IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 6, 2016, accepted January 1, 2017, date of publication January 10, 2017, date of current version March 8, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2651170

LATMAPA: Load-Adaptive Throughput-MAximizing Preamble Allocation for Prioritization in 5G Random Access

MIKHAIL VILGELM[1], (Student Member, IEEE), H. MURAT GÜRSU[1], (Student Member, IEEE), WOLFGANG KELLERER[1], (Senior Member, IEEE), AND MARTIN REISSLEIN[2], (Fellow, IEEE)

[1]Deptartment of Electrical and Computer Engineering, Technical University of Munich, 80290 Munich, Germany
[2]School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85287-5706, USA

Corresponding author: M. Reisslein (reisslein@asu.edu)

This work was supported in part by the European Research Council under the European Unions Horizon 2020 Research and Innovation Program (under Grant 647158 FlexNets) and in part by a Friedrich Wilhelm Bessel Research Award from the Alexander von Humboldt Foundation.

ABSTRACT Persistently high traffic loads and heterogeneous quality of service (QoS) requirements arising from machine-to-machine communication in wireless 5G systems require effective random access prioritization. 5G systems will likely evolve from mature wireless technologies, e.g., long term evolution (LTE). LTE conducts random access through preamble contention based on slotted Aloha principles. Prior studies have mainly examined random access prioritization for addressing temporary traffic bursts through manipulating the access contention procedure on a given set of preambles, such as adapting the number of permitted transmission attempts and back off windows. We conduct a detailed study of random access prioritization through separating (splitting) the random access preambles into non-overlapping priority classes. Based on the obtained insights, we develop the Load-Adaptive Throughput-MAximizing Preamble Allocation (LATMAPA). LATMAPA automatically adjusts the preamble allocation to the priority classes according to the random access load and a priority tuning parameter. Extensive analytical and simulation evaluations indicate that LATMAPA provides effective QoS differentiation across a wide range of random access loads, which are expected in 5G systems.

INDEX TERMS 5G wireless system, LTE connection establishment, machine-to-machine (M2M) traffic, priority classes, preamble separation, random access.

I. INTRODUCTION

The Evolution of cellular networks towards 5G brings an eminent shift in design objectives. While the previous generations of wireless systems have mainly focused on data rate hungry applications, 5G must be designed to satisfy the diverse needs of Machine-to-Machine (M2M) applications [1]–[6]. The M2M market is becoming more important for cellular network operators also due to saturating growth of the personal communication market [7], [8]. The diverse communication requirements of M2M applications range from delay-tolerant smart metering in smart grid networks [9]–[11] to highly delay-sensitive and loss-intolerant applications in Intelligent Transportation Systems [12]–[15].

In order to efficiently support the wide variety of 5G application requirements, it is natural that the development of 5G systems will rely on the evolution of existing mature cellular technologies, such as 3GPP Long Term Evolution (LTE) and its recent version LTE-Advanced (LTE-A). LTE and LTE-A already support Quality-of-Service (QoS) differentiation in radio access scheduling [16] and in tunneling through the core network. However, M2M applications pose significant challenges for future 5G systems that build on LTE technologies. A key M2M challenge is the large number of devices in a single cell [17], which is expected to create high signaling overheads. Thus, 5G systems require not only the design of new physical layer and radio frame structures [18], but also the evolution of the procedures for *obtaining* and *maintaining* the radio resources [19].

In an LTE-A network, a well-known bottleneck for communication in the uplink direction is the connection establishment between a User Equipment (UE) and Evolved NodeB (eNB) [20]. Unlike the schedule-based operation of a

connected UE, initial connection establishment is performed using Random Access (RA), and, thus, prone to collisions and degraded performance in the overload region [21], [22]. As Gerasimenko *et al.* [23] show, RA alone can cause delays of more than 150 ms in densely populated cells. Due to the sporadic nature of message transmissions in many MTC applications, such as emergency Vehicle-to-Infrastructure (V2I) messages, it is not prudent to keep radio resources continuously reserved. Instead, most MTC devices must complete the RA procedure before sending a message. Thus, in order to support differentiated QoS for different classes of M2M traffic, effective random access prioritization is needed.

A. CONTRIBUTIONS

Extensive prior work has examined the mitigation of temporary, non-persistent UE request traffic bursts through manipulation of the random access procedure on a given set of preambles, as detailed in Section I-B. In contrast, we consider persistently high (constant) UE request traffic loads. For the constant traffic setting, we examine the effects of preamble separation on the LTE random access throughput, delay, and request drop ratio for two UE request classes. Class I represents delay-intolerant UE requests and class II represents delay-tolerant UE requests. One can imagine a class mapping to QoS Class Identifier (QCI) classes [24], or a mapping to Human-to-Human (H2H) and M2M devices [25]–[27]. We quantify the throughput, delay, and drop ratio trade-offs of separating the preambles into two disjoint sets. For underloaded systems, we find that there is a ''safe'' allocating region, where class I prioritization is relatively harmless for class II. Also, we quantify an allocation region where the overall throughput is increased due to preamble separation.

Based on these insights, we develop the Load-Adaptive Throughput-MAximizing Preamble Allocation (LATMAPA). LATMAPA is based on a throughput maximization principle and automatically adapts the number of preambles allocated to the high- and low-priority classes according to their load levels. Our evaluations indicate that LATMAPA effectively ensures high throughput as well as low delays and drop probabilities for the high priority class across a wide load range.

B. RELATED WORK

Random access in LTE-A has been studied from a variety of angles. In the following, we briefly review the categories most closely related to our random access prioritization study. We first review studies on the general problem area of controlling random access overload due to M2M traffic in LTE random access. Then we review studies on QoS provisioning and prioritization in LTE random access.

1) M2M TRAFFIC IN LTE RANDOM ACCESS

As a 3GPP approved solution, Access Class Barring (ACB) has been introduced in LTE release 8 [28]. ACB defines a specific barring probability parameter, which is used by every UE to decide probabilistically whether or not to attempt a transmission. Building upon the standardized solution, studies [29], [30] have proposed methods for cooperatively changing barring parameters among the neighboring base stations, while dynamic adjustments of the ACB have been studied in [31] and [32]. The ACB has similarly been exploited for accommodating M2M traffic in [33]. The study [34] has examined dynamic adjustments of the contention window and retransmission limit based on the current load, while the authors in [35] used future load predictions to update the access barring parameters. In order to limit the cross-influence of M2M and H2H devices, M2M-specific back-off or variable access cycles for M2M-devices can be employed [36]. Self-optimizing methods for PRACH resource allocation have been proposed in [37] and [38]. Pratas *et al.* [39] and Condoluci *et al.* [40], [41] have investigated an expansion of the random access contention space through a combination of conventional preambles and access code words.

Due to the similarity of the LTE RACH problem with slotted ALOHA (s-ALOHA), many methods from earlier studies on s-ALOHA have been adopted for LTE. A prominent example is tree-based collision resolution, first introduced for s-ALOHA in [42] and [43]. Madueno *et al.* have analyzed tree-based collision resolution for LTE [44]. Madueno *et al.* have also examined splitting the RACH cycle into a phase for estimating the number of arrivals, followed by a phase for serving the arrivals with tree algorithms [45].

2) PRIORITIZATION THROUGH RANDOM ACCESS PROCEDURE MANIPULATION

Several studies, e.g., [46]–[49], have investigated random access prioritization through manipulations of the random contention procedures or parameters, such as transmission attempt limit and backoff window duration, on a given set of preambles. Moreover, as a refinement of ACB, Extended Access Class Barring (EAB) has been introduced by 3GPP in Release 11 [28]. EAB enables prioritization through assigning different barring probabilities to the different UE classes [50], [51]. The adjustment of the random access contention, e.g, through EAB, on a given set of preambles is complementary to our approach of conducting the random access contention of the different priority classes on separate sets of preambles. In particular, the random access contention could be differentiated within a given preamble set to achieve further QoS differentiation. Generally, methods that manipulate the random access contention, such as EAB, are designed for non-persistent temporary UE request traffic bursts [31]; whereas we focus on persistently high UE request traffic loads.

3) PRIORITIZATION THROUGH PREAMBLE SEPARATION

A few prior studies have examined different forms of preamble separation. In particular, some studies have split the preambles into distinct sets for contention-based random access and for non-contention (dedicated)

access [38], [52], [53]. Chu *et al.* have developed a general model of resource allocation in slotted ALOHA (whereby a preamble can be considered a resource) through a matrix representation [54]. Complementary to these studies, we focus on contention-based random access.

Initial studies of the prioritization of contention-based random access through separating preambles have been conducted by Lee *et al.* [55], Kalalas *et al.* [56], and Lin *et al.* [57], [58]. (The prioritization through preamble separation has also been covered in the patent [59]). These initial studies have only examined throughput for pre-configured fixed static preamble separation. In contrast, we consider dynamic adaptive preamble separation according to the traffic loads for the priority classes according to the LATMAPA approach introduced in this study. Moreover, we conduct an in-depth evaluation of LATMAPA that considers throughput, delay, and drop probabilities.

Zhao *et al.* [60] have proposed a heuristic load-adaptive preamble allocation rule, which we consider as a comparison benchmark in our evaluations, see Section VI-C. Zhao *et al.* have incorporated the heuristic preamble allocation rule into an overall protocol with a variant of binary exponential backoff. In this study, we focus on examining the effects of preamble allocation for prioritizing LTE random access. We do not vary the backoff process; rather we consider the standard LTE uniform random backoff throughout.

Du *et al.* [61] have proposed an approach for PRACH resource allocation, aiming at minimizing the contention resolution time. The approach relies on real-time knowledge of the number of contending UEs in every PRACH slot, and on numerical solvers for calculating the optimal split for certain load values. In contrast, LATMAPA requires only average load as an input, and provides a closed-form expression for the optimal split. We compare LATMAPA to the approach by Du *et al.* [61] in Section VI-C.

4) OTHER RELATED LTE RANDOM ACCESS STUDIES

The impact of limitations of the Physical Downlink Control Channel (PDCCH) on the LTE random access procedure and the proper dimensioning of the PDCCH has been studied in [62]–[64]. We assume that the PDCCH is properly dimensioned and does not limit random access. For completeness, we note that the connection establishment via LTE random access has also been studied in the context of heterogeneous access networks [65], [66] and the discovery procedure in Device-to-Device (D2D) [67].

C. PAPER STRUCTURE

The remainder of the paper is organized as follows. Sections II and III give background on LTE random access and the concept of preamble separation. Section IV analyzes how the number of allocated preambles affects the RACH performance for individual classes and for the entire system over a range of UE request loads. Section V examines preamble allocation methods that strive to meet a delay target or strive to maximize throughput; the throughput maximization approach results in the Load-Adaptive Throughput-MAximizing Preamble Allocation (LATMAPA) approach. Sec. VI evaluates LATMAPA through analysis, simulations, and benchmark comparisons. Section VII summarizes the paper.

II. LTE-A RACH BACKGROUND

We review the basic operation of the LTE-A random access in this section. A UE needs to go through the Random Access (RA) procedure in order to obtain initial synchronization with an eNB. The RA procedure is necessary in case of the transition from RRC-IDLE state to RRC-CONNECTED, or in other cases of lost synchronization, such as handover. An outcome of a successful RA procedure is the acquisition of resources for an uplink transmission on the Physical Uplink Shared Channel (PUSCH). There are two modes of the RA procedure: contention-free or contention-based RA. In the contention-free mode, the eNB can uniquely identify the UE by a received preamble sequence, due to the fact that the preamble to be used has been communicated to the UE in advance. Such a scenario is possible in certain cases, e.g., during a handover between two eNBs. In the following we consider only the contention-based RA procedure. The LTE RA procedure uses a dedicated Physical Random Access Channel (PRACH). An example location of the PRACH on the resource grid is depicted in Fig. 1.

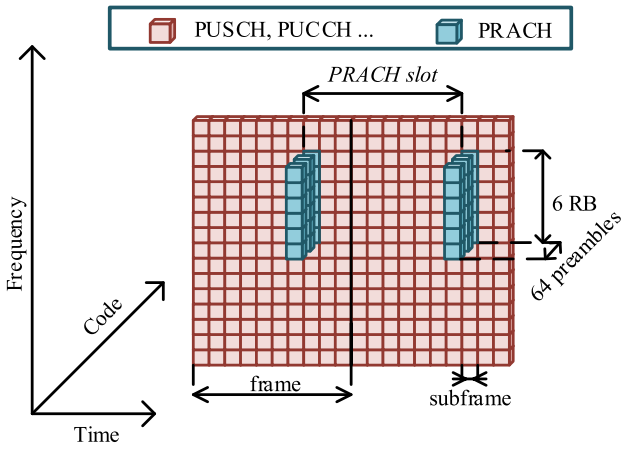


FIGURE 1. Exemplary PRACH allocation on the resource grid. System parameters: 3 MHz bandwidth corresponding to 16 Resource Blocks (RBs) for all channels, PRACH configuration index 5, frequency offset 7, preamble length in frequency domain 6 RBs [68].

The random access procedure starts with a UE listening for the System Information Block 2 (**SIB2**) message advertised by the eNB on the broadcast channel, see Fig. 2. The SIB2 message contains the PRACH configuration index and the frequency offset. These two parameters inform the UE about the sub-frames and Resource Blocks (RBs) that are reserved for RA in the next frame. Depending on the configuration index, one or more sub-frames can be reserved for PRACH. We define a *PRACH slot* or *slot* as the time between the beginning time instants of two consecutive PRACH sub-frames.

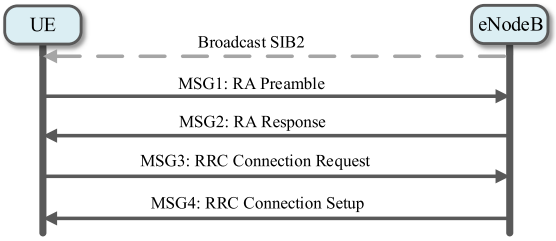


FIGURE 2. Illustration of steps of contention-based LTE Random Access (RA) procedure.

The length of the PRACH slot depends on the PRACH configuration index (see Section VI-E), and can vary from 1 ms up to 20 ms. For simplicity, we assume that all four protocol steps illustrated in Fig. 2 are completed within one PRACH slot.

Afterwards, the UE selects a preamble from the available set, and sends the selected preamble sequence to the eNB as Message 1 (**MSG1**). In a typical configuration, the LTE RACH has 64 available preambles, whereby 10 preambles are reserved for contention-free access, and $M = 54$ preambles are available for contention-based access. Note that MSG1 does not contain any information about the UE's identity. Thus, a possible collision could not be identified upon the reception of MSG1. The eNB can only detect whether a particular preamble has been selected or not, but not how many UEs have selected the preamble [44]. After the reception of MSG1, the eNB replies with **MSG2**, containing a set of preambles accepted for transmission as well as the corresponding C-RNTI and timing advance values. At this point, all accepted UEs send **MSG3** in the scheduled slot, and, if more than one UE had selected the same preamble for MSG1, a collision will occur and none of the collided UEs will be granted access. If, however, a UE had selected a unique preamble for sending MSG1, no collision occurs at this step and the UE will receive the necessary connection setup response as **MSG4**.

III. PREAMBLE SEPARATION

In this section we introduce the considered prioritization through preamble separation and define the model of random access system with preamble separation.

A. PREAMBLE ASSIGNMENT OPTIONS

In general, several options of allocating preambles can be considered. Conventionally, there is **no separation**, meaning all devices compete in the entire set of preambles and can collide with each other. Another option is fixed, **non-overlapping assignment**, where both classes have their own preamble set, thus, competing only with the devices from the same class. **Overlapping assignment** [27], [57], [58] assumes that prioritized UEs can compete in the entire set, whereas non-prioritized UEs can only use a predefined fraction of the preambles.

In this paper, we compare the steady-state performance of the system for the no separation and non-overlapping assignment allocation options. The separation of the preambles into two sets involves a number of trade-offs. By allocating more preambles to class I, we degrade the performance of class II.

B. MODELING LTE-A RACH WITH PREAMBLE SEPARATION

Generally, the LTE RACH can be represented as a multi-channel slotted Aloha system, with a slot representing one time-domain RACH opportunity, and a channel representing one RACH preamble [69], [70]. In our model, we consider two device classes, both with an infinite number of UEs and constant request arrival rates. That is, the numbers of arriving requests per slot are modeled by independent Poisson distributions, with the expected values λ_I and λ_{II} for class I (delay-intolerant devices) and class II (delay-tolerant devices), respectively.

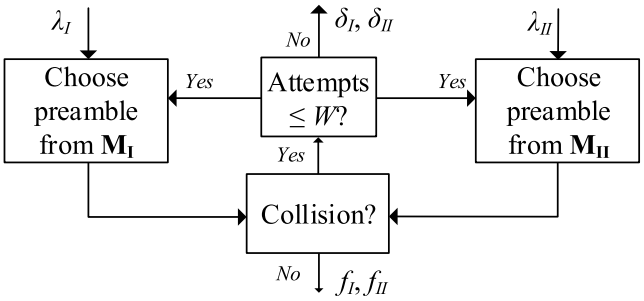


FIGURE 3. Illustration of two-class fixed-assignment RACH model with preamble sets $\mathbf{M_I}$, $\mathbf{M_{II}}$: UE requests arrive with rates λ_I and λ_{II} for the two classes and select preambles from their respective fixed-assigned sets $\mathbf{M_I}$ and $\mathbf{M_{II}}$. Preamble transmissions without a collision result in successes. Collided preamble transmissions are retransmitted until W attempts are reached and then dropped (if the Wth attempt collides).

The UEs of both classes attempt to send a RACH MSG1, which consists of a RACH preamble chosen uniformly out of the available sets $\mathbf{M_I}$ and $\mathbf{M_{II}}$, respectively, as illustrated in Fig. 3. Any request that has collided in a first transmission attempt is retransmitted again up to the maximum of W transmission attempts. The re-transmission proceeds after a back-off time that is uniformly chosen from the interval 0 to $B_{\max}$. If a request has collided W times, it is considered as dropped. We denote δ for the request drop probability. We denote f for the probability of success in one attempt. The average delay D measures the average number of slots from the first request transmission attempt until the successful reception of the request. Note that the delay D does not take the unsuccessful (dropped) requests into account.

Since we consider infinite sets of devices, the arrival rates of the initial (new) requests remain constant, while the retransmissions increase the total number of UEs attempting access up to x for the steady-state [70]. A summary of system model notations in presented in the Tab. 1. We note that some MAC and physical layer considerations have not been captured in our system model, since we focus on the preamble contention aspect. We acknowledge that, in general, the neglected parameters, such as UE location [71], inter-cell

TABLE 1. Summary of model notations.

$\mathbf{M}$	Set of all preambles available in each slot
M	Total number of preambles available in each slot ($= 54$ if not stated otherwise)
$m_I,\ m_{II}$	Numbers of preambles allocated for class I, class II
W	Maximum number of allowed transmission attempts ($= 8$ if not stated otherwise)
$B_{\max}$	Max. back-off value in slots ($= 20$, default)
$\lambda_I,\ \lambda_{II}$	Poisson arrival rates of class I, class II UE req./slot
$\rho = \frac{\lambda}{M}$	Poisson proc. arrival rate, normalized for one preamble
f	Steady-state UE req. success probability in one attempt
x	Expected number of UE req. (incl. initial arrivals + retransmissions) contending for preambles in a slot.
T	Steady-state throughput of UE req. per preamble per slot
D	Average steady-state delay (in slots)
δ	Steady-state UE req. drop prob., after max. of W attempts
$\hat{\rho}$	Normalized Poisson process arrival rate achieving the max. throughput, referred to as *peak throughput load*
$\hat{D}$	Steady-state del. (in slots) for successful UE req. if $\rho = \hat{\rho}$
$\hat{\delta}$	Steady-state ratio of dropped UE req. if $\rho = \hat{\rho}$

interference, or access barring [72], can influence the RACH behavior.

IV. ANALYSIS OF RANDOM ACCESS SYSTEM

In this section, we analyze the influence of the numbers of preambles assigned to the two classes on the key performance metrics throughput, delay, and drop ratio. Initially, as groundwork, we analyze the random access system without preamble separation. Then, we proceed to examine preamble separation.

A. OVERALL SYSTEM WITHOUT PREAMBLE SEPARATION

1) REVIEW OF STEADY STATE ANALYSIS

Utilizing the notation summarized in Table 1, we first briefly review the steady-state analysis of the system without preamble separation [70], [73]. In steady-state,

$$f = e^{-\frac{x}{M}} \text{ and} \tag{1}$$

$$\frac{x}{\lambda} = \frac{1-(1-f)^W}{f}. \tag{2}$$

As there is no closed-form solution for Eqn. (2) with respect to f, numerical methods have to be used to obtain f and x from the system of Eqns. (1), (2). The obtained f and x values are used to calculate the performance metrics as [70], [73]:

- Throughput T: ratio of successfully received requests to the total number of transmission opportunities:

$$T = \frac{\lambda}{M}(1-\delta). \tag{3}$$

- Delay D: time period from the first transmission attempt until the request is successfully received by the eNB. Since the number of PRACH slots in a given LTE frame depends on the PRACH configuration, we measure the delay in units of PRACH slots:

$$D = \left(1 + \frac{B_{\max}}{2}\right)\frac{1}{f-1} \times \times \frac{1+(W-1)(1-f)^W - W(1-f)^{W-1}}{1-(1-f)^W}. \tag{4}$$

- Drop ratio δ: ratio of the requests that did not succeed in any of the W transmission attempts to the total number of initial requests transmitted:

$$\delta = (1-f)^W. \tag{5}$$

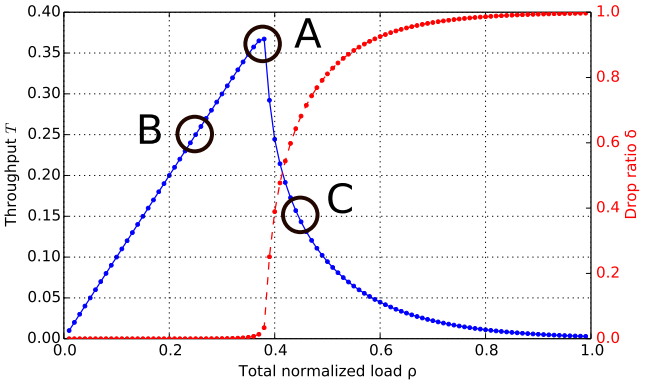


FIGURE 4. Throughput T and drop ratio δ vs normalized arrival rate ρ in the system without preamble separation, $W = 8$ transmission attempts.

