This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2015.2480760, IEEE Transactions on Wireless Communications

A Performance Evaluation Framework for Interference-Coupled Cellular Data Networks

Henrik Klessig, *Member, IEEE*, David Öhmann, *Member, IEEE*, Albrecht Fehske, *Member, IEEE*, and Gerhard Fettweis, *Fellow, IEEE*

***Abstract*—In regard to the continuing network densification as a part of the solution to the mobile data traffic demand explosion, managing future 5G ultra-dense networks is becoming increasingly complex. Moreover, the problem of (partly) limited capacity in time and space requires the joint treatment of spatio-temporal data traffic and inter-cell interference dynamics. Concerning this matter, we propose a performance evaluation framework, which is capable of estimating various cell-specific and user-specific key performance metrics considering the complex spatio-temporal interaction of traffic and interference dynamics. We provide methods to obtain these metrics with low complexity, making the framework attractive to be applied to self-organizing network solutions for future (ultra-)dense networks. We stress on the framework's broad applicability and demonstrate the effects of internal flow and external interference dynamics on network performance under various conditions. In particular, we highlight the dominance of these dynamics over the impact of the speed of the variation of inter-cell interference, the scheduler, the file size distribution, and fast fading.**

***Index Terms*—performance estimation; flow level modeling; interference coupling; network densification; wireless networks**

I. Introduction

During the last decade, the focus of cellular network planning and optimization has changed dramatically. On the one hand, key performance indicators (KPIs) for network evaluation have changed from rather technical metrics, such as the signal-to-interference-plus-noise ratio (SINR), to more quality-of-service (QoS)- and quality-of-experience (QoE)-related metrics. On the other hand, network optimization has profited a lot by the development of very useful and detailed network models that, again, allow the use of more advanced performance metrics. In the following, we want to shed light on the aforementioned changes and try to focus on inter-cell interference and mobile data traffic dynamics, which should be the key aspects in the derivation of the performance evaluation framework proposed.

A. The Need for Scalable Network Models

The ever-increasing mobile traffic demand imposes tough challenges on network planning and research on novel radio access technologies. Due to this trend, today's mobile networks are usually capacity-limited, which can be experienced most intensively at traffic hot spots, where radio resources for individual users are scarce. A limitation on capacity generally relates to a highly loaded system and, therefore, to an interference-limited rather than a noise-limited system. In addition, since mobile traffic demand is likely to fluctuate strongly in time and space, see e. g. [1], limited capacity and the impact of inter-cell interference are anything but static phenomena, neither in space nor in time. Hence, both, the spatio-temporal behavior of the traffic demand and effects on network performance induced by dynamic inter-cell interference, have to be considered jointly.

In the future, fifth generation (5G) cellular networks will only be able to cope with the *limited capacity*-problem through a densification in space (ultra-dense small cell deployments, massive MIMO (multiple-input multiple-output)) and a densification in frequency (novel waveform design, millimeter wave communications), see e. g. [2], [3]. In addition, spatio-temporal data traffic dynamics call for very flexible and adaptive radio network solutions, as well as autonomous self-planning, self-optimization, and self-healing capabilities. However, the rapidly increasing number of cells and mobile devices per unit area demand smart network models that enable an efficient performance estimation and, based thereon, an effective network optimization with low complexity. Consequently, such network models require the following three features:

1) Scalability and low complexity for the examination of large-scale networks with hundreds of nodes,
2) Performance evaluation with a paradigm shift from rather technical metrics to user-specific metrics (QoE),
3) Ability to capture capacity-related aspects, including dynamic inter-cell interference and mobile data traffic fluctuations, rather than coverage-related aspects.

B. Mobile Data Traffic and Interference Modeling

The main idea of developing more advanced network, traffic, and interference models is to be able to evaluate and predict network performance, develop optimization algorithms, and describe wireless network characteristics more accurately. In regard to a more accurate description of user-specific performance indicators, the notion of *elastic data flows* [4], [5], in combination with each base station (BS) representing a server in a queuing system, has proved to be very effective. There has been extensive work on modeling of and investigating on the effects of flow level dynamics in queuing systems with processor sharing discipline in a mobile cellular context: For example, the authors in [6] and [7] investigate the effect of different schedulers and user mobility at flow level, respectively. Capacity gains of frequency reuse schemes have been analyzed in [8] and flow-level performance of buffered streaming services have been modeled and investigated in [9].

A first attempt to capture the complex interactions of flow-level dynamics and inter-cell interference has been made

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2015.2480760, IEEE Transactions on Wireless Communications

in [10], where second-degree approximations on the mean number of active flows have been introduced. More specifically, the authors propose using upper and lower performance bounds (with respect to the speed of variation of interference) and assume, that interfering BSs provide either maximum data rates or minimum data rates by experiencing either no or full interference, respectively. However, the drawback of this solution is that only the performance of one cell under consideration can be approximated neglecting its own impact on the performance of surrounding BSs.

Recently, modeling the mutual coupling among interfering BSs has gained momentum again due to the reasons explained in the previous section. The authors in [11]–[13] introduced and analyzed a fixed point solution for deriving the average cell loads under time-averaged interference conditions based on interference function calculus [14]. This cell load model has been extended for systems with admission control in [15] and with non-stationary traffic conditions in [16].

C. Contributions of this Article and Novel Aspects

In this article, we make the following contributions:

1) We provide a generic framework to model cellular wireless networks, which can be adapted to a wide range of wireless technologies, architectures, deployment strategies, and dynamic mechanisms. In addition to the network model, we recommend using a traffic model based on the common notion of *data flows*.
2) Using the aforementioned network and traffic models, we derive a comprehensive set of KPIs that capture the effects of flow level dynamics within, as well as, dynamic interference among cells. We provide two techniques to compute this set of KPIs with reduced computational complexity.
3) We study the impact of the traffic load within the cell under consideration and in surrounding cells, as well as, the effect of different schedulers through system level simulations. Further, we compare the results to the theoretical model proposed and highlight the impact of the speed of the variation of inter-cell interference.

Parts of this work have been published in [15]–[18]. We summarize the results therein and add value to them by the following aspects: (1) We collect different modeling aspects from the aforementioned papers and formulate the mathematical derivation in a more rigorous manner in order to provide more theoretical insights. (2) We derive stability conditions, under which different performance computation methods are valid. (3) Three additional metrics are introduced, namely, a general description of the cell capacity, throughput-dependent network coverage, and user throughput statistics. (4) The general applicability of the framework to various network technologies is emphasized more elaborately and shown by a comparison with simulation results under various conditions. (5) We analyze the impact of admission control, the speed of variation of inter-cell interference, the scheduler, the files size distribution, and fast fading on the accuracy of the model. More specifically, we highly that the flow dynamics predominate the network performance in most cases.

D. Notations and Organization of the Article

Throughout this article, we use calligraphic symbols, such as $\mathcal{N}$, for sets, $\mathbb{N}_{>0}$ for the set of strictly positive integers, $\mathbb{R}$, $\mathbb{R}_{+}$, and $\mathbb{R}_{>0}$ for the sets of real numbers, positive real numbers, and strictly positive real numbers, respectively. The variables $t \in \mathbb{R}_{+}$ and $u \in \mathbb{R}^2$ denote the time variable and a spatial coordinate, respectively. The expectation operator is $\mathbb{E}[\cdot]$ and the probability operator is $\mathbb{P}[\cdot]$. Further, we use the sign function $\mathrm{sgn}(\cdot)$, as well as, the unit vector e_n, with the n^{th} element equal to one. We also use a vector $\mathbf{1}_k := (1, \ldots, 1)^{\mathrm{T}}$ of ones with length $|\mathbf{1}_k| = k$.

In Section II, we present the underlying network and traffic models, and elaborate on their wide applicability. A detailed derivation of KPIs follows in Section III and methods to obtain them are given in Section IV. Simulation results validate the models in Section V and conclusions are given in Section VI.

II. General System Model

In the following, we provide a general definition of a cellular network in the downlink along with important assumptions on the dynamics of mobile data traffic and inter-cell interference.

A. Cellular Network Model

We consider a region $\mathcal{L} \subset \mathbb{R}^2$, in which $N \in \mathbb{N}_{>0}$ ($N < \infty$), BSs with indices $i \in \mathcal{N} := \{1, \ldots, N\}$ are located. Each BS i covers a region $\mathcal{L}_i$, depending on its type with respect to transmit power, antenna type, height mounted etc. (which are known to be macro, micro, pico, or femto BSs), or depending on environmental conditions, such as deployment density, deployment at indoor or outdoor locations, and shadowing. We assume that, at some given point in time t, a mobile user at location $u \in \mathcal{L}$ is uniquely associated to one specific BS, such that we have $\mathcal{L}_i \cap \mathcal{L}_j = \varnothing$ for all $i, j \in \mathcal{N}, i \neq j$, and $\bigcup_{i \in \mathcal{N}} \mathcal{L}_i = \mathcal{L}$. In the remainder, we will use the terms *cell* and *base station* interchangeably. Note that this definition of a cellular network is rather general, which allows to apply it to a variety of different technologies, such as multi-layer and multi-RAT (radio access technology) networks, heterogeneous networks, or to adopt different user association policies.

B. Traffic Model and Time Scale Considerations

In general, it is very important to make assumptions on and to draw conclusions from specific time scale considerations in order to evaluate network performance properly. This crucially depends on what is to be investigated. For example, to analyze user scheduling, one has to take into account a proper modeling of fast fading and link level performance, which may be cumbersome when analyzing performance of networks of hundreds of nodes. The time scales considered in this article are that of data flow dynamics, which are useful for investigating long-term system performance statistics.

1) Data Flow and User Session Model: Usually, mobile users initiate data transfers randomly in time with random file sizes. Upon arrival of such a data transfer, the download process starts and lasts until the file has been downloaded completely. A common notion of this behavior is that of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2015.2480760, IEEE Transactions on Wireless Communications

elastic data flows and *user sessions* [4], which proves to be more appropriate than investigations on packet level when it comes to quantifying mobile user experience and network performance. A user session is a series of consecutive data flows of one user with arbitrary thinking time between the individual flows. In general, elastic data flows are packets that belong to specific objects like mp3 files, web pages, buffered streaming files, etc. Provided that user sessions arrive independently as Poisson, and that there is a reasonably large user population, as well as, a specific user behavior in case of admission control, the arrival of data flows can also be modeled as a Poisson process, see e. g. [5], [19]–[21]. In the remainder, we adopt these insights and consider data flows, the arrival of which is modeled as Poisson, only. Thus, we will use the terms *user* and *data flow* interchangeably.

Let us consider an integrable, normalized, and time-dependent user density distribution $\delta(u,t)$ with $\int_{\mathcal{L}} \delta(u,t)\,\mathrm{d}u = 1$, and an overall arrival rate $\lambda(t)$ of data flows that arrive to the entire system. Then the product of both, the user density and the overall arrival rate, represents the spatially distributed rate that determines the arrival process of flows in the region $\mathcal{L}$. Integration over cell areas $\mathcal{L}_i$ yields

$$\lambda_i(t) := \int_{\mathcal{L}_i} \lambda(t)\delta(u,t)\,\mathrm{d}u, \tag{1}$$

which describes the time-variant flow arrival rate to BS i in flows per second. Another common assumption is that data flow sizes ω are exponentially distributed. Let Ω be the mean size of a flow in bits. Typically, flow sizes range from a few hundreds of kilobits to several tens of Megabits, such that transfers last from a few hundreds of milliseconds to several seconds considering today's cellular technologies and applications.

2) User Mobility, Slow Fading, and Handover Mechanisms: We assume that the majority of the traffic is generated by stationary (e. g. indoor) and slow pedestrian users. Since handovers are mainly caused by user mobility along with the slow fading process, we omit modeling handover procedures, which is motivated by the following example: Consider a mobile user experiencing a rather low throughput of 2 Mbps and downloading a file of about 1 Mbyte. Assuming an average pedestrian speed of 1.4 meters per second, then the user will travel about 5.6 meters during the download process, which is far smaller than typical shadowing correlation distances in suburban and urban environments (25 - 100 m) [22]. This motivates to omit the effect of varying shadowing on receive power conditions during individual data transfers, as well.

3) Rayleigh Fading Effects: Using the same example from above and assuming a carrier frequency of 1.9 GHz, which is typical for Long Term Evolution (LTE) networks in urban regions, we obtain a channel coherence time of about 48 milliseconds. Compared to the total duration of the download, the channel coherence time is much smaller, such that we can consider the impact of fast fading as a time-average effect on the average receiving conditions and, hence, on the flow level performance. A detailed modeling of the impact of Rayleigh fading on flow level performance can be found in [7]. In addition, the authors in [9] observe that fast fading has a minor effect on buffered streaming performance than flow level dynamics. The aforementioned thoughts finally lead us to the following important assumption:

Assumption 1 (Static receiving conditions)**.** During the transmission of a flow, receiving conditions, i. e., the signal strength, remain constant.

As a result, we use location-dependent, but time-independent, receive powers $p_i(u)$ incorporating path loss effects, shadowing, antenna patterns, penetration losses, etc. We will support this assumption by simulation results in Section V-B.

C. Radio Link Quality and Cell Capacity

In contrast to slow and fast fading effects, the performance of the transfer of one data flow is strongly affected by other competing data flows in the same cell, as well as, dynamic inter-cell interference through surrounding cells, which is driven by the random presence of data flows. In order to characterize both, we make the assumption of best effort service through and the activity of a BS:

Assumption 2 (Best effort service)**.** If there is at least one active flow in a cell, the corresponding BS is said to be *active*. Then, it allocates all radio resources and, therefore, transmits with full power, as long as all flows have been served.

Let us consider a set $\mathcal{Y} := \{0,1\}^N$, where the vector $y \in \mathcal{Y}$ denotes the activities of the BSs. The i^{th} element of the vector y, y_i, is one if the i^{th} BS is active, otherwise zero. Then, we can collect the indices of all inactive and active BSs in the sets $\mathcal{N}_0(y) := \big\{i = 1,\dots,N \mid y_i = 0\big\}$ and $\mathcal{N}_1(y) := \big\{i = 1,\dots,N \mid y_i = 1\big\}$, respectively. Now, given a certain interference scenario y and using the Shannon-Hartley theorem, we can compute the SINR γ_i and the maximum achievable rate c_i by

$$\begin{aligned} \gamma_i(u,y) &:= \frac{p_i(u)}{\sum_{j\in\mathcal{N}_1(y)\setminus\{i\}} p_j(u) + N_0} \text{ and} && (2)\\ c_i(u,y) &:= aB\min\Big\{\log_2\big(1 + b\gamma_i(u,y)\big), c_{\text{max}}\Big\}, && (3)\end{aligned}$$

respectively. The noise power is denoted as N_0 and the bandwidth is given by B. Bandwidth and SINR efficiencies [23] are incorporated in the quantities a and b, respectively. The quantity c_{max} provides the maximum data rate determined by the highest modulation and coding scheme of the wireless technology at hand.

So far, we are able to quantify the maximum data rate $c_i(u,y)$ experienced by a user at location u if connected to BS i. In order to characterize the user throughput including the fact that a user shares radio resources with other competing users, we need to know the average data rate provided by BS i. Let us first take into consideration the normalized user distribution in cell i given by

$$\delta_i(u,t) := \frac{\delta(u,t)}{\int_{\mathcal{L}_i}\delta(u,t)\,\mathrm{d}u}. \tag{4}$$

Then, provided that there are active users in the cell, the average rate provided by BS i can be formulated as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2015.2480760, IEEE Transactions on Wireless Communications

the user density-weighted harmonic mean of the achievable rates $c_i(u, y)$, i. e.,

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

For a more stringent explanation for and elaborate derivation of the cell capacity as the weighted harmonic mean, we refer to [6]. This definition of the cell capacity $C_i(y, t)$ is a very important step to characterize the system performance, in particular with respect to queuing-theoretic analyses, which will be described in Section III-A.

D. User Scheduling and Admission Control

In this article, we stick to a resource-fair scheduler. The most popular resource-fair scheduler is a Round Robin scheduler, where radio resources are shared fairly among active users, e. g., by allocating time resources in a repetitive manner. Though using a Round Robin scheduler appears to be very restrictive, the model and results can be extended to scenarios with different schedulers easily. In particular, incorporating scheduling gains in the flow model itself is shown in [6]. Another, less intricate approach has been considered in [24], where the authors include scheduling gains in the bandwidth and SINR efficiencies, a and b, respectively. Nevertheless, we will compare the resource-fair and throughput-fair schedulers in Section V by system level simulations.

Since the user arrival and departure processes are random, the number of concurrently active users connected to a BS is also random. There may be cases, where this number becomes very high and individual user throughput degrades strongly. As a result, network operators employ admission control mechanisms in their BSs. We use a simple, yet realistic, admission control policy by restricting the maximum number of active flows in cell i to $K_i \in \mathbb{N}_{>0}$, and explicitly include the case where $K_i \to \infty$, i. e., where all flows are admitted service. For an overview of more advanced (adaptive) admission control mechanisms, we refer the interested reader to [25].

E. Wireless Technologies Applicable

Since the network model presented is rather general, it can be extended or modified easily to different cellular technologies, in particular to any OFDMA (Orthogonal Frequency Division Multiple Access)-, TDMA (Time Division Multiple Access)-, or FDMA (Frequency Division Multiple Access)-based radio access technology. It is applicable to frequency reuse one networks or to fractional frequency reuse patterns, and it is generic with respect to the network topology.

Flow level models as the one presented in this article have been shown to be very useful for algorithm design, in particular for bandwidth allocation schemes [26], load balancing [27], antenna optimization [28], user association [29], data offloading with small cells [30], [31], or interference coordination [32], [33]. Further use of flow level models has been made in [34] and [35]. The aforementioned promising algorithms give rise for expecting analyses from the flow-level perspective to be part of planning and (self-)optimization of future 5G cellular networks.

III. Key Performance Indicators

In the next section, we derive important network-specific and user-specific KPIs based on a queuing-theoretic modeling of BSs with the cell capacities given in Eq. (5) as starting point.

A. Queuing-Theoretic Preliminaries

In the following, we consider one cell in isolation and in a fixed interference scenario y. Given the mean file size Ω, we compute the average rate, with which users in cell i are served, or in other words, leave the system, by

$$\mu_i(y,t) = \frac{C_i(y,t)}{\Omega}. \tag{6}$$

We can now interpret each BS i as a queuing system with arrival rate $\lambda_i(t)$ and service rate $\mu_i(y, t)$. Considering the fact that we stick to the Round Robin scheduler, we choose the Egalitarian Processor Sharing (EPS) service discipline as its queuing-theoretic counterpart. Usually, the service rate $\mu_i(y, t)$ varies according to dynamic inter-cell interference induced by surrounding BSs, such that the queue lengths are mutually coupled. However, let us first formulate the queuing-theoretic basics for decoupled queues in a static interference regime.

1) Decoupled Queues through Static Interference: In particular, we are interested in the probability distribution of the number of concurrently active flows in each of the cells. Let $X_i(y, t)$ be a random process describing the number of active flows in cell i under interference scenario y, that is

$$X_i(y,t) \quad \in \quad \{0, \dots, K_i\} . \tag{7}$$

Given the aforementioned queuing-theoretic assumptions, we can compute the time-variant state probabilities

$$\pi_i(x_i, y, t) := \mathbb{P}\left[X_i(y,t) = x_i \mid y\right] \tag{8}$$

by solving the following differential equation system, also called system of *balance equations*,

$$\begin{aligned} \frac{\mathrm{d}\pi_i(x_i,y,t)}{\mathrm{d}t} \quad = \quad & \lambda_i(t)\pi_i(x_i-1,y,t) \\ & - \big(\lambda_i(t)+\mu_i(y,t)\big)\,\pi_i(x_i,y,t) \\ & + \mu_i(y,t)\pi_i(x_i+1,y,t). \end{aligned} \tag{9}$$

The system above basically describes the evolution of the state probabilities of a Markov process as depicted in Fig. 1.

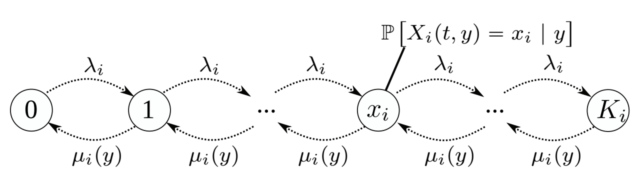


Fig. 1. Markov process for a fixed interference scenario y. Here, for the stationary traffic process.

In the next step, we are interested in a steady-state analysis and assume a stationary user arrival process with $\lambda_i(t) := \lambda_i$ and $\mu_i(y, t) := \mu_i(y)$ for all t. We have the following definition:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2015.2480760, IEEE Transactions on Wireless Communications

Definition 1 (y-Stability of a Base Station)**.** *A BS is said to be* y-stable *if the number of active flows does not diverge, i. e.,*

$$\limsup_{t\to\infty} X_i(y,t) < \infty. \tag{10}$$

Note that, for finite K_i, BS i is stable by definition, since admission control ensures a finite number of concurrently active flows. However, we introduce this definition, because we explicitly consider the case, where no admission control is employed, as well. In the steady state, that is for $\frac{\mathrm{d}\pi_i(x_i,y,t)}{\mathrm{d}t} = 0$, Eqs. (9) reduce to the well-known M/M/1/K-EPS product form

$$\lim_{t\to\infty} \pi_i(x_i,y,t) = \frac{\big(1-\rho_i(y)\big)\,\rho_i(y)^{x_i}}{1-\rho_i(y)^{K_i+1}}, \tag{11}$$

where $\rho_i(y) := \lambda_i/\mu_i(y)$ is referred to as the load of a BS. If BS i does not employ admission control, meaning $K_i \to \infty$, Eq. (11) is only valid if this BS is y-stable. In this case, the stationary state probabilities become

$$\lim_{t\to\infty} \pi_i(x_i,y,t) = \big(1-\rho_i(y)\big)\,\rho_i(y)^{x_i}. \tag{12}$$

In general, instability of such a system is caused by the fact that the service rate $\mu_i(y)$ is not large enough to serve the incoming data requests given by the rate λ_i, in particular when $\mu_i(y) \leqslant \lambda_i$. From the derivations in Eqs. (2)-(6), we observe that $\mu_i(y)$ is minimized for $y = \mathbf{1}_N$. In this regard, we have the following definition of strict stability of the network:

Definition 2 (Strict Stability of a Network)**.** *A network composed of BSs is said to be* strictly stable *if, for each BS* $i \in \mathcal{N}$*, one of the following two conditions hold, either*

1) K_i *is finite, or*
2) *for* $K_i \to \infty$*, BS* i *is* y*-stable for* $y = \mathbf{1}_N$*.*

Otherwise, the system is said to be (potentially) unstable.

We will need the definition of strict stability, when we solve the steady-state probabilities for the case of coupled queues, which are discussed next.

2) Coupled Queues through Dynamic Interference: So far, the implications of Assumption 2 with respect to inter-cell interference dynamics have not been considered. Let $\mathcal{X} := \{0,\dots,K_1\} \times \dots \times \{0,\dots,K_N\}$ be the state space of the network. Then, the vector process $X(t)$ describes the number of active flows in all BSs, that is

$$X(t) := \big(X_1(t),\dots,X_N(t)\big) \in \mathcal{X}. \tag{13}$$

Note that, for the moment, we omitted the dependency of X_i on y. Introducing $Y(t) \in \mathcal{Y}$ as the random vector process describing the interference dynamics in the network and following Assumption 2, we can write $Y(t) = \mathrm{sgn}\big(X(t)\big)$. We will use the vector x as the realization of the process $X(t)$ and write y instead of $\mathrm{sgn}(x)$ for better readability. Now, it becomes clear, why we omitted the interference scenario y in Eq. (13), that is the interference situation is implicitly determined by the vector process $X(t)$ itself.

The rate of serving a particular data flow by a certain BS is *modulated* by the interference processes of surrounding BSs. The fact that this also affects the interference process of the BS under consideration causes the same effect in surrounding BSs, and so on. In other words, the evolution of the state probabilities with respect to one BS is not only dependent on the flow dynamics within the corresponding cell, but also on the flow dynamics in all other surrounding cells through the process $Y(t)$. This is the essence of the interference coupling behavior, which we want to capture in our derivations.

The problem at hand relates to performance evaluation of so-called queues with modulated service rates. In general, it is cumbersome to obtain the state probabilities of such queues and only few results, for example for M/MM/1-FCFS (MM: Markov-modulated; FCFS: first come first serve) queues exist, see e. g. [36], [37]. Particularly in our case, it appears to be extremely difficult to obtain the exact state probabilities due to two reasons. First, we consider the EPS service discipline, such that the rate modulation applies to multiple data flows at different locations and at the same time. Second, in contrast to [37], where the authors consider one independent environment process modulating the service rate, we have N processes with *mutual* service rate modulation.

Let us, for now, assume that we have already obtained the state probabilities

$$\pi(x,t) := \mathbb{P}\big[X(t) = x\big] \tag{14}$$

of the vector process $X(t)$. Network- and user-specific KPIs can only be obtained meaningfully for stable systems. In this regard, we give the following definition for the stability of the network for a stationary user arrival process:

Definition 3 (Stability of a Network)**.** *A network is said to be* stable, *if*

$$\limsup_{t\to\infty} X(t) < \infty. \tag{15}$$

Otherwise, the system is said to be unstable.

B. Network-Specific Key Performance Indicators

In what follows, we assume stationarity in a reasonably large interval around time instant t. Then, for a stable network, the PASTA (Poisson Arrivals See Time Averages) property [38] along with Assumption 2 allow for the computation of the expected BS resource utilization, the expected probability of blocking a data flow request, and the expected mean number of concurrently active data flows, i. e.

$$\eta_i(t) := \quad \mathbb{P}\big[X_i(t) > 0\big] \quad = \sum_{\substack{x\in\mathcal{X}\\ x_i \geqslant 1}} \pi(x,t), \tag{16}$$

$$P_{\mathrm{b},i}(t) := \quad \mathbb{P}\big[X_i(t) = K_i\big] \quad = \sum_{\substack{x\in\mathcal{X}\\ x_i = K_i}} \pi(x,t), \tag{17}$$

$$n_i(t) := \quad \sum_{j=1}^{K_i} j\,\mathbb{P}\big[X_i(t) = j\big] \quad = \sum_{j=1}^{K_i} \sum_{\substack{x\in\mathcal{X}\\ x_i = j}} j\,\pi(x,t), \tag{18}$$

respectively. Further, using Little's law [39], we have the expected time duration a user spends in the system, i. e. the mean sojourn time,

$$T_i(t) = \frac{n_i(t)}{\lambda_i(t)\,\big(1-P_{\mathrm{b},i}(t)\big)}. \tag{19}$$

Let us consider the quantity $C_i(t)$ to be the expected cell capacity (average data rate provided) with respect to the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2015.2480760, IEEE Transactions on Wireless Communications

varying interference. Given the fact that, the cell throughput is the utilized cell capacity, i. e. the product $\eta_i(t)C_i(t)$, and that it is equivalent to the product of effective arrival rate and average flow size, $\lambda_i(t)\left(1 - P_{\mathrm{b},i}(t)\right)\Omega$, we have the expression

$$C_i(t) = \lambda_i(t)\left(1 - P_{\mathrm{b},i}(t)\right)\Omega/\eta_i(t). \tag{20}$$

Note that for BSs without admission control ($P_{\mathrm{b},i}(t) = 0$), the cell capacity relates to the maximum traffic demand $\lambda_i(t)\Omega$, for which the BS is stable ($\eta_i(t) < 1$).

As a last step, we want to quantify the average flow throughput in a cell, which can be defined as the expectation of the ratio of the file size and the corresponding time spent in the system. Let $\xi_i^\omega(t)$ be the random variable for the sojourn time of a flow with size ω. The average flow throughput would be $\mathbb{E}\left[\frac{\omega}{\xi_i^\omega(t)}\right]$. However, obtaining the distribution of sojourn time is cumbersome even for the well-studied M/G/1-PS queue [40]. Instead, we consider the approximation $\mathbb{E}\left[\frac{\omega}{\xi_i^\omega(t)}\right] \approx \frac{\mathbb{E}[\omega]}{\mathbb{E}\left[\xi_i^\omega(t)\right]}$, where the term $\mathbb{E}\left[\xi_i^\omega(t)\right]$ is equivalent to the mean sojourn time given in Eq. (19). Consequently, we have the flow throughput

$$R_i(t) := \frac{\Omega}{T_i(t)} = \frac{\eta_i(t)C_i(t)}{n_i(t)}. \tag{21}$$

Remark 1. Though, expression (21) is an approximation, it helps quantifying the average user performance in a cell. Note that, for the M/G/1-PS queue, R_i has been shown to be a lower bound for the exact value $\mathbb{E}\left[\frac{\omega}{\xi_i^\omega(t)}\right]$ and to fit well for reasonably large file sizes, see [5], [40].

C. *User-Specific Key Performance Indicators*

In addition to network-specific metrics and cell-average performance indicators in Eqs. (16)-(21), the framework proposed also allows for user-specific performance evaluation. In particular, this is of high importance, since users experience different network quality across the entire cell with decreased performance at cell edges, which is reflected in the location-dependent achievable rates $c_i(u, y)$. Since an exact investigation is very cumbersome, we consider two regimes that account for the extreme cases with respect to the speed of variation of the interference dynamics, similar to the analyses in [10]. In the so-called *quasi-stationary* regime, interference dynamics induced by surrounding BSs are infinitely slow compared to the flow dynamics within the cell under consideration. In the *fluid* regime, the speed of interference variations are assumed to be infinitely fast.

Let us collect all network states x for a given interference scenario y in an aggregate set $\mathcal{X}(y)$, more formally $\mathcal{X}(y) = \left\{x \in \mathcal{X} \mid \operatorname{sgn}(x) = y\right\}$. Then, the probability that an active flow, connected to BS i, observes interference situation y is

$$\zeta_i(y,t) = \mathbb{P}\left[\operatorname{sgn}\left(X(t)\right) = y \mid X_i(t) > 0\right] \tag{22}$$

$$= \frac{1}{\eta_i(t)} \sum_{\substack{x \in \mathcal{X}(y) \\ x_i \geqslant 1}} \pi(x,t). \tag{23}$$

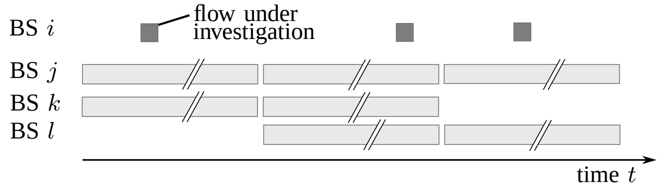


Fig. 2. Data flow observing dynamic interference in the quasi-stationary regime.

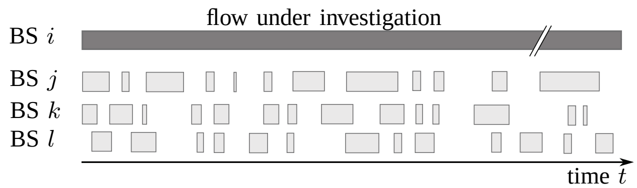


Fig. 3. Data flow observing dynamic interference in the fluid regime.

1) Quasi-Stationary Regime: As illustrated in Fig. 2, we assume that each data flow observes one interference scenario during the entire transmission process but time-average conditions in its own cell, such that its expected (with respect to the flow size) sojourn time at location u is $\frac{n_i(t)\Omega}{\eta_i(t)c_i(u,y)}$. Let $\mathcal{A}_i = \left\{y \in \mathcal{Y} \mid y_i = 1\right\}$ be the set collecting all interference scenarios, where BS i is active. Now, taking the expectation with respect to the interference vector process $Y(t)$, conditioned on the existence of at least one active flow in cell i, results in

$$\tau_i^{\mathrm{qs}}(u,t) = \sum_{y \in \mathcal{A}_i} \frac{n_i(t)\Omega}{\eta_i(t)c_i(u,y)} \zeta_i(y,t). \tag{24}$$

Following the same approach as for the expected user throughput in a cell, see Eq. (21), we obtain the expected flow throughput at location u as $r_i^{\mathrm{qs}}(u,t) := \Omega/\tau_i^{\mathrm{qs}}(u,t)$.

2) Fluid Regime: In the fluid regime, we consider flows observing all interference scenarios infinitely often, see Fig. 3. Hence, we approximate their average achievable rates by

$$c_i^{\mathrm{fl}}(u,t) = \sum_{y \in \mathcal{A}_i} c_i(u,y)\zeta_i(y,t), \tag{25}$$

such that we write for the expected flow sojourn time

$$\tau_i^{\mathrm{fl}}(u,t) = \frac{n_i(t)\Omega}{\eta_i(t)c_i^{\mathrm{fl}}(u,t)}, \tag{26}$$

and for the expected flow throughput $r_i^{\mathrm{fl}}(u,t) := \Omega/\tau_i^{\mathrm{fl}}(u,t)$, respectively. Note that, by the definitions given above, we have $\tau_i^{\mathrm{qs}}(u,t) \geqslant \tau_i^{\mathrm{fl}}(u,t)$ and $r_i^{\mathrm{qs}}(u,t) \leqslant r_i^{\mathrm{fl}}(u,t)$, where equality holds if, for all $j \neq i$, $Y_j(t)$ is almost surely constant or if $c_i(u,y)$ is independent of y, for example in the cell center.

Remark 2. The quasi-stationary regime equations may be used, when there are only very few changes in the interference scenario during the transmission of a flow. This is for example the case when the flow sizes in neighboring cells are much larger than that of the flow under consideration. In contrast, the fluid regime equations may be more appropriate, when there are many changes of the interference scenario during the transmission of a flow.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2015.2480760, IEEE Transactions on Wireless Communications

D. Throughput Statistics and Network Coverage

Network operators are specifically interested in user throughput statistics and network coverage. Though, the classical definition of coverage is defined as area or traffic covered through a minimum receive power, another interesting definition is coverage with respect to a minimum throughput threshold r_{min}. Let us formulate the spatial region, in which the minimum throughput requirement is fulfilled, by

$$\mathcal{L}_{\mathcal{C}}(t, r_{\text{min}}) = \left\{ u \in \mathcal{L} \mid r_i^{\text{qs/fl}}(u,t) \geqslant r_{\text{min}}; i \text{ s.t. } u \in \mathcal{L}_i \right\}. \tag{27}$$

The fraction of traffic demand covered is given by

$$\mathcal{C}(t, r_{\text{min}}) = \int_{\mathcal{L}_{\mathcal{C}}(t, r_{\text{min}})} \delta(u,t)\, \mathrm{d}u. \tag{28}$$

The relation between r_{min} and $\mathcal{C}$ can also be expressed as follows. The $\big(100(1-\mathcal{C})\big)^{\text{th}}$ percentile of the user throughputs, averaged with respect to the flow and interference dynamics, equals r_{min}. Finally, this interpretation allows us to analyze the entire throughput statistics in the network or per cell, respectively. We will show corresponding model and simulation results in Section V.

IV. Methods to Obtain State Probabilities

So far, we have defined a variety of important KPIs provided that we know the state probabilities $\pi(x,t)$. However, obtaining $\pi(x,t)$ is not a trivial task. In the following, we will provide two techniques, called *state aggregation* and *average interference*, and elaborate on special cases, where particular techniques are not suited.

A. General Approach

Both methods consist of two steps. The first step, in which both techniques differ, is the approximation of the probability $\zeta_i(y,t)$ that an active flow sees interference scenario y (see Eq. (23)) using the formulas for the decoupled state probabilities from Section III-A1. We will show this step in Sections IV-B and IV-C for the state aggregation and the average interference methods, respectively.

The second step is to approximate the average service rate $\mu_i(t)$ for each of the BSs i from the approximated probabilities $\tilde{\zeta}_i(y,t)$, that is

$$\mu_i(t) := \frac{1}{\Omega} \left[\int_{\mathcal{L}_i} \frac{\delta_i(u,t)}{\sum_{y \in \mathcal{A}_i} \tilde{\zeta}_i(y,t) c_i(u,y)} \, \mathrm{d}u \right]^{-1}. \tag{29}$$

Note that the service rate $\mu_i(t)$ is computed as if all flows in the cell i would observe the external interference process in the fluid regime. Now, we substitute the interference scenario-dependent service rates $\mu_i(y,t)$ in Eq. (9) by $\mu_i(t)$ to obtain decoupled state probabilities $\pi_i(x_i,t)$, which are now independent of the interference process, as well. Finally, we approximate the state probabilities of the vector process $X(t)$ by

$$\pi(x,t) \approx \prod_{i \in \mathcal{N}} \pi_i(x_i,t). \tag{30}$$

B. Approximation by State Aggregation

Here, we only sketch the state aggregation approach and present the main result. We refer to [17], [18] for a more elaborate derivation of the approximate state probabilities from so-called balance equations that capture the evolution of the state probabilities $\pi(x)$. Let us again consider the state aggregates with fixed interference scenario y, $\mathcal{X}(y)$, from Section III-C and recall that the service rates $\mu_i(y,t)$ remain constant within one aggregate. Assuming much more state transitions within aggregates than among different aggregates, we consider flow dynamics within and among aggregates separately. On this basis, we presume that flow dynamics in one cell evolve independently of the dynamics within other cells and focus on the transitions among aggregates.

The transition rates $p(y,y',t)$ between each pair of aggregates, $\mathcal{X}(y)$ and $\mathcal{X}(y')$, can be described as

$$p(y,y',t) = \begin{cases} \lambda_i(t) & \text{for } y' = y + e_i, \\ \mu_i(y,t) \frac{\pi_i(1,y,t)}{1-\pi_i(0,y,t)} & \text{for } y' = y - e_i, \\ 0 & \text{else.} \end{cases} \tag{31}$$

As can be seen, the transition rate from one aggregate to another by activating a BS n still remains the arrival rate λ_i; however, the rate, with which a BS is deactivated, becomes the service rate $\mu_i(y,t)$ multiplied by the probability of exactly one flow active conditioned on the fact that there is at least one flow active in cell n. Fig. 4 depicts the approach, which targets to collapse the entire state space of cardinality $|\mathcal{X}|$ into another state space of cardinality $|\mathcal{Y}|$.

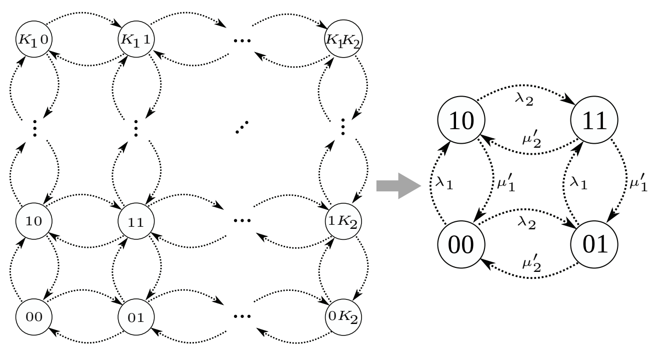


Fig. 4. Reducing the state space by state aggregation, with $\mu_i'(y,t) = \mu_i(y,t) \frac{\pi_i(1,y,t)}{1-\pi_i(0,y,t)}$. Here for an example with two coupled BSs, i. e., a two-dimensional Markov chain.

Let $\sigma(y,t)$ be the aggregate probabilities. Now, consider the transition matrix $P(t) := \big(p_{ij}(t)\big)$ with arbitrary ordering $i \rightarrow y(i)$, $p_{ij}(t) := p\,\big(y(i), y(j), t\big)$, and $p_{ii}(t) := -\sum_{\substack{i=1 \\ j \neq i}}^{2^N} p_{ij}(t)$. Then, we solve the system of differential equations

$$\frac{\mathrm{d}\sigma(y,t)}{\mathrm{d}t} = \sigma(y,t) P(t) \tag{32}$$

for the aggregate probabilities $\sigma(y,t)$, from which we finally derive the approximate probabilities of an active flow observing interference scenario y by

$$\tilde{\zeta}_i(y,t) := \frac{\sigma(y,t)}{\sum_{y' \in \mathcal{A}_i} \sigma(y',t)}. \tag{33}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2015.2480760, IEEE Transactions on Wireless Communications

C. Approximation by Time-Averaged Interference

Another elegant way of approximating the state probabilities is by assuming time-averaged interference. The approach is the following: Presume knowledge about the approximate probabilities of BS activity, $\bar{\eta}_j(t) :\approx \mathbb{P}\left[X_j(t) > 0\right]$. Then, the time-averaged interference power is the product of $\bar{\eta}_j(t)$ and the receive power $p_j(u)$, such that the SINR becomes

$$\bar{\gamma}_i\left(u,t,\bar{\eta}(t)\right) = \frac{p_i(u)}{\sum_{j\neq i}\bar{\eta}_j(t)p_j(u) + N_0} \tag{34}$$

and the achievable data rate becomes

$$\bar{c}_i\left(u,t\right) = aB\min\left\{\log_2\left(1+b\bar{\gamma}_i\left(u,t,\bar{\eta}(t)\right)\right), c_{\max}\right\}, \tag{35}$$

respectively, where $\bar{\eta}(t) := \left(\bar{\eta}_1(t),\dots,\bar{\eta}_N(t)\right)^{\mathrm{T}}$. Following the exact same approach as in Section II-C to derive the cell capacity, we obtain

$$\bar{C}_i\left(t\right) := \left[\int_{\mathcal{L}_i}\frac{\delta_i(u,t)}{\bar{c}_i\left(u,t\right)}\,\mathrm{d}u\right]^{-1} \tag{36}$$

and

$$\bar{\mu}_i(t) = \frac{\bar{C}_i\left(t\right)}{\Omega}. \tag{37}$$

Now, substituting $\bar{\mu}_i(t)$ into Eq. (9), we obtain a system of differential equations for each BS i, i. e.

$$\begin{aligned}\frac{\mathrm{d}\bar{\pi}_i(x_i,t)}{\mathrm{d}t} &= \lambda_i(t)\bar{\pi}_i(x_i-1,t)\\ &\quad -\left(\lambda_i(t)+\bar{\mu}_i(t)\right)\bar{\pi}_i(x_i,t)\\ &\quad +\bar{\mu}_i(t)\bar{\pi}_i(x_i+1,t).\end{aligned} \tag{38}$$

Since we have $\bar{\eta}_i(t) = 1-\bar{\pi}_i(0,t)$ and $\bar{\mu}_i$ is a function of $\bar{\eta}$ through Eqs. (34)-(37), the systems (38) are non-linearly coupled. However, we can overcome the problem of the coupling easily by solving Eq. (38) iteratively for each BS i and by substituting $\bar{\eta}_i(t) = 1-\bar{\pi}_i(0,t)$ in each iteration time step, as proposed in [16]. Again, as a last step, we compute the approximate probabilities of an active flow observing interference scenario y by

$$\tilde{\zeta}_i(y,t) := \sum_{\substack{x\in\mathcal{X}(y)\\ x_i\geqslant 1}}\prod_{i\in\mathcal{N}}\bar{\pi}_i(x_i,t)\Big/\left(1-\bar{\pi}_i(0,t)\right). \tag{39}$$

Note that, compared to the proposed average interference solutions in [15] and [16], the approach presented here in combination with the second step described in Section IV-A is a novel. As will be shown in the numerical results, this enables a much more accurate performance evaluation.

D. Feasibility and Steady-State Simplifications

Depending on the desired results and the system considered, certain methods to compute the state probabilities cannot be applied or, in contrast, reduce to rather elegant solutions. With respect to the desired results, we distinguish between transient and steady-state analyses.

1) Transient Analyses: Transient analyses require solving multiple, systems (aggregation and average interference approximations) of differential equations. However, if there exists at least one BS i that does not employ admission control, i. e. $K_i\to\infty$ obtaining the transients becomes infeasible. One exception is the analysis of the transient of unbounded systems provided that the queues are decoupled (which is the case for the aggregation and average interference model), converge to a steady state, and provided that the starting state is known [18].

2) Steady-State Analysis with State Aggregation: Assume the system dynamics reach a steady-state for stationary traffic demand, meaning for $\lambda_i(t)=\lambda_i$ and $\delta_i(u,t)=\delta_i(u)$ for all t. The steady-state state probabilities can be computed by the state-aggregation method with low complexity. In a first step, we substitute the time-dependent state probabilities $\pi_i(x_i,y,t)$ in Eq. (31) by their steady-state counterparts $\lim_{t\to\infty}\pi_i(x_i,y,t)$ from Eq. (11) or Eq. (12). In a second step, we solve the resulting system of equations (32) by letting $\frac{\mathrm{d}\sigma(y,t)}{\mathrm{d}t} = 0$. Since this approach contains the computation of steady-state probabilities for all $y\in\mathcal{Y}$ including the case where $y=\mathbf{1}_N$, it is only applicable to strictly stable systems according to Definition 2.

3) Steady-State Analysis with Time-Averaged Interference: In order to carry out steady-state analyses with the time-averaged interference assumption, we let $\frac{\mathrm{d}\bar{\pi}_i(x_i,t)}{\mathrm{d}t} = 0$ in Eq. (38). Recall that $\bar{\mu}_i$ is a function of $\bar{\eta}$, such that we can define the load of a BS by

$$\bar{\rho}_i(\bar{\eta}) := \frac{\lambda_i}{\bar{\mu}_i(\bar{\eta})}. \tag{40}$$

It is a well-known result that Eq. (38) reduces to the product form for the state probabilities, similar to Eqs. (11) and (12). This fact along with Eq. (40) and considering that $\bar{\eta}_i(t) = 1-\bar{\pi}_i(0,t)$, we obtain the BS utilization for BSs with admission control

$$\bar{\eta}_i = f_i(\bar{\eta}) := \bar{\rho}_i(\bar{\eta})\frac{1-\bar{\rho}_i(\bar{\eta})^{K_i}}{1-\bar{\rho}_i(\bar{\eta})^{K_i+1}}, \tag{41}$$

and for BSs without admission control

$$\bar{\eta}_i = f_i(\bar{\eta}) := \min\left\{\bar{\rho}_i(\bar{\eta}),1\right\}, \tag{42}$$

respectively. We observe that $\bar{\eta}_i$ is given in an implicit form; however, from the fact that, for all $i\in\mathcal{N}$, the mapping $f_i:\mathbb{R}_+^N\to\mathbb{R}_{>0}$ is a standard interference function, see [14], [15], [17], we have the following proposition:

Proposition 1. *Let $f(\cdot) := \left(f_1(\cdot),\dots,f_N(\cdot)\right)^{\mathrm{T}}$. For any initial utilization vector $\bar{\eta}^0\in\mathbb{R}_+^N$ with $\bar{\eta}^0 < \mathbf{1}_N$, the sequence $\bar{\eta}^{k+1} := f\left(\bar{\eta}^k\right)$ for $k=0,1,2,\dots$ converges to a unique fixed point $\breve{\eta}$.*

Though, we are able to compute the resource utilizations, or equivalently the probabilities of BS activity $\breve{\eta}$, the computation of state probabilities may fail for instable networks. In this regard, we make the following definition of stability under time-averaged interference:

Definition 4 (Stability of a Network with Time-Averaged Interference)**.** *A network is said to be* stable under time-averaged interference *if, for each BS $i\in\mathcal{N}$, one of the following two conditions hold, either*

1) K_i *is finite, or*
2) *for* $K_i \to \infty$, $\bar{\rho}_i(\breve{\eta}) < 1$.

Otherwise, the system is said to be unstable under time-averaged interference.

Now, for every stable network according to Definition 4, the stationary state probabilities of BS i are given by

$$\lim_{t\to\infty} \bar{\pi}_i(x_i, t) = \frac{\left(1 - \bar{\rho}_i(\breve{\eta})\right) \bar{\rho}_i(\breve{\eta})^{x_i}}{1 - \bar{\rho}_i(\breve{\eta})^{K_i+1}}, \tag{43}$$

or, if we have $K_i \to \infty$, by

$$\lim_{t\to\infty} \bar{\pi}_i(x_i, t) = \left(1 - \bar{\rho}_i(\breve{\eta})\right) \bar{\rho}_i(\breve{\eta})^{x_i}. \tag{44}$$

Due to the strictly monotonic function $\bar{\rho}_i(\breve{\eta})$, every strictly stable system is also stable under time-averaged interference. This also implies that, if we are able to obtain steady-state probabilities through the state-aggregation method, we can use the less complicated time-average approximation. It is important to note that the computation of the resource utilizations $\breve{\eta}$ using so-called interference function calculus, which is adopted here as well, has been studied for systems without admission control already in [11]–[13]. The extension to networks with admission control in [15] and in this article, constitutes a novel aspect of studying the performance of cellular networks using interference function calculus.

V. Numerical Validation

In this section, we present simulation results obtained from discrete event simulations and compare them with the results obtained with the framework proposed.

A. Simulation Scenario and Benchmark

We use the model of a cellular network consisting of seven sectors in a hexagonal layout as depicted in Fig. 5. The inter-site distance equals 500 m. We restrict the total number of BSs to seven and the maximum number of concurrently active flows in each cell to two, respectively. Traffic is generated according to a Poisson process and is uniformly distributed across each cell. If not otherwise stated, flow sizes are exponentially distributed with mean 16 Mbit. W. l. o. g., the mean traffic in the central and neighboring cells is set to 11 Mbps and 16 Mbps, respectively. Antenna patterns and pathloss calculation are compliant with standard 3GPP models [41]; further parameters are listed in Table I.

As a benchmark, we use two types of bounds with respect to the interference coupling: The first type considers either full or no interference. The second type are so-called second-degree bounds from [10], which have been adapted here to systems with admission control. Using the second-degree bounds, it is assumed that the cell under consideration experiences infinitely fast (resp. slow) interference dynamics from neighboring BSs whose rates are calculated assuming no (resp. full) interference. We will restrict the computations and simulations to stationary traffic arrival processes only in order to highlight the essential behavior of the performance metrics and the accuracy of the framework.

TABLE I
Scenario Configuration

Maximum transmit power	49 dBm
Path loss model [41]	$128.1\,\text{dB} + 37.6 \log_{10}(d/\text{km})\,\text{dB}$
Carrier frequency [41]	2.0 GHz
Bandwidth B	20 MHz
Bandwidth efficiency a	0.63
SINR efficiency b	0.4
Fast fading margin [42]	-2 dB
Antenna diversity gain [42]	3 dB
BS noise figure [43]	-5 dB
BS mounted on height [41]	32 m
UE noise figure [43]	-9 dB
UE height [41]	1.5 m
Thermal noise	-174 dBm/Hz
Scheduler	Round Robin or throughput-fair
Average flow size	16 Mbit

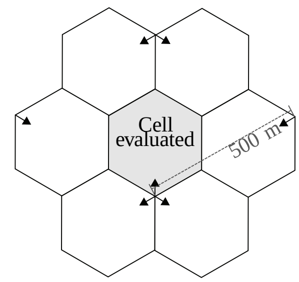


Fig. 5. Simulation scenario with one central cell and six interfering neighbors, and inter-site distance of 500 m.

B. Results

In the following, we analyze the impact of internal and external flow (resp. interference) dynamics on network-specific performance metrics, user-specific performance, as well as, the impact of the speed of the interference process on both, network- and user-specific KPIs.

1) Impact of Internal and External Data Flow Dynamics: Figs. 6 and 7 depict network-specific KPIs introduced in Section III-B for increasing traffic in neighboring cells and in the cell under consideration, respectively. An increased interference level (Fig. 6) drastically decreases the performance. For instance, a rise in traffic demand from 5 Mbps to 15 Mbps in neighboring cells basically halves the average flow throughput in the affected cell by reducing the transmission efficiency by about 35 % (represented by the drop of the cell capacity and the increased BS utilization). In fact, the external interference processes partly have a much higher impact on the cell-related performance than the actual flow dynamics. The model proposed (red curves) exhibits a remarkable accuracy in all cases shown here. Additionally, assuming average interference is a slightly optimistic approximation for lower cell utilizations and less intensive inter-cell interference.

2) User-Specific Performance: Figs. 8 and 9 show the cumulative distribution functions of the flow throughput and the sojourn time, respectively. As can be seen, the user-specific KPIs in the quasi-stationary (qs) regime (solid red curve) yield a good lower performance bound for lower throughputs and a very tight bound for low sojourn times. Since data flows at the cell edge exhibit longer sojourn times due to lower rates, they are likely to observe many changes of the interference scenario. As a result, the accuracy of the model in the qs regime is lower for such data flows. Note that the gap between

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2015.2480760, IEEE Transactions on Wireless Communications

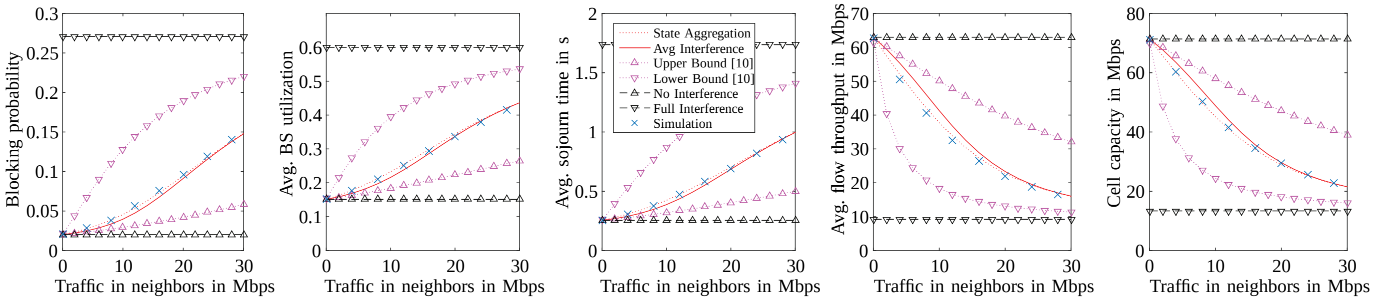


Fig. 6. Cell-specific KPIs for increasing traffic and interference level in neighboring cells. The traffic demand in the cell evaluated is fixed to 11 Mbps.

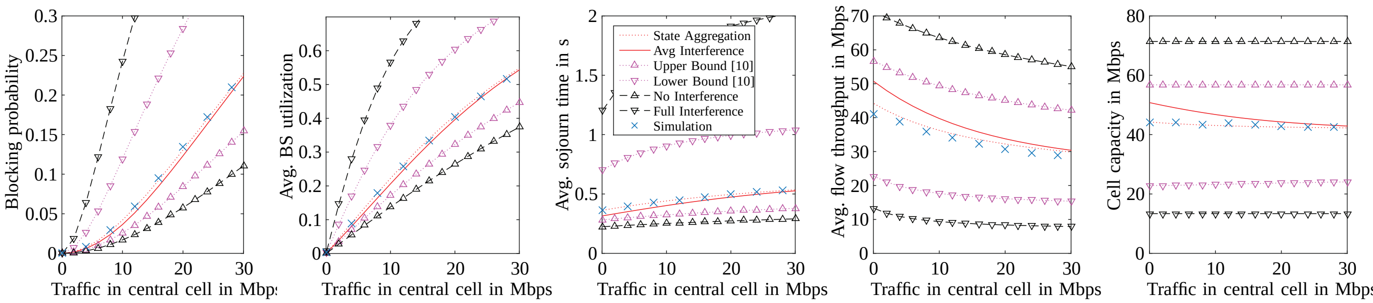


Fig. 7. Cell-specific KPIs for increasing traffic level in central cell. The traffic demand in the neighboring cells is fixed to 11 Mbps.

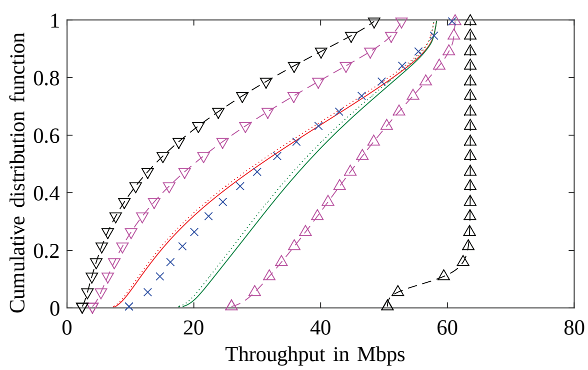


Fig. 8. Throughput statistics within the cell under consideration.

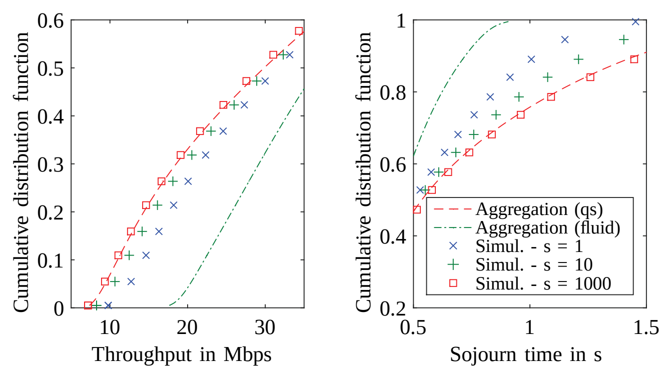


Fig. 10. Impact of the speed of the external interference process on the throughput and sojourn time distributions.

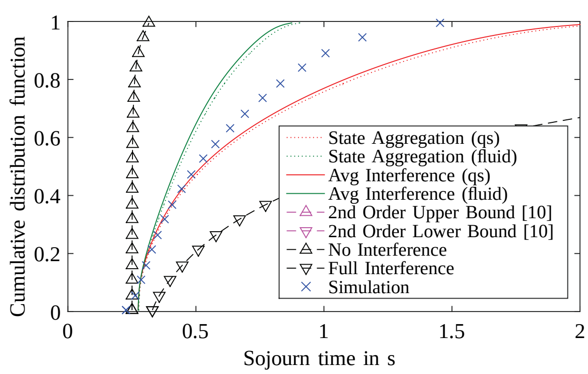


Fig. 9. Flow sojourn time statistics within the cell under consideration.

the qs and fluid regimes for low throughputs and high sojourn times is a corroboration of the fact that interference dynamics have the greatest impact at the cell edges.

3) Impact of the Speed of Variation of Interference: We now configure our simulation such that the speed of the external flow processes is decreased by a factor s, which means that the flow arrival rates (resp. flow sizes) in neighboring cells are divided (resp. multiplied) by s. The total traffic demand remains the same. Fig. 10 shows the distributions of the flow sojourn times and the user throughputs for different factors $s \in \{1, 10, 1000\}$. The information that can be gleaned from the figures is that the simulation results converge to the quasi-stationary model results for decreasing speed of the external processes. In fact, each flow in the cell under consideration sees the external interference dynamics in a qs regime, meaning that it is unlikely that the interference scenario changes during the transmission of a flow.

4) Impact of Admission Control and the Scheduler: The average BS utilization and the average sojourn time are depicted in Fig. 11 for increasing interference level, two different schedulers, and various configurations of admission control. A higher admission control parameter leads to higher BS utilizations and lower user performance. Independent of the admission control setting, the model provides sufficiently accurate results. It can also be observed that a throughput-fair scheduler, which allocates resources to concurrently active users in a way, such that they have the same data throughput, is slightly more resource-efficient. However, this comes at the cost of a higher mean flow sojourn time. Interestingly, the scheduler has a minor impact on the KPIs compared to interference dynamics that are captured well by the model.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2015.2480760, IEEE Transactions on Wireless Communications

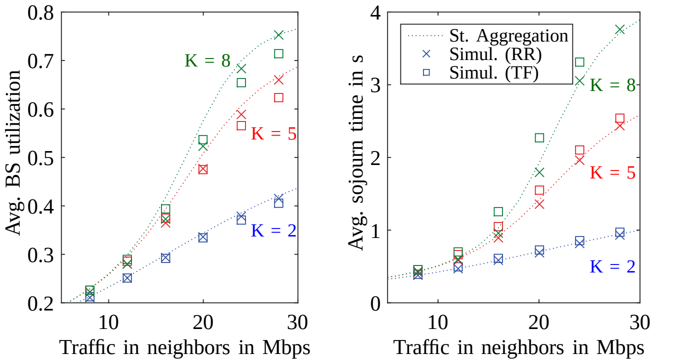


Fig. 11. Impact of the admission control parameter K_i and the scheduler on the BS utilization and average sojourn time.

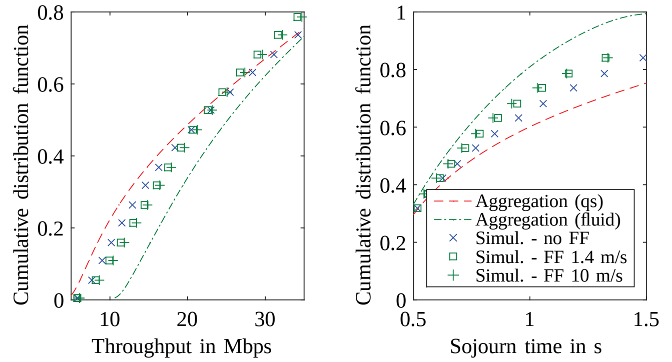


Fig. 13. Impact of fast fading on the sojourn time and throughput statistics. The admission control parameter is set to $K_i = 5$.

5) Impact of the File Size Distribution: In order to prove the broad applicability of the model, we consider deterministic (fixed) file sizes, and Weibull distributed file sizes in addition to the exponential distribution used before. The configurations "Weibull A" and "Weibull B" have shape parameters $k = 2$ and $k = 0.5$, respectively. The mean file size is 16 Mbit. In Fig. 12, we observe that the user throughputs and sojourn times approach the quasi-stationary model results, when the skewness of the distribution becomes larger. This effect can be explained by the fact that a higher skewness of the distribution generates more data flows with smaller sizes, which ultimately observe the interference process in a quasi-stationary regime rather than a fluid regime. Additionally, we see that the file size distribution has again a minor impact on the performance.

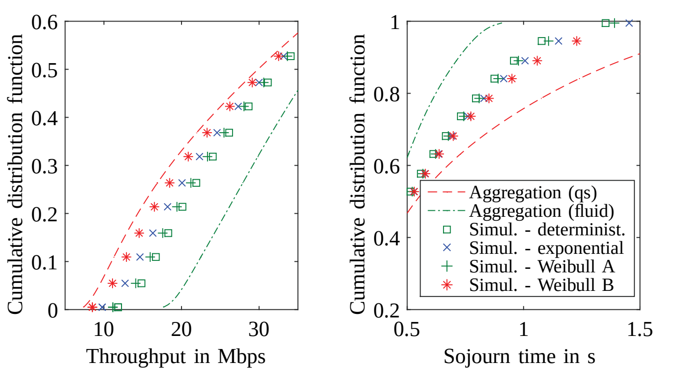


Fig. 12. Impact of the file size distribution on the sojourn time and throughput statistics.

6) Impact of Fast Fading: The framework proposed makes the assumption of considering the fast fading process as a time-average impact on the receive powers. Fig. 13 presents simulation results for different fast fading profiles given by the user speed (0 m/s, 1.4 m/s, 10 m/s) and compares them to the model results. Adding fast fading seems to improve the user performance slightly. However, there is almost no difference between the performance of users with speeds of 1.4 m/s and 10 m/s. Only a very minor impact on cell-specific KPIs have been observed (not shown here). From the simulations, we can conclude that the assumption is valid when flow-level dynamics and inter-cell interference dynamics are in the focus.

VI. Conclusions and Future Work

In this article, we proposed an evaluation framework for cellular networks, which can be applied to a variety of radio access technologies, deployment strategies, and network architectures. With the help of this framework, it is possible to estimate a comprehensive set of important network-specific and user-specific key performance indicators as a function of the spatio-temporal dynamics of the data traffic demand. This set includes base station resource utilizations, average cell throughputs, and cell capacities, as well as, user throughput statistics. In particular, we have shown that dynamic inter-cell interference induced by neighboring base stations can have a much higher impact on these metrics than the actual data flow dynamics in the cell under consideration. The framework proposed is able to estimate these effects with a higher precision compared with existing works. We also have shown that the impact of flow and inter-cell interference dynamics predominate scheduler settings, the flow size distribution, as well as, fast fading effects. Studying the interference coupling behavior of cellular systems, which (partly) employ admission control, constitutes a novel aspect of wireless network performance evaluation using interference function calculus.

Future work may comprise different adaptations of the model proposed to more specific (future) network architectures including distributed antenna systems, the phantom cell concept with a separation of the control plane and the user plane, or multi-RAT (radio access technology) networks. This would enable a more effective management, performance evaluation, and performance estimation of such upcoming cellular network approaches. Additionally, three rather theoretical aspects could be the following. First, modeling the impact of inter-cell interference dynamics on buffered or live streaming services would extend the framework by adding video or voice traffic-related key performance indicators. Second, incorporating backhaul links and backhaul nodes as potential bottlenecks would lead to a more holistic view on the entire network performance. In this regard, the model could be extended using theory of queuing networks (instead of queuing systems only). Finally, since 5G networks are expected to facilitate ultra-low latency applications, potential latency bottlenecks from a network point of view could be identified by a more elaborate modeling of flow sojourn times.

ACKNOWLEDGMENT

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

REFERENCES

[1] D. Rose, J. Baumgarten, and T. Kürner, "Spatial Traffic Distributions for Cellular Networks with Time Varying Usage Intensities per Land-Use Class," in *2014 IEEE 80st Vehicular Technology Conference*, Vancouver, Canada, 2014, pp. 1–5.

[2] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What Will 5G Be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[3] N. Bhushan, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. Sukhavasi, C. Patel, and S. Geirhofer, "Network densification: the dominant theme for wireless evolution into 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 82–89, Feb. 2014.

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

[5] S. B. Fred, T. Bonald, A. Proutiere, G. Régnié, and J. W. Roberts, "Statistical bandwidth sharing," 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.

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

[7] T. Bonald, S. Borst, and A. Proutiere, "How mobility impacts the flow-level performance of wireless data systems," in *IEEE INFOCOM 2004*, vol. 3, 2004, pp. 1872–1881.

[8] T. Bonald and N. Hegde, "Capacity Gains of Some Frequency Reuse Schemes in OFDMA Networks," in *GLOBECOM 2009 - 2009 IEEE Global Telecommunications Conference*, Nov. 2009, pp. 1–6.

[9] Y. Xu, S. E. Elayoubi, E. Altman, and R. El-Azouzi, "Impact of flow-level dynamics on QoE of video streaming in wireless networks," in *2013 Proceedings IEEE INFOCOM*, Apr. 2013, pp. 2715–2723.

[10] T. Bonald, S. Borst, N. Hegde, and A. Proutiére, "Wireless data performance in multi-cell scenarios," *ACM SIGMETRICS Performance Evaluation Review*, vol. 32, no. 1, p. 378, Jun. 2004.

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

[12] I. Siomina, "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.

[13] R. L. G. Cavalcante, S. Stanczak, M. Schubert, A. Eisenblaetter, and U. Tuerke, "Toward Energy-Efficient 5G Wireless Communications Technologies: Tools for decoupling the scaling of networks from the growth of operating power," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 24–34, Nov. 2014.

[14] 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.

[15] H. Klessig, A. Fehske, and G. Fettweis, "Admission control in interference-coupled wireless data networks: A queuing theory-based network model," in *2014 12th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, May 2014, pp. 151–158.

[16] D. Öhmann, A. J. Fehske, and G. P. Fettweis, "Non-Stationary Traffic Conditions in Transient Flow Level Models for Cellular Networks," in *Proceedings of the 20th European Wireless Conference*, Barcelona, Spain, 2014, pp. 1–6.

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

[18] D. Öhmann, A. Fehske, and G. Fettweis, "Transient flow level models for interference-coupled cellular networks," in *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Oct. 2013, pp. 723–730.

[19] T. Bonald and A. Proutière, "Insensitivity in processor-sharing networks," *Performance Evaluation*, vol. 49, no. 1-4, pp. 193–209, Sep. 2002.

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

[21] 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.

[22] J. Weitzen and T. Lowe, "Measurement of angular and distance correlation properties of log-normal shadowing at 1900 MHz and its application to design of PCS systems," *IEEE Transactions on Vehicular Technology*, vol. 51, no. 2, pp. 265–273, Mar. 2002.

[23] 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, Apr. 2007, pp. 1234–1238.

[24] J. Baumgarten and T. Kuerner, "Lte downlink link-level abstraction for system-level simulations," in *European Wireless 2014; 20th European Wireless Conference; Proceedings of*, May 2014, pp. 1–5.

[25] M. Ghaderi and R. Boutaba, "Call admission control in mobile cellular networks: a comprehensive survey," *Wireless Communications and Mobile Computing*, vol. 6, no. 1, pp. 69–93, Feb. 2006.

[26] J. Bartelt, A. Fehske, H. Klessig, G. Fettweis, and J. Voigt, "Joint Bandwidth Allocation and Small Cell Switching in Heterogeneous Networks," in *2013 IEEE 78th Vehicular Technology Conference (VTC Fall)*. IEEE, Sep. 2013, pp. 1–5.

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

[28] A. J. Fehske, H. Klessig, J. Voigt, and G. P. Fettweis, "Concurrent Load-Aware Adjustment of User Association and Antenna Tilts in Self-Organizing Radio Networks," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 5, pp. 1974–1988, Jun. 2013.

[29] 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.

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

[31] L. You, L. Lei, and D. Yuan, "Range assignment for power optimization in load-coupled heterogeneous networks," in *Communication Systems (ICCS), 2014 IEEE International Conference on*, Nov 2014, pp. 132–136.

[32] R. Combes, Z. Altman, and E. Altman, "Interference coordination in wireless networks: A flow-level perspective," in *2013 Proceedings IEEE INFOCOM*, Apr. 2013, pp. 2841–2849.

[33] A. Khlass, T. Bonald, and S. E. Ayoubi, "Flow-Level Performance of Intra-site Coordination in Cellular Networks," in *Modeling & Optimization in Mobile, Ad Hoc & Wireless Networks (WiOpt), 2013 11th International Symposium on*, Tsukuba Science City, 2013, pp. 216–223.

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

[35] R. Combes and E. Altman, "Flow-level performance of random wireless networks," *arXiv preprint arXiv:1307.7375*, Jul. 2013.

[36] Y. Zhou and N. Gans, "A single-server queue with Markov modulated service times," 1999.

[37] S. Mahabhashyam and N. Gautam, "On queues with Markov modulated service rates," *Queueing Systems*, no. 814, pp. 1–24, 2005.

[38] R. W. Wolff, "Poisson Arrivals See Time Averages," *Operations Research*, vol. 30, no. 2, pp. 223–231, Apr. 1982.

[39] J. D. C. Little, "A Proof for the Queuing Formula: L = λ W," *Operations Research*, vol. 9, no. 3, pp. 383–387, Jun. 1961.

[40] A. A. Kherani and A. Kumar, "Performance Analysis of TCP with Nonpersistent Sessions," in *Workshop on the Modeling of Flow and Congestion Control Mechanisms*, Paris, France, 2000, pp. 1–14.

[41] Technical Specification Group Radio Access Network, "TR 36.814 v9.0.0 - Evolved Universal Terrestrial Radio Access (E-UTRA) - Further advancements for E-UTRA: Physical layer aspects," 3rd Generation Partnership Project, Tech. Rep., 2010.

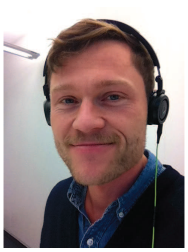
[42] A. J. Fehske, F. Richter, and G. P. Fettweis, "Energy efficiency improvements through micro sites in cellular mobile radio networks," in *Proc. IEEE GLOBECOM Workshops*, 2009, pp. 1–5.

[43] Technical Specification Group Radio Access Network, "TR 36.942 v9.0.1 - Evolved Universal Terrestrial Radio Access (E-UTRA) - Radio frequency (RF) system scenarios," 3rd Generation Partnership Project, Tech. Rep., 2010.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2015.2480760, IEEE Transactions on Wireless Communications

Henrik Klessig Henrik Klessig received his diploma from the Department of Electrical and Computer Engineering of Dresden University of Technology, Germany, in 2012. Within his studies he visited Alcatel-Lucent, Bell Labs, Germany, for a six-month internship, where he was engaged in LTE base station power modeling and research on integrated energy-efficient hardware and resource management solutions for wireless base stations. His research interests are cellular heterogeneous networks, self-organizing networks, and energy efficiency optimization. Since May 2012 he is with the Vodafone Chair Mobile Communications Systems, University of Dresden, Germany.

David Öhmann David Öhmann received the B.Eng. degree in Information Technology from the DHBW Karlsruhe, Germany, in 2010 and the M.Sc. degree in Electrical Engineering and Information Technology from the Technical University of Dortmund, Germany, in 2012. In his master thesis, he investigated the potentials of cooperative information for vertical handover decision algorithms in cooperation with BMW Research & Technology, Munich, Germany. Since January 2013, he is with the Vodafone Chair Mobile Communications Systems at TