CHALMERS
UNIVERSITY OF TECHNOLOGY

Unsourced Multiple Access for Heterogeneous Traffic Requirements

Citation for the original published paper (version of record):
Ngo, K., Durisi, G., Graell I Amat, A. et al (2022). Unsourced Multiple Access for Heterogeneous Traffic Requirements. Conference Record - Asilomar Conference on Signals, Systems and Computers, 2022-October: 687-691. http://dx.doi.org/10.1109/IEEECONF56349.2022.10051987

(article starts on next page)

Unsourced Multiple Access for Heterogeneous Traffic Requirements

Khac-Hoang Ngo*, Giuseppe Durisi*, Alexandre Graell i Amat*,
Petar Popovski†, Anders E. Kalør†, and Beatriz Soret‡
*Department of Electrical Engineering, Chalmers University of Technology, 41296 Gothenburg, Sweden
†Department of Electronic Systems, Aalborg University, 9220 Aalborg Øst, Denmark
‡Telecommunication Research Institute (TELMA), Universidad de Málaga, 29010 Málaga, Spain

***Abstract*—We investigate the coexistence of critical and massive Internet of Things (IoT) services in the context of the unsourced multiple access (UMA) framework, introduced by Polyanskiy (2017). We consider the standard UMA setup in which all users employ a common codebook and the receiver returns an unordered list of codewords. To model the critical IoT service, we assume that the users can also communicate a common alarm message. We further assume that the number of active users in each transmission attempt is random and unknown. We derive a random-coding bound for the Gaussian multiple access channel and demonstrate that orthogonal network slicing enables the two traffic types to coexist with high energy efficiency.**

I. Introduction

Massive Internet of Things (IoT) is a communication paradigm that targets a large number of low-cost, battery-limited, uncoordinated devices that intermittently transmit small data volumes [1]. Such use case is referred to as massive machine-type communications (mMTC) in the fifth-generation (5G) wireless cellular standard. The characteristics of massive IoT are captured by the unsourced multiple access (UMA) model [2], where all users transmit their messages using the same codebook, the decoder returns an unordered list of messages, the error event is defined on a per-user basis, and the error probability is averaged over all users. In [2], a random-coding bound on the energy efficiency achievable on the Gaussian multiple access channel (MAC) was derived. Modern random access schemes [3] exhibit a large gap to this bound. This gap has been later reduced in, e.g., [4]–[7]. An extension to the case of random and unknown number of active users was reported in [8], where both misdetections (MDs),i.e., transmitted messages that are not included in the decoded list, and false positives (FPs), i.e., decoded messages that were not transmitted, were considered.

Another segment of the IoT traffic covers critical IoT services, which aim to deliver data with strict latency and reliability guarantees [1]. Critical IoT services are mapped to the ultra-reliable and low-latency communication (URLLC) use case in 5G. Such use case can be analyzed using tools from finite-blocklength information theory [9], [10].

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 101022113, and from the Swedish Research Council under grant 2021-04970.

This paper aims to investigate the coexistence of massive and critical IoT. In [11], the authors proposed to leverage reliability diversity to perform simultaneous transmission of different traffic types (also referred to as nonorthogonal network slicing) followed by successive interference cancellation (SIC). They showed that this approach leads to significant gains over orthogonal slicing when mMTC and enhanced mobile broadband (eMBB) traffic are present, or when URLLC and eMBB traffic are present. However, they noted that nonorthogonal slicing between URLLC and mMTC may be problematic due to the need to ensure reliability for URLLC devices in the presence of random interference patterns caused by mMTC transmissions. A first attempt to incorporate critical IoT traffic into the UMA model was reported in [12]. There, on top of standard messages, the users communicate a common alarm message that needs to be decoded with higher reliability than the standard messages. The authors assumed that a user drops the standard message in favor of the alarm message when both messages are available, and that the total number of active users transmitting either messages is known. They showed that, in nonorthogonal network slicing, the FP probability of alarm messages dominates and significantly reduces the energy efficiency when the total number of users is large.

In this paper, we study an orthogonal network slicing scheme where orthogonal resources are allocated to standard alarm messages. Differently from [12], we consider a random and unknown number of active users for both traffic types. Furthermore, both messages are transmitted if they are available. We provide a random-coding achievability bound for orthogonal network slicing on the Gaussian MAC. We assume that the standard traffic operates with a given additional energy per bit, which we call a backoff, on top of the minimum energy per bit required when the alarm traffic is not present. We then use our bound to analyze the energy efficiency of the alarm traffic. Through numerical results, we show that a limited backoff is sufficient to transmit the alarm traffic with high energy efficiency, provided that i) a large number of users transmit the alarm message and ii) the power at which the alarm message is transmitted is much smaller than that of the standard message. We also show that the bottleneck of nonorthogonal network slicing is the residual interference from the alarm signal when decoding the standard messages.

Notation: We denote system parameters by sans-serif letters, e.g., K, scalar random variables by upper case letters, e.g., X, and their realizations by lower case letters, e.g., x. Vectors are denoted likewise with boldface letters, e.g., a random vector $\boldsymbol{X}$ and its realization $\boldsymbol{x}$. We denote the $n \times n$ identity matrix by $\boldsymbol{I}_n$, and the all-zero vector by $\mathbf{0}$. The Euclidean norm is denoted by $\|\cdot\|$. We use $\mathfrak{P}(\mathcal{A})$ to denote the set of all subsets of $\mathcal{A}$; $[m:n] \triangleq \{m, m+1, \ldots, n\}$; $[n] \triangleq [1:n]$; $\mathbb{1}\{\cdot\}$ is the indicator function. We denote the Gamma function by $\Gamma(x) \triangleq \int_0^\infty z^{x-1}e^{-z}\mathrm{d}z$, and the upper incomplete Gamma functions by $\Gamma(x,y) \triangleq \int_y^\infty z^{x-1}e^{-z}\mathrm{d}z$. The complement of an event A is denoted by $\bar{A}$. We denote the Binomial distribution with parameters (n,p) by $\mathrm{Bino}(n,p)$.

