m\prime=0}^{n_c+n_{\bar{c}}}$ can be then written as

$$\begin{aligned} q_c &= \Pr[B = c | N_c = n_c, N_{\bar{c}} = n_{\bar{c}}, \{M_{m'}\}_{m'=0}^{n_c+n_{\bar{c}}}] \\ &= \gamma_K\Big(\sum_{m'=n_c+1}^{n_{\bar{c}}+n_c} M_{m'}, \epsilon_2\Big) \\ &\quad \cdot \sum_{m=1}^{n_c} \sum_{j=1}^{M_m} \binom{M_m}{j} (1-\epsilon_2)^j \epsilon_2^{\sum_{\substack{m'=0 \\ m'\neq m}}^{n_c} M_{m'}+M_m-j}, \end{aligned} \tag{34}$$

where the first sum is over all possible CS messages, the second sum is over all combinations of APs that have the CS message m, and the third sum at the exponent is over all APs that have a CS message $m' \neq m$. The CS throughput can be computed by averaging (34) over all conditioning variables as

$$R_c = \mathbb{E}_{N_c, N_{\bar{c}}, \{M_m\}_{m=0}^{N_c+N_{\bar{c}}}}[q_c]. \tag{35}$$

In a similar manner, the conditional probability of receiving a NCS message at the BS can be written as

$$\begin{aligned} q_{\bar{c}} &= \Pr[B = \bar{c} | N_c = n_c, N_{\bar{c}} = n_{\bar{c}}, \{M_{m'}\}_{m'=0}^{n_c+n_{\bar{c}}}] \\ &= \sum_{m=n_c+1}^{n_c+n_{\bar{c}}} \sum_{j=1}^{M_m} \binom{M_m}{j} (1-\epsilon_2)^j \epsilon_2^C, \end{aligned} \tag{36}$$

where

$$C = \sum_{\substack{m''=n_c+1 \\ m'' \neq m}}^{n_c+n_{\bar{c}}} M_{m''} + M_m - j + \sum_{m'=1}^{n_c} M_{m'} \tag{37}$$

where the first sum in (36) is over all possible NCS messages m; the second sum is over all possible combinations of APs that have message m. The first and second sums in (37) are over all APs that have a different NCS message and a CS message respectively. The NCS throughput can be then obtained by averaging over the conditioning RVs as

$$R_{\bar{c}} = \mathbb{E}_{N_c, N_{\bar{c}}, \{M_m\}_{m=0}^{N_c+N_{\bar{c}}}}[q_{\bar{c}}]. \tag{38}$$

The packet success rate under the superposition model for CS and NCS can be obtained by fixing m to one and substituting n_c and $n_{\bar{c}}$ by n'_c and $n'_{\bar{c}}$ in (34) and (36).

B. Examples

In Fig. 4, we plot the throughput region for non-orthogonal resource allocation and inter-service TDMA under the superposition model. Comparing the regions of the collision model and the superposition model, it is clear that the latter provides a larger throughput region being able to leverage transmissions of the same packets from multiple APs as compared to the collision model. This can also be seen as function of K in Fig. 4.

In Fig. 5, we explore the effect of the number of APs L on the CS and NCS throughputs. In practice, in NTN, the number of LEO satellites available can be tuned by properly designing their orbit, or by varying their speed of rotation in order to slow them down above areas of high devices density. To capture separately the effects of the radio access and the backhaul channel

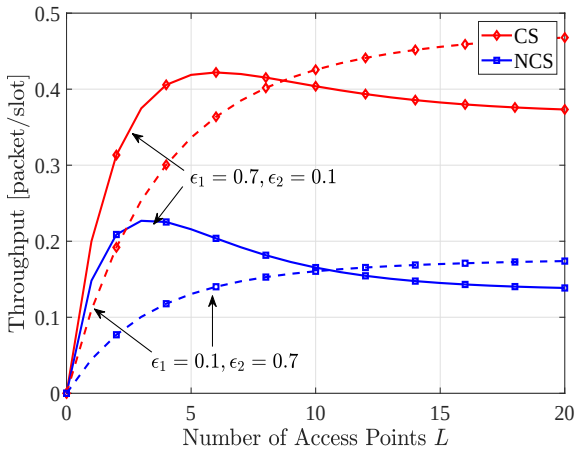
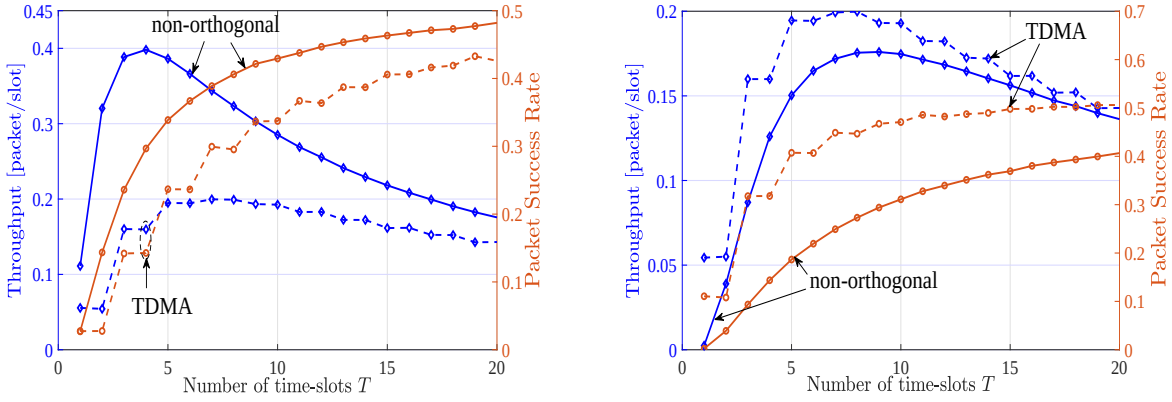


