Increasing the Capacity of Large-Scale HetNets through Centralized Dynamic Data Offloading

Henrik Klessig, Michael Günzel, and Gerhard Fettweis
Vodafone Chair Mobile Communications Systems
Dresden University of Technology, Germany
Email: {henrik.klessig, michael.guenzel, fettweis}@tu-dresden.de

Abstract*—Typically, mobile users cluster around points of interest in dense urban environments such as city centers forming so-called data traffic hot spots and hot zones. To provide capacity to such users efficiently, mobile operators deploy small cells. However, the deployment of heterogeneous networks, which consist of overlaying macro cells and many co-channel small cells, entails many problems. One typical problem is that, more often than not, hot spot users are not covered by the small cells due to the spatially fluctuating nature of the traffic demand. Data offloading, meaning actively shifting macro cell users to small cells, is a promising approach to address this issue. In this paper, we extend a queuing-theoretic model based on the notion of elastic data flows in order to model data offloading, or more specifically, cell range expansion along with inter-cell interference coordination. The model explicitly considers mutual co-channel interference and enables predicting the performance of networks consisting of hundreds of cells with very low computational effort. Based on this model, we present a heuristic centralized data offloading algorithm, which, for a certain traffic demand, is able to increase the 5th percentile of the data flow throughput by a factor of 4.5 and to halve the probability of service unavailability. Moreover, we show that the network capacity can be increased by about 41.3 % if data offloading is performed.

***Index Terms*—wireless network optimization; data offloading; inter-cell interference coordination; cell range expansion; flow level modeling; queuing theory**

I. Problems due to Unequal Load Balances in Heterogeneous Networks

Mobile traffic demand is inevitably increasing strongly [1]. Moreover, users often gather together at some specific places, especially in urban environments such as city centers. Then they form so-called hot spots and hot zones, where the majority of mobile data traffic is generated.

A cost- and energy-efficient way to serve such hot spot users is to deploy small cells. Although, they provide large capacities in some limited areas [2], more often than not, they do not cover all hot spot users. Thus they leave a lot of traffic load for the overlaying macro cells, which causes overload situations, congestion, or low throughputs for macro cell users. Further, due to the higher utilization of the overlaying macro cells, a lot of inter-cell interference is generated within the small cells leading to worse performance experienced by the users therein. The major reason for such load imbalances is that the locations and sizes of, and the traffic demand generated in hot spots or hot zones often change during the course of a day. This fact calls for dynamic, automated network management solutions that adapt network parameters to varying traffic conditions.

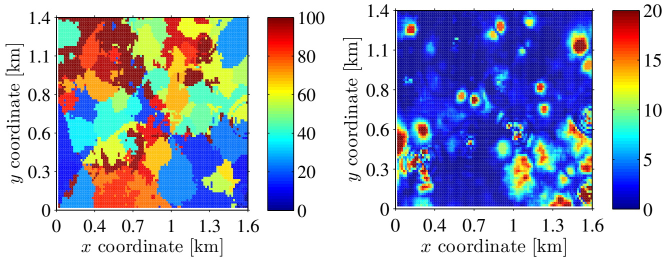
A. Example Network Evaluation

Throughout the paper, we illustrate our findings on the basis of the realistic model of a real deployed network in a dense-urban North American city and the model assumptions presented in Section II. The total area is about 2.24 km^2. The network consists of 29 macro cells (with transmit power of 46 dBm) and 19 micro cells (38 dBm) deployed at traffic hot spots. All base stations employ the same frequency band of width 10 MHz. Receive powers are affected by the path loss computed with the COST-Hata model at a carrier frequency of 2.6 GHz and by slow fading with clutter-dependent standard deviations ranging from 1 dB to 9 dB. Admission control mechanisms ensure that each base station admits data transfers of at most ten concurrent users. The traffic demand varies strongly in space with differences up to a factor of 10^5.

Fig. 1 illustrates the network for a snapshot, where the mean traffic demand is 90 Mbps/km^2. Many macro cells are highly loaded, which is indicated by the red shaded cell areas in Fig. 1(a); however, there are also many cells (especially small cells) that are lowly loaded. Such load imbalances usually bring along large differences in data throughputs achievable in different cells, since the base station utilizations mainly account for the amount of radio resources that can be allocated to the users (see Fig. 1(b)). As a consequence, there exist a large number of users that experience very poor service. More specifically, the worst 5 % of the users have a maximum throughput of only 0.28 Mbps and around 7.3 % of data requests are not admitted. In the following, we will show how large heterogeneous networks can be modeled accurately and how such a model can be used to tackle the problems mentioned above efficiently and effectively.

B. Data Offloading through CRE and ICIC

One promising method to counter the problem of load imbalances in heterogeneous networks and to improve their performance is dynamic data offloading, meaning Cell Range Expansion (CRE) along with Inter-Cell Interference Coordination (ICIC). Thereby, users in overloaded or congested macro cells are actively forced to connect to smaller base stations; however, they have to be protected from strong inter-cell interference generated by the macros. Mechanisms to protect resources, that is, ICIC, can either be performed by muting frequency resources (Fractional Frequency Reuse, FFR) [3] or time slots (Almost Blank Sub-frames, ABS) [4].



(a)

(b)

Fig. 1. (a) Base station utilizations in % (corresponding cell areas are color-coded) and (b) data flow throughput in Mbps for a mean traffic demand of 90 Mbps/km^2 and without offloading.