II. System Model

We consider a MAC in which K users are given access opportunity over a frame of n uses of a stationary memoryless additive white Gaussian noise (AWGN) channel. Let $\boldsymbol{S}_k \in \mathbb{R}^{\mathsf{n}}$ be the signal transmitted by user k, which may be $\mathbf{0}$ if the user is inactive. This signal is subject to the power constraint $\|\boldsymbol{S}_k\|^2/\mathsf{n} \le \mathsf{P}$, $\forall k \in [\mathsf{K}]$. The corresponding received signal is given by $\boldsymbol{Y} = \sum_{k=1}^{\mathsf{K}} \boldsymbol{S}_k + \boldsymbol{Z}$, where $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{\mathsf{n}})$ is the AWGN, which is independent of $\{\boldsymbol{S}_k\}_{k=1}^{\mathsf{K}}$.

1) Message Generation: We let $\mathcal{M}_{\mathrm{a}}$ denote the set of alarm messages and $\mathcal{M}_{\mathrm{s}}$ the set of standard messages; both sets are common to all users. Let $\mathsf{M}_{\mathrm{a}} \triangleq |\mathcal{M}_{\mathrm{a}}|$ and $\mathsf{M}_{\mathrm{s}} \triangleq |\mathcal{M}_{\mathrm{s}}|$. We assume that M_{s} is much larger than M_{a}. In a frame, if an alarm event has occurred, let W_0 be the corresponding alarm message, drawn uniformly from $\mathcal{M}_{\mathrm{a}}$. Each user transmits this message with probability ρ_{d}. On top of that, with probability ρ_{s}, user k, $k \in [\mathsf{K}]$, generates a standard message W_k uniformly over $\mathcal{M}_{\mathrm{s}}$ and independently of the other users. To summarize, each user either transmits an alarm message, a standard message, or both messages, or is inactive. Fig. 1 illustrates the message generation rule. For convenience, we denote by w_{e} the "null message", mapped to the all-zero codeword (no transmission).

Remark 1: ρ_{d} is the product of the probability that a user detects the alarm event and the probability that, upon detecting the alarm event, the user decides to transmit the alarm message. The former probability represents the sensitivity of the devices, while the latter probability is a design choice.

Remark 2: The alarm message needs to be decoded with much higher reliability than the standard messages since reporting the alarm event is crucial for the system operation.

We assume that the number of users transmitting an alarm message and/or a standard message is unknown to the receiver.

2) Random-Access Code: Similar to [2], for the standard traffic, all users employ the same codebook and the receiver decodes up to a permutation of messages. Furthermore, as in [8], to address a *random* and *unknown* number of active users, we need to account for both MD and FP of the standard messages, referred to as SMD and SFP, respectively. We also need to consider MD and FP of the alarm message, referred to as AMD and AFP, respectively. We define the probabilities of these events and the random-access code in the following.

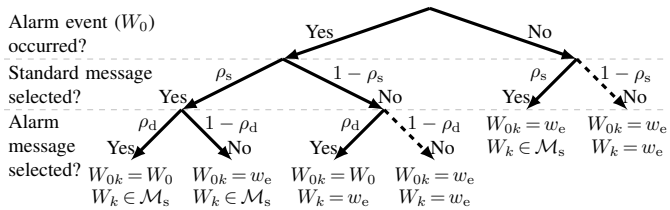

Fig. 1. A tree representation of the message generation of user k. The user generates an alarm message W_{0k} and a standard message W_k indicated by the leaves.

Definition 1 (Random-access code): Consider the Gaussian MAC with both standard and alarm traffic described above. An $(\mathsf{M}_{\mathrm{a}}, \mathsf{M}_{\mathrm{s}}, n, \epsilon_{\mathrm{amd}}, \epsilon_{\mathrm{afp}}, \epsilon_{\mathrm{smd}}, \epsilon_{\mathrm{sfp}})$ random-access code for this channel, where M_{a} and M_{s} are respectively the sizes of the alarm and standard message sets, n is the framelength, and $\epsilon_{\mathrm{amd}}, \epsilon_{\mathrm{afp}}, \epsilon_{\mathrm{smd}}, \epsilon_{\mathrm{sfp}} \in (0,1)$, consists of:

- A random variable U defined on a set $\mathcal{U}$ that is revealed to both the users and the receiver before the transmission. This random variable acts as common randomness and allows for the use of randomized coding strategies.
- An encoding function $f\colon \mathcal{U} \times (\mathcal{M}_{\mathrm{a}} \cup \{w_{\mathrm{e}}\}) \times (\mathcal{M}_{\mathrm{s}} \cup \{w_{\mathrm{e}}\}) \to \mathbb{R}^n$ that produces the transmitted codeword $\boldsymbol{S}_k = f(U, W_{0k}, W_k)$ for user k, for a given alarm message W_{0k} and standard message W_k.
- A decoding function $g\colon \mathcal{U} \times \mathbb{R}^n \to (\mathcal{M}_{\mathrm{a}} \cup \{w_{\mathrm{e}}\}) \times (\mathfrak{P}(\mathcal{M}_{\mathrm{s}}) \cup \{w_{\mathrm{e}}\})$ that provides an estimate $\widehat{W}_0$ of the common alarm message W_0 and an estimate $\widehat{\mathcal{W}} = \{\widehat{W}_1, \ldots, \widehat{W}_{|\widehat{\mathcal{W}}|}\}$ of the list of transmitted standard messages. That is, $(\widehat{W}_0, \widehat{\mathcal{W}}) = g(U, \boldsymbol{Y})$.