Fig. 5: CS and NCS throughputs as function of the number of APs L under the superposition model under non-orthogonal resource allocation ($G = 8$ [packet/frame], $T = 4$ [time-slot/frame], $\gamma_c = 0.5$ and for $\epsilon_1 \neq \epsilon_2$).

erasures, we consider different values for the channel erasure probabilities ϵ_1 and ϵ_2. We highlight two different regimes: the first is when ϵ_1 is large and ϵ_2 is small, and hence larger erasures occur on the access channel; while the second covers the complementary case where ϵ_1 is small and ϵ_2 is large. In the first regime, increasing the number of APs is initially beneficial to both CS and NCS messages in order to provide additional spatial diversity for the radio access, given the large value of ϵ_1; but larger values of L eventually increase the probability of collisions at the BS on the backhaul due to the low value of ϵ_2. In the second regime, when $\epsilon_1 = 0.1$ and $\epsilon_2 = 0.8$ much lower throughputs are generally obtained due to the significant losses on the backhaul channel. This can be mitigated by increasing the number of APs, which increases the probability of receiving a packet at the BS.

Finally, we consider the interplay between the throughputs and packet success rate levels for both non-orthogonal resource allocation and TDMA as function of the number of time-slots T. These are plotted in Fig. 6 for $G = 15$ [packet/frame], $\epsilon_1 = \epsilon_2 = 0.5, L = 3$ APs, $\alpha = 0.5$ and $\gamma_c = 0.5$. For both services, following the discussion around Fig. 2 we observe that the packet success rate level under both allocation schemes increases as function of T. This is because larger value of T decrease chances of packet collisions. However, this is not the case for the throughput, since large values of T may cause some time-slots to be left unused, which penalizes the throughput. For the CS in Fig. 6a, it is seen that non-orthogonal resource allocation



(a) CS throughput and packet success rate

(b) NCS throughput and packet success rate

Fig. 6: CS and NCS throughputs and packet success rate levels as function of the number of time-slots T for non-orthogonal resource allocation (solid lines) and inter-service TDMA (dashed lines) ($G = 15$ [packet/frame], $\epsilon_1 = \epsilon_2 = 0.5, L = 3$ APs, $\alpha = 0.5$ and $\gamma_c = 0.5$).

outperforms TDMA in both throughput and packet success rate level due to the larger number of available resources. In contrast, Fig. 6b shows that TDMA provides better NCS throughput and packet success rate level than non-orthogonal resource allocation. The main reason for this is that the lower number of resources in TDMA is compensated by the absence of inter-service interference for NCS messages.

VII. Throughput and Packet Success Rate Analysis Under Fading Channels

The binary erasure channel model discussed in the previous sections offers a tractable set-up that facilitates the analysis of the throughput and packet success rate, enabling the derivation of closed-form expressions in various cases of interest. It is also of practical interest as a simplified model for mmwave channels [51] and satellite communications scenarios represented in Fig. 1. In this section, we briefly study a more common scenario that accounts for fading channels in both radio and backhaul channels. This typically represents terrestrial scenarios as shown in Fig. 1b. More complex models that include both fading and erasures [52] can also be analyzed following the same steps presented below (see Section VIII for some details). We first detail the channel and signal models, and then we derive the throughput and packet success rate metrics.

A. Channel and Signal Models

At any time-slot t, the channels between each user m and AP l and between each AP l and the BS are assumed to follow the standard Rayleigh fading model, and are denoted as $h_{l,m}(t) \sim \mathcal{CN}(0, \alpha^2)$ and $g_l(t) \sim \mathcal{CN}(0, \beta^2)$, respectively. Ensuring consistency with the erasure model, we assume that all channels are independent and that the average channel gains α^2 and β^2 are fixed. Furthermore, as detailed below, we assume that each AP and the BS decode at most one packet in each slot. Finally, we denote the transmission rates of CS and NCS messages as r_c and $r_{\bar{c}}$ bit/s/Hz, respectively. Assuming that both access and backhaul channels are allocated the same amount of radio resources, the transmission rates are the same for both channels.

Given the numbers $N_c(t) = n_c$ and $N_{\bar{c}}(t) = n_{\bar{c}}$ of CS and NCS messages in the given time-slot t, the signal received at the l-th AP as time-slot t can be written as

$$y_l(t) = \sum_{m=1}^{n_c} h_{lm}(t) x_m(t) + \sum_{m'=n_c+1}^{n_c+n_{\bar{c}}} h_{lm'}(t) x_{m'}(t) + n_l(t), \tag{39}$$

where $n_l(t) \sim \mathcal{CN}(0, 1)$ denotes complex white Gaussian noise at the l-th AP. The powers of CS and NCS devices are respectively denoted as

$$\mathbb{E}[|x_m(t)|^2] = P_c \text{ and } \mathbb{E}[|x_{m'}(t)|^2] = P_{\bar{c}}, \tag{40}$$

where we take $P_c \geq P_{\bar{c}}$ to capture the generally larger transmission power of CS transmissions. The Signal-to-Interference-plus-Noise Ratio (SINR) of message m at AP l is given as

$$\text{SINR}_{l,m}^{AP} = \frac{|h_{lm}(t)|^2 P_m}{1 + \sum_{\substack{m'=1 \\ m' \neq m}}^{n_c+n_{\bar{c}}} |h_{lm'}(t)|^2 P_{m'}}, \tag{41}$$

where $P_m = P_c$ for a CS message $m \in \{1, \ldots, n_c\}$ and $P_m = P_{\bar{c}}$ for a NCS message $m \in \{n_c + 1, \ldots, n_c + n_{\bar{c}}\}$. Let $m_l^\star$ denote the message with the largest SINR at the l-th AP, i.e.,

$$m_l^\star = \underset{m \in \{1, \ldots, n_c + n_{\bar{c}}\}}{\text{argmax}} \text{SINR}_{l,m}^{AP}. \tag{42}$$

