An Analytical Model for Flow-level Performance in Heterogeneous Wireless Networks

George Arvanitakis, Thrasyvoulos Spyropoulos and Florian Kaltenberger
EURECOM, 06410, Biot, France, firstname.lastname@eurecom.fr

***Abstract*—Modern cellular networks are becoming denser, less regularly planned, and increasingly heterogeneous, making performance analysis challenging. We develop a flexible and accurate model of such heterogeneous networks (HetNets) consisting of K tiers of randomly located Base Stations (BSs), with different densities, transmit powers and Radio Access Technologies (RATs). Our main goal is to understand the impact of flow level dynamics on such a system, assuming non-saturated users that randomly generate download requests ("flows"). We do so by deriving analytically the per flow delay, the load, the utilization and the congestion probability of BSs in different tiers.**

We base our analysis on stochastic geometry, to understand the impact of topological randomness and intra- and inter-tier interaction, and queueing theory, to model the competition between concurrent flows within the same BS, for each RAT. This allows us to model the interference more realistically as a function of network load. We apply our model to the case of a 2-tier network based on LTE and WiFi and study different user inter-tier association criteria, such as Off-load, Max-SINR association, and Min-Delay association. Our results provide some interesting qualitative and quantitative insights about the impact of these association policies and different traffic intensities.

***Index Terms*—Stochastic Geometry; Queueing; HetNets; LTE; WiFi; User Association; Load-based Interference;**

I. Introduction

MOBILE data traffic has been increasing exponentially, and this trend is expected to continue for the foreseeable future [1]. To alleviate the overloaded macro-cell network, operators are additionally deploying small cells to capture traffic in hot spots. One promising scenario for such heterogeneous networks (HetNet) is the combination of LTE macro cells with WiFi small cells. Already today it is possible to integrate WiFi access points into the core network of cellular systems, and perform off-loading of traffic from LTE to WiFi. In future releases of 3GPP (release 13 and above) a tighter integration of WiFi and LTE technologies is foreseen that will also allow the aggregation of the two technologies.

HetNet architectures offer numerous advantages, but they also lead to denser, irregular, and more heterogeneous deployments, due to the often unplanned and incremental deployment of new (small cell) BSs [2], as well as the potentially different Radio Access Technologies (RAT). As a result, analyzing such networks, e.g., for protocol comparison or network planning, becomes increasingly challenging. What is more, the usually considered metrics in such analyses, like SINR or capacity, often fail to capture the actual user experience, because they do not take into account the heavy load of modern cellular networks [3], [4]. A better metric might be latency, as one of the key goals of 5G technologies is to minimize latency at the user or application level [5], [6].

To this end, we have developed a flexible and accurate model for the performance of future HetNets, in order to understand the impact of important network parameters. Our model consists of K orthogonal tiers of randomly located Base Stations (BSs), with different densities, transmit powers and Radio Access Technologies (RATs), as well as randomly placed users. Users are assumed to be non-saturated, randomly generating requests for new file/flow downloads of varying sizes, and they perceive performance in terms of the average delay to finish such a download. In other words, we are interested in the flow-level dynamics or flow-level performance of this system [4], [7]–[9]. Furthermore, BSs are modeled as queueing systems, that schedule concurrently arriving user flows according to the respective RAT scheduler, and network-related performance is measured in terms of the stationary load imposed on each BS, and the probability (or percentage) of BS being congested.

Our analysis is based on the combination of two key theoretical tools that have recently provided many insights on cellular network performance: (i) We use *queueing theory* to model the performance of dynamic flow arrival and service via the respective scheduler, at the level of a single BS; (ii) We utilize *stochastic geometry*, in order to understand the impact of topological randomness and interaction/competition between BSs at the network level, in order to derive statistics about the *number of users associated with a base station* at each tier, and the *modulation and coding schemes (MCS)* offered at each BS. Both these quantities serve as key inputs to the BS queueing model: the former to define the total traffic intensity (in terms of flow arrivals) a given BS has to serve, and the latter to define the average service rate (in terms of flow departures) that a BS is able to offer. We combine these two mathematical tools in order to provide an analytical framework that analyzes the flow level dynamics in large, random placed, multi-tier heterogeneous networks. Summarizing, our main contributions are:

(a) We derive the probability of coverage of a typical randomly located user in a randomly placed wireless network without assuming saturated BS or users.

(b) We propose an analytical model for LTE/WiFi HetNets, in a system where interference from nearby base stations is constant, as usually assumed in most related work. Our model captures both physical layer performance, providing statistics for coverage maps and MCS distributions, as well as flow-level performance as perceived by the user (mean flow delay) and the network operator (network's utilization, congestion probability).

(c) We extend this model to the challenging, yet more realis-

tic setup, where interference from nearby base stations depends on their load (e.g., a BS with no active transmissions will not interfere), and demonstrate that modeling this load-based interference has an important impact on both quantitative and qualitative conclusions.

(d) We compare our results to the current state-of-the-art performance frameworks [10], [11] and show that our model offers significantly higher accuracy in the medium and high load regimes (our scheme has up to 40 times better accuracy), while being at least as accurate for all loads and scenarios.

(e) We use our analytical framework to study the impact of popular user association policies like *Off-load* (all users within range of a WiFi AP are associated to the WiFi network), *Max-SINR* (a user is associated with the BS offering the best SINR, among any tier), and *Min-Delay* (a user is associated with the tier which offering the best combination of throughput and load in order to minimize the average delay [9]). Our results provide some interesting qualitative and quantitative insights.

The rest of the paper is organized as follows. In the next section, we provide a brief discussion of related work. Then, in Section III we present our model for performance at the BS level. In Section IV, we discuss our PHY Layer model, and in Section V, we derive the user cardinality distribution for our topology, which plays a key role in our model, and we compute the arrival rate. Section VI presents the analytical steps to specify the service rate, which includes both pure analytical formulas and technical details for each one of the chosen RAT. Section VII considers some scenarios of interest and applies our analytical results to obtain insights. Section VIII presents the future steps of our work.

II. Related Work

There are a number of seminal works employing stochastic geometry. The distribution of the coverage areas is studied in [12], and the distribution of the interference in [13]. The same framework has been widely used for studies of large and heterogeneous networks, because it not only avoids the problem of ideal and simplistic hexagonal or linear topologies but also it provides closed form expressions. As some such examples, [14] models the K-Tier downlink of heterogeneous cellular networks, [15] analyzes Carrier Aggregation in heterogeneous cellular networks, [16] tackles the problem of off-loading. Additionally, [17] models the downlink coverage probability in MIMO HetNets, [18] studies the problem of fractional frequency reuse for heterogeneous cellular networks, and [19] further considers the backhaul network. The main drawback of these works is the unrealistic assumptions of saturated users (i.e., not considering flow dynamics) and saturated BSs (i.e., assuming that all BSs interference at full power, all the time). Regarding the latter shortcoming two notable exceptions are [20] and [21], where the authors consider variable cell loads and load-aware interference models, in order to calculate the feasible network throughput w.r.t. cell load. However, these works also consider saturated users.

In a different research thread, there are a lot of works that have studied the flow-level dynamics in cellular systems. [8] and [22] use queuing models to take into account the random nature of traffic arrivals and departures, in order to obtain the flow-level performance of different schedulers. [4] and [7] apply such queueing results to model wireless network systems such as 3G/3G+ and derive expressions for the flow-level performance. However, these only consider simple cellular topologies (e.g., line networks, or small hexagonal topologies), and assume a known rate distribution and always ON interference. [23] attempts to take into account the performance dependency between nearby BSs, when one considers load-based interference, and propose a methodology to derive some performance bounds. Nevertheless, this work also considers simple topologies.

To the best of our knowledge, the following recent works are closest to ours, attempting to also combine more sophisticated topologies with flow-level performance. [24] provides an analytical framework that calculates the stability of a Poisson Bipolar network. However, in a Poisson Bipolar network, each BS has a dedicated receiver at a random distance, so it cannot be used to examine the impact of users with different rates, associated to the same BS. Additionally, this work does not consider other flow-level metrics (e.g. delay) besides stability. In [25] the authors assume homogeneous PPP topologies for both BS and users in order to capture uplink performance, considering flow-level traffic dynamics. However, the authors assume a saturated interference scenario, not capturing the interplay between load-based interference and flow-level performance. In [10] and [26] the authors also model flow-level performance in a randomly placed network (h-PPP) using results from queuing theory as well. However, the BS and user spatial coordinates are assumed as input (rather than considering any specific model, stochastic or not). What is more, the user MCS distribution is further assumed as another input to the problem, in order to avoid the key coupling problem between the MCS distribution and network load, which is at the core of this analytical problem. Hence, while useful, this framework can only be considered as a helpful tool that may accelerate the simulation process, rather than an analytical framework for an arbitrary, randomly placed network. Finally, in [26] and [11] the same authors perform a mean-cell analysis towards deriving analytically the mean BS load of a randomly placed network. The mean-cell approximation assumes that all BSs serve exactly the same number of users, so all BSs produce the same amount of interference. Due to the simplicity of the mean-cell approximation, the framework is accurate only for the low load case (as we will see in the result section). As a remark, we note that all the aforementioned works [10], [11], [26] assume that the number of users in a cell is constant (uniform distribution of users) which does not allow to capture load distribution statistics and congestion probability for a BS. Unlike our work, [27] models the system at packet-level rather than flow-level, and would lead to a non work-conserving (and thus inefficient) system if applied to flows. Furthermore, the coupling of BS queues that interfere with each other only when they are active, is handled using a simple first order approximation, where the load and performance of a BS in question is affected by load dependent interference from the rest $N-1$, but the performance of these $N-1$ is assumed as independent (and equal to the best case).

