Admission Control in Interference-Coupled Wireless Data Networks: A Queuing Theory-Based Network Model

Henrik Klessig, Albrecht Fehske, and Gerhard Fettweis
Vodafone Chair Mobile Communications Systems
Dresden University of Technology, Germany
Email: {henrik.klessig, albrecht.fehske, fettweis}@tu-dresden.de

***Abstract*—Mobile traffic demand varies significantly in time and space. Hence, wireless radio resources in hotspot areas and at peak traffic hours may be scarce. Consequently, special attention has to be paid to effects induced by admission control, i. e., blocking of data requests by base stations in case of high utilization or overload. Moreover, rising traffic demand requires denser deployments and frequency reuse one. Due to the resulting inter-cell interference, the base stations' utilizations have to be considered mutually dependent, which affects the admission control performance. In this paper, we extend a flow level model for elastic traffic, which explicitly takes into account the dynamic mutual inter-cell interference among base stations, by admission control. The model presented allows computing exact values for the average base station resource utilization, flow throughputs, and blocking probabilities. To analyze large networks containing many cells, we extend two approximation techniques, a state aggregation and an average interference approach, and compare them with the exact solution. Both techniques require far less computational effort and show remarkable accuracy. We believe that the extended flow level model is a positive step towards a more accurate, flexible, and holistic framework for network analysis and planning, and self-organizing network techniques.**

***Index Terms*—blocking probability; wireless network; interference; admission control; flow level modeling; queuing theory**

I. Introduction

Advanced wireless data network planning tools and optimization algorithms require models, which accurately predict network performance metrics such as base station (BS) resource utilizations, user throughputs, or data request blocking probabilities. This, in turn, requires a fundamental understanding of data traffic and the network behavior in case of admission control and interference caused by neighboring BSs.

In contrast to investigating network performance on a packet level, a notation using *flows* and *sessions* is far more advantageous when describing Internet data traffic, which has already been highlighted by Roberts [1] in 2001. This notation allows a more in-depth analysis of (wireless) networks based on queuing theory and leads to simple expressions for network key performance indicators (KPI) such as flow sojourn times or throughputs. Applying flow level data traffic models, Fredj et al. pointed out the importance of admission control on session and flow level in [2]. Especially in case of high resource utilization, the users' perception of network performance in terms of throughputs, may deteriorate drastically, if other incoming data requests are not blocked. Further, Delcoigne et al. [3] and Benameur et al. [4], [5] studied the integration of streaming and elastic traffic and, also, explicitly stressed the advantage of admission control when applied to both streaming and elastic flows. Introducing blocking mechanisms prevents excessive flow sojourn times, which would negatively affect the users' quality of experience. Later, Bonald [6] showed that, under certain conditions, blocking models, such as the popular Erlang model, are not affected by higher moments of the service time distribution than the mean even in the case of non-Poisson flow arrivals (insensitivity property). This insight has been extended by Wu et al. in [7] to wireless data networks with admission control and processor sharing service disciplines.

Due to today's stringent capacity requirements, current wireless network technologies, such as 3GPP Long Term Evolution, apply the same frequency bands in each cell, along with rate adaptive transmission. The use of these techniques induce complex interactions between the number of flows being served by individual BSs. In this regard, Bonald et al. [8] provided first and second order bounds for the mean flow throughputs, and transfer delays under mutual interference. Recently, studying this cell load-coupling behavior has regained momentum, e. g., Majewski and Koonert [9] introduced a conservative coupled-cell load model for network planning, which has been further studied by Siomina et al. and Fehske et al. in [10] and [11], respectively, and has been applied in self-organizing network algorithms in [12] and [13].

So far, current work considers either admission control mechanisms on isolated links in wireless network models with some fixed interference conditions, or coupled cell load behavior in cellular data networks without admission control, e. g. [11]. In this paper, we combine both concepts, coupled cell loads and admission control, and thereby generalize the results of [11]. We derive the mean flow throughputs in the so-called fluid and quasi-stationary regimes, the flow blocking probabilities, and mean base station resource utilizations. Furthermore, we provide two techniques, which allow for accurate approximation of these quantities with low computational effort. In order to detail theoretical results for elastic traffic, we leave investigations on streaming, and application to network planning and self-organizing networks for future publications.

978-3-901882-63-0/2014 - Copyright is with IFIP

II. Flow Level Model with Admission Control

We consider the downlink of a wireless network consisting of N base stations (BS) deployed in the coverage region defined by a nonempty compact set $\mathcal{L} := \{u \in \mathbb{R}^2 \mid \exists i : p_i(u) \geqslant p_{\min}\}$. The terms $p_i(u) > 0$, and $p_{\min} > 0$ denote the receive powers w. r. t. BS i at location $u \in \mathcal{L}$, and the minimum receive power necessary to connect to the network, respectively. The latter is usually determined by the receiver sensitivity of mobiles. Further, we define the partition $\mathcal{P}$ of the coverage area $\mathcal{L}$ into N cells $\mathcal{L}_i$ as $\mathcal{P} := \{\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_N\}$. The partition $\mathcal{P}$ may be given by Voronoi tessellation or specific user-base station association rules that define the cell shapes.

A. Traffic Model and Base Station Activity

1) Flows and Sessions: We model traffic and data requests by the users in terms of *elastic flows* and *sessions*. A data flow represents data packets that belong to one specific object like a web page, an email, or a general data file, whereas, a session consists of a number of consecutive flows initiated by a user during a specific time interval, e. g., while browsing the Internet. Similar to the arrival of telephone calls, the arrival of sessions can be described by a Poisson process [2].

