Modeling Simple HetNet Configurations with Mixed Traffic Loads

Modeling Simple HetNet Configurations With Mixed Traffic Loads

Marco Ajmone Marsan
Politecnico di Torino, Italy, and
IMDEA Networks Institute, Spain
marco.ajmone@polito.it

Michela Meo
Politecnico di Torino, Italy
michela.meo@polito.it

Matteo Sereno
Università di Torino, Italy
matteo.sereno@unito.it

***Abstract*—In this paper we consider radio access network configurations comprising cells of different size and the simultaneous presence of elastic and inelastic services. The system is modeled as a queuing network, and we examine the system performance for variable parameter values, showing that some of the emerging behaviors can be unexpected, and provide insight into the effective deployment of small cells. In particular, we see that a large fraction of elastic traffic, together with an area of the small cell corresponding to a significant portion of the macro cell area are important aspects for the effective exploitation of the small cell. These behaviors can significantly impact the deployment of small cells, which are expected to become increasingly popular because of the need to provide additional capacity through densification of the cell layout.**

I. Introduction

Network densification is the approach to radio access network (RAN) evolution that allowed mobile network operators (MNOs) to cope with the exponential increase of network traffic observed in recent years, and now further accelerated by the lockdowns due to the COVID-19 pandemic[1]. Network densification implies the coverage of the RAN service area with progressively more dense deployments of cells of smaller size. This brings the advantage of a higher level of reuse of the portions of spectrum licensed to an MNO, as well as shorter distances between base stations (BSs) and User Equipments (UEs) carried by end users, with the possibility to achieve higher data rates due to better signal to interference and noise ratio (SINR). Densification often implies the addition of small cell BSs (briefly small cells – SCs – with a reach of the order of 100 m) in areas already covered by existing macro cell BSs (briefly MCs – with a reach of the order of 1 km), producing the so-called heterogeneous network (HetNet) cell layouts [1]. This typically occurs when a new generation of RAN technology is introduced, as it is happening now with 5G. HetNets often comprise two (or more) layers of BSs overlaid in a same area, possibly using different radio access technologies (RATs). Planning a HetNet, and dimensioning BS resources in a HetNet scenario, requires more complex approaches with respect to the traditional techniques used for network planning and dimensioning, that proved effective in contexts where all cells had similar characteristics and parameters [2]. In particular, there are two key aspects of HetNets that make dimensioning extremely complex. First, given the different coverage areas, how effective is the presence of a SC in alleviating congestion on the MC? Second, given the heterogeneity of services, is the presence of a SC effective in the same way on different kinds of traffic, e.g., elastic and inelastic traffic? Only by properly answering these questions it is possible to dimension and position SCs in such a way that densification brings the desired benefit.

In this paper we focus on the interaction of one or two SCs deployed within the coverage area of a MC, and we show that the behaviors emerging from the interaction of the cells are not trivial. As a result, the insight generated by our observations provides interesting guidelines for the deployment of SCs in HetNet scenarios.

The main contributions of this paper are the following.

- We propose and solve a model consisting in a queuing network serving customers corresponding to two different types of services: i) inelastic services, that require a fixed data rate for their entire duration, and ii) elastic services, that require the transfer of a fixed amount of data at the highest data rate that the can be provided.
- We investigate the impact of SCs in alleviating congestion on the MC considering the respective coverage areas and the different density of users in the considered areas.
- We investigate the impact of SCs on both elastic and inelastic services.

II. Mixing Elastic and Inelastic Services

We propose in this section a model to study the behaviour of BSs loaded with traffic deriving from a realistic mix of services; in particular, we account for two different classes of services: inelastic and elastic services.

[1]An Italian MNO reported a 30 to 40% growth of daily traffic at the start of lockdown, with a six-fold increase of voice conferencing traffic.

Inelastic services require a continuous flow of data with constant rate and correspond to voice or video conversations, as well as live video streaming. The service data rate is fixed by the rate at which data are produced at the source, and cannot be increased or decreased by the network. Hence, an inelastic service instance is assumed to require a fixed data rate for the whole duration of the service. If the BS is not capable of providing the required data rate, the service instance cannot be activated, and is blocked by the admission control algorithm. Note that, while we consider inelastic services with the same data rate requirements, our approach can be generalized to the case of different data rates (e.g., for video or voice) by defining user classes, like in [3].

Elastic services instead require the transfer of a given quantity of data at the maximum possible speed. In this case, data is available at the source, and the data rate is constrained by the network capabilities, as well as the source and destination capabilities of transmitting/receiving data, and the characteristics of the network protocols. Examples are file transfers, web browsing, content download, as well as several streaming services. An elastic service instance is thus assumed to require a minimum data rate, and, on top of that, to evenly share, together with all other elastic service instances, all the BS capacity which is not used by inelastic services.

The study of the performance of a BS under a mix of inelastic and elastic services requires the development of non-standard modeling tools, since this case was not previously tackled in the scientific literature, with the exception of [3], where however only one cell is considered, user mobility is not accounted for, and only an approximate solution is obtained. We will first describe the case of an individual BS, and then move on to configurations with one macro cell and one or two small cells. Note however that the generalization of our approach to scenarios with numbers of small cells in the order of ten is straightforward.

A. Modeling a BS Supporting Elastic and Inelastic Services

A BS supporting both elastic and inelastic services can be described by a queue like the one sketched in Fig. 1. We will use the terms *inelastic customers* and *elastic customers* to refer to customers modeling respectively inelastic and elastic service requests.

The dynamics of the queue are described by a discrete-state continuous-time stochastic process $Q(t) = (I(t), E(t))$, whose state at time t is the pair of values indicating the numbers of inelastic and elastic customers in service at time t. We are interested in the computation of performance metrics that can be derived from the limiting probabilities of this process, $\pi_{n_i,n_e} = \lim_{t\to\infty} P\{I(t) = n_i, E(t) = n_e\}$. The queue state is thus $\underline{s} = (n_i, n_e)$.

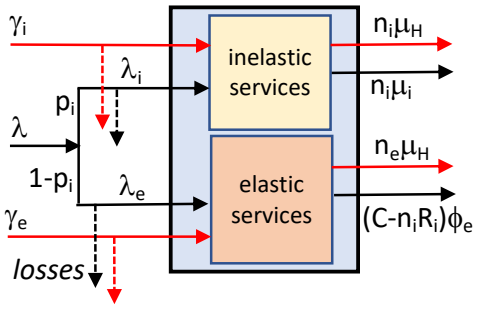

Fig. 1. Queuing model of a BS with elastic and inelastic traffic.

