

Article

Spectral Coexistence of QoS-Constrained and IoT Traffic in Satellite Systems

Andrea Munari  and Federico Clazzer * 

German Aerospace Center (DLR), Institute of Communications and Navigation, 82234 Weßling, Germany; andrea.munari@dlr.de

* Correspondence: federico.clazzer@dlr.de

Abstract: The flourishing of Internet of Things (IoT) applications, characterized by vast transmitter populations and the sporadic transmission of small data units, demands innovative solutions for the sharing of the wireless medium. In this context, satellite connectivity is an important enabler for all scenarios in which terminals are under-served by terrestrial communications and are thus fundamental for providing worldwide coverage. In turn, the design of medium access policies that attain efficient use of the scarce spectrum and can cope with flexible yet unpredictable IoT traffic is of the utmost importance. Starting from these remarks, we investigate in this work the coexistence of a quality of service (QoS)-constrained service with IoT traffic in a shared spectrum as alternative to a more traditional orthogonal allocation among the two services, with an eye on satellite applications. Leaning on analytical tools, we provide achievable rate regions, assuming a slotted ALOHA access method for IoT terminals and accounting for practical aspects, such as the transmission of short packets. Interesting trends emerge, showcasing the benefit of an overlay allocation with respect to segregating the resources for the two services.

Keywords: machine-type communications; grant-free access; spectrum-sharing; IoT via satellite



Citation: Munari, A.; Clazzer, F. Spectral Coexistence of QoS-Constrained and IoT Traffic in Satellite Systems. *Sensors* **2021**, *21*, 4630. <https://doi.org/10.3390/s21144630>

Academic Editors: Riccardo De Gaudenzi and Beatriz Soret

Received: 18 May 2021

Accepted: 28 June 2021

Published: 6 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Massive machine-type communications (MMTC) and the Internet of Things (IoT) are attracting steadily growing research and industry attention, emerging as a fundamental component for next-generation wireless systems. Driven by a blooming number of IoT applications, this novel communications paradigm aims at serving vast populations of often low-power, low-complexity terminals that generate sporadic traffic in the form of short data packets. Examples of practical relevance span a wide set of scenarios, ranging from smart agriculture or industry, where sensors may collect data (e.g., temperature, pressure, and presence of chemical substances) and deliver status information to a common gateway, to environmental monitoring or asset tracking [1–3].

Support for this multitude of use cases is already provided by terrestrial networks in a number of well-established commercial products [4], e.g., LoRa [5–7], SigFox [8], Ingenu [9], as well as by standardized approaches, such as NB-IoT and LTE-M [10]. In parallel to this, satellite-based solutions have recently gained traction as a key enabler to provide global coverage to mMTC services [11]. Favored, among other factors, by the significant reduction in launch cost and by the use of off-the-shelves components, such as reprogrammable software defined radios (SDRs), a revived interest toward the deployment of low-Earth orbit (LEO) satellites has characterized the past few years. From this viewpoint, large constellations such as Starlink, OneWeb or the planned Amazon Kuiper are flanked by a growing number of smaller LEO networks focusing on specialized commercial services, e.g., [12–14]. Satellite IoT connectivity is one of the key scenarios included in the non-terrestrial networks (NTN) standardization efforts within the 3GPP ecosystem, and is aimed to become part of the standard already from Release 17.

The increasing interest toward mMTC has spurred significant research efforts to tackle a number of challenges that span the whole communications protocol stack. In particular, the intermittent transmission of short packets from IoT terminals calls for innovative approaches. At the physical layer, for instance, the design of the channel codes that operate efficiently over blocks in the order of a few tens to hundreds of bits and with possibly limited channel state information is fundamental [15]. Moreover, the grant-based medium access control policies encountered in traditional human-centric communication systems is largely ineffective for the traffic profile encountered in machine-type applications. As a matter of fact, the overhead needed to coordinate resource sharing in the presence of a massive population of sporadically active transmitters is highly inefficient. In this perspective, a flourishing line of research is emerging, with the proposal of a number of novel *modern* random access protocols [16–19]. These solutions lean on a joint design of coding and medium access, allowing to achieve spectral efficiencies comparable to those of coordinated schemes under a truly grant-free paradigm, and offer a promising way forward for next-generation mMTC systems. At the same time, random access is already at the core of current IoT communications, as variations of the basic ALOHA strategy [20] are employed by many widely used mMTC solutions [5,8,11].

From this standpoint, grant-free and coordinated access policies offer complementary characteristics, and are commonly employed side by side. Indeed, many communications systems host and provide support to a variety of use cases, ranging from sporadic and often lower-priority mMTC traffic to services with more stringent demands in terms of data rate, reliability and quality of service (QoS). This combination is typically achieved by assigning blocks of orthogonal resources (slices)—in time, frequency, or a combination thereof—to applications with distinct requirements. Relevant examples in this direction are the ETSI S-MIM standard for S-band mobile interactive multimedia [21], focusing on the satellite uplink, or the mobile communications standard 5G from 3GPP [22]. In the former case, the transmission of regular or high data rate traffic is served via demand-assignment procedures, whereas delivery of IoT messages can be attempted on dedicated resource blocks via a variation of spread spectrum ALOHA [23]. Similarly, in 5G, a large amount of the spectrum is dedicated to the enhanced mobile broadband (eMBB) traffic, while IoT messages are sent via the NB-IoT waveform in dedicated sub-bands or unused guard bands [24].

By construction, such approaches avoid interference between different services, easing the provision of proper QoS levels and simplifying waveform design. On the other hand, the reservation of orthogonal sets of resources may not be fully efficient in the presence of machine-type communications. In fact, the unpredictable nature of mMTC traffic inherently leads to significant load fluctuations, hindering the precise tuning of the amount of bandwidth to be allocated and often resulting in an either heavily congested or under-utilized channel. From a different angle, instead, a controlled level of interference may be tolerated by non-mMTC services without violating the target QoS requirements. Indeed, information theoretic results on the multiple access channel show that spectrum sharing and interference cancellation allow to achieve the corner points of the capacity region, suggesting that the coexistence of the two services over a shared band may be beneficial.

Starting from these remarks, we explore in this work the possibility to serve different types of traffic over the same set of resources concurrently. Specifically, we tackle the need to provide uplink support to traffic with a specific QoS target as well as to enable transmission for a best-effort service to IoT devices. Leaning on information theoretic tools, we derive the maximum aggregate rate that can be granted to mMTC traffic without violating the rate and reliability requirements of the coexisting service. We then compare the performance of such an *overlay* configuration to a benchmark setup where resources are orthogonally split, highlighting promising gains.

1.1. Related Work

The first research direction considered spectrum coexistence for geographically spaced systems [25,26]. More recently, the idea of sharing bandwidth among non-homogeneous services has started to receive some attention in the context of 5G cellular systems, aiming to go beyond the inefficiencies of orthogonal slicing. In [27], the possibility to multiplex eMBB and ultra-reliable low-latency communications (URLLC) was studied following an information theoretic approach, pinpointing the potential of the idea in a multi-cell cloud radio access network (C-RAN). Additional insights are provided in [28], where the two aforementioned services are flanked by mMTC traffic. Among other scenarios, the work studied a heterogeneous non-orthogonal multiple access solution for the OFDM 5G uplink, where resources are shared between eMBB and mMTC with different reliability requirements. Assuming packets that are long enough to justify an asymptotic information-theoretic analysis, the authors explored via numerical solutions some key trade-offs in terms of rates that can be granted to the distinct services, showing how an overlay allocation can be beneficial in certain regimes. Providing efficient connectivity while guaranteeing the target QoS among several concurrent services was also thoroughly investigated above layer two, e.g., [29]. Finally, a survey on inter-system spectrum sharing focusing on services with equal access rights to the resources can be found in [30].

1.2. Main Contributions and Structure of the Paper

Within this line of study, we aim to shed light on the potential of a non-orthogonal resource distribution, as well as to trigger additional research on the topic in a satellite-IoT context. In particular, in our work, we do the following:

- We characterize some fundamental trends for the performance of an overlay allocation of a QoS-constrained service and of mMTC traffic, composed of short packets. Based on practical considerations, we assume the receiver to first attempt decoding the former traffic. In the case of success, its contribution is removed (interference cancellation), and retrieval of the underlying mMTC packets is attempted;
- In this setup, we derive the exact analytical expressions for the maximum aggregate rate that can be granted to mMTC traffic as a function of the requirements in terms of the rate and packet error rate set for the QoS-constrained service;
- Going beyond the approach followed in [28], the analysis relies on information theoretic arguments to capture the impact of short packets transmission, as well as the impact of the length of frames in which the uplink communication is organized.

The remainder of this paper is structured as follows. After introducing the system model in Section 2, we provide in Section 3 the initial insight by comparing the performance of the overlay and orthogonal allocations in a setting characterized by packets long enough to justify the use of asymptotic information theoretic tools, considering both an ergodic and a non-ergodic case. The study is then complemented in Section 4 by exploring the impact of transmission of short packets commonly encountered in mMTC applications, relying on the normal approximation [31,32]. The numerical results are presented and discussed in Section 5, highlighting some fundamental trends and comparing the effectiveness of orthogonal and overlay allocations in all of the considered cases. Finally, Section 6 draws the conclusions, offering some relevant open issues and future research directions.

1.3. Notation

We denote random variables (vectors) by uppercase (bold) letters, while we refer to their realizations in lowercase, e.g., X and x ; $\mathbf{X}$ and $\mathbf{x}$. The probability mass function (PMF) of a discrete r.v. is denoted as $p_X(x) := \mathbb{P}\{X = x\}$ and that of a discrete random vector as $p_{\mathbf{X}}(\mathbf{x}) := \mathbb{P}\{\mathbf{X} = \mathbf{x}\}$. Moreover, we indicate the conditional PMFs $\mathbb{P}\{X = x | Y = y\}$ as $p_X(x|y)$ and $\mathbb{P}\{\mathbf{X} = \mathbf{x} | Y = y\}$ as $p_{\mathbf{X}}(\mathbf{x}|y)$. Finally, the expectation operator is denoted as $\mathbb{E}[\cdot]$, while $\mathbf{I}_k$ is the identity matrix of size k .

