Deploying Small Cells in Traffic Hot Spots: Always a Good Idea?

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

Fatemeh Hashemi
Politecnico di Torino, Italy
fatemeh.hashemi@studenti.polito.it

***Abstract*—We look at a very simple RAN (Radio Access Network) configuration comprising one macro cell and one small cell, the latter being strategically positioned to absorb the traffic peaks that occur in some time periods in a portion of the area covered by the macro cell. We study this two-cell system with a simple model based on a network of two queues, and we examine the system performance for variable parameter values, showing that some of the emerging behaviors can be critical. In particular, we see that when the handover rate out of the small cell increases, the blocking probability in the macro cell also increases, quickly reaching unacceptable levels. This can be a problem, since high handover rates correspond to limited dimensions of the small cell with respect to the macro cell, which is what is normally expected, unless the small cell is deployed in an area of very slow end user mobility. These behaviors (although possibly not applicable to all small cell scenarios) can have an important impact on the deployment of small cells, which are expected to become increasingly popular because of the need to provide additional capacity in RANs through densification of the cell layout.**

I. Introduction

The increasingly rapid succession of mobile network generations, and the availability of different types of equipment, has given rise to a situation where Base Stations (BSs) of different types and generations coexist within a Radio Access Network (RAN), and offer services to end users, on either the same or different frequency bands. Macro BSs typically use higher power, hence have longer reach (of the order of one or few kilometers) and higher coverage. The corresponding cells are normally termed macro cells. On the contrary, micro or pico BSs have shorter reach and cover areas with radius of the order of 100 meters. The corresponding cells are consequently termed small cells. They are expected to be the main instrument to achieve network densification, with the objective of providing capacity increases in areas where the end user request for telecommunication services is high, and to allow for better spectrum reuse, hence better spectral efficiency. Additionally, femto BSs have even lower reach, and are typically used indoor, possibly as a substitute for WiFi.

In dense urban environments, where traffic peaks are extremely high in the most busy areas, it is forecasted that RANs will evolve toward layouts in which small cells overlap with macro cells, and absorb a large fraction of the traffic generated in hot spots. The network architectures that will result from these complex layouts of different BS types are often termed Heterogeneous Networks (HetNets [1]), and comprise several layers of BSs overlaid in a same area, possibly using different Radio Access Technologies (RAT). 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 this paper, we look at the problem of small cell deployment and cell resource dimensioning, and we show that even in simple scenarios, where a small cell is used to absorb the peaks of traffic generated in a hot spot within a macro cell, the impact of the parameters of the small cell on the performance of the macro cell can be quite substantial. More specifically, we look at a simple configuration of two cells, where one small cell is present within a macro cell, and is positioned so as to absorb a high peak of traffic which in a particular time interval is generated in a hot spot. We model this two-cell configuration with a simple network of two queues, and we show that the presence of the small cell has a significant impact on the performance of the macro cell, not always in an easily predictable way.

The rest of this paper is structured as follows. In Section II we present the two-cell layout that we consider in this paper. In Section III we illustrate the queuing network that we use to model the two-cell layout, and we discuss the model solution. In Section IV we present and discuss numerical results. In Section V we discuss some previous work related to our approach. Finally, Section VI concludes the paper.

II. The Two-Cell Layout

We consider the very simple two-cell layout of Fig. 1. One macro cell, named cell A, is overlaid to one small cell, named cell B. The small cell is strategically positioned to absorb the traffic peak that in the considered time interval is generated in a hot spot. Of course, in mobile networks, traffic varies largely both in space and

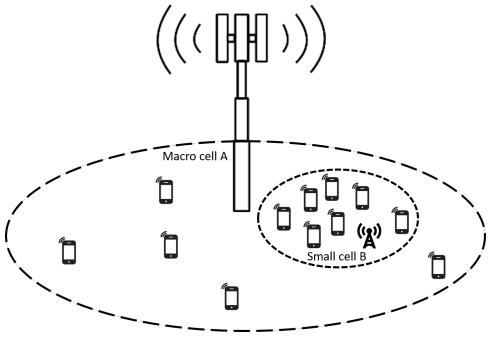

Fig. 1. The two-cell layout that we consider in this paper: one macrocell is overlaid to a small cell, which is positioned so as to absorb the peaks of traffic generated in a hot spot.