The contribution of the paper is threefold: 1) We extend a queing-theoretic model based on the notion of *elastic data flows* in order to incorporate ICIC and CRE algorithms. The model enables evaluating data offloading on a large scale, meaning in large networks with many macro base stations and many small cells, with very low computational effort. Thus, sophisticated Monte-Carlo simulations can be avoided. It is important to note that the model considers dynamic inter-cell interference explicitly (induced by the dynamic nature of the arrival and departure of data flows), which constitutes an extension of the work in [5]. 2) We present a simple heuristic centralized data offloading algorithm, which is able to improve network performance with respect to user data throughputs and service availability. 3) Based on the example network described above, we illustrate that the algorithm yields considerable gains in network capacity, which is defined as the maximum traffic demand satisfied while fulfilling certain quality of experience constraints.

II. Modeling Data Offloading in HetNets

In the following, we propose a simple yet flexible and scalable model, which considers the dynamic behavior of user data requests and, what comes along with it, dynamic mutual inter-cell interference among cells.

A. Wireless Network Model

In the remainder, we use the terms *macro cell* and *small cell* as the coverage areas of a macro base station and a low power node, respectively. Further, we assume that macro base stations and low power nodes serve their users in the same frequency band and, thereby, generate mutual co-channel interference.

1) Network Layout: We consider the downlink of a heterogeneous network consisting of N_M macro cells and N_S small cells covering an area $\mathcal{L} \subset \mathbb{R}^2$. We collect the indices of macro base stations and low power nodes in the sets $\mathcal{B}_\mathrm{M} := \{1, \ldots, N_\mathrm{M}\}$ and $\mathcal{B}_\mathrm{S} := \{N_\mathrm{M}+1, \ldots, N_\mathrm{M}+N_\mathrm{S}\}$, respectively. We denote each of the macro cell areas as $\mathcal{L}_i \subset \mathcal{L}$ with $i \in \mathcal{B}_\mathrm{M}$. In contrast to macro cells, we divide each small cell area into two distinct regions, a center and an edge region. Hereafter, the edge region models the area, in which users originally were served by a macro base station but are offloaded to a low power node. We define the small cell center and edge regions by $\mathcal{L}_j$ and $\mathcal{L}_{j,\mathrm{E}}$ with $j \in \mathcal{B}_\mathrm{S}$, respectively.

A user at location $u \in \mathcal{L}$ can either be located within a macro cell or a small cell (either within the center or the edge region). The cell areas then form a partition $\mathcal{P}$ over $\mathcal{L}$, that is, $\mathcal{P}(\mathcal{L}) = \{\mathcal{L}_i, \mathcal{L}_j, \mathcal{L}_{j,\mathrm{E}}\}|_{i \in \mathcal{B}_\mathrm{M}, j \in \mathcal{B}_\mathrm{S}}$. We define the user association rule that assigns a user at each location $u \in \mathcal{L}$ to a physical base station $i \in \mathcal{B}_\mathrm{M} \cup \mathcal{B}_\mathrm{S}$ by the bijective function $g : \mathcal{L} \to \mathcal{B}_\mathrm{M} \cup \mathcal{B}_\mathrm{S}$.

2) Mobile Traffic and Radio Link Model: We model mobile traffic based on the notion of elastic data flows and sessions [6], where each session is composed of a number of subsequent data flows that are initiated by the mobile user. This notion has become a common way to describe flow dynamics and network performance (that users experience) derived from them, see e. g. [7], [8]. It also allows to implicitly consider mutually coupled cells through dynamic inter-cell interference and, thereby, to estimate the performance of real networks quite accurately [9].

One reasonable, and helpful, assumption that stems from the traffic model is that flow dynamics happen on a time-scale (seconds) that is significantly longer than fast fading effects (milliseconds) and considerably shorter than the impact of slow fading (several tens of seconds). Consequently, both effects can be included in the location-dependent but otherwise constant receive powers $p_i(u)$, that is, the signal strength that a user at location u receives from base station $i \in \mathcal{B}_\mathrm{M} \cup \mathcal{B}_\mathrm{S}$. Further, we incorporate the path loss, antenna patterns, user equipment and base station noise figures, and penetration losses in the functions $p_i(\cdot)$.

3) Incorporating Cell Range Expansion: The decision, to which base station a user connects to, is usually driven by rather simple cell selection and re-selection processes [10], which are based on maximum receive signal strengths. For data offloading purposes it is possible to make connecting to a low power node more attractive to a macro cell user by adding a bias, or *cell individual offset* CIO_j, to the physical signal strength yielding a *virtual* receive power $p_{j,\mathrm{virt}}$, that is,

$$p_{j,\mathrm{virt}}(u) := \begin{cases} p_j(u) & \text{if } j \in \mathcal{B}_\mathrm{M}, \\ p_j(u) \cdot \mathrm{CIO}_j & \text{if } j \in \mathcal{B}_\mathrm{S}. \end{cases} \tag{1}$$

A user at location $u \in \mathcal{L}$ connects to a base station $g(u)$ if that base station provides the highest virtual receive power. Hence, the user association rule $g(u)$ is formulated as

$$g(u) := \underset{j \in \mathcal{B}_\mathrm{M} \cup \mathcal{B}_\mathrm{S}}{\operatorname{argmax}} \; p_{j,\mathrm{virt}}(u). \tag{2}$$

