Why your smartphone doesn't work in very crowded environments

Why Your Smartphone Doesn't Work in Very Crowded Environments

Paolo Castagno*, castagno@di.unito.it
Vincenzo Mancuso†, vincenzo.mancuso@imdea.org
Matteo Sereno*, matteo@di.unito.it
Marco Ajmone Marsan‡†, ajmone@polito.it

*Università di Torino, Turin, Italy
†IMDEA Network Institute, Madrid, Spain
‡Politecnico di Torino, Turin, Italy

***Abstract*—An experience common to smartphone users is the difficulty in accessing services in crowded scenarios, such as a rock concert or a football match. In these cases, to (partially) mitigate frustration, users generically claim that network congestion is occurring, and try again and again to access the network with their smartphones: the result is that user frustration and network congestion reinforce each other! This paper investigates the root causes of poor performance of cellular networks in crowded environments, and shows that the commonly adopted random access procedure can prevent full utilization of wireless resources. We develop a simple, yet accurate analytical model, to analyze why attempting random access to wireless resources can become a problem even when access congestion avoidance is enforced, e.g., with the Access Class Barring (ACB) technique. The model we propose suggests that cluster-based network access, leveraging device-to-device communications, significantly alleviates access problems. Moreover, it sheds light on scalability laws that govern network utilization and quality of experience, in terms of cell capacity, number of access channels, and cluster size.**

I. Introduction

Our common experience is that wireless access networks perform poorly in very crowded environments. When we enjoy a soccer match or a rock concert in an extremely crowded stadium, and we would like to share our emotions with friends, we discover that placing a phone call or sending a short video, even posting a picture, is not possible, due to network congestion. When large numbers of networking experts gather at top international conferences in their field to discuss the latest research results, reading emails during the occasionally uninteresting talk is a problem, because the wireless access network is not able to sustain the very large number of email clients. These phenomena were quantitatively observed in [1], by collecting measurements over a tier-1 cellular network in the US during crowded events, and showing substantial performance degradations with respect to normal conditions.

The problem can only get worse. The Cisco Visual Networking Index forecast 2015-2020 [2] estimates that by 2020 the number of devices connected to IP networks will be more than three times as high as the world's population, generating an overall traffic of 2.3 ZB (equal to $2.3 \cdot 10^{21}$ B). Two thirds of this traffic will come from wireless devices, and 30% of the total will be generated by smartphones. The total mobile data traffic in 2020 will reach 30.5 EB (over $3 \cdot 10^{19}$ B) per month, with the highest volume in the Asia Pacific region, and the highest growth in the Middle East Africa region.

The 5G Infrastructure Public Private Partnership, in short 5G PPP, initiated by the European Commission, together with companies and research institutions of the field, shares those extreme visions [3]. Among the key challenges for 5G, a prominent position is given to the connection of over 7 trillion wireless devices serving over 7 billion people, and to the service of extremely crowded environments, such as a stadium, providing capacities of the order of 0.75 Tb/s over the stadium area, and an automated factory, comprising terminal densities up to 100 devices per m^2, and requiring sub-ms latency.

This paper looks at the performance of wireless access networks in extremely crowded environments, focusing as an example on the case of a group of LTE cells covering a stadium. The main contributions are the following:

- We develop a simple analytical model that captures the key aspects of the behaviour of a cell and we use it to understand the main sources of poor performance.
- We validate the analytical model with detailed simulations, which prove the validity of the assumptions introduced for analytical tractability.
- We show how the model can be instrumental for a correct dimensioning of crowded cellular systems.
- We propose the adoption of device-to-device (D2D) communications [4] as a means to improve performance in extremely crowded environments, and we quantify the benefits that can be achieved with the D2D approach, showing that D2D clusters of size k are more beneficial to system performance than a costly increase of system capacity by a factor k (e.g., through the deployment of k more cells).

The rest of this paper is structured as follows. Section II discusses the stadium scenario that we consider in this work; Section III overviews resource allocation request procedures. Section IV presents the analytical model. Section V uses the model to illustrate the system behavior. Section VI discusses numerical results. Section VII addresses the related work and Section VIII concludes the paper.

II. Scenario

The reference scenario that we use in our analysis is a large stadium, with capacity roughly comprised between 50 and 100 thousand spectators. Many such structures exist around the

world, including, e.g.: the Maracana in Rio de Janeiro, the Rose Bowl in Pasadena, and the Camp Nou in Barcelona.

Of course, such extraordinary numbers of people (terminals) imply a wide variety of services: spectators may want to send to their friends short videos, or pictures of the event, may receive all sort of messages, as well as phone calls, and at the same time terminals may be involved in content downloads.

We primarily focus on services which imply the human intervention, such as the transmission of a picture with a messaging application. In this case, the human user is in the service loop, so that the basic sequence of the service operations is made of a request for the radio access network resources, possibly repeated several times, until resources are granted, then the use of the network resources, followed by a think time before the next service request.

We will see that in some cases the system bottleneck is in the request for the radio access network resources, mostly because cellular systems use an Aloha-like contention-based scheme for this operation. It may thus happen that, while the network resources are available, request collisions do not allow their allocation. Under these circumstances, a reduction of the number of requests is mandatory to restore acceptable network performance. This can be obtained by reducing the number of users who are allowed to issue requests, or by forcing users to *coalesce* during the request phase. This is where D2D comes into play. If end user terminals are allowed to form clusters (or are instructed to form clusters by the network, through appropriate commands issued by the BS), only one request is issued whenever multiple terminals of the same cluster require access to the network resources, as proposed in [5] for opportunistic scenarios.

III. Accessing Resources in LTE

In LTE and LTE-A, end user terminals (called User Equipments – UEs in the LTE jargon), to access data channels, if not already connected to the BS (called evolved NodeB – eNodeB), have to proceed through the RACH (Random Access CHannel) procedure. Two types of random access procedures are defined: contention-based and contention-free [6]. In each LTE cell a fixed number (64) of orthogonal preamble signatures (PSs) are available, and the operation of the two types of RACH procedure depends on a partitioning of these PSs between those for contention-based access and those reserved for allocation to specific UEs on a contention-free basis. The contention-free RACH procedure is reserved to delay-sensitive cases, such as incoming traffic and handovers [7]. A contention-based random access PS is chosen at a UE to send a random access signal to the eNodeB. A conflict occurs if more than one UE use the same PS and time-frequency resources, resulting in undecodable messages at the eNodeB. The contention-based procedure consists of an exchange of four messages to set up a connection among UE and eNodeB.

Step 1: UE $\rightarrow$ eNodeB (Random Access Preamble) The first message conveys the randomly chosen RACH PS. The UE selects one of the available PSs and transmits it in a time-frequency slot. Several UEs may choose the same PS and the eNodeB may not be able to decode it. After the PS transmission, UE begins to monitor the downlink control channel (PDCCH) looking for an answer.

Step 2: UE $\leftarrow$ eNodeB (Random Access Response – RAR) This message is sent by the eNodeB on the PDCCH , and addressed with an ID identifying the time-frequency slot in which the PS was decoded. Whether multiple UEs have collided or not, if no RAR matching message has been received within the RAR window, they must repeat the RACH procedure, after a backoff delay. The duration of such backoff is randomly chosen in the range $(0, B]$ where B is the maximum number of subframes in a backoff period, and varies in $(0 - 960]$ ms.

Step 3: UE $\rightarrow$ eNodeB (Scheduled Transmission) The UE that receives the RAR message responses a scheduled transmission request that includes the ID of the device and a radio resource control (RRC) connection request message on the uplink shared channel (UL-SCH).

Step 4: UE $\leftarrow$ eNodeB (Content Resolution) Contention resolution is released from the eNodeB on the PDSCH. This identifies that no conflict on the access procedure exists. The UE can transfer data to eNodeB.

Once a UE has successfully performed the RACH procedure, it owns an active duplex connection and is in the `RRC_CONNECTED` state. Keeping a connection running requires that the eNodeB reserves physical resources devoted to this connection, even if there is no traffic available for the intended UE. Therefore the eNodeB can handle only a limited number of connected devices. Such devices incur in high battery consumption.

As long as the communication is alive, the UE remains in the `RRC_CONNECTED` state, but after an inactivity period, it begins to perform sleep cycles, from which it can return to the `RRC_CONNECTED` state without performing the contention-based RACH procedure. Sleeping UEs are not counted toward the maximum that can be handled by eNodeB.

Since the above-described access mechanism is based on a multichannel slotted Aloha, each PS representing an Aloha channel, its performance degrade beyond the threshold of 1 request/slot per PS. Hence, in dense scenarios, congestion can happen and become a system bottleneck. To alleviate congestion, state of the art solutions adopt the Access Class Barring (ACB) mechanism, which segments devices in several classes [8]. Devices within each class are managed through two parameters: the access barring probability and the barring time. With ACB, devices that are ready to attempt a random access are probabilistically *barred*, and barred devices wait for a barring time before making another barring decision, i.e., a device can be barred multiple times in a row. ACB is effective in smoothing peaks of access requests, but it does not change the RACH load under steady-state conditions. Moreover, ACB introduces a stochastic delay.

IV. Analytical Model

We model the operations of n end user terminal devices located in the same cell, under the coverage of one base station (BS or eNodeB). The notation used in this paper is summarized in Table I.

TABLE I
NOTATION AND CELL PARAMETERS USED IN SECTION VI

Quantity	Symbol	Value
Number of devices (or clusters)	n	
BS capacity	C	150–1500 [Mb/s]
Network Max Accepted Requests	M	200
Number of Random Access Preambles	N	54
Slot Time	τ	0.01 [s]
Back-off time RACH	B_0	av. 0.15 [s]
Back-off time Network	B_1	av. 1 [s]
ACB access probability	p_a	0.05–0.95
ACB barring time	B_a	av. 4-512s
Transmitted data volume	F_S	av. 1.5 [MB]
Transmission time	S	
Think time	T_{TH}	av. 30 [s]
Device uplink speed limit	R	
Probability to skip RACH procedure	p_J	≤ 0.5
Access delay	A_T	
Thinking subsystem throughput	λ	
Network subsystem throughput	ξ	
Random Access subsystem Input	γ	
Arrival Rate at Network subsystem	σ	
Collision Probability	p_C	
Rejection Probability	p_B	

Each device generates uplink transmission requests according to the 3GPP contention-based RACH procedure briefly described in Section III to obtain a transmission grant from the BS. We account for the fact that the establishment of downlink flows might provide the devices with extra opportunities to obtain transmission grants, skipping contention through the contention-free RACH procedure.

In the following, we derive a model for access requests and service operation in the cell, and show how to compute network utilization, access delay, and in general how to assess the behavior of the system as a function of the number of devices in the cell, for a given BS configuration (in terms of capacity, number of RACH channels, RACH slot duration, backoffs experienced upon failed RACH procedures, etc.).

A. Closed representation of the system

The BS has uplink capacity C, in bits per second, and can share its capacity among at most M devices at a time (i.e., there can be up to M devices in state `RRC_CONNECTED`). The number of RACH channels (i.e., orthogonal preamble signatures - PS) available for Random Access is N and the interval between two consecutive Random Access Opportunities (RAOs) is τ. If during τ a single device selects a given RACH channel, then the RACH procedure is successful, otherwise the RACH channel is either unused or a collision happens with multiple devices attempting to use the same PS.

A RACH collision results in a random backoff B_0, after which a RACH retry follows. In case of successful RACH procedure, the device is granted transmission only if there are less than M devices under service at the BS, otherwise the device goes through a random backoff B_1 followed by another RACH procedure. The model also considers ACB with uniform access probability p_a for all classes, and barring time with average duration $E[B_a]$.

For what concerns the traffic generated by end user terminals, we consider human-operated wireless devices, and assume that each device produces a new transmission request, with random data volume F_S, only after its previous request has been served. More specifically, upon service completion, we assume that the devices enters a "think time" period with random duration T_{TH} before generating the next request. Unless otherwise specified, the average service time $E[S]$ only depends on C, M and the average value $E[F_S]$, i.e., we assume that the serving speed is fixed and equal to C/M, so that $E[S] = \frac{M \cdot E[F_S]}{C}$. However, we will also show how to account for the equal sharing of the BS capacity among the actual number of devices under service in the system, and for service speeds limited by a device uplink speed R.

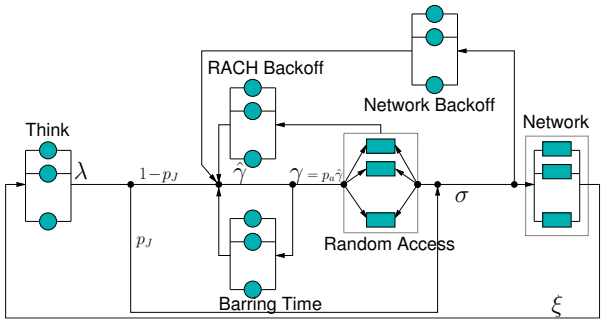

Fig. 1. Closed queueing network model of a cell.

The resulting system model is depicted in Fig. 1. The model comprises 6 main components: *i)* Think, representing the end user think time between the end of a service and the generation of a new access request; this is modeled with an infinite server queue with exponential service rate $\frac{1}{E[T_{TH}]}$; *ii)* Random Access, representing the RACH contention-based procedure; this is modeled as a set of N parallel slotted Aloha channels, receiving each $\frac{1}{N}$ of the total load offered to the RACH; the slot duration for any of the N slotted Aloha channels is τ; *iii)* Barring Time, which models ACB operation as an infinite server with average service time $E[B_a]$ affecting a portion $1-p_a$ of the flow directed to the Random Access; *iv)* Network, representing the BS resources, modeled as an M/G/M/0 queue with average service time $E[S]$. The Network queue is fed by the output of the Random Access subsystem and by the requests that skip the Random Access because of transmission opportunities generated by downlink traffic requests; these are modeled by means of the "jump probability" p_J, which is the probability to skip the contention-based RACH procedure, and access directly the BS resources. *v)* Network Backoff, and *vi)* RACH Backoff, representing the two backoffs, which are modeled by means of infinite server queues with exponential service times, with rates $\frac{1}{E[B_0]}$ and $\frac{1}{E[B_1]}$, respectively.

Fig. 1 also shows that the system is closed, i.e., the population is finite, with the number of customers fixed to n. We denote by λ the output of the Think subsystem, and by ξ the output of the Network subsystem. Because of the closed structure of the system, $\lambda = \xi$. We indicate with γ the total arrival rate at the N RACH channels in the Random Access subsystem, and we assume that RACH requests follow N parallel and i.i.d. Poisson processes with intensity $\frac{\gamma}{N}$. Although devices decide to send RACH requests asynchronously, such requests are cumulated over τ seconds and physically sent at the same time over the same frequency band. Thus, the

successful output of each of the N RACH channels is that of a slotted Aloha system with $\frac{\gamma\tau}{N}$ arrivals per slot, which is given by $\frac{\gamma\tau}{N}e^{-\frac{\gamma\tau}{N}}$ successes per slot, as known from the the standard analysis of multichannel slotted Aloha [9]. The maximum throughput per slot of such multichannel slotted Aloha system is $\frac{N}{e}$, which is achieved for $\gamma\tau = N$.

With the above, the arrival rate at the network service is $\sigma = \gamma e^{-\frac{\gamma\tau}{N}} + p_J\lambda$, the arrival rate at the RACH backoff B_0 is $\gamma\left(1-e^{-\frac{\gamma\tau}{N}}\right)$, and the one at the Network backoff B_1 is $p_B\sigma$, where p_B is the blocking probability, given by the Erlang-B formula with M servers and load $\rho = E[S]\sigma$. The load accepted and served by the network service is $\xi = (1-p_B)\sigma$. For analytical tractability, we introduce the simplifying assumption that all arrival processes are homogeneous and independent Poisson processes.

In the described system, quantities λ, σ, and ξ (and therefore also ρ and p_B) are functions of γ. It is possible to write a recursive equation in γ by considering that γ is $\hat{\gamma}$ minus what enters the Barring Time block. $\hat{\gamma}$ results from the sum of four arrival rates: $\lambda(1-p_J)$ from the Think subsystem, the output of backoffs B_0 and B_1, plus the recycle caused by ACB:

$$\hat{\gamma} = \lambda(1-p_J) + \gamma\left(1-e^{-\frac{\gamma\tau}{N}}\right) + p_B\left(\gamma e^{-\frac{\gamma\tau}{N}} + p_J\lambda\right) + (1-p_a)\hat{\gamma},$$

which, combined with $\gamma = p_a\hat{\gamma}$, yields a recursive expression for γ, which does not depend on ACB operation at all:

$$\gamma = \lambda(1-p_J) + \gamma\left(1-e^{-\frac{\gamma\tau}{N}}\right) + p_B\left(\gamma e^{-\frac{\gamma\tau}{N}} + p_J\lambda\right). \quad (1)$$

The recursive expression (1) has two unknowns: γ and λ (note that p_B can be written as function of ξ, and $\xi = \lambda$). Unfortunately, this expression is not enough to identify the operating point of the system, because it contains no dependency on the population size n. However, to introduce n in the loop, and remove λ, we can apply Little's law to different blocks in the modeled system, as presented in the following.

Solving system equations requires iteration, whose proof of convergence is omitted due to lack of space.

B. Dependence on the population size n

From the model described in the previous subsection, we can easily derive the expressions for the network utilization, the number of devices under service and in any of the system blocks depicted in Fig. 1, the time of a complete cycle between two transmissions, and the delay to access the service. All these quantities can be expressed as function of γ, and γ can be expressed as function of the population size n.

Utilization and distribution of devices. The network utilization ξ is equal to $\sigma(1-p_B) = \left(\gamma e^{-\frac{\gamma\tau}{N}} + p_J\lambda\right)(1-p_B)$. Therefore, since $\xi = \lambda$, it is immediate to obtain the following expressions for ξ, λ, σ and ρ:

$$\xi = \lambda = \frac{\gamma e^{-\frac{\gamma\tau}{N}}(1-p_B)}{1-p_J(1-p_B)}; \quad (2)$$

$$\sigma = \frac{\xi}{1-p_B} = \frac{\gamma e^{-\frac{\gamma\tau}{N}}}{1-p_J(1-p_B)}; \quad (3)$$

$$\rho = E[S]\sigma = \frac{E[S]\gamma e^{-\frac{\gamma\tau}{N}}}{1-p_J(1-p_B)}. \quad (4)$$

Note that, since ρ in (4) only depends on γ and p_B, we have that p_B actually depends only on γ. Thus, all the quantities representing arrival rates in the system model are functions of γ only, for fixed values of the other system parameters.

The number of devices under service, that cannot exceed M, is computed by applying Little's law at the Network, i.e., $n_S = \xi E[S] \leq M$, which also implies that utilization cannot exceed $M/E[S]$. Similarly, the average number of devices in Think is proportional to the average number of devices under service, i.e., $n_{TH} = \xi E[T_{TH}] = n_S\frac{E[T_{TH}]}{E[S]}$.