Independent of the session structure given by, e. g., a particular distribution of the average time a user takes to think, the flow arrival can also be modeled as a Poisson process, when individual data requests (flows) are blocked by the BS [6]. The Poisson arrival assumption holds as long as the users show some specific behavior, i. e., either such a flow is ignored and the user is inactive for some rethinking time (so-called *jump-over blocking*), or the transmission of a blocked flow is re-tried after some rethinking time with a certain probability (blocking with random re-trial). We denote the flow arrival intensity to BS i (in s^{-1}) by λ_i with flow sizes ω following an exponential distribution with common mean Ω (in Mbits). Further, we introduce $\delta_i(\cdot)$, the normalized spatial user distribution in cell $\mathcal{L}_i$ with $\int_{\mathcal{L}_i} \delta_i(u)\,\mathrm{d}u = 1$.

2) Activity of a Base Station: The fact that there is at least one active flow in a cell determines whether the corresponding BS is active, and whether it generates interference in neighboring cells. In this regard, for ease of mathematical tractability, we make the following assumption:

Assumption 1. If there is at least one active flow being served in a cell, all available transmission resources are allocated to that flow and the corresponding BS transmits at full power.

Remark 1. Assumption 1 might be understood as a rather severe restriction compared to real mobile systems in practice, in which users may be allocated only a fraction of frequency resources. However, it is used here to account for varying interference with a simplified modeling of radio resource scheduling. Moreover, it reflects a best-effort service discipline, which is reasonable for elastic traffic. As explained further below (Section IV), relaxing Assumption 1 by considering time-average interference, i. e., scheduling maintaining an average frequency resource utilization and, thereby, representing the other extreme case, results in similar cellular network performance with respect to the impact of interference.

To capture the dynamic interference variations originating from flow dynamics, we denote a vector by $y \in \mathcal{Y} := \{0,1\}^N$, the elements of which assume the value one if the corresponding BSs are active and zero otherwise. We introduce the set of interference scenarios, where BS i is active, by $\mathcal{A}_i = \{y \in \mathcal{Y} \mid y_i = 1\}$ and collect the indices of active and inactive BSs in the sets

$$\mathcal{N}_0(y) := \{i = 1, \ldots, N \mid y_i = 0\} \quad \text{and} \tag{1}$$

$$\mathcal{N}_1(y) := \{i = 1, \ldots, N \mid y_i = 1\}. \tag{2}$$

B. Radio Link Model

We assume fast and slow fading effects to evolve on much shorter, and, much longer time scales, respectively, than flow durations, so that they are contained either as averages, or as instantaneous values in location-dependent but otherwise constant functions $p_i(\cdot)$.

1) Radio Link Quality: Further, the collection of BSs that interfere at any given point in time, strongly affects the radio link quality. In contrast to slow and fast fading effects, these interference scenarios evolve on the same time scale as the flow dynamics and, as a result, data rates as well as the utilizations of all BSs are strongly interconnected.

Considering various interference scenarios y, we define the signal-to-interference and noise ratio (SINR) experienced by a data flow at location $u \in \mathcal{L}$ w. r. t. BS i as

$$\gamma_i(u,y) := \frac{p_i(u)}{\sum_{j \in \mathcal{N}_1(y)\setminus\{i\}} p_j(u) + N_0}, \tag{3}$$

with $N_0 > 0$ denoting the noise power, and the data rate as

$$c_i(u,y) := aB \min\left\{\log_2\left(1 + b\gamma_i(u,y)\right), c_{\max}\right\}, \tag{4}$$

where B, a, b denote the bandwidth, bandwidth efficiency, and SINR efficiency, respectively. Following the approach in [14], we incorporate an average packet scheduling gain via the parameters a and b, and choose corresponding values for more spectrally efficient scheduling mechanisms, MIMO techniques, or system specific overheads. The term $c_{\max}$ denotes the maximum bit rate achievable, given by the highest modulation and coding scheme of the system at hand.

2) Cell Capacity: In order to model important network key performance indicators (KPIs) on flow level, we take the metric *cell capacity* C_i as a starting point, which we define as the rate provided to a user at any given point in time and averaged over all locations within a cell. Hence, it becomes the harmonic mean of the location-dependent rates $c_i(u,y)$ achieved within the serving area $\mathcal{L}_i$, i. e.,

$$C_i(y) := \left[\int_{\mathcal{L}_i} \frac{\delta_i(u)}{c_i(u,y)}\,\mathrm{d}u\right]^{-1}. \tag{5}$$

It is of high importance to understand, why the weighted harmonic mean instead of an arithmetic mean is applied here. Usually, a user at location u requests a specific amount of data in Mbit to be transferred. This data volume along with his or her rate $c_i(u, y)$ ultimately determines the time for transmission (data amount divided by the rate). Therefore, the lower the rate $c_i(u, y)$, the longer it will take to transfer the data. As a result, users with a lower rate at cell edges are more likely to be active for longer periods of time thereby decreasing the cell's capacity. This fact in explicitly considered, since the harmonic mean is more sensitive to outliers. This problem usually occurs when averaging rates or speeds for given quantities, e. g., data to be processed or transferred, or distances to be traveled. A more elaborate derivation of the cell capacity as the harmonic mean of rates can be found in [15].

Note further that, since the individual achievable rates $c_i(u, y)$ within a cell highly depend on the collection of active neighboring BSs, the cell capacity $C_i(y)$ is a function of the interference scenario y, as well, and is lower under mutual interference.

3) Handover Mechanisms: Usually, the users' mobility along with the slow fading process triggers handover events. Here, we assume that a flow remains connected to a single serving BS during its entire lifetime, since user mobility and slow fading commonly occur on longer time scales. We, therefore, disregard modeling handover mechanisms in detail.