The resulting dependency of the total throughput and drop ratio on the total normalized load ρ is depicted in Fig. 4. We observe that there are two distinct operating regions: an underloaded region (to the left of point **A** in Fig. 4), and an overloaded region (to the right of point **A**). The underloaded region is characterized by a linear increase of the throughput and a steady low drop ratio. On the other hand, in the overloaded region, the drop ratio increases rapidly as the throughput drops.

Our hypothesis is that the preamble separation into two device classes has different effects and involves different trade-offs depending on whether the total system load is in the underloaded or overloaded region. Hence, it is important to exactly know the load value at the border between these two regions. Therefore, we find in the next subsection the normalized load value $\hat{\rho}$ corresponding to the maximum throughput at point **A** in Fig. 4.

2) PEAK THROUGHPUT LOAD

Considering the total normalized load $\rho = \lambda/M$, we evaluate the load value $\hat{\rho}$ that achieves the peak throughput, i.e., corresponds to point **A** in Fig. 4, by analyzing the function $T(\rho)$ Eqn. (3). After solving Eqn. (1) for x and substituting it in Eqn. (2), considering that $\rho = \lambda/M$, we obtain:

$$\rho = \frac{f\ln(f)}{(1-f)^W - 1}. \tag{6}$$

From Eqns. (3) and (6):

$$T = -f\ln(f). \tag{7}$$

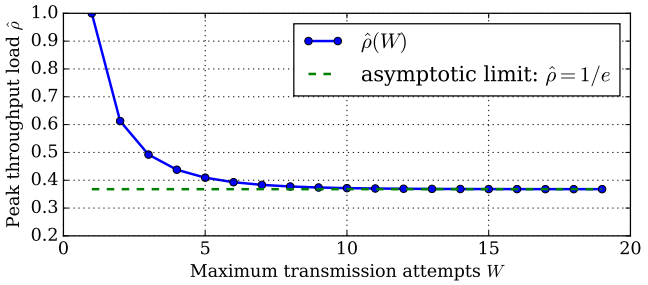


FIGURE 5. Peak throughput load value $\hat{\rho}$ achieving maximum throughput as a function of maximum number of transmission attempts W.

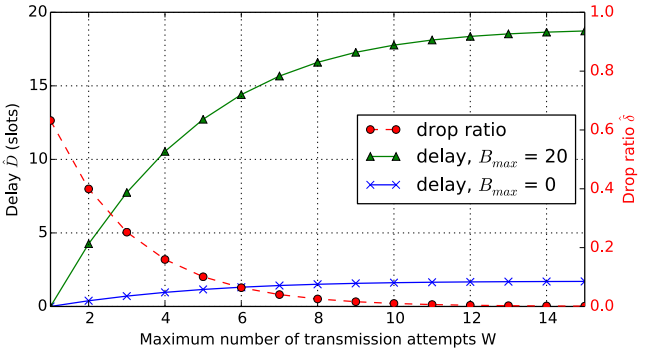


FIGURE 6. Delay $\hat{D}$ and drop ratio $\hat{\delta}$ achieved for the maximum throughput (peak throughput load $\hat{\rho}$) vs. maximum number of transmission attempts W, for different values of $B_{\max}$.

Now, we can find the value of f maximizing the throughput through differentiation

$$\frac{dT}{df} = f\frac{d(\ln(f))}{df} + \ln(f) = \ln(f) + 1 \tag{8}$$

and setting Eqn. (8) to zero. Thus,

$$f = 1/e \tag{9}$$

attains the maximum throughput. By substituting $1/e$ for f in Eqn. (6), we obtain the *peak throughput load*:

$$\hat{\rho} = \frac{1}{e(1-(1-1/e)^W)}. \tag{10}$$

We observe that $\hat{\rho}$ depends only on the number of allowed transmission attempts W, and asymptotically reaches $1/e$ for $W \to +\infty$, see Fig. 5. Consequently, substituting f with (9) in Eqns. (4), (5), and (7), we obtain for the peak throughput load:

$$\hat{D} = (1+\frac{B_{\max}}{2})(e-1) \times \times \frac{1+(W-1)(1-1/e)^W - W(1-1/e)^{W-1}}{1-(1-1/e)^W} \tag{11}$$

$$\hat{\delta} = (1-1/e)^W, \tag{12}$$

$$\hat{T} = 1/e. \tag{13}$$

The resulting dependencies are presented in Fig. 6. It is intuitively clear that increasing W increases the delay, while decreasing the drop ratio. Note that the steady-state throughput does not depend on the $B_{\max}$ with the model assumptions in [70]. (This observation does not hold for the general case with finite number of UEs in the cell or with a varying arrival rate λ.)

From Fig. 6, we also observe that, if the parameters $B_{\max}$ and W are properly chosen (e.g., $W = 8$, $B_{\max} = 0$), the system performance at the peak load point is characterized by moderate delays and low drop probability.

B. PRIORITIZATION WITH PREAMBLE SEPARATION

We now proceed to analyze the fixed-assignment preamble separation, i.e., we examine the split of the preambles into two non-overlapping sets $\mathbf{M_I}$ and $\mathbf{M_{II}}$. Since the two sets are non-overlapping, we can consider them as two independent systems. Thus, their performance metrics can be obtained via Eqns. (1)–(5), whereby we replace M in Eqns. (1) and (3) by m_I and m_{II}, respectively.

In the next subsections, we examine the separation effects in the underloaded and overloaded regions with two example cases of the total initial arrival rate: $\rho = (\lambda_I + \lambda_{II})/M = 0.25$ (point **B**) and $\rho = (\lambda_I + \lambda_{II})/M = 0.45$ (point **C** in Fig. 4). For ease of illustration, we set the absolute arrival rates of both classes to be equal, i.e., $\lambda_I = \lambda_{II}$.

1) UNDERLOADED REGION

The plots in Fig. 7 represent the performance of the RACH for class I and class II with a fixed preamble assignment (separation), where m_I (on the x-axis) is the number of preambles assigned to class I; thus, $m_{II} = M - m_I$ preambles are assigned to class II. The throughput is normalized with respect to all $M = |\mathbf{M}|$ available preambles. We observe from Fig. 7(a), that for the underloaded case there exists a region $m_l \le m \le m_r$ where the total throughput with preamble separation matches exactly the total throughput without separation. This region is bounded by the number of preambles m_l and m_r achieving the peak throughput of class I and II, respectively:

$$m_l = \left\lceil \frac{\lambda_I}{\hat{\rho}} \right\rceil \text{ and } m_r = \left\lfloor M - \frac{\lambda_{II}}{\hat{\rho}} \right\rfloor. \tag{14}$$

Hence, the width Δm of this region is:

$$\Delta m = m_r - m_l = M - \left\lceil \frac{\lambda_{II}}{\hat{\rho}} \right\rceil - \left\lceil \frac{\lambda_I}{\hat{\rho}} \right\rceil. \tag{15}$$

The region width Δm is zero, if $(\lambda_I + \lambda_{II})/M = \hat{\rho}$, i.e., when the total load equals the peak throughput load; which corresponds to point **A** in Fig. 4. Figs. 7(b) and 7(c) indicate that prioritization within this region moderately decreases the delay in one class, while keeping the drop ratio very low. Even though underloaded systems do not pose performance challenges for the RACH in practice, our analysis shows that an efficient delay-targeted prioritization for class I can be performed within the region $m_l \le m \le m_r$ without a significant performance degradation for class II.

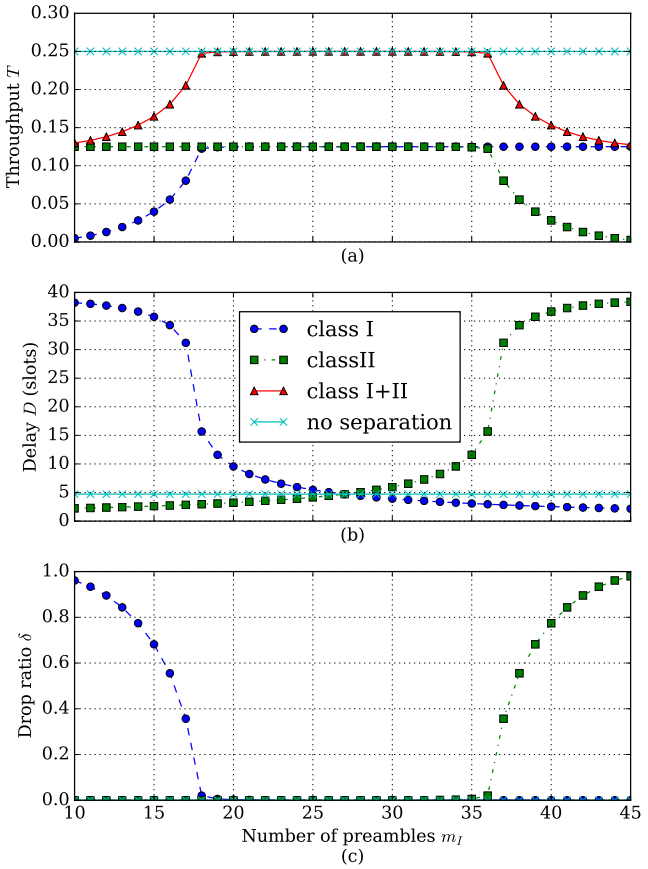


FIGURE 7. System performance vs. number m_I of preambles allocated to class I, for underloaded $\lambda = \lambda_I + \lambda_{II} = 0.25M$ scenario (point **B** in Fig. 4). Fig. (a) shows throughput T, Fig. (b) delay D, and Fig. (c) shows drop probability δ. System parameters: $M = 54$ preambles, $B_{\max} = 20$ slots, $W = 8$ transmission attempts.