Let $\widetilde{\mathcal{W}} = \{\widetilde{W}_1, \ldots, \widetilde{W}_{|\widetilde{\mathcal{W}}|}\}$ be the set of distinct elements of $\mathcal{W} = \{W_k\colon W_k \neq w_{\mathrm{e}}, k \in [\mathsf{K}]\}$. We assume that the decoding function satisfies the following constraints on the AMD, AFP, SMD, and SFP probabilities, respectively:

$$P_{\mathrm{amd}} \triangleq \mathbb{P}\Big[\widehat{W}_0 \neq W_0 \big| A\Big] \le \epsilon_{\mathrm{amd}}, \quad (1)$$

$$P_{\mathrm{afp}} \triangleq \mathbb{P}\Big[\widehat{W}_0 \neq w_{\mathrm{e}} \big| \bar{A}\Big] \le \epsilon_{\mathrm{afp}}, \quad (2)$$

$$P_{\mathrm{smd}|B} \triangleq \mathbb{E}_{|\widetilde{W}|}\left[\frac{1}{|\widetilde{\mathcal{W}}|}\sum_{i=1}^{|\widetilde{W}|}\mathbb{P}\Big[\widetilde{W}_i \notin \widehat{\mathcal{W}} \big| B\Big]\right] \le \epsilon_{\mathrm{smd}}, \quad (3)$$

$$P_{\mathrm{sfp}|B} \triangleq \mathbb{E}_{|\widehat{\mathcal{W}}|}\left[\frac{1}{|\widehat{\mathcal{W}}|}\sum_{i=1}^{\widehat{\mathcal{W}}}\mathbb{P}\Big[\widehat{W}_i \notin \widetilde{\mathcal{W}} \big| B\Big]\right] \le \epsilon_{\mathrm{sfp}}. \quad (4)$$

Here, we used the convention that $0/0 = 0$ to circumvent the cases $|\widetilde{\mathcal{W}}| = 0$ or $|\widehat{\mathcal{W}}| = 0$. Furthermore, A denotes the event that an alarm has occurred, and (3) and (4) hold for both $B = A$ and $B = \bar{A}$.

III. Heterogeneous Orthogonal Multiple Access

We propose an orthogonal slicing strategy referred to as heterogeneous orthogonal multiple access (H-OMA). Each frame is split into two blocks containing respectively n_{a} channel uses dedicated to the alarm traffic, and $\mathsf{n}_{\mathrm{s}} = \mathsf{n} - \mathsf{n}_{\mathrm{a}}$ channel uses dedicated to the standard traffic. We next describe the signal model in each block.

A. Signal Model

1) Alarm Block: If an alarm event has occurred, the common alarm message W_0 is sent in the alarm block by every user that detects the alarm and decides to transmit. The received signal is $\boldsymbol{Y}_\mathrm{a} = K_\mathrm{a}\boldsymbol{X}_0 + \boldsymbol{Z}_\mathrm{a}$, where $K_\mathrm{a} \geq 0$ is the number of users transmitting the common alarm codeword $\boldsymbol{X}_0 \in \mathbb{R}^{\mathsf{n}_\mathrm{a}}$, and $\boldsymbol{Z}_\mathrm{a} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{\mathsf{n}_\mathrm{a}})$ is the AWGN. If no alarm event occurs, $K_\mathrm{a} = 0$; otherwise, $K_\mathrm{a} \sim \mathrm{Bino}(\mathsf{K}, \rho_\mathrm{d})$. We impose the power constraint $\|\boldsymbol{X}_0\|^2/\mathsf{n}_\mathrm{a} \leq \mathsf{P}_\mathrm{a}$. This model is equivalent to a single-user AWGN channel with random signal-to-noise ratio (SNR) $K_\mathrm{a}^2\mathsf{P}_\mathrm{a}$. The average energy per bit of alarm traffic is upper-bounded by $(E_\mathrm{b}/N_0)_\mathrm{a} \triangleq \frac{\mathsf{n}_\mathrm{a}\mathsf{P}_\mathrm{a}\rho_d\mathsf{K}}{2\log_2 \mathsf{M}_\mathrm{a}}$.

2) Standard Block: The standard block resembles the UMA channel with random and unknown number of active users considered in [8]. The number of active users in this block is $K_\mathrm{s} \sim \mathrm{Bino}(\mathsf{K}, \rho_\mathrm{s})$. We assume without loss of generality that the first K_s users transmit. The received signal is then given by $\boldsymbol{Y}_\mathrm{s} = \sum_{k=1}^{K_\mathrm{s}} \boldsymbol{X}_k + \boldsymbol{Z}_\mathrm{s}$, where $\boldsymbol{X}_k \in \mathbb{R}^{\mathsf{n}_\mathrm{s}}$ is the standard codeword transmitted by user k, and $\boldsymbol{Z}_\mathrm{s} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{\mathsf{n}_\mathrm{s}})$ is the AWGN. We impose a power constraint $\|\boldsymbol{X}_k\|^2/\mathsf{n}_\mathrm{s} \leq \mathsf{P}_\mathrm{s}$, $k \in [K_\mathrm{s}]$. The average energy per bit of standard traffic is upper-bounded by $(E_\mathrm{b}/N_0)_\mathrm{s} \triangleq \frac{\mathsf{n}_\mathrm{s}\mathsf{P}_\mathrm{s}}{2\log_2 \mathsf{M}_\mathrm{s}}$.

In accordance with Definition 1, the output of the overall encoding function is the concatenation of an alarm codeword and a standard codeword. To satisfy the overall power constraint, we set $(\mathsf{P}_\mathrm{a}, \mathsf{P}_\mathrm{s})$ such that $\mathsf{n}_\mathrm{a}\mathsf{P}_\mathrm{a} + \mathsf{n}_\mathrm{s}\mathsf{P}_\mathrm{s} \leq \mathsf{n}\mathsf{P}$. We note that, for fixed P_a and P_s, orthogonality implies that the MD and FP in the standard traffic are independent of the alarm event, i.e., $P_{\mathrm{smd}|\bar{A}} = P_{\mathrm{smd}|A}$ and $P_{\mathrm{sfp}|\bar{A}} = P_{\mathrm{sfp}|A}$.

B. Random-Coding Bound

In the following, for a given frame split $(\mathsf{n}_\mathrm{a}, \mathsf{n}_\mathrm{s})$ and power allocation $(\mathsf{P}_\mathrm{a}, \mathsf{P}_\mathrm{s})$, we derive a random-coding bound on the AMD, AFP, SMD, and SFP probabilities in (1)–(4).

1) Alarm Block: For the encoder, we fix $\mathsf{P}'_\mathrm{a} \leq \mathsf{P}_\mathrm{a}$ and draw M_a alarm codewords $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{\mathsf{M}_\mathrm{a}}$ independently from $\mathcal{N}(\boldsymbol{0}, \mathsf{P}'_\mathrm{a}\boldsymbol{I}_{\mathsf{n}_\mathrm{a}})$. To convey an alarm message W_0, the active users transmit $\boldsymbol{c}_{W_0}$ provided that $\|\boldsymbol{c}_{W_0}\|^2 \leq \mathsf{n}_\mathrm{a}\mathsf{P}_\mathrm{a}$. Otherwise, they transmit the all-zero codeword, i.e., $\boldsymbol{X}_0 = \boldsymbol{c}_{W_0}\mathbb{1}\{\|\boldsymbol{c}_{W_0}\|^2 \leq \mathsf{n}_\mathrm{a}\mathsf{P}_\mathrm{a}\}$. Given a realization $\boldsymbol{y}_\mathrm{a}$ of the received signal, the decoder proceeds in two steps. First, it estimates the number of active users K_a as

$$K'_\mathrm{a} = \underset{k \in \{0\} \cup [k_{\mathrm{a},\ell}:k_{\mathrm{a},u}]}{\arg\max}\ m_\mathrm{a}(\boldsymbol{y}_\mathrm{a}, k) \tag{5}$$

where $m_\mathrm{a}(\boldsymbol{y}_\mathrm{a}, k)$ is a suitably chosen metric, and $k_{\mathrm{a},\ell} \geq 1$ and $k_{\mathrm{a},u} \leq \mathsf{K}$ are lower and upper estimation limits chosen based on the distribution of K_a, i.e., based on ρ_d. If $K'_\mathrm{a} = 0$, the decoder returns the null message w_e. Otherwise, given $K'_\mathrm{a} = k'_\mathrm{a} > 0$, the decoder uses k'_a to establish an interval for minimum-distance decoding of the alarm message as

$$(\widehat{W}_0, \widehat{K}_\mathrm{a}) = \underset{w \in \mathcal{M}_\mathrm{a},\ k \in \{0\} \cup [\underline{k'_\mathrm{a}}:\overline{k'_\mathrm{a}}]}{\arg\min} \|\boldsymbol{y}_\mathrm{a} - k\boldsymbol{c}_w\|^2 \tag{6}$$

where $\underline{k'_\mathrm{a}} \triangleq \max\{k_{\mathrm{a},\ell}, k'_\mathrm{a} - r_\mathrm{a}\}$, $\overline{k'_\mathrm{a}} \triangleq \min\{k_{\mathrm{a},u}, k'_\mathrm{a} + r_\mathrm{a}\}$. Here, r_a is a chosen nonnegative integer, which we call the *alarm-message decoding radius*. Finally, the decoder returns $\widehat{W}_0$ if $\widehat{K}_\mathrm{a} > 0$, or returns w_e if $\widehat{K}_\mathrm{a} = 0$.

Remark 3: The first step (5) results in an AFP if $K_\mathrm{a} = 0$ but $K'_\mathrm{a} > 0$. To avoid this AFP, we include $k = 0$ in the refined estimation of K_a in the second step (6).

An error analysis of this random-coding scheme leads to the following bounds on the AMD and AFP probabilities.

Theorem 1 (Random-coding bound for the alarm block): Fix M_a, r_a, $\mathsf{n}_\mathrm{a} \in [\mathsf{n}]$, $k_{\mathrm{a},\ell} \in [0 : \mathsf{K}]$, $k_{\mathrm{a},u} \in [k_{\mathrm{a},\ell} + 1 : \mathsf{K}]$, P_a, and $\mathsf{P}'_\mathrm{a} < \mathsf{P}_\mathrm{a}$. The AMD and AFP probabilities achieved by the random-coding scheme just described are upper-bounded by ϵ_amd and ϵ_afp, respectively, where

$$\epsilon_\mathrm{amd} = \sum_{k_\mathrm{a}=k_{\mathrm{a},\ell}}^{k_{\mathrm{a},u}} P_{K_\mathrm{a}}(k_\mathrm{a}) \sum_{k'_\mathrm{a}=k_{\mathrm{a},\ell}}^{k_{\mathrm{a},u}} \zeta(k_\mathrm{a}, k'_\mathrm{a})\gamma_\mathrm{amd}(k_\mathrm{a}, k'_\mathrm{a}) + \hat{p}, \tag{7}$$

$$\epsilon_\mathrm{afp} = \sum_{\hat{k}_\mathrm{a}=\max\{k_{\mathrm{a},\ell}, k'_\mathrm{a}\}}^{k_{\mathrm{a},u}} \min\left\{1, \sum_{k'_\mathrm{a}=\max\{k_{\mathrm{a},\ell}, \hat{k}_\mathrm{a}-r_\mathrm{a}, 1\}}^{\min\{k_{\mathrm{a},u}, \hat{k}_\mathrm{a}+r_\mathrm{a}\}} \zeta(0, k'_\mathrm{a})\right\} \cdot \gamma_\mathrm{afp}(\hat{k}_\mathrm{a}), \tag{8}$$

with $P_{K_\mathrm{a}}(k_\mathrm{a}) = \binom{\mathsf{K}}{k_\mathrm{a}}\rho_\mathrm{d}^{k_\mathrm{a}}(1-\rho_\mathrm{d})^{\mathsf{K}-k_\mathrm{a}}$ and

$$\hat{p} \triangleq \frac{\Gamma\left(\frac{\mathsf{n}_\mathrm{a}}{2}, \frac{\mathsf{n}_\mathrm{a}\mathsf{P}_\mathrm{a}}{2\mathsf{P}'_\mathrm{a}}\right)}{\Gamma(\mathsf{n}_\mathrm{a}/2)} + 1 - \sum_{k=k_{\mathrm{a},\ell}}^{k_{\mathrm{a},u}} P_{K_\mathrm{a}}(k), \tag{9}$$