The rest of the devices $n - n_S - n_{TH}$ are attempting access, either waiting for the next RACH opportunity (including after a barring event) or in one of the backoff queues, so applying again Little's law we obtain:

$$n - n_S - n_{TH} = \gamma\left(\frac{\tau}{2} + \frac{1-p_a}{p_a}E[B_a]\right) + \\ + \gamma\left(1-e^{-\frac{\gamma\tau}{N}}\right)E[B_0] + \frac{p_B\gamma e^{-\frac{\gamma\tau}{N}}}{1-p_J(1-p_B)}E[B_1],$$

where the average delay incurred in a RACH attempt is computed as half of the slot duration because of the Poisson arrival assumption. The total number of devices in the network can therefore be expressed as a function of γ:

$$n = \gamma\left(\frac{\tau}{2} + \frac{1-p_a}{p_a}E[B_a]\right) + \gamma\left(1-e^{-\frac{\gamma\tau}{N}}\right)E[B_0] + E[B_1] \\ \cdot \underbrace{\frac{p_B\,\gamma e^{-\frac{\gamma\tau}{N}}}{1-p_J(1-p_B)}}_{\frac{p_B}{1-p_B}\xi} + (E[S] + E[T_{TH}])\underbrace{\frac{\gamma e^{-\frac{\gamma\tau}{N}}(1-p_B)}{1-p_J(1-p_B)}}_{\xi} \quad (5)$$

This is a monotonic relation between n and γ, which can be inverted (although not in closed form) to express γ as a function of n. However, we have seen that all quantities of interest in the system are functions of γ, so that we can conclude that they are eventually functions of n only, i.e., of the device population's size.

Cycle duration. The average time for a complete cycle in the system (e.g., the cycle between two consecutive service completions) is denoted by $E[T_{cycle}]$ and can be easily computed from the model of Fig. 1, by considering that: *i)* the probability to collide on a slotted Aloha representing the RACH channel with Poisson arrivals of intensity $\frac{\gamma\tau}{N}$ arrivals per slot is $p_C = 1 - e^{-\frac{\gamma\tau}{N}}$, and *ii)* collisions are assumed to be independent. Hence, we can write that:

$$E[T_{cycle}] = \frac{p_B}{1-p_B}E[B_1] + E[S] + E[T_{TH}] + \left(\frac{1}{1-p_B} - p_J\right) \\ \cdot \left[e^{\frac{\gamma\tau}{N}}\left(\frac{1-p_a}{p_a}E[B_a] + \frac{\tau}{2} + E[B_0]\right) - E[B_0]\right]. \quad (6)$$

The term in brackets in (6) is the average time spent in the loop formed by the RACH and the RACH backoff blocks, which has to be counted $\frac{1}{1-p_B}$ times on average (i.e., the average number of Bernoulli trials before a success, including the success that occurs when a device finds the Network available), except for the case in which a request skips the RACH, which occurs with probability p_J. The quantity $\frac{1-p_a}{p_a}E[B_a] + \frac{\tau}{2} +$

$E[B_0]$ is the time to complete one of such RACH loops—which includes, on average, $\frac{1-p_a}{p_a}$ passages through the ACB backoff—and there are, on average, $\frac{p_C}{1-p_C} = e^{\frac{\gamma\tau}{N}} - 1$ collisions before a successful RACH attempt (in which case the RACH backoff does not occur). The network backoff is traversed only after a failed network access (i.e., $\frac{p_B}{1-p_B}$ consecutive times, on average), whilst the Network and Think subsystems are traversed only once per cycle. $E[T_{cycle}]$ depends on γ since we have shown that p_B also depends on γ. So, using (5) we conclude that $E[T_{cycle}]$ can be written as function of n.

Access delay. The access delay, indicated as $E[A_T]$, is the time spent in a cycle, excluding the think time and the service, and is therefore easily obtained from (6):

$$E[A_T] = E[T_{cycle}] - E[S] - E[T_{TH}]. \tag{7}$$

As for $E[T_{cycle}]$, this is an expression that depends on γ, and therefore on n. An alternative expression for $E[A_T]$ is obtained by applying Little's law to the part of the system that excludes network service and think time:

$$E[A_T] = \frac{n - n_S - n_{TH}}{\lambda}. \tag{8}$$

Since $\lambda = \xi$, (8) reveals that the access delay is (practically) linear with the population size if ξ is (roughly) constant in a range of n, so that also n_S and n_{TH} are constant. As we will show later, such range exists if the Network saturates before the Random Access. That range is very relevant, because any point in it leads to maximal utilization.

C. QoE indexes

We use two indexes to express the quality of experience (QoE) for the end user. The first index η_S compares the service time with the time spent waiting before service starts, and it decreases with the access delay:

$$\eta_S := \frac{E[S]}{E[S] + E[A_T]}. \tag{9}$$

The second index is η_A, which is inversely proportional to the service time and fades exponentially with the access delay. Service time and access delay used in η_A are normalized to their values obtained with the smallest population n that causes the presence of M devices under service (denoted by n'):

$$\eta_A := \frac{E[S]|_{n=n'}}{E[S]} e^{-\frac{E[A_T]}{E[A_T]|_{n=n'}}}. \tag{10}$$

Differently form η_S, Index η_A is very sensitive to relative increases of delay rather than to absolute increases.

D. Analysis with D2D support

When D2D is used to alleviate RACH contention problems, terminal clusters come into play, each of them behaving as a single device. Thus, we can use the same formulas as above, with n, n_S, n_{TH} denoting the number of clusters in the system, under service and in think time, respectively. Similarly, all arrivals and services refer to clusters. The main effect of clusters is the reduced load to the Random Access. The impact is non-linear because γ does not scale linearly with n.

Cluster formation. Clusters form either spontaneously, when a device announces its willingness to wait for other users to join in a random access attempt, or under the control of the BS, when RACH collision probability becomes problematic.

Cluster service time. If k is the average cluster size, i.e., the average number of devices in a cluster, the service time becomes k times higher than for the case without clusters.

Cluster think time. In the case of clustered RACH access, the think time increases as well. Indeed, for a cluster, the think time corresponds to the think time of the device that initiates the cluster, plus the time needed for the other members to join. However, assuming a very high density of devices, forming a cluster of a few units is very quick. For instance, clusters of k devices have an average think time of $E[T_{TH}] + \sum_{i=1}^{k-1} \frac{E[T_{TH}]}{m-i} \simeq E[T_{TH}]\left(1 + \frac{k-1}{m}\right)$, where $m \gg k$ is the number of devices that can join the cluster. In practice, the think time increase is negligible in crowded environments, and so we ignore it in the model.

E. Impact of resource sharing under non-saturated conditions

If we consider that the BS resources can be shared by active connections, it is obvious that underloaded systems offer higher rates to the active devices. Therefore, the analysis proposed so far is valid in the region in which the Network subsystem is fully loaded, while it contains an approximation elsewhere. To fix this approximation, let us consider a Network subsystem that shares equally its resources among the connected devices, up to a rate R that can be interpreted as the maximum rate achievable by a device or as the maximum rate agreed in the user's SLA. In such case, the service time becomes $E[S] = E[F_S] \max\left\{\frac{1}{R}, \frac{1}{C/n_S}\right\}$.

Note that $E[S]$ is equal to $\frac{E[F_S]}{R}$ when the number of devices under service is not enough to saturate the Network subsystem. The adaptation of $E[S]$ to the number of devices under service introduces a further element of dependence on n, and a non-linearity. However, the impact on system performance is quite limited and can be neglected, as shown in the next section.

V. System Behavior

Here we study the bottlenecks of the system, point out some notable points in the performance curves, and analyze how performance is affected by the number of devices present in the cell and by the introduction of D2D-based clusters.

A. Bottlenecks

The model depicted in Fig. 1 has two potential bottlenecks: the Random Access and the Network subsystems. The former filters network access attempts, and asymptotically prevents any network request as γ grows with the population n. The Network subsystem has finite capacity, and therefore cannot serve more than M simultaneous requests.

Fig. 2 shows a typical case in which the maximum throughput of the Random Access is below the capacity of the Network subsystem, and thus is the only bottleneck, for all population sizes. In this case, the Network subsystem throughput ξ and the input σ of the Network subsystem are equal, since the blocking probability p_B is negligible. From (7), the access delay becomes a linear affine function

of $e^{\frac{\gamma\tau}{N}}$, and therefore grows with e^n. However, as shown in Fig. 2, a system in which the Random Access saturates before the Network subsystem does not suffer high delay. The range of device populations that roughly maximizes network utilization is quite narrow, and corresponds to a rather small interval around the peak efficiency of a multichannel slotted Aloha system, i.e., to values of n close to the one that yields $\gamma\tau = N$ (about 400 in the figure). This is the context that was previously analysed in the literature for the case of machine to machine (M2M) communications [10], with a model similar to ours. Here we focus on the more complex two-bottleneck case, in which the Network subsystem saturates before the Random Access, which is typical for the stadium scenario.

Fig. 3 shows an example of the model behaviour when both the Random Access and the Network subsystem can become the system bottleneck. Indeed, the Network subsystem is a bottleneck for lower values of population size, until the Random Access reaches success probabilities so low to starve the Network subsystem. In the figure we can identify three operational regions. In the first region (low number of devices and low load: roughly below 550 devices for the specific example), p_B and p_C are close to zero, $\sigma \simeq \xi$, and the delay is practically negligible. In the second region (shaded in the figure, roughly from 550 to 7500 users), the throughput of the Network subsystem is constant, while σ follows the familiar bell-shaped curve of slotted Aloha, and the delay grows linearly with the user population, as visible from (8). In the third region, p_B is negligible again, so that $\sigma \simeq \xi$ like in the first region, but the delay now grows exponentially with γ, and therefore with n. Out of such three regions, only the second one is desirable for system operation, since the Network subsystem resources are not wasted, and delay scales linearly with the number of devices in the cell.

B. Notable operational points

Random Access saturates first. In this case, $p_B \simeq 0$, so that $\xi \simeq \frac{\gamma e^{-\frac{\gamma\tau}{N}}}{1-p_J}$, the average number of devices in service is $E[n_S] = \xi E[S] < M$, and the average service time has to be constant (as remarked in Section IV-E, $E[S]$ must be equal to $\frac{E[F_S]}{R}$, otherwise the Network subsystem saturates). The network throughput is maximal when the output of the Random Access is maximal. This occurs for a number n^* of users that results in $\gamma = \frac{N}{\tau}$. From (5) we obtain the approximation (linear in N):

$$n^* \simeq \frac{N}{\tau}\left[\frac{E[S]+E[T_{TH}]}{e(1-p_J)} + \left(1-e^{-1}\right)E[B_0] + \frac{1-p_a}{p_a}E[B_a]\right] + \frac{N}{2}.$$

With Network saturation. In this case, to characterize the behavior of the system in the three operational regions shown in Fig. 3, in addition to n^* we characterize n' and n'', i.e., the values of n that correspond to the first and the second knee of the curve representing ξ vs. n.

Note that $n' \leq n^* \leq n''$, and the throughput of the Network subsystem is constant and equal to $\frac{C}{E[F_S]}$ for all values in the interval $[n', n'']$. Therefore, (1) reduces to:

$$\gamma e^{-\frac{\gamma\tau}{N}} = \frac{C}{E[F_S]}\frac{1-p_J(1-p_B)}{1-p_B}, \quad \forall\gamma \mid n \in [n', n''].$$

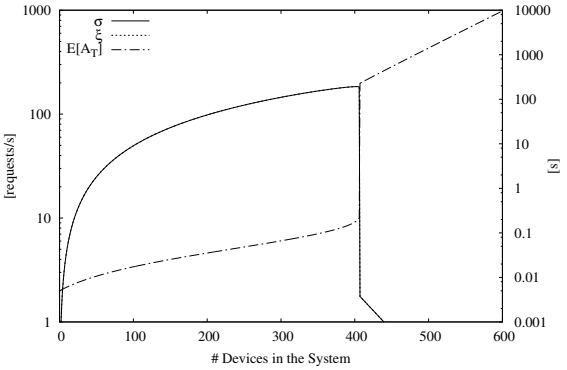

Fig. 2. Random Access-limited model behavior. Left scale for σ and ξ, right scale for access delay.

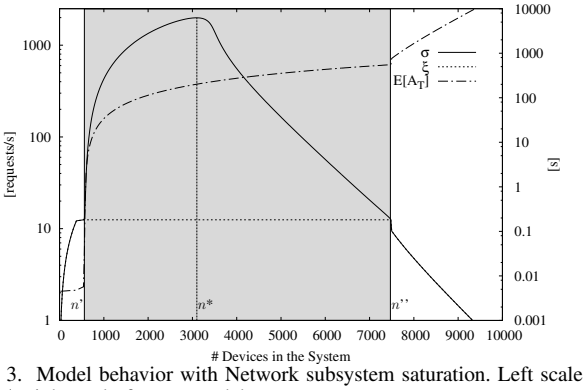

Fig. 3. Model behavior with Network subsystem saturation. Left scale for σ and ξ, right scale for access delay.

At the extremes of the considered interval $[n', n'']$, the Network subsystem has exactly enough resources to satisfy the demand, so that we can consider $p_B \simeq 0$:

$$\gamma e^{-\frac{\gamma\tau}{N}} \simeq \frac{C}{E[F_S]}(1-p_J), \quad \gamma \mid n \in \{n', n''\}. \tag{11}$$

Considering that the l.h.s. of (11) is a non-negative continuous function of γ that starts from 0, grows until it reaches the value $\frac{N}{e\tau}$ at $\gamma = \frac{N}{\tau}$ and then decreases asymptotically to 0, expression (11) admits two (possibly coinciding) real solutions only if $\frac{C}{E[F_S]}(1-p_J) \leq \frac{N}{e\tau}$. So, a range of values of n such that the throughput of the Network subsystem is constant and maximal exists, if and only if

$$N \geq \frac{e\tau C}{E[F_S]}(1-p_J). \tag{12}$$

The distance between the zeros of γ in (11) decreases logarithmically with C increasing (and with p_J decreasing). Since γ is monotone with respect to n, this means that the interval $[n', n'']$ becomes smaller with larger capacities C (and with smaller probabilities p_J), and $n' = n'' = n^*$ when (12) holds as equality. If (12) does not hold, the Network subsystem cannot saturate, and we fall back to the Random Access-limited scenario of Fig. 2.