In the cellular environment, users in different positions within a cell experience different SINR values, and they must use the modulation and coding scheme that is compatible with such SINR value. This implies that for each time slot that the BS allocates to a user, the corresponding number of bits transmitted by the user depends on the user location. We denote with $1/\sigma_j$ the number of bits per slot transmitted by a user in location j, corresponding to coordinates (x_j, y_j) in a specified Cartesian plane.

We assume that the BS can allocate to users up to D slots per second in total. An inelastic service requires R_i bits per second, hence an inelastic user in position j must be allocated $R_i\sigma_j$ slots per second in total. The total number of slots allocated to inelastic users in state (n_i, n_e) is

$$\sum_{j=1}^{n_i} R_i\sigma_j \le D \tag{1}$$

Assuming an equal time scheduler for elastic users, considering that inelastic services require a minimum bit rate equal to R_e b/s, in state (n_i, n_e) with $n_e \ge 1$, an elastic user in position k receives a data rate

$$\frac{D - \sum_{j=1}^{n_i} R_i\sigma_j}{n_e\sigma_k} \ge R_e \qquad \forall k = 1, \cdots, n_e \tag{2}$$

Given a state (n_i, n_e) and a spatial distribution of users, the state is admissible if the two above inequalities are satisfied.

In case of user mobility, as considered in this paper, user movement makes the SINR perceived by each user change along the user trajectory. For this reason, instead of static values of $1/\sigma_j$, we use the average value of the quantities $1/\sigma_j$, which is denoted by $1/\sigma$. The BS capacity in bits per second is thus $C = D/\sigma$.

The value of C determines the admissible values for n_i and n_e, and thus defines the process state space:

$$\mathcal{S} = \{\underline{s} = (n_i, n_e) | n_i \ge 0, n_e \ge 0, n_i R_i + n_e R_e \le C\} \tag{3}$$

The queue receives four streams of customer arrivals. New service requests, which are generated when UEs are within the cell, arrive with rate λ. Since a new request can either be for an inelastic or an elastic service, the

TABLE I
TRANSITION RATES OUT OF STATE (n_i, n_e) OF THE MARKOV CHAIN MODELING THE BS WITH HETEROGENEOUS TRAFFIC

Destination	Rate
(n_i+1, n_e)	$\lambda_i + \gamma_i$
(n_i, n_e+1)	$\lambda_e + \gamma_e$
(n_i-1, n_e)	$n_i(\mu_H + \mu_i)$
(n_i, n_e-1)	$n_e\mu_H + (C - n_iR_i)\phi_e$

flow is split into two flows with rates, respectively, equal to $\lambda_i = p_i\lambda$ and $\lambda_e = (1-p_i)\lambda$, where the parameter p_i is the probability of a service being inelastic. In addition, the cell receives handover flows from neighboring cells, with rates γ_i and γ_e, for the two kinds of service. Handover flows are represented in red in Fig. 1. Arrivals are assumed to follow Poisson processes.

The service of inelastic customers requires a bit rate R_i for the whole duration of a service that is assumed to be a random variable with exponential distribution and average $1/\mu_i$ seconds. In state (n_i, n_e), the flow of customers leaving the cell due to termination of an inelastic service is, hence, given by $n_i\mu_i$.

Elastic customers share the capacity that is not used by inelastic customers, with a *minimum* service rate of R_e bits per second. The duration of an elastic service depends on the amount of data to be transferred and the data rate that inelastic services leave to elastic services. We assume that elastic services require the transmission of files whose size has an exponential distribution with rate ϕ_e, hence an average of $1/\phi_e$ bits. This means that the average duration of the file transfer is comprised between a minimum $1/(\phi_e C)$ seconds (if the elastic service can use the whole BS capacity C) and a maximum $1/(\phi_e R_e)$ seconds (if the elastic service can only use the minimum admissible data rate R_e). In state (n_i, n_e), the flow of customers leaving the cell due to termination of an elastic service is, hence, given by: $(C - n_iR_i)\phi_e$, where the term $(C - n_iR_i)$ represents the bandwidth that is available for elastic services and that is shared among the n_e customers according to the Processor-Sharing discipline.

In addition to service termination, customers leave the queue when a handover out of the cell occurs. To represent handovers, we assume that the user spends in the cell an amount of time, called *dwell time*, that can be described by an exponentially distributed random variable with rate μ_H, equal for both inelastic and elastic customers. In state (n_i, n_e), the flows of outgoing handovers have rate $n_i\mu_H$ and $n_e\mu_H$, for inelastic and elastic customers, respectively.

Exponential assumptions make the stochastic process $Q(t)$ a continuous-time Markov chain. The transition rates out of state $\underline{s} = (n_i, n_e)$ are summarized in Table I; clearly, a transition is possible in the Markov chain only if the destination state is admissible according to the definition of the state space given in (3). Blocking occurs for inelastic services whenever at an arrival $(n_i+1)R_i+n_eR_e > C$, and for elastic services whenever at the arrival $n_iR_i+(n_e+1)R_e > C$; i.e., whenever there is not enough capacity to serve an additional inelastic request at rate R_i or an elastic request at the minimum rate R_e.

B. Performance Indicators

Once the limiting probabilities of all states (n_i, n_e), denoted by π_{n_i,n_e}, have been computed, it is possible to derive aggregate performance metrics.

The average number of active inelastic and elastic services are computed as

$$E[N_i] = \sum_{\substack{n_i \\ (n_i,n_e)\in\mathcal{S}}} \sum_{n_e} n_i \, \pi_{n_i,n_e} \tag{4}$$

$$E[N_e] = \sum_{\substack{n_i \\ (n_i,n_e)\in\mathcal{S}}} \sum_{n_e} n_e \, \pi_{n_i,n_e} \tag{5}$$

The average fraction of the cell capacity used by inelastic and elastic services are

$$E[U_i] = \sum_{\substack{n_i \\ (n_i,n_e)\in\mathcal{S}}} \sum_{n_e} \frac{n_iR_i}{C} \, \pi_{n_i,n_e} \tag{6}$$

$$E[U_e] = \sum_{\substack{n_i \\ (n_i,n_e)\in\mathcal{S},\, n_e>0}} \sum_{n_e} \frac{C - n_iR_i}{C} \pi_{n_i,n_e} \tag{7}$$

The resulting average load of the cell is

$$\rho = \sum_{\substack{n_i \\ (n_i,0)\in\mathcal{S}}} \frac{n_iR_i}{C} \, \pi_{n_i,0} + \sum_{\substack{n_i \\ (n_i,n_e)\in\mathcal{S},\, n_e>0}} \sum_{n_e} \pi_{n_i,n_e} \tag{8}$$