Summarizing, the technical novelty of our work compared to the few related attempts to derive flow-level performance for large random networks consist of one or more of the following: (i) Our framework models the number of users in a cell as a random variable, allowing us to derive the probability distributions for performance metrics of interest (load, utilization, congestion probability, delay) across BSs in the network, rather than just a "mean" BS. (ii) We provide an explicit analysis and formula for both the mean delay and load of a base station, that is significantly more accurate even in the average sense, compared to mean-cell analysis. (iii) The state-of-the-art models of [11] assume that the MCS distribution, necessary to derive flow-level performance metrics, is actually given. We derive this MCS distribution analytically. (iv) Finally, unlike related works, we consider different queueing models for LTE and WiFi BSs to capture the respective MAC better.

As a final note, the seminal work of [9] addresses the optimal user association problem in a single tier from a flow-level dynamics point of view, proposing a load-based association algorithm. While optimal user association within a tier is not considered in this paper, we use a similar load-based approach for choosing between tiers, during our evaluation.

III. Performance at the BS level

We assume that each BS experiences a *dynamic* traffic load and we would like to study the performance at *flow-level*. We state here our assumptions regarding a single randomly chosen BS, and comment where necessary.

A.1: Each *connected* user to a BS generates new *flow* requests randomly, and independently of other users, according to a Poisson Process with density λ_f.

A.2: A flow is a sequence of packets corresponding to the same user or application request (e.g., a file or web page download). Each flow has a random size, in terms of bits, drawn from a *generic* distribution with mean value $\langle s \rangle$.

A.3: The number of users n associated with a BS is a *random* variable with probability mass function (pmf) $f_N(n)$ that depends on the density of the BSs, the density of users, and the association criteria. This pmf will be derived in Section V.

The following Lemma follows easily, by using a simple Poisson merging argument [28].

Lemma 3.1: If n users are associated with a given BS, the aggregate flow arrival process to that BS is Poisson$(n\lambda_f)$.

Remark: While a Poisson arrival model is pretty standard in related literature, note that if the number of users n at a BS is relatively large, assumption (A.1) can be relaxed to more general traffic arrivals, and we can then use the Palm-Khintchine theorem [28] to support Lemma 3.1 as an approximation.

A.4: In the absence of other flows, *a single flow will be served at full rate*, with the maximum Modulation and Coding Scheme (MCS) that the BS can offer to that UE, which in turns depends on the SINR-BLER (block error rate) specifications for that RAT. The rate of the arbitrary user could be assumed as a random variable and the corresponding pmf, $f_{\mathrm{R}}(r)$, is derived in Section VI.

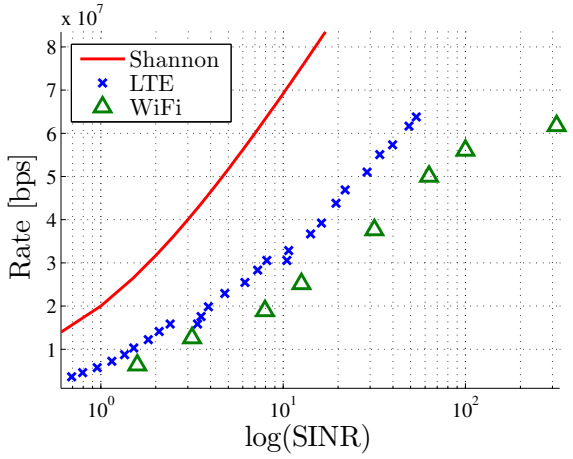


Fig. 1: Comparison between MCS-base rates per RAT and Shannon's limit for an AWGN channel

We note here that it is common, when analyzing wireless networks, to use the Shannon's theorem to derive SINR-rate relationship, as actual RATs do not provide an elegant way to calculate the user's MCS and related rate. When a single network is analyzed, this assumption does not affect the validity of the qualitative results. However, in the case of HetNets, and especially HetNets operating with different RAT, this assumption does not hold, as the offered user rate does not scale the same with respect to SINR, for different RATs. For instance, Fig. 1 presents the outcome of the PHY modeling procedure for the case of LTE and WiFi, as will be described in [29]. It is evident that LTE is, on average, 37% closer to the Shannon rate compared to WiFi, for their common operating SINR range. Hence, if we were to model the SINR-rate relation for both LTE and WiFi according to the Shannon formula, we would significantly overestimate the performance of the WiFi tier.

We will assume a single MIMO layer and a single carrier in our analysis. Increased rates due to spatial multiplexing and carrier aggregation can be included in the model with a proper physical abstraction models.

A. Queueing Model for BS Schedulers

When more than one flows are served in parallel by a BS, the BS operates as a *queueing system*. The service rate for a flow is generally smaller than what assumption (A.4) predicts, and depends on the number of active flows (BS load), and the centralized scheduler (in the case of 3G/4G) or distributed media access control (MAC) protocol (in the case of WiFi) which decides how the available resources will be distributed between flows. While a number of different scheduling algorithms exist, the majority of them tries to allocate the available resources between competing flows (e.g. LTE resource blocks, WiFi channel) in a fair or proportionally fair manner.

1) Resource Fair Scheduler: Assume the BS allocates the same amount of resources to all flows, and they are served simultaneously, e.g., with a round robin, TDMA-like algorithm.

If the service time slot is small (e.g., of packet size) compared to the total size of a flow, the flow level performance at that BS can be approximated by a multi-class M/G/1 Processor Sharing (PS) system. This model has already been used to analyzed 3G/3G+ BS performance [4], [7]. While each flow shares the channel for the same amount of time (hence "resource fair"), during that time it might transmit at a different rate, depending on its SINR and resulting MCS (hence the "multi-class" service).

LTE schedulers are significantly more complex, allocating competing flows both time and frequency resources (Resource Blocks), possibly taking into account the queue backlog of each flow and flow priority, and also attempting to take advantage of instantaneous SINR variations in time and frequency to achieve further multi-user diversity [30]. While a large number of algorithms have been proposed (see e.g., [31] for an extensive survey), in the lack of special priority traffic, most implemented schedulers lead to a proportionally fair throughput allocation between flows [30] and can also be approximated by a similar multi-class M/G/1 PS queue. The following is a direct application of the multi-class M/G/1/PS result [32].

Lemma 3.2: For a BS with n users generating flows of mean size $\langle s \rangle$, with instantaneous transmission rates drawn from distribution $f_{\mathrm{R}}(r)$, and allocated resources by a resource fair scheduler, the effective service rate of the cell is

$$\langle \mu \rangle_{\text{rf}} = \left(\sum_i \frac{f_R(r_i) \cdot \langle s \rangle}{r_i} \right)^{-1} \text{ flows/sec}, \tag{1}$$

and the mean flow delay is given by

$$E[T]_{\text{rf}} = \frac{1}{\langle \mu \rangle_{\text{rf}} - n\lambda_f} \ . \tag{2}$$

We further define the BS's load as

$$\rho = \frac{\text{input job rate}}{\text{service job rate}} = \frac{n\lambda_f}{\langle \mu \rangle_{\text{rf}}} \ , \tag{3}$$

when the system is stable $\rho < 1$. Performance gains from opportunistic scheduling can be included in the above equation as a multiplicative factor in front of $\langle \mu \rangle_{\text{rf}}$.

2) Throughput Fair Scheduler: Some schedulers attempt to achieve fairness more aggressively, by trying to equalize per flow throughput for all nodes. For example, if two concurrent flows experience different channel conditions (say one being "far" and one being "near" the BS) a throughput fair scheduler will attempt to give more resources to the flow with the worse channel (e.g., more resource blocks in the case of LTE, or schedule the far flow more often in the case of 3G). This can be seen as a Generalized or Discriminatory Processor Sharing system (a generalized version of the M/G/1/PS) [22], with different weights per flow that, for throughput-fair systems, can be taken as inversely proportional to the average rate experienced by that flow.

It is known that throughput fair schedulers perform poorly compared to proportionally fair ones, and thus are not often considered [8]. Nevertheless, throughput fair scheduling turns out to be a good approximation of how the 802.11 WiFi MAC allocates resources between flows [33]. In WiFi, all nodes compete for the channel and when they do get access, in the basic implementation, they send a single frame and then have to retry. WiFi like LTE supports rate adaptation, therefore each frame might be transmitted at a different rate, depending on the maximum MCS that can be offered to the respective node. Nevertheless, due to the random access MAC, each node gets access with equal chance, regardless of their distance from the AP. If each flow corresponds to a large number of frames (usually a good assumption given the small max size of a frame), this essentially equalizes the long-term throughput of each flow, regardless of its MCS. Hence, the WiFi scheduler for a single BS could be seen as throughput-fair, and can be modeled as a Discriminatory Processor Sharing (DPS) queue.

The following lemma presents the mean service time ($E[T]$) for such a throughput-fair scheduler in a system with rate adaptation.