C. Resource Sharing and Admission Control

In wireless networks, the cell's capacity is typically bounded by some limiting bandwidth and is shared among several users. Particularly, in high traffic regimes with many users initiating various data flow requests, congestion may occur in some of the cells. As a result, network performance, in terms of throughput and delay, deteriorates drastically. Therefore, it is beneficial to block single user requests in order to maintain the performance of ongoing data transmissions of other users. As a consequence, network operators introduce admission control in order to prevent congestion. In the following, we assume that, in each cell i, resources are shared *equally* among at most L_i users according to the egalitarian processor sharing (EPS) discipline [16], i. e., BS i admits at most L_i concurrent flows and blocks all further requests. Note that applying the EPS discipline is equivalent to the assumption of a blind scheduling policy, for instance the well-known Round Robin scheduler.

D. Flow Level KPIs under Static Interference

In order to reflect the network's quality of service more precisely, we concentrate on flow level metrics, in particular on flow throughputs and on the number of requests rejected, as opposed to classical network performance metrics, such as SINR and rate distributions. For this purpose, we use basic queuing theoretic concepts and thereby avoid extensive Monte Carlo simulations.

We interpret some BS i as a single server with a queue of maximum length L_i and an interference-dependent mean service rate $\mu_i(y) := C_i(y)/\Omega$. Let $X_i(t) \in \{0, \ldots, L_i\}$ denote a continuous time random process representing the number of active flows in a cell i at time t. Then, for some *fixed* interference scenario y, a BS i can be described by an M/M/1/L_i EPS queuing system, and its load $\rho_i(y)$ is given as the ratio of arrival and service rates

$$\rho_i(y) := \frac{\lambda_i}{\mu_i(y)} = \frac{\lambda_i \Omega}{C_i(y)}. \tag{6}$$

Note that in contrast to the derivations in [11], we clearly distinguish between the *cell load* and the *resource utilization* of the serving BS. Since there may exist flow requests that are not admitted and thus blocked, we have to exclude them when computing the BSs' utilization. Hence, the resource utilization of a BS i is given by

$$\eta_i(y) = \rho_i(y)\left(1 - P_{\mathrm{b},i}(y)\right) = \rho_i(y)\frac{1 - \rho_i(y)^{L_i}}{1 - \rho_i(y)^{L_i+1}} \tag{7}$$

where the flow blocking probability $P_{\mathrm{b},i}$ in cell i depends on the maximum number L_i of flows being served concurrently and on the interference scenario y, as well:

$$P_{\mathrm{b},i}(y) = \frac{\left(1 - \rho_i(y)\right)\rho_i(y)^{L_i}}{1 - \rho_i(y)^{L_i+1}}. \tag{8}$$

The mean number n_i of active flows in a system with finite buffer size L_i can be given as

$$n_i(y) = \frac{\rho_i(y)}{1 - \rho_i(y)} - \frac{(L_i + 1)\rho_i(y)^{L_i+1}}{1 - \rho_i(y)^{L_i+1}}, \tag{9}$$

and used in the equation to compute the average flow throughput $r_i(y)$ w. r. t. BS i, under interference scenario y, i. e.,

$$r_i(y) = \frac{\eta_i(y)C_i(y)}{n_i(y)}, \tag{10}$$

which is basically the average utilized capacity divided by the mean number of active flows. [15]–[17]

In the following, we drop the assumption of y being fixed.

E. Flow Level KPIs under Variable Interference

Let $X(t) := \left(X_1(t), \ldots, X_N(t)\right) \in \mathcal{S}$ denote the vector process of the number of active flows in N cells with state space $\mathcal{S}$ and transition rates from state x to state x'

$$q(x, x') = \begin{cases} \lambda_i & \text{for } x' = x + e_i,\ x'_i \leqslant L_i, \\ \mu_i(y) & \text{for } x' = x - e_i,\ x_i \leqslant L_i \\ 0 & \text{else,} \end{cases} \tag{11}$$

where e_i denotes the N-dimensional unit vector with the i^{th} component equal to one. With the definitions above along with Assumption 1, we are able to restate the definition of the vector of active BSs as $y := \operatorname{sgn}(x)$. As a result, the interference scenario y itself becomes a random process, capturing the dynamic effect of inter-cell interference and, hence, the correlation between the BSs' quality of service provided to the mobile users.

Since many system KPIs, such as a BS's resource utilization, flow throughput, or blocking probability, can directly be deduced from the state probabilities, we are interested in characterizing the joint stationary state distribution $\pi(x)$ of the vector process $X(t)$. It can be obtained by solving the system of balance equations [11]

$$\left(\sum_{n\in\mathcal{N}_0(y)}\lambda_n+\sum_{n\in\mathcal{N}_1(y)}\lambda_n+\mu_n(y)\right)\pi(x)= \\ =\sum_{n\in\mathcal{N}_0(y)}\mu_n(y+e_n)\pi(x+e_n)+ \\ +\sum_{n\in\mathcal{N}_1(y)}\big(\lambda_n\pi(x-e_n)+\mu_n(y)\pi(x+e_n)\big)\,. \tag{12}$$

Remark 2. Note that, in accordance with the admission control policy, the state probabilities $\pi(x)$ in Eqs. (12) are zero, whenever any element x_n is larger than the maximum number L_n of flows concurrently being served.

Unfortunately, the service rates $\mu_i(y)$ vary among the states in a manner such that the queuing network is not partially reversible. As a result, techniques used to obtain a closed product form of the stationary probability distribution $\pi(x)$ are not applicable. However, since the state space is finite, as opposed to the system without admission control as in [11], the system of balance equations (12) can be solved for stationary state probabilities $\pi(x)$ with standard numerical tools. In the following, we provide the derivation of important network KPIs from the stationary state probabilities along with approximations obtainable with low computational effort.