The data rate used by an elastic service in state (n_i, n_e) with $n_e \geq 1$ is $r_e(n_i, n_e) = \frac{C-n_iR_i}{n_e}$, and the corresponding probability is

$$P\{r_e(n_i, n_e)\} = \frac{\pi_{n_i,n_e}}{1 - \sum_{\substack{n_i \\ (n_i,0)\in\mathcal{S}}} \pi_{n_i,0}} \tag{9}$$

Hence, the average data rate used by an elastic service is computed as

$$E[r_e] = \sum_{\substack{n_i \\ (n_i,n_e)\in\mathcal{S},n_e>0}} \sum_{n_e} r_e(n_i, n_e)P\{r_e(n_i, n_e)\} \tag{10}$$

The loss probabilities for inelastic and elastic services are

$$P\{\text{loss}_i\} = \sum_{\substack{n_i \\ (n_i,n_e)\in\mathcal{S},C-(n_iR_i+n_eR_e)<R_i}} \sum_{n_e} \pi_{n_i,n_e} \tag{11}$$

$$P\{\text{loss}_e\} = \sum_{\substack{n_i \\ (n_i,n_e)\in\mathcal{S},C-(n_iR_i+n_eR_e)<R_e}} \sum_{n_e} \pi_{n_i,n_e} \tag{12}$$

The average time devoted by the BS to an elastic service before either completion or handover can be obtained from Little's result as:

$$E[T_e] = \frac{E[N_e]}{(\lambda_e + \gamma_e)(1 - P\{\text{loss}_e\})} \tag{13}$$

Similarly, the average time devoted by the BS to an inelastic service before either completion or handover can be expressed as:

$$E[T_i] = \frac{E[N_i]}{(\lambda_i + \gamma_i)(1 - P\{\text{loss}_i\})} = \frac{1}{\mu_i + \mu_H} \tag{14}$$

The probability that an inelastic service leaves the cell because of completion is simply

$$P\{\text{comp}_i\} = \frac{\mu_i}{\mu_i + \mu_H} \tag{15}$$

The total average service completion rate of elastic customers is equal to

$$\mu_{Te} = \sum_{\substack{n_i \\ (n_i,n_e)\in\mathcal{S},n_e>0}} \sum_{n_e} (C - n_i R_i)\phi_e \, \pi_{n_i,n_e} \tag{16}$$

The probability that an elastic service leaves the cell because of completion is

$$P\{\text{comp}_e\} = \frac{\mu_{Te}}{\mu_{Te} + \sum_{\substack{n_i \\ (n_i,n_e)\in\mathcal{S},n_e>0}} \sum_{n_e} n_e \mu_H \, \pi_{n_i,n_e}} \tag{17}$$

It is interesting to observe that the queuing model we have described is a combination of an M/M/K/K queue and an M/M/1-PS queue with a limit on the number of customers in service. Since both models are insensitive to the service time distribution, it can be conjectured that also this more complex queuing model enjoys the same type of insensitivity.

C. Impact of mobility

Mobility has an important impact on the performance of the queue. In order to investigate it, we separately consider elastic and inelastic customers. As previously mentioned, we assume that mobility is a characteristic of the users which is independent of their service demand and can be represented by the dwell time being exponentially distributed with rate μ_H, whatever the class of customer is.

For inelastic customers, mobility has the effect of making the average activity time in the cell shorter than the average service time; indeed, the service rate in the cell is $(\mu_H + \mu_i)$, as indicated in Table I. The case of elastic customers is more complex. Based on the load, the queue with elastic customers and mobility exhibits two working regimes: *processor sharing* regime, appearing at low load, and *infinite server* regime, at high load. When the load is low, the service rate of elastic customers is very high, the time to complete the service is short and typically shorter than the dwell time; the service is most of the time completed before exiting the cell. The queue behaves like a PS queue. Conversely, when the load is high, the service rate is low (equal to R_e or slightly higher) which means that the time to complete the service is long and typically longer than the dwell time. The customers tend to leave the queue before service completion. The queue behaves similarly to an $M/M/\infty$ queue.

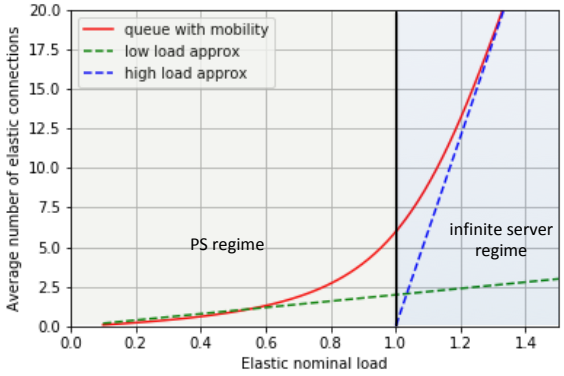

Fig. 2. Effect of mobility on the queue with elastic customers and mobility: average number of elastic connections versus nominal elastic load

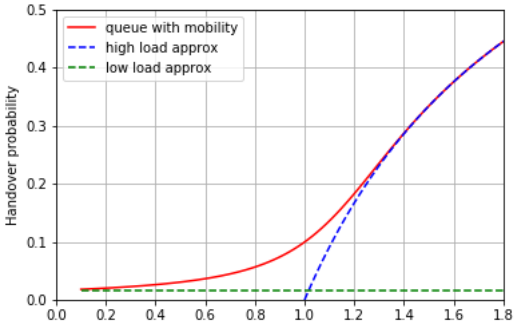

Fig. 3. Effect of mobility on the queue with elastic customers and mobility: handover probability versus nominal elastic load

The two regimes are represented in Fig. 2, that reports, in the red solid curve, the average number of elastic connections versus the nominal elastic load, given by λ_e/μ_e, in a queue with average dwell time equal to $1/\mu_H = 120$ s, and μ_e the maximum customer service rate given by C/ϕ_e, with $C = 50$ Mbps, $1/\phi_e = 100$ Mbit. On the right, the shaded light blue area identifies the infinite server regime; on the left, the green area identifies the low load processor sharing regime. The two regimes can be approximated by the following limiting behaviors. At low load the dashed green line is given by the line crossing the point $(0, 0)$ (no connection at zero load) and the point for which, under load $\rho = \lambda_e/\mu_e = 1/2$, the PS queue has an average number of customers equal to 1. At high load, under the infinite server regime, the average number of connections tends to the dashed blue line that is the curve of and ideal $M/M/\infty$ queue with service rate μ_H, where

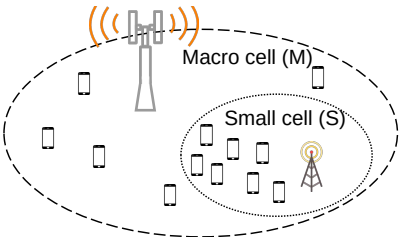

Fig. 4. The two-cell layout that we consider in this paper: one macro cell is overlaid to a small cell, which is positioned so as to absorb the peaks of traffic generated in a hot spot.