A user, which is connected to a low power node $i \in \mathcal{B}_\mathrm{S}$, is an offloaded user, if it was connected to a macro base station in case there was no bias ($\mathrm{CIO}_j = 1$), otherwise it is a regular small cell user. Provided that $g(u) = i \in \mathcal{B}_\mathrm{S}$, the fact that such a user is an offloaded user (or not) can be written as

$$u \in \begin{cases} \mathcal{L}_{i,\mathrm{E}} & \text{if } \exists j \in \mathcal{B}_\mathrm{M} : p_j(u) > p_i(u), \\ \mathcal{L}_i & \text{otherwise.} \end{cases} \tag{3}$$

Usually, a meaningful bias is computed by the base station and reported to the user’s hand set. Within one cell, each hand set receives the same bias from its base station. The resulting virtual receive power is computed by the hand set, based on which it initiates the cell (re-)selection process.

4) Incorporating Inter-Cell Interference Coordination: Instead of assuming some fixed ICIC scheme, such as Fractional Frequency Reuse [3] or Almost Blank Sub-frames [4], which coordinate the scheduling of resources either in the frequency or time domain, respectively, we model ICIC based on a general blanking of radio resources. Blanked radio resources may correspond to frequency or time slots, or to a mixture of both. We assume a fraction $m \in [0, 1)$ of resources not to be accessible by users that are located in macro cells or small cell centers. Hence, these resources are exclusively reserved for small cell edge users. We call these resources *muted* resources. Note that the definitions above imply that the scheduling decision (meaning whether a user is allocated muted/protected resources or not) is based on the physical and virtual receive powers.

In order to model the low power nodes with two different sets of resources (each of which are either designated to center or edge users), we introduce another set of base station indices $\mathcal{B}_\text{E} := \{N_\text{M} + N_\text{S} + 1, \dots, N_\text{M} + 2N_\text{S}\}$. We interpret the set $\mathcal{B}_\text{S}$ as the collection of *physical* low power node indices and the set $\mathcal{B}_\text{E}$ as the collection of *virtual* low power node indices. Each of the virtual low power nodes uniquely relates to a physical base station. Formally, we write this relation as a bijective function $v : \mathcal{B}_\text{S} \to \mathcal{B}_\text{E}$ and we assume $p_i(u) = p_{v(i)}(u)$, meaning that the receive power per resource experienced at location u is equal for the two radio resource sets provided by a low power node. For ease of notation, we also use $\mathcal{L}_{v(i)} := \mathcal{L}_{i,\text{E}}$ for all $i \in \mathcal{B}_\text{S}$ and write $\mathcal{B} := \mathcal{B}_\text{M} \cup \mathcal{B}_\text{S} \cup \mathcal{B}_\text{E}$.

We model the signal-to-interference and noise ratio (SINR) based on the *average interference* assumption initially proposed in [11]. According to this assumption, the transmission of data to a user is affected by time-average interference conditions that are based on the average resource utilization of the base stations and, hence, on the traffic conditions and the users' behavior. Let η be a vector of length $N_\text{M} + 2N_\text{S}$, the elements $\eta_i \in [0, 1)$ of which describe the individual average utilizations of reserved resources of all physical and virtual base stations. To be more elaborate, an element η_i with $i \in \mathcal{B}_\text{M} \cup \mathcal{B}_\text{S}$ models the utilization of the fraction $1 - m$ of the total resources of base station i. If $i \in \mathcal{B}_\text{E}$, the element η_i models the utilization of the fraction m of the total resources of the virtual base station i. The SINR at location u with respect to base station i is

$$\gamma_i(u,\eta) := \begin{cases} p_i(u)\left(\sum\limits_{j\in\mathcal{B}_\text{M}\cup\mathcal{B}_\text{S}\setminus\{i\}} \eta_j p_j(u) + N_0\right)^{-1} & \text{for } i \in \mathcal{B}_\text{M} \cup \mathcal{B}_\text{S}, \\ p_i(u)\left(\sum\limits_{j\in\mathcal{B}_\text{E}\setminus\{i\}} \eta_j p_j(u) + N_0\right)^{-1} & \text{for } i \in \mathcal{B}_\text{E}, \end{cases} \tag{4}$$

with N_0 denoting the noise power. The explanation for the definition of the SINR $\gamma_i(u, \eta)$ is as follows: If a user is connected to a macro base station or to a low power node while located in the small cell center region, it will experience inter-cell interference from all other physical macro base stations and low power nodes. If the user is an offloaded user, that is, it is connected to a virtual low power node or equivalently located in a small cell edge region, it only receives interference from other virtual low power nodes. According to the definition of an offloaded user in Eq. (3), there will be a macro base station providing a larger receive power. However, since the edge user transmits its data by utilizing the share m of resources not accessible by macro base stations and physical low power nodes, it is protected from their inter-cell interference. This is also justified by the fact that hand sets using modern radio access technologies, such as Long Term Evolution (Rel. 10/11), are capable of eliminating reference signals or pilots by methods such as Common Reference Signal-Interference Cancellation [12].