2. System Model and Preliminaries

Throughout our discussion, we focus on the uplink of a wireless satellite system which supports two distinct services, labeled as $\mathcal{S}_a$ and $\mathcal{S}_b$. The former grants specific requirements in terms of data-rate and error probability, and is embodied in our setting by a single user, s_a , which sends data to the receiver under an average power constraint, ρ_a whenever it is granted access to the medium. On the other hand, $\mathcal{S}_b$ foresees that a large number of terminals share the wireless channel under a slotted ALOHA contention policy [33] to attempt packet delivery in a best-effort fashion, e.g., for mMTC. Specifically, over any time slot allotted to $\mathcal{S}_b$, a variable number of users U access the channel, each transmitting toward the receiver for the whole slot duration with average power ρ_b . We note that, while the single user for the transmission of a data packet is subject to an average power constraint ρ_b , the *overall* power injected in the system by service $\mathcal{S}_b$ varies both with the number of terminals accessing the channel over a single slot and the number of allocated slots. Such a working assumption is representative of the practical implementations of machine-type communications. Following a common approach, U is modeled as a Poisson r.v. of parameter λ , independent and identically distributed across slots, so that the following holds:

$$p_U(u) = \frac{\lambda^u e^{-\lambda}}{u!}.$$

The assumption is especially representative for traffic generated by a multitude of low duty cycle nodes monitoring and/or reporting data generated by heterogeneous systems and phenomena [34], as often encountered in mMTC satellite systems. Furthermore, we consider for $\mathcal{S}_b$ no feedback on the outcome of sent packets nor retransmission policies. This approach is in line with the best-effort nature of many IoT applications, where packets losses may be tolerated, awaiting the delivery of a successive update from the monitored devices.

To characterize the performance of the system, we tackle two operation modes, considering an *orthogonal* and an *overlay* allocation for $\mathcal{S}_a$ and $\mathcal{S}_b$. In both cases, having in mind the uplink of a satellite system, we assume for the link between a transmitter and the receiver a line-of-sight connection with perfect power control. Accordingly, the channel coefficient at the receiver is a constant and known value for all users, assumed to be unitary for the sake of simplicity, and all users of a given service are received with the same power level.

2.1. Orthogonal Allocation

In this configuration, resources are split between services so that no mutual interference between $\mathcal{S}_a$ and $\mathcal{S}_b$ arises. Without loss of generality, we assume orthogonality to be achieved in time and focus on the setup exemplified by Figure 1a (the reported analysis and results would hold for an orthogonal allocation in the frequency domain as well). Time is divided into successive frames of equal duration. Within each frame, a fraction α of the time is granted exclusively to s_a for data delivery. During this period, the channel output at the receiver over the ℓ -th channel use takes the following form:

$$Y_\ell = X_\ell + W_\ell$$

where $X_\ell \sim \mathcal{CN}(0, \rho_a/\alpha)$ is the complex symbol transmitted by the user, and $W_\ell \sim \mathcal{CN}(0, \sigma^2)$ is a zero mean, σ^2 variance, circularly-symmetric complex Gaussian additive noise component. In view of the fraction α of resources available for transmission, the average signal-to-noise ratio (SNR) at the receiver is thus, as follows:

$$\gamma_a := \frac{\rho_a}{\alpha\sigma^2}. \quad (1)$$

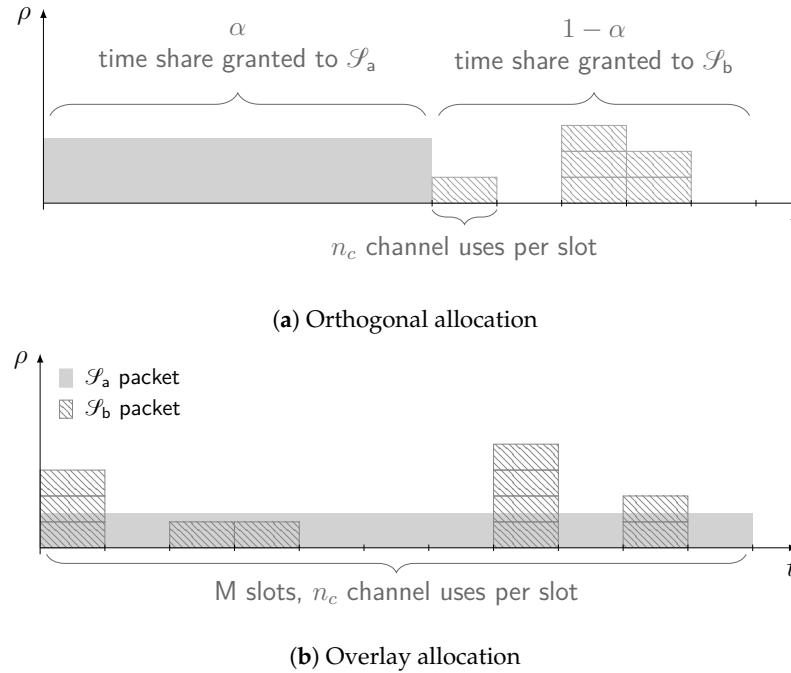


Figure 1. Reference timeline for the orthogonal and overlay resource allocations.

As for users of the best-effort service $\mathcal{S}_b$, channel access is only permitted during the remaining fraction $(1 - \alpha)$ of the time. To instantiate a slotted ALOHA policy, the corresponding portion of the frame is assumed to be divided in slots of equal duration, each allowing the transmission of a single packet of $\mathcal{S}_b$. Denoting by n_s the number of channel uses over a slot, the input–output relation over the ℓ -th slot conditioned on having $U = u$ users concurrently transmitting can be expressed as follows:

$$\mathbf{Y}_\ell = \sum_{j=1}^u \mathbf{X}_\ell^{(b,j)} + \mathbf{W}_\ell. \quad (2)$$

In (2), $\mathbf{X}_\ell^{(b,j)}$ is the codeword transmitted by the j -th user of $\mathcal{S}_b$ active in slot ℓ , whose n_s components are modeled as i.i.d. circularly-symmetric normal r.v. with zero mean and variance ρ_b , i.e., $\mathbf{X}_\ell^{(b,j)} \sim \mathcal{CN}(\mathbf{0}, \rho_b \mathbf{I}_{n_s})$. In turn, $\mathbf{W}_\ell \sim \mathcal{CN}(\mathbf{0}, \sigma^2 \mathbf{I}_{n_s})$ is the additive Gaussian noise, leading to the incoming signal vector at the receiver, $\mathbf{Y}_\ell \in \mathbb{C}^{n_s}$. Note that when $U = 0$, no user of $\mathcal{S}_b$ transmits over the slot, and the summation in (2) brings no contribution so that solely the noise component $\mathbf{W}_\ell$ is observed. As discussed in Section 1, this event, occurring with probability $e^{-\lambda}$, represents a waste of resources, as bandwidth granted to mMTC is not employed to attempt data delivery.

2.2. Overlay Allocation

When the system is operated in overlay mode, both services are allowed to access the channel concurrently and without time restrictions, as illustrated in Figure 1b. Such a policy allows $\mathcal{S}_a$ and $\mathcal{S}_b$ to enjoy the whole share of resources at the cost of suffering from mutual interference, due to the loss of orthogonality. In this case, the whole frame is assumed to be split in M slots, each consisting of n_s channel uses. Using all the resources, s_a transmits then a codeword $\mathbf{X}^{(a)}$ of size $n_s M$ channel uses. In turn, users of service $\mathcal{S}_b$ can send a packet over any of the M available slots. To properly capture the signal model, although s_a encodes its message across all the available channel uses, it is convenient to express $\mathbf{X}^{(a)}$ as the concatenation of the M sub-codewords transmitted over the slots as follows:

$$\mathbf{X}^{(a)} = [\mathbf{X}_1^{(a)}, \dots, \mathbf{X}_M^{(a)}]$$

where $\mathbf{X}_\ell^{(a)} \sim \mathcal{CN}(\mathbf{0}, \rho_a \mathbf{I}_{n_s})$, $\ell = 1, \dots, M$. Following this notation, the channel output at the receiver over the ℓ -th slot conditioned on having $U = u$ users of $\mathcal{S}_b$ access the channel takes the following form:

$$\mathbf{Y}_\ell = \mathbf{X}_\ell^{(a)} + \sum_{j=1}^u \mathbf{X}_\ell^{(b,j)} + \mathbf{W}_\ell. \quad (3)$$

A packet of a best-effort user is thus always affected by the interference coming from $\mathcal{S}_a$, and possibly by other transmissions of $\mathcal{S}_b$. Conversely, any sub-codeword sent by s_a is received interference-free with probability $e^{-\lambda}$, i.e., if no mMTC user has transmitted over the corresponding slot. It is worth stressing that the number of interfering packets from $\mathcal{S}_b$ is not known a priori, so the coding rate of s_a cannot be dynamically adapted on a slot-by-slot basis.

For both the orthogonal and overlay configurations, we are interested in the maximum sum rate that the system can offer to the best-effort service while granting the QoS requirements of $\mathcal{S}_a$. Specifically, denoting by R_a the rate in bits per channel use of s_a and by R_b the average aggregate number of bits per channel use decoded at the receiver for $\mathcal{S}_b$, we aim at deriving the rate region given by all pairs (R_a, R_b) that allow s_a to experience an error probability that is less than or equal to a target value.

3. Asymptotic Analysis

In order to gather preliminary insight on the performance of the system, we first consider an asymptotic setting in which the codewords transmitted by s_a , as well as those by users of $\mathcal{S}_b$ over a slot, can be made long enough to approach the *classical* information theoretic results.

3.1. Orthogonal Allocation

When resources are orthogonally split among services, s_a delivers data over an additive white gaussian noise (AWGN) channel with SNR γ_a defined in (1), yet is allowed to transmit only for a fraction α of the time. In this setting, the user can then communicate with vanishing small error probability for any rate as follows:

$$R_a < \alpha \log_2(1 + \gamma_a). \quad (4)$$

Let us focus instead on the share granted to $\mathcal{S}_b$, and indicate as $R_{b,\ell}$ the communication rate achieved by the service over the ℓ -th slot. If no user accesses the channel (i.e., $U = 0$), then clearly $R_{b,\ell} = 0$. Conversely, if a single transmission is performed ($U = 1$, with probability $\lambda e^{-\lambda}$), data are sent over an AWGN channel with capacity as follows:

$$C_b = \log_2\left(1 + \frac{\rho_b}{\sigma^2}\right). \quad (5)$$

Finally, when more than one user becomes active over the slot, a *collision* takes place, and the channel output at the receiver takes the general form of (2). As will be discussed later (see Remark 1), we assume terminals of $\mathcal{S}_b$ to encode information at a rate approaching C_b , as would be done in the absence of interference. Accordingly, in the event of a collision, the actual channel capacity falls below the employed rate, and we regard all packets sent over the slot of interest to be lost. In other words, following a common modeling approach for random access schemes (e.g., [20]), collisions are of a *destructive* nature, and $R_{b,\ell} = 0$ whenever more than one user accesses the channel concurrently ($U > 1$). In summary, (i) for $U \neq 1$, $R_{b,\ell} = 0$, and (ii) for $U = 1$, $R_{b,\ell} = C_b$. Combining these observations, the average rate for $\mathcal{S}_b$ in bits per channel use over a slot readily follows:

$$\mathbb{E}[R_{b,\ell}] = \lambda e^{-\lambda} C_b. \quad (6)$$

Equation (6) clarifies how, for any given power level ρ_b , the slotted ALOHA best-effort channel shall be operated at a load $\lambda = 1$ (pkt/slot) in order to maximize the average

aggregate rate. Recalling that only a fraction $(1 - \alpha)$ of the resources can be leveraged, the achievable sum-rates for $\mathcal{S}_b$ are then characterized as follows:

$$R_b < (1 - \alpha) e^{-1} \cdot \log_2 \left(1 + \frac{\rho_b}{\sigma^2} \right). \quad (7)$$

Remark 1. The widely-employed assumption of destructive collisions reflected in (6) is of practical relevance for the considered mMTC setting. Indeed, given that the number of competing terminals over a slot cannot generally be predicted before transmission, a coding scheme that effectively and dynamically takes into account interfering packets may be difficult to develop and goes beyond the complexity of the available terminals. On the other hand, a backoff on the transmission rate to increase resiliency to interference may not be efficient in lightly loaded channels, where sporadic access may lead to having many slots without contention, resulting in significance performance loss for the tight link budget configurations that are typical of mMTC links. Finally, we note that many machine-type applications attempt the delivery of updates that are often repeated or very much correlated in content over time (e.g., reporting of sensed data), positively trading off higher packet loss rates due to interference for a larger aggregate channel sum rate achieved in the absence of coordination among nodes.

3.2. Overlay Allocation

When the system is operated in overlay configuration, the incoming signal at the receiver is the superposition of transmissions performed by both services. Throughout our study, we assume decoding to start from the message sent by s_a , treating the interference component of $\mathcal{S}_b$ packets as noise. In the case of successful decoding, ideal interference cancellation is performed, removing the contribution of s_a from the overall waveform and presenting to the receiver the M-slot frame populated by packets of $\mathcal{S}_b$ alone for further processing. Ideal interference cancellation of the packet of user s_a can be regarded as a reasonable assumption, especially taking into account that its transmission can enjoy a large number of channel uses, and the data-aided channel estimation can be properly tuned. Moreover, we also note that non-ideal interference cancellation would yield a minor impact on the model by effectively increasing the noise power suffered by the transmissions of service $\mathcal{S}_b$ since both the signal and the noise are modeled as Gaussian distributed r.v.s..

Remark 2. The assumption on the order of decoding for the supported services stems from two practical considerations. First, as will be further discussed in Sections 4 and 5, mMTC traffic is typically composed of very short packets—often in the order of a few hundreds of bits—confining pilot symbols to few channel uses per transmitted message in order to avoid excessive overhead. Accurate channel estimation needed for decoding packets of $\mathcal{S}_b$, especially in the presence of interference from s_a , may thus not be viable. On the other hand, the longer data units that characterize $\mathcal{S}_a$ render the cost of stronger pilot sequences manageable, and, together with its more regular and predictable traffic patterns, can lead to estimates of the channel parameters that are good enough to retrieve information and perform accurate interference cancellation. In addition to this, the choice of initiating decoding from $\mathcal{S}_a$ is driven by QoS arguments. It is indeed reasonable to overlay best-effort mMTC traffic to an existing service only as long as the latter does not experience a significant loss in performance. Along this line of reasoning, the receiver shall be able to decode $\mathcal{S}_a$ even in the presence of underlying interference, i.e., treating the whole of $\mathcal{S}_b$ traffic as noise.

Under this assumption, s_a experiences the block-interference channel described in (3), with the interference level remaining constant over a coherence period of one slot duration, and changing to an independent realization during the subsequent n_s channel uses. Specifically, conditioned on the number U_ℓ of users of $\mathcal{S}_b$ accessing the channel over the ℓ -th slot, the r.v. describing the signal-to-interference and noise ratio (SINR) seen at the receiver for s_a takes the following form:

$$\text{SINR}_\ell^{(a)} = \frac{\rho_a}{\sigma^2 + U_\ell \rho_b}.$$

Within this framework, to better grasp the effect of the bursty overlaid mMTC traffic, we first analyze an ergodic setting, later extending our results to the more practically-relevant non-ergodic setup.

3.2.1. Ergodic Case

When the number of slots within a frame is sufficiently large (i.e., $M \rightarrow \infty$), the codeword of s_a experiences, in the limit, all the possible interference values, making it possible to define the ergodic channel capacity.

$$\begin{aligned} C_a^{(e)} &= \mathbb{E} \left[\log_2 \left(1 + \frac{\rho_a}{\sigma^2 + U\rho_b} \right) \right] \\ &= \sum_{u=0}^{\infty} \frac{\lambda^u e^{-\lambda}}{u!} \log_2 \left(1 + \frac{\rho_a}{\sigma^2 + u\rho_b} \right). \end{aligned} \quad (8)$$

In this case, a vanishingly small error probability can still be granted to the user for any rate $R_a < C_a^{(e)}$. Clearly, from a system design perspective, the overlay operation mode becomes meaningful—and possibly convenient—if the QoS requirements of $\mathcal{S}_a$ in terms of data rate are met, even in the presence of interference. We are thus interested in configurations for which the achievable rate in overlay matches the one of an orthogonal allocation, i.e., for which, leaning on (4) and (8),

$$\alpha \log_2(1 + \gamma_a) = \sum_{u=0}^{\infty} \frac{\lambda^u e^{-\lambda}}{u!} \log_2 \left(1 + \frac{\rho_a}{\sigma^2 + u\rho_b} \right). \quad (9)$$

In particular, for any value of α employed to configure a dedicated-resource allocation, (9) imposes a constraint on the tolerable level of interference from $\mathcal{S}_b$ in overlay mode, limiting the average number of transmissions per slot and thus the intensity of the mMTC traffic. Denoting this value as λ_m , the message of s_a is retrieved and its contribution on the incoming waveform is perfectly canceled with probability approaching one for any channel load $\lambda < \lambda_m$ of service $\mathcal{S}_b$. We note that, while a closed-form expression of λ_m is elusive due to the transcendental nature of (9), its value can easily be determined numerically. If the system is operated satisfying this constraint, the receiver can then process packets sent by mMTC users via slotted ALOHA after having performed interference cancellation, and the average aggregate rate for $\mathcal{S}_b$ is once again captured by (6). In order to maximize performance while meeting the requirements of the QoS-driven user, $\mathcal{S}_b$ shall thus be operated at a channel load as follows:

$$\lambda^* := \min\{\lambda_m, 1\}.$$

We recall indeed that values of λ larger than 1 [pkt/slot], even if tolerable by the QoS requirements of $\mathcal{S}_a$, would be ineffective from a throughput perspective for the mMTC traffic in view of the detrimental effect of destructive collisions.

Combining these remarks, and recalling that both services have now access to the whole set of resources, the rate region (R_a, R_b) for the ergodic case finally evaluates to the following:

$$\begin{aligned} R_a &< \sum_{u=0}^{\infty} \frac{\lambda_m^u e^{-\lambda_m}}{u!} \log_2 \left(1 + \frac{\rho_a}{\sigma^2 + u\rho_b} \right) \\ R_b &< \lambda^* e^{-\lambda^*} \log_2 \left(1 + \frac{\rho_b}{\sigma^2} \right). \end{aligned}$$

3.2.2. Non-Ergodic Case

Albeit insightful in pinpointing a fundamental trade-off between the achievable rate for $\mathcal{S}_a$ and the amount of mMTC traffic that can be served, the ergodic setup fails to capture some important aspects of practical relevance. In particular, the assumption of

having an asymptotically large number of slots over which user s_a can encode its message can become highly inaccurate for typical systems, characterized by frames of a length of few hundreds of slots, e.g., [21]. In such conditions, each codeword of service $\mathcal{S}_a$ experiences a finite number M of interference realizations, and a vanishingly small error probability can no longer be granted. The QoS requirements of the service are therefore more appropriately characterized by a rate-reliability pair, where R_a is complemented by a tolerable error probability for the sent codeword. In the non-ergodic setting, the latter is captured using information theory tools by the outage probability P_{out} . Specifically, for the block-interference channel under study, the two quantities are related as follows:

$$P_{\text{out}} = \mathbb{P} \left\{ \frac{1}{M} \sum_{\ell=1}^M C_{a,\ell} < R_a \right\} \quad (10)$$

where $C_{a,\ell}$ denotes the instantaneous capacity for s_a over the ℓ -th slot, and can be expressed as a function of the number U_ℓ of $\mathcal{S}_b$ users transmitting over the slot as follows:

$$C_{a,\ell} := \log_2 \left(1 + \frac{\rho_a}{\sigma^2 + U_\ell \rho_b} \right). \quad (11)$$

Due to the discrete levels of interference that can be experienced, the r.v. $C_{a,\ell}$ takes values in the alphabet as follows:

$$\mathcal{A} = \left\{ c_\ell \in \mathbb{R} \mid c_\ell = \log_2 \left(1 + \frac{\rho_a}{\sigma^2 + u \rho_b} \right), u \in \mathbb{N} \right\}$$

and its PMF can readily be derived from (11), leading to the following:

$$p_{C_{a,\ell}}(c_\ell) = \mathbb{P} \left\{ u = \frac{\rho_a}{\rho_b(2^{c_\ell} - 1)} - \frac{\sigma^2}{\rho_b} \right\} = \frac{e^{-\lambda} \lambda^{\beta(c_\ell)}}{\beta(c_\ell)!} \quad \forall c_\ell \in \mathcal{A} \quad (12)$$

where the final expression resorts to the ancillary quantity as follows:

$$\beta(c_\ell) := \frac{\rho_a}{\rho_b(2^{c_\ell} - 1)} - \frac{\sigma^2}{\rho_b}.$$

Leaning on this, let us define for compactness the r.v. as follows:

$$V := \sum_{\ell=1}^M C_{a,\ell}.$$

taking values in the set $\mathcal{A}_V = \left\{ v \in \mathbb{R} \mid v = \sum_{\ell=1}^M c_\ell, c_\ell \in \mathcal{A} \right\}$. Observing that the instantaneous capacity values over different slots are i.i.d., the PMF of V can simply be computed as the M -fold convolution of (12), to obtain for any $v \in \mathcal{A}_V$ the following:

$$p_V(v) = \underbrace{(p_{C_{a,\ell}} * p_{C_{a,\ell}} * \cdots * p_{C_{a,\ell}})}_M(v).$$

Accordingly, an exact expression of the outage probability in (10) can finally be derived as follows:

$$P_{\text{out}} = \mathbb{P}\{V < MR_a\} = \sum_{\substack{v \in \mathcal{A}_V \\ v < MR_a}} p_V(v). \quad (13)$$

This result offers a useful system design tool. Indeed, for any pair (R_a, P_{out}) , (12) and (13) provide an exact characterization of the maximum channel load λ_m that can be granted to mMTC traffic without violating the QoS requirements of $\mathcal{S}_a$.

Taking such a constraint into account, let us now focus on $\mathcal{S}_b$, aiming to derive the aggregate rate R_b that can be achieved for the maximum arrival intensity sustainable by $\mathcal{S}_a$. In this perspective, we observe that two conditions have to be met for an mMTC packet to be retrieved over a slot in a non-ergodic setup: (i) the packet of s_a is decoded, and (ii) no other user of $\mathcal{S}_b$ concurrently accesses the channel over the slot of interest. Note that condition (i) is consistent with the system model assumptions. Indeed, should the message of s_a not be decoded, its interference contribution cannot be removed, affecting any underlying mMTC packet. Recalling the discussion of Remark 1, this would prevent the successful retrieval of information for users of $\mathcal{S}_b$. Furthermore, let us denote by W the r.v. taking values in $\{0, \dots, M\}$ and counting the number of slots characterized by the transmission of a single user (named also *singleton slots*) of $\mathcal{S}_b$, which, for the Poisson traffic under study, follows a binomial distribution of parameters M and $\lambda e^{-\lambda}$. Following this notation, for any arrival intensity λ , the aggregate rate of $\mathcal{S}_b$ can be conveniently expressed as follows:

$$R_b = \frac{1}{M} \sum_{w=0}^M \left[w \log_2 \left(1 + \frac{\rho_b}{\sigma^2} \right) \cdot \left(1 - \mathbb{P} \left\{ \sum_{\ell=1}^M C_{a,\ell} < MR_a \mid W = w \right\} \right) \right] \cdot p_W(w) \quad (14)$$

where, recalling (10), the term within square brackets takes into account how all w singleton slots bring a contribution of $\log_2(1 + \rho_b/\sigma^2)$ bits per channel use if, and only if, s_a is not in outage. To complete the calculation, we therefore need to derive the outage probability for $\mathcal{S}_a$ conditioned on the r.v. W . To this aim, it is useful to rearrange the conditional expression as follows:

$$\mathbb{P} \left\{ \sum_{\ell=1}^M C_{a,\ell} < MR_a \mid W = w \right\} = \mathbb{P} \left\{ w \log_2 \left(1 + \frac{\rho_a}{\sigma^2 + \rho_b} \right) + \sum_{\ell, U_\ell \neq 1} C_{a,\ell} < R_a M \mid W = w \right\} = \quad (15)$$

$$\mathbb{P} \left\{ \sum_{\ell, U_\ell \neq 1} C_{a,\ell} < R_a M - w \log_2 \left(1 + \frac{\rho_a}{\sigma^2 + \rho_b} \right) \mid W = w \right\} \quad (16)$$

where the first addend on the right-hand side of (15) accounts for the sum of the instantaneous capacity of all w slots with a single transmission of service $\mathcal{S}_b$, whereas the summation takes into consideration the contributions of all other $M - w$ slots (i.e., those with $U_\ell \neq 1$). In turn, the PMF of the instantaneous capacity $C_{a,\ell}$ over a *non-singleton* slot can be directly derived from (12), to obtain for any value in the alphabet $\{c_\ell \in \mathbb{R} \mid c_\ell = \log_2(1 + \frac{\rho_a}{\sigma^2 + u\rho_b}), u \in \mathbb{N} \setminus \{1\}\}$,

$$\mathbb{P}\{C_{a,\ell} = c_\ell \mid U_\ell \neq 1\} = \frac{e^{-\lambda} \lambda^{\beta(c_\ell)}}{\beta(c_\ell)!} \frac{1}{1 - \lambda e^{-\lambda}}. \quad (17)$$

where the normalization factor readily follows from the Poisson distribution of the mMTC traffic. Recalling now that also the r.v. $C_{a,\ell}$ over non-singleton slots are i.i.d., the PMF of the sum of $M - w$ such values can be once more derived by taking the $(M - w)$ -fold convolution of (17), allowing to compute (16). Leaning on this result, and plugging the binomial distribution of W into (14), R_b can be finally computed for any value of λ as follows:

$$R_b = \frac{1}{M} \log_2 \left(1 + \frac{\rho_b}{\sigma^2} \right) \sum_{w=0}^M w \binom{M}{w} (\lambda e^{-\lambda})^w (1 - \lambda e^{-\lambda})^{M-w} \times \left(1 - \mathbb{P} \left\{ \sum_{\ell, U_\ell \neq 1} C_{a,\ell} < R_a M - w \log_2 \left(1 + \frac{\rho_a}{\sigma^2 + \rho_b} \right) \mid W = w \right\} \right). \quad (18)$$

The presented approach provides thus the sought pairs of sustainable rates. As discussed, the maximum sustainable arrival intensity λ_m for $\mathcal{S}_b$ can be computed from (13) for any target QoS requirements (R_a, P_{out}) . In turn, mMTC traffic can be operated effectively at any channel load lower than $\lambda^* = \min\{1, \lambda_m\}$, with the corresponding aggregate rate R_b obtained by evaluating (18) for $\lambda = \lambda^*$.

4. Finite Blocklength Analysis

Going beyond the asymptotic setting discussed in Section 3, we complement our study by delving into a more practical scenario that closely relates to mMTC applications, where packets transmitted by users can be in the order of few hundreds bits. To capture this aspect we leverage tools of finite-length information theory, and focus in particular on the normal approximation [31,32] to characterize the rates achievable by $\mathcal{S}_a$ and $\mathcal{S}_b$.

4.1. Orthogonal Allocation

Let us first focus on the orthogonal configuration, and denote by n the (finite) number of channel uses available over the whole transmission frame. Following the notation of Section 2, the transmission of s_a spans $n_a := \alpha n$ channel uses, and takes place over an AWGN channel of SNR γ_a defined in (1), i.e., we implicitly focus on values of α such that $\alpha n \in \mathbb{N}$. In this case, a vanishingly small error probability cannot be granted, even in the absence of interference, and the performance of $\mathcal{S}_a$ is properly described by the couple (R_a^*, P_e) , where R_a^* denotes the maximum rate that can be supported for a codeword error probability P_e . Specifically, the two quantities are related as follows [31]:

$$R_a^*(P_e) = C_a - \sqrt{\frac{\mathcal{V}_a}{n_a}} Q^{-1}(P_e) + \mathcal{O}\left(\frac{\log n_a}{n_a}\right) \quad (19)$$

where the channel capacity C_a and channel dispersion $\mathcal{V}_a$ are defined by the following expressions:

$$C_a = \log_2(1 + \gamma_a)$$

$$\mathcal{V}_a = \frac{\gamma_a(2 + \gamma_a)}{2(1 + \gamma_a)^2} (\log_2 e)^2$$

and $Q^{-1}(\cdot)$ denotes the inverse Q function. If we now recall that the user s_a transmits only for a fraction α of the time, and approximates (19) by disregarding the terms of order $\mathcal{O}(\log n_a/n_a)$, the set of rates achievable by $\mathcal{S}_a$ for a target error probability P_e^* can be characterized as follows:

$$R_a < \alpha \left(C_a - \sqrt{\frac{\mathcal{V}_a}{n_a}} Q^{-1}(P_e^*) \right). \quad (20)$$

Consider now service $\mathcal{S}_b$. In line with what was discussed in Section 3, collisions are once more regarded as destructive, so slots in which more than one user transmits lead to the loss of all the packets sent. In the setup under study, however, a non-vanishing error probability is to be expected for any transmission rate R_b even over a *singleton* slot, in view of the finite number n_s of available channel uses. Accordingly, introducing the channel dispersion,

$$\mathcal{V}_b = \frac{\frac{\rho_b}{\sigma^2} \left(2 + \frac{\rho_b}{\sigma^2} \right)}{2 \left(1 + \frac{\rho_b}{\sigma^2} \right)^2} (\log_2 e)^2.$$

and recalling the channel capacity $\mathcal{C}_b$ in (5), the maximum rate R_b^* that can be granted to $\mathcal{S}_b$ under a codeword error probability ϵ can be approximated as follows:

$$R_b^*(\epsilon) \cong \mathcal{C}_b - \sqrt{\frac{\mathcal{V}_b}{n_s}} Q^{-1}(\epsilon). \quad (21)$$

In terms of system design, (21) allows then to tune the rate-reliability requirements of $\mathcal{S}_b$. From this standpoint, we assume within our study mMTC traffic to be of the best-effort nature, i.e., without stringent requirements in terms of reliability, and select the transmission rate in order to maximize the attained spectral efficiency. Specifically, we consider $\mathcal{S}_b$ to be operated at an error rate ϵ^* that maximizes the information bits per channel use retrieved over a singleton slot:

$$\epsilon^* = \underset{\epsilon}{\operatorname{argmax}} \{ (1 - \epsilon) R_b^*(\epsilon) \}. \quad (22)$$

Finally, recalling that mMTC traffic can enjoy a share $(1 - \alpha)$ of the resources and that the fraction of singleton slots is bounded by e^{-1} (achieved for $\lambda = 1$), the set of aggregate rates achievable by $\mathcal{S}_b$ in the orthogonal configuration can be described as follows:

$$R_b < (1 - \alpha) e^{-1} (1 - \epsilon^*) R_b^*(\epsilon^*).$$

4.2. Overlay Allocation

As discussed in Section 2, when the system is operated in overlay mode, the total number of channel uses, n , available over a frame is split into M time slots of n_s channel uses each, i.e., $n = Mn_s$. Let us denote by $\mathbf{U} = [U_1, \dots, U_M]$ the random vector whose components describe the number of users of $\mathcal{S}_b$ that transmit over each of the slots. Recalling the i.i.d. Poisson distribution of the r.v. U_ℓ , the joint PMF of $\mathbf{U}$ can be expressed for any arrival intensity λ as follows:

$$p_{\mathbf{U}}(\mathbf{u}) = \prod_{\ell=1}^M \frac{\lambda^{u_\ell} e^{-\lambda}}{u_\ell!}. \quad (23)$$

Consider now the transmission of s_a , and condition on a specific realization $\mathbf{U} = \mathbf{u}$. In this case, each portion of the sent message experiences a distinct (and independent) interference level, allowing to model the problem as the transmission of an n channel uses codeword over M parallel AWGN channels. Following the approach presented in [35] in the context of block-fading, the relation between maximum sustainable rate R_a^* and codeword error probability P_e we can then be approximated as follows:

$$R_a^*(P_e, \mathbf{u}) \cong \frac{1}{M} \mathcal{C}_{a,ov}(\mathbf{u}) - \sqrt{\frac{\mathcal{V}_{a,ov}(\mathbf{u})}{n_s M^2}} Q^{-1}(P_e) \quad (24)$$

where

$$\begin{aligned} \mathcal{C}_{a,ov}(\mathbf{u}) &= \sum_{\ell=1}^M \log_2(1 + \gamma_{a,\ell}) \\ \mathcal{V}_{a,ov}(\mathbf{u}) &= \sum_{\ell=1}^M \frac{\gamma_{a,\ell}(2 + \gamma_{a,\ell})}{2(1 + \gamma_{a,\ell})^2} (\log_2 e)^2 \end{aligned}$$

and we introduce for convenience the SINR experienced by s_a over the ℓ -th slot as follows:

$$\gamma_{a,\ell} := \frac{\rho_a}{\sigma^2 + u_\ell \rho_b}.$$

Solving (24) with respect to P_e , we can easily obtain the error probability when a given rate R_a is employed by s_a and a specific realization of the interference pattern $\mathbf{U} = \mathbf{u}$ is experienced:

$$P_e(\mathbf{u}) = Q\left(\frac{\frac{C_{a,ov}(\mathbf{u})}{M} - R_a}{\sqrt{\frac{\mathcal{V}_{a,ov}(\mathbf{u})}{n_s M^2}}}\right). \quad (25)$$

We now observe that, in view of the random nature of the interference of service $\mathcal{S}_b$, it is more meaningful to evaluate the *average* error probability $\bar{P}_e$ as follows:

$$\bar{P}_e = \mathbb{E}[P_e(\mathbf{u})] = \sum_{\mathbf{u}} P_e(\mathbf{u}) p_{\mathbf{U}}(\mathbf{u}) \quad (26)$$

by employing (23) and (25). The QoS requirements of $\mathcal{S}_a$ are properly specified in this case by the pair $(R_a, \bar{P}_e)$.

In this perspective, the result in (26) offers a relevant system design tool. Indeed, for any targeted rate R_a , an inspection of the equation allows to derive the maximum traffic intensity λ_m of $\mathcal{S}_b$ that can be supported without violating the average error probability $\bar{P}_e$.

Taking the lead from this, let us then focus on $\mathcal{S}_b$ for which we need to compute the maximum aggregate rate that can be achieved under the constraint on λ_m . In this case, three conditions have to be met for a mMTC message to be retrieved: (i) the message of s_a has to be decoded and its interference canceled; (ii) the packet of $\mathcal{S}_b$ has to be sent over a singleton slot; and (iii) the n_s -channel use codeword has to be decoded. To account for all these conditions, we follow an approach similar to the one discussed in Section 3.2.2, and condition on the number of singleton slots W that take place over the observed frame, writing the aggregate average rate achieved for a traffic intensity λ as follows:

$$R_b = \frac{1}{M} \sum_{w=0}^M w R_b^*(\epsilon^*)(1 - \epsilon^*) \left(\sum_{\mathbf{u}} (1 - P_e(\mathbf{u})) p_{\mathbf{U}}(\mathbf{u} | w) \right) p_W(w). \quad (27)$$

where we assume that the coding scheme employed by $\mathcal{S}_b$ has been tuned following (22), so that each of the w singleton slots brings a contribution of $R_b^*(\epsilon^*)(1 - \epsilon^*)$ information bits per channel use. Recalling that W follows a binomial distribution of parameters $(M, \lambda e^{-\lambda})$, the complete evaluation of (27) simply requires to specify the joint PMF of the number of users of $\mathcal{S}_b$ that transmitted over each of the M slots, conditioned on having $W = w$ singleton ones. The distribution can be computed by considering two cases. First, for any vector $\mathbf{u}$ whose number of components with value 1 (i.e., the number of singleton slots in the considered frame realization) is different from w , we clearly have $p_{\mathbf{U}}(\mathbf{u} | w) = 0$. For any other $\mathbf{u}$, instead.

$$p_{\mathbf{U}}(\mathbf{u} | w) = \frac{\prod_{\ell=1}^M \frac{\lambda^{u_\ell} e^{-\lambda}}{u_\ell!}}{\binom{M}{w} (1 - \lambda e^{-\lambda})^{M-w} (\lambda e^{-\lambda})^w} = \frac{\prod_{\ell=1}^{M-w} \frac{\lambda^{u_\ell} e^{-\lambda}}{u_\ell!}}{\binom{M}{w} (1 - \lambda e^{-\lambda})^{M-w}}$$

where the numerator follows from the i.i.d. Poisson distribution of the number of transmissions over a slot, whereas the normalization factor accounts for the probability of having exactly w singleton slots out of the available M .

In conclusion, the set of admissible rates for $\mathcal{S}_b$ can be computed by taking into account the constraint λ_m imposed by the QoS requirements of $\mathcal{S}_a$, and by evaluating (27) for all traffic intensities smaller than $\lambda^* = \min\{1, \lambda_m\}$.

5. Numerical Results

In this section, we present and discuss some numerical results obtained from the general framework developed in Sections 3 and 4. Bearing in mind the uplink of a satellite IoT system, we target two relevant scenarios. In the former, packets of both service $\mathcal{S}_a$

and $\mathcal{S}_b$ reach the receiver with the same power level so that $\rho_a/\sigma^2 = \rho_b/\sigma^2 = 0$ dB. This setting reflects the coexistence of two services whose transmitters have a similar hardware (amplifier, antenna) and thus are received with comparable signal strength. In the latter, the QoS-constrained traffic of $\mathcal{S}_a$ is transmitted with higher power, and we consider the configuration $\rho_a/\sigma^2 = 10$ dB, $\rho_b/\sigma^2 = 0$ dB. The second setting, instead, assumes that the QoS-constrained service features transmitting units equipped with a better and possibly more costly hardware, thus resulting in a more favorable link budget. The selected configurations are in line with LEO satellite systems targeting IoT applications, e.g., [36].