Lemma 3.3: The mean per flow delay for a throughput fair system with input flow rate λ, and flows being served with rates drawn from a pmf $f_R(r_k)$, can be calculated according to

$$\begin{aligned} E[T]_{\text{tf}} = \sum_k f_R(r_k) \Big[& \frac{\langle s \rangle / r_k}{1 - \lambda / \langle \mu \rangle_{\text{tf}}} \\ & + \frac{\sum_j f_R(r_j) \lambda (1 - \frac{r_j}{r_k}) (\langle s \rangle / r_j)^2}{2(1 - \lambda / \langle \mu \rangle_{\text{tf}})^2} \Big] \ , \end{aligned} \tag{4}$$

where $\langle s \rangle$ is the mean flow size and

$$\langle \mu \rangle_{\text{tf}} = \left(\sum_k \frac{f_R(r_k) \cdot \langle s \rangle}{r_k} \right)^{-1} \ . \tag{5}$$

Proof: Let us first consider a throughput fair system, and derive the mean service rate, Eq. (5). Consider a long time interval during which N packets get transmitted, corresponding to different flows. Assume each packet is of equal size S (e.g., the max WiFi frame size) but is transmitted with a possibly different rate r drawn from pmf $f_R(r)$ with K discrete values, depending on the MCS used for transmitting that packet. Assume that out of these N packets, N_i are transmitted with rate r_i, ($\sum_i N_i = N$). Hence, the *average* transmission rate in terms of bits/sec for these N packets is

$$\frac{\text{bits in } N \text{ pkts}}{\text{transmission time for } N \text{ pkts}} = \frac{N \cdot S}{N_1 \frac{S}{r_1} + \cdots + N_K \frac{S}{r_K}} \ . \tag{6}$$

However, as N goes to infinity, the N_i converges to its mean value $f_R(r_i) \cdot N$ by the law of large numbers, hence the denominator of Eq. (6) converges to

$$\lim_{N \to \infty} (N_1 \frac{S}{r_1} + N_2 \frac{S}{r_2} + \cdots + N_K \frac{S}{r_K}) = \sum_i f_R(r_i) \cdot N \cdot \frac{S}{r_i} \ . \tag{7}$$

Since $\frac{1}{x}$ is continuous and all $r_i > 0$, we can use the Continuous Mapping Theorem [34](Th. 5.23) to show that Eq. (6) converges to

$$\frac{1}{\sum_i f_R(r_i) \cdot \frac{1}{r_i}} \ . \tag{8}$$

Eq. (8) gives the average transmission rate of the scheduler over a sufficiently long sample path of packets. Since the

system is ergodic, we can divide with the mean flow size $\langle s \rangle$ to get Eq. (5).

To go beyond the mean load and derive the mean delay for this system, we use the approximation from Avrachenkov *et al.* [35] for DPS systems, which to our best knowledge, provides the most accurate solution, assuming large enough flow sizes. Specifically, for a given network load, the expected delay for flows of class k having size x, denoted as $E[T_k(x)]$, asymptotically converges to

$$\lim_{x\to\infty}\left(E[T_k(x)] - \frac{x}{1-\lambda/\langle\mu\rangle_{\text{tf}}}\right) = \frac{\sum_j \lambda_j(1-\frac{w_k}{w_j})E[X_j^2]}{2(1-\lambda/\langle\mu\rangle_{\text{tf}})^2}. \quad (9)$$

We applied the equation above to our system, by having classes corresponding to different MCS and weights $w_i = 1/r_i$ inversely proportional to the service rate. $E[X_i^2]$ is the second moment of service requirement (flow sizes normalized in seconds) for flows of class i, which we approximate with $E[X_i^2] \approx (\langle s \rangle / r_i)^2$. The incoming job rate λ_i of class i: assuming that the probability of an incoming job to be of class i is $f_R(r_i)$ then $\lambda_i = f_R(r_i) \cdot \lambda$.

Putting everything together gives us Eq. (4). ■

Note that the above analysis, when applied to 802.11, ignores the impact of collisions and RTS/CTS frames, analyzed in [36], and thus is an upper bound. Nevertheless, in light of the high speeds and features of 802.11n\ac, such as frame aggregation or block of ACK transmissions (by a single node), implies that the impact of such overhead can be safely ignored. (we refer the interested reader to [29].)

It is interesting to observe that the above result implies that *the mean service rate, in the long run, for a WiFi system with rate adaptation, turns out to be the same as that of a resource-fair system (Eq.* (1)). Nevertheless, this does *not* imply that the mean flow delay is also the same, as the scheduling discipline is different (DPS instead of PS). Unfortunately, there does not exist a closed form solution for the mean flow delay of a throughput fair system. We will therefore consider the following two approximations.

Approximation 1: When the BS load is low (i.e., $\frac{\lambda}{\langle\mu\rangle} \to 0$), flows rarely "compete" with each other, and it is easy to see that the mean delay is approximately equal to the resource fair case, i.e., Eq. (2). This is also *a lower bound* on the delay, for higher load values. Furthermore, the observed poor performance of WiFi [33] has led researchers to propose slight modifications of 802.11, taking advance of the new feature of frame aggregation [37], in order for WiFi to operate closer to a resource fair scheduler.

Approximation 2: For general loads, we can use Avrachenkov's approximation as presented in Eq. (4). As we mentioned, this result is an asymptotic as the service rate (flow size) is going to infinity, but even for small flow sizes our simulation results show that the approximation is decent.

Fig. 2 shows the simulated mean delay of a throughput fair system as well as the two analytical approximations (resource fair and [35]) for flow size $\langle s \rangle = 12.5$ MBytes, respectively. For small flow sizes the performance of a throughput fair system lies in between the two approximations and as the size is increasing the performance of the system is approaching the approximation of [35], that we use in this paper.

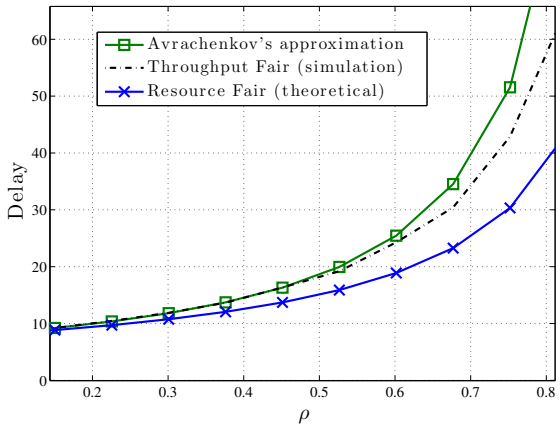


Fig. 2: Delay vs load for a WiFi throughput fair system

B. Network-wide Performance

So far we have focused on a single BS. Our goal in this paper is to consider a network of a large number of such BSs, possibly belonging to different orthogonal HetNet tiers, and understand the performance of this network along to main dimensions:

- *Stability (congested probability)*: We would like to know the percentage of BS whose input load $n \cdot \lambda_f$ exceeds the available service capacity $\langle\mu\rangle$ (i.e., $\rho > 1$) thus exhibiting per flow delays that grow to infinity.
- *Utilization*: Network utilization expresses the probability of a randomly chosen BS to be active at a random time instance or the percentage of network's active BS at an arbitrary moment. It can be defined as the average utilization over all BSs, $\mathcal{U} = E[\mathcal{U}_{\mathcal{BS}i}]$, where $\mathcal{U}_{\mathcal{BS}i} = \min(\rho_i, 1)$ is the percentage of time that i-th BS is active, ρ_i according to Eq. (3).
- *Per flow delay*: we would like to know the expected network-wide delay for a randomly chosen user flow, when this flow is served by a stable BS.

Based on the previous discussion, for a single tier it is clear that these metrics depend on the same two key parameters:

1) The cardinality n of the users associated at a BS, which is a random variable with pmf $f_N(n)$ that depends on the topology of BS and user density.
2) The probability that each user is served with a given rate r, namely the rate distribution $f_R(r)$ for this BS that depends on the topology and interference between nearby BSs.

We derive $f_N(n)$ in Section V. We then derive $f_R(r)$ in Section VI. In the case of multiple tiers, the above quantities also depend on the association policies between tiers, Section VII.

IV. PHY Layer Modeling

Before we proceed with the derivation of the cardinality and rate probability distributions, we state here our assumptions about the network topology and physical layer model.

A.5: Users are distributed according to an independent Poisson Point Process with density λ_u.

A.6: We assume a network with k independent tiers of BSs. The number of BSs of tier j inside an area S follows a homogeneous Poisson Point Process (PPP), Φ_{BS_j}, with density λ_{BS_j}, and independently of other tiers. Hence, the number of BSs of tier j in an are S

$$P(N_j = n \mid S) = \frac{(\lambda_{\text{BS}_j} S)^n e^{-\lambda_{\text{BS}_j} S}}{n!}, \quad n = 0, 1, \dots \quad . \quad (10)$$

We assume that each tier operates at different frequency. Thus, **tiers** are orthogonal, i.e., interference at each BS originates only from BSs of the same tier [1]. It can be argued that the above model does not exactly capture current cellular networks, consisting mostly of macro eNodeBs that are usually carefully planned to maximize coverage, and could perhaps be better modeled by standard hexagonal or grid topologies. Nevertheless, in the case of WiFi access points or future, considerably more dense networks consisting mostly of pico- or femto-cells, topologies are expected to be considerably more random and uncoordinated, with BSs having a non-zero probability to be very close. We can consider the aforementioned two topologies as ideal and worst case scenarios respectively, in terms to interference. As shown in [13], the coverage probability in terms of the SINR threshold, in real BS deployments, lies in most cases roughly midway between the coverage probability in the two extreme cases above.

A.7: A standard power loss propagation model is used. We assume a path loss exponent $\alpha > 2$ (for $\alpha \leq 2$ the denominator of SINR goes to infinity), Rayleigh fading at the channel with mean 1 and constant transmit power of P_{tx}. So, the received power at distance d from the BS is given by $P_{\text{rx}} = hd^{-\alpha}$ where h follows an exponential distribution, $h \sim \exp(P_{\text{tx}})$. Hence, the SINR is given by

$$SINR_i = \frac{P_{\text{rx}_i}}{\sum\limits_{n \neq i} P_{\text{rx}_n} + \sigma^2}, \quad (11)$$

where σ^2 is calculated w.r.t the bandwidth (BW) from $\sigma^2_{\text{dBm}} = -174 + 10\log_{10}(BW)$ [38].

