On Flow Level Modeling of Multi-Cell Wireless Networks

Albrecht J. Fehske and Gerhard P. Fettweis
Vodafone Stiftungslehrstuhl, Technische Universität Dresden
Email: {albrecht.fehske, fettweis}@ifn.et.tu-dresden.de

***Abstract*—In order to meet increasing traffic demands, future generations of cellular networks are characterized by decreasing cell sizes at full frequency reuse. Due to inevitable inter-cell interference, load conditions in neighboring cells can no longer be considered independent. This paper provides a flow level modeling framework for cellular networks, where the coupling of flow level dynamics due to intercell interference is specifically taken into account. Since an adequate queueing model renders analytically intractable, we review different methods from the literature to bound and approximate the stationary behavior of network performance measures. Numerical investigations of a typical wireless scenario reveal, that in high and low load regimes first as well as second order bounds may be quite loose, depending on the type of bound. Especially for design of network optimization algorithms, bounds do not appear to suitably reflect network performance, and approximation techniques must be considered instead. In this regard, a suitable tradeoff between computational complexity and accuracy over the whole traffic range is provided by a model based on the notion of average interference.**

I. Introduction

Flow level models that capture the dynamic nature of arrivals and service periods of data requests are crucial to characterize the performance of cellular networks above the physical layer. In such models, a cell is commonly represented by a queue and the main object of study is a random process describing the number of data flows being serviced in a cell at any point in time.

The ability of these techniques to characterize not just flow delays and throughputs but also utilization and energy consumption of wireless base stations (BSs) statistically over longer periods of time appears particularly useful in the area of self-organizing networks (SON), where these KPIs are addressed by different use cases. Several applications of flow level models to SON-type optimization of cellular systems are already reported in the literature. Among others, the authors of [1] use them to maximize the cell capacity via adaptation of pilot powers. In [2], [3], Kim et al. provide a framework to selectivly optimize either flow throughput, flow delay, BS energy consumption, or maximal BS loads via adjustment of the user association policy. This framework is extended to address said KPIs via joint optimization of antenna tilts and the user association rule in [4].

Scenarios with multiple interfering cells are naturally represented by a network of queues where the number of active flows is a vector process. Unfortunately, the adequate queueing model, a so called *processor sharing model*, renders analytically intractable. Since analytical or straightforward numerical computation is not viable, most studies tend to ignore the coupling effect by assuming that data flows always observe some given *constant* interference during their service. Such an assumption leads to a network model with independent queues. These assumptions, however, are accurate only in a fully loaded or a completely empty system. Realistically, each flow observes only a single interference condition and, as a consequence, the service rates in each cell depend on the network state, i.e., the number of active flows in all other cells.

In the context of wireless networks, these observations are first made by Bonald et al. in [5], where the authors provide techniques to bound and approximate average flow throughputs in individual cells based on minimal and maximal interference assumptions in all other cells. Additionally, numerical studies demonstrate that the constant interference assumption, as mentioned above, provides only very coarse approximation of the actual flow throughput. A later work [6], applies the concept of *aggregation of variables* to the same multi-cell scenario to obtain very accurate approximation of the BS utilizations. Since this accuracy comes at the cost of considerable complexity, the same paper also provides a second approximation technique based on the idea that flows observe *average interference*. The latter model is indepently proposed in [7], the only difference being that the algorithm proposed to compute BS utilizations is considerably more cumbersome compared to the one proposed in [6].

This paper consolidates the work presented in above references. The contributions are threefold. First we briefly overview and discuss a framework to represent radio networks as multi-cell flow level models in Section III and Section IV. Based on which in Section IV, we present four approaches to approximate flow level KPIs in multi-cell scenarios. We then compare the accuracy of these approximation techniques to results obtained from simulation of a wireless network and discuss tradeoffs between different methods in Section V and Section VI.

II. Preliminaries: Traffic Modeling

Data traffic can be categorized as being either *streaming* or *elastic* traffic. The latter typically originates from transferring digital documents like web pages or different types of files being downloaded. In contrast to streaming, the transmission rate of elastic traffic can be adapted to network conditions.

Here, we assume all traffic to be elastic. This assumption reflects the traffic situation today, which is dominated by TCP controlled data transfers. It is anticipated, that streaming video applications will rapidly develop and have a much larger share in the future [8]. Extension of the framework presented here to include streaming-type traffic is left for future work.

Flow Level versus Packet Level

The dynamics of elastic internet traffic at packet level are notoriously difficult to analyze. Their temporal statistics exhibit self-similarity and multi-fractal behavior, which are induced by the heavy tailed distribution of document sizes as well as the mechanisms of TCP congestion control (see, e.g. [9], [10] and references therein). As a result, network performance on packet level is hardly tractable. In particular, established models from queueing theory, which require (at least) the assumption of Poisson arrival processes, are not directly applicable.

In [11], [12], Bonald et al. put forward a traffic modeling paradigm, that considers data traffic on *flow* and *session* rather than on packet level. In this regard, a flow represents a continuous stream of packets pertaining to a particular content, like a web page or any kind of file. A session is a collection of flows whose statistical properties are independent of flows of another session. In particular, a session is associated with an individual user. It is further assumed that users generate sessions independently and no user accounts for an excessive amount of sessions.

Performance Evaluation

A key observation is, that users perceive network performance on flow and session, rather than on packet level and, consequently, performance modeling is most naturally done on flow level.