With $\hat{c}_i(u, \eta) := aB \min\{\log\left(1 + b\gamma_i(u, \eta)\right), c_\text{max}\}$ being the maximum data rate achievable at location u if all resources were available, the effective rate incorporating muting resources by ICIC amounts to

$$c_i(u,\eta,m) = \begin{cases} (1-m)\cdot \hat{c}_i(u,\eta) & \text{for } i \in \mathcal{B}_\text{M} \cup \mathcal{B}_\text{S}, \\ m \cdot \hat{c}_i(u,\eta) & \text{for } i \in \mathcal{B}_\text{E}. \end{cases} \tag{5}$$

The quantities B, a, b, and c_max denote the total system bandwidth, the bandwidth and SINR efficiencies according to [13], and the maximum data rate achievable by the highest modulation and coding scheme, respectively.

B. Deriving Network Key Performance Indicators

In order to compute network performance indicators, we begin with deriving the capacity $C_i(\eta, m)$ of a base station or cell $i \in \mathcal{B}$ in Mbps, from which all relevant performance metrics will be derived. It is important to note that the metrics are with regard to all types of base stations introduced, macro base stations, physical as well as virtual low power nodes. The cell capacity can be defined as the weighted harmonic mean of the individual location-dependent rates $c_i(u, \eta, m)$ [7],

$$C_i(\eta,m) := \left(\int_{\mathcal{L}_i} \delta_i(u)\left(c_i(u,\eta,m)\right)^{-1} \mathrm{d}u\right)^{-1}. \tag{6}$$

A very important fact is that the weighting factors $\delta_i(u)$ with $\int_{\mathcal{L}_i} \delta_i(u)\,\mathrm{d}u = 1$ account for the heterogeneous user distribution within cell $\mathcal{L}_i$, such that the capacity $C_i(\eta, m)$ models the average data rate, a base station i provides to the users.

The definition of the cell capacity in Eq. (6) along with the definitions of the SINRs and achievable rates in Eqs. (4)-(5), contain two very crucial aspects which are the following: 1) The capacities of macro base stations and physical low power nodes ($\mathcal{B}_\text{M} \cup \mathcal{B}_\text{S}$) are mutually coupled through inter-cell interference, and the same is true among virtual low power nodes ($\mathcal{B}_\text{E}$). However, there is no interference-coupling among any two base stations, each of which belongs to a different set ($\mathcal{B}_\text{M} \cup \mathcal{B}_\text{S}$ or $\mathcal{B}_\text{E}$). In general, cell capacities decrease if the utilization of any interfering base station increases.

2) Nevertheless, the splitting of resources among physical base stations (macro base stations and low power nodes) and virtual low power nodes by the parameter m causes a coupling of cell capacities between both sets.

Following the modeling approach in [14], we assume that

1) Data flows arrive at any base station i according to a Poisson process with rate λ_i in s^{-1},[1]
2) Data flows have an exponential file size distribution with finite mean Ω in Mbit,
3) At most L_i flows are served concurrently in cell i and all further requests are blocked by admission control mechanisms [14], and
4) Radio resources are shared equally among these concurrently active flows according to the Egalitarian Processor Sharing (EPS) service discipline, which is the limiting case of the Round Robin scheduler.

The assumptions above allow each base station to be modeled as an M/M/1/L_i EPS queuing system with load

$$\rho_i(\eta) = \frac{\lambda_i \Omega}{C_i(\eta, m)}. \tag{7}$$

According to the definition of an M/M/1/L_i EPS queuing system, we can compute the average utilization η_i of base station i by

$$\eta_i = \rho_i(\eta)\left(1 - P_i(\eta)\right) =: f_i(\eta) \tag{8}$$

with the average probability P_i of blocking data flows,

$$P_i(\eta) = \frac{\left(1 - \rho_i(\eta)\right)\rho_i(\eta)^{L_i}}{1 - \rho_i(\eta)^{L_i+1}}. \tag{9}$$

Note that the resource utilization η_i of base station i in Eq. (8) is given in an implicit form. According to Theorem 1 in [14], the system of equations $\eta = f(\eta)$, with vector function $f(\cdot) := \left(f_1(\cdot), \ldots, f_{N_\mathrm{M}+2N_\mathrm{S}}(\cdot)\right)$ can be solved numerically with very low effort by a fixed point iteration.

For network evaluation, we use the average (with respect to the interference) data flow throughputs r_i that users experience,

$$r_i(u, \eta, m) = \frac{\eta_i c_i(u, \eta, m)}{n_i(\eta)}, \tag{10}$$

and the 5th percentile R_5 derived from them. The quantity n_i is the average number of active flows in cell i and given by

$$n_i(\eta) = \frac{\rho_i(\eta)}{1 - \rho_i(\eta)} - \frac{(L_i + 1)\,\rho_i(\eta)^{L_i+1}}{1 - \rho_i(\eta)^{L_i+1}}. \tag{11}$$

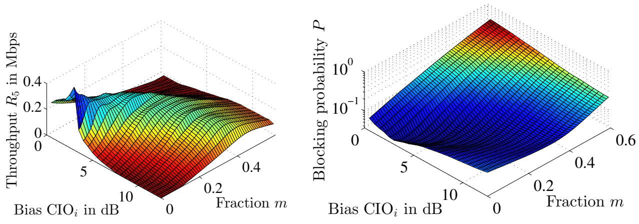
Further, we use the probability P_b that any data request is blocked in the entire network,

$$P(\eta) := \frac{\sum_{i\in\mathcal{B}} \lambda_i P_i(\eta)}{\sum_{i\in\mathcal{B}} P_i(\eta)}. \tag{12}$$

For further information about the *average interference* model, we refer to [14] and the work referenced therein.