A.8: We assume that all BS of the same tier have equal transmit power, but transmit power might differ between tiers. Similarly, all BSs at the same tier implement the same scheduling policy (either Resource Fair or Throughput Fair), but different tiers might have different policies.

A.9: When more than one tiers exist in the network, a user association policy decides which tier a given user will be sent to. We consider the following association policies:

- *Off-loading:* In the simplest scenario, the operator might steer to a preferred tier (e.g. WiFi or 4G) all users that can connect (i.e., receive a sufficiently high SINR) to this tier. Such off-loading can be achieved by broadcasted *Absolute Priorities* to all nodes in RRC_IDLE state or dynamically through *Dedicated Priorities* indicated to nodes in RRC_CONNECTED state [30], [39].

[1]In the case of two tiers sharing the same frequency, we could model this as a single tier with possibly different transmission powers or rates (e.g., for small cells and macro-cells).

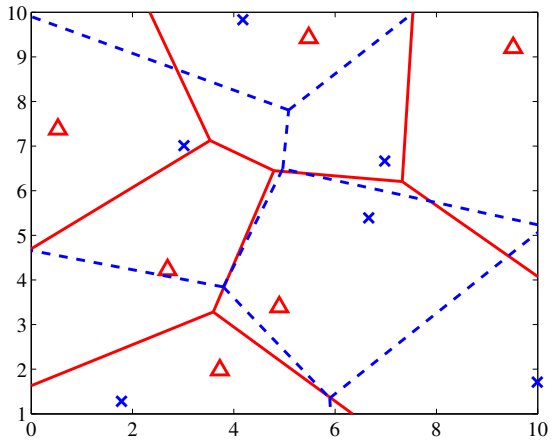


Fig. 3: Voronoi Tessellation example, 2-tiers, solid lines correspond to tessellations in respect to △ network and dash lines to × network

- *Max-SINR:* A user choses to associate with the tier that provides the best SINR.
- *Min-Delay Association:* The load of each tier is also taken into account when associating, in order to minimize the average delay of the system. While a number of load-based association algorithms have been considered, here we assume a simple version of the association rule proposed in [9].

Within a given tier, we assume the user association criterion is maximum SINR, which is standard. A number of recent works [9], [40] have shown that this criterion is sub-optimal and more sophisticated criteria (e.g., load-based, as in the case of inter-tier association) could be applied for intra-tier association, in order to improve performance. Nevertheless, these results are equally applicable to every tier, and our focus in this paper is relative impact of using multiple tiers, rather than the optimal performance of each tier itself.

Assuming all of the above and additionally, that on average, the received power is monotonic in respect to distance, our criterion is simplified to the closest distance criterion, so, the BSs's coverage areas could be represented by Voronoi Regions (Tessellations), Fig. 3 shows two orthogonal networks and their Voronoi regions.

V. Cardinality of Associated Users

We are now ready to consider the pmf of the user cardinality per BS, $f_N(n)$, which as explained earlier decides the total input traffic to each BS. We first consider a single BS tier. Even in this case, deriving this cardinality for an arbitrary cell is not trivial. Observe that the size of an arbitrary cell is a random variable, depending on the random BS topology, and the number of users given a specific cell size is also a random variable. The proof for the following theorem can be found in [41] or [42]. Additionally [41] presents an approximation that is accurate and significantly less computationally demanding in terms complexity and memory allocation.

Theorem 5.1: Consider a single tier of BSs distributed in 2D as a homogeneous PPP with density λ_{BS}, and offering coverage

to a set of users distributed as another PPP with density λ_{u}. Assume further that user association within this tier is done using the closest-distance rule, as explained in Section IV. The ratio between users and BS density is defend as $\zeta = \frac{\lambda_{\mathrm{u}}}{\lambda_{\mathrm{BS}}}$. Then, the probability of having exactly n users in an arbitrary cell, $f_N(n)$, is given by:

$$f_N(n) = \frac{343}{n!15}\sqrt{\frac{7}{2\pi}}\frac{\zeta^n}{(\zeta+\frac{7}{2})^{n+\frac{7}{2}}}\Gamma(n+\frac{7}{2}) \ , \qquad (12)$$

where Γ is the gamma distribution. The first and second moments of Eq. (12) are

$$\langle n \rangle = \zeta \qquad \text{and} \qquad \mathrm{var}_n = \zeta + \frac{2}{7}\zeta^2 \ . \qquad (13)$$

Finally, if we take into account the asymptotic behavior of the Gamma function, $\lim_{n\to\infty}\frac{\Gamma(n+\alpha)}{\Gamma(n)n^\alpha} = 1$, and apply the definition $\Gamma(k) = (k-1)!$, Eq. (12) can be significantly simplified to

$$f_N(n) = A_\zeta u_\zeta^n n^{5/2} \ , \qquad (14)$$

where $u_\zeta = \frac{\zeta}{\zeta+7/2}$. Due to this asymptotic step, we need to renormalize the pdf, with A_ζ serving as the new normalization factor, which depends on ζ as well.

We should mention that because of the memoryless property of h-PPP there is no correlation in users cardinality among adjacent cells.

We next move to the calculation of the MCS-rate distribution, and the impact of inter-tier association on both user cardinality and rate distribution.

VI. MCS Distribution for each RAT

We are interested in the maximum rate (or MCS) a user can receive data at from the BS it is associated with, given a desired BLER. Our goal is to derive the rate distribution $f_R(r)$ in order to calculate the service rate $\langle\mu\rangle$ in terms of flows/sec for the average BS. This rate depends on the SINR for that user. For a given SINR, the offered MCS is well defined in the specifications of a given RAT. The SINR in turn depends on both the distance of the user to the serving BS and the interference from other nearby BS of the same tier. Furthermore, a nearby BS might not interfere if it is actually not transmitting at that time, which further complicates analysis. For this reason, we will first consider a "saturated" scenario where interfering BS are assumed to always be ON and interfering. We will then consider the case of load-based interference, where a BS only interferes if it is currently *active* serving at least one user.

A. Rate Distribution for Always ON Interference

We will assume again that BSs and users are distributed according to independent homogeneous PPPs, and focus on a specific network tier. In [13], the authors present an approach to derive the "coverage probability" of a randomly located user, i.e., the probability that the user's SINR is above a certain threshold. In doing so, it is assumed that interfering BSs always transmit with a power P_{tx}. This assumption is a good approximation when the load of the system is high, in which case the utilization of most BS is close to 1 (i.e., are serving users most of the time). It can also be a valid assumption if the SINR at the user is measured with respect to Reference Signals (i.e., "pilots") that are transmitted at specific times slots by all BS, regardless of whether a BS is serving users or not at that time [30]. Nevertheless, this is not always the case. As a result, in scenarios where BS utilization is lower, this assumption might lead to fairly pessimistic results. We consider this in Section VI-C.

For the sake of completeness, we mention here again the results from [13] that are applicable to our problem. Given a BS density λ_{BS}, and path loss constant α, the coverage probability for an SINR threshold T is

$$\begin{aligned} p_c\left(T,\lambda_{BS},\alpha\right) &\triangleq \mathbb{P}\left[SINR > T\right] \\ &= \pi\lambda_{BS}\int_0^\infty e^{-\pi\lambda_{BS}u(1+\beta(T,\alpha))-\frac{1}{\mu}T\sigma^2u^{\alpha/2}}du \ , \end{aligned} \qquad (15)$$

where $\beta\left(T,\alpha\right) = T^{2/\alpha}\int_{T^{-2/\alpha}}^\infty \frac{1}{1+u^{\alpha/2}}du$.

If we assume that additive noise is negligible w.r.t. interference (a reasonable assumption for the dense modern networks) Eq. (15) can be significantly simplified as $p_c\left(T,\lambda_{BS},\alpha\right) = 1/\left(1+\beta\left(T,\alpha\right)\right)$. Furthermore, if we assume that $\alpha = 4$, we obtain

$$p_c\left(T,\lambda_{BS},4\right) = \frac{1}{1+\sqrt{T}\left(\pi/2 - \arctan\left(1/\sqrt{T}\right)\right)} \ . \qquad (16)$$

Finally, assuming and SINR threshold τ_i for each MCS (mcs_i), the pmf of the MCS $f_{MCS}(mcs)$ can be obtained at Eq. (17) through the coverage probability.

$$f_{\mathrm{MCS}}(mcs_i) = p_c\left(\tau_i,\lambda,\alpha\right) - p_c\left(\tau_{(i+1)},\lambda,\alpha\right) \ . \qquad (17)$$

Given the MCS, the actual rate can be easily calculated based on the total bandwidth of the system in question. Increased rates due to spatial multiplexing and independent carriers can be included in the model with a proper physical abstraction models.

B. Multi-tier Association

In the case of a multi-tier network, the user density λ_u and pmf of MCS $f_{MCS}(mcs)$ of each tier also depend on the inter-tier association policy. We present here how to calculate those two parameters for the basic association schemes considered. We should clarify that the association rules below denote the tier that the user will be associated with and not the BS. Given the tier, the user is associated with its closest BS. Let's assume there are two tiers, with $f^i_{\mathrm{MCS}}(mcs)$ and $p^i_c\left(T,\lambda_{BS},\alpha\right)$ be the MCS distribution and the coverage probability of each tier $i = \{1,2\}$ as derived in Eq. (15) and Eq. (17) respectively.