For a reasonably sized user population, session arrivals resemble a Poisson process. The session structure in terms of number, size, and inter-arrival time of flows, however, does inherit the more complex statistical properties from the packet level. It is argued in [11], [12] that, despite these intricacies, the system is well represented by a Markovian model, as long as session arrivals can be considered Poisson. A prerequisite is that transmission resources are shared somewhat equally among contending flows. At least partially, we can observe the validity of this result in Section VI, where quite a close approximation of a system with heavy tailed service time distribution is achieved by Markovian models.

III. Radio Network Modeling

In the following, we discuss major assumptions regarding the representation of a wireless network as a flow level model.

A. Network Layout

Throughout the paper, we consider the downlink of a cellular network consisting of N base stations covering a compact region $\mathcal{L} \subseteq \mathbb{R}^2$. BS locations and types can be perfectly arbitrary. We assume users to be spatially distributed according to some distribution $\delta(\cdot)$ with $\int_{\mathcal{L}} \delta(u)\,\mathrm{d}u = 1$. Users can in principle be mobile with certain restrictions as explained subsequently. We denote the serving area or cell of BS i by $\mathcal{L}_i \subset \mathcal{L}$.

B. Traffic Model

For the analytical part, we assume that the arrival of flow requests to the network takes place according to a Poisson process with intensity λ. The overall intensity is then split among individual cells according to their associated coverage area. Flow sizes are assumed to be exponentially distributed with common mean denoted by Ω. The terms λ, Ω, and $\delta(u)$ determine the traffic intensity distribution $\sigma(u) := \lambda\Omega\delta(u)$ in Mbps per km^2, which we use in the remainder of the paper.

C. Radio Link and Resource Sharing

The radio link quality generally depends on the users' distance to the serving and the surrounding base stations, on the number of active BSs generating interference, as well as their transmit powers.

The data rates, in addition, depend on the amount of transmission resources allocated to a particular flow. In this regard, we make the following important assumptions.

Assumption 1 (Resource Sharing)**.** We assume that transmission resources are shared evenly among all flows in the cell. Further, if there is at least one active flow within a cell, the corresponding base station utilizes all available transmission resources and thus decidedly transmits at full power.

We call a BS *active* if it serves at least one flow. We denote by $y \in \mathcal{Y} := \{0,1\}^N$ a vector with $y_i = 1$ if BS i is active and $y_i = 0$ otherwise. Further, we collect the indices of inactive and active BS for each y in the sets

$$\mathcal{N}_0(y) := \left\{ i \in \mathbb{N}_N \,|\, y_i = 0 \right\}, \tag{1}$$

$$\mathcal{N}_1(y) := \left\{ i \in \mathbb{N}_N \,|\, y_i = 1 \right\}. \tag{2}$$

In addition, we define the set $\mathcal{A}_i$, collecting all vectors y for which BS i is active:

$$\mathcal{A}_i := \{ y \in \mathcal{Y} \,|\, y_i = 1 \}. \tag{3}$$

Assuming BS i to be active, i.e., $y_i = 1$ the signal-to-interference-and-noise-ratio (SINR) and the data rate achievable at location $u \in \mathcal{L}_i$ are given by

$$\gamma_i(u,y) = \frac{p_i(u)}{\sum_{\substack{j \in \mathcal{N}_1(y) \\ j \neq i}} p_j(u) + \theta} \quad \text{and} \tag{4a}$$

$$c_i(u,y) = a \cdot w \cdot \log_2\big(1 + b \cdot \gamma_i(u,y)\big), \tag{4b}$$

where θ, p_i, and w denote the noise power, the power received from BS i at location u, and the bandwidth available for transmission, respectively. The inclusion of all path loss and fading related effects is discussed subsequently. The purpose of parameters a and b is explained in Section III-C4 below.

1) Fast and Slow Fading: We presume here, that serving a data flow takes much longer than the coherence time of a wireless channel, and thus, data flows observe fast fading by its average. Similarly, we presume that shadowing effects happen on a much larger time scale and are constant over the duration of many flows. Consequently, we assume fast and slow fading effects to be contained in the location-dependent terms $p_i(\cdot)$.

2) User Mobility and Handovers: While users can be mobile in principle, here we require that the received powers $p_i(\cdot)$ do not change during a flow duration. Flow durations are usually rather short periods, e.g., less than a second, and path gains can be assumed to be constant over radii of few meters [13]. As a consequence, user mobility, even though not necessarily zero, is restricted to a few meters per second, i.e., typical pedestrian speed. Such "quasi-stationarity" is fairly realistic because as of today about 80% of all data traffic originates from indoor locations (e.g., [14]).

Handover events are commonly triggered by user mobility and the slow fading process. Since both happen on a much larger time scale than a typical flow duration, we do not model handover events explicitly here and assume, that a flow remains connected to a single serving base station during its entire lifetime.

3) Fast Packet Scheduling: Following the approach in [15], we incorporate an *average* packet scheduling gain into the model via the parameters a and b, i.e., choosing larger parameters for a more spectrally efficient scheduling mechanism. The main reason for doing so is simplicity. Since fast scheduling mechanisms explicitly adapt to fast fading conditions, the approach is justified when flow durations are much longer than the channel coherence time. In this case, each flow experiences the effects of fast fading and fast scheduling only by their averages.

4) Adapting the Link Model with Parameters a and b: The parameters a and b in Eq. (4b) are used to further tailor the data rate achievable with a certain SINR γ_i and bandwidth w to the system under study. The same model is already proposed by Mogensen et al. in [15] and used to capture – for instance – the effects of packet scheduling (as discussed above), MIMO techniques, or system specific overheads, which individually increase or decrease the average bitrates. In this regard, we can think of the products aw and $b\gamma_i$ as the effective bandwidth and effective SINR, respectively.