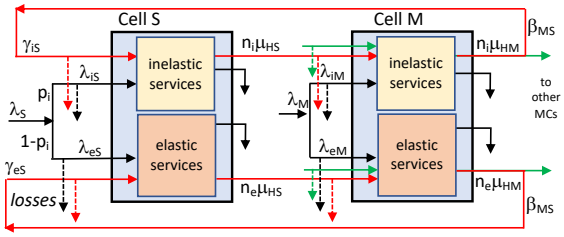

Fig. 5. A network of two queues that models a macro base station and a small cell base station with elastic and inelastic traffic.

we subtract to the number of customers the term λ_e/μ_e that represents the customers that manage to complete the service at the queue.

The probability to handover, i.e., to move out of the cell before call termination, is shown in Fig. 3. Consistently with what done before, we approximate the low load condition with the case in which there is one customer only in the queue. In this case, reported as the green dashed line in the figure, the probability to handover is given by $\mu_e/(\mu_e+\mu_H)$. At high load, the flow of handovers is given by $\lambda_e-\mu_e$, since all customers handover, except those who complete service; the probability to handover is, hence, given by $(\lambda_e-\mu_e)/\lambda_e$.

An optimal cell deployment should work under the PS regime, so as to reduce service times, and avoid a large number of handovers.

III. The Multiple-Cell Layout

The queuing model presented in the previous section can be used as an elementary block to study multiple-cell layouts by composing several queues in a queuing network. Since here we are interested in understanding the benefit that one or two small cells can carry into a macro cell, we consider for starters the simple two-cell layout of Fig. 4. The case of two small cells is obtained by simply adding one more small cell within the macro cell. One SC, named S, is overlaid to one MC, named M. The total area served by the two cells is denoted by A. The area served by S is denoted by A_S, while the area served by M is denoted by $A_M = A - A_S$. Cell S is positioned so as to absorb the traffic peak that is generated in a hot spot, and relieve the possible congestion occurring in the MC. For this reason, when a user is in the coverage area of both M and S, it is associated with S.

The queuing network representing the system of Fig. 4 is reported in Fig. 5. The end user terminals which are associated with either cell can generate inelastic or elastic service requests, and those requests are served, provided the cell has available resources. If a service request cannot be accepted because of lack of resources, a loss occurs. Those service requests that do start, can either complete before the terminal moves out of the cell, or request a handover from the cell in which they are served, toward the new cell under whose coverage the terminal moves. If a handover request cannot be accommodated because of lack of resources in the new cell, a loss occurs.

Given the topological layout of our system, services active within M can either request a handover toward S or toward other neighboring (macro) cells. These two events happen for fractions β_{MS} and $1-\beta_{MS}$ of the handovers, respectively. Conversely, services active within S can only request handovers toward M ($\beta_{SM}=1$). Moreover, M can receive handover requests from several neighboring macro cells at rates γ_{iM} and γ_{eM}. Red lines in Fig. 5 refer to handovers that flow between M and S, while green lines indicate handovers in or out of M, involving neighboring cells.

We denote by δ_M and δ_S the density of users, either active or inactive, in number of users per square meter in cells M and S, respectively. Since S was deployed to absorb a peak of traffic, we assume $\delta_S = a\delta_M$, with $a \geq 1$ describing the larger user density in S with respect to M. The average numbers of users in M and S can be written as $\delta_M A_M$ and $\delta_S A_S = a\delta_M A_S$, respectively.

Denote by λ_M and λ_S the service request rate in number of requests per second in cells M and S, respectively. Assuming that the need for communication services is the same for the users in the small and in the macro cell, the request arrival rate is proportional to the number of users in the areas, and, hence, $\lambda_S = a\lambda_M A_S/A_M$.

We denote by $1/\mu_{HM}$ and $1/\mu_{HS}$ the average dwell times in M and S, respectively. Since the numbers of customers within cells cannot grow to infinity, at steady-state the mobility from S to M, expressed as the average number of users going from S to M in the time unit, balances that from M to S; therefore:

$$\beta_{MS}\delta_M A_M \mu_{hM} = a\delta_M A_S \mu_{hS} \tag{18}$$

from this, we can relate the mobility parameters in the two cells:

$$\mu_{hS} = \frac{\beta_{MS}}{a}\frac{A_M}{A_S}\mu_{hM} \tag{19}$$

These flows include both active and inactive users. Active users, that generate handover flows might, instead,

TABLE II
TRANSITION RATES OUT OF STATE $(n_{iM}, n_{eM}, n_{iS}, n_{eS})$ OF THE MARKOV CHAIN MODELING THE TWO-CELL LAYOUT WITH HETEROGENEOUS TRAFFIC; ONLY STATE VARIABLES THAT CHANGE VALUE ARE INDICATED

Destination	Rate
$(n_{iM}+1, \cdot, \cdot, \cdot)$	$\lambda_{iM}+\gamma_{iM}$
$(\cdot, n_{eM}+1, \cdot, \cdot)$	$\lambda_{eM}+\gamma_{eM}$
$(n_{iM}-1, \cdot, \cdot, \cdot)$	$n_{iM}[\mu_H(1-\beta_{MS})+\mu_i]$
$(n_{iM}-1, \cdot, n_{iS}+1, \cdot)$	$n_{iM}\mu_H\beta_{MS}$
$(\cdot, n_{eM}-1, \cdot, \cdot)$	$(C_M-n_{iM}R_i)\phi_e+n_{eM}\mu_H(1-\beta_{MS})$
$(\cdot, n_{eM}-1, \cdot, n_{eS}+1)$	$n_{eM}\mu_H\beta_{MS}$

not be balanced by effect of blocking probabilities and performance of the cell in serving elastic users.

As regards the capacity of the two cells, it can be reasonable to assume that either M belongs to an earlier generation (for example, M could be in 4G technology) with respect to S (which could be in 5G technology), or the two cells are implemented with the same technology, but their peak powers are trimmed to obtain the desired coverage areas. The capacity of the two cells depends on the portion of licensed spectrum activated in each cell, and is an important design parameter. We denote by C_M and C_S the capacity in bits per second in M and S, respectively.