The l-th AP only attempts to decode message $m_l^\star$. Decoding is correct if the standard Shannon capacity condition $\text{SINR}_{l,m_l^\star} \geq 2^{r_m} - 1$ is satisfied, where $r_m = r_c$ if $m_l^\star \in \{1, \ldots, n_c\}$ and $r_m = r_{\bar{c}}$ if $m_l^\star \in \{n_c, \ldots, n_c + n_{\bar{c}}\}$.

Each l-th AP transmits the decoded message $m_l^\star$, if any, to the BS over the wireless backhaul channel with transmission power $P_{m_l^\star}^{AP} = P_c^{AP}$ if $m_l^\star \in \{1, \ldots, n_c\}$ and $P_{m_l^\star}^{AP} = P_{\bar{c}}^{AP}$ if $m_l^\star \in$

$\{n_c+1,\ldots,n_c+n_{\bar{c}}\}$. Consequently, the signal $y_{BS}(t+1)$ received at the BS in time-slot $t+1$ can be written as the sum of messages sent by all APs as

$$y_{BS}(t+1)=\sum_{l=1}^{L} g_l(t+1)x_{m_l^\star}(t)+n_{BS}(t+1), \tag{43}$$

where $n_{BS}(t)\sim\mathcal{CN}(0,1)$ denotes the white Gaussian noise at the BS. Let $\mathcal{L}_m=\{l: m_l^\star=m\}$ denote the set of indices of APs that decoded a message $m\in\mathcal{M}^\star$, where $\mathcal{M}^\star=\{m:\exists\ l=1,\ldots,L \text{ s.t. } m=m_l^\star\}$ denotes the set of messages decoded by at least one AP in time-slot t. The SINR of a message $m\in\mathcal{M}^\star$ received at the BS can be written as

$$\mathrm{SINR}_m^{BS}=\frac{|\sum_{l\in\mathcal{L}_m} g_l(t+1)|^2 P_m^{AP}}{1+\sum_{m'\in\mathcal{M}^\star\setminus\{m\}}|\sum_{l\in\mathcal{L}_{m'}} g_l(t+1)|^2 P_{m'}^{AP}}. \tag{44}$$

In a manner similar to APs, the BS attempts decoding only of the message $m_{BS}^\star$ with the highest SINR, namely

$$m_{BS}^\star=\underset{m\in\mathcal{M}^\star}{\mathrm{argmax}}\,\mathrm{SINR}_m^{BS}. \tag{45}$$

Message $m_{BS}^\star$ is decoded correctly if the standard Shannon capacity condition $\mathrm{SINR}_{m_{BS}^\star}\geq 2^{r_{m_{BS}^\star}}-1$ is satisfied, where $r_{m_{BS}^\star}=r_c$ if $m_{BS}^\star\in\{1,\ldots,n_c\}$ and $r_{m_{BS}^\star}=r_{\bar{c}}$ if $m_{BS}^\star\in\{n_c+1,\ldots,n_c+n_{\bar{c}}\}$.

B. Performance Analysis

The analysis follows the same steps as in Section VI, as long as one properly redefines the probabilities p_c and $p_{\bar{c}}$ of decoding correctly a CS or a NCS message at any given AP, as well as the probabilities q_c and $q_{\bar{c}}$ of decoding correctly a CS or NCS message at the BS. According to the discussion in Section VII-A, the former probabilities can be respectively written as

$$p_c=\Pr[m_l^\star\in\{1,\ldots,n_c\}\text{ and }\mathrm{SINR}_{l,m_l^\star}^{AP}>2^{r_c}-1] \tag{46a}$$

$$\text{and } p_{\bar{c}}=\Pr[m_l^\star\in\{n_c+1,\ldots,n_c+n_{\bar{c}}\} \tag{46b}$$

$$\text{and } \mathrm{SINR}_{l,m_l^\star}^{AP}>2^{r_{\bar{c}}}-1],$$

where $m_l^\star$ is defined in (42), while the latter probabilities can be redefined as

$$q_c=\Pr[m_{BS}^\star\in\{1,\ldots,n_c\}\text{ and }\mathrm{SINR}_{m_{BS}^\star}^{BS}>2^{r_c}-1] \tag{47a}$$

$$\text{and } q_{\bar{c}}=\Pr[m_{BS}^\star\in\{n_c+1,\ldots,n_c+n_{\bar{c}}\} \tag{47b}$$

$$\text{and } \mathrm{SINR}_{m_{BS}^\star}^{BS}>2^{r_{\bar{c}}}-1]. \tag{47c}$$

While closed-form expressions for (46) and (47) appear prohibitive to derive (see, e.g., [53]), these probabilities can be easily estimated via Monte Carlo Simulations. Having computed probabilities (46)-(47), the throughputs of CS and NCS messages can be respectively obtained using (35) and (38). The packet success rate can be computed by redefining q_c and $q_{\bar{c}}$ to take into account a single message sent by a single user (for instance, the first one) as follows:

$$q_c = \Pr[m^{\star}_{BS} = 1 \text{ and } \mathrm{SINR}^{BS}_{m^{\star}_{BS}} > 2^{r_c} - 1 | N_c(t) \geq 1] \tag{48a}$$

$$\text{and } q_{\bar{c}} = \Pr[m^{\star}_{BS} = 1 \text{ and } \mathrm{SINR}^{BS}_{m^{\star}_{BS}} > 2^{r_{\bar{c}}} - 1 | N_{\bar{c}}(t) \geq 1]. \tag{48b}$$