With regard to Assumption 1 and the PASTA (*Poisson Arrivals See Time Averages*) property of systems with Poisson arrivals [18], the probability that at least one flow is active in some cell i is equivalent to the BS's resource utilization and the blocking probability corresponds to the probability of exactly L_i flows being active. Hence, both can be computed by

$$\eta_i:=\sum_{\substack{x\in\mathcal{S}\\x_i\geqslant 1}}\pi(x)\quad\text{and}\quad P_{\mathrm{b},i}:=\sum_{\substack{x\in\mathcal{S}\\x_i=L_i}}\pi(x), \tag{13}$$

and the mean number of active flows in cell i becomes

$$n_i=\sum_{j=1}^{L_i}\sum_{\substack{x\in\mathcal{S}\\x_i=j}}j\pi(x). \tag{14}$$

In order to determine flow throughputs, we have to consider the speed with which interference varies and, hence, the number of interference scenarios experienced during the entire transmission of a flow. The two limiting cases, i. e., a flow observes infinitely many state transitions, or just only one interference scenario y, is provided by the *fluid* and *quasi-stationary* regimes, respectively, where the former overestimates and the latter underestimates system performance in terms of throughputs [3]. Introducing $\sigma_i(y)$ as the probability of a flow in cell i experiencing interference scenario y as

$$\sigma_i(y)=\frac{\sum_{\substack{x\in\mathcal{S}\\\mathrm{sgn}(x)=y}}\pi(x)}{\sum_{\substack{x\in\mathcal{S}\\x_i\geqslant 1}}\pi(x)}, \tag{15}$$

the average flow throughput r_i^{qs} (in the quasi-stationary setup) can be computed by the harmonic mean of interference-dependent throughputs given by Eq. (10), i. e.,

$$r_i^{\mathrm{qs}}=\left[\sum_{y\in\mathcal{A}_i}r_i(y)^{-1}\sigma_i(y)\right]^{-1}. \tag{16}$$

Since, in the fluid regime, a flow observes all interference scenarios y infinitely often, it experiences a mean rate $\sum_{y\in\mathcal{A}_i}c_i(u,y)\sigma_i(y)$ and the cell capacity becomes

$$C_i^{\mathrm{fl}}=\left[\int_{\mathcal{L}_i}\frac{\delta_i(u)}{\sum_{y\in\mathcal{A}_i}c_i(u,y)\sigma_i(y)}\,\mathrm{d}u\right]^{-1}. \tag{17}$$

The mean flow throughput in the fluid regime can then be calculated by inserting n_i, η_i and C_i^{fl} into Equation (10).

Note that the argument for taking the harmonic rather than the arithmetic mean previously given for Eq. (5) is also true for Eqs. (16) and (17).

III. Approximation Techniques

For a large state space $\mathcal{S}$, i. e., a large number of BS N and flows L_i allowed concurrently, the computation of network KPIs as described above may be cumbersome, since the order of the system of balance equations (12) grows fast with N and L_i (see III-C). We, therefore, provide approximation techniques in the following.

A. Approximation by State Aggregation

The service rates $\mu_i(y)$ are state-dependent in general, but are equal in states, where the same BSs are active, i. e., where we can observe the same interference situation. Based on the definition of the interference situation $y:=\mathrm{sgn}(x)$, we can aggregate these states by $\mathcal{S}(y):=\{x\in\mathcal{S}\mid\mathrm{sgn}(x)=y\}$.

Conditioned on observing the network being in states within one aggregate, the system essentially behaves like a network of independent queues, since the service rates do not vary in the specific states. Therefore, we consider a separate investigation of the transition rates within and among aggregates as in [11]. We apply and extend the procedure in [11] by admission control, i. e., by limiting the state space $\mathcal{S}$. However, to ensure brevity, the approach in [11] is not elaborated upon. The approximate state probabilities $\widetilde{\pi}(x)$, conditioned on being in aggregate $\mathcal{S}(y)$, become

$$\widetilde{\pi}(x)=\begin{cases}\prod_{n\in\mathcal{N}_1(y)}\frac{\left(1-\rho_n(y)\right)\rho_n^{x_n-1}(y)}{1-\rho_n^{L_n}(y)}\sigma(y) & \text{for } y\neq 0,\\ \sigma(0) & \text{for } y=0,\end{cases} \tag{18}$$

where $\sigma(y)$ denotes the probability of observing interference scenario y and, hence, of being in aggregate $\mathcal{S}(y)$. In order to determine $\sigma(y)$, we proceed by substituting the exact state probabilities $\pi(x)$ with the approximate state probabilities $\widetilde{\pi}(x)$ in Eqs. (12) and by summing up terms that correspond to individual aggregates. We then obtain the transition rates among aggregates:

$$p(y,y') = \begin{cases} \lambda_n & \text{for } y' = y + e_n, \\ \mu_n(y) - \lambda_n \frac{1-\rho_n^{L_n-1}(y)}{1-\rho_n^{L_n}(y)} & \text{for } y' = y - e_n, \\ 0 & \text{else.} \end{cases} \quad (19)$$

Introducing the matrix $P = [p_{ij}]$ with arbitrary ordering $i \mapsto y(i)$, $p_{ij} := p\big(y(i), y(j)\big)$, and $p_{ii} = -\sum_{j=1}^{2^N} p_{ij}$, we can compute the probabilities $\sigma(y)$ by solving the system of equations given by

$$P\sigma = 0, \quad (20)$$

where the vector σ contains individual aggregate probabilities $\sigma(\cdot)$ in corresponding order. Once the System (20) has been solved for $\sigma(y)$, the approximate probabilities $\widetilde{\pi}(x)$ are computed via Eq. (18). Network KPIs of interest can now be calculated as outlined in Section II-E by substituting $\pi(x)$ with $\widetilde{\pi}(x)$.

B. M/M/1/L Approximations and Performance Bounds

1) Fluid Approximation based on Average Interference: In order to compute the mean flow throughputs and blocking probabilities for large networks with many BSs, we resort to another approach (introduced in [9]) for systems without blocking and further studied in [10] and [11]. We extend these approaches by introducing admission control. We assume, that each data flow experiences *time-average interference conditions*, such that the individual queues are decoupled. The notion of time-average interference thereby relaxes Assumption 1.

The SINR and the achievable rate become

$$\overline{\gamma}_i(u,\eta) := \frac{p_i(u)}{\sum_{j \neq i} \eta_j p_j(u) + N_0} \quad \text{and} \quad (21)$$

$$\overline{c}_i(u,\eta) = aB \min\left\{\log_2\left(1 + b\overline{\gamma}_i(u,y)\right), c_{\max}\right\}, \quad (22)$$

respectively. Using Eqs. (6) and (7), the load and resource utilization are

$$\overline{\rho}_i(\eta) := \frac{\lambda_i \Omega}{\overline{C}_i(\eta)}, \quad \text{with} \quad \overline{C}_i(\eta) := \left[\int_{\mathcal{L}_i} \frac{\delta(u)}{\overline{c}_i(u,\eta)}\, \mathrm{d}u\right]^{-1} \quad \text{and} \quad (23)$$

$$\eta_i(\overline{\rho}_i) := h_i(\overline{\rho}_i) = \overline{\rho}_i \frac{1 - \overline{\rho}_i^{L_i}}{1 - \overline{\rho}_i^{L_i+1}}, \quad (24)$$

respectively.

Since the cell load $\overline{\rho}_i$ is a function of the BS utilization η through the interference coupling effect, the definition of the utilization η_i can be rewritten as $\eta_i := f_i(\eta) = (h_i \circ \overline{\rho}_i)(\eta)$, where $(h_i \circ \overline{\rho}_i)$ denotes the composition of the functions $h_i(\cdot)$ and $\overline{\rho}_i(\cdot)$. Further, we define the corresponding vector function $f(\cdot) = \big(f_1(\cdot), \ldots, f_i(\cdot), \ldots, f_N(\cdot)\big)$. Note that the BS resource utilization η_i is given only in an implicit form; however, the system of equations $\eta = f(\eta)$ can be solved numerically via a fixed point iteration, which is stated in the following theorem (the proof is detailed in the appendix):

Theorem 1. *For any initial utilization vector $\eta^0 \in \mathbb{R}_+^N$ with $\eta_i^0 < 1$, the sequence $\eta^{k+1} := f(\eta^k)$ for $k = 0, 1, 2, \ldots$ converges to a unique fixed point $\overline{\eta}$.*

Once, the resource utilization $\overline{\eta}$ has been obtained, the blocking probability and the mean number of active flows can be computed via Eqs. (8) and (9) with loads $\overline{\rho}_i(\overline{\eta})$ as arguments. Finally, inserting the mean number of active flows, the resource utilizations $\overline{\eta}$, and the cell capacity $\overline{C}_i(\overline{\eta})$ into Eq. (10), yields the mean flow throughput under the average interference assumption.

2) First Order Bounds: For comparison, we also state upper and lower bounds for the performance metrics of interest and assume either *minimum* or *maximum* interference conditions, respectively. In these cases, there is no dynamic load coupling among the cells, and in order to compute network KPIs, it is sufficient to insert either $y = 0$ or $y = 1$ into Eqs. (5) - (10), respectively. These bounds correspond to the first order bounds derived in [2] and [19] for systems without admission control.

3) Second Order Bounds: Following [19], it is assumed that the process X_i depends on all other processes X_j, $j \neq i$, but the latter evolve either under minimum or maximum interference scenarios. Suppose $\hat{\eta}_i$ and $\check{\eta}_i$ are derived from the first order upper and lower performance bound, respectively. Then the probability that a flow in cell i sees interference situation y is

$$\widehat{\sigma}_i(y) = \prod_{\substack{j \in \mathcal{N}_0(y) \\ j \neq i}} \left(1 - \hat{\eta}_j\right) \prod_{\substack{j \in \mathcal{N}_1(y) \\ j \neq i}} \hat{\eta}_j \quad (25)$$

for the first case. In the second, $\breve{\sigma}_i(y)$ can be calculated by using the equation above and by substituting $\hat{\eta}_i$ by $\check{\eta}_i$. The *second order upper bound* combines the fluid regime and the probabilities $\widehat{\sigma}_i(y)$ by inserting the latter into C_i^{fl} and by using M/M/1/L EPS equations (Eqs. (6) - (10)). The *second order lower bounds* for the mean flow throughput and blocking probability can be determined via the quasi-stationary throughput Eq. (16) combined with $\breve{\sigma}_i(y)$ and by

$$\breve{P}_{\text{b},i} = \sum_{y \in \mathcal{A}_i} P_{\text{b},i}(y) \breve{\sigma}_i(y). \quad (26)$$

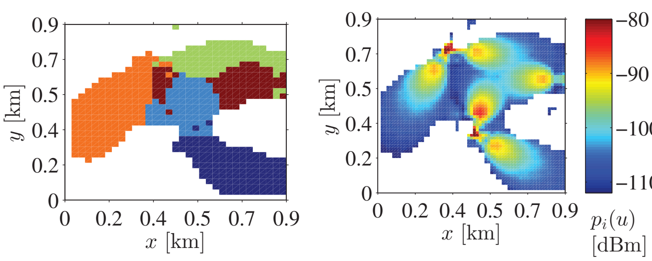
C. Complexity Analysis

Clearly, the finiteness of the state space $\mathcal{S}$ (finite L_i) allows for an exact computation of the steady state probabilities $\pi(x)$; however, the order of the System (20) to be solved is $\prod_{i=1}^N (L_i + 1)$. To compute the KPIs for a standard two-tier hexagonal network with 57 cells and with at most nine flows being served in each cell concurrently, the order becomes 10^{57}, which essentially renders the computation infeasible.