The above condition also tells that the number of RACH channels needed to allow network saturation scales linearly with the capacity of the network and with $(1-p_J)$.

The notable points described above and the asymptotic behavior of γ vs. n can be approximated by means of following closed form expressions that can be readily derived:

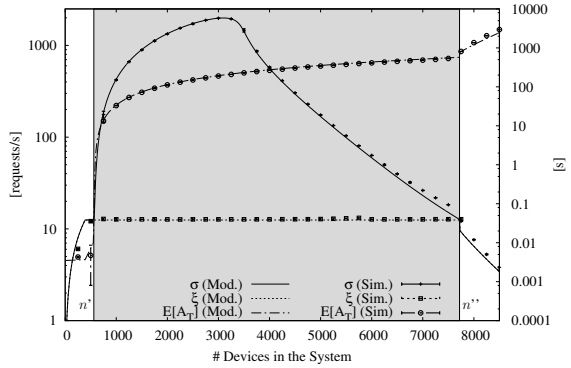

Fig. 4. Model validation for a cell with 150 Mb/s capacity and $p_a = 1.0$. Left scale for σ and ξ, right scale for access delay.

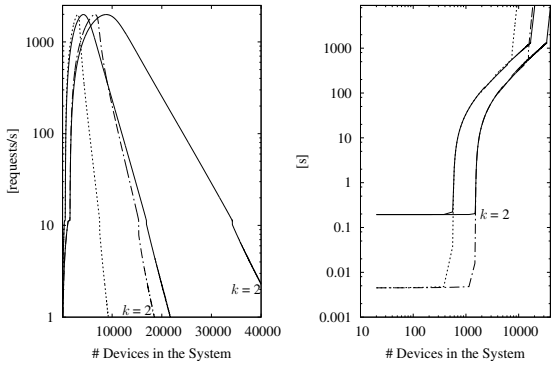

Fig. 5. Throughput (left) and access delay (right): impact of ACB with $[p_a = 0.95, E[B_a] = 4]$ (solid lines) and clustering (where k is specified) in the stadium scenario.

$$n' \simeq \frac{C}{E\left[F_S\right]}\left[E[S]+E\left[T_{TH}\right]+(1-p_J)\left(\frac{\tau}{2}+\frac{1-p_a}{p_a}E[B_a]\right)\right];$$

$$n^* \simeq M+E\left[T_{TH}\right]\frac{C}{E\left[F_S\right]}+\frac{N}{2}+\frac{N}{\tau}\left(1-e^{-1}\right)E\left[B_0\right]$$
$$+\frac{N}{\tau}\frac{1-p_a}{p_a}E[B_a]+\left[\frac{N}{e\tau}-\frac{C}{E\left[F_S\right]}(1-p_J)\right]E\left[B_1\right].$$

With clusters. Clustering k devices results in transferring $kE\left[F_S\right]$ bits per network access, hence the cluster service time $E[S]$ becomes k times longer. So, n' decreases with increasing cluster size. However, the number of devices within clusters becomes kn'. Denoting by $E[S|1]$ the service time without clusters, we have:

$$kn' \simeq \frac{C}{E\left[F_S\right]}\left[kE[S|1]+E\left[T_{TH}\right]+(1-p_J)\left(\frac{\tau}{2}+\frac{1-p_a}{p_a}E[B_a]\right)\right]$$

which includes $(k-1)\frac{CE[S|1]}{E[F_S]}$ more devices w.r.t. the case without clusters. Similarly, we can observe that kn^* grows by M plus a number of devices proportional to N for each increase of 1 in the cluster size k.

The interval $n''-n'$ increases with the cluster size, because a factor k appears in the denominator of the r.h.s. of (11) when clusters are used. Therefore, the increase of the size of the network saturation region, in terms of devices, becomes $k(n''-n')$, which is more than a k-fold increase.

We can conclude that the beneficial impact of clustering is larger than the one obtained by increasing cell capacity, which is linear, and it comes at a much lower deployment cost.

C. Delay

The access delay $E\left[A_T\right]$ is negligible when the Random Access saturates first, and for $n < n'$ when the Network subsystem also saturates, unless ACB introduces high delay by using low values for p_a and/or high values for $E[B_a]$. When the Network subsystem is saturated, we know from (8) that $E\left[A_T\right]$ is proportional to n with coefficient $\frac{1}{\lambda} = \frac{E[F_S]}{C}$. For $n > n''$, the delay explodes exponentially. Therefore, the desirable range of population sizes goes from n' to $n' + \Delta n$, where Δn is such that the delay $\frac{\Delta n E[F_S]}{C}$ is bearable by the applications running at the devices in the network.

So, in practice, the study of n' and its approximation are key to tune system parameters properly during network design.

D. Validation through packet-level simulation

In order to validate the simplifying assumptions that we had to introduce for the analytical tractability of the model, we developed a packet-level simulator that reproduces the behaviour of the closed model in Fig. 1. However, in the simulator we used uniformly distributed (rather than exponentially) file sizes; in addition, the output of the Random Access subsystem is not a Poisson process, rather an impulsive process in which all successful RACH attempts are brought at the Network subsystem ingress at the same time. Of course, in the simulator, the assumption that all arrival processes are Poisson, homogeneous and independent does not hold.

Fig. 4 reports an example of the simulated results for σ and ξ, together with the analytical results. Specifically, we report numerical results for a cell with $C = 150$Mb/s, $R = 10$Mb/s, $N = 54$, $M = 200$, and $\tau = 0.01$s, which are typical values for LTE BSs. Moreover, we used $E\left[T_{TH}\right] = 30$s, $E\left[B_0\right] = 0.15$s, $E\left[B_1\right] = 1$s, $E\left[F_S\right] = 1.5$MB, to account for typical upload of pictures and small videos during crowded events by using applications like Whatsapp, with automatic file upload retry. We run each experiment a sufficient number of times to obtain small 95% confidence intervals. The figure clearly shows that the model is extremely accurate. We tested a wide range of values for all relevant parameters, and found very similar model accuracy in all cases.