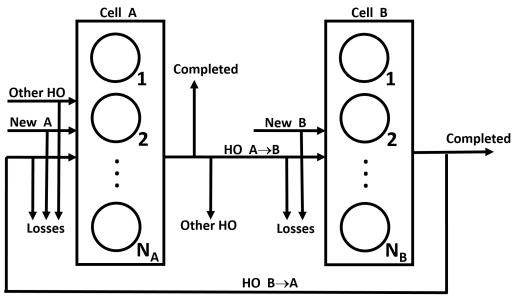

Fig. 2. The network of two queues that is used to model the two-cell layout that we consider in this paper.

in time, and variability becomes more pronounced when small areas are considered, so that the considerations that we develop in this paper are valid only when the small cell is actually absorbing a peak of traffic.

The end user terminals which are associated with either cell can generate service requests, and those requests are served, provided the cell has resources available. If a service request cannot be accepted because of lack of resources, a *loss* occurs. Those service requests that do start, can either terminate 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 two cells, services active within macro cell A can either request a handover toward small cell B or toward other neighboring cells. On the contrary, services active within small cell B can only request handovers toward macro cell A. Symmetrically, the macro cell A can receive handover requests from several neighboring cells, including small cell B, whereas small cell B can only receive handover requests from terminals associated with macro cell A.

As regards the capacity of the two cells, it can be reasonable to assume that the macro cell belongs to an earlier generation (for example, the macro cell could be in 4G technology) with respect to the small cell (which could soon be in 5G technology). We can conclude that the capacity of the small cell is likely higher than the one of the macro cell in terms of number of connected users and possibly bandwidth.

We can also observe that, for a given mobility pattern of end users, it is reasonable to assume that the time users spend within the area covered by a small cell is shorter than the time spent under the coverage of the macro cell. This implies that, under equal traffic load of the two cells, the rate of handovers out of a small cell is higher than the handover rate out of a macro cell. This, of course, under the assumption that the type of service accessed by end user and their mobility pattern do not depend on the cell they are in.

III. The Two-Queue Model

We model the two-cell system with the very simple network of two queues shown in Fig. 2. Both queues are $\cdot/M/N_K/0$, i.e., they have no waiting line, a number of equivalent servers equal to N_K (N_A for queue A and N_B for queue B), exponentially distributed service times (although this type of queues in isolation is known to be invariant with respect to service time distributions), and an arrival process that depends on exogenous arrivals as well as departures from the other queue.

The number of servers at each queue models the maximum number of users that can be accepted in service within the cell.

Arrivals at queue A comprise three components: i) exogenous arrivals that model the new service access requests generated by end users within the macro cell; ii) handover request coming from the small cell in our two-cell system; iii) handover requests coming from cells that are outside the two-cell system. We assume that new call arrivals follow a Poisson process with rate λ_A. Similarly, we assume that handover arrivals from the cells out of our two-cell system follow a Poisson process with rate λ_{hA}. Handover arrivals from the small cell correspond to a fraction of the departures from queue B, and are explicitly represented in the model.

Arrivals at queue B comprise instead just two components: i) exogenous arrivals that model the new service access requests generated by end users within the small cell; ii) handover request coming from the macro cell in our two-cell system. In this case we assume that new call arrivals follow a Poisson process with rate λ_B. Handover arrivals from the macro cell correspond to a fraction of the departures from queue A, and are explicitly represented in the model.

Service times at each one of the two queues depend on two aspects: i) the time necessary to satisfy the request for telecommunication service issued by the end user, that we denote as T_S, and ii) the time spent by the end user within the area covered by the cell, called *dwell*

time and denoted by T_D. Both of these times in both cells are assumed to be random variables with negative exponential distribution, with rates μ_A and μ_B for T_{SA} and T_{SB}, respectively, and μ_{hA} and μ_{hB} for T_{DA} and T_{DB}, respectively.