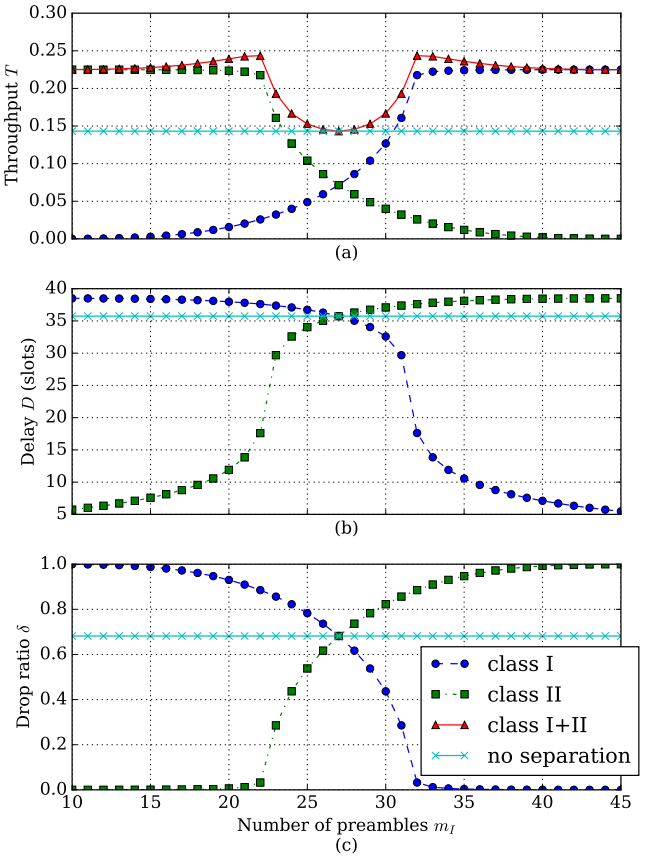


FIGURE 8. System performance vs. number of preambles allocated to class I m_I, for overloaded $\lambda = \lambda_I + \lambda_{II} = 0.45M$ scenario (point **C** in Fig. 4). Fig. (a) shows throughput T, Fig. (b) delay D, and Fig. (c) shows drop probability δ. System parameters: $B_{\max} = 20$ slots, $W = 8$.

2) OVERLOADED REGION

We first observe for the overloaded region in Fig. 8(a) that the total (aggregate) system throughput (of both class I and class II) is higher or equal to the throughput without preamble separation. There are two throughput peaks on the plot, corresponding to $m_I = 22$ and $m_I = 32$. These peaks correspond to preamble allocations maximizing the throughput of class I and class II respectively (i.e., point **A** in Fig. 4). Since $\lambda_I = \lambda_{II}$, the peaks are of equal magnitude.

For a general case, the total throughput is calculated from the throughputs of class I T_I and class II T_{II} as follows:

$$T = T_I m_I + T_{II}(M - m_I). \tag{16}$$

The maximum total throughput depends on the λ_I/λ_{II} ratio. If $\lambda_I/\lambda_{II} > 1$, then the maximum total throughput corresponds to the point when the number of preambles allocated to class I maximizes its performance. The maximum achievable total throughput is $T^{\max} = 1/e$, and is possible whenever $T_I = \hat{T}$, and $\lambda_I = \hat{\rho}M$.

Now we turn to the delay and drop ratio of the overloaded region. From Fig. 8(b), we conclude that any prioritization of class I (for $m_I > 27$) results in a significant delay decrease for class I, with a slight delay increase for class II. Importantly, the delay reduction for class I comes at the expense of an increased drop probability for class II, as shown in Fig. 8(c).

V. ALLOCATION METHODS

The goals of prioritization on the RACH can be both to increase the number accepted UEs (throughput), as well as to decrease the access delay. In this section, we consider two approaches for calculating the number m_I of preambles for the prioritized class I: based on delay requirement matching and based on throughput maximization.

A. MATCHING THE TARGET AVERAGE DELAY

If the devices in the delay-intolerant class have a common delay requirement, then it can be beneficial to dimension the RACH according to this requirement. In this study, we consider delay in slots, therefore translation into the actual time domain requires knowledge of the PRACH configuration parameters. For instance, the PRACH configuration index 7 [68], results in one RACH opportunity per frame; thus, the length of one slot is 10 ms. Following the analysis in Sec. IV, we can calculate the required minimum number of

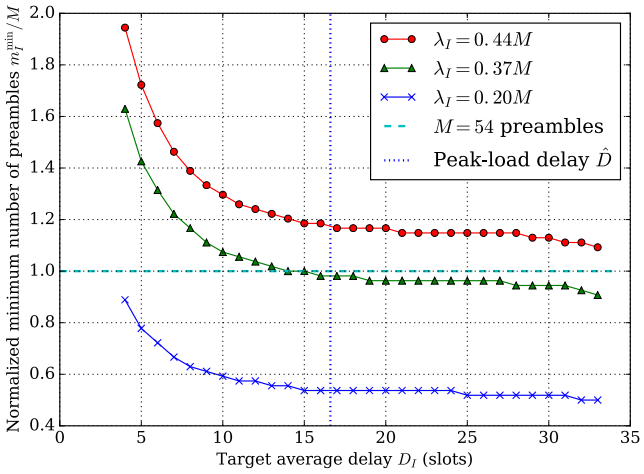


FIGURE 9. Normalized minimum number $m_I^{\min}/M$ of preambles necessary for allocation to class I (y-axis) in order to meet the delay requirement (x-axis, in slots) for different load values λ_I expressed as a fraction of the total number $M = 54$ of available preambles. The region above the $M = 54$ line is not achievable due to the insufficient available preambles. The region to the right of the peak-load delay $\hat{D}$ line is achieved in the overloaded system with high drop ratio δ. System parameters: $B_{\max} = 20$ slots, $W = 8$.

preambles $m_I^{\min}$ (see Fig. 9) in order to achieve a target average delay. Specifically, substituting x obtained from Eqn. (2) into Eqn. (1) gives

$$m_I^{\min} = \left\lceil \lambda_I \frac{(1-f)^W - 1}{f \ln(f)} \right\rceil. \tag{17}$$

There is no closed-form relation between f and a given delay requirement D. However, for a given delay requirement D, Eqn. (4) can be solved numerically for f.

The method of using target delay for allocating preambles suffers from several drawbacks. The delay parameter does not account for dropped requests δ, and, thus, does not represent a good standalone metric for the performance: for given system parameters W and $B_{\max}$, the target delay requirement can be located in the overloaded region (to the right of the peak-load delay $\hat{D}$ in the Fig. 9) and, thus, can be accompanied by a high drop ratio. If $m_I^{\min} > M$ (see $M = 54$ line in Fig. 9) the target delay cannot be achieved at all for a given λ_I. Moreover, since $B_{\max}$ has no influence on the throughput or drop ratio (see Sec. IV-A.2), a better adjustment for the average delay can be achieved through a proper $B_{\max}$ setting.

B. THROUGHPUT MAXIMIZATION: LATMAPA

Alternatively, the preamble-based prioritization can target the throughput (and corresponding drop ratio) as performance metric. The goal for setting the minimum necessary number of preambles $m^{\min}$ is to keep the throughput of the corresponding class at its highest value. The maximum throughput per preamble is achieved if λ normalized by the number of allocated preambles $m^{\min}$ is equal to the peak throughput load $\hat{\rho}$ (10), i.e., if

$$m^{\min} = \left\lceil \frac{\lambda}{\hat{\rho}} \right\rceil. \tag{18}$$

Algorithm 1 Load-Adaptive Throughput-MAximizing Preamble Allocation (LATMAPA)

procedure LATMAPA
UE req. arrival rates: λ_I for high prior. class I; λ_{II} for low prior. class II;
RACH parameters: W transm. attempts, M preambles;
Prioritization factor r, $r \in [0, 1]$;
Calculate $m_I^{\min}$, $m_{II}^{\min}$ for λ_I, λ_{II} via Eqn. (19)
  if $m_I^{\min} \leq M - m_{II}^{\min}$ **then**
    $m_{II} \leftarrow m_{II}^{\min}$; $m_I \leftarrow M - m_{II}$
  else
    $m_{II} \leftarrow \max\left(\left\lceil \frac{r M m_{II}^{\min}}{m_I^{\min} + m_{II}^{\min}} \right\rceil, M - m_I^{\min}\right)$
    $m_I \leftarrow M - m_{II}$
  end if
return Preamble numbers for classes I and II: m_I, m_{II}
end procedure

From Eqn. (18) and (10):

$$m^{\min} = \lceil \lambda e (1 - (1 - 1/e)^W) \rceil. \tag{19}$$

As shown in Figs. 4 and 6, the drop ratio can be kept low as long as the throughput of class I remains less than or equal to the peak throughput. However, if we allocate more than $m_I^{\min}$ preambles to class I, the overall throughput may decrease while having almost no effect on the throughput and drop ratio of class I. Thus, by choosing $m_I > m_I^{\min}$, the performance of class II may be unnecessarily degraded.

Following these observations, we propose the Load-Adaptive Throughput MAximizing Preamble Allocation (LATMAPA) for determining the necessary amount of preambles (see Algorithm 1). LATMAPA requires UE request arrival rate estimates which can be obtained with combinations of existing short [45] and long [74], [75] timescale prediction techniques. The core idea of LATMAPA is that for the given arrival rates λ_I, λ_{II} we calculate the respective necessary number of preambles $m_I^{\min}$, $m_{II}^{\min}$ using Eqn. (19). If there are enough resources to meet the demand of both classes (underloaded case, Sec. IV-B.1), i.e., if $M \geq m_I^{\min} + m_{II}^{\min}$, then we allocate to class II its required number of preambles, i.e., $m_{II} = m_{II}^{\min}$, and allocate the remaining preambles to class I:

$$m_I = M - m_{II}^{\min}. \tag{20}$$

Thus, the number m_I of preambles allocated to class I is a least as large as necessary ($m_I^{\min}$). Hence, class I is prioritized compared to class II.

Next, consider the overloaded case (see Section IV-B.2) when there are not enough preambles to satisfy the demand of both classes, i.e., when $M < m_I^{\min} + m_{II}^{\min}$. In order to maintain a prescribed level of performance for class II, we introduce a prioritization factor r, $r \in [0, 1]$, that regulates the minimum number of preambles allocated to class II. In particular, we allocate to class II the portion r of the proportional allocation

of the M preambles according to the ratio $m_{II}^{\min}/(m_I^{\min}+m_{II}^{\min})$ of the required preambles for classes I and II, i.e., we allocate $rMm_{II}^{\min}/(m_I^{\min}+m_{II}^{\min})$ preambles to class II. On the other hand, if the prioritization factor r is so low that the allocation according to r would gives less preambles to class II than are left after allocating $m_I^{\min}$ preambles to class I, then we allocate the remaining $M-m_I^{\min}$ preambles to class II. Thus, overall, we allocate the number of preambles specified in Step 9. of Algorithm 1 to class II. As specified in Step 10. of Algorithm 1, we then allocate the remaining $M-m_{II}$ preambles to class I.