VI. Stadium: Numerical Results

We consider a stadium covered by a set of LTE cells. The system parameters are as reported in Table I (right-most column). Fig. 5 illustrates the impact of ACB and clustering in the specified scenario. We only report the results obtained with the ACB configuration that causes less delay, and one example of clustering ($k = 2$). The figure shows that either ACB or clustering makes it possible to significantly increase the number of users in the system. In particular, clustering as few as groups of 2 users is very effective in increasing n'. ACB suffers large delays, so as to make it quite undesirable even for limited user population sizes. However, Fig. 5 also shows that ACB and clustering *in combination* achieve low delay and guarantee access to very large user populations.

Fig. 6 reports the values for the two QoE indexes η_S and η_A we defined in Section 4B. Both indexes try to capture the user

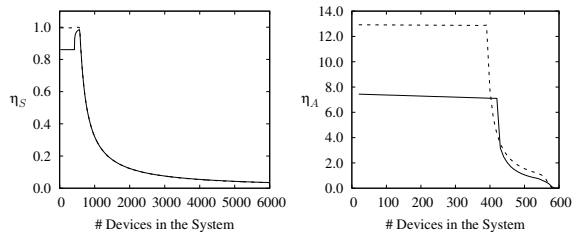

Fig. 6. End user QoE indicators η_S and η_A. Quality degrades with ACB $[p_a = 0.95, E[B_a] = 4]$ (solid lines) w.r.t. scenario without ACB (dashed lines), because of the additional delay it causes.

satisfaction, combining the service time and the access delay. In the first case we just compute the ratio between the service time and the sum access delay plus service time. In the second case we define a more elaborate parameter, which is inversely proportional to the service time, normalized to the service time when M users are under service, and exponentially fades with the access delay normalized to the access delay value at population value equal to n', thus being very sensitive to relative increases of delay rather than to absolute increases.

The curves of the QoE parameters show qualitatively similar trends. As regards η_S, with a low population of UEs, the network access time $E[T_A]$ is very low and mostly depends on RACH transit and ACB operation. Each device in service is guaranteed a rate equal to R, keeping η_S close to 1, unless ACB is used and $E[A_T]$ cannot be neglected. When the BS can no longer provide the maximum rate R to each one of the n_S devices in service, $E[S]$ starts to increase, while $E[A_T]$ is practically constant (without ACB) or slowly increasing (with ACB), so that its weigh in η_S diminishes as the population increases. However, when the number of devices reaches the value n', the access delay $E[A_T]$ starts increasing fast (and linearly) causing a hyperbolic decrease of η_S towards zero. The QoE parameter starts dropping around 550 devices in the cell. In general, the figure shows that using ACB *is detrimental* in terms of quality experience in steady state conditions, especially with small populations, when the ACB delay is the most prominent component of the access delay.

For what concerns η_A, the figure shows that, without ACB, it starts from the value $10/0.75 = 13.33$. This is the ratio between the data rate cap for each individual device, and the data rate given by the BS to each user once the maximum number of users (200) is reached (150 Mb/s divided by 200 users means 0.75 Mb/s per user). With ACB, the additional delay due to barring decreases the initial value of η_A. In all cases, the curve stays close to the initial value as long as the access delay remains negligible, then it rapidly drops. Also in this case, the QoE parameter starts dropping around 500 devices. Note that this means that a coverage of the $50,000$ users in the stadium with good QoE would require about 100 cells, if each user carries just one device, 200 cells if each users carries two devices, and so on.

Of course, one possibility to improve performance is to use cells with higher capacity. In Fig. 7 we plot curves of $E[A_T]$ for cell capacities in the range 150-$1,500$ Mb/s. The critical element for QoE is given by the points where the access delay starts increasing significantly. This means about 550 devices with capacity 150 Mb/s and about $4,000$ devices with capacity

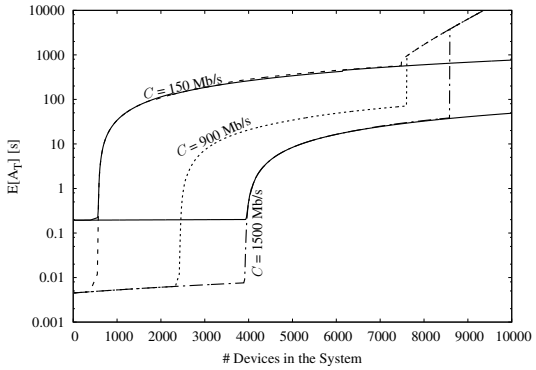

Fig. 7. Access Delay for variable cell capacity with ACB $[p_a = 0.95, E[B_a] = 4]$ (solid lines) and without (dashed lines).

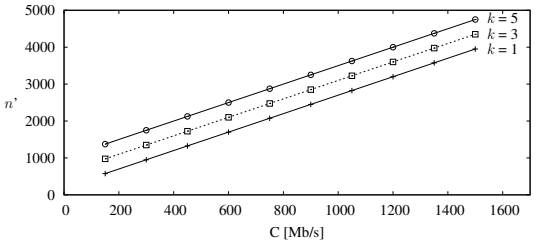

Fig. 8. Values of n' versus the cell capacity, for variable cluster sizes. ACB curves are practically superposed to curves without ACB.

$1,500$ Mb/s. The latter translates into 12 cells for $50,000$ devices, 25 in the case each spectator carries 2 devices.

In addition, Fig. 7 clearly shows that the operating area where both end users and network operators wish "to be" is just before the curve's first knee. In such neighbourhood, ACB does not play any significative role, and $E[A_T]$ is a fraction of $E[S]$, before starting to rapidly move to bigger values. It is important to recall that this phase change in the access delay is pinpointed by n'. The second knee of the curves corresponds to n'', and both knees change with the cell capacity. It is very important to notice that in the whole interval $[n', n'']$ the system bottleneck is the Network due to the limitation of M `RRC_CONNECTED` devices. When the number of devices in the cell becomes larger than n'', we see a switch in the bottlenecks, and only from this point on the RACH subsystem becomes unstable, and the access time explodes, going asymptotically to infinite.

Increasing the cell capacity or the number of cells is quite costly, and may not be the most desirable solution to achieve good QoE in crowded environments. A much simpler option can be to allow users to coalesce in their network access attempts through the formation of clusters. Fig. 8 shows the values of n' as a function of the cell capacity, for variable cluster sizes. We immediately appreciate the advantages of clusters: the adoption of coalitions brings a gain comparable to the one obtained increasing C with a negligible cost (if any) to the network provider. Indeed, a gain equal to or larger than that obtained by doubling the cell capacity can be achieved by adopting a cluster size $k = 3$.

Finally, to evaluate the importance of reducing the load of the Random Access in presence of downlink traffic, we repeat our tests with different values of p_J. Skipping the contention-based RACH procedure introduces a small improvement in

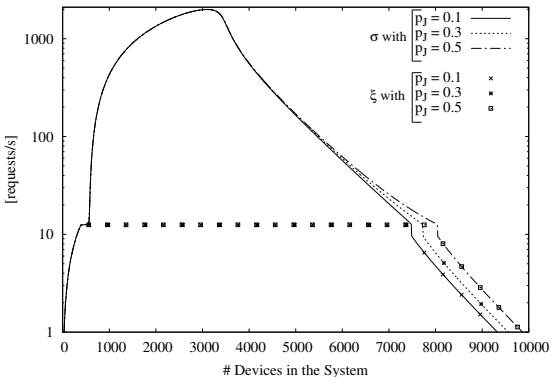

Fig. 9. Impact of skipping the contention-based RACH procedure.

terms of the height of the point at n^*, allowing the RACH to sustain a slightly higher arrival frequency. However, as can be seen in Fig. 9 for the case with no ACB, the main impact of p_J on the performance of the system is reflected in the value of n'', which is moved towards larger values of n. It must be noted that the increase in height at n^* is so small not to be visible on the graphs, and that the increase in the value of n'' is not relevant from the point of view of applications, because at those numbers of users per cell, performance (e.g., in terms of access delay) is intolerably bad.

VII. Related Work

In [11], 3GPP has identified the random access mechanism as a possible problem when the number of connected devices rises to tens of thousands. For this reason, MAC overload control has been investigated, and a broad literature exists on this topic. See [12] for a comprehensive overview. Simple models to estimate the probability of preamble collision in the PRACH channel are presented in a few 3GPP standard documents (e.g., [11]), and in the literature (e.g. [10], [12], [13], [14]). The conclusions of most of these studies point out that for Mobile-Type Communications (MTC) applications, the PRACH procedure can drastically limit network performance. Possible approaches to modify the PRACH access procedure have been proposed in [15], [16].

Most of the previous studies on dense cellular environments have focused on MTC scenarios, and [12] shows that the differences between the human-based and the MTC scenarios are substantial. Nevertheless, the PRACH access mechanism, and its interactions with the other phases of the network usage cycle play an important role also in case of human-based scenarios. This was shown in [1], through a measurement-based study of cellular network performance during crowded events, showing that network access failures become orders of magnitude higher than those observed on routine days, and the interaction between access and transmission phases generates behaviors difficult to predict. The simple analytical model presented in this paper provides a tool to understand the root causes of the behaviors measured in [1], and to quantify the impact of the crowd size on network performance, also indicating possible approaches to correctly dimension the network and to mitigate the negative impacts of crowds.

VIII. Conclusions

This paper presents a model to capture the key aspects of the behaviour of a cellular networks in crowded environments. The main merit of the model lies in the insight that it brings on cellular system operations in very crowded environments, and in the possibility to use it to drive the correct dimensioning of the cellular system in very crowded environments. As an example, the model allows the assessment of the benefits achievable through the adoption of D2D communications to reduce the congestion on the RACH much effectively than with ACB, thus significantly improving performance and QoE. For example, our model shows that, instead of serving 50,000 terminals with 100 cells of capacity 150 Mb/s each, it is possible to use 25 cells, each of capacity 300 Mb/s, provided that clusters of 5 devices are formed to access the RACH.

References

[1] M. Zubair Shafiq, L. Ji, A. X. Liu, J. Pang, S. Venkataraman, and J. Wang, "A First Look at Cellular Network Performance During Crowded Events," in *Proc. of the ACM SIGMETRICS*, ser. SIGMETRICS '13. New York, NY, USA: ACM, 2013, pp. 17–28.

[2] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2010-2015," Tech. Rep., February 2011. [Online]. Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper/c11-520862.pdf

[3] W. Mohr, "5G empowering vertical industries," Cisco, Tech. Rep., April 2016. [Online]. Available: https://ec.europa.eu/digital-single-market/en/blog/5g-empowering-vertical-industries-0

[4] A. Asadi, Q. Wang, and V. Mancuso, "A Survey on Device-to-Device Communication in Cellular Networks," *IEEE Communications Surveys & Tutorials*, 2014.

[5] A. Asadi, V. Mancuso, and R. Gupta, "An SDR-based Experimental Study of Outband D2D Communications," in *Proceedings of IEEE INFOCOM*, Apr. 2016.

[6] "Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification," 3GPP, TS 36.321 Release 13 V13.1.0, April 2016.

[7] S. Sesia, I. Toufik, and M. Baker, *LTE, The UMTS Long Term Evolution: From Theory to Practice*. Wiley Publishing, 2009.

[8] I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, and V. Casares-Giner, "Performance Analysis of Access Class Barring for Handling Massive M2M Traffic in LTE-A Networks."

[9] M. Ajmone Marsan, D. Roffinella, and A. Murru, "ALOHA and CSMA protocols for multichannel broadcast networks," in *Proc. of Canadian Commun. Energy Conf.*, Montreal, P.Q., Canada, Oct. 1982.

[10] J. J. Nielsen, D. M. Kim, G. C. Madueno, N. K. Pratas, and P. Popovski, "A Tractable Model of the LTE Access Reservation Procedure for Machine-Type Communications," in *2015 IEEE Global Communications Conference (GLOBECOM)*, Dec 2015, pp. 1–6.

[11] "Study on RAN Improvements for Machine-type Communications," 3GPP, TR 37.868 Release 11 V11.0.0, Septemeber 2011.

[12] L. A. Andres Laya and J. Alonso-Zarate, "Is the Random Access Channel of LTE and LTE-A Suitable for M2M Communications? A Survey of Alternatives," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 4,16, April 2011.

[13] O. Arouk and A. Ksentini, "General Model for RACH Procedure Performance Analysis," *IEEE Communications Letters*, vol. 20, no. 2, pp. 372–375, Feb 2016.

[14] G. C. Madueno, J. J. Nielsen, D. M. Kim, N. K. Pratas, C. Stefanovic, and P. Popovski, "Assessment of LTE Wireless Access for Monitoring of Energy Distribution in the Smart Grid," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 675–688, March 2016.

[15] T. P. C. de Andrade, C. A. Astudillo, and N. L. S. da Fonseca, "Random access mechanism for RAN overload control in LTE/LTE-A networks," in *Proc. of IEEE International Conference on Communications (ICC)*. IEEE, 2015, pp. 5979–5984.

[16] Y.-C. Yuan-Chi Pang, G.-Y. Lin, and H.-Y. Wei, "Context-Aware Dynamic Resource Allocation for Cellular M2M Communications," *IEEE Internet of Things Journal*, vol. 3, no. 3, pp. 318–326, 2016.