$$\gamma_\mathrm{amd}(k_\mathrm{a}, k'_\mathrm{a}) \triangleq \begin{cases} 1, \text{ if } k'_\mathrm{a} \in \big\{[0 : \max\{k_\mathrm{a} - r_\mathrm{a}, k_{\mathrm{a},\ell}\}] \\ \qquad\qquad \cup\ [k_\mathrm{a} + r_\mathrm{a} : k_{\mathrm{a},u}]\big\}, \\ \sum_{\hat{k}_\mathrm{a} \in \{0\} \cup [\underline{k'_\mathrm{a}}:\overline{k'_\mathrm{a}}]} q(\hat{k}_\mathrm{a}), \quad \text{otherwise}, \end{cases} \tag{10}$$

$$q(\hat{k}_\mathrm{a}) \triangleq \min_{s>0} \mathbb{P}\left[\sum_{i=1}^{\mathsf{n}_\mathrm{a}} \imath_s(\hat{k}_\mathrm{a}, X'_i; Y'_i) \leq \ln\frac{\mathsf{M}_\mathrm{a}-1}{V}\right], \tag{11}$$

$$\gamma_\mathrm{afp}(\hat{k}_\mathrm{a}) \triangleq \min_{s>0} \mathbb{E}\left[\frac{1}{\Gamma(\mathsf{n}_\mathrm{a}/2)} \cdot \Gamma\left(\frac{\mathsf{n}_\mathrm{a}}{2}, \frac{\frac{\mathsf{n}_\mathrm{a}}{2}\ln(1+2\hat{k}_\mathrm{a}^2\mathsf{P}'_\mathrm{a}s) - \ln\frac{\mathsf{M}_\mathrm{a}}{V}}{2s(1-(1+2\hat{k}_\mathrm{a}^2\mathsf{P}'_\mathrm{a}s)^{-1})}\right)\right], \tag{12}$$

$$\zeta(k_\mathrm{a}, k'_\mathrm{a}) \triangleq \min_{\substack{k \in \{0\} \cup [k_{\mathrm{a},\ell}:k_{\mathrm{a},u}] \\ k \neq k'_\mathrm{a}}} \mathbb{P}[m_\mathrm{a}(\boldsymbol{Y}'_\mathrm{a}, k'_\mathrm{a}) > m_\mathrm{a}(\boldsymbol{Y}'_\mathrm{a}, k)], \tag{13}$$

In (11) and (12), V is uniformly distributed over $[0, 1]$. In (11), $[X'_1 \ldots X'_{\mathsf{n}_\mathrm{a}}]^\intercal \sim \mathcal{N}(\boldsymbol{0}, \mathsf{P}'_\mathrm{a}\boldsymbol{I}_{\mathsf{n}_\mathrm{a}})$. Given $X'_i = x'_i$, we have that $Y'_i \sim \mathcal{N}(k_\mathrm{a}x'_i, 1)$, and $\imath_s(\hat{k}_\mathrm{a}, X'_i; Y'_i)$ is the generalized information density given by $\imath_s(\hat{k}_\mathrm{a}, x; y) \triangleq -s(y - k_\mathrm{a}x)^2 + \frac{sy^2}{1+2s\hat{k}_\mathrm{a}^2P'} + \frac{1}{2}\ln\left(1+2s\hat{k}_\mathrm{a}^2P'\right)$. Finally, in (13), $\boldsymbol{Y}'_\mathrm{a} \sim \mathcal{N}(\boldsymbol{0}, (1 + k_\mathrm{a}^2\mathsf{P}'_\mathrm{a})\boldsymbol{I}_{\mathsf{n}_\mathrm{a}})$.

Proof 1: The proof relies on the random-coding union bound with parameter s [13]. It is omitted due to the space limit and will be provided in an extended version of this paper.

Remark 4: The quantity $\zeta(k_\mathrm{a}, k'_\mathrm{a})$ is an upper bound on the probability that given $K_\mathrm{a} = k_\mathrm{a}$, the estimation step (5)

returns $k'_{\rm a}$. Closed-form expressions of $\zeta(k_{\rm a}, k'_{\rm a})$ for the maximum likelihood estimation and energy-based estimation of $K_{\rm a}$ can be deduced from [8, Th. 2].

2) Standard Block: We consider the random-coding scheme proposed in [8, Sec. III-A]. Specifically, for the encoder, we fix $\mathsf{P}'_{\rm s} < \mathsf{P}_{\rm s}$ and generate the $\mathsf{M}_{\rm s}$ standard codewords $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{\mathsf{M}_{\rm s}}$ independently from the distribution $\mathcal{N}(\boldsymbol{0}, \mathsf{P}'_{\rm s}\boldsymbol{I}_{\mathsf{n}_{\rm s}})$. To convey a standard message W_k, the corresponding active user transmits $\boldsymbol{X}_k = \boldsymbol{c}_{W_k}\mathbb{1}\{\|\boldsymbol{c}_{W_k}\|^2 \leq \mathsf{n}_{\rm s}\mathsf{P}_{\rm s}\}$. Given the channel output $\boldsymbol{y}_{\rm s}$, the decoder first estimates the number of active users as

$$K'_{\rm s} = \underset{k\in[k_{{\rm s},\ell}:k_{{\rm s},u}]}{\arg\max}\ m_{\rm s}(\boldsymbol{y}_{\rm s}, k)\,, \tag{14}$$

where $m_{\rm s}(\boldsymbol{y}_{\rm s}, k)$ is a suitably chosen metric, and $k_{{\rm s},\ell}$ and $k_{{\rm s},u}$ are limits on $K'_{\rm s}$, chosen based on the distribution of $K_{\rm s}$, i.e., based on $\rho_{\rm s}$. Then, given $K'_{\rm s} = k'_{\rm s}$, the decoder chooses the best list size within an interval around $k'_{\rm s}$ as

$$\widehat{\mathcal{W}} = \underset{\mathcal{W}'\subset[\mathcal{M}_{\rm s}]:\ |\mathcal{W}'|\in[\underline{k'_{\rm s}}:\overline{k'_{\rm s}}]}{\arg\min} \left\|\boldsymbol{y}_{\rm s} - \textstyle\sum_{i\in\mathcal{W}'}\boldsymbol{c}_i\right\|^2, \tag{15}$$

where $\underline{k'_{\rm s}} \triangleq \max\{k_{{\rm s},\ell}, k'_{\rm s} - r_{\rm s}\}$, $\overline{k'_{\rm s}} \triangleq \min\{k_{{\rm s},u}, k_{\rm s} + r_{\rm s}\}$, and $r_{\rm s}$ is a chosen nonnegative integer, which we call the *standard-message decoding radius*. Bounds on the SMD and SFP probabilities achieved by this random-coding scheme follow from [8, Th. 1].