The queue service time is in both cases the minimum between T_D and T_S (with the appropriate subscripts A or B). Since the minimum of two independent random variables with negative exponential distribution is known to be exponentially distributed with a rate which is equal to the sum of the rates of the two original distributions, we obtain that the average service times at the two queues are:

$$E[S_K] = \frac{1}{\mu_K + \mu_{hK}} = \frac{1}{\mu_{tK}} \tag{1}$$

where K is the queue (or cell) index, and can thus be either A or B.

The nominal loads of the two queues are defined as:

$$\rho_K = \frac{\lambda_K}{\mu_K} \tag{2}$$

while the normalized nominal loads of the two queues are defined as:

$$\rho_{nK} = \frac{\lambda_K}{N_K \mu_K} \tag{3}$$

After service at queue A, a customer in the network of queues has three options: 1) leave the network of queues (because the telecom service completed within the macro cell); 2) move to queue B (this means that the service handovers to cell B); 3) again leave the network of queues (in the case of a handover to a cell which is not within the two-queue system). The corresponding probabilities are:

$$P_{1A} = \frac{\mu_A}{\mu_{tA}} \qquad P_{2A} = \frac{\beta \mu_{hA}}{\mu_{tA}} \qquad P_{3A} = \frac{(1-\beta)\mu_{hA}}{\mu_{tA}} \tag{4}$$

where β is the probability that a handover from the macro cell is directed to the small cell.

After service at queue B, a customer has two options: 1) leave the network of queues (because the telecom service completed within the small cell); 2) move to queue A (this means that the service handovers to cell A). The corresponding probabilities are:

$$P_{1B} = \frac{\mu_B}{\mu_{tB}} \qquad P_{2B} = \frac{\mu_{hB}}{\mu_{tB}} \tag{5}$$

The network of queues of Fig. 2 does not admit a product form solution, because of losses. However, it is easy to translate the queuing network model into a Continuous-Time Markov Chain (CTMC), which can be solved with standard techniques. From the CTMC solution we can compute different performance metrics. However, in this paper we will look at the loss probability in the two cells as our main performance indicator. These loss probabilities can be expressed as:

$$P_{lossA} = \frac{\sum_{k_B=0}^{N_B} (\lambda_A + \lambda_{hA} + k_B \mu_{hB}) \pi_{N_A, k_B}}{\sum_{k_A=0}^{N_A} \sum_{k_B=0}^{N_B} (\lambda_A + \lambda_{hA} + k_B \mu_{hB}) \pi_{k_A, k_B}} \tag{6}$$

$$P_{lossB} = \frac{\sum_{k_A=0}^{N_A} (\lambda_B + k_A \mu_{hA} \beta) \pi_{k_A, N_B}}{\sum_{k_A=0}^{N_A} \sum_{k_B=0}^{N_B} (\lambda_B + k_A \mu_{hA} \beta) \pi_{k_A, k_B}} \tag{7}$$

where π_{k_A, k_B} is the steady-state probability of k_A services in progress in macro cell A, and k_B services in progress in small cell B.

Note finally that the model parameters that correspond to maximum number of users in service in each cell (N_A and N_B), new service request rates (λ_A and λ_B), service duration (μ_A and μ_B), and dwell times (μ_{hA} and μ_{hB}), must be estimated and plugged into the model. On the contrary, λ_{hA}, the rate of handover arrivals to cell A from cells other than the small cell B, which are not represented in the model, with an equilibrium assumption can be taken equal to the rate of handover out of cell A towards those same cells, and thus estimated from the model itself. This requires a fixed point iteration, which can be shown to converge in a small number of steps.

IV. Numerical Results

For starters, we look at the case in which the two-cell system comprises a macro cell that can serve up to $N_A = 16$ simultaneous access requests, and a small cell that can serve up to $N_B = 64$ simultaneous access requests. These values are smaller than the actual values allowed in cellular systems (the maximum number of terminals in the RRC_CONNECT state at a state-of-the-art BS is decided by the equipment manufacturer, and can be of the order of 200), but are sufficient for a description of the phenomena of interest. Choosing $N_B > N_A$ reflects the increase in capacity typical of the newer generations of equipment (soon likely 5G) used for small cells.