VI. EVALUATION

A. SIMULATION SET-UP

We implemented the simulation models with an event-based OMNeT++ framework (C++) [76]. We collected and processed the statistics with Python-based open-source SciPy [77] libraries. We simulated the random access procedure at the level of detail corresponding to our model. In particular, we simulated one eNB with either the infinite-source (UE) assumption (i.e., back-logged request do not reduce the arrival rate) or with a finite large number $N \in \{1000, 5000, 10000, 30000\}$ of UEs. An RA request is considered as collided if two or more UEs select the same preamble in the same time slot. No propagation or interference effects are considered. The 95 % confidence intervals are less than 3 % of the corresponding sample means and are not plotted to avoid visual clutter.

B. LATMAPA: ANALYSIS VS. SIMULATION

Fig. 10 shows the LATMAPA performance as a function of the normalized class I arrival rate $\rho_I = \lambda_I/M$ for a fixed class II arrival rate $\rho_{II} = \lambda_{II}/M = 0.15$. We have set the prioritization factor to the relatively small value $r = 0.02$ so as to initially consider a scenario with pronounced prioritization. We examine the impact of r in detail in Section VI-D. In Fig. 10, we compare our analytical model for an infinite UE population with simulations for finite UE populations.

From Fig. 10(a), we observe that for increasing class I traffic load ρ_I, LATMAPA sustains a nearly linearly increasing class I throughput almost up to the load point $\rho_I = \hat{\rho} \approx 0.37$. Note that at the $\hat{\rho}$ load point, the number $m_I^{\min}$ of preambles required for class I reaches the total number of available preambles M. We observe from Fig. 10(c) that the class II throughput starts to drop when the class I load approaches $\hat{\rho} - \rho_{II} = 0.22$. This is because the pronounced prioritization for the considered small $r = 0.02$ adaptively takes preambles from the low-priority class II and assigns the preambles to the high-priority class I as the class I traffic load increases.

Similarly, we observe from Fig. 10(c) that LATMAPA maintains a nearly constant high class II throughput until the total required number of preambles $m_I^{\min} + m_{II}^{\min}$ exceeds the number of available preambles M, i.e., until the RACH becomes overloaded.

We observe a positive side effect of prioritization with LATMAPA in Fig. 10e, which shows the total throughput for both classes. In the overloaded region, we observe that prioritizing class I leads to an increase of the total throughput with LATMAPA compared to the total throughput without separation (which is plotted as the "No separation, inf, ana" curve). This throughput increase achieved with LATMAPA prioritization corresponds to the throughput increase achieved with preamble separation in the overloaded region (see Sec. IV-B.2 and Fig. 8(a)). We also observe a slight "dip" (decrease) in the total throughput in the load range between $\hat{\rho} - \rho_{II} = 0.22$ and $\hat{\rho}$. This dip effect is due to different slopes to the left and right of the maximum throughput region (**A** in Fig. 4): class II throughput decreases faster (slope to the right of **A**) than class I gains throughput (slope to the left of **A**).

Regarding the accuracy of the analysis, we observe from Fig. 10 that the simulation for the infinite UE population model essentially coincides with the analysis for the infinite UE population model. We also observe from Fig. 10(a), (c), and (e) that the finite UE population throughputs are approximated by the infinite UE population analysis. The discrepancy in throughputs between simulation and analysis increases with decreasing number of UEs. However, the analysis gives a meaningful approximation and lower throughput bound down to 10,000 UEs. We observe from Figs. 10(b) and (d) that the drop ratios from the finite-UE simulations deviate significantly from the analytical infinite-UE results. However, the infinite-UE analysis provides an upper bound of the drop ratios.

Importantly, LATMAPA inherently excludes any cross impact between the two UE classes, i.e., the QoS levels of the two request classes are isolated from each other. Therefore, QoS, i.e., low drop ratios and, hence, low delays, can be guaranteed for class I as long as there are enough preambles (i.e., for low r, we need $M \geq m_I$). The QoS level isolation achieved with preamble separation is fundamentally different from prioritization methods that manipulate the random access on a given set of preambles, e.g., methods that manipulate the access barring, backoff window, or number of transmission attempts, because these prioritization methods do not eliminate contention of the different classes for the same set of preambles. Also, the preamble separation approach allows for effective prioritization during long periods of overload and for steady-state operation, where the access barring based approaches fail [31].

C. LATMAPA: COMPARISON WITH OTHER ALLOCATION METHODS

We compare LATMAPA with the two existing load adaptive preamble allocation mechanisms in [60] and [61]. With ρ_I and ρ_{II} denoting the normalized loads of high-priority class I and low-priority class II UE requests, respectively, the Zhao2014 allocation mechanism [60] allocates $m_I = \min\{\lfloor 1.5\rho_I M \rfloor, \lfloor Mw\rho_I/(\rho_I+\rho_{II})\rfloor\}$ preambles to the high-priority class I. The weight parameter w is varied in the range (0, 10]. The remaining $m_{II} = M - m_I$ preambles are allocated to the low-priority class II.

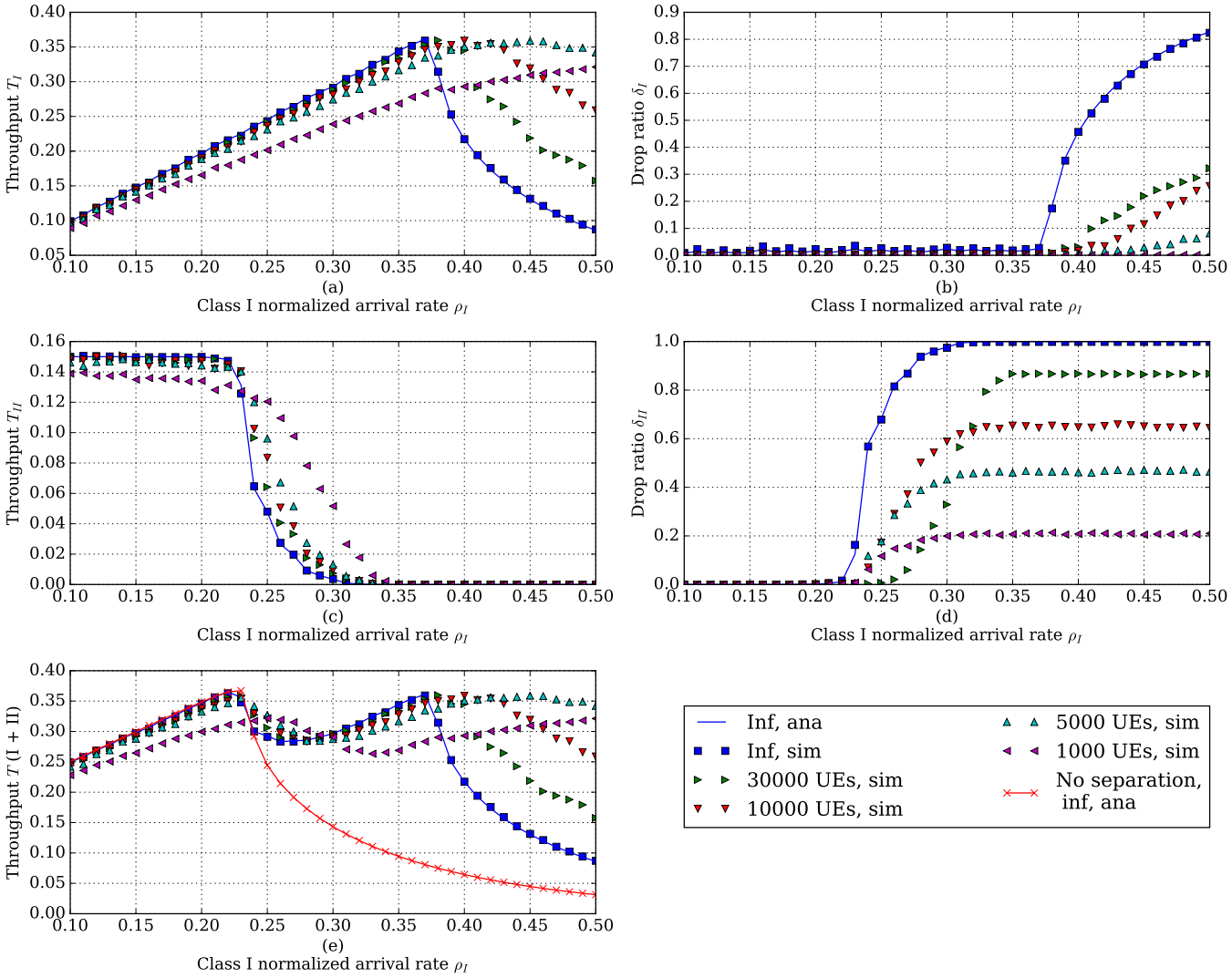


FIGURE 10. LATMAPA throughput T and drop ratio δ for class I (a, b), class II (c, d) and the total throughput (e) as a function of normalized class I arrival rate $\rho_I = \lambda_I/M$; $\rho_{II} = \lambda_{II}/M = 0.15$, fixed; System parameters: $M = 54$ preambles, $W = 8$, $r = 0.02$. LATMAPA is used to calculate the number of preambles m_I and m_{II}. Model verification for infinite and finite number of UEs.

The Du2016 allocation mechanism [61] considers the access barring factor b_I, and the number x_I of contending UEs in a given slot for class I. The Du2016 approach calculates an optimal split $\beta^\star = m_I/m_{II}$ as follows:

$$\beta^\star = \begin{cases} \dfrac{x_I(1-b_I)}{M\log(x_I(1-b_I)) - x_I(1-b_I)} \\ \qquad \text{if } x_I(1-b_I) \in [3, +\infty) \\ \dfrac{x_I(1-b_I)}{M\log(x_I(1-b_I)/2) - x_I(1-b_I)} \\ \qquad \text{if } x_I(1-b_I) \in (1, 3). \end{cases} \tag{21}$$

From the optimal split $\beta^\star$, the Du2016 approach allocates $m_I = M\beta^\star/(1+\beta^\star)$ preambles to the high-priority class I and $m_{II} = M - m_I$ preambles to the low-priority class II. Note that the Du2016 approach utilizes information about the exact number of contending UE requests in the upcoming slot. It is not realistic to obtain this number for every slot; however, the expected number of contending UE requests can be obtained as a function of the arrival rate λ_I by numerically solving Eqns. (1) and (2). For a fair comparison with LATMAPA and Zhao2014, we use this expected number of arrivals x_I for obtaining the optimal split as in Eqn. (21), and set the barring factor $b_I = 0$.