Theorem 2 (Random-coding bound for the standard block): Fix $\mathsf{M}_{\rm s}$, $r_{\rm s}$, $\mathsf{n}_{\rm s} \in [\mathsf{n}]$, $k_{{\rm s},\ell} \in [0:\mathsf{K}]$, $k_{{\rm s},u} \in [k_{{\rm s},\ell}+1:\mathsf{K}]$, $\mathsf{P}_{\rm s}$, and $\mathsf{P}'_{\rm s} < \mathsf{P}_{\rm s}$. The SMD and SFP probabilities achieved by the random-coding scheme just described are upper-bounded by $\epsilon_{\rm smd}$ and $\epsilon_{\rm sfp}$, obtained by adapting $\epsilon_{\rm MD}$ and $\epsilon_{\rm FA}$ given in [8, Th. 1] to the real-valued case.

3) Overall Random-Coding Bound: By combining Theorem 1 and Theorem 2, we obtain the following random-coding bound for H-OMA.

Theorem 3 (Random-coding bound for H-OMA): Fix $r_{\rm a}$, $r_{\rm s}$, $\mathsf{n}_{\rm a} \in [0:\mathsf{n}]$, $k_{{\rm a},\ell} \in [0:\mathsf{K}]$, $k_{{\rm a},u} \in [k_{{\rm a},\ell}+1:\mathsf{K}]$, $k_{{\rm s},\ell} \in [0:\mathsf{K}]$, $k_{{\rm s},u} \in [k_{{\rm s},\ell}+1:\mathsf{K}]$, $(\mathsf{P}_{\rm a}, \mathsf{P}_{\rm s})$ such that $\mathsf{n}_{\rm a}\mathsf{P}_{\rm a} + \mathsf{n}_{\rm s}\mathsf{P}_{\rm s} \leq \mathsf{n}\mathsf{P}$, $\mathsf{P}'_{\rm a} < \mathsf{P}_{\rm a}$, and $\mathsf{P}'_{\rm s} < \mathsf{P}_{\rm s}$. For the considered Gaussian MAC with both standard and alarm traffic, there exists an $(\mathsf{M}_{\rm a}, \mathsf{M}_{\rm s}, \mathsf{n}, \epsilon_{\rm amd}, \epsilon_{\rm afp}, \epsilon_{\rm smd}, \epsilon_{\rm sfp})$ random-access code where $\epsilon_{\rm amd}$ and $\epsilon_{\rm afp}$ are given in Theorem 1, and $\epsilon_{\rm smd}$ and $\epsilon_{\rm sfp}$ are given in Theorem 2.

Remark 5: Our bounding techniques can be used to generalize the analysis of the performance of the heterogeneous nonorthogonal multiple access (H-NOMA) scheme proposed in [12] to the case of unknown number of active users. In H-NOMA, both standard and alarm codewords are transmitted in the whole frame. They are generated as in H-OMA, but with $\mathsf{n}_{\rm a} = \mathsf{n}_{\rm s} = \mathsf{n}$. The receiver first decodes the alarm message and estimates the number of alarm users similarly to the alarm block in H-OMA by treating the standard codewords as noise. Thus, bounds on $P_{\rm amd}$ and $P_{\rm afp}$ for H-NOMA can be obtained by adapting the bounds in (7) and (8) to the effective noise variance $1 + K_{\rm s}\mathsf{P}'_{\rm s}$ and by averaging over the distribution of $K_{\rm s}$. Next, exploiting reliability diversity, the receiver removes the decoded alarm codeword from the received signal and proceeds to decode the standard messages. If there is no alarm, the received signal is similar to that in the standard block in H-OMA. By assuming that the standard messages can be decoded only if an AFP has not occurred, one can bound $P_{{\rm smd}|\bar{A}}$ and $P_{{\rm sfp}|\bar{A}}$ in a similar manner as for H-OMA, upon accounting for the AFP probability. If there is an alarm, the residual interference plus noise after SIC is $\boldsymbol{Z} + K_{\rm a}\boldsymbol{c}_{W_0} - \widehat{K}_{\rm a}\boldsymbol{c}_{\widehat{W}_0}$. Thus, $P_{{\rm smd}|A}$ and $P_{{\rm sfp}|A}$ can be bounded by adapting the bounds for H-OMA to the effective noise variance, which is $1+(K_{\rm a}-\widehat{K}_{\rm a})^2\mathsf{P}'_{\rm a}$ if $W_0 = \widehat{W}_0$ and $1+(K_{\rm a}^2+\widehat{K}_{\rm a}^2)\mathsf{P}'_{\rm a}$ otherwise.