C. Examples

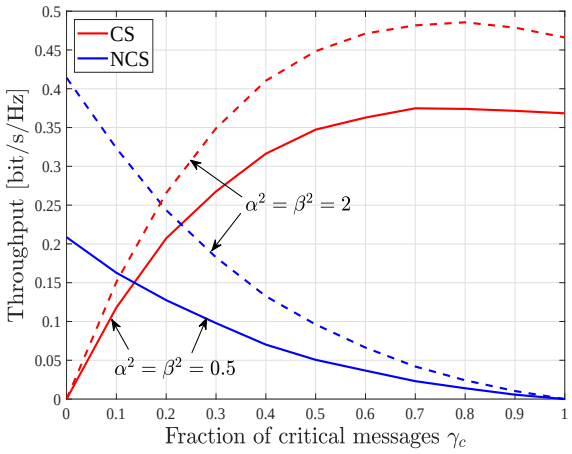


Fig. 7: CS and NCS throughput as function of the fraction of CS messages γ_c under the fading channels model and using non-orthogonal resource allocation ($G = 20$ [packet/frame], $T = 4$ [time-slot/frame], $P_c = P_c^{AP} = 10$, $P_{\bar{c}} = P_{\bar{c}}^{AP} = 4$ and $L = 3$ APs).

We now consider the fading channels model discussed in Sec. VII with the main aim of relating the insights obtained from the analysis of erasure channels to the more common Rayleigh fading setup. We fix $P_c = P_c^{AP} = 10$ and $T = 4$ [time-slot/frame]. In an analogy to Fig. 3, we start in Fig. 7 by investigating the throughput of CS and NCS messages as function of the fraction of CS messages γ_c for different values of the average channel powers α^2 and β^2. In general, we observe similar trends as in Fig. 3. Most notably, the throughput of CS messages peaks at a value of γ_c that strikes the best balance between the combining gains due to the transmission

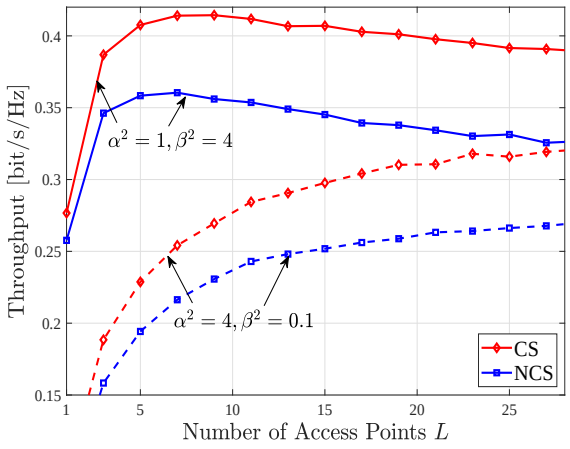


Fig. 8: CS and NCS throughputs as function of the number of APs L under the fading model and under non-orthogonal resource allocation ($G = 10$ [packet/frame], $T = 4$ [time-slot/frame], $\gamma_c = 0.5$, $P_c = P_c^{AP} = 10$, and $P_{\bar{c}} = P_{\bar{c}}^{AP} = 9$).

of a message from multiple APs and the interference created by concurrent AP transmissions of different messages. However, in contrast to the erasure model in which increasing the erasure rate can be advantageous, the throughput of both services improves as the average channel strengths α^2 and β^2 are increased. This is because interference from concurrent transmissions has a more deleterious effect under the collision model assumed when considering erasures than under the SINR model. For the latter, reducing both channel strengths α^2 and β^2 has the net effect of reducing the SINRs despite the decrease in interference power.

Finally, to compare some of the design insights from the erasure model, we plot in Fig. 8 the throughput of both services as function of the number of APs L. We can see that, in a manner similar to Fig. 5, when α^2 is high and β^2 is low, which is akin to lower ϵ_1 and high ϵ_2 in Fig. 5, the throughput of both services increases as function of L. This is because low values of the channel power β^2 in the backhaul channel imply that the SINR is limited by the signal power and not by interference. Therefore, increasing the space diversity via a larger L can be advantageous in this regime. Furthermore, as evinced from Fig. 5, the SINR of the backhaul channel may be limited by the level of interference and hence when α^2 is low and β^2 is high, which is akin to high ϵ_1 and low ϵ_2 in Fig. 5, increasing the number of APs beyond a given threshold reduces the throughput.

VIII. Conclusions And Extensions

This paper studies grant-free random access for coexisting CS and NCS in IoT systems with shared wireless backhaul and uncoordinated APs. Non-orthogonal and orthogonal inter-service resource sharing schemes based on random access are considered. From the CS perspective, it was found that non-orthogonal sharing is preferable to a standard inter-service TDMA protocol in terms of both throughput and packet success rate level. In contrast, this is not the case for the NCS, since inter-service orthogonal resource allocation eliminates interference from the larger-power CS. Furthermore, both erasure and fading channels models are considered to capture satellite and terrestrial applications respectively. Similarities were found between these models which proves the suitability of the erasure model for such type of analysis due to its mathematical tractability properties. Through extensive numerical results, the impact of both spatial and time resources is investigated, revealing trade-offs between throughput and packet success rate for both services.

Our model could be directly extended to cater for multiple antennas satellites, known as multi-beam satellites, and more realistic inter-satellite interference. For the former, different beams could be used per-antenna to cover different areas, with the proposed grant-free NOMA applied separately to each area. As for the latter, it is more realistic to assume different direct and interference channel gains to account for the revolution of LEO satellites around the planet. This can be directly considered by modeling interference channels with a different channel gain than the direct channel. The analysis could be directly extended to this case at the price of a more cumbersome notation.