As a starting point, let us focus on the asymptotic, ergodic setting introduced in Section 3. In this case, a vanishingly small error probability can always be granted to $\mathcal{S}_a$, whose QoS requirements are solely specified in terms of a target data rate R_a . The corresponding results are reported in Figure 2. The two solid lines represent the boundary of the rate pairs achievable by orthogonal and overlay allocations respectively, according to (4), (7) and (8). The shadowed light-red area comprises rate pairs achievable only with an overlay allocation, whereas the shadowed light-blue area denotes rates that are non-achievable by any of the two schemes. The two SNR scenarios, i.e., $\rho_a/\sigma^2 = \rho_b/\sigma^2 = 0$ dB and $\rho_a/\sigma^2 = 10$ dB, $\rho_b/\sigma^2 = 0$ dB are depicted in Figure 2a,b, respectively. In both configurations, irrespective of the target rate of user s_a , an overlay allocation is always beneficial to service $\mathcal{S}_b$ in terms of maximum achievable rates. Incidentally, we note that a slightly different trend was observed in [28], where, even in the asymptotic setting, orthogonal allocation may be beneficial for a small rate region. The discrepancy stems from two main factors. First, different channel models, i.e., AWGN with perfect power control vs. Rayleigh fading, are considered. Second, distinct decoding condition on the mMTC service are assumed. Indeed, while in our case we consider destructive collisions, ref. [28] relies on capture effect and interference cancellation (IC) also for service $\mathcal{S}_b$ data units.

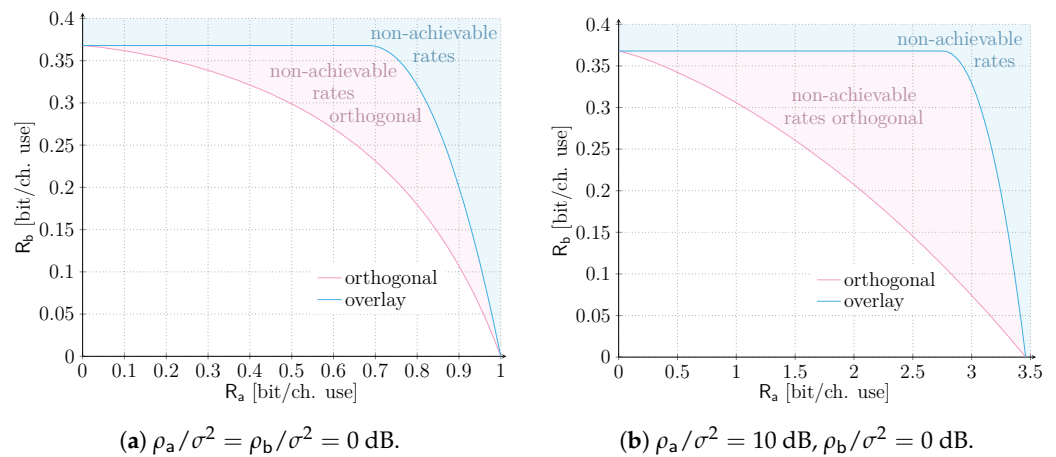


Figure 2. Asymptotic ergodic rate regions for an orthogonal and overlay allocation of resources of services $\mathcal{S}_a$ and $\mathcal{S}_b$. Two scenarios are considered: $\rho_a/\sigma^2 = \rho_b/\sigma^2 = 0$ dB and $\rho_a/\sigma^2 = 10$ dB, $\rho_b/\sigma^2 = 0$ dB.

From Figure 2, we can also infer that, when the two services are operated in overlay, $\mathcal{S}_b$ can achieve its maximum rate of $e^{-1} \cong 0.368$ (bit/ch. use) up to $R_a \leq 0.68$ (bit/ch. use) for $\rho_a/\sigma^2 = 0$ dB, and up to $R_a \leq 2.76$ (bit/ch. use) for $\rho_a/\sigma^2 = 10$ dB. In other words, increasing the target rate for the QoS-constrained service $\mathcal{S}_a$ does not impact service $\mathcal{S}_b$ in this region. Especially for the setting of Figure 2b, there is a very large range of rates for service $\mathcal{S}_a$, where the IoT traffic is limited by the poor performance of a slotted ALOHA access method. Indeed, the channel code protection of the data unit of user s_a could allow a larger channel traffic of service $\mathcal{S}_b$, beyond $\lambda = 1$, but cannot be reaped due to the limitation in the medium access. Such a remark hints at how advanced alternatives relying on packet repetition, e.g., [37], can be beneficial to expand the achievable rate region,

fostering the need for additional research in this direction. Finally, for $R_a > 0.68$ (bit/ch. use), or $R_a > 2.76$ (bit/ch. use) the maximum rate achievable for service $\mathcal{S}_b$ sees a larger and steeper degradation with respect to the orthogonal allocation as $R_a \rightarrow \log_2(1 + \rho_a/\sigma^2)$. As a consequence, more caution on the tuning of the traffic intensity for service $\mathcal{S}_b$ shall be devoted for such rates.

Let us now take a further step and consider the practical constraint imposed by having a finite number of slots within the communications frames (Section 3.2.2). Under these conditions, the transmission of service $\mathcal{S}_a$ fails to experience all possible interference levels from service $\mathcal{S}_b$, and a non-ergodic setting has to be considered for the overlay allocation. The corresponding results for a number of time slots in the set $M \in \{25, 100, 200\}$ are shown in Figure 3 together with the ergodic benchmark as reference. The set of time slots considered is in line with the literature on random access targeting satellite uplink scenarios, e.g., [37,38]. For our discussion, we set the target (average) codeword error probability for service $\mathcal{S}_a$ to 10^{-3} , and focus on the $\rho_a/\sigma^2 = \rho_b/\sigma^2 = 0$ dB scenario. Similar trends were also found for the unbalanced SNR scenario, and are not reported here for the sake of compactness. The impact of finite-length frames is clearly visible already for a moderately large number of slots, and the reduction of the achievable rate region becomes even more pronounced as the system operates with smaller values of M . For example, if user s_a targets a rate of $R_a = 0.8$ (bit/ch. use), the maximum achievable rate for service $\mathcal{S}_b$ is reduced by $\sim 51\%$ with respect to the ergodic setting, when $M = 25$. The rate R_b contraction becomes even more relevant for larger values of R_a . Conversely, increasing the number of time slots to 100 or 200 mitigates the trend. For the same target rate of $R_a = 0.8$ (bit/ch. use), the maximum achievable rate for service $\mathcal{S}_b$ is reduced by only $\sim 18\%$ with respect to the asymptotic setting, when 200 slots are considered. On the other hand, it is relevant to remark that, also in the non-ergodic setting, there exists a significant range of rate values for service $\mathcal{S}_a$ that allows to operate $\mathcal{S}_b$ at the maximum aggregate rate. Such a result confirms the potential of the overlay approach.

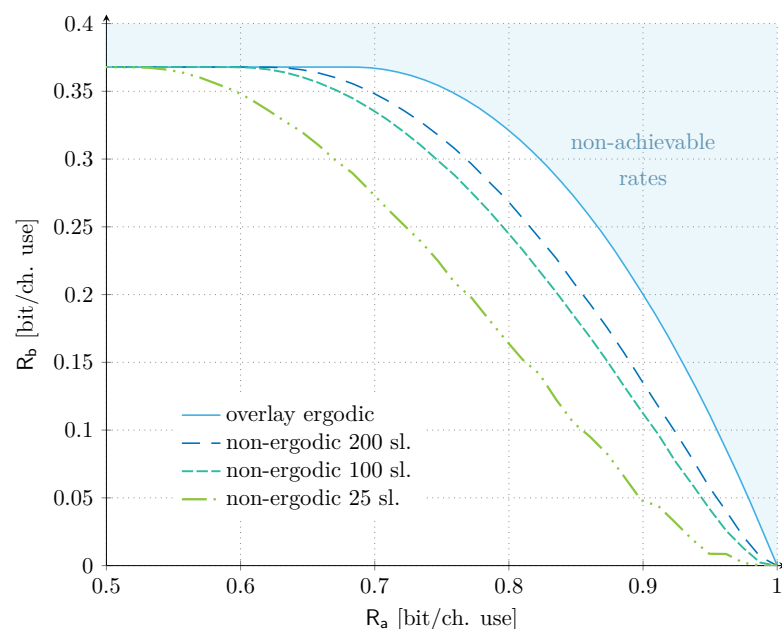


Figure 3. Asymptotic non-ergodic rate regions for overlay allocation. We fix $\rho_a/\sigma^2 = \rho_b/\sigma^2 = 0$ dB. The number of time slots are in the set $M \in \{25, 100, 200\}$ and the target outage probability for service $\mathcal{S}_a$ is $P_{\text{out}}^* = 10^{-3}$. The asymptotic ergodic rate region for the overlay allocation is also provided for reference.