The state of the network of queues is described by the number of inelastic and elastic customers in service at each queue: $\underline{\mathbf{n}} = (n_{iM}, n_{eM}, n_{iS}, n_{eS})$.

The analysis of this queuing network requires the solution of a continuous-time Markov chain with a state space $\mathcal{N}$:

$$\begin{aligned}\mathcal{N} \;=\; \{&\underline{\mathbf{n}} = (n_{iM}, n_{eM}, n_{iS}, n_{eS}) \mid \\ &n_{iM}R_i + n_{eM}R_e \le C_M, \\ &n_{iS}R_i + n_{eS}R_e \le C_S\}\end{aligned} \tag{20}$$

whose cardinality is of the order of

$$\frac{1}{2}\frac{C_M}{R_i}\frac{C_M}{R_e}\frac{C_S}{R_i}\frac{C_S}{R_e} \tag{21}$$

Transition rates out of state $\underline{\mathbf{n}}$ are reported in Table II. For each transition, in the destination state, only the state variables that change value are indicated. Handovers between M and S are explicitly modeled in the Markov chain (red lines of Fig. 5); these correspond to transitions for which a customer leaves a cell and enters the other one, i.e., the number of customers in one cell decreases by one while the number of customers in the other cell increases by one. The flow of handovers between the MC and neighboring cells (green lines of the figure), not explicitly modeled in the queuing network, is instead derived through a fixed-point approximation procedure. Indeed, in equilibrium conditions, the handover flow from macro cell M to other macro cells balances the flow of incoming handovers from other macro cells to M. By iteratively solving the model, the flow intensity can be numerically computed.

TABLE III
DEFAULT PARAMETER VALUES

Parameter	Values	Parameter	Values
C_M	50 Mb/s	C_S	50 Mb/s
A	3.14 km^2	a	1
R_i	500 kb/s	R_e	300 kb/s
p_i	0.5	$1/\phi_e$	100 Mb
$1/\mu_{HM}$	$\{120, 1200\}$ s	$1/\mu_i$	180 s
λ	0.6 request/s		

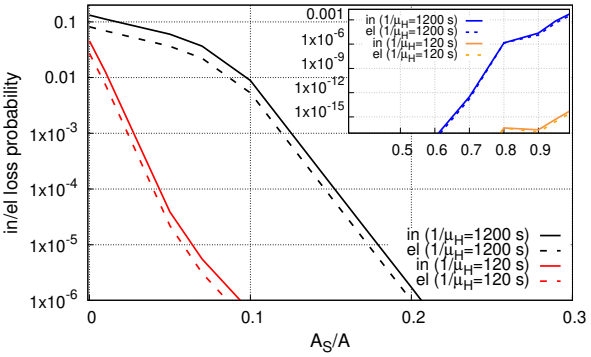

Fig. 6. Loss probabilities in the macro cell and in the small cell (inset) versus cell area ratio for elastic (el) and inelastic (in) services, with $1/\mu_{HM} = 1200$ or 120 s

Once the model is solved and the steady-state probability for each state $\underline{\mathbf{n}}$ is computed, performance indicators can be derived in a similar way to what was described in Sec. II-B. Formulas are not reported here for the sake of brevity.

More complex layouts, comprising a larger number of cells, can be modeled by composing several queues in a queuing network in which customers move among queues according to users mobility. If the state spaces generated by such larger network configurations are too large for an exact solution, approximate approaches are possible.

IV. INTERACTION BETWEEN THE SMALL CELLS AND THE MACRO CELL

Numerical results that illustrate the interaction between a macro cell M and one or two small cells (S when only one cell is present; S_1 and S_2 in the case of two small cells) are obtained with the analytical model described in the previous section, for the values of parameters listed in Table III, unless otherwise specified.

A. Impact of the small cell area and capacity

In Fig. 6 we show the loss probabilities in M and S (inset) versus the ratio A_S/A, with $R_e = 300$ kb/s, for $1/\mu_{HM} = 1200$ s (slow mobility, black curves) or 120 s (fast mobility, red curves). We can see that the loss probabilities for the two types of services are quite close (as expected, since the difference between R_i and R_e is small) and that, for slow mobility, the impact of S

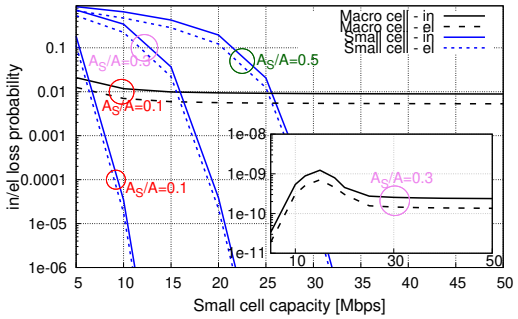

Fig. 7. Loss probabilities in the macro cell and in the small cell versus the small cell capacity, with three different cell area ratios, with $1/\mu_{HM} = 1200$ s

becomes significant when its area is over 10% of area A. Fast mobility increases the effectiveness of S, because more frequent handovers allow a better exploitation of its capacity, since they bring more traffic to S. The loss probability in S remains quite low until its area becomes close to the full service area. Note that the curves are not symmetrical with respect to the area ratio 0.5 because M receives, in addition to all handovers from S, incoming handovers from neighboring macro cells. These results suggest that under a low mobility scenario, the impact of S is marginal, unless the cell becomes significantly large with respect to M. The cost of deploying a small cell might thus be worth only if mobility is high.

Fig. 7 shows how the loss probabilities vary as functions of the small cell capacity, when the area of S, A_S is 10%, 20%, and 50% of the total area A. When the small cell capacity is very low (5 Mb/s), losses at S are very high. Increasing the capacity of S implies fewer losses, hence initially more handovers toward M, whose loss probability increases. Further increases in the capacity of S imply a large reduction of the loss probability at the small cell. The effect on M is a small decrease and then a stabilization (see inset). This is due to the fact that additional capacity in S does not improve the performance of M because the traffic that can be served by S is little (due to its limited coverage). These observations lead to conclude that if the small cell has a limited size, it is not useful to equip it with large capacity if the objective is to improve the performance of the macro cell.