IV. Numerical Experiments

We consider $\mathsf{n} = 30000$ and $(\mathsf{M}_{\rm s}, \mathsf{M}_{\rm a}) = (2^{100}, 2^3)$. We set $\mathsf{K} \in [1000:30000]$ and $\rho_{\rm s} = 0.01$, so that $\mathbb{E}[K_{\rm s}] \in [10:300]$, similar to the setting in [2], [8]. We consider the mild target reliability $\max\{P_{\rm smd}, P_{\rm sfp}\} \leq 10^{-1}$ for the standard traffic, and the stringent target reliability $\max\{P_{\rm amd}, P_{\rm afp}\} \leq 10^{-5}$ for the alarm traffic. Let $(E_{\rm b}/N_0)^*_{\rm s}$ be the minimum required energy per bit for the standard traffic if the alarm traffic is not present, i.e., if $\mathsf{n}_{\rm s} = \mathsf{n}$. We address the following question: *Let the standard traffic operate at $(E_{\rm b}/N_0)^*_{\rm s} + \delta$ (dB) for a fixed backoff $\delta > 0$. What is the minimum required $(E_{\rm b}/N_0)_{\rm a}$?* To this end, we first find $(E_{\rm b}/N_0)^*_{\rm s}$ in a similar manner as in [8]. We then find the minimum blocklength $\mathsf{n}_{\rm s,min}$ required to satisfy $\max\{\epsilon_{\rm smd}, \epsilon_{\rm sfp}\} \leq 10^{-1}$ at $(E_{\rm b}/N_0)^*_{\rm s}+\delta$ dB. The number of available channel uses for the alarm traffic is thus $\mathsf{n}_{\rm a,max} = \mathsf{n} - \mathsf{n}_{\rm s,min}$. Finally, we minimize the required $(E_{\rm b}/N_0)_{\rm a}$ using a golden-section search over $\mathsf{n}_{\rm a} \in [\mathsf{n}_{\rm a,max}]$, where for each value of $\mathsf{n}_{\rm a}$, the required $(E_{\rm b}/N_0)_{\rm a}$ is minimized over $\rho_{\rm d}$ and $\mathsf{P}_{\rm a}$. In the following, we set $\delta = 0.1$ dB, for which $\mathsf{n}_{\rm a,max} \in \{7584, 964, 526\}$ for $\mathsf{K} \in \{10000, 20000, 30000\}$, respectively, and report the minimum $(E_{\rm b}/N_0)_{\rm a}$.

1) Impact of Device Sensitivity: As noted in Remark 1, $\rho_{\rm d}$ is upper-bounded by the probability that a user detects the alarm event, denoted by $\rho_{\rm d,max}$, which indicates the device sensitivity. To study the impact of $\rho_{\rm d,max}$ on the alarm-traffic energy efficiency, we vary its value and show in Fig. 2 the corresponding minimum $(E_{\rm b}/N_0)_{\rm a}$ as a function of K. We observe that a low $(E_{\rm b}/N_0)_{\rm a}$ can be achieved, especially for high $\rho_{\rm d,max}$. This indicates that in H-OMA, the alarm message can be transmitted at a high energy efficiency, at a cost of only a marginal backoff in the standard-traffic energy efficiency. A higher $(E_{\rm b}/N_0)_{\rm a}$ is required as $\rho_{\rm d,max}$ decreases, i.e., when the devices become less sensitive. Furthermore, we found that it is optimal to set $\rho_{\rm d}$ to its maximum value $\rho_{\rm d,max}$ and then minimize $\mathsf{P}_{\rm a}$ and $\mathsf{n}_{\rm a}$. That is, one should let every user that detects the alarm event transmit at a low power using only few channel uses. The reason is that, to increase the effective SNR $K_{\rm a}^2\mathsf{P}_{\rm a}$ while keeping a low total energy $K_{\rm a}\mathsf{P}_{\rm a}\mathsf{n}_{\rm a}$, one should increase $K_{\rm a}$ (via increasing $\rho_{\rm d}$) and reduce $\mathsf{P}_{\rm a}$ and $\mathsf{n}_{\rm a}$.

2) Impact of Dynamic Range: Keeping $\mathsf{P}_{\rm a}$ as low as possible leads to a large difference in the transmitted power between the two blocks. For example, for the parameters used in Fig. 2, the ratio $\mathsf{P}_{\rm s}/\mathsf{P}_{\rm a}$ is between 30 dB and 70 dB. This power imbalance may not be compatible with the limited dynamic range of IoT devices. To account for this, we impose

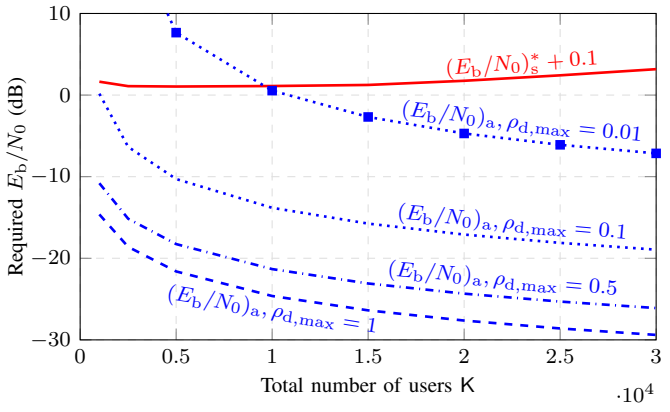

Fig. 2. The minimum $(E_b/N_0)_a$ required to satisfy $\max\{\epsilon_{amd}, \epsilon_{afp}\} \leq 10^{-5}$ when $(E_b/N_0)_s = (E_b/N_0)_s^* + 0.1$ dB for different device's sensitivity $\rho_{d,max}$. Here, $\mathsf{M}_a = 2^3$, $\mathsf{M}_s = 2^{100}$, and $\rho_s = 0.01$.