The average time to complete a service is taken to be the time unit of our model, so that we set $\mu_A = \mu_B = 1$. No difference is considered between the two cells, since the available bandwidth is assumed to be equal, and the same telecommunication services are assumed to be selected in both the macro and the small cell.

The rate of new requests for telecommunication services in the macro cell is varied between 1 and 15 requests per time unit, so that the normalized nominal load ρ_{nA} for the macro cell varies between $1/16 \simeq 6\%$ and $15/16 \simeq 94\%$. The new request rate in the small cell is set to 20, 40, or 60 requests per time unit. This range of values, higher than for the macro cell, reflects the fact that the small cell is installed so as to absorb a traffic peak in a hot spot. The corresponding normalized nominal load ρ_{nB} of the small cell is thus varied between $20/64 \simeq 31\%$ and $60/64 \simeq 94\%$.

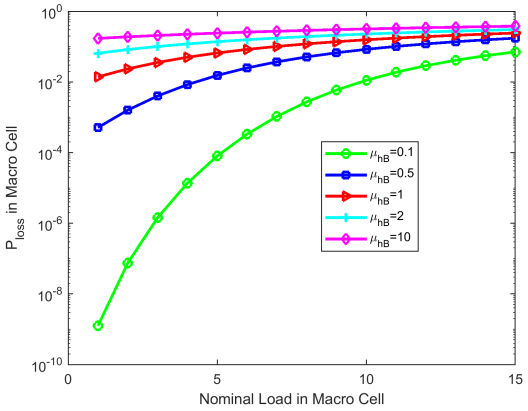

Fig. 3. Loss probability in the macro cell versus the new request arrival rate in the macro cell, for $\lambda_B = 20$, and for different values of μ_{hB}.

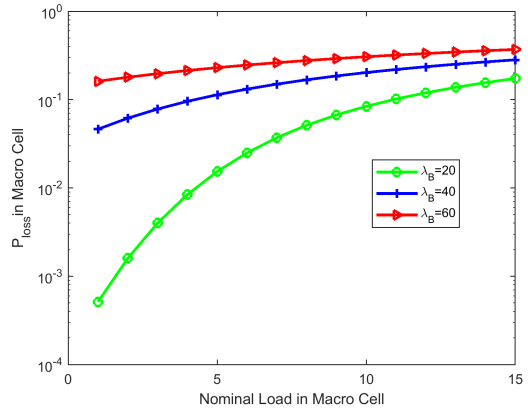

Fig. 4. Loss probability in the macro cell versus the new request arrival rate in the macro cell, for three different values of λ_B, and for $\mu_{hB} = 0.5$.

Given the difference in coverage between the macro and the small cell, assuming that mobility is the same in the two environments, it is natural to choose shorter dwell times in the small cell with respect to the macro cell. However, we must also take into consideration that a small cell could be installed to cover a reduced mobility area, such as a shopping mall or a crowded pedestrian city square, so that, in some contexts, dwell times in the small cell could even be longer than for the macro cell. In order to reflect these possibilities, we set $\mu_{hA} = 1$ (i.e., we select the average dwell time in the macro cell equal to the average time to complete a telecom service), and we vary μ_{hB}.

Finally, we select $\beta = 1/3$, so that one third of the handovers out of the macro cell go to the small cell.

The blocking probability in the macro cell in the scenario we just described is reported in Fig. 3, for $\lambda_B = 20$ (so that $\rho_{nB} \simeq 30\%$), and values of μ_{hB} between 0.1 and 10. The value $\mu_{hB} = 10$ is consistent with the assumption of equal mobility in the two cells,