C. Some Insights from Numerical Experiments

In order to illustrate some findings by means of the example scenario from Section I, we vary the biases CIO_i of all low power nodes and the fraction m of muted resources. We observe the flow throughput 5 %-ile R_5 and the network flow blocking probability P. Fig. 2 depicts both quantities for a mean traffic density of 90 Mbps/km^2. As can be seen, there is a relatively strong correlation between the blocking probability and the throughput. If the parameters are well adjusted, cell capacities increase and transmission resources can be utilized more efficiently, which increases the flow throughput. As a consequence, fewer flows are concurrently active because the data transmissions can be finished more quickly, which reduces the chance of incoming data transfers to be blocked.



(a) (b)

Fig. 2. (a) Flow throughput 5 %-iles R_5 and (b) overall network flow blocking probabilities P for varying biases CIO_i (the same applied to all low power nodes) and fractions m of muted resources. The average traffic demand is 90 Mbps/km^2.

In that scenario, data offloading is beneficial in the sense that choosing a bias of $\mathrm{CIO}_i = 2.5\,\mathrm{dB}$ yields an increased flow throughput and a reduced blocking probability. However, stronger data offloading (by choosing a higher bias), e. g., to reduce the energy consumption of the large macro base stations, is only recommended, if a certain fraction $m > 0$ of radio resources is reserved for the small cell edge users, such that they do not suffer from the strong inter-cell interference from the macro base stations.

Given a certain bias CIO_i, say 6 dB, varying the fraction m has the following effect: Increasing m beyond 0.3 would reserve more resources for the protected small cell edge users. Although these users would benefit, the users served by the macro base stations and physical low power nodes suffer since fewer resources are left over, which worsens their throughput and blocking probability, likewise. The opposite argument is true for decreasing the bias below 0.3. Given a certain fraction m of resources reserved for the protected users, say 0.3, varying the bias would have the following effect: Decreasing the bias below 2 dB (and finally yielding 0 dB) would make reserved resources useless and there are no users that benefit from it. Eventually, the network performance decreases with respect to both, the throughput and the blocking probability. Increasing the bias beyond 6 dB still yields acceptable results; however, it can be seen that an increase of the bias along with an increase of the fraction m of muted resources is more beneficial. The behavior of the 5 %-ile user throughput has been studied in [4] and is in accordance with our results.

[1] Note that λ_i models the temporal arrival of data flows and that the impact of the spatial inhomogeneity of the user distribution is considered in the cell capacity C_i already.

It can be expected that low power node-individual biases would yield higher flow throughputs and lower blocking probabilities than shown in this section. Therefore, we propose a centralized data offloading algorithm incorporating the optimization of the bias of each low power node individually.

III. A Centralized Data Offloading Algorithm

For the data offloading algorithm presented below, we assume that information about the hand sets' receive powers $p_i(u)$ as well as about the traffic conditions $\delta_i(u)$, Ω, and λ_i is available at a central network entity. For example, this can be accomplished by so-called call traces [15], where hand sets report their receive powers, signal delays, and other information to base stations. Amongst others, the position of the user can be determined from this data establishing concrete values for λ_i and $\delta_i(u)$.

A. Cost Function and Algorithm Description

In order to increase the network performance by means of the flow throughput and blocking probability, we introduce a cost function $\mathcal{C}$ based on these two metrics, i. e.,

$$\mathcal{C}(P, R_5; d) = d \cdot e^{400 \cdot P - 10} + (1-d) \cdot e^{30 - 20 \cdot R_5/\text{Mbps}} \quad (13)$$

The cost function $\mathcal{C}$ is chosen such that network blocking probabilities P larger than 5 % and flow throughput 5 %-iles (or cell edge throughputs) smaller than 1 Mbps/km^2 are unfavorable and cause large costs (denoted as quality of experience (QoE) constraints in the remainder). Furthermore, the cost function is parametrized by a quantity $d \in [0, 1]$, by which the focus can be shifted either towards blocking probability (by increasing d) or throughput percentile (by decreasing d). The goal of the algorithm described subsequently is to minimize the cost $\mathcal{C}$. The algorithm carries out a heuristic search for an appropriate global fraction m of protected resources and appropriate low power node-individual biases CIO$_i$. The procedure is the following: We sweep over possible values for the fraction m in steps of 1 %. For each value, we sweep over all physical low power nodes (three times) and vary the bias CIO$_i$ of each low power node in a range of 3 dB around the initial value (compare taxi cab method as a special case of Powell's method [16]). The bias and the fraction m are chosen, for which the cost $\mathcal{C}$ is minimized.

We allow a maximum number of ten flows to be served concurrently by each base station, such that the corresponding ten available slots have to be split among small cell edge and center users. In general, we assume that $L_i = 7$ slots are available for small cell center users and $L_{v(i)} = 3$ slots are available for offloaded users, if cell range expansion is performed by the low power node $i \in \mathcal{B}_\text{S}$. Without offloading, the number of slots provided by the low power node to the small cell users is ten, the same number as for macro base stations. In order to characterize the impact of a dynamic allocation of the available slots in the case of data offloading, we also compare the performance if, in addition, the split of slots was also optimized according to the cost $\mathcal{C}$, see AAC (*adaptive admission control*) method below. We apply four different flavors of the algorithm to the network described in Section I,

1) $d = 1$: This optimizes the service availability solely by reducing the flow blocking probability.
2) $d = 0$: This optimizes the flow throughput solely.
3) $d = 0.5$: This optimizes both, the flow throughput and the service availability.
4) $d = 0.5$+AAC: In addition to adjusting the parameters m and CIO$_i$, the allocation of available slots among concurrently active small cell center and edge users is optimized, meaning the split of 10 slots in total into L_i and $L_{v(i)}$. The improvement of both, the flow throughput and the service availability, is considered.