Appendix

A. *Proof of Proposition 3*

An AP successfully retrieves a non-critical packet when only one of the $N_{\bar{c}}$ non-critical packets transmitted arrives unerased and all critical transmitted packets N_c do not reach the AP due to erasures. This happens with probability

$$p_{N_{\bar{c}}} = \underbrace{n_{\bar{c}} \left(1-\epsilon_1\right) \epsilon_1^{n_{\bar{c}}-1}}_{(a)} \underbrace{\epsilon_1^{n_c}}_{(b)}, \tag{49}$$

where term (a) is the probability that one of the $n_{\bar{c}}$ non-critical messages is successfully received at an AP and term (b) is the probability all critical messages are erased. On the backhaul, a

non-critical packet reaches the BS via one of the APs when the packet is not erased over the backhaul channel and no other packet (critical and non-critical) is successfully received from any of the remaining $L-1$ APs. The overall probability of successful reception at the BS is thus $q_{\bar{c}} = Lq_{n_{\bar{c}}}(1-q_{n_c}-q_{n_{\bar{c}}})^{L-1}$, with $q_{n_{\bar{c}}} = p_{n_{\bar{c}}}(1-\epsilon_2)$ and q_{n_c} defined in (3). The non-critical throughput can then be written by averaging $q_{\bar{c}}$ over n_c and $n_{\bar{c}}$ as

$$R_{\bar{c}} = \mathbb{E}_{N_c,N_{\bar{c}}}[q_{\bar{c}}] = \sum_{n_{\bar{c}}=0}^{\infty}\sum_{n_c=0}^{\infty} \frac{G_{\bar{c}}^{n_{\bar{c}}}e^{-G_{\bar{c}}}}{n_{\bar{c}}!}\frac{G_c^{n_c}e^{-G_c}}{n_c!} \cdot L\, q_{n_{\bar{c}}}(1-q_{n_c}-q_{n_{\bar{c}}})^{L-1}. \tag{50}$$

Following similar steps as the one detailed in the proof of *Proposition 1* and after some tedious yet straightforward rearrangements, the non-critical throughput in (50) can be written in closed form as detailed in (15).

We now derive the non-critical packet success rate. Similar to the single-service case, the probability of receiving a non-critical packet at an AP from a given user u given that u is active is given by

$$p_u = (1-\epsilon_1)\epsilon_1^{n'_{\bar{c}}-1}\epsilon_1^{n_c}. \tag{51}$$

The probability that the packet from user u is received successfully at the BS is then computed as

$$q_u = Lp_u(1-\epsilon_2)(1-q)^{L-1}, \tag{52}$$

where $q = (1-\epsilon_2)[p_{n_c} + p_{n'_{\bar{c}}}]$ is the probability that the BS successfully receives a critical message or a non critical message from any of the remaining $L-1$ APs. The non-critical packet success rate $\Gamma_{\bar{c}}$ can be obtained by averaging q_u over $n'_{\bar{c}}$ and n_c, where $P[N'_{\bar{c}} = n'_{\bar{c}}|N'_{\bar{c}} \geq 1] = (1-e^{-G_{\bar{c}}})^{-1}(e^{-G_{\bar{c}}}G_{\bar{c}}^{n'_{\bar{c}}})/(n'_{\bar{c}}!)$ is the distribution of non-critical packets given that user u is active. The non-critical packet success rate $\Gamma_{\bar{c}}$ can be obtained in closed form by following similar steps in the proof of *Proposition 1*.

B. Proof of Proposition 4

The probability of receiving a non-critical message at the BS can be written as

$$q_{\bar{c}} = Lq_{n_{\bar{c}}}(1-q_{n_c}-q_{n_{\bar{c}}})^{L-1}, \tag{53}$$

where $q_{n_c} = p_{n_c}\gamma_K(n_{\bar{c}},\epsilon_1)(1-\epsilon_2)$ and $q_{n_{\bar{c}}} = p_{n_{\bar{c}}}(1-\epsilon_2)$ are the probabilities of receiving successfully a critical or non-critical packet respectively at the BS. The non-critical throughput

$R_{\bar{c}}$ can be then obtained by averaging $q_{\bar{c}}$ in (53) over all values of n_c.

Moving to non-critical packet success rate $\Gamma_{\bar{c}}$, the probability of receiving successfully the packet of a given user u at an AP is defined in the same way as in (51). The probability of receiving this packet at the BS can be written as

$$q_{\bar{c}} = L p_u (1-\epsilon_2)(1-q)^{L-1}, \tag{54}$$

where q is the probability of receiving any critical or non-critical packet from the remaining $L-1$ APs. This can be written as $q = (1-\epsilon_2)[p_{n_c}\gamma_{K-1}(n'_{\bar{c}}, \epsilon_1) + n'_{\bar{c}} p_u]$, where the first part corresponds to receiving any critical message at the BS while the second part to receiving any non-critical message at the BS. The non-critical packet success rate $\Gamma_{\bar{c}}$ can be then obtained by averaging $q_{\bar{c}}$ over $N_c = n_c$ and $N_{\bar{c}} = n_{\bar{c}}$.