The average number of connections per service type in the two cells is reported in Fig. 8 as a function of the ratio A_S/A. The beneficial effect of S on the performance of M is clearly visible, even for small values of A_S, in the fast drop of the number of elastic services. With no small cell, the macro cell is overloaded, and the number of elastic services grows to large values because their data rate requirement is lower than for inelastic services. As soon as S absorbs a sizable portion of traffic, the congestion in M is reduced, elastic connections are served at higher data rate, and their average number

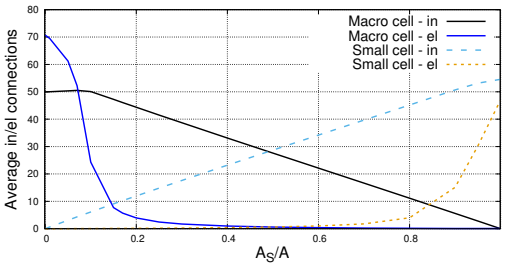

Fig. 8. Average number of inelastic and elastic connections versus cell area ratio, with $1/\mu_{HM} = 1200$ s

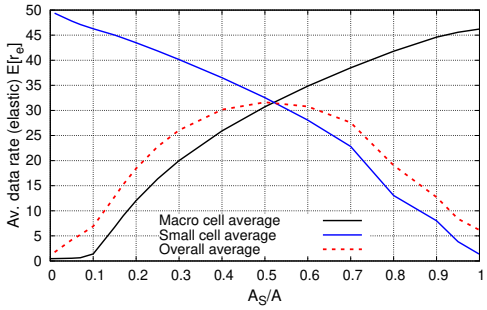

Fig. 9. Average data rate used by an elastic service versus cell area ratio, with $1/\mu_{HM} = 1200$ s

quickly decreases, until it becomes very small because they are served very fast. The different classes of service perceive a different benefit from the introduction of S: elastic services get larger advantage from the capacity of S with respect to inelastic services.

The same conclusion can be drawn from Fig. 9, that shows the average data rate used by an elastic service in Mb/s versus the ratio A_S/A, with $R_e = 300$ kb/s, and $1/\mu_{HM} = 1200$ s. When the area of S is low, the (few) elastic services at the small cell receive a high data rate, while M only provides a low data rate, close to the minimum, to the many elastic connections it serves. However, as soon as S absorbs more traffic, M becomes less congested and the data rate increases to reasonable levels. The most fair case is close to area ratio 0.5, but high data rates of at least 20 Mb/s are achieved in both the macro and the small cell with area ratios between about 0.3 and 0.7.

B. Impact of the service request rate

The dependence of the loss probability of inelastic service requests on the traffic intensity is shown in Fig. 10 for the case in which the two cell service areas are the same, i.e., $A_S = A_M = A/2$, and for both fast and slow mobility. Despite the areas being the same, as previously observed, there is an asymmetry in the behavior of the two cells due to fact that M receives handover traffic from neighboring macro cells in addition to all handovers from S.

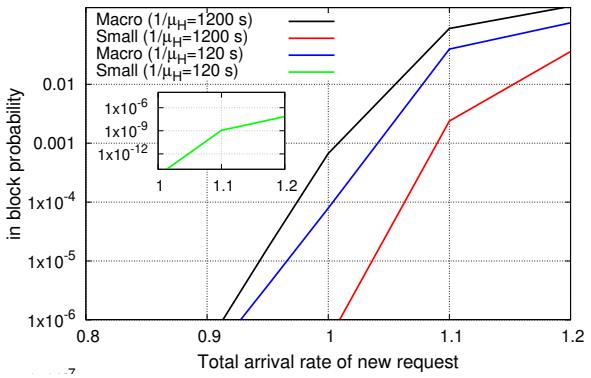

Fig. 10. Loss probability for inelastic traffic in the macro cell and in the small cell versus the total arrival rate λ, with $A_S/A = 0.5$, and $1/\mu_{HM} = 1200$ s or 120 s

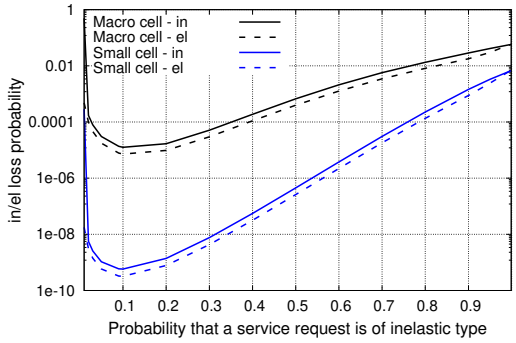

Fig. 11. Loss probability for elastic and inelastic services in the two cells versus p_i in the case $A_S = A_M$, with $1/\mu_{HM} = 1200$ s

C. Impact of the service request mix

The effect of the traffic mix is visible in Fig. 11, where we plot the loss probabilities for elastic and inelastic services in the two cells versus the probability that a service request is of inelastic type, with $\lambda = 1$. The minimum loss probabilities in the two cells are obtained with large percentages of elastic traffic (close to 90%). Since the service request arrival rate is equal to 1, with all elastic services the two cells are at saturation (1 service request per second, each amounting to 100 Mb on average, and two cells with 50 Mb/s data rate each). The addition of some inelastic services reduces the load, since each elastic service requires only 90 Mb (0.5 Mb/s for 180 s on average). This makes the loss probability initially decrease. However, since the system remains close to saturation, we must also consider that elastic services receive close to their minimum data rate (0.3 Mb/s), about half the one of inelastic services. When their arrival rate decreases, more inelastic services enter the system, each consuming more bandwidth. Being the system close to saturation, the cells become progressively full of inelastic services, each using a larger share of resources, thus making the loss probability grow.

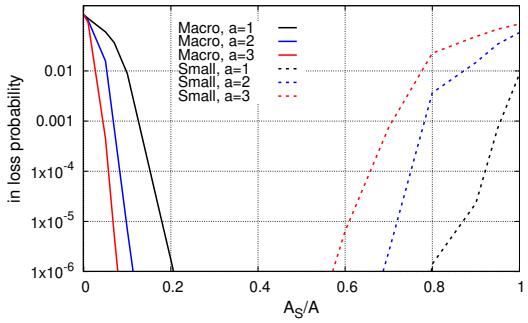

Fig. 12. Loss probability for inelastic traffic in the macro cell and in the small cell versus cell area ratio, with $1/\mu_{HM} = 1200$ s for different values of amplification factor a

D. Impact of the user density in the small cell

We now consider the case of a traffic hot spot, i.e., we look at a scenario in which users are not uniformly distributed over the area A. We assume that in the small cell area A_S the traffic per unit area is a times higher than in M. Fig. 12 reports the loss probability of inelastic services in M and S for $a = 1, 2, 3$ (the loss probability of inelastic services is similar and omitted here for the sake of readability). The total service request arrival rate λ is the same in all cases (λ=0.6), but the partition of requests over the two cells changes, increasing the density of traffic in S with respect to M. As a increases, the loss probability in M decreases due to the smaller density of traffic in M. When a is large, the beneficial effect of S on the performance of M is visible even for low small cell area. As expected, in presence of hot spot behaviors, S can be very effective for the improvement of the system performance. The asymmetry in the curves, as the ratio A_S/A goes to 1, is due mainly due to the uneven distribution of the traffic.