Lemma 6.1 (Off-load): For the Off-load case, the user is associated with tier-2, if the achieved SINR for that tier is higher than a coverage threshold τ_0. Then, the pmf of MCS and the density of users for tier-1 are given by

$$\begin{aligned} p_1^{'}(mcs) &= f^1_{\mathrm{MCS}}(mcs)\left(1-p_c^2\left(\tau_0,\lambda_{BS},\alpha\right)\right) \ , \\ \lambda_u^1 &= \lambda_u\left(1-p_c^2\left(\tau_0,\lambda_{BS},\alpha\right)\right) \ , \end{aligned} \qquad (18)$$

where the term $\left(1-p_c^2\left(\tau_0,\lambda_{BS},\alpha\right)\right)$ denotes the non-coverage probability, meaning the probability that the user's

SINR in tier 2 is less than the threshold τ_0. Due to the tiers' orthogonality, the probability of achieving MCS_i at one tier and the non-coverage probability at the second are independent. Finally, due to Poisson thinning, the density of first tier is the initial density λ_u thinned by the non-coverage probability at tier two.

Lemma 6.2 (Max SINR): For the Max-SINR case, the user is associated with the tier providing the maximum SINR. Thus, pmf of the MCS and the corresponding density of users for tier-1 are as follows (similar for tier-2)

$$\begin{aligned} p_1^{'}(mcs) &= \int_{\tau_i}^{\tau_{i+1}} p_c^1(\tau)\left(1-p_c^2\left(\tau_0,\lambda_{BS},\alpha\right)\right)d\tau\ , \\ \lambda_u^1 &= \lambda_u \int_0^{\infty} p_c^1(\tau)\left(1-p_c^2\left(\tau_0,\lambda_{BS},\alpha\right)\right)d\tau\ . \end{aligned} \tag{19}$$

The above tier-association rules are the two most common ones considered. However, these rules purely depend on coverage statistics and ignore the load of the network which plays an equally, if not more, important role on performance. We therefore propose a third tier-association rule that takes this load into account. The distributed association algorithm of [9], proves that each user, in a single tier, in order to minimize the delay should be associated with the BS j that maximizes the metric $C_j(1-\rho_j)^2$, were C_j is the user's throughput on BS j and ρ_j is the corresponding BS load. Here we provide a modified version for the multi-tier association.

Lemma 6.3 (Min Delay): Let us assume that the arbitrary user could operate with any of N different MCS at tier 1, according to pmf $f^1_{\text{MCS}}(mcs)$, and M different MCS at tier 2, according to $f^2_{\text{MCS}}(mcs)$. We can form a new set $\mathcal{L}$ of $N\times M$ values of the combinations of the two initial pmfs. Each of those $N\times M$ possible sub-sets of users will associate to the tier i according to the following criterion:

$$i(x)=\underset{j\in\mathcal{B}}{\text{argmax}}\ c_j\left(1-\mathcal{U}_j\right)^2\ ,\forall x\in\mathcal{L}\ , \tag{20}$$

where $\mathcal{B}$ is the set of all tiers (in our case $\{1,2\}$), c_j is the rate that tier j is able to provide at users of type $x\in\mathcal{L}$ and $\mathcal{U}_j$ is the utilization of each tier. This association rule is applied iteratively among all classes, until convergence (for our RATs and topology, two iterations are needed).
It is easy to see that the above lemmas can be easily extend to more than two tiers.

C. Rate Distribution for Load-based Interference

As mentioned earlier, the previous results assume that all BS are interfering all the time. In practice, when the load ρ of a BS A is low, e.g., $\rho=0.5$, then BS A would be transmitting and causing interference only 50% of the time [2]. This implies that another nearby BS B will be actually serving users at higher rates than the ones predicted in the saturated case. This, in turn, means that BS B will also have a higher $\langle\mu\rangle$ and thus lower utilization than the one predicted, which in turn creates less interference for BS A. At flow level, this creates a system of dependent PS queues, which is notoriously hard to analyze at Markov chain level (see e.g. [23] for an attempt to derive some performance bounds). We choose to take here a different approach and use an iterative algorithm in order to calculate $\langle\mu\rangle$ of those dependent BSs.

Let us assume that we knew the correct $\langle\mu\rangle_{lb}$ and expected utilization $\mathcal{U}$ of the network, assuming a load-based interference as described above. Then, the following lemma extends the previous analysis based on stochastic geometry, in order to approximate the coverage probability for this load-based interference scenario (Due to space limitation, the proof of this lemma can be found in [43]).

Lemma 6.4: The coverage probability of an arbitrary user in a random cellular network (assuming thermal noise negligible compared to interference), when the average utilization of BSs is $\mathcal{U}$, and each BS is interfering only for the amount of time that it is serving users (i.e., for a percentage of time $\mathcal{U}\leq 1$) is given by

$$\begin{aligned} p_c^{lb}\left(T,\lambda,\alpha\right) = \sum_{n=0}^{N_{max}-1}&\left(f_N(n)\frac{1}{1+\mathcal{A}_{\mathcal{U}}}\right) \\ &+\overline{F_N}\left(N_{max}\right)\frac{1}{1+\mathcal{A}_{\mathcal{U}=1}}\ . \end{aligned} \tag{21}$$

Where $\mathcal{A}_{\mathcal{U}}=(T\mathcal{U})^{2/\alpha}\int_{(T\mathcal{U})^{2/\alpha}}^{\infty}\frac{1}{1+u^{\alpha}/2}du$, we remind that $\mathcal{U}=\min(n\lambda_f/\langle\mu\rangle,1)$ and $\overline{F_N}$ is the ccdf of users' cardinality, Section V. Assuming path loss exponent $\alpha=4$, Eq. (21) could further be simplified by replacing $\mathcal{A}_{\mathcal{U}}$ and $\mathcal{A}_{\mathcal{U}=1}$ with

$$\begin{aligned} \mathcal{A}_{\mathcal{U}} &= \sqrt{\frac{T}{N_{max}}n}\cdot arccot\left(\frac{1}{\sqrt{\frac{T}{N_{max}}n}}\right)\ , \\ \mathcal{A}_{\mathcal{U}=1} &= \sqrt{T}\cdot arccot\left(\frac{1}{\sqrt{T}}\right)\ . \end{aligned}$$

Combining Eq. (21) with the aforementioned Eq. (17) we obtain again the user's MCS distribution. But, taking into account that $N_{max}=\langle\mu\rangle/\lambda_f$, we can observe from Eq. (21) that in contrast to the always ON case the coverage probability depends on service rate $\langle\mu\rangle$. Thus, MCS distribution depends on the $\langle\mu\rangle$ as well. On the other hand, $\langle\mu\rangle$ is dependent on the MCS distribution as we can see from Eq. (5).

Due to the aforementioned dependencies we can re-write Eq. (5) as

$$\langle\mu\rangle=\left(\sum_{mcs_i}\frac{f_R(mcs_i|\langle\mu\rangle)\cdot\langle s\rangle}{r(mcs_i)}\right)^{-1}\ . \tag{22}$$

To prove analytically the convergence of the *fixed point problem* of Eq. (22) is not trivial. A study about the coupling of load and rate distribution is presented in [21] where authors provide results related to computational aspects for numerically approaching the solution. In the cases of our interest, the relationship between MCS thresholds and rates defined according to LTE and WiFi systems or Shannon's formula. For those cases the left part of equation Eq. (22) is a strictly increasing function with derivative equal to one, and the right part is again a strictly increasing function with respect to $\langle\mu\rangle$ but its derivative is strictly smaller than one

[2]Even if the SINR estimate is based on the pilot signals, which are always transmitted at the designated LTE resource elements, the *actual* interference experienced during transmission will be lower in practice, leading to better effective rates (e.g., due to fewer HARQ retransmissions required).

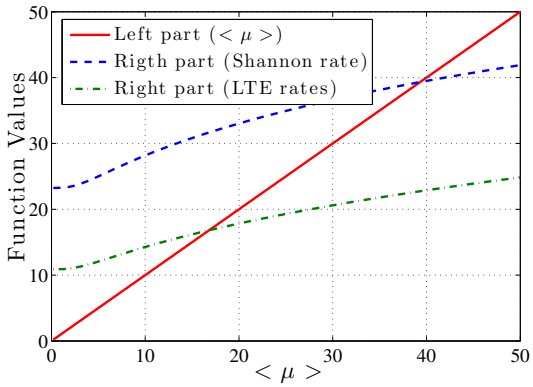


Fig. 4: Left and right part of the Eq. (22) for the cases were the relationship between SINR and the corresponding data rate calculating according to i) LTE RAT and ii) Shannon's formula

(calculated computationally, [44]), Fig 4. Thus there is exactly one solution for the Eq. (22) and could be approached by simple gradient methods.

D. Rate for each RAT

Two parameters are missing in order to derive $\langle\mu\rangle$. Firstly, we need SINR thresholds τ_i for each MCS mode to calculate f_{MCS} from Eq. (17) and secondly, the corresponding rate of each MCS.

The supported MCS modes are RAT depended and always defined at the corresponding protocol description documents [45], [46]. The operation threshold for each mode is not always defined in the protocol since it heavily depends on the receiver implementation characteristics. For the 2-tier example that we demonstrate in this paper we will need one SINR table for the LTE modes and one for the WiFi. The reference receivers which were used are the OpenAirInterface platform (for more details see [47], [29]). The results of this procedure depicted in Fig. 1, where every marker corresponds to an MCS and the x-coordinates of them are the SINR threshold τ_i and the y-coordinates are the corresponding rates $rate_i$. For the WiFi case we consider the effective throughput of the access points, taking into account the overhead of this RAT like sniff, ack/nak, RTS/CTS, coalitions, etc [29].