C. Proof of Proposition 5

The event that a transmitted critical packet is received at the BS passing through one of the APs occurs if the AP successfully decodes one critical packet, and the packet is not erased over the backhaul channel. Conditioned on $N_c = n_c$ and $N_{\bar{c}} = n_{\bar{c}}$, this event has the probability $q'_{n_c} = p_{n_c}\gamma_K(n_{\bar{c}}, \epsilon_1)\,(1-\epsilon_2)$. The number of incoming backhaul critical packets over a slot follows the distribution $\mathrm{Bin}(L, q'_{n_c})$. Hence, the critical throughput can be written as

$$R_c = \sum_{n_c=0}^{\infty}\sum_{n_{\bar{c}}=0}^{\infty} \frac{G_c^{n_c}e^{-G_c}}{n_c!}\frac{G_{\bar{c}}^{n_{\bar{c}}}e^{-G_{\bar{c}}}}{n_{\bar{c}}!} \cdot L\, q'_{n_c}(1-q'_{n_c})^{L-1}. \tag{55}$$

Now we split the sum over the non-critical packets transmitted $n_{\bar{c}}$ in two parts, the first considers a number of non-critical packets not exceeding $n_{\bar{c}} < K$, while the second part corresponds $n_{\bar{c}} \geq K+1$. In the first part $\gamma_K(n_{\bar{c}}, \epsilon_1) = 1$ by definition, so $q'_{n_c} = q_{n_c}$. Consequently, (55) can be written as the sum of two terms

$$\begin{aligned} R_c = &\sum_{n_{\bar{c}}=0}^{K} \frac{G_{\bar{c}}^{n_{\bar{c}}}e^{-G_{\bar{c}}}}{n_{\bar{c}}!}\sum_{n_c=0}^{\infty}\frac{G_c^{n_c}e^{-G_c}}{n_c!} \cdot L\, q_{n_c}(1-q_{n_c})^{L-1} \\ &+ \sum_{n_{\bar{c}}=K+1}^{\infty} \frac{G_{\bar{c}}^{n_{\bar{c}}}e^{-G_{\bar{c}}}}{n_{\bar{c}}!}\sum_{n_c=0}^{\infty}\frac{G_c^{n_c}e^{-G_c}}{n_c!} \cdot L\, q'_{n_c}(1-q'_{n_c})^{L-1}. \end{aligned} \tag{56}$$

The first term in (56) is the product between the cumulative distribution function (CDF) of a Poisson distribution of parameter $G_{\bar{c}}$ computed in K and the same expression of the throughput

for the single service case found in Section IV in (4). As for the second term, following a simple yet tedious mathematical derivation it can be written as

$$\begin{aligned}&\sum_{n_{\bar{c}}=K+1}^{\infty}\frac{G_{\bar{c}}^{n_{\bar{c}}}e^{-G_{\bar{c}}}}{n_{\bar{c}}!}\sum_{n_c=0}^{\infty}\frac{G_c^{n_c}e^{-G_c}}{n_c!}\cdot L\,q'_{n_c}(1-q'_{n_c})^{L-1}\\&=\sum_{\ell=0}^{L-1}(-1)^{\ell}\,L\binom{L-1}{\ell}\left[\frac{(1-\epsilon_1)(1-\epsilon_2)}{\epsilon_1}\right]^{\ell+1}e^{-G_c}\\&\qquad\cdot\ \mathcal{H}_{\ell+1}\left(G_c\,\epsilon_1^{\ell+1}\right)\cdot\xi(K,G_{\bar{c}},\ell),\end{aligned}\tag{57}$$

where

$$\xi(K,G_{\bar{c}},\ell)=\sum_{n_{\bar{c}}=K+1}^{\infty}\frac{G_{\bar{c}}^{n_{\bar{c}}}\,e^{-G_{\bar{c}}}}{n_{\bar{c}}!}\left[\sum_{i=0}^{K}\binom{n_{\bar{c}}}{i}(1-\epsilon_1)^i\,\epsilon_1^{n_{\bar{c}}-i}\right]^{\ell+1}.\tag{58}$$

Finally, putting together (56) and (57) the lemma can be concluded.

Moving to the packet success rate, the probability of receiving a given critical packet from a user u at an AP is

$$p_u=(1-\epsilon_1)\epsilon_1^{n'_c-1}\gamma_K(n_{\bar{c}},\epsilon_1).\tag{59}$$

The probability of receiving this critical packet at the BS is given by

$$q_c=Lp_u(1-\epsilon_2)(1-q_{n'_c})^{L-1}\tag{60}$$

where $q_{n'_c}=n'_c(1-\epsilon_1)\epsilon_1^{n'_c-1}\gamma_K(n_{\bar{c}},\epsilon_1)(1-\epsilon_2)$ is the probability of receiving any critical packet at the BS. Finally the critical packet success rate can be obtained by averaging q_c over n'_c and $n_{\bar{c}}$.

D. Asymptotic Throughput and Packet Success Rate for Single Service

We now state two theorems regarding the asymptotic behaviour of the throughput and packet success rate for large number of APs.

Theorem 1: The critical throughput tends to zero for large number of APs, i.e., $\lim_{L\to+\infty}R_c=0$.

Proof: Let us first start by defining $R_c(m)=\sum_{n_c=1}^{m}\frac{G_c^{n_c}e^{-G_c}}{n_c!}\cdot Lq_{n_c}(1-q_{n_c})^{L-1}$, which represents

the summation in (5) up to m terms. Note that the summation starts at $n_c = 1$ as the throughput is null otherwise. Consequently, we have

$$\begin{aligned}\lim_{L\to\infty} R_c &= \lim_{L\to\infty}\lim_{m\to\infty} R_c(m)\\ &\underset{(a)}{=} \lim_{m\to\infty}\lim_{L\to\infty} R_c(m)\\ &= \lim_{m\to\infty}\sum_{n_c=1}^{m}\frac{G_c^{n_c}e^{-G_c}}{n_c!}\cdot\lim_{L\to\infty}\left[Lq_{n_c}(1-q_{n_c})^{L-1}\right]\\ &\underset{(b)}{=} 0,\end{aligned}$$

where (a) follows from Moore-Osgood theorem [54, Theorem 1] for interchanging limits using the fact that $R_c(m)$ converges for each value of L and (b) is due to the fact that the second limit is null because $(1 - q_{n_c}) < 1$. □

Theorem 2: The critical packet success rate tends to zero for large number of APs, i.e., $\lim_{L\to+\infty} \Gamma_c = 0$.