In contrast, the aggregation model only requires the solution of a linear system of 2^N equations, which is feasible for a moderate number N of BSs. Far less computational effort is needed by the average interference model, where N nonlinear equations are to be solved iteratively. Hence, the approximation with the aggregation model can be very beneficial for performance evaluation of networks with a moderate number of BSs, whereas the average interference approach allows analysis of large networks with several tens or hundreds of cells.

IV. Numerical Evaluation

In order to quantify the accuracy of the approximations, we use a toy network scenario, which is depicted in Fig. 1. We utilize five cells and restrict the number of flows being served concurrently in each cell to three. We apply a bandwidth B of 10 MHz and uniform spatial traffic distributions $\delta_i(\cdot)$. Receive powers $p_i(\cdot)$ are generated by the Okumura-Hata path loss model [20] and cell areas $\mathcal{L}_i$ are formed by a user association rule according to strongest receive power. We evaluate the light blue central cell in Fig. 1(a), since it is surrounded by the other four cells and strongly affected by inter-cell interference.



(a) (b)

Fig. 1. (a) Cell areas and (b) receive powers of the network used for numerical evaluation.

Remark 3. The reason why to choose such a small network is that the exact computation of the network KPIs requires high computational effort as stated in Section III-C. However, from our experience of using the model, we can affirm that the following observations also hold for larger networks.

A. General Observations

Flow throughputs and blocking probabilities, in general, highly depend on the traffic demand, which is shown in Figs. 2(a) and 2(b). We observe that the exact solution and both its approximations, the aggregation and the average interference model, approach the first and second order bounds for low and high traffic, respectively. In contrast, for moderate traffic situations, the interference-coupling behavior can be clearly identified: The higher the traffic demand, larger the resources that have to be allocated by a BS and higher the interference generated in the surrounding cells. This, in turn, decreases the cell capacity of other BSs. As a consequence, more resources have to be allocated to serve a given demand and fewer resources are left for the individual users, thereby, decreasing their data throughputs.

B. Accuracy Considerations

Figs. 2(c) -2(e) depict the errors of the approximations and second order bounds w. r. t. the exact flow blocking probabilities (absolute errors), and the flow throughputs in the fluid and quasi-stationary regimes (relative errors), respectively. Computing the blocking probabilities, the aggregation model is superior to the average interference approach for very high traffic. It is more likely to find the system in a state, where all BSs are active, i. e., $y = 1$, and where blocking most likely occurs. Therefore, the transitions among the aggregates play a minor role, and the blocking probabilities are well approximated by the system conditioned on being in aggregate described by $y = 1$ and the decoupled queue therein. The aggregation model approximates the flow throughput well, both in the fluid and the quasi-stationary setup. The low effort average interference approach provides a good approximation of flow blocking probabilities and throughputs, as well. Since the throughput approximation lies in between the fluid and quasi-stationary regimes, which constitute limiting cases with regard to the variation speed of the interference, the average interference model can be accepted as a very accurate approximation.

V. Summary and Fields of Application

In this paper, we extend a wireless network flow level model for elastic traffic by introducing admission control functionality into base stations. In contrast to other existing work, the queuing theory-based model presented explicitly takes into account the dynamic mutual interference among all base stations instead of assuming some fixed interference level. As opposed to the system without admission control, the extended model allows for computing exact values for flow blocking probabilities in all cells and for mean flow throughputs in fluid and quasi-stationary regimes. In addition to the exact solution, we adapt two approximation techniques, a state aggregation and an average interference approach, to a system with admission control. Both approximations allow analyzing wireless networks containing several tens or even hundreds of cells, with low computational effort, and show remarkable accuracy. Specifically, the average interference approach is a convenient trade-off between accuracy and low complexity for performance evaluation of very large networks.

In high traffic scenarios, special attention has to be paid to effects induced by admission control, i. e., blocking of data requests and users, as well as, varying guaranteed throughputs by limiting the number of users. Such effects may worsen the users' perception of the network quality drastically. In this regard, we believe that the flow level model with blocking is a positive step towards a more accurate and more flexible analysis of quality of experience metrics of wireless multi-cell networks. Also, we think that it will contribute to more effective network planning and (self-) optimization techniques, such as adaptive service-type dependent admission control.

Another area to apply an extended version of the model may be planning of future 5G networks, in which sporadic low rate-low latency and human initiated high data rate applications will run concurrently. The model is capable of computing data flow sojourn times (as the inverse of the throughput) and, hence, of characterizing latency. The model may help ensuring low latency by freeing resources, which can be achieved by blocking high data rate users.