IV. Flow Level Dynamics

Based on the assumptions regarding the radio network, this section defines the corresponding flow level models considered.

A. A Single Base Station

Consider a network of N base stations and let $X_i(t) \in \mathbb{N}_0$ denote the number of flows in some cell i at time t. With the above definitions, we can understand $X_i(t)$ as a continuous time random process. For a fixed vector y of active BSs and considering all points in its serving area, we can represent a single BS by an M/M/1 PS queuing model (e.g., [16]). The mean serving rate for flows at BS i is given as

$$\mu_i(y) = \frac{C_i(y)}{\Omega} \quad \text{with} \quad C_i(y) := \left(\int_{\mathcal{L}_i} \frac{\delta_i(u)}{c_i(u,y)} \, \mathrm{d}u \right)^{-1}, \tag{5}$$

where $\delta_i(u) := \frac{\delta(u)}{\int_{\mathcal{L}_i} \delta(u) \, \mathrm{d}u}$ denotes the distribution of users conditioned on being in cell i. For each i the utilization of the queue, for given $y > 0$, is defined as

$$\rho_i(y) := \min\left(\frac{\lambda_i}{\mu_i(y)}, \, 1 \right) = \min\left(\frac{\lambda_i \Omega}{C_i(y)}, \, 1 \right). \tag{6}$$

The term y indicates which BSs are transmitting and thus represents the *interference scenario*. Focussing on the dynamics of a single BS in this section, we assume y to be fixed, i.e., other BSs are either always active or always inactive, depending on the choice of y. Subsequent sections explicitely address the case when y is not fixed but a random process itself.

B. Flow Level KPIs Considered

Before proceeding to analyze the dynamics of coupled BSs, this section introduces flow level performance metrics. Since subsequent sections introduce different techniques that reduce the analysis of mulitple interfering BSs to variants of an M/M/1 PS system described above, we provide concrete KPI definitions for the M/M/1 PS queueing model. Derivations of these definitions can be found in standard textbooks such as [16]. As before, we assume some interference scenario y to be fixed.

With Assumption 1, the *average utilization of time and frequency resources* of BS i is equivalent to the probabilitiy that there is at least one flow in cell i and is stated in Eq. (6). The utilization itself may not be of direct interest, however, a variety of flow level metrics (for instance all metrics considered here) are strictly monotonic functions of the utilization. For that reason, the BS load itself is a basic quantity of this study.

Let T_i denote the random overall time that a flow spends in the cell i from its arrival until the completion of service. The *average sojourn time* of flows in cell i is given as the expectation of T_i, i.e.,

$$\tau_i(y) := \mathrm{E}(T_i) = \frac{\Omega}{\big(1 - \rho_i(y)\big) C_i(y)}. \tag{7}$$

The *average flow throughput* is defined as the ratio of average flow size and average sojourn time, i.e.,

$$r_i(y) := \frac{\Omega}{\mathrm{E}(T)} = \big(1 - \rho_i(y)\big) C_i(y). \tag{8}$$

Note that, strictly speaking, the average flow throughput is the expectation of the ratio $\frac{S}{T}$, i.e., $\mathrm{E}\left(\frac{S}{T}\right)$, where S denotes the random flow size. Since this latter expectation is difficult to obtain, r_i is commonly used for performance evaluation (see [17] for a more detailed discussion).

C. Multiple Interfering Base Stations

In the following, we extend the flow level model of a single access point as presented in the previous Section IV-A to the case of several interfering BSs. We make the same assumptions regarding arrival and service time processes as before.

Let $X(t) := \big(X_1(t), \ldots, X_N(t)\big) \in \mathcal{S} := \mathbb{N}_0^N$ for $t \geqslant 0$ denote the vector process of the number of flows in N cells with state transition rates

$$q(x, x') = \begin{cases} \lambda_i & \text{for } x' = x + e_i, \\ \mu_i(y) & \text{for } x' = x - e_i, \\ 0 & \text{else,} \end{cases} \tag{9}$$

where the term e_i denotes an N-dimensional vector with the i'th component equal to one and all other components equal to zero. The vector $x \in \mathcal{S}$ collects the number of flows in all cells.

Note that using Assumption 1, we can identify the interference scenario y with the sign of the state x, i.e., $y = \operatorname{sgn}(x)$. In the following, let $\sigma(y)$ denote the probability of finding the network in state y. Let further ρ_i denote the utilization of BS i with

$$\rho_i = \sum_{y \in \mathcal{A}_i} \sigma(y). \tag{10}$$

1) Tractability of the Stationary Behavior: The service rates in Eq. (9) vary among states, which implies that the queueing network is not partially reversible. Techniques to obtain a product form stationary distribution (from which $\sigma(y)$, ρ_i and other performance metrics could be derived) as proposed, e.g., in [18], are not applicable here. In fact, in [19], Fayolle et al. observe that, in case of two queues a product form solution exists if and only if $\mu_1(0,1)+\mu_2(1,0) = \mu_1(1,1) + \mu_2(1,1)$. In case of a wireless network, this condition requires that the service rate provided by one of the BSs must be larger in case of interference, which is contradictory to the definition of the SINR and service rate in Eq. (4) and Eq. (5), respectively.

D. Performance Bounds

Following the framework proposed by Bonald et al. in [5], this section defines first and second order performance bounds on the average BS resource utilization. Note that in [5], the authors focus on the corresponding bounds for the flow throughput, rather than the resource utilization itself.

In the following, *lower bound on performance* means an actual lower bound on the flow throughput, but an upper bound on the BS utilization and the sojourn time. Similarly, *upper bound on performance* means an actual upper bound on the flow throughput but a lower bound on the BS utilization and flow sojourn time.