Proof: The proof follows similar steps as the proof for Theorem 1 detailed above. □

REFERENCES

[1] 3GPP, "Study on new radio (NR) access technology physical layer aspects," TR 38.802, Mar. 2017.

[2] M. Shafi *et al.*, "5G: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE Journ. on Selected Areas in Communi.*, vol. 35, no. 6, pp. 1201–1221, 2017.

[3] A. A. Esswie and K. I. Pedersen, "Opportunistic Spatial Preemptive Scheduling for URLLC and eMBB Coexistence in Multi-User 5G Networks," *IEEE Access*, vol. 6, pp. 38 451–38 463, 2018.

[4] P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. Navarro-Ortiz, and J. M. Lopez-Soler, "Narrowband IoT Data Transmission Procedures for Massive Machine-Type Communications," *IEEE Network*, vol. 31, no. 6, pp. 8–15, November 2017.

[5] 3GPP, "3GPP Release-17," Tech. Rep. [Online]. Available: https://www.3gpp.org/release-17

[6] T. Quek, M. Peng, O. Simeone, and W. Yu, *Cloud Radio Access Networks: Principles, Technologies, and Applications*. Cambridge University Press, 2017.

[7] [Online]. Available: https://www.cnbc.com/2019/04/04/amazon-project-kuiper-broadband-internet-small-satellite-network.html

[8] [Online]. Available: https://www.spacex.com/news/2019/05/24/starlink-mission

[9] A. Azari, P. Popovski, G. Miao, and C. Stefanovic, "Grant-Free Radio Access for Short-Packet Communications over 5G Networks," in *IEEE GLOBECOM 2017*, pp. 1-7, Singapore, Singapore, Dec. 2017.

[10] R. Kassab, O. Simeone, and P. Popovski, "Information-Centric Grant-Free Access for IoT Fog Networks: Edge vs Cloud Detection and Learning," *arXiv preprint arXiv:1907.05182*, 2019.

[11] M. Masoudi, A. Azari, E. A. Yavuz, and C. Cavdar, "Grant-Free Radio Access IoT Networks: Scalability Analysis in Coexistence Scenarios," in *IEEE International Conference on Communications (ICC)*, pp. 1-7, Kansas city, MO, USA, May. 2018.

[12] N. Abramson, "THE ALOHA SYSTEM: another alternative for computer communications," in *Proceedings of the November 17-19, 1970, fall joint computer conference*. ACM, pp. 281-285, 1970.

[13] Sigfox, "SIGFOX: The Global Communications Service Provider for the Internet of Things," *www.sigfox.com*.

[14] LoRa Alliance, "The LoRa Alliance Wide Area Networks for Internet of Things," *www.lora-alliance.org*.

[15] [Online]. Available: https://www.orbcomm.com/eu/networks/satellite/orbcomm-og2

[16] [Online]. Available: https://myriota.com/

[17] Y. . E. Wang, X. Lin, A. Adhikary, A. Grovlen, Y. Sui, Y. Blankenship, J. Bergman, and H. S. Razaghi, "A Primer on 3GPP Narrowband Internet of Things," *IEEE Commun. Magazine*, vol. 55, no. 3, pp. 117–123, March 2017.

[18] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-Orthogonal Multiple Access (NOMA) for Cellular Future Radio Access," in *IEEE Veh. Technol. conf. (VTC Spring)*, pp. 1-5, Dresden, Germany, Jun. 2013.

[19] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (noma)," in *2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Sep. 2013, pp. 611–615.

[20] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. Kwak, "Power-Domain Non-Orthogonal Multiple Access (NOMA) in 5G Systems: Potentials and Challenges," *IEEE Communications Surveys Tutorials*, vol. 19, no. 2, pp. 721–742, Secondquarter 2017.

[21] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the Performance of Non-Orthogonal Multiple Access in 5G Systems with Randomly Deployed Users," *IEEE Sig. Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

[22] R. Kassab, O. Simeone, P. Popovski, and T. Islam, "Non-Orthogonal Multiplexing of Ultra-Reliable and Broadband Services in Fog-Radio Architectures," *IEEE Access*, vol. 7, pp. 13 035–13 049, Jan. 2019.

[23] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55 765–55 779, Sept. 2018.

[24] R. Kassab, O. Simeone, and P. Popovski, "Coexistence of URLLC and eMBB services in the C-RAN Uplink: An Information-Theoretic Study," in *Proc. IEEE GLOBECOM*, Abu Dhabi, UAE, Dec. 2018.

[25] E. Balevi, F. T. A. Rabee, and R. D. Gitlin, "ALOHA-NOMA for Massive Machine-to-Machine IoT Communication," in *IEEE International Conference on Communications (ICC)*, pp. 1-5, Kansas city, MO, USA, May 2018.

[26] E. Paolini, G. Liva, and M. Chiani, "Coded Slotted ALOHA: A Graph-Based Method for Uncoordinated Multiple Access," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6815–6832, Dec 2015.

[27] A. Munari, F. Clazzer, G. Liva, and M. Heindlmaier, "Multiple-Relay Slotted ALOHA: Performance Analysis and Bounds," *arXiv preprint arXiv:1903.03420*, 2019.