The algorithms are detailed in pseudo-code in the Appendix.

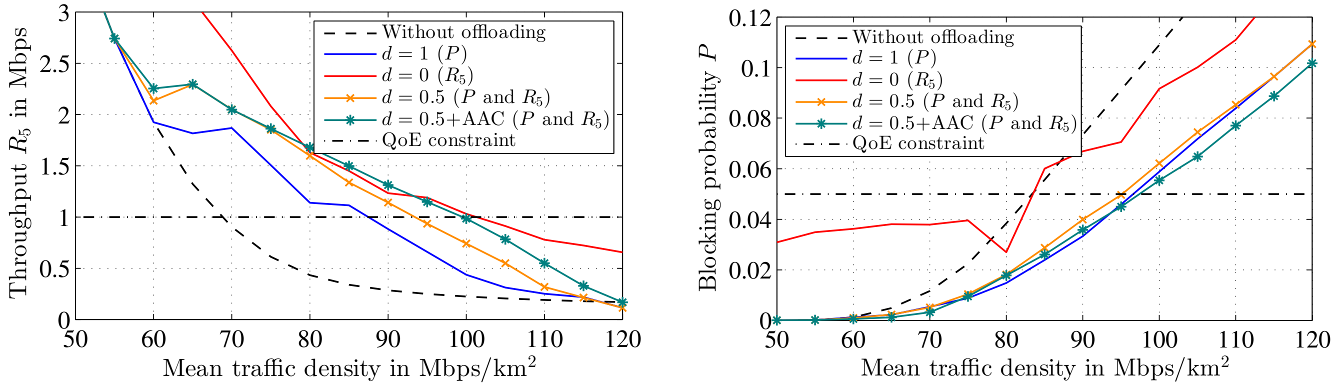
B. Algorithm Performance

Fig. 3 depicts the flow throughput percentiles R_5 and the network blocking probabilities P for various configurations of the cost function $\mathcal{C}$ and for increasing mean traffic density.

Initial network performance (dashed lines) is rather poor; however, it can be improved by data offloading, that is, by increasing the coverage area of smalls cells and protecting resources for small cell edge users. For instance, for a mean traffic demand of 90 Mbps/km^2, the throughput percentile can be increased by a factor of around 4.5 for $d = 0$ and for $d = 0.5$ (with AAC). The blocking probability can be reduced from 7.3 % down to 3.3 % for $d = 1$. The rather large difference between the curves for $d = 0$ and $d = 1$ (for optimizing either the throughput or the blocking probability solely) suggests that there might be a trade-off in optimizing both metrics. However, we can see that the joint optimization ($d = 0.5$) leads to almost the same performance with regard to both, the improved throughput ($d = 0$) and blocking probability ($d = 1$), at the same time. Interestingly, there is only a slight improvement due to the additionally optimized allocation of free slots among small cell center and small cell edge data flows, meaning by the adaptive admission control scheme AAC. Furthermore, improving the blocking probability appears to be tougher than the data flow throughput, which can be seen by comparing the gains in both metrics in the very high traffic regime.

Fig. 4(a) illustrates the fraction m of muted resources chosen by the algorithms. All algorithms (except for $d = 0$ - improving throughput solely) have in common that there is a sudden increase in protected resources for a traffic demand, where data offloading becomes beneficial. After the peak fraction of muted resources of around 20 %, the amount slowly declines. This indicates that the gain through protecting resources and, hence, improving the SINRs and throughputs of small cell edge users, is being compensated by a loss caused by the reduction of the amount of resources provided by macro base stations and physical low power nodes.

Interestingly, the algorithm that improves the flow throughput solely ($d = 0$), offloads traffic to the small cells reserving only very few resources for the offloaded users. Thereby, it considerably improves the throughput of small cell center and



(a) (b)

Fig. 3. (a) Throughput 5 %-iles R_5 and (b) network blocking probabilities P for differently configured cost functions $\mathcal{C}$ and for increasing traffic demand.