1) First order bounds: First order lower and upper performance bounds on the utilization of BS i are obtained by assuming that the process X_i of the number of flows in cell i evolves like in an M/M/1 PS system (i.e., independently of all processes X_j) under the best case ($y = 0$) and worst case ($y = 1$) interference scenarios, respectively. Specifically, first order upper and lower performance bounds for the utilization of BS i are given as

$$\hat{\rho}_i' := \min\left(\frac{\lambda_i \Omega}{C_i(0)},\ 1\right) \quad \text{and} \quad \check{\rho}_i' := \min\left(\frac{\lambda_i \Omega}{C_i(1)},\ 1\right),$$

respectively. Above, $C_i(0)$ and $C_i(1)$ denote the capacity of cell i from Eq. (5) when all other BSs are always idle and always active, respectively. The corresponding expressions for flow sojourn times and flow throughputs are obtained by inserting $\rho_i(0)$, $C_i(0)$, $\rho_i(1)$ and $C_i(1)$, into Eqs. (7) and (8), respectively.

2) Second Order Bounds: Second order performance bounds on the utilization of BS i are obtained by assuming that the process X_i depends on the state of all other processes X_j, with $j \neq i$, however, the latter evolve independently under either best case or worst case interference conditions. In addition, the processes X_j are assumed to evolve either much faster or much slower than the process X_i, which, in the limit, leads to the so-called *fluid regime* and *quasi-stationary regime* for the processes X_j, respectively. It is shown in [20] that network performance is overestimated by the former and underestimated by the latter.

a) Second Order Upper Bound: Let $\hat{\sigma}_i(y)$ denote the probability that BS i sees interference scenario y, assuming all processes X_j evolve independently and without any inter cell interference, which is given as

$$\hat{\sigma}_i(y) := \prod_{\substack{j \in \mathcal{N}_0(y) \\ j \neq i}} \big(1 - \rho_j(0)\big) \prod_{\substack{j \in \mathcal{N}_1(y) \\ j \neq i}} \rho_j(0). \tag{11}$$

In the fluid regime, the data rate achievable at any location $u \in \mathcal{L}_i$ is given as the average with respect to the random interference scenario y, which leads to the second order performance upper bound on the utilization

$$\hat{\rho}_i'' := \min\left(\frac{\lambda_i \Omega}{\widehat{C}_i},\ 1\right) \quad \text{with} \tag{12a}$$

$$\widehat{C}_i := \left(\int_{\mathcal{L}_i} \frac{\delta_i(u)}{\sum_{y \in \mathcal{A}_i} c_i(u,y)\, \hat{\sigma}_i(y)}\, \mathrm{d}u\right)^{-1}. \tag{12b}$$

Second order upper bounds on the sojourn times and flow throughputs are obtained by inserting both $\hat{\rho}_i''$ and $\widehat{C}_i$ into Eqs. (7) and (8).

b) Second Order Lower Bound: Replacing $\rho_j(0)$ with $\rho_j(1)$ in Eq. (11) yields the probability that BS i sees the interference scenario y assuming that all other cells permanently see maximum interference. Let us denote this probability by $\check{\sigma}_i(y)$. As opposed to the fluid regime, in the quasi-stationary case the interference scenario evolves very slowly such that the process X_i reaches stationarity before other BSs changes states. As a result, the utilization of BS i and the sojourn time are defined as the arithmetic mean with respect to the distribution $\check{\sigma}_i(y)$. Due to its definition, the flow throughput is obtained as the corresponding harmonic mean. For $\rho(1) < 1$,

we write utilization, sojourn time and throughput as:

$$\check{\rho}_i'' := \min\left(\sum_{y\in\mathcal{A}_i}\rho_i(y)\check{\sigma}_i(y),\ 1\right), \tag{13a}$$

$$\check{\tau}_i'' := \sum_{y\in\mathcal{A}_i}\tau_i(y)\check{\sigma}_i(y), \quad \text{and} \tag{13b}$$

$$\check{r}_i'' := \left[\sum_{y\in\mathcal{A}_i} r_i(y)^{-1}\check{\sigma}_i(y)\right]^{-1}, \tag{13c}$$

respectively. The set $\mathcal{A}_i$ is defined in Eq. (3). For $\rho_i(1) = 1$ the KPIs are given by

$$\check{\rho}_i'' := 1, \quad \check{\tau}_i'' := \infty, \quad \text{and} \quad \check{r}_i'' := 0. \tag{13d}$$

E. Second Order Approximations of Flow Level Performance

As stated before, the fluid and quasi-stationary regimes are known to respectively overestimate and underestimate the performance of a queueing system [20]. Thus, pairing the best and worst case interference assumptions with the fluid and quasi-stationary regimes, respectively, lead us to upper and lower bounds on performance as explained above.

In addition to these bounds, *approximations* of the BS utilization, flow throughput, and sojourn time are obtained by cross-pairing the fluid and quasi-stationary regimes with full and zero interference assumptions, respectively. Based on the definitions in the previous section, these approximations are obtained by interchanging the probabilities $\hat{\sigma}$ and $\check{\sigma}$ in Eqs. (13) and (12).

F. Approximating Flow Level Performance via Aggregation

Aggregation of variables is a versatile tool to reduce the complexity of analyzing systems with a large state space. While studying complex interactions in economics, Simon and Ando lay foundations for these principles in [21]. The general idea behind aggregation is to decompose the overall state space into groups or aggregates where strong interactions occur, and then characterize transitions within and amongst aggregates separately.

1) Partitioning the State Space: With the definition $y := \operatorname{sgn}(x)$, the state-dependent transition rates $\mu_i(\cdot)$ are given in Eq. (5). These rates are state dependent in general, but are equal in states where *the same* BSs are active and are strictly smaller in states where *additional* BSs are active, i.e.,

$$\begin{aligned} \operatorname{sgn}(x) = \operatorname{sgn}(x') \quad &\Longrightarrow \quad \mu_i(x) = \mu_i(x'), \\ \operatorname{sgn}(x) > \operatorname{sgn}(x') \quad &\Longrightarrow \quad \mu_i(x) < \mu_i(x'), \end{aligned} \tag{14}$$

where, the inequality is taken component wise.

We now partition the set of all possible states $\mathcal{S}$ into 2^N disjoint subsets by grouping states corresponding to the same interference scenario, i.e.,

$$\mathcal{S}(y) := \big\{x \in \mathbb{N}_0^N \mid \operatorname{sgn}(x) = y\big\}.$$

Considering the sets $\mathcal{S}(y)$, we can think of y as representing a collection or an *aggregate* of states $x \in \mathcal{S}(y)$. It follows from Eqs. (14) that, conditioned on being in one of the states in any aggregate $\mathcal{S}(y)$, the network essentially behaves as a network of independent queues. This property of the process X can be exploited to approximate the stationary behavior of the process $X(t)$ by considering transitions within and between aggregates separately.

2) Approximate Aggregate Probabilities: Let the terms $\tilde{\sigma}(y)$ denote the approximate probabilities of being in an aggregated state y. Following the derivations in [6], these probabilities are obtained as solutions to the system

$$P\tilde{\sigma} = 0, \tag{15}$$

where, for any ordering $i \mapsto y(i)$, the elements of the matrix $P = [p_{ij}]$ are defined as

$$p_{ij} := \begin{cases} \lambda_n & \text{for } y(i) = y(j) + e_n, \\ \max\Big[\mu_n\big(y(j)\big) - \lambda_n,\ 0\Big] & \text{for } y(i) = y(j) - e_n, \\ 0 & \text{else,} \end{cases}$$

and $\tilde{\sigma}$ denotes a vector collecting all probabilities $\tilde{\sigma}(\cdot)$ in corresponding order.

3) A Fluid Approximation With Aggregate Probabilities: The partitioning of the state space inherently presumes a quasi-stationary setup, in which an infinite number of flows pass through any active BS before the aggregated state y changes again. Numerical investigations by Fischer et al. in [22] show, that a finer approximation of flow level KPIs is obtained by considering the probabilities $\tilde{\sigma}$ in the fluid regime, where transitions between aggregates y do not happen infinitively slowly but arbitrarily fast.

Consequently and analogous to the fluid approximations befoseen earlier, we define the utilization of any BS i for the aggregation model as the ratio

$$\tilde{\rho}_i = \frac{\lambda_i \Omega}{\widetilde{C}_i} \quad \text{with} \tag{16}$$

$$\widetilde{C}_i(y) := \left(\int_{\mathcal{L}_i} \frac{\delta_i(u)}{\sum_{y\in\mathcal{A}_i} c_i(u,y)\,\tilde{\sigma}(y)}\,\mathrm{d}u\right)^{-1}. \tag{17}$$

The corresponding expressions for flow sojourn time and throughput are then obtained from Eqs. (7) and (8).

G. An Approximation Based on Average Interference

The link model in Eqs. (4), (5), and (6) provides different BS utilizations for each aggregated state y, corresponding to all possible interference conditions. This section considers a link model and the respective utilization, which occurs when data flows experience the *average interference over all aggregates* y. Assuming average interference, we write the SINR as

$$\bar{\gamma}_i(u,\sigma) := \frac{p_i(u)}{\sum_{y\in\mathcal{Y}}\sigma(y)\sum_{\substack{j\in\mathcal{N}_1(y)\\ j\neq i}} p_j(u) + \theta},$$

where σ denotes a vector collecting the $\sigma(y)$ in some order. By re-ordering the terms in the denominator and using Eq. (10) we can express the mean SINRs as functions of $\rho = (\rho_1, \ldots, \rho_N)^{\mathrm{T}}$ in the form

$$\bar{\gamma}_i(u,\rho) := \frac{p_i(u)}{\sum_{j\neq i} p_j(u)\rho_j + \theta}. \tag{18}$$

Based on mean interference, the SINRs can be expressed as functions of N BS utilizations ρ_i instead of 2^N aggregate probabilities $\sigma(y)$. Moreover, if exposed to average interference, BSs behave as a network of independent queues and all utilizations are given by the expression

$$f_i(\rho) = \min\left(\frac{\lambda_i \Omega}{\bar{C}_i(\rho)},\ 1\right), \quad \text{with}$$

$$\bar{C}_i(\rho) = \left(\int_{\mathcal{L}_i} \frac{\delta_i(u)}{\bar{c}_i(u,\rho))}\,\mathrm{d}u\right)^{-1},$$

where $\bar{c}_i$ is the data rate corresponding to the SINR in Eq. (18). We note, that under average interference the utilizations are given only implicitly, since the SINR w.r.t. BS i now depends on the loads ρ_j of the interferers. The above formulation suggests computation of the BS utilization via the fixed point iteration

$$\rho^{k+1} := f\big(\rho^k\big), \tag{19}$$

where $f(\cdot) = \big(f_1(\cdot), \ldots, f_N(\cdot)\big)^{\mathrm{T}}$. In this regard, we state the following result (refer to [6] for the proof):

Theorem 1. *For any initial load vector $\rho^0 \in \mathbb{R}_+^N$, the sequence $\rho^{k+1} := f\big(\rho^k\big)$ for $k = 0, 1, 2, \ldots$ converges to a unique fixed point $\bar{\rho} = (\rho_1, \ldots, \rho_N)^{\mathrm{T}}$.*

According to Theorem 1, the BS utilization under average interference is well defined and given as the fixed point of Eq. (19), which we denote by $\bar{\rho}$. The corresponding flow sojourn times and throughputs are, once again, obtained by inserting $\bar{\rho}$ into Eqs. (7) and (8).

V. Numerical Studies

The previous section presents two types of bounds as well as four different techniques to approximate the BS utilizations in a network of interfering BSs. This section is concerned with comparing the accuracy of the approximations and tightness of bounds.

A. Simulation Setup

We study a network of six macro cells, located at four sites with a common inter-site distance of 500 meters. One central site features three sectors, the three remaining sites are positioned around the center and feature one sector each pointing toward the center. In addition, there is a single micro cell placed in the central area. For evaluation, we select one of the three central sectors as depicted in Fig.1.

We consider the downlink of an LTE 10 MHz system, where all aspects related to the radio environment, transmit powers, etc. are modeled according to 3GPP recommendations in [23].

The system dynamics are simulated as follows: Flows arrive at each cell according to Poisson processes with equal and gradually increasing intensities. The users are uniformly distributed within each cell. The flow size S follows a heavy tailed Pareto distribution, i.e., the CDF is given as $\mathrm{P}[\text{size} > x] = \left(\frac{\theta}{x}\right)^{\alpha}$ with shape parameter $\alpha = 1.5$ and scale parameter $\theta = 2.67$ Mbit. The resulting average flow size is $\mathrm{E}(S) = \frac{\alpha\,\theta}{\alpha - 1} = 8$ Mbit, i.e., 1 Mbyte. The underlying spatial grid consists of 5000 pixels with side length of about 14 m each. For every simulation run, we observe about 1000 departures from every pixel.

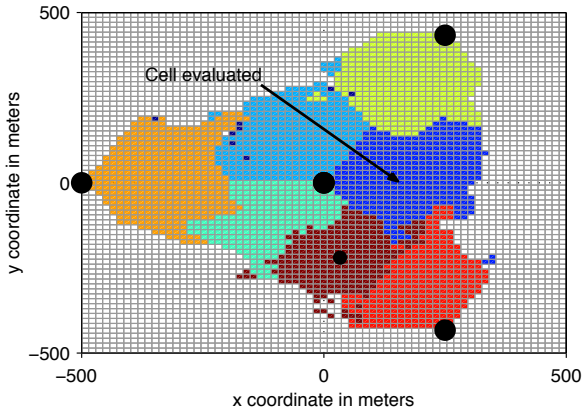


Fig. 1: Network layout used for numerical studies

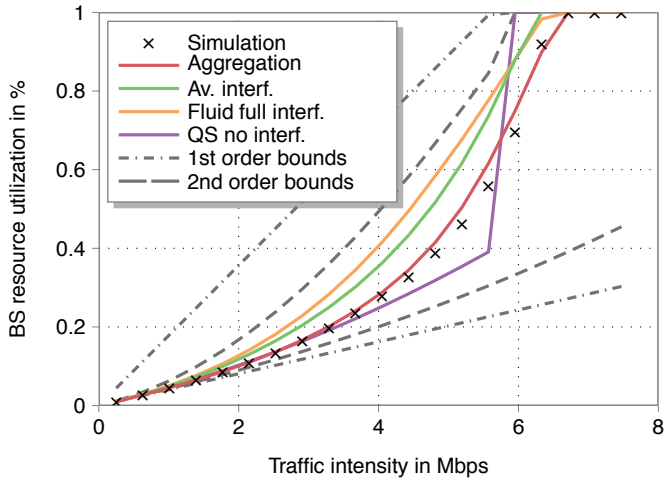
B. Simulation Results

Besides simulating the BS utilization, sojourn time, and flow throughput, we compute KPI estimates according to the second order approximations in Section IV-E, the aggregation model in Section IV-F, and the average interference model in Section IV-G, respectively. In addition, we show first and second order bounds defined in Section IV-D.

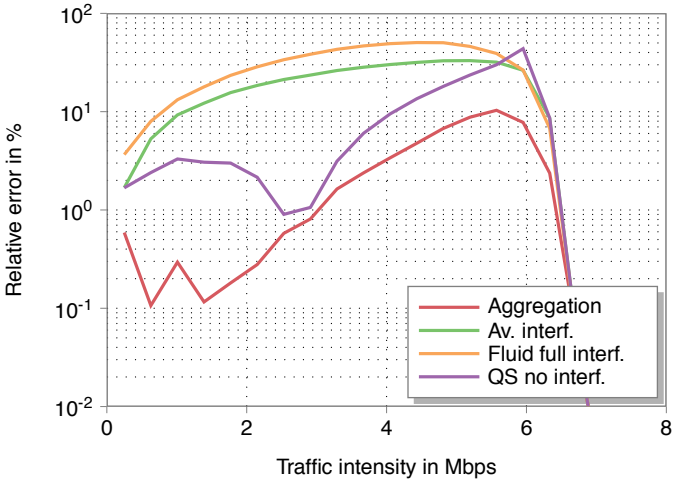
1) Base Station Utilization: First we observe the BS utilization itself, depicted in Fig.2a. The considerable gap between the first order lower and upper bound indicates a strong influence of interference for this scenario. We also observe that both first but also the second order bounds approximate the actual BS resource utilization quite coarsely, in particular for low and high load regimes, depending on the type of bound.

A closer approximation over the complete range is obtained by the approximation techniques. The quasi-stationary regime of the zero interference case ("QS no interf.") provides a very close approximation for low to medium loads, which is also observed for the flow throughput in [5]. However, for high loads the quasi-stationary regime is goverend by the behavior of $\rho_i(1)$ (compare the subequations in Eq. (13)), which leads to a very coarse approximation in high load conditions.

The closest approximation is provided by the aggregation technique. Fig.2b displays the relative error of the load estimate compared to the simulated load for all techniques. The relative error appears to be largest for medium loads, which, unfortunately, are of most practical interest. The average interference model appears to be slightly more accurate than the one combining the fluid regime with full interference ("Fluid full interf."). The maximum deviations observed for this setup are about 8 % and 30 % for the aggregation and mean interference model, and about 45 % for both second order approximations, respectively.

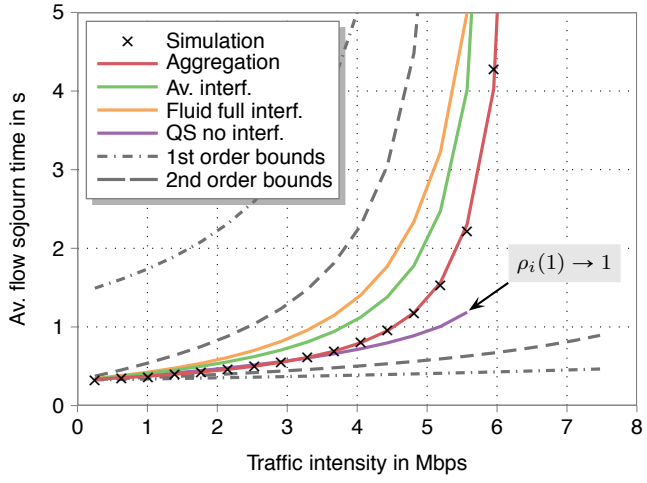


(a) BS utilization

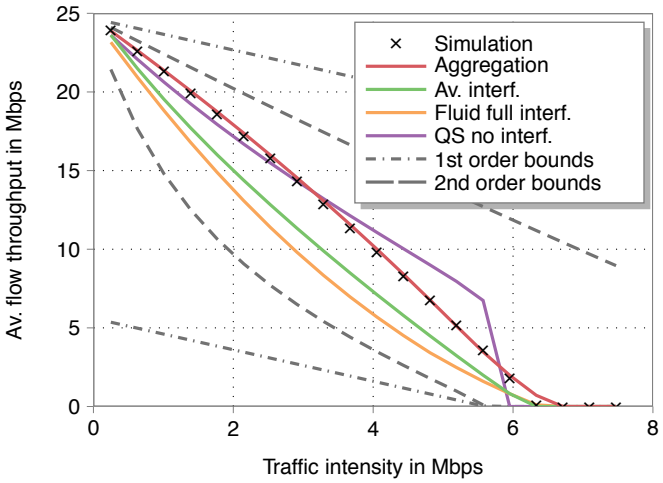


(b) Relative error of utilization estimates

Fig. 2: BS utilization: Simulation, estimation techniques, and bounds.



(a) Flow sojourn time



(b) Flow throughput

Fig. 3: Flow sojourn time and flow throughput: Simulation, estimation techniques, and bounds.

2) Flow Sojourn Time and Flow Throughput: Simulated and estimated flow sojourn times and throughputs as well as bounds are displayed in Fig.3a and Fig.3b. Since both KPIs are strictly monotonic functions of the utilization, we observe similar behavior as for the load itself. In particular, we observe a stark over- and under-estimation of both KPIs by the upper bounds in the low load regime and by the lower bounds in the high load regime, respectively. The need for approximation techniques is clearly illustrated. We observe in Fig.3a, that the quasi-stationary approximation of the sojourn time ("QS no interf.") is well defined only for a limited range of the traffic intensity until the utilization under full interference becomes equal to one.

VI. Summary and Discussion

We provide a flow level modeling framework for cellular networks, where the coupling of flow level dynamics due to intercell interference is specifically taken into account. Since the adequate queueing model renders analytically intractable, we state different methods from the literature to approximate the stationary behavior of the system. In particular, we consider the BS utilization, flow sojourn time, and flow throughput as measure of network performance.

A. Performance Bounds

Numerical investigations of a typical wireless scenario consisting of six macro and one micro cell reveal that in high and low load regimes first as well as second order bounds may be quite loose, depending on the type of bound (i.e., upper or lower).

Especially for design of network optimization algorithms, first order bounds (as used, e.g., in [2], [3]) do not appear suitable to represent any KPI considered here and approximation techniques must be considered instead.

B. Performance Approximation Techniques

A method based on aggregation of variables showed the closest approximation of all KPIs over the complete range of traffic. Unfortunately, this method also requires the highest computational effort: Calcualtion of aggregate probabilities for a network of N BSs requires solving the System (15) of size 2^N, which becomes infeasible for larger N

Close approximation of all KPIs for low to medium loads is provided by the quasi-stationary approximation with a zero interference assumption. The accuracy, however, decreases if the utilization under full interference approaches one, i.e., for $\rho_i(1) = 1$.

A suitable tradeoff between complexity and accuracy over the whole traffic range is, however, provided by the average interference model outlined in Section IV-G. Computation of the BS utilization requires computation of a fixed point, which grows like N in complexity. Moreover, it showes a closer approximation of all KPIs than both second order approximation methods. This model is already used for network optimization in a few publications such as [24], [4].

ACKNOWLEDGEMENT

This work was supported by the European Commission in the framework of the FP7 Network of Excellence in Wireless COMmunications NEWCOM# (contract n.318306).

REFERENCES

[1] R. Combes, Z. Altman, and E. Altman, "Self-organization in wireless networks: A flow-level perspective," in *2012 Proceedings IEEE INFOCOM*. IEEE, Mar. 2012, pp. 2946–2950.

[2] K. Son, H. Kim, Y. Yi, and B. Krishnamachari, "Base Station Operation and User Association Mechanisms for Energy-Delay Tradeoffs in Green Cellular Networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 8, pp. 1525–1536, Sept. 2011.

[3] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Distributed Alpha-Optimal User Association and Cell Load Balancing in Wireless Networks," *IEEE ACM Transactions on Networking*, vol. 20, no. 1, pp. 177–190, Feb. 2012.

[4] A. Fehske, H. Klessig, J. Voigt, and G. Fettweis, "Concurrent Load-Aware Adjustment of Cell Selection and Antenna Tilts in Self-Organizing Radio Networks," *Transaction on Vehicular Technology*, vol. to appear, 2013.

[5] T. Bonald, S. Borst, N. Hegde, and A. Proutiére, "Wireless data performance in multi-cell scenarios," in *SIGMETRICS'04, Proceedings of the joint international conference on Measurement and modeling of computer systems*, vol. 32, no. 1, June 2004, p. 378.

[6] A. J. Fehske and G. P. Fettweis, "Aggregation of Variables in Load Models for Cellular Data Networks," in *Proceedings of the International Conference on Communication*, Ottawa, 2012.

[7] I. Siomina and D. Yuan, "Analysis of Cell Load Coupling for LTE Network Planning and Optimization," *IEEE Transactions on Wireless Communications*, vol. 11, no. 6, pp. 2287–2297, June 2012.

[8] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2009-2014," 2011.

[9] A. Feldmann, A. C. Gilbert, and W. Willinger, "Data networks as cascades: Investigating the multifractal nature of Internet WAN traffic," *ACM SIGCOMM Computer Communication Review*, vol. 28, no. 4, pp. 42–55, Oct. 1998.

[10] A. Feldmann, A. C. Gilbert, P. Huang, and W. Willinger, "Dynamics of IP traffic: A study of the role of variability and the impact of control," *ACM SIGCOMM Computer Communication Review*, vol. 29, no. 4, pp. 301–313, Oct. 1999.

[11] T. Bonald and J. W. Roberts, "Insensitivity results in statistical bandwidth sharing," in *Proceedings of the 17th International Test Conference (ITC)*, 2001.

[12] S. B. Fredj, T. Bonald, A. Proutiere, G. Régnié, and J. W. Roberts, "Statistical bandwidth sharing: A study of congestion at flow level," in *Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications - SIGCOMM '01*, vol. 31, no. 4. New York, New York, USA: ACM Press, Aug. 2001, pp. 111–122.

[13] Z. Wang, E. Tameh, and A. Nix, "Joint Shadowing Process in Urban Peer-to-Peer Radio Channels," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 1, pp. 52–64, Jan. 2008.

[14] Internet Business Solutions Group, "Connected Life Market Watch," 2011.

[15] P. Mogensen, W. Na, I. Z. Kovács, F. Frederiksen, A. Pokhariyal, K. I. Pedersen, T. Kolding, K. Hugl, and M. Kuusela, "LTE Capacity compared to the Shannon Bound," in *IEEE Vehicular Technology Conference (VTC Spring)*, no. 1, 2007, pp. 1234–1238.

[16] R. Nelson, *Probability, stochastic processes, and queueing theory / the mathematics of computer performance modeling*. New York: Springer, 1995.

[17] A. Kherani and A. Kumar, "Stochastic models for throughput analysis of randomly arriving elastic flows in the Internet," in *Proceedings.Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 2. IEEE, 2002, pp. 1014–1023.

[18] T. Bonald, "Insensitive queueing models for communication networks," in *Proceedings of the 1st international conference on Performance evaluation methodolgies and tools*. New York, New York, USA: ACM Press, Oct. 2006, p. 57.

[19] G. Fayolle and R. Iasnogorodski, "Two coupled processors: The reduction to a Riemann-Hilbert problem," *Zeitschrift fuer Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 47, no. 3, pp. 325–351, 1979.

[20] F. Delcoigne, A. Proutière, and G. Régnié, "Modeling integration of streaming and data traffic," *Performance Evaluation*, vol. 55, no. 3-4, pp. 185–209, Feb. 2004.

[21] H. A. Simon and A. Ando, "Aggregation of Variables in Dynamic Systems," *Econometrica*, vol. 29, no. 2, pp. 111–138, Nov. 1961.

[22] E. Fischer, A. Fehske, and G. P. Fettweis, "A Flexible Analytic Model for the Design Space Exploration of Many-Core Network-on-Chips Based on Queueing Theory," in *SIMUL 2012, The Fourth International Conference on Advances in System Simulation*, 2012, pp. 119–124.

[23] 3GPP, *TR36.814 "Further advancements for {E-UTRA} physical layer aspects"*, 3rd Generation Partnership Project Std., 2010.

[24] I. Siomina and D. Yuan, "Load Balancing in Heterogeneous LTE : Range Optimization via Cell Offset and Load-Coupling Characterization," in *Proceedings of the International Colnference on Communications*, Ottawa, Canada, 2012.