We compare the preamble allocation for class I resulting from LATMAPA, Zhao2014 [60], and Du2016 [61] in Fig. 11, with the fixed class II arrival rate $\rho_{II} = 0.2$. We observe that both Zhao2014 and Du2016 do not allocate enough preambles to the high priority class I. For Zhao2014 [60], we observe that changing the weight parameter w only influences the allocation for high loads $\rho_I \geq 0.35$. In contrast to the Zhao2014 and Du2016 allocation methods, LATMAPA allocates the required minimum number m_I^{min} of preambles to the high-priority class I as long as the available number of preambles M and traffic load ρ_I

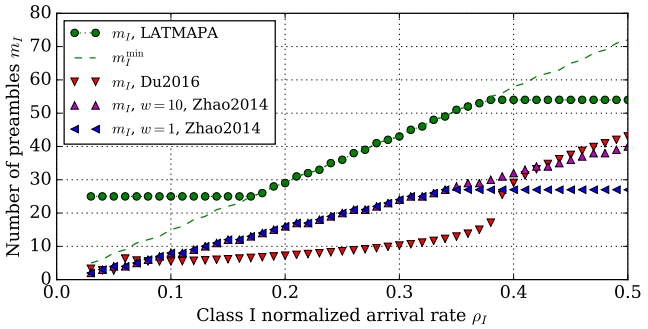


FIGURE 11. Comparison of the number of preambles m_I allocated to the high priority class I by LATMAPA, Zhao2014 [60], and Du2016 [61] as a function of the normalized class I UE request arrival rate ρ_I. The minimum required number of preambles $m_I^{\min}$ from Eqn. (17) is plotted as a reference. Class II load is kept constant at $\rho_{II} = 0.2$.

permit; hence, LATMAPA more effectively prioritizes the high-priority class I traffic than the prior Zhao2014 and Du2016 approaches.

D. LATMAPA: IMPACT OF PRIORITIZATION FACTOR r

The prioritization factor $r \in [0, 1]$ controls the minimum level of service provided to class II. It only plays a role if the overall number of preambles is insufficient to satisfy the traffic load of both classes. That is, if $r = 1$, the available preambles are allocated proportionally to two classes. If $0 < r < 1$, class II only obtains an r portion of the proportional preamble allocation. In the other extreme case, if $r = 0$, class I gets all available preambles.

Fig. 12(a) shows the number of allocated preambles as a function of the class I normalized arrival rate (load) ρ_I, with class II arrival rate fixed at $\rho_{II} = 0.2$. We observe that the preamble allocation is static until the arrival rate reaches the point where $m_I^{\min} + m_{II}^{\min} = M$ at $\rho_I = 0.17$: class II gets only the necessary number of $m_{II} = m_{II}^{\min}$ preambles, and class I receives the remaining $m_I = M - m_{II}^{\min}$ preambles.

The prioritization factor starts playing a role once the arrival rate of class I increases above $\rho_I = 0.18$. For $r = 1$, both classes are treated equally and share the available M preambles proportionally to their loads ρ_I and ρ_{II}. For $r = 0.6$, we observe a shift in the preamble allocation towards class I: class I is prioritized, hence, the gap between m_I and m_{II} is larger than for $r = 1$. For $r = 0.0$, class I is first allocated its required minimum number of preambles $m_I^{\min}$ (up to the available M preambles), and any remaining preambles are allocated to class II. Thus, the setting $r = 0$ corresponds to strict prioritization.

E. TUNING PRACH CONFIGURATION INDEX

In the preceding sections, we have analyzed and evaluated scenarios for a prescribed fixed PRACH configuration index. However, practical scenarios require the tuning of the PRACH configuration index, which corresponds to the number of PRACH slots available in a given frame [38], [78]. Our model can be readily extended to tune the PRACH

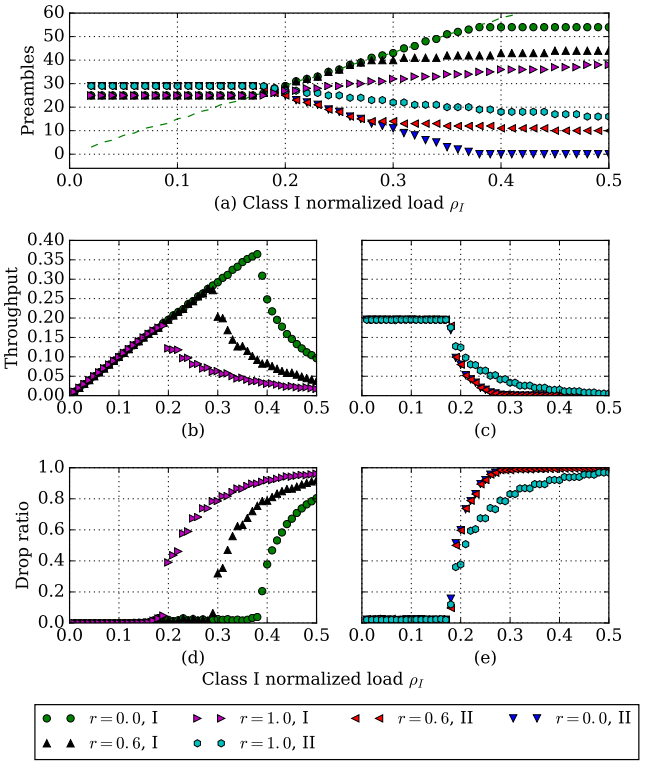


FIGURE 12. LATMAPA performance with different prioritization factors r: (a) number of preambles allocated to classes I and II; (b), (c) throughput of class I and II respectively; (d), (e) drop ratio for class I and II respectively. X-axis is class I normalized arrival rate ρ_I; class II traffic load $\rho_{II} = 0.2$, fixed.

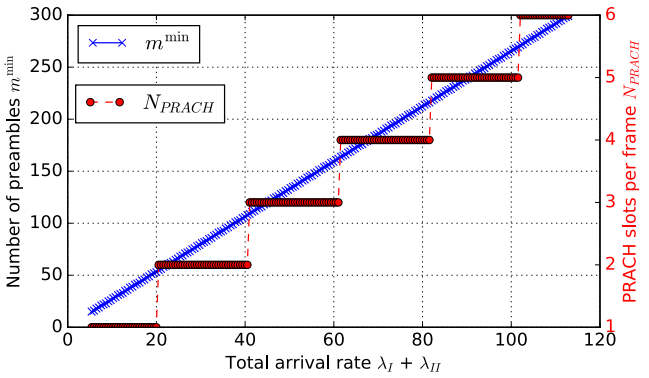


FIGURE 13. Minimum number of required preambles $m^{\min}$ and PRACH slots per frame N_{PRACH} as a function of total arrival rate $\lambda_I + \lambda_{II}$ [arrivals per frame].

configuration index. The tuning allows to choose the optimal index in order to properly provision the channel. In particular, if $m^{\min} = m_I^{\min} + m_{II}^{\min} \geq M$, then a larger number of PRACH slots per frame N_{PRACH} is needed:

$$N_{\text{PRACH}} = \left\lceil \frac{m_I^{\min} + m_{II}^{\min}}{M} \right\rceil. \tag{22}$$

Fig. 13 shows the required minimum number of preambles $m^{\min}$ and the required number of PRACH slots per frame N_{PRACH} as a function of the total arrival rate $\lambda_I + \lambda_{II}$. The number of PRACH slots can be used to determine the

PRACH configuration index, e.g., index 0 for $N_{\text{PRACH}} = 1$ or index 12 for $N_{\text{PRACH}} = 5$.

F. REMARK ON COMPARISON WITH RANDOM ACCESS PROCEDURE MANIPULATION METHODS

Generally, LTE random access performance has been studied for two main settings: constant (steady-state) traffic, where the system behavior is studied for long periods of constant UE request load, and bursty traffic, where the system is studied for temporary (sudden) overload periods. The constant traffic studies have mainly focused on evaluating steady-state performance aspects and influencing parameters [70]. On the other hand, the bursty traffic studies have focused on methods for the efficient resolution of large amounts of simultaneous (one-shot) or nearly simultaneous (mostly modeled as beta-distributed with a prescribed activation time) UE request arrivals [31]. State-of-the-art methods for prioritizing random access through the manipulation of the random access procedures on a given set of preambles, such as Extended Access Barring (EAB), belong to the category of bursty traffic studies. That is, random access parameter manipulation methods, such as EAB, have been developed for temporary, non-persistent overload conditions. Thus, these random access parameter manipulation methods are not suitable for addressing persistent, constant overload conditions (which are the focus of this present study). For instance, studies [70], [72] have demonstrated that neither access barring or tuning of the back-off parameters change the steady-state throughput or drop ratio of systems with constant traffic loads. Therefore, LATMAPA, which has been developed for steady-state (constant) traffic, can not be directly compared with prioritization methods that manipulate the random access procedures on a given set of preambles so as to address non-persistent traffic bursts. However, future research can explore combinations of these two types of approaches so as to address non-persistent traffic bursts that are superimposed on steady-state traffic.

VII. CONCLUSION

For the setting of steady-state (constant) UE request arrival (load) traffic in future 5G wireless systems that have evolved from LTE-A, we have examined the separation of the preambles into into two classes, a high-priority class I and a low-priority class II. For underloaded traffic conditions we have determined a safe prioritization region Δm, within which delay decreases for class I are not accompanied by noticeable performance degradations for class II. For overloaded traffic conditions we have demonstrated that preamble separation can increase the total (aggregate) throughput. Prioritization of class I in the overloaded region comes at the cost of increasing the ratio of dropped requests for class II, but can significantly decrease the delay and increase the throughput for class I.

We have further investigated two possible preamble allocation methods for prioritization. The first approach matches the average access delay of the prioritized class, but turned out to be not practical. The second method, Load-Adaptive Throughput-MAximizing Preamble Allocation (LATMAPA) strives to maximize the system throughput. We demonstrated that LATMAPA gives favorable performance up to the exhaustion of available preambles by the prioritized class I.

Future research can investigate the combination of our LATMAPA preamble separation approach for steady-state (constant) overload traffic conditions with methods that manipulate the random access on a given set of preambles, such as Extended Access Barring, for mitigating temporary arrival bursts. Moreover, developing an analytical model to account for a finite number of UEs in the cell can bring additional insights into the LATMAPA dynamics. There is also a need to design a practical protocol for informing UEs about the available $\mathbf{M_I}$ and $\mathbf{M_{II}}$ preamble sets, which could be achieved by adding the preamble set information to the broadcasted system information blocks. More broadly, the preamble separation could be explored in future research as a dimension for virtualizing wireless network services with differentiated grades of service [79]–[82] in emerging software defined access networks [83]–[86].

REFERENCES

[1] C. Bockelmann *et al.*, "Massive machine-type communications in 5G: Physical and MAC-layer solutions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 59–65, Sep. 2016.
[2] A. Gupta and R. K. Jha, "A survey of 5G network: Architecture and emerging technologies," *IEEE Access*, vol. 3, pp. 1206–1232, 2015.
[3] Y. Liu, Z. Yang, R. Yu, Y. Xiang, and S. Xie, "An efficient MAC protocol with adaptive energy harvesting for machine-to-machine networks," *IEEE Access*, vol. 3, pp. 358–367, 2015.
[4] N. K. Pratas, C. Stefanovic, G. C. Madueño, and P. Popovski, "Random access for machine-type communication based on Bloom filtering," in *Proc. IEEE GLOBECOM*, 2016.
[5] N. K. Pratas, S. Pattathil, C. Stefanovic, and P. Popovski. (Oct. 2016). "Massive machine-type communication (MMTC) access with integrated authentication." [Online]. Available: https://arxiv.org/abs/1610.09849
[6] Y. Qi, A. U. Quddus, M. A. Imran, and R. Tafazolli, "Semi-persistent RRC protocol for machine-type communication devices in LTE networks," *IEEE Access*, vol. 3, pp. 864–874, 2015.
[7] M. Polese, M. Centenaro, A. Zanella, and M. Zorzi, "M2M massive access in LTE: RACH performance evaluation in a smart city scenario," in *Proc. IEEE ICC*, May 2016, pp. 1–6.
[8] G. Wu, S. Talwar, K. Johnsson, N. Himayat, and K. D. Johnson, "M2M: From mobile to embedded Internet," *IEEE Commun. Mag.*, vol. 49, no. 4, pp. 36–43, Apr. 2011.
[9] V. C. Gungor *et al.*, "Smart grid technologies: Communication technologies and standards," *IEEE Trans. Ind. Informat.*, vol. 7, no. 4, pp. 529–539, Nov. 2011.
[10] A. A. Khan, M. H. Rehmani, and M. Reisslein, "Cognitive radio for smart grids: Survey of architectures, spectrum sensing mechanisms, and networking protocols," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 860–898, 1st Quart., 2016.
[11] Y. Zhang, R. Yu, M. Nekovee, Y. Liu, S. Xie, and S. Gjessing, "Cognitive machine-to-machine communications: Visions and potentials for the smart grid," *IEEE Netw.*, vol. 26, no. 3, pp. 6–13, May/Jun. 2012.
[12] G. Araniti, C. Campolo, M. Condoluci, A. Iera, and A. Molinaro, "LTE for vehicular networking: A survey," *IEEE Commun. Mag.*, vol. 51, no. 5, pp. 148–157, May 2013.
[13] G. Fodor *et al.*, "An overview of device-to-device communications technology components in METIS," *IEEE Access*, vol. 4, pp. 3288–3299, 2016.
[14] P. Kela, J. Turkka, and M. Costa, "Borderless mobility in 5G outdoor ultra-dense networks," *IEEE Access*, vol. 3, pp. 1462–1476, 2015.
[15] A. Osseiran *et al.*, "Scenarios for 5G mobile and wireless communications: The vision of the METIS project," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 26–35, May 2014.

[16] A. Asheralieva and Y. Miyanaga, "Effective resource block allocation procedure for quality of service provisioning in a single-operator heterogeneous LTE-A network," *Comput. Netw.*, vol. 108, pp. 1–14, Oct. 2016.

[17] C. Kalalas, L. Thrybom, and J. Alonso-Zarate, "Cellular communications for smart grid neighborhood area networks: A survey," *IEEE Access*, vol. 4, pp. 1469–1493, 2016.

[18] T. A. Levanen, J. Pirskanen, T. Koskela, J. Talvitie, and M. Valkama, "Radio interface evolution towards 5G and enhanced local area communications," *IEEE Access*, vol. 2, pp. 1005–1029, 2014.

[19] A. Ijaz *et al.*, "Enabling massive IoT in 5G and beyond systems: PHY radio frame design considerations," *IEEE Access*, vol. 4, pp. 3322–3339, 2016.

[20] M. S. Ali, E. Hossain, and D. I. Kim, "LTE/LTE-A random access for massive machine-type communications in smart cities," *IEEE Commun. Mag.*, vol. 55, no. 1, pp. 73–83, Jan. 2017.

[21] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the random access channel of LTE and LTE-A suitable for M2M communications? A survey of alternatives," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 4–16, 1st Quart., 2014.

[22] K. Zheng, S. Ou, J. Alonso-Zarate, M. Dohler, F. Liu, and H. Zhu, "Challenges of massive access in highly dense LTE-advanced networks with machine-to-machine communications," *IEEE Wireless Commun.*, vol. 21, no. 3, pp. 12–18, Jun. 2014.

[23] M. Gerasimenko, V. Petrov, O. Galinina, S. Andreev, and Y. Koucheryavy, "Impact of machine-type communications on energy and delay performance of random access channel in LTE-advanced," *Trans. Emerg. Telecommun. Technol.*, vol. 24, no. 4, pp. 366–377, 2013.

[24] *Policy and Charging Control Architecture*, 3GPP document 23.203, 3rd Generation Partnership Project, Jun. 2015.

[25] T. P. de Andrade, C. A. Astudillo, and N. L. da Fonseca, "Allocation of control resources for machine-to-machine and human-to-human communications over LTE/LTE-A networks," *IEEE Internet Things J.*, vol. 3, no. 3, pp. 366–377, Jun. 2016.

[26] G. Foddis, R. G. Garroppo, S. Giordano, G. Procissi, S. Roma, and S. Topazzi, "On RACH preambles separation between human and machine type communication," in *Proc. IEEE ICC*, May 2016, pp. 1–6.

[27] K.-D. Lee, S. Kim, and B. Yi, "Throughput comparison of random access methods for M2M service over LTE networks," in *Proc. IEEE GLOBECOM Workshops (GC Wkshps)*, Dec. 2011, pp. 373–377.

[28] *Study on RAN Improvements for Machine-Type Communications (Release 11) 3GPP Technical Specification Group Radio Access Network*, 3GPP document 37.868, 3rd Generation Partnership Project, Sep. 2011.

[29] Y.-H. Hsu, K. Wang, and Y.-C. Tseng, "Efficient cooperative access class barring with load balancing and traffic adaptive radio resource management for M2M communications over LTE-A," *Comput. Netw.*, vol. 73, pp. 268–281, Nov. 2014.

[30] S.-Y. Lien, T.-H. Liau, C.-Y. Kao, and K.-C. Chen, "Cooperative access class barring for machine-to-machine communications," *IEEE Trans. Wireless Commun.*, vol. 11, no. 1, pp. 27–32, Jan. 2012.

[31] S. Duan, V. Shah-Mansouri, Z. Wang, and V. Wong, "D-ACB: Adaptive congestion control algorithm for bursty M2M traffic in LTE networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 9847–9861, Dec. 2016.

[32] L. Song, W. Zhou, Y. Hou, and M. Gao, "Load-aware ACB scheme for M2M traffic in LTE-A networks," in *Proc. 11th Int. Conf. Broad-Band Wireless Comput., Commun. Appl. (BWCCA)*, Nov. 2016.

[33] F. Morvari and A. Ghasemi, "Two-stage resource allocation for random access M2M communications in LTE network," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 982–985, May 2016.

[34] X. Yang, A. Fapojuwo, and E. Egbogah, "Performance analysis and parameter optimization of random access backoff algorithm in LTE," in *Proc. IEEE Veh. Technol. Conf. (VTC-Fall)*, Sep. 2012, pp. 1–5.

[35] J. Park and Y. Lim, "Adaptive access class barring method for machine generated communications," *Mobile Inf. Syst.*, vol. 2016, Jul. 2016, Art. no. 6923542.

[36] M. Hasan, E. Hossain, and D. Niyato, "Random access for machine-to-machine communication in LTE-advanced networks: Issues and approaches," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 86–93, Jun. 2013.

[37] A. Lo, Y. W. Law, M. Jacobsson, and M. Kucharzak, "Enhanced LTE-advanced random-access mechanism for massive machine-to-machine (M2M) communications," in *Proc. 27th WWRF Meeting*, 2011, pp. 1–5.

[38] O. N. C. Yilmaz, J. Hämäläinen, and S. Hämäläinen, "Self-optimization of random access channel in 3rd generation partnership project long term evolution," *Wireless Commun. Mobile Comput.*, vol. 11, no. 12, pp. 1507–1517, 2011.

[39] N. K. Pratas, H. Thomsen, Č. Stefanović, and P. Popovski, "Code-expanded random access for machine-type communications," in *Proc. IEEE Globecom Workshops*, Dec. 2012, pp. 1681–1686.

[40] M. Condoluci, G. Araniti, M. Dohler, A. Iera, and A. Molinaro, "Virtual code resource allocation for energy-aware MTC access over 5G systems," *Ad Hoc Netw.*, vol. 43, pp. 3–15, Jun. 2016.

[41] M. Condoluci, M. Dohler, G. Araniti, A. Molinaro, and J. Sachs, "Enhanced radio access and data transmission procedures facilitating industry-compliant machine-type communications over LTE-based 5G networks," *IEEE Wireless Commun.*, vol. 23, no. 1, pp. 56–63, Feb. 2016.

[42] B. S. Tsybakov and V. A. Mikhailov, "Free synchronous packet access in a broadcast channel with feedback," *Problemy Peredachi Inf.*, vol. 14, no. 4, pp. 32–59, 1978.

[43] J. Capetanakis, "Tree algorithms for packet broadcast channels," *IEEE Trans. Inf. Theory*, vol. 25, no. 5, pp. 505–515, Sep. 1979.

[44] G. C. Madueño, S. Stefanovic, and P. Popovski, "Efficient LTE access with collision resolution for massive M2M communications," in *Proc. IEEE Globecom Workshops*, Dec. 2014, pp. 1433–1438.

[45] G. C. Madueño, N. K. Pratas, Č. Stefanović, and P. Popovski, "Massive M2M access with reliability guarantees in LTE systems," in *Proc. IEEE ICC*, Jun. 2015, pp. 2997–3002.

[46] M. C. Chuah, O.-C. Yue, and Q. Zhang, "Methods and apparatus for random backoff based access priority in a communications system," U.S. Patent 6 594 240, Jul. 15, 2003.

[47] M. E. Rivero-Ángeles, D. Lara-Rodríguez, and F. A. Cruz-Pérez, "Differentiated backoff strategies for prioritized random access delay in multiservice cellular networks," *IEEE Trans. Veh. Technol.*, vol. 58, no. 1, pp. 381–397, Jan. 2009.

[48] J.-B. Seo and V. C. Leung, "Design and analysis of backoff algorithms for random access channels in UMTS-LTE and IEEE 802.16 systems," *IEEE Trans. Veh. Technol.*, vol. 60, no. 8, pp. 3975–3989, Oct. 2011.

[49] Y. Xiao, "Performance analysis of priority schemes for IEEE 802.11 and IEEE 802.11e wireless LANs," *IEEE Trans. Wireless Commun.*, vol. 4, no. 4, pp. 1506–1515, Jul. 2005.

[50] Z. Alavikia and A. Ghasemi, "Overload control in the network domain of LTE/LTE-A based machine type communications," in *Proc. Wireless Netw.*, 2017, pp. 1–16.

[51] N. Zangar, S. Gharbi, and M. Abdennebi, "Service differentiation strategy based on MACB factor for M2M communications in LTE-A networks," in *Proc. IEEE Annu. Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2016, pp. 693–698.

[52] N. Hu, X.-L. Li, and Q.-N. Ren, "Random access preamble assignment algorithm of TD-LTE," in *Advances in Computer, Communication, Control and Automation*. Berlin, Germany: Springer, 2011, pp. 701–708.

[53] D. Kim, W. Kim, and S. An, "Adaptive random access preamble split in LTE," in *Proc. Int. Conf. Wireless Commun. Mobile Comput. (IWCMC)*, Jul. 2013, pp. 814–819.

[54] Y.-Y. Chu, R. Harwahyu, R.-G. Cheng, and C.-H. Wei, "Study of generalized resource allocation scheme for multichannel slotted ALOHA systems," in *Proc. IEEE PIMRC*, Sep. 2015, pp. 1702–1706.

[55] K.-D. Lee, M. Reisslein, K. Ryu, and S. Kim, "Handling randomness of multi-class random access loads in LTE-advanced network supporting small data applications," in *Proc. IEEE Globecom Workshops*, Dec. 2012, pp. 436–440.

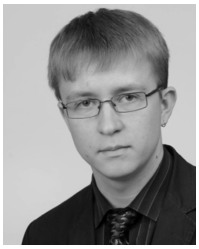
[56] C. Kalalas, F. Vazquez-Gallego, and J. Alonso-Zarate, "Handling mission-critical communication in smart grid distribution automation services through LTE," in *Proc. IEEE Int. Conf. Smart Grid Commun.*, Nov. 2016, pp. 399–404.

[57] J.-P. Cheng, C.-H. Lee, and T.-M. Lin, "Prioritized random access with dynamic access barring for RAN overload in 3GPP LTE-A networks," in *Proc. IEEE GLOBECOM Workshops*, Dec. 2011, pp. 368–372.

[58] T.-M. Lin, C.-H. Lee, J.-P. Cheng, and W.-T. Chen, "PRADA: Prioritized random access with dynamic access barring for MTC in 3GPP LTE-A networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 5, pp. 2467–2472, May 2014.

[59] C. Chun-Yuan, Y.-H. Chen, Y.-X. Zheng, and F. Yu-Chuan, "Prioritized random access method," U.S. Patent 8 705 352, Apr. 22, 2014.

[60] X. Zhao, J. Zhai, and G. Fang, "An access priority level based random access scheme for QoS guarantee in TD-LTE-A systems," in *Proc. IEEE Veh. Technol. Conf. (VTC-Fall)*, Sep. 2014, pp. 1–5.
[61] Q. Du, W. Li, L. Liu, P. Ren, Y. Wang, and L. Sun, "Dynamic RACH partition for massive access of differentiated M2M services," *Sensors*, vol. 16, no. 4, pp. 455.1–455.19, 2016.
[62] P. Osti, P. Lassila, S. Aalto, A. Larmo, and T. Tirronen, "Analysis of PDCCH performance for M2M traffic in LTE," *IEEE Trans. Veh. Technol.*, vol. 63, no. 9, pp. 4357–4371, Nov. 2014.
[63] J. J. Nielsen, D. M. Kim, G. C. Madueño, N. K. Pratas, and P. Popovski, "A tractable model of the LTE access reservation procedure for machine-type communications," in *Proc. IEEE GLOBECOM*, Dec. 2015, pp. 1–6.
[64] G. C. Madueño, J. J. Nielsen, D. M. Kim, N. K. Pratas, Č. Stefanović, and P. Popovski, "Assessment of LTE wireless access for monitoring of energy distribution in the smart grid," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 675–688, Mar. 2016.
[65] D. Niyato, P. Wang, and D. I. Kim, "Performance modeling and analysis of heterogeneous machine type communications," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2836–2849, May 2014.
[66] M. Vilgelm and W. Kellerer, "Impact of request aggregation on machine type connection establishment in LTE-Advanced," in *Proc. IEEE WCNC*, Mar. 2017.
[67] D. Tsolkas, N. Passas, and L. Merakos, "Device discovery in LTE networks: A radio access perspective," *Comput. Netw.*, vol. 106, pp. 245–259, Sep. 2016.
[68] *Evolved Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation (Release 12)*, document 3GPP 36.211, 3rd Generation Partnership Project, Aug. 2015.
[69] O. Arouk and A. Ksentini, "General model for RACH procedure performance analysis," *IEEE Commun. Lett.*, vol. 20, no. 2, pp. 372–375, Feb. 2016.
[70] R. Tyagi, F. Aurzada, K.-D. Lee, S. Kim, and M. Reisslein, "Impact of retransmission limit on preamble contention in LTE-advanced network," *IEEE Syst. J.*, vol. 9, no. 3, pp. 752–765, Sep. 2015.
[71] K. S. Ko, M. J. Kim, K. Y. Bae, D. K. Sung, J. H. Kim, and J. Y. Ahn, "A novel random access for fixed-location machine-to-machine communications in OFDMA based systems," *IEEE Commun. Lett.*, vol. 16, no. 9, pp. 1428–1431, Sep. 2012.
[72] R. Tyagi, F. Aurzada, K.-D. Lee, and M. Reisslein, "Connection establishment in LTE-A networks: Justification of Poisson process modeling," *IEEE Syst. J.*, to be published.
[73] J. H. Sarker and S. J. Halme, "An optimum retransmission cut-off scheme for slotted ALOHA," *Wireless Pers. Commun.*, vol. 13, nos. 1–2, pp. 185–202, 2000.
[74] S. Choi, W. Lee, D. Kim, K.-J. Park, S. Choi, and K.-Y. Han, "Automatic configuration of random access channel parameters in LTE systems," in *Proc. IFIP Wireless Days*, Oct. 2011, pp. 1–6.
[75] G.-Y. Lin, S.-R. Chang, and H.-Y. Wei, "Estimation and adaptation for bursty LTE random access," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2560–2577, Apr. 2016.
[76] A. Varga, "The OMNeT++ discrete event simulation system," in *Proc. Eur. Simulation Multiconf. (ESM)*, vol. 9. 2001, no. S185, p. 65.
[77] E. Jones *et al. SciPy: Open Source Scientific Tools for Python*, accessed on Oct. 12, 2016. [Online]. Available: http://www.scipy.org/
[78] J.-H. Yun, "Cross-layer analysis of the random access mechanism in universal terrestrial radio access," *Comput. Netw.*, vol. 56, no. 1, pp. 315–328, 2012.
[79] A. Blenk, A. Basta, M. Reisslein, and W. Kellerer, "Survey on network virtualization hypervisors for software defined networking," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 655–685, 1st Quart., 2016.
[80] M. Li, F. R. Yu, P. Si, E. Sun, and Y. Zhang, "Random access optimization for M2M communications in VANET with wireless network virtualization," in *Proc. ACM Symp. Develop. Anal. Intell. Veh. Netw. Appl.*, Nov. 2016, pp. 1–7.
[81] C. Liang and F. R. Yu, "Wireless network virtualization: A survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 358–380, 1st Quart., 2015.
[82] V.-G. Nguyen, T.-X. Do, and Y. Kim, "SDN and virtualization-based LTE mobile network architectures: A comprehensive survey," *Wireless Pers. Commun.*, vol. 86, no. 3, pp. 1401–1438, 2016.
[83] K. Liang, L. Zhao, X. Chu, and H. Chen, "An integrated architecture for software defined and virtualized radio access networks with fog computing," *IEEE Netw.*, vol. 86, no. 3, pp. 1401–1438, Feb. 2016.
[84] R. Mijumbi, J. Serrat, J.-L. Gorricho, N. Bouten, F. De Turck, and R. Boutaba, "Network function virtualization: State-of-the-art and research challenges," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 236–262, 1st Quart., 2016.
[85] A. Thyagaturu, A. Mercian, M. P. McGarry, M. Reisslein, and W. Kellerer, "Software defined optical networks (SDONs): A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2738–2786, 4th Quart., 2016.
[86] M. Yang, Y. Li, D. Jin, L. Zeng, X. Wu, and A. V. Athanasios, "Software-defined and virtualized future mobile and wireless networks: A survey," *Mobile Netw. Appl.*, vol. 20, no. 1, pp. 4–18, Feb. 2014.



MIKHAIL VILGELM (S'16) received the Diplom-Engineer degree from the Ural Federal University, Yekaterinburg, Russia, in 2011, and the M.Sc. degree from the Technical University of Munich in 2013, with master's thesis completed at NTT DOCOMO's European research laboratories (DOCOMO Euro-Labs) on mobility management in cellular networks. He is currently pursuing the Ph.D. degree with the Technical University of Munich. He has been with the Chair of Communication Networks, Department of Electrical and Computer Engineering, Technical University of Munich, since 2013. His research interests include wireless resource management for machine-to-machine communication and cyber-physical systems.

H. MURAT GÜRSU (S'16) was born in Istanbul, Turkey, in 1989. He received the B.Sc. degree in electrical and electronics engineering from Bogazici University in 2012, and the M.Sc. degree in communication engineering from the Technical University of Munich (TUM) in 2014, where he is currently pursuing the Ph.D. degree. He has been with the Chair of Communication Networks, Department of Electrical and Computer Engineering, TUM, since 2015. His research interests include low latency high reliability communication, wireless resource management, and wireless sensor networks.

WOLFGANG KELLERER (M'96–SM'11) has been the Director and the Head of Wireless Technology and Mobile Network Research with NTT DOCOMO's European research laboratories, DOCOMO Euro-Labs, for over ten years. He is currently a Full Professor with the Technical University of Munich, and also has been heading the Chair of Communication Networks with the Department of Electrical and Computer Engineering, since 2012. His research resulted in over 200 publications. He holds 29 granted patents in the areas of mobile networking and service platforms. His research focuses on concepts for the dynamic control of networks (Software Defined Networking), network virtualization and network function virtualization, and application-aware traffic management, wireless networks the emphasis is on machine-to-machine communication, device-to-device communication, and wireless sensor networks, with a focus on resource management toward a concept for fifth generation mobile communications. He is a member of the ACM and the VDE ITG.

MARTIN REISSLEIN (A'96–S'97–M'98–SM'03–F'14) received the Ph.D. in systems engineering from the University of Pennsylvania in 1998. He is currently a Professor with the School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe. He currently serves as an Associate Editor for the *IEEE Transactions on Education*, *Computer Networks*, and *Optical Switching and Networking*.