We also investigate the trends under finite codeword length in Figure 4. In this case, not only the time slots, but also the number of channel uses per time slot are finite and we thus rely on the analysis provided in Section 4. As for the previous scenario,

we set the target average codeword error probability of service $\mathcal{S}_a$ to 10^{-3} , and consider $\rho_a/\sigma^2 = \rho_b/\sigma^2 = 0$ dB. The set of time slots is $M \in \{25, 100, 200\}$ and the set of channel uses per time slot $n_s \in \{100, 1000\}$. The latter choice is representative of practical mMTC applications, e.g., LoRa [5] and SigFox [8], which enable the transmission of payloads of up to 96 and 2000 bits, respectively. Accordingly, the set of total channel uses in the orthogonal allocation evaluates to $n \in \{2500, 10,000, 20,000, 25,000\}$. In the main plot of Figure 4, the orthogonal allocation and the overlay allocation for $M \in \{25, 100, 200\}$ and $n_s = 100$ channel uses per time slot are compared. As we can observe, increasing the number of time slots has a more beneficial impact on the overlay allocation than on the orthogonal one. In the former case, the increase in the number of time slots allows supporting a larger channel load for service $\mathcal{S}_b$ for the same target outage probability. In the latter case instead, it only impacts the correction term in (20) with respect to capacity, and thus results in a minor benefit on the achievable rate pairs. It is also worth noting that, as expected, the finite number of channel uses strongly affects the maximum achievable rate of service $\mathcal{S}_b$, reducing it by $\sim 18\%$ with respect to the asymptotic case. More interestingly, in contrast with what was discussed in the asymptotic setting, a region where an orthogonal allocation is superior to the overlay configuration emerges. Focusing on the $M = 25$ case, such a region is well highlighted in the subplot, where both scenarios with $n_s \in \{100, 1000\}$ are depicted. Although rather limited and only for rates of service $\mathcal{S}_a$ larger than 0.8 (bit/ch. use), such inversion suggests that reserving a portion of time for the IoT service alone is beneficial instead of letting the two services compete completely.

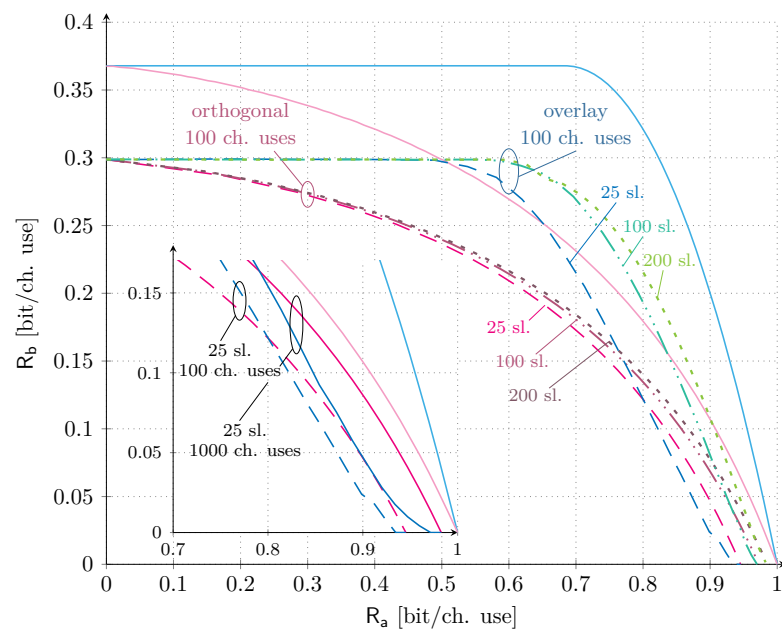


Figure 4. Finite length results for orthogonal and overlay allocations. We fix $\rho_a/\sigma^2 = \rho_b/\sigma^2 = 0$ dB. The number of time slots are in the set $M \in \{25, 100, 200\}$, the number of channel uses per time slot are $n_s \in \{100, 1000\}$ and the target average codeword error probability for service $\mathcal{S}_a$ is 10^{-3} . The asymptotic ergodic rate region for the both orthogonal and overlay allocations are also provided for reference.

In order to further investigate this aspect, we provide in Figure 5a different angle on the considered results. In particular, we analyze the ratio of the maximum rate achievable by service $\mathcal{S}_b$ with an overlay allocation to the one obtained in the orthogonal case as a function of R_a . Values larger than 1 thus identify regions of R_a where an overlay allocation can outperform the orthogonal one, so that a more aggregate mMTC rate is achieved while granting the same performance to the QoS-constrained service. To highlight the impact of short-packet transmissions, we compare the asymptotic, ergodic setting with finite length scenarios. In the latter case, we set the target average codeword error prob-

ability of service $\mathcal{S}_a$ to 10^{-3} . Let us focus first on the setting $\rho_a/\sigma^2 = \rho_b/\sigma^2 = 0$ dB. As already shown in Figure 2, under the asymptotic, ergodic setting, the overlay allocation is always beneficial. However, Figure 5a reveals the presence of an optimal operating point at $R_a \cong 0.92$ (bit/ch. use) for which the benefit achieved by the overlay allocation is maximized. A similar trend can also be observed for both $M = 100$ and $M = 25$, although for progressively smaller values of rate R_a . Moreover, the improvement reduces as well with the number of available time slots, going from a peak 87.5% increase in rate R_b for the overlay allocation with respect to the orthogonal one in the asymptotic setting, to a 48.7% improvement for $M = 100$ and 32.7% for $M = 25$. The plot also confirms that, in practical finite-length setups, there exist values of R_a for which an orthogonal allocation is convenient (i.e., the ratio falls below 1). Nonetheless, such a region drastically reduces by increasing the number of slots available over a frame. Finally, the effect of the number of channel uses appears to have a minor impact. Taking the lead from this, we explore in Figure 5b the configuration $\rho_a/\sigma^2 = 10$ dB, $\rho_b/\sigma^2 = 0$ dB, focusing only on $n_s = 100$. Increasing the SNR enjoyed by the QoS-constrained service drastically increases the advantage perceived by the IoT traffic when an overlay allocation is adopted. Rate exceeding a six-fold improvement for the IoT traffic can be achieved in the asymptotic ergodic setting, while, even for as few as 25 slots, more than a two-fold increase in R_b is expected when in overlay allocation. Interestingly, the region where the orthogonal allocation is superior for the equal SNR scenario appears to vanish in this scenario.

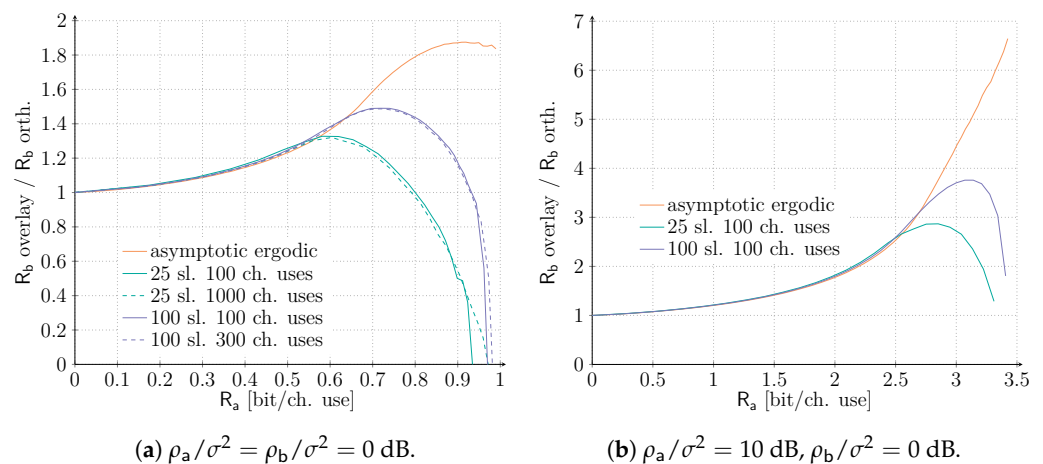


Figure 5. Ratio of the maximum rate achievable by service $\mathcal{S}_b$ with an overlay allocation (R_b overlay) over the one with an orthogonal allocation (R_b orth.) as a function of R_a . Both the ergodic asymptotic setting and the finite length scenarios are presented. In the latter, we set the target average codeword error probability of service $\mathcal{S}_a$ to 10^{-3} for both SNR setups. $M = 25$ and $M = 100$ slots together with $n_s = 100$ and $n_s = 300$ channel uses are investigated for $\rho_a/\sigma^2 = \rho_b/\sigma^2 = 0$ dB, while only $n_s = 100$ channel uses for $\rho_a/\sigma^2 = 10$ dB, $\rho_b/\sigma^2 = 0$ dB are shown.

6. Conclusions and Outlook

In this paper, we investigated the potential of letting two services, a QoS-constrained ($\mathcal{S}_a$) and a mMTC ($\mathcal{S}_b$), share a common spectrum by overlaying their transmissions in an AWGN scenario modeling the uplink of a satellite communication system. The receiver attempts decoding of $\mathcal{S}_a$ and, if successful, removes its contribution by means of IC, possibly allowing to retrieve data units of the mMTC traffic transmitted with a slotted ALOHA policy. Leveraging analytical tools, we have shown that an overlay allocation is beneficial in most situations, compared to a more traditional orthogonal allocation among the two services. Starting with an asymptotic scenario, where both the codewords and the number of time slots are very large, we delved into a non-ergodic setting (finite number of time slots) and the more practical finite-length regime (both time slots and codewords are finite). Achievable rate regions and expected gains for the IoT aggregate rate when

the overlay allocation is adopted are presented for all scenarios. Furthermore, a rate tuple (R_a, R_b) maximizing the improvement on the aggregate rate for the mMTC service is identified, showcasing a possible optimal operating point for the overlay system.

The presented work aims at stimulating further research in the context of the spectral coexistence between mMTC and QoS-constrained traffic in satellite scenarios. The possibility to upgrade the medium access policy of the mMTC service to more advanced solutions and exploring the benefits of *modern* random access schemes is an interesting direction. For example, the use of repetition-based solutions, e.g., [37], may unleash further benefits of the overlay allocation for low enough rates of the QoS-constrained service. In particular, when strong forward error correction is adopted on service $\mathcal{S}_a$, the mMTC service is not able to fully exploit it since high channel load values are detrimental in slotted ALOHA (cf. Figures 2 and 4, for example). A repetition-based scheme instead, can reap the rewards of a stronger interference rejection in $\mathcal{S}_a$ (lower rate) by heavily loading the physical layer with packet copies. Furthermore, we focused in this paper on perfect power control, i.e., all terminals of service $\mathcal{S}_b$ are received with the same power. In practical scenarios, in turn, fading and topology trigger variability in the received power levels so that the *capture* of packets is viable at the receiver. The impact of this aspect is also worth exploring, as it may trigger other relevant benefits and trade-offs, all the more so when coupled with IC, becoming especially relevant in the context of modern random access policies for IoT. The decoding order of services—first $\mathcal{S}_a$, then IC and subsequently $\mathcal{S}_b$ —can also be further investigated when power variability is present, along the lines of [28]. Finally, the applicability of the overlay allocation shall be investigated in a real-world scenario, entailing details such as the link budget, topology of the transmitters and effective error correcting code, among others.