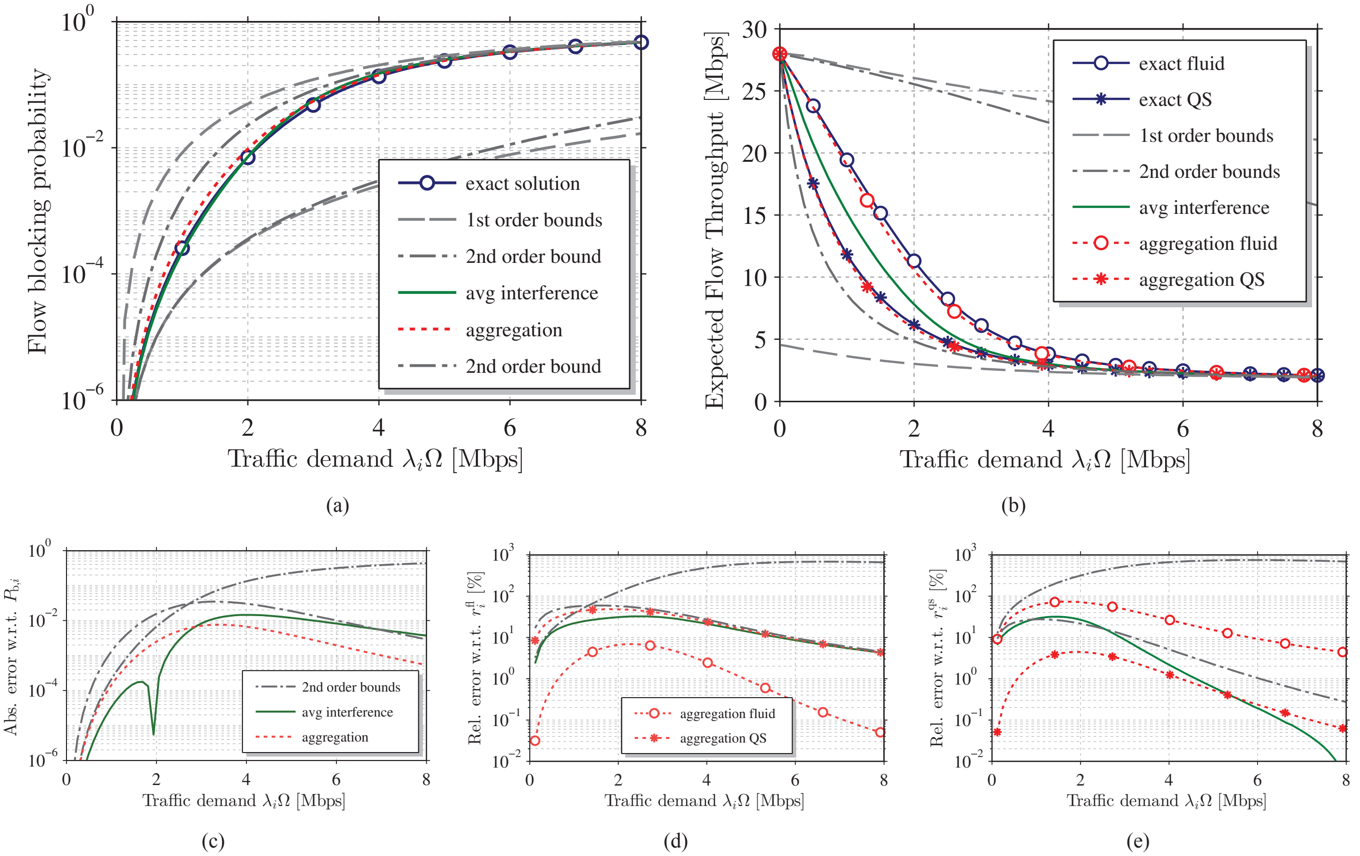


Fig. 2. (a) Comparison of the flow blocking probability and (b) mean flow throughput computations. (c) Absolute errors w. r. t. exact flow blocking probability, and (d)-(e) relative errors w. r. t. fluid regime and quasi-stationary regime flow throughputs for increasing traffic demand. Note that both legends in figures (c) and (d) apply jointly for figures (c)-(e).

Acknowledgment

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

Appendix

A. Proof of Theorem 1

For better readability, we omit the bar and write $\rho(\eta)$ instead of $\bar{\rho}(\eta)$; however, by ρ we always mean the load derived with the average interference approach, see Eq. (23). We also omit the indices i at some places and write h, ρ, and L instead. First, we show properties of the individual functions $\rho_i(\cdot)$ and $h_i(\cdot)$, which imply positivity, the nondecreasing property, and concavity of their compositions $f_i := (h_i \circ \rho_i)$.

a) Properties of the function $\rho_i(\eta)$: The loads $\rho_i(\eta)$ are *nondecreasing* and *strictly concave* in the utilization vector η. This has already been shown in the proof of Theorem 1 in [11].

b) Properties of the function $h_i(\rho_i)$: The second derivative of the function $h(\rho)$ (see Eq. (24)) is

$$h''(\rho) = \overbrace{\rho^{L-1}(1-\rho^{L+1})(L+1)}^{a} \times \frac{\Big[\overbrace{\rho^{L+1}(L\rho - L - 2)}^{b} + \overbrace{(L+2)\rho - L}^{c}\Big]}{\underbrace{(1-\rho^{L+1})^4}_{d}} \tag{27}$$

We show concavity by studying $h''(\rho)$ for three cases:
$\underline{\rho = 0}$: Substituting, we observe that $h''(0) = 0$, which implies concavity of the function $h(\rho)$ for $\rho = 0$.
$\underline{\rho \to 1}$: In the limit, we obtain $\lim_{\rho \to 1} h''(\rho) = -\frac{L(L+2)}{6(L+1)} < 0, \forall L > 0$. Hence, the function $h(\rho)$ is strictly concave for $\rho \to 1$.
$\underline{\rho \neq 0, \rho \neq 1}$: We further observe

i) $a \lesseqgtr 0$ for $\rho \gtreqless 1$,
ii) $b \lesseqgtr 0$ for $\rho \lesseqgtr \frac{L+2}{L}$ and $b = 0$ for $\rho = \frac{L+2}{L}$,
iii) $c \lesseqgtr 0$ for $\rho \lesseqgtr \frac{L}{L+2}$ and $c = 0$ for $\rho = \frac{L}{L+2}$.