VII. Results

Already today it is possible to integrate WiFi networks into the core networks of cellular systems, and perform off-loading of traffic to WiFi access points. In future releases of 3GPP, a tighter integration of WiFi and LTE technologies is expected. For this reason, we choose a heterogeneous RAT scenario consisting of LTE and WiFi orthogonal tiers, as a case study. We will consider the following "fixed" parameters for the two networks: (i) pathloss $\alpha = 4$, (ii) thermal noise $\sigma^2 = -100$ dBm (iii) $BW_{LTE} = BW_{WiFi} = 20$ MHz, (iv) one antenna per eNodeB and one spatial stream per WiFi AP. Finally, we should mention that if the thermal noise is much smaller than the interference (which is the case in our system), the value of P_{tx} does not affect the results, as shown in [13]. The rest of the parameters will act as variables, and we'll discuss their value range per scenario.

For the WiFi network, a number of different setups and 802.11 standards could be considered. The traditional WiFi protocol is tuned to a roughly 20 MHz channel. The newer versions of WiFi (n/ac) have the capability of channel bonding in order to operate with 40 to 160 MHz. Larger bandwidths could also be considered via carrier aggregation in LTE. All of those additional channel capabilities are orthogonal to our model and out of the scope of this paper, so we assume for simplicity and fairness that both networks operate with 20 MHz.

Finally, as explained earlier, current WiFi implementation operate closer to a throughput fair scheduler. However, as mentioned in Section III, the WiFi scheduler could be modified to avoid the "WiFi anomaly" problem and operate as resource fair [33]. We will therefore consider WiFi with both types of schedulers, in order to better understand their impact.

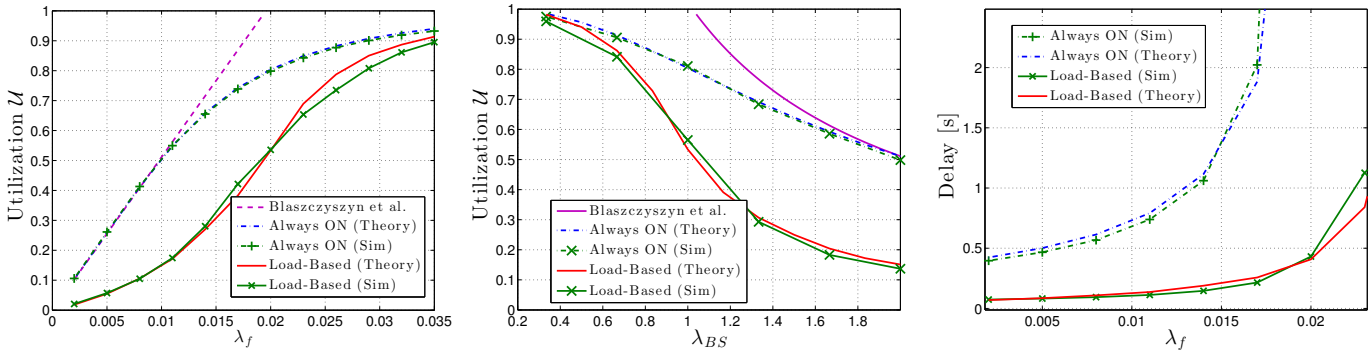
A. Model Validation

As a first step, we would like to validate our basic theoretical results for a single tier, against simulation results, in both saturated and load-based interference scenarios. Additionally, we compare our method with the analytical approach of [11] that provides closed form results about the network load for the saturated case. We remind the reader that [11] assumes that the MCS distribution is somehow known, we fill this gap by using the MCS distribution as it calculated in our framework. An LTE network is considered for this purpose (but can be expanded to any RAT with the proper model of its PHY and MAC characteristics). The performance metrics from the simulated scenarios that are used for the comparison are (i) network's utilization[3] and (ii) average flow delay of the median BS[4].

Our packet-level simulator generates BSs and users randomly placed in a large surface with given densities (λ_{BS}, λ_u). Users are associated with the closest BS and generate flows according to a Poisson distribution with density λ_f and average flow size $\langle s\rangle = 5$ Mbits (625 Kbytes). The flows are forwarded to the corresponding BS which is modeled as a multi-class M/G/1/PS. The service rate of each flow for every time quantum is calculated via the SINR-MCS relation for LTE. We will consider two interference scenarios: (1) always ON case, where all the neighboring BS are contribute to the interference (corresponding to Section VI), (2) load-based case, where we calculate interference by taking into account only the base stations that are ON at this time quantum (corresponding to Section VI-C). We further consider only the users whose SINR is higher than the threshold of the lowest MCS for the always ON case (otherwise a user connected in one quantum might be outside of coverage in the next). Fig. 5 (a) and (b), present network's utilization $\mathcal{U}$ w.r.t. λ_f and

[3]It turns out that congestion probability, average delay, interference, etc. depends more on utilization than on load.

[4]The latter is computed by calculating the mean delay for each BS in the simulation and then taking the median among the BSs. We choose the simulated median rather than the average, as the latter grows to infinity even if a single BS is congested.



a: Load ρ w.r.t flow density λ_f, BS density $\lambda_{BS} = 1$

b: Load ρ w.r.t BS density λ_{BS}, flow density $\lambda_f = 0.02$

c: Delay w.r.t flow density λ_f, BS density $\lambda_{BS} = 1$

Fig. 5: Comparison of theory and simulation results for the case of single tier LTE network

λ_{BS} respectively, for both scenarios $\lambda_u = 200$. Three general comments from those plots are: (i) The Blaszczyszyn *et al* framework, because of the mean value approximation is able to predict accurately the performance of the system only when it is under-utilized. The prediction error depends on the BS congestion probability, which for low loads is $P(\rho > 1) \approx 0$. Mean value analysis essentially fails due to Jensen's inequality, using the mean load directly, and implicitly averaging some congested BSs (with load $\rho > 1$). Due to the concavity of the function, the utilization turns out to be higher than the average utilization of a cell. For example, in Fig. 5 (a) for a $\lambda_f = 0.2$ the prediction error of utilization according to [11] is 20% and ours model is 0.5%, roughly 40 times less. While the utilization is a relatively simple metric, it is easy to see that mean value analysis can have an equally important (if not bigger) impact on delay, especially in the case of load-based interference. Clearly, the amount of interference a neighboring BS contributes depends on the percentage of time it is active, i.e. it's utilization. Hence, overestimating this utilization will overestimate the neighboring interference and underestimate the respective service rates (of the coupled queues), hence further failing to predict delays. Hence, mean value analysis like the one used in [11] could be seen as a first order approximation useful for low loads (and relatively large networks) only. (ii) Both of our theoretical results match the simulation results quite well. (iii) The gap between the always ON and load-based interference scenarios are extremely high, underlining the importance of the latter.

In Fig. 5 (a), for $\lambda_f = 0.02$ the always ON prediction is that the network is 70% loaded instead of 30% of the load-based. That means that the network could be much more robust w.r.t. data traffic than the studies that assume saturated BSs predict.

In Fig. 5 (b), for high density of BS always ON model predicts 50% utilized network, while load-based only 15%. The gap between always ON and load-based prediction increases w.r.t. density of the network. This happens because saturated analysis is able to capture only the gain coming from the fact that an arbitrary BS on average serves less users at a denser network, but not the gain coming from the fact that surroundings BSs will be less loaded, and therefore will cause less interference. Thus, the gain to deploy a denser network is much higher than predicted by an analysis that does not take the load-depended interference into account.

Fig. 5 (c), shows the median delay of the packet-level simulator as well as the theoretical predictions for saturated and load-based cases. Again it can be seen that the theoretical predictions are quite accurate, and that always ON interference over-estimates delay by orders of magnitude.

B. Comparing Different RATs

Having validated our theoretical results, we proceed now with their direct application to different scenario of interest, in order to obtain insights regarding the congestion probability of a BS (the probability that a BS's load is $\rho > 1$) and flow delay statistics in large random topologies. We first study the impact of the following elements on flow-level performance: (a) the MCS-SINR relation (which differs between WiFi and LTE), (b) the scheduler (throughput fair and resource fair), and (c) the type of interference (always-on or load-based). We do this first for single-tier systems, before moving on to multi-tier systems. This will also facilitate our subsequent discussion of 2-tier systems, where multiple factors affect performance concurrently.

At this point, is useful to introduce the following abbreviations for the legends in all figures: (*ON*) or (*LB*) refer to the always ON or load-based interference case respectively. Additionally, for WiFi tiers (*app1*) refers to the resource-fair and (*app2*) to the throughput-fair scheduling policy approximation (see section III).

Fig. 6 and 7 present compactly the performance (congestion probability and delay) for the two networks of interest (LTE, WiFi) for the same density of connected users ($\lambda_u = 100$) and for average flow size $\langle s \rangle = 12.5$ Mbytes. As already mentioned at Section III, two different approximations for the MAC performance of WiFi are assumed: (i) similar MAC performance with LTE (i.e., resource fair scheduler), which is the best case scenario for WiFi; this is a valid approximation for very low loads or assuming a modified WiFi scheduler, and (ii) an asymptotic approximation which is accurate for real, throughput-fair WiFi schedulers, as flow sizes increase.

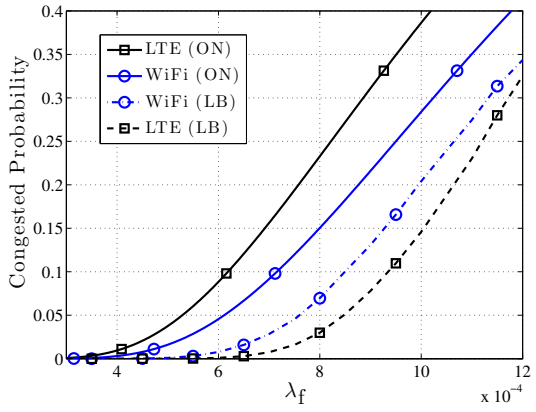


Fig. 6: Congestion probability w.r.t flow density λ_f

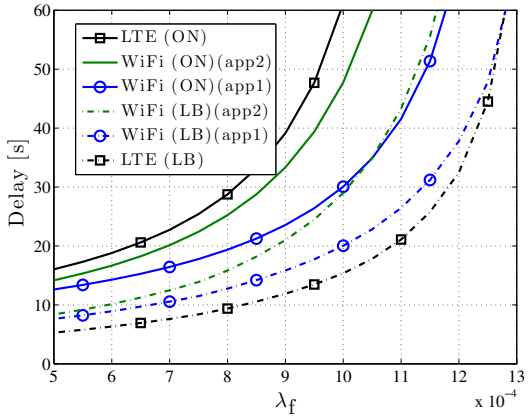


Fig. 7: Delay of each network w.r.t flow density λ_f

We stress here that the respective congestion probabilities are the same for both WiFi schedulers, as this only depends on the incoming job rate and the average service rate $\langle\mu\rangle$, which are the same in both cases, as we showed in Section III.

Looking at the mean delay, in Fig. 7, and comparing the two different cases of the WiFi schedulers, it is clear that resource-fair version of WiFi outperforms the throughput fair one, as expected. For the load-based case, when the network load causes 10% of congestion probability, the average flow delay of the throughput fair WiFi system is 25% higher compared to the resource fair one.

Focusing now on the saturated (Always ON) case, it seems somewhat surprising, at first, that the LTE network performs worse than resource fair WiFi, if we take into account that for the same SINR, LTE tends to operate with higher rate (See also [29]). The reason for this are the edge users. A number of users with low SINR that are regarded as "out of service" and not taken into consideration for the WiFi network, are instead covered by a similar density LTE network. E.g., for the always ON case, the coverage area was 0.67 and 0.47 for the LTE and WiFi networks respectively[5]. Hence, the "edge" users in a WiFi BS end up getting better rates than the "edge" users in LTE (This is also evident from Fig. 1, where the lowest MCS for WiFi - the lowest green triangle - provides higher rates than the lowest MCSs for LTE - the lowest blue crosses).

The above low values for the coverage area originate from two previous-mentioned worst case assumptions: (i) the random BS placement; in the PPP model, it is possible that a BS ends up asymptotically close to another; (ii) the interference is calculated assuming that neighboring BSs are saturated. By examining the load-based case, we notice how critical the second assumption is, as for low load values the coverage area was almost 1 for both RATs, while for a load around $\rho = 0.5$ coverage areas were 0.9 and 0.7 for LTE and WiFi, respectively.

Nevertheless, when one considers the more realistic case of load-based interference, LTE outperforms WiFi (both the throughput-fair and resource-fair versions). This is mainly due to the LTE's smaller granularity between the MCS, thus being able to better take advantage of the SINR improvement, due to lower average interference, for the same input traffic.

C. Cooperative 2-tier HetNets

In this last section, we move on to multi-tier HetNets which is the main focus of this paper. Here, we are interested in understanding the impact of coexistence of different RATs in orthogonal frequencies (the case of coexistence in the same band could also be handled with some modifications by our model, but is part of future work). Particularly, our goal is to capture the impact of different types of association criteria between different tiers, on the performance gains by introducing a 2nd tier. As mentioned before, the tier-association criteria of interest are:

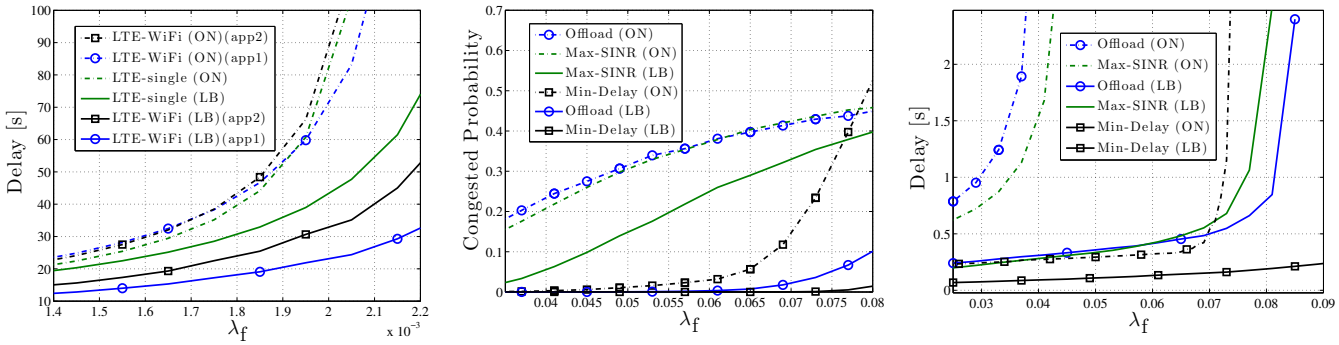
Off-load: This is the simplest (and most aggressive off-load) policy, where the user, if is able to establish connection with the WiFi network, does it without any further criterion.

Max-SINR: Here, the user choses to associate with the tier that provides the best SINR, thus attempting to improve his channel conditions in order to achieve the highest possible throughput.

Min-Delay: Here in this scenario, the user choses to associate with load related criteria in order to minimize the average delay of the system. In our case the association criterion, between tiers, is the modified version of [9] as proposed in Lemma 6.3.

First, we examine the simple Off-load policy. We assume that LTE is the primary network and WiFi the secondary one with the same density. Fig. 8 (a) presents two different cases of this 2-tier HetNet. The difference between those two cases is about the WiFi scheduler: one were the WiFi AP operate as an "ideal", resource-fair scheduler and one as throughput-fair. Additionally, for comparison reasons, we include as "baseline" plot a single-tier LTE network with double BS density (i.e. the total number of LTE BS is equal to the sum of LTE

[5]This is also the reason why we had to normalize the user densities to ensure that the *absolute* number of connected users per BS is the same for each network.



a: Comparison of the association schemes, w.r.t. flow density λ_f

b: Comparison of the association schemes, w.r.t. flow density λ_f

c: Comparison of the association schemes, w.r.t. flow density λ_f

Fig. 8: Cooperative association schemes

BS and WiFi AP in the other two scenarios). Interestingly, for the saturated case almost all scenarios perform the same. What is particularly surprising is that the Off-loading case performs almost the same as the single-tier LTE case: while the total density of the BSs is the same for both cases, the Off-loading scenario uses double the spectrum than the single-tier LTE scenario. In order to sketch the explanation we mention: 1) Off-load association does not affect the MCS distribution of each tier, 2) on the always ON interference the MCS distribution of each tier does not depend on the BS's density (the gain of the probabilistic higher received power is equal with the loss of higher interference). 3) In always ON case, WiFi captures slightly less than the 50% of traffic. Taking into account the aforementioned, the gain of the second tier is only that the cardinality of the users to the BS decreasing, just like the single-tier with higher BS density. The picture totally changes in the load based case, because the extra spectrum means less users per band and therefore, less interference.

Nevertheless, if we turn our attention to the load-based interference cases, we see that: (i) the WiFi scheduler highly affects the overall performance; (ii) the 2-tier network outperforms the LTE-only for both schedulers, which is more in-line with what we would have expected. This further underlines the importance of load-based analysis, which in this case not only has a quantitative, but also a clear qualitative impact.

For the rest of this section, we only consider a best case WiFi network (i.e., resource fair), in order to focus our attention on association policies, and understand the limits of performance improvements by introducing a WiFi tier. To be more realistic we assume now denser secondary network than the primary one. More precise, a secondary WiFi network with $\lambda_{WiFi} = 5$, and a primary LTE network with $\lambda_{LTE} = 1$. Congestion probability and per flow delay for different traffic input rates λ_f, are depicted in Fig. 8 (b) and (c), respectively.

Looking at the Off-load and Max-SINR criteria for the saturated case, the congestion probability of both cases is almost equal and Max-SINR performs better, with respect to mean delay. However, considering the load-based interference cases, the Off-load policy is much more robust with respect to congestion probability and outperforms Max-SINR with respect to delay, as well.

This discrepancy between the saturated and load-based cases originates from fact that saturated analysis is able to capture only one side of the gain stemming from increasing network density. On the one hand, the saturated case correctly captures the fact that an "arbitrary" BS on average has to serve fewer users, in a denser network (thus dealing with a smaller ρ due to a decrease in the numerator, i.e., the input traffic). On the other hand, it fails to capture that the surrounding BSs will be less loaded as well, and therefore cause less interference, which in turn, leads to even better performance for the (fewer) users served (due to an increase in $\langle \mu \rangle$ and a resulting further decrease in ρ). As a consequence, the saturated model underestimates association schemes that tend to utilize the denser (WiFi) network more.

Last but not least, it is clear from Fig. 8 (c) that the Min-Delay association policy significantly outperforms the other two, by up to an order of magnitude or more, for high loads. Unlike the other two policies that only consider PHY layer performance (MAX-SINR) or naively try to reduce the load of the primary network (Off-load), Min-Delay directly takes into account the actual load experienced, which plays the key role on the per-flow performance. It is also interesting to note that, especially for low loads, the Min-Delay policy is also quite stable in terms of congestion probability Fig. 8 (b). While the considered association policies are admittedly abstracted version of real detailed policies, we believe these results make a strong case for sophisticated, load-based association mechanisms in future HetNets, in order to better balance the loads between tiers and ensure the best user experience.

VIII. Conclusion / Future work

We have presented an analytical framework for an accurate prediction of the flow-level performance of a large randomly placed network. This analysis considers both the case of always ON interfering neighboring BSs, as well as the case of load-depended interference. It turns out that the performance gap between the aforementioned cases could be rather high, affecting not only quantitative insights, but often qualitative conclusions as well, and thus should be carefully taken into account during network design. Additionally, we have considered multi-tier topologies, modeling some common association

criteria, and evaluating their impact on both user- and network-centric performance. This initial study was not meant to be complete, given the large range of parameters and degrees of freedom in such multi-tier scenarios, but rather to provide some initial representative insights, and demonstrate the utility of our proposed framework.

Three general conclusions should be mentioned: i) The scheduling policy could strongly effect system's flow-level performance, even if the PHY characteristics are the same. ii) The two different interference approaches, always ON and load-based, change totally the performance of the system (single-tier or multi-tier), so we should be very careful about this assumption when model a system. iii) The gain of the load related association policy is surprisingly high comparing to the more traditional ones (Off-load Max-SINR).

As future work, we plan to apply our framework, together with different association criteria, in Carrier Aggregation scenarios. Additionally, as mentioned earlier, we believe our framework could be modified to analyze scenarios of LTE and WiFi coexistence, in the same band, as well as to consider advanced LTE-A features such as Cell Range Expansion and Almost Blank Subframes which are expected to play a key role in future HetNets.

REFERENCES

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2014–2019," *Whitepaper*, 2015.

[2] A. Ghosh, N. Mangalvedhe, R. Ratasuk, B. Mondal, M. Cudak, E. Visotsky, T. Thomas, J. Andrews, P. Xia, H. Jo, H. Dhillon, and T. Novlan, "Heterogeneous cellular networks: From theory to practice," *Comm. Magazine, IEEE*, 2012.

[3] J. Andrews, S. Singh, Q. Ye, X. Lin, and H. Dhillon, "An overview of load balancing in HetNets: old myths and open problems," *IEEE TWC*, 2014.

[4] T. Bonald and A. Proutiere, "Wireless downlink data channels: User performance and cell dimensioning," in *ACM MOBICOM*, 2003.

[5] Nokia, "5G use cases and requirements," *White Paper*, 2015.

[6] Ericsson, "5G radio access," *White Paper*, 2015.

[7] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," *IEEE/ACM ToN*, 2005.

[8] T. Bonald and J. Roberts, "Scheduling network traffic," in *ACM SIGMETRICS*, 2007.

[9] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Distributed α -optimal user association and cell load balancing in wireless networks," *IEEE/ACM ToN*, 2012.

[10] M. Karray and M. Jovanovic, "A queueing theoretic approach to the dimensioning of wireless cellular networks serving variable bit-rate calls," *IEEE TVT*, 2013.

[11] B. Błaszczyszyn, M. Jovanovic, and M. K. Karray, "How user throughput depends on the traffic demand in large cellular networks," in *WIOPT-SPASWIN*, 2014.

[12] F. Baccelli, B. Blaszczyszyn, and F. Tournois, "Spatial averages of downlink coverage characteristics in CDMA networks," in *INFOCOM IEEE*, 2002.

[13] J. Andrews, F. Baccelli, and R. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE TCOM*, 2011.

[14] H. Dhillon, R. Ganti, F. Baccelli, and J. Andrews, "Modeling and analysis of K-tier downlink heterogeneous cellular networks," *IEEE JSAC*, 2012.

[15] X. Lin, J. Andrews, and A. Ghosh, "Modeling, analysis and design for carrier aggregation in heterogeneous cellular networks," *IEEE TCOM*, 2013.

[16] S. Singh, H. Dhillon, and J. Andrews, "Offloading in heterogeneous networks: Modeling, analysis, and design insights," *IEEE TWC*, 2013.

[17] H. Dhillon, M. Kountouris, and J. Andrews, "Downlink coverage probability in MIMO HetNets," in *ASILOMAR*, 2012.

[18] T. Novlan, R. Ganti, A. Ghosh, and J. Andrews, "Analytical evaluation of fractional frequency reuse for heterogeneous cellular networks," *IEEE TCOM*, 2012.

[19] G. Zhang, Q. S. Quek, A. Huang, M. Kountouris, and H. Shan, "Backhaul-aware base station association in two-tier heterogeneous cellular networks," *SPAWC*, 2015.

[20] H. Dhillon, R. Ganti, and J. Andrews, "Load-aware modeling and analysis of heterogeneous cellular networks," *IEEE TWC*, 2013.

[21] I. Siomina and D. Yuan, "Analysis of cell load coupling for LTE network planning and optimization," *IEEE TWC*, 2012.

[22] S. Aalto, U. Ayesta, S. Borst, V. Misra, and R. Núñez Queija, "Beyond processor sharing," in *ACM SIGMETRICS*, 2007.

[23] T. Bonald and A. Proutière, "On performance bounds for the integration of elastic and adaptive streaming flows," in *ACM SIGMETRICS*, 2004.

[24] Y. Zhong, M. Haenggi, T. Q. S. Quek, and W. Zhang, "On the stability of static poisson networks under random access," *IEEE TCOM*, 2016.

[25] O. Galinina, S. Andreev, M. Gerasimenko, Y. Koucheryavy, N. Himayat, S. P. Yeh, and S. Talwar, "Capturing spatial randomness of heterogeneous cellular/WLAN deployments with dynamic traffic," *IEEE JSAC*, 2014.

[26] B. Błaszczyszyn, M. Jovanovic, and M. K. Karray, "Performance laws of large heterogeneous cellular networks," in *WIOPT*, 2015.

[27] Y. Zhong, T. Q. S. Quek, and X. Ge, "Heterogeneous cellular networks with spatio-temporal traffic: Delay analysis and scheduling," *IEEE JSAC*, 2017.

[28] M. Harchol-Balter, *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*.

[29] G. Arvanitakis and F. Kaltenberger, "Stochastic analysis of two-tier HetNets employing LTE and WiFi," in *EuCNC*, 2016.

[30] S. Sesia, I. Toufik, and M. Baker, *LTE - the UMTS long term evolution : from theory to practice*. Wiley, 2009.

[31] F. Capozzi, G. Piro, L. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey," *Communications Surveys Tutorials, IEEE*, 2013.

[32] G. Fayolle, I. Mitrani, and R. Iasnogorodski, "Sharing a processor among many job classes," *J. ACM*, 1980.

[33] M. Heusse, F. Rousseau, G. Berger-Sabbatel, and A. Duda, "Performance anomaly of 802.11b," in *INFOCOM IEEE*, 2003.

[34] A. Karr, *Probability*. Springer, 1993.

[35] K. Avtachenkov, U. Ayesta, P. Brown, and R. Núñez Queija, "Discriminatory processor sharing revisited," *INFOCOM*, 2005.

[36] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE JSAC*, 2000.

[37] Y. Lin and V. Wong, "Frame aggregation and optimal frame size adaptation for IEEE 802.11n WLANs," in *GLOBECOM IEEE*, 2006.

[38] A. F. Molisch, *Wireless Communications*. Wiley, 2010.

[39] P. Fotiadis, M. Polignano, L. Chavarria, I. Viering, C. Sartori, A. Lobinger, and K. Pedersen, "Multi-layer traffic steering," in *VTC IEEE*, 2013.

[40] N. Sapountzis, S. Sarantidis, T. Spyropoulos, N. Nikaein, and U. Salim, "Reducing the energy consumption of small cell networks subject to QoE constraints," in *GLOBECOM IEEE*, 2014.

[41] G. Arvanitakis and F. Kaltenberger, "PHY layer modeling of LTE and WiFi RATs," Eurecom, Tech. Rep. RR-16-317.

[42] S. M. Yu and S. L. Kim, "Downlink capacity and base station density in cellular networks," in *WiOpt*, 2013.

[43] G. Arvanitakis, T. Spyropoulos, and F. Kaltenberger, "An analytical model for flow-level performance of large, randomly placed small cell networks," *GLOBECOM IEEE*, 2016.

[44] G. Arvanitakis and F. Kaltenberger, "Energy vs QoE tradeoff of dense mobile networks," in *submited to ICNC 2018*.

[45] *LTE Specifications, http://www.3gpp.org/DynaReport/36-series.htm*.

[46] *802.11 Specifications, http://standards.ieee.org/about/get/802/802.11.html*.

[47] *OpenAir Interface, www.openairinterface.org*.

Arvanitakis George holds a Diploma in Physics from the National Kapodistrian University of Athens, Greece, and a MSc. in Radio Engineering from the same university. He is currently a Ph.D Student at EURECOM / Telecom-ParisTech.

Thrasyvoulos Spyropoulos received the Diploma in Electrical and Computer Engineering from the National Technical University of Athens, Greece, and a Ph.D degree in Electrical Engineering from the University of Southern California. He was a post-doctoral researcher at INRIA and then, a senior researcher with the Swiss Federal Institute of Technology (ETH) Zurich. He is currently an Assistant Professor at EURECOM, Sophia-Antipolis. He is the recipient of the best paper award in IEEE SECON 2008, and IEEE WoWMoM 2012, and runner-up award for ACM MobiHoc 2011, and IEEE WoWMoM 2015.

Florian Kaltenberger (S05-M08) received his Dipl.-Ing. degree and his Ph.D. degree both in technical mathematics from the Vienna University of Technology in 2002 and 2007, respectively. He is an assistant professor at the Communication Systems Department of Eurecom, Sophia-Antipolis, France, and part of the management team of the real-time open-source 5G platform OpenAirInterface.org. From 2003 to 2007 he was with the Wireless Communications Group of the Austrian Research Centers, where he was developing a real-time MIMO channel emulator. His research interests include 5G and MIMO systems at large, software defined radio, signal processing for wireless communications, as well as channel modeling and simulation. Dr. Kaltenberger is a co-chair of the experimental radio access working group of the COST IRACON action. He received the Neal Shepherd best propagation paper award in 2013, best project award for the Celtic-Plus project SPECTRA in 2015, and best demo award for the project ADEL at the European Conference on Networks and Communications in 2016.