Author Contributions: Conceptualization, A.M. and F.C.; methodology, A.M. and F.C.; software, A.M. and F.C.; validation, A.M. and F.C.; formal analysis, A.M. and F.C.; investigation, A.M. and F.C.; writing—original draft preparation, A.M. and F.C.; writing—review and editing, A.M. and F.C.; visualization, A.M. and F.C.; supervision, A.M. and F.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank Gianlugi Liva, German Aerospace Center (DLR), for the insightful discussions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Villa-Henriksen, A.; Edwards, G.; Pesonen, L.; Green, O.; Sorensen, C. Internet of Things in arable farming: Implementation, applications, challenges and potential. *Biosyst. Eng.* **2020**, *191*, 60–84. [CrossRef]
2. Aceto, G.; Persico, V.; Pescapé, A. A Survey on Information and Communication Technologies for Industry 4.0: State-of-the-Art, Taxonomies, Perspectives, and Challenges. *IEEE Commun. Surv. Tuts.* **2019**, *21*, 3467–3501. [CrossRef]
3. Dhingra, S.; Madda, R.; Gandomi, A.; Patan, R.; Daneshmand, M. Internet of Things Mobile-Air Pollution Monitoring System (IoT-Mobair). *IEEE Internet Things J.* **2019**, *6*, 5577–5584. [CrossRef]
4. Raza, U.; Kulkarni, P.; Sooriyabandara, M. Low Power Wide Area Networks: An Overview. *IEEE Commun. Surv. Tuts.* **2017**, *19*, 855–873. [CrossRef]
5. LoRa Alliance. The LoRa Alliance Wide Area Networks for Internet of Things. Available online: www.lora-alliance.org (accessed on 2 July 2021).
6. Georgiou, O.; Raza, U. Low Power Wide Area Network Analysis: Can LoRa Scale? *IEEE Wirel. Commun. Lett.* **2017**, *6*, 162–165. [CrossRef]
7. Adelantado, F.; Vilajosana, X.; Tuset-Peiro, P.; Martinez, B.; Melia-Segui, J.; Watteyne, T. Understanding the Limits of LoRaWAN. *IEEE Commun. Mag.* **2017**, *55*, 34–40. [CrossRef]

8. Sigfox. SIGFOX: The Global Communications Service Provider for the Internet of Things. Available online: www.sigfox.com (accessed on 2 July 2021).
9. Ingenu. Available online: www.ingenu.com (accessed on 2 July 2021).
10. Höglund, A.; Lin, X.; Liberg, O.; Behravan, A.; Yavuz, E.; Van der Zee, M.; Sui, Y.; Tirronen, T.; Ratilainen, A.; Eriksson, D. Overview of 3GPP Release 14 Enhanced NB-IoT. *IEEE Netw.* **2017**, *31*, 16–22. [\[CrossRef\]](#)
11. Marchese, M.; Moheddine, A.; Patrone, F. IoT and UAV integration in 5G hybrid terrestrial-satellite networks. *Sensors* **2019**, *19*, 3704. [\[CrossRef\]](#) [\[PubMed\]](#)
12. exactEarth. Available online: www.exactearth.com (accessed on 2 July 2021).
13. KLEO. KLEO Connect GmbH. Available online: www.kleo-connect.com (accessed on 2 July 2021).
14. Kepler Communications Inc. Available online: www.keplercommunications.com (accessed on 2 July 2021).
15. Coşkun, M.; Durisi, G.; Jerkovits, T.; Liva, G.; Ryan, W.; Stein, B.; Steiner, F. Efficient error-correcting codes in the short blocklength regime. *Elsevier Phys. Commun.* **2019**, *34*, 66–79. [\[CrossRef\]](#)
16. Berlioli, M.; Cocco, G.; Liva, G.; Munari, A. *Modern Random Access Protocols*; NOW Publisher: Boston, MA, USA; Delft, The Netherlands, 2016.
17. Polyanskiy, Y. A Perspective on Massive Random-Access. In Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 25–30 June 2017.
18. Fengler, A.; Jung, P.; Caire, G. SPARCs and AMP for unsourced random access. In Proceedings of the 2019 IEEE International Symposium on Information Theory (ISIT), Maison de la Mutualité, Paris, France, 7–19 July 2019.
19. Amalladinne, V.; Chamberland, J.F.; Narayanan, K. A coded compressed sensing scheme for unsourced multiple access. *IEEE Trans. Inf. Theory* **2020**, *66*, 6509–6533. [\[CrossRef\]](#)
20. Abramson, N. The throughput of packet broadcasting channels. *IEEE Trans. Commun.* **1977**, *25*, 117–128. [\[CrossRef\]](#)
21. ETSI. TS 102 721-3: *Satellite Earth Stations and Systems; Air Interface for S-band Mobile Interactive Multimedia (S-MIM); Part 3: Physical Layer Specification, Return Link Asynchronous Access*; Technical Report; ETSI: Sophia Antipolis, France, 2012.
22. Lien, S.; Shieh, S.; Huang, Y.; Su, B.; Hsu, Y.; Wei, H. 5G new radio: Waveform, frame structure, multiple access, and initial access. *IEEE Commun. Mag.* **2017**, *55*, 64–71. [\[CrossRef\]](#)
23. del Rio Herrero, O.; De Gaudenzi, R. High efficiency satellite multiple access scheme for machine-to-machine communications. *IEEE Trans. Aerosp. Electron. Syst.* **2012**, *48*, 2961–2989. [\[CrossRef\]](#)
24. Wang, Y.; Lin, X.; Adhikary, A.; Grovlen, A.; Sui, Y.; Blankenship, Y.; Bergman, J.; Razaghi, H. A primer on 3GPP narrowband internet of things. *IEEE Commun. Mag.* **2017**, *55*, 117–123. [\[CrossRef\]](#)
25. Gastpar, M. On capacity under receive and spatial spectrum-sharing constraints. *IEEE Trans. Inf. Theory* **2007**, *53*, 471–487. [\[CrossRef\]](#)
26. Lee, J.; Andrews, J.; Hong, D. Spectrum-sharing transmission capacity. *IEEE Trans. Wirel. Commun.* **2011**, *10*, 3053–3063. [\[CrossRef\]](#)
27. Kassab, R.; Simeone, O.; Popovski, P. Coexistence of URLLC and eMBB services in the C-RAN uplink: An information-theoretic study. In Proceedings of the 2018 IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, United Arab Emirates, 9–13 December 2018.
28. Popovski, P.; Trillingsgaard, K.F.; Simeone, O.; Durisi, G. 5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view. *IEEE Access* **2018**, *6*, 55765–55779. [\[CrossRef\]](#)
29. Trakadas, P.; Karkazis, P.; Leligou, H.; Zahariads, T.; Vicens, F.; Zurita, A.; Alemany, P.; Soenen, T.; Parada, C.; Bonnet, J.; et al. Comparison of management and orchestration solutions for the 5G era. *J. Sens. Actuator Netw.* **2020**, *9*, 1–18. [\[CrossRef\]](#)
30. Voicu, A.; Simic, L.; Petrova, M. Survey of spectrum sharing for inter-technology coexistence. *IEEE Commun. Surv. Tuts.* **2019**, *21*, 1112–1144. [\[CrossRef\]](#)
31. Polyanskiy, Y.; Poor, H.; Verdú, S. Channel coding rate in the finite blocklength regime. *IEEE Trans. Inf. Theory* **2010**, *56*, 2307–2359. [\[CrossRef\]](#)
32. Durisi, G.; Koch, T.; Popovski, P. Toward massive, ultrareliable, and low-latency wireless communication with short packets. *Proc. IEEE* **2016**, *104*, 1711–1726. [\[CrossRef\]](#)
33. Roberts, L. ALOHA packet systems with and without slots and capture. *ACM SIGCOMM Comput. Commun. Rev.* **1975**, *2*, 28–42. [\[CrossRef\]](#)
34. Metzger, F.; Hoßfeld, T.; Bauer, A.; Kounev, S.; Heegaard, P.E. Modeling of aggregated IoT traffic and its application to an IoT cloud. *Proc. IEEE* **2019**, *107*, 679–694. [\[CrossRef\]](#)
35. Lancho, A.; Koch, T.; Durisi, G. On single-antenna Rayleigh block-fading channels at finite blocklength. *IEEE Trans. Inf. Theory* **2020**, *66*, 496–519. [\[CrossRef\]](#)
36. Conti, M.; Guidotti, A.; Amatetti, C.; Vanelli-Coralli, A. NB-IoT over non-terrestrial networks: Link budget analysis. In Proceedings of the 2020 IEEE Global Communications Conference (GLOBECOM), Taipei, Taiwan, 7–11 December 2020.
37. Liva, G. Graph-based analysis and optimization of contention resolution diversity slotted ALOHA. *IEEE Trans. Commun.* **2011**, *59*, 477–487. [\[CrossRef\]](#)
38. Casini, E.; Gaudenzi, R.D.; del Rio Herrero, O. Contention Resolution Diversity Slotted ALOHA (CRDSA): An Enhanced Random Access Scheme for Satellite Access Packet Networks. *IEEE Trans. Wirel. Commun.* **2007**, *6*, 1408–1419. [\[CrossRef\]](#)