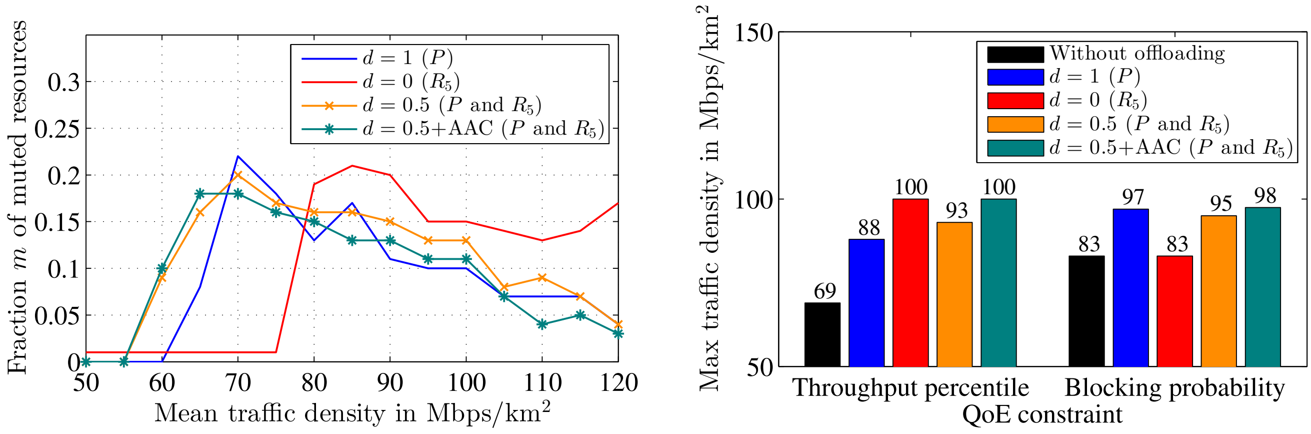
macro cell users at the cost of high blocking probabilities for the data transfers of the small cell center users (due to very few available resources). However, for higher traffic demand, the larger number of offloaded users cannot be neglected anymore such that resources have to be reserved for them.

Taking both QoE constraints, the blocking probability P being smaller than 5 % and the throughput percentile R_5 being greater than 1 Mbps, into account, the network capacity can be defined as the maximum traffic demand in Mbps/km^2, which can be served by the network while fulfilling either one of the constraints or both of them at the same time. Fig. 4(b) depicts the maximum traffic densities, for which either of the two constraints are fulfilled. All flavors of the algorithm are able to increase the network capacity. As expected, optimizing the flow throughput yields lower improvements of the capacity with respect to the flow blocking probability constraint and vice versa. Taking the maximum traffic demand of approx. 69 Mbps/km^2, for which both constraints are fulfilled simultaneously for the initial network configuration, as a baseline, the algorithms achieve capacity improvements of 27.5 % (for $d = 1$), 20.3 % ($d = 0$), 34.8 % ($d = 0.5$), and 41.3 % ($d{=}0.5$+AAC).

Finally, Fig. 5 depicts the base station utilizations and flow throughputs in the network with offloading ($d = 0.5$+AAC) for the same traffic demand as in Section I. They indicate a much more balanced distribution of the network load among all base stations. As a consequence, service availability has been increased (not shown here) and there is a much more equal distribution of provided data throughputs across the users. As a result, radio resources can be allocated more fairly to the individual users and there are less users suffering poor quality of service.

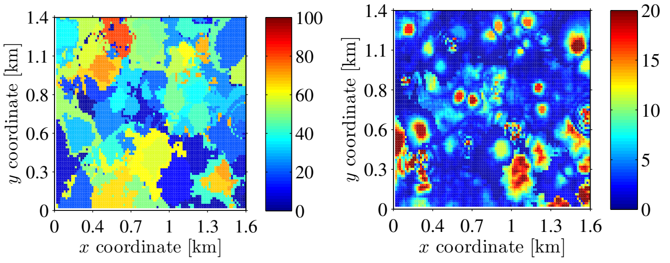
IV. Conclusions and Outlook

The cellular flow level model, which has been extended to capture inter-cell interference and cell range expansion, may be used in centralized entities that trigger or manage self-organization use cases to improve the network performance. Specifically, considerable network capacity gains are achievable through dynamic data offloading compared with classical heterogeneous network operation. Although, there is a strong correlation between increasing data flow throughputs and service availability, both metrics should be considered jointly to achieve the best quality of experience.



(a) (b)

Fig. 4. (a) Fraction m of muted resources for various configurations of the cost function $\mathcal{C}$ and (b) maximum traffic demand, which can be served by the network while QoE constraints are fulfilled.



(a) (b)

Fig. 5. (a) Base station utilizations in % (corresponding cell areas are color-coded) and (b) data flow throughput in Mbps for a mean traffic demand of 90 Mbps/km^2 and with data offloading ($d = 0.5$+AAC).

For future work, there may be two extensions to further improve the model’s accuracy and the algorithms’ achievements:

1) A more advanced model for opportunistic scheduling could be incorporated, especially with regard to choosing the resources that have to be protected, as e. g. in [17]. For instance, it may be beneficial to mute (protect) radio resources, where macro base stations interfere more strongly due to slow and fast fading effects.
2) The fraction of muted resources, either through Almost Blank Sub-frames, Fractional Frequency Reuse, or a combination of both, may be adjusted per macro base station individually. Though, this would make the network model more complex, the procedure may be straightforward and may produce promising gains in network capacity.

Note that, in order to guarantee that the model is still applicable, these enhancements have to be carried out considering the fixed point problem $\eta = f(\eta)$.

Acknowledgment

The work presented in this paper was sponsored by the government of the Free State of Saxony, Germany, and by the European Regional Development Fund (ERDF) within the *Cool Silicon Cluster of Excellence* under contract 31529/2794.