E. Resource utilization

Up to now we have taken the viewpoint of the end user, since we have examined loss probability, which is the most relevant QoS index for inelastic services, and received data rate, which is the key performance indicator for elastic services. Let us now take the point of view of the network operator, and look at the utilization of network resources.

In Figs. 13, 14 and 15 we plot the curves of the utilization of the macro and small cell capacity by the two types of services versus i) the cell areas ratio, ii) the total service request rate, and iii) the fraction of inelastic service requests, with $R_e = 300$ kb/s and $1/\mu_{HM} = 1200$ s. The utilization of capacity at both cells varies almost linearly with the cell areas (see Fig. 13), but at load 0.6 the total bandwidth of the two cells is not well utilized, since they collectively receive as input

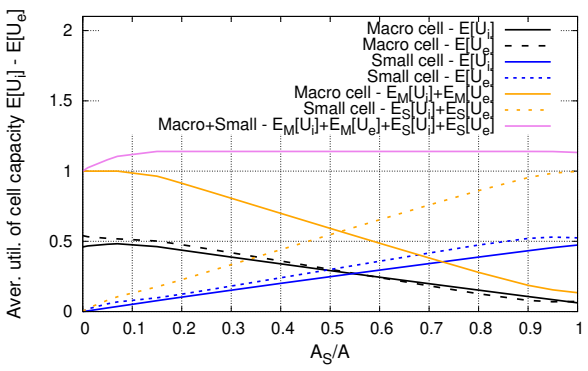

Fig. 13. Average capacity utilization at the macro and small cell versus the small cell area, with $\lambda = 0.6$, and $1/\mu_{HM} = 1200$ s

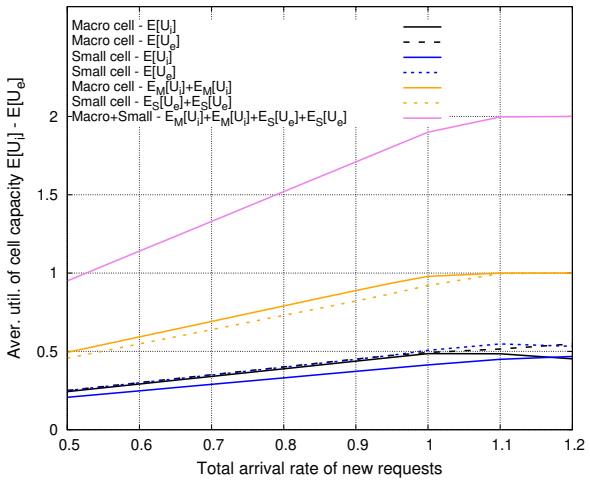

Fig. 14. Average capacity utilization at the macro and small cell versus arrival rate, with $A_S/A = 0.01$, and $1/\mu_{HM} = 1200$ s

57 Mb per second to transfer[2], and each cell data rate is 50 Mb/s. Note that the ratio 57/50 equals 1.14, which is the plateau of the pink curve in Fig. 13.

Progressively increasing the load (see Fig. 14) with equal cell area, it is possible to fully exploit the bandwidth of the two cells, but only at very high loads. In particular, with overall arrival rate equal to 1, equal areas and $p_i = 0.5$, each cell receives 47.5 Mb to transfer every second. Both cells are thus close to overload.

The behavior for variable values of p_i (see Fig. 15) for an overall arrival rate equal to 1 shows that the presence of elastic services is beneficial for the exploitation of the installed capacity, especially at S.

Finally, Fig. 16 shows that, in presence of a traffic hot spot, a small cell of adequate size can be quite effective in absorbing a significant fraction of the load.

[2]The total input is computed as load times elastic service file size, times fraction of elastic services plus load times duration of inelastic services times data rate times fraction of inelastic services: $0.6 \times 100 \times 0.5 + 0.6 \times 180 \times 0.5 \times 0.5 = 30 + 27 = 57$.

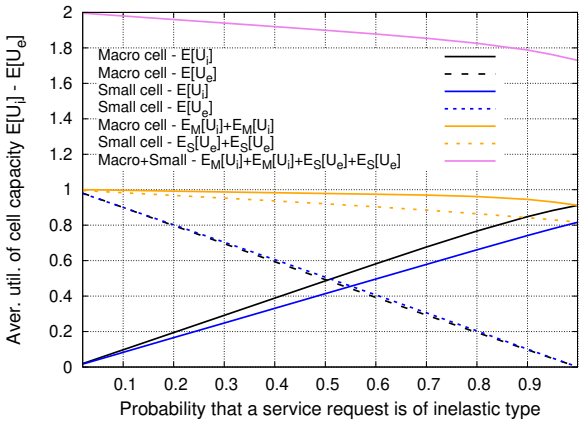

Fig. 15. Average capacity utilization at the macro and small cell versus fraction of inelastic service requests, with $\lambda = 1$, $A_S/A = 0.01$, and $1/\mu_{HM} = 1200$ s

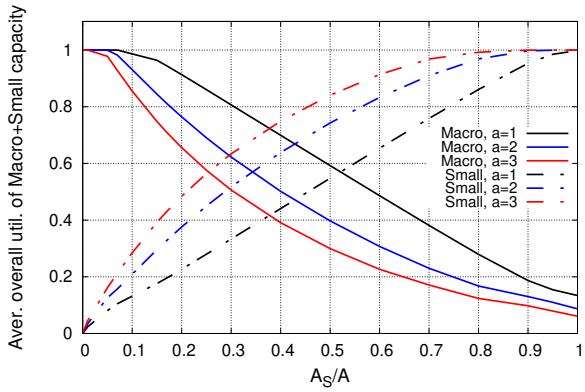

Fig. 16. Average capacity utilization at the macro ($E_M[U_i] + E_M[U_e]$) and small cell ($E_S[U_i] + E_S[U_e]$) versus arrival rate, with $\lambda = 0.6$, $A_S/A = 0.01$, and $1/\mu_{HM} = 1200$ s for different values of hot spot traffic amplification factor a

F. Two Small Cells

A configuration with one macro cell M and two small cells S_1 and S_2 can be studied with a queuing network comprising three queues properly interconnected.

In Fig. 17 we report loss probabilities for inelastic and elastic customers in the cases of a circular area with 1 km radius covered by either just M, or M and S_1, or M, S_1 and S_2. The small cells areas are either 1% or 10% of the macro cell area. The average dwell time in the macro cell is $1/\mu_{HM} = 1200$, and those in the small cells are set proportionally. Small cells are not contiguous, and traffic is uniform over the area served by M.

We can observe that when small cells are very small, they bring small benefit, mostly because of traffic uniformity, and elastic services benefit more than inelastic services, as we already observed. When small cells are larger, they produce larger benefit, and two small cells are better than one, because they absorb more traffic in the area. If S_1 and S_2 are contiguous, so that handovers between small cells are possible, loss probabilities (not shown) are even lower, but the difference is not huge. This also happens when traffic over S_1 and S_2 is

proportionally higher than over M. As expected, the more peaked the traffic over the small cells is, the larger the improvement is (not shown, for the sake of brevity).

V. Discussion

The analysis of the numerical results generated by our model leads to many interesting considerations.

First of all, we can see that, in case of uniform distribution of users in the area of interest, for good performance the area covered by the small cell must not be too small (at least 20-30% of the overall area). This means that in the idealized case in which the macro cell covers a circular region of radius 1 km, the small cell should cover a circular region with radius of the order of 500 m. This means that cells that are small due to propagation constraints, coming from the combination of physical location of the BS and the transmission power, might not be that effective in improving the overall performance. Second, the small cell capacity must be comparable to the macro cell capacity and higher capacity of the small cell is not needed, unless the small cell covers a hot spot with much higher density of users and service demand. Hence, in the scenarios that are often proposed for the transition from 4G to 5G, with 4G macro cells and high capacity 5G small cells, the large capacity of the small cell might turn out to be little effective and under-utilized.

The non-uniformity in traffic load, corresponding to higher traffic density in the area of the small cell has beneficial effects on the interaction between the macro and the small cell, as expected. While under uniform traffic distribution, the presence of a small cell can be beneficial for the macro cell only if it is large enough to absorb a significant amount of traffic, the presence of a small cell in correspondence of a hot spot can be very effective.

VI. Related Work

Research in planning, design and performance evaluation of cellular networks has attracted huge attention in the last decades, stimulated by the explosive success of mobile communications, and by the continuing evolution of cellular network generations and architectures. Now, with the arrival of 5G and the corresponding novelties in architectural features, cellular network planning and design issues are again in the hot spot.

A recent survey of cellular network planning issues can be found in [2]. Specific works in cellular network performance analysis and design are for example [4]–[6]

Particularly relevant to this work is a previous paper focusing on the same HetNet layout, but only considering inelastic services [7]. The coexistence of inelastic and elastic services was considered in [9], aiming at the design of admission control strategies, and in [3], with the objective of designing channel-aware scheduling algorithms.

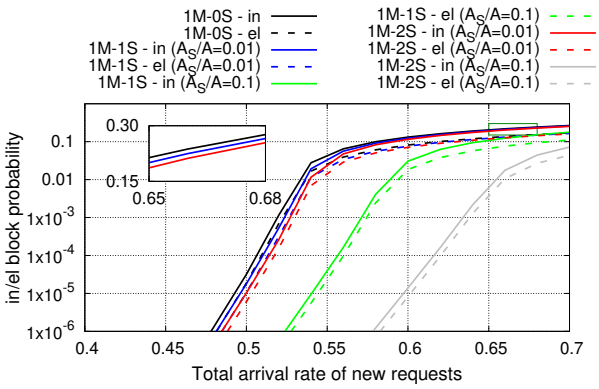

Fig. 17. Loss probability for inelastic and elastic services in the macro cell and in the small cell(s), for 0, 1, or 2 small cells, versus the total arrival rate λ, for two sizes of small cells, with $1/\mu_{HM} = 1200$ s

VII. Conclusions

We studied a HetNet scenario, where one macro cell is overlaid to one or two small cells deployed so as to absorb traffic from the macro cell and where elastic and inelastic services coexist. Performance results indicate that, with standard system parameters, the presence of the small cell can be effective and significantly improve the performance of the served area, only if the small cell is large enough to absorb a good amount of the macro cell traffic or under an uneven distribution of traffic such that the small cell covers a hot spot where quite some traffic is generated. Our analysis shows also that elastic services take larger advantage of the presence of a (possibly little loaded) small cell.

References

[1] M. A. Khan, S. Leng, W. Xiang, K. Yang, "Architecture of heterogeneous wireless access networks: A short survey," TENCON 2015, Macao, pp. 1-6.

[2] A. Taufique, M. Jaber, A. Imran, Z. Dawy, E. Yacoub, "Planning Wireless Cellular Networks of Future: Outlook, Challenges and Opportunities," IEEE Access, vol.5, 2017.

[3] S. Borst, N. Hegde, "Integration of Streaming and Elastic Traffic in Wireless Networks," IEEE INFOCOM 2007, Barcelona, 2007, pp. 1884-1892.

[4] G. Harine, R. Marie, R. Puigjaner, K. Trivedi, "Loss formulas and their application to optimization for cellular networks," IEEE Trans. on Vehicular Technology, vol.50, n.3, May 2001.

[5] M. Ajmone Marsan, S. Marano, C. Mastroianni, M. Meo, "Performance analysis of cellular mobile communication networks supporting multimedia services," 6th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, Montreal, Que., 1998, pp. 274-281.

[6] M. Ajmone Marsan, G. de Carolis, E. Leonardi, R. Lo Cigno, M. Meo, "Efficient estimation of call blocking probabilities in cellular mobile telephony networks with customer retrials," IEEE JSAC, Vol.19, n.2, pp. 332-346, 2001.

[7] M. Ajmone Marsan, F. Hashemi, "Deploying Small Cells in Traffic Hot Spots: Always a Good Idea?," PIMRC 2018, Bologna, Italy.

[8] M. Sidi, D. Starobinski, "New call blocking versus handoff blocking in cellular networks," Wireless Networks, vol.3, pp. 15-27, 1997.

[9] N. Benameur, S. B. Fredj, F. Delcoigne, S. Oueslati-Boulahia, J. Roberts, "Integrated admission control for steaming and elastic traffic," QoFIS 2001, Coimbra, Portugal.