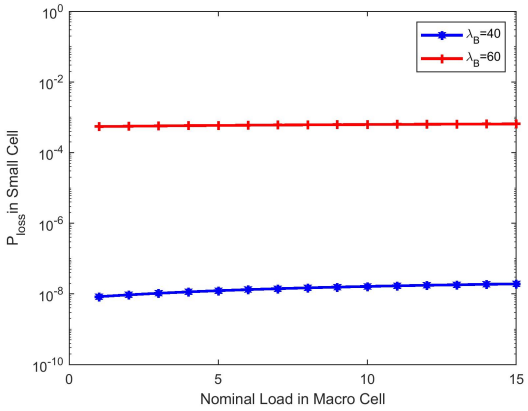

Fig. 5. Loss probability in the small cell versus the new request arrival rate in the macro cell, for three different values of λ_B, and for $\mu_{hB} = 0.5$.

with a ratio between the radiuses equal to 10 (1 km for the macro cell and 100 m for the small cell). Lower values for μ_{hB} reflect larger small cell sizes, or slower mobility within the small cell, and lead to significantly lower loss probabilities in the macro cell. In the same scenario, the blocking probability in the small cell always remains less than 10^{-12}, with higher values again for lower values of μ_{hB}.

In Figs. 4 and 5 we show the blocking probability in the macro cell and in the small cell, respectively, versus λ_A, for variable values of λ_B, and for $\mu_{hB} = 0.5$.

The results in Figs. 3, 4, and 5, show that the impact of the dwell time in the small cell on the loss probability in the macro cell is big. This is expected, since the small cell absorbs a large quantity of traffic, but, if the dwell time in the small cell is short (high values of μ_{hB}, which mean that the cell is indeed small and mobility is the same as in the macro cell), with high probability this traffic handovers to the macro cell, and consumes resources there. If we assume a target loss probability in the macro cell of the order of 0.01, we see that, with the chosen parameters, this is achievable only with $\mu_{hB} = 0.5$ or 0.1, i.e., with slower mobility in the small cell with respect to the macro cell (much slower, accounting for the fact that the small cell is small). This result is not a surprise, in light of the results in [3], [4], which showed how critical it can be to have very high handover rates. As regards the small cell, with large values of μ_{hB}, loss probabilities (not plotted) are extremely small, since new service requests absorbed by the small cell consume resources for a short time before moving to the macro cell. Even with larger values of μ_{hB}, with $\lambda_B = 20$, loss probabilities in the considered scenario are always less than 10^{-12}.

Higher values of exogenous traffic in the small cell (i.e., new service access requests, with rate λ_B) obviously make the loss probability in the small cell increase;

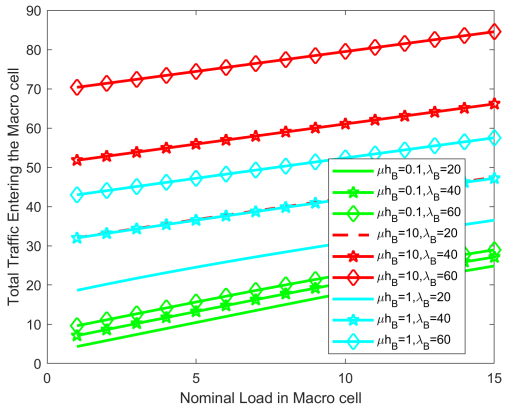

Fig. 6. Total traffic entering the macro cell versus the new request arrival rate in the macro cell, for three different values of λ_B = 20, 40, 60, and for three values of μ_{hB} = 0.1, 1, 10.

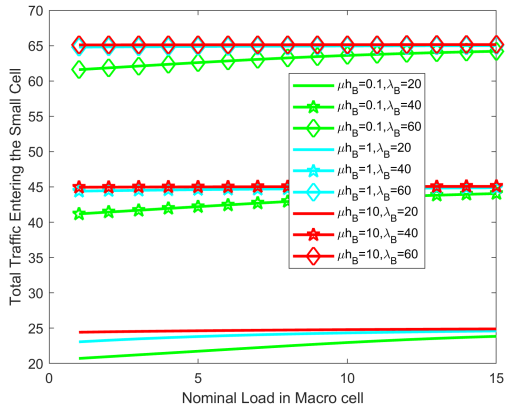

Fig. 7. Total traffic entering the small cell versus the new request arrival rate in the macro cell, for three different values of λ_B = 20, 40, 60, and for three values of μ_{hB} = 0.1, 1, 10.

however, they also significantly impact the performance of the macro cell. The increase from $\lambda_B = 20$ to $\lambda_B = 60$ makes the loss probability in the macro cell at $\lambda_A = 1$ (normalized nominal load of the macro cell around 6%) grow from about 0.0005 to over 0.1, with $\mu_{hB} = 0.5$, as can be seen in Fig. 4.

This large influence of the behavior of the small cell on the performance of the macro cell is not coupled with an equivalent influence of the macro cell on the small cell, where the loss probability is rather insensitive to the value of λ_A, as can be seen in Fig. 5.

To summarize the message that can be extracted from the results we saw so far, we can say that if we install a small cell within a macro cell to absorb a traffic peak, and as a result of this, the normalized nominal load of the macro cell is about 6%, while the load of the small cell is about 30%, with the parameters we considered, the loss probability in the small cell is negligible (as expected, since load is low), but the loss probability in the macro cell can be anywhere between 10^{-9} and 0.1, depending on the parameter μ_{hB} (and we are just considering values of μ_{hB} between 0.1 and 10), which jointly reflects the size of the small cell and the end user mobility in the small cell. This means that a (very) small cell positioned in an area with not very slow end user mobility can be a critical element for the performance of the macro cell.

In order to understand the reason of this not so obvious interplay between the small cell and the macro cell, in Figs. 6 and 7 we show the total traffic entering the macro cell and the small cell, respectively, (including new requests as well as *all* handovers) versus the new request arrival rate in the macro cell, for three different values of λ_B = 20, 40, 60, and for three values of μ_{hB} = 0.1, 1, 10. The results show that the total traffic entering the macro cell heavily depends on μ_{hB}, in addition to λ_A (note that the two curves for $\lambda_b = 40, \mu_{hB} = 1$ and for $\lambda_b = 20, \mu_{hB} = 10$ overlap), while the total traffic entering the small cell marginally depends on λ_A, and shows a limited dependence on μ_{hB}.

As we mentioned before, variations in the values of dwell times and new call request rates in the two cells can reflect changes in the relative size of the two cells, and in the user population density. A further step of interest in our analysis can relate to the selection of the most effective respective capacities of the two cells. For this reason, we keep fixed to 80 the total maximum number of services in progress at the two cells ($N_A + N_B = 80$), but we vary the values of N_A and N_B in the range from 8 to 72, with $\mu_A = \mu_B = 1$, $\mu_{hA} = 1$, and in a first case $\lambda_A = 10$, $\lambda_B = 20$, $\mu_{hB} = 2$, in a second case $\lambda_A = 20$, $\lambda_B = 10$, $\mu_{hB} = 2$, and in a third case $\lambda_A = 10$, $\lambda_B = 20$, $\mu_{hB} = 10$. Note that the overall input traffic to be served ($\lambda_A + \lambda_B$) is kept constant at 30 requests per time unit, and that in the first two cases $\mu_{hB} = 2\mu_{hA}$, while in the third case $\mu_{hB} = 10\mu_{hA}$. The first case corresponds to a small cell that absorbs 2/3 of the total load of the area, and an average dwell time equal to half the one of the macro cell. Instead, the second case corresponds to a small cell that absorbs 1/3 of the total load of the area, and an average dwell time again equal to half the one of the macro cell. Finally, the third case corresponds to a small cell that absorbs 2/3 of the total load of the area, and an average dwell time equal to 1/10 of the one of the macro cell. This means that case 1 could correspond to a small cell that is not so small, and absorbs a large fraction of the macro cell coverage area traffic. Case 2 corresponds to a smaller cell that absorbs less than half of the area traffic, and is positioned so as to serve users with limited mobility. Case 3 refers to a very small cell that absorbs a majority of the area traffic, where mobility is higher due to the cell size (this is the paradigmatic case for the use of small cells). Also consider the fact that, if we could use only the macro cell with 80 parallel services to serve all

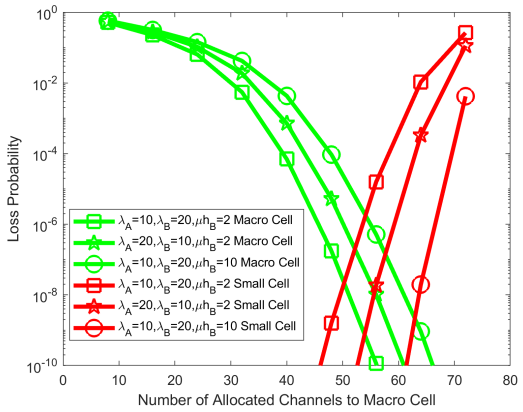

Fig. 8. Loss probabilities in the macro cell and in the small cell with $N_A + N_B = 80$, versus N_A, with $\mu_A = \mu_B = 1$, $\mu_{hA} = 1$, and $\lambda_A = 10$ or 20, $\lambda_B = 20$ or 10, $\mu_{hB} = 2$ or 10.

input traffic, the loss probability would be negligible.

Fig. 8 shows the loss probability in the macro cell and in the small cell in those cases. The results show that, in order to obtain acceptable loss probability in both cells (i.e., lower than 10^{-3}), in all three cases, it is necessary to allocate to the macro cell at least half of the total amount of resources (i.e., N_A must be chosen between 40 and 48). In addition, we see that going from case 2 to case 1, i.e., reducing the fraction of the total load associated to the small cell, while keeping the same mobility, we can reduce the quantity of resources in the macro cell, as intuitively expected. Instead, if we reduce the fraction of the total load associated to the small cell, while increasing mobility to account for the smaller cell size, we need to increase the quantity of resources in the macro cell for a given performance target. This conveys the message that for a small cell to have a positive impact on performance, the small cell capacity must be less than the capacity of the macro cell, and the good values of the ratio between the macro and small cell capacity increase with the rate of mobility in the small cells. In a nutshell, this tells us that small cells of very small size with high capacity to serve traffic hot spots may not be a good idea, unless they are deployed in very low mobility areas.

V. Related Work

The research topics evolving around cellular network planning, design and performance evaluation have received a huge attention in the last decades, due to the increasing success and relevance of mobile communications, and to the rapid evolution of cellular network generations and architectures.

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 [5]–[8].

Particularly relevant to this work are the already cited papers [3], [4], that show that for very high mobility the performance of cellular networks degrades significantly. This is a phenomenon similar to what we observed, which has the potential of becoming critical in the case of cells of very small dimension, unless those small cells cover areas where end user mobility is limited.

VI. Conclusions

We studied a very simple HetNet scenario, where one macro cell is overlaid to one small cell deployed so as to absorb a traffic hot spot. Performance results indicate that, with standard system parameters, the presence of the small cell can be very critical in terms of loss probability at the macro cell. The root of the problem is in the short end user dwell time in the small cell, which makes traffic absorbed by the small cell quickly return to the macro cell. The main takeaway message of our study is that cells of very small size with high capacity overlaid to a macro cell to serve traffic hot spots may not be a good idea, unless they are deployed in very low mobility areas, so that they can actually handle a significant portion of the hot spot traffic, before the end users handover out of the small cell.

In our study we did not consider the impact of handover procedures, which on the one hand imply an overhead, but on the other could implement a smart choice of those users that are transferred to the small cell, selecting only the ones with low mobility, so as to avoid the effect that we observed. The results we presented are not necessary valid for all small cell deployments. The investigation of more complex scenarios is necessary to see what happens in the case of multiple contiguous small cells and of real end user mobility patterns.

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, pp.4821-4845, 2017.

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

[4] K.Sohraby, "Blocking and forced termination in pico-cellular wireless networks: An asymptotic analysis," 9th Annual Workshop on Computer Communications, 1994.

[5] 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, pp.664-673, May 2001.

[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, 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.

[8] M.Ajmone Marsan, G.Ginella, R.Maglione, M. Meo, "Performance analysis of hierarchical cellular networks with generally distributed call holding times and dwell times," IEEE Trans. on Wireless Communications, vol.3, n.1, pp.248-257, Jan. 2004.