Appendix

Algorithm 1 Offloading Algorithm

```
Input: $d$, $k := 0$, $\forall i \in \mathcal{B} : p_i(u)$, $\lambda_i \Omega \delta_i(u)$, $\forall i \in \mathcal{B}_\mathrm{S} : \mathrm{CIO}_i$
for all $m \in \{0, 0.01, \ldots, 0.99\}$ do
    while $(k < 3)$ do
        for all $i \in \mathcal{B}_\mathrm{S}$ do
            for all $\mathrm{CIO}' \in \{\mathrm{CIO}_i - 3\ \mathrm{dB}, \ldots, \mathrm{CIO}_i + 3\ \mathrm{dB}\}$ do
                $\mathrm{CIO}_i := \mathrm{CIO}'$
                Call: Algorithm 2
                solve $\eta = f(\eta)$, compute $P(\eta), R_5(\eta), \mathcal{C}(P, R_5; d)$
            end for
            choose and apply $m$ and $\mathrm{CIO}_i$ that minimize $\mathcal{C}$
        end for
        $k := k + 1$
    end while
end for
return $m$, $\forall i \in \mathcal{B}_\mathrm{S} \cup \mathcal{B}_\mathrm{E} : L_i$, $\forall i \in \mathcal{B}_\mathrm{S} : \mathrm{CIO}_i$
```

Algorithm 2 AAC Algorithm

```
if $\mathrm{CIO}_i > 0$ dB then
    if AAC active then
        for all $L' \in \{1, \ldots, 9\}$ do
            $L_i := L'$, $L_{v(i)} = 10 - L'$
            solve $\eta = f(\eta)$, compute $P(\eta), R_5(\eta)$, and $\mathcal{C}(P, R_5; d)$
        end for
        choose and apply $L_i$ and $L_{v(i)}$ that minimize $\mathcal{C}$
    else
        $L_i := 7$, $L_{v(i)} = 3$
    end if
else
    $L_i := 10$
end if
return $L_i$, $L_{v(i)}$
```

References

[1] Cisco, “Cisco Visual Networking Index: Global Mobile Data Taffic Forecast Update 2012-2017,” Tech. Rep., 2012.

[2] A. Fehske, F. Richter, and G. Fettweis, “Energy efficiency improvements through micro sites in cellular mobile radio networks,” in *GLOBECOM Workshops, 2009 IEEE*, Nov 2009, pp. 1–5.

[3] N. Himayat, S. Talwar, A. Rao, and R. Soni, “Interference management for 4g cellular standards [wimax/lte update],” *Communications Magazine, IEEE*, vol. 48, no. 8, pp. 86–92, August 2010.

[4] Y. Wang and K. Pedersen, “Performance analysis of enhanced inter-cell interference coordination in lte-advanced heterogeneous networks,” in *Vehicular Technology Conference (VTC Spring), 2012 IEEE 75th*, May 2012, pp. 1–5.

[5] S. Akbarzadeh, R. Combes, and Z. Altman, “Self-organizing femtocell offloading at the flow level,” *International Journal of Network Management*, vol. 23, no. 4, pp. 259–271, 2013.

[6] J. W. Roberts, “Traffic Theory and the Internet,” *IEEE Communications Magazine*, vol. 39, no. 1, pp. 94–99, 2001.

[7] T. Bonald, “Flow-level performance analysis of some opportunistic scheduling algorithms,” *European Transactions on Telecommunications*, vol. 16, no. 1, pp. 65–75, Jan. 2005.

[8] A. J. Fehske and G. P. Fettweis, “On Flow Level Modeling of Multi-Cell Wireless Networks,” in *WiOpt*, Tsukuba City, 2013.

[9] A. Fehske, H. Klessig, J. Voigt, and G. Fettweis, “Flow-level models for capacity planning and management in interference-coupled wireless data networks,” *Communications Magazine, IEEE*, vol. 52, no. 2, pp. 164–171, February 2014.

[10] 3GPP, “3GPP TS 36.304 User Equipment (UE) procedures in idle mode and procedures for cell reselection in connected mode,” *www.3gpp.org*.

[11] K. Majewski and M. Koonert, “Conservative Cell Load Approximation for Radio Networks with Shannon Channels and its Application to LTE Network Planning,” in *2010 Sixth Advanced International Conference on Telecommunications*. IEEE, 2010, pp. 219–225.

[12] B. Soret, Y. Wang, and K. I. Pedersen, “Crs interference cancellation in heterogeneous networks for lte-advanced downlink,” in *Communications (ICC), 2012 IEEE International Conference on*. IEEE, 2012, pp. 6797–6801.

[13] P. Mogensen, W. Na, I. Z. Kovacs, F. Frederiksen, A. Pokhariyal, K. I. Pedersen, T. Kolding, K. Hugl, and M. Kuusela, “LTE Capacity Compared to the Shannon Bound,” in *2007 IEEE 65th Vehicular Technology Conference - VTC2007-Spring*, no. 1. IEEE, Apr. 2007, pp. 1234–1238.

[14] H. Klessig, A. Fehske, and G. Fettweis, “Admission control in interference-coupled wireless data networks: A queuing theory-based network model,” in *12th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (Wi-Opt ’14)*, Hammamet, Tunesia, May 2014, accepted for publication.

[15] 3GPP, “3GPP TS 32.421 Subscriber and equipment trace; Trace concepts and requirements,” *www.3gpp.org*.

[16] M. J. D. Powell, “An efficient method for finding the minimum of a function of several variables without calculating derivatives,” *The Computer Journal*, no. 7(2), pp. 155–162, 1964.

[17] F. Tesema, P. Zanier, I. Viering, A. Fehske, and G. Fettweis, “Simplified scheduler model for son algorithms of eicic in heterogeneous networks,” in *European wireless conference, 2014 IEEE International Conference on*, May 2014, accepted for publication.