[28] D. Jakovetić, D. Bajović, D. Vukobratović, and V. Crnojević, "Cooperative Slotted Aloha for Multi-Base Station Systems," *IEEE Trans. on Commun.*, vol. 63, no. 4, pp. 1443–1456, April 2015.

[29] A. Munari and F. Clazzer, "Modern Random Access for Beyond-5G Systems: a Multiple-Relay ALOHA Perspective," *arXiv preprint arXiv:1906.02054*, 2019.

[30] R. Kassab, O. Simeone, A. Munari, and F. Clazzer, "Space Diversity-Based Grant-Free Random Access for Critical and Non-Critical IoT Services," *arXiv preprint arXiv:1909.10283*, 2019.

[31] A. Ivanov, M. Stoliarenko, S. Kruglik, S. Novichkov, and A. Savinov, "Dynamic resource allocation in leo satellite," in *2019 15th International Wireless Communications Mobile Computing Conference (IWCMC)*, 2019, pp. 930–935.

[32] C. Jin, X. He, and X. Ding, "Traffic analysis of leo satellite internet of things," in *2019 15th International Wireless Communications Mobile Computing Conference (IWCMC)*, 2019, pp. 67–71.

[33] B. Di, L. Song, Y. Li, and H. V. Poor, "Ultra-Dense LEO: Integration of Satellite Access Networks into 5G and Beyond," *IEEE Wireless Communications*, vol. 26, no. 2, pp. 62–69, 2019.

[34] Y. Su, Y. Liu, Y. Zhou, J. Yuan, H. Cao, and J. Shi, "Broadband leo satellite communications: Architectures and key technologies," *IEEE Wireless Communications*, vol. 26, no. 2, pp. 55–61, 2019.

[35] T. Takeda and K. Higuchi, "Enhanced User Fairness Using Non-Orthogonal Access with SIC in Cellular Uplink," in *IEEE Vehicular Technology Conference (VTC Fall)*, Sept 2011, pp. 1–5.

[36] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5g," *IEEE Communications Surveys Tutorials*, vol. 20, no. 3, pp. 2294–2323, 2018.

[37] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the random access channel of lte and lte-a suitable for m2m communications? a survey of alternatives," *IEEE Communications Surveys Tutorials*, vol. 16, no. 1, pp. 4–16, 2014.

[38] H. Han, Y. Li, W. Zhai, and L. Qian, "A grant-free random access scheme for m2m communication in massive mimo systems," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 3602–3613, 2020.

[39] 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 *2014 IEEE Globecom Workshops (GC Wkshps)*, 2014, pp. 900–905.

[40] M. B. Shahab, R. Abbas, M. Shirvanimoghaddam, and S. J. Johnson, "Grant-free non-orthogonal multiple access for iot: A survey," *IEEE Communications Surveys Tutorials*, vol. 22, no. 3, pp. 1805–1838, 2020.

[41] P. Billingsley, *Probability and measure*. John Wiley & Sons, 2008.

[42] S. Azimi, O. Simeone, and R. Tandon, "Content delivery in fog-aided small-cell systems with offline and online caching: An information—Theoretic analysis," *Entropy*, vol. 19, no. 7, p. 366, 2017.

[43] A. Vahid and R. Calderbank, "Two-User Erasure Interference Channels With Local Delayed CSIT," *IEEE Trans. Inf. Theory*, vol. 62, no. 9, pp. 4910–4923, Sep. 2016.

[44] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User Association for Load Balancing in Heterogeneous Cellular Networks," *IEEE Trans. on Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, June 2013.

[45] X. Jia, T. Lv, F. He, and H. Huang, "Collaborative data downloading by using inter-satellite links in leo satellite networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1523–1532, March 2017.

[46] A. Munari, F. Clazzer, G. Liva, and M. Heindlmaier, "Multiple-Relay Slotted ALOHA: Performance Analysis and Bounds," *arXiv preprint arXiv:1903.03420*, 2019.

[47] J. del Prado Pavon, S. S. Nandagopalan, S. Choi, T. Sato, and J. Bennet, "System and method for performing clock synchronization of nodes connected via a wireless local area network," Oct. 10 2006, US Patent 7,120,092.

[48] E. Perron, M. Rezaeian, and A. Grant, "The On-Off Fading Channel," in *Proc. IEEE ISIT*, 2003.

[49] A. Munari, F. Clazzer, and G. Liva, "Multi-Receiver Aloha Systems - a Survey and New Results," in *Proc. IEEE ICC Workshop on Uncoordinated Massive Access Protocols*, 2015.

[50] F. Formaggio, A. Munari, and F. Clazzer, "On Receiver Diversity for Grant-Free Based Machine Type Communications," *Accepted for publications in the Ad Hoc Networks Journal*, 2020.

[51] Y. Ghasempour, N. Prasad, M. Khojastepour, and S. Rangarajan, "Managing analog beams in mmWave networks," in *2017 51st Asilomar Conference on Signals, Systems, and Computers*, Oct 2017, pp. 1212–1218.

[52] I. Moreira, C. Pimentel, F. P. Barros, and D. P. B. Chaves, "Modeling Fading Channels With Binary Erasure Finite-State Markov Channels," *IEEE Trans. Veh. Tech.*, vol. 66, no. 5, pp. 4429–4434, May 2017.

[53] M. Zorzi and S. Pupolin, "Outage probability in multiple access packet radio networks in the presence of fading," *IEEE Transactions on Vehicular Technology*, vol. 43, no. 3, pp. 604–610, Aug 1994.

[54] Z. Kadelburg and M. Marjanovic, “Interchanging two limits,” *Enseign. Math*, vol. 8, pp. 15–29, 2005.