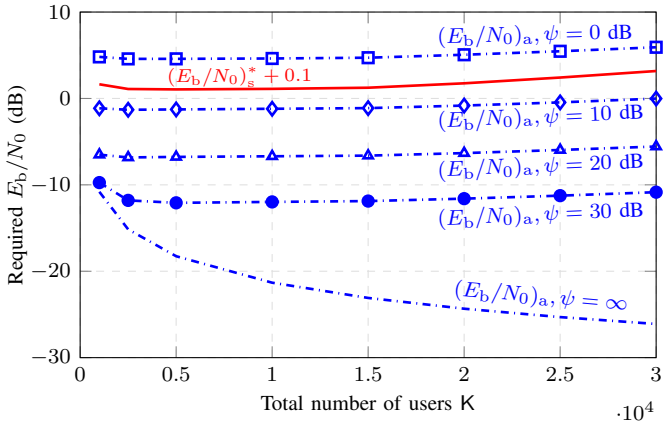

Fig. 3. The minimum $(E_b/N_0)_a$ required to satisfy $\max\{\epsilon_{amd}, \epsilon_{afp}\} \leq 10^{-5}$ when $(E_b/N_0)_s = (E_b/N_0)_s^* + 0.1$ dB for different dynamic range ψ. Here, $\mathsf{M}_a = 2^3$, $\mathsf{M}_s = 2^{100}$, $\rho_s = 0.01$, and $\rho_{d,max} = 0.5$.

the additional constraint $\mathsf{P}_s/\mathsf{P}_a \leq \psi$ in the minimization of $(E_b/N_0)_a$. In Fig. 3, we plot the minimum required $(E_b/N_0)_a$ for $\rho_{d,max} = 0.5$ and $\psi \in \{0, 10, 20, 30, \infty\}$ dB. The case $\psi = \infty$ corresponds to the setting in Fig. 2. We see that a narrower dynamic range leads to a higher required $(E_b/N_0)_a$. If the users transmit at equal power over the two blocks, i.e., $\psi = 0$ dB, the alarm traffic requires a higher energy per bit than the standard traffic. Furthermore, for a finite dynamic range, the required $(E_b/N_0)_a$ increases with K, which is in contrast with the case of infinite dynamic range. For each n_a, the required $(E_b/N_0)_a$ is minimized by setting P_a to its minimum value P_s/ψ and then minimizing ρ_d.

3) Comparison with H-NOMA: As mentioned in Remark 5, our bounding techniques can be used to extend the random-coding bound for H-NOMA reported in [12] to the case of unknown number of active users. We found that the bottleneck for H-NOMA is to satisfy the target SMD and SFP probabilities when there is an alarm. Specifically, although the alarm message W_0 can be reliably decoded, the number of alarm users K_a can be estimated incorrectly with significant probability. For example, with $K_s = 100$, the probability of wrongly estimating the number of alarm users given no AMD, computed as $\mathbb{P}\big[\arg\min_k \|\boldsymbol{Y} - k\boldsymbol{c}_{W_0}\|^2 \neq K_a\big]$, is 0.276 for $\psi = 20$ dB and 0.426 for $\psi = 30$ dB. This leads to a high residual interference $K_a\boldsymbol{c}_{W_0} - \widehat{K}_a\boldsymbol{c}_{\widehat{W}_0}$ when decoding the standard messages even if $\widehat{W}_0 = W_0$. Because of this bottleneck, H-NOMA cannot satisfy the reliability requirements for both traffic types with the same $(E_b/N_0)_s$ backoff $\delta = 0.1$ dB unless $\rho_d = 1$, in which case $K_a = \mathsf{K}$, or P_a is high, i.e., comparable to P_s, so that the estimation of K_a is reliable. In both cases, however, the required $(E_b/N_0)_a$ is high. For the setting in Fig. 3, if one sets $\mathsf{P}_a = \mathsf{P}_s$, the required $(E_b/N_0)_a$ is around 28–30 dB. This shows that reliability diversity [11] is hard to exploit in nonorthogonal network slicing between massive and critical IoT.

V. Conclusions

We investigated massive and critical IoT in a setting with both a standard UMA traffic and an alarm traffic. Considering a random and unknown number of active users, we accounted for both misdetections and false positives. For the Gaussian MAC, our results show that the both traffic types can coexist with high energy efficiency by means of orthogonal network slicing, while nonorthogonal network slicing is inefficient.

References

[1] A. Zaidi, A. Brännerby, A. Nazari, M. Hogan, and C. Kuhlins, "Ericsson white paper: Cellular IoT in the 5G era," Stockholm, Sweden, Tech. Rep., Feb. 2020. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/white-papers/cellular-iot-in-the-5g-era

[2] Y. Polyanskiy, "A perspective on massive random-access," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 2523–2527.

[3] M. Berioli, G. Cocco, G. Liva, and A. Munari, "Modern random access protocols," *Foundations and Trends in Networking*, vol. 10, no. 4, pp. 317–446, Nov. 2016.

[4] A. Fengler, P. Jung, and G. Caire, "SPARCs for unsourced random access," *IEEE Trans. Inf. Theory*, vol. 67, no. 10, pp. 6894–6915, Oct. 2021.

[5] V. K. Amalladinne, A. K. Pradhan, C. Rush, J.-F. Chamberland, and K. R. Narayanan, "Unsourced random access with coded compressed sensing: Integrating AMP and belief propagation," *IEEE Trans. Inf. Theory*, vol. 68, no. 4, pp. 2384–2409, Apr. 2022.

[6] A. K. Pradhan, V. K. Amalladinne, K. R. Narayanan, and J.-F. Chamberland, "Polar coding and random spreading for unsourced multiple access," in *Proc. IEEE Int. Conf. Communications (ICC)*, Dublin, Ireland, Jun. 2020.

[7] Z. Han, X. Yuan, C. Xu, S. Jiang, and X. Wang, "Sparse Kronecker-product coding for unsourced multiple access," *IEEE Wireless Commun. Lett.*, vol. 10, no. 10, pp. 2274–2278, Oct. 2021.

[8] K.-H. Ngo, A. Lancho, G. Durisi, and A. Graell i Amat, "Unsourced multiple access with random user activity," *IEEE Trans. Inf. Theory*, vol. 69, no. 7, pp. 4537–4558, Jul. 2023.

[9] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[10] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Sep. 2016.

[11] 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, Aug. 2018.

[12] K. Stern, A. E. Kalør, B. Soret, and P. Popovski, "Massive random access with common alarm messages," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Paris, France, Jul. 2019.

[13] A. Martinez and A. Guillén i Fàbregas, "Saddlepoint approximation of random-coding bounds," in *Proc. Inf. Theory Applicat. Workshop (ITA)*, La Jolla, CA, USA, Feb. 2011.