Therefore, the function $h(\rho)$ is strictly concave for $\rho \in \left(0, \frac{L}{L+2}\right]$ and $\rho \in \left[\frac{L+2}{L}, \infty\right)$. For $\rho \in \left(\frac{L}{L+2}, 1\right)$, $a > 0$

holds, and for strict concavity, $b + c < 0$ is to be shown, which can be deduced from the assumption $\frac{L}{L+2} < \rho < 1$. For $\rho \in \left(1, \frac{L+2}{L}\right)$, $a < 0$ holds, and for strict concavity, $b + c > 0$ is to be shown, which can be proved by using induction over L. Hence, the function $h(\rho)$ is *concave* for $\rho = 0$ and for $\rho \to \infty$, and *strictly concave* for $\rho \in (0, \infty)$. The first derivative of the function $h(\rho)$ is

$$h'(\rho) = \frac{1 - (L+1)\rho^L + L\rho^{L+1}}{(1 - \rho^{L+1})^2}, \tag{28}$$

which, in the limit, becomes $\lim_{\rho \to \infty} h'(\rho) = 0$. Since the function $h(\cdot)$ is nondecreasing in the limit and strictly concave for $\rho \in (0, \infty)$, it must be *increasing* for $\rho \in (0, \infty)$. Moreover, for $\rho = 0$, we obtain $h'(0) = 1 > 0$.

c) Properties of the function $f_i(\eta)$: For $\lambda_i, \Omega \geqslant 0$ and $C_i(\eta) > 0$, all functions $\rho_i(\eta)$ produce positive results for $\eta \geqslant 0$. By inspection, one can see that the functions $h_i(\rho_i)$ are positive for any $\rho_i \geqslant 0$. Hence, their composition f_i is *positive* for $\eta \geqslant 0$. Further, for all η' satisfying $\rho_i(\eta') = 1$, the limit becomes $\lim_{\eta \to \eta'} f_i(\eta) = \frac{L_i}{L_i+1} > 0, \forall L_i$. Since the functions $h_i(\rho_i)$ are nondecreasing and the functions $\rho_i(\eta)$ are positive and nondecreasing for $\eta \geqslant 0$, the compositions $f_i(\eta)$ must be *nondecreasing* for $\eta \geqslant 0$. This follows from $f_i'(\eta) = (h_i \circ \rho_i)'(\eta) = h_i'\left(\rho_i(\eta)\right) \cdot \rho_i'(\eta)$. Moreover, since the functions $h_i(\rho_i)$ are increasing and concave for finite $\rho_i \geqslant 0$, and the functions $\rho_i(\eta)$ are strictly concave for $\eta \geqslant 0$, their compositions $f_i(\eta)$ must be *strictly concave*, which is a consequence of $f_i''(\eta) = h_i''\left(\rho_i(\eta)\right) \cdot \left(\rho_i'(\eta)\right)^2 + h_i'(\rho_i(\eta)) \cdot \rho_i''(\eta)$.

d) Convergence to the unique fixed point: If $f(\eta)$ is a so-called standard interference function (SIF), then Theorem 1 holds [21], which is true if it has the following properties:

1) Positivity: $f(\eta) \geqslant 0$ for $\eta \geqslant 0$,
2) Monotonicity: $\eta \geqslant \eta' \implies f(\eta) \geqslant f(\eta')$,
3) Scalability: $\alpha f(\eta) > f(\alpha\eta)$ for $\alpha \in \mathbb{R}, \alpha > 1$.

Positivity and monotonicity of the functions $f_i(\eta)$ imply the same properties for the vector function $f(\eta)$. Proving scalability, we use the same argument as in the proof of Theorem 1 in [11], i.e., positivity and concavity of the elements $f_i(\cdot)$ imply scalability of the vector function $f(\cdot)$. ■

REFERENCES

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

[2] S. Ben Fredj, T. Bonald, A. Proutiere, G. Régnié, and J. W. Roberts, "Statistical bandwidth sharing: A study of congestion on 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.

[3] 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.

[4] N. Benameur and S. B. Fredj, "Integrated admission control for streaming and elastic traffic," *Quality of Future Internet . . .*, 2001.

[5] N. Benameur, S. Ben Fredj, S. Oueslati-Boulahia, and J. Roberts, "Quality of service and flow level admission control in the Internet," *Computer Networks*, vol. 40, no. 1, pp. 57–71, Sep. 2002.

[6] T. Bonald, "The Erlang model with non-poisson call arrivals," *ACM SIGMETRICS Performance Evaluation Review*, vol. 34, no. 1, p. 276, Jun. 2006.

[7] Y. Wu, C. Williamson, and J. Luo, "On processor sharing and its applications to cellular data network provisioning," *Performance Evaluation*, vol. 64, no. 9-12, pp. 892–908, Oct. 2007.

[8] T. Bonald, S. Borst, N. Hegde, A. Proutiére, and A. Proutière, "Wireless data performance in multi-cell scenarios," Tech. Rep. 1, Jun. 2004.

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

[10] 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, Jun. 2012.

[11] A. J. Fehske and G. P. Fettweis, "Aggregation of variables in load models for interference-coupled cellular data networks," *2012 IEEE International Conference on Communications (ICC)*, pp. 5102–5107, Jun. 2012.

[12] I. Siomina and D. Yuan, "Load balancing in heterogeneous LTE: Range optimization via cell offset and load-coupling characterization," *2012 IEEE International Conference on Communications (ICC)*, pp. 1357–1361, Jun. 2012.

[13] A. Fehske, H. Klessig, J. Voigt, and G. Fettweis, "Concurrent load-aware adjustment of user association and antenna tilts in self-organizing radio networks," *Vehicular Technology, IEEE Transactions on*, vol. 62, no. 5, pp. 1974–1988, 2013.

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

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

[16] L. Kleinrock, *Queueing Systems Vol. II: Computer Applications*. Wiley Interscience, 1975.

[17] J. Virtamo, "Queuing theory lecture notes," Online: http://www.netlab.tkk.fi/opetus/s383143/kalvot/english.shtml, last visited on Jan 13, 2014.

[18] R. W. Wolff, "Poisson arrivals see time averages," *Operations Research*, vol. 30, no. 2, pp. 223 – 231, 1982.

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

[20] COST Hata pathloss model, Online: http://www.lx.it.pt/cost231/ final_report.htm, last visited on 6/27/2013.

[21] R. Yates, "A framework for uplink power control in cellular radio systems," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 7, pp. 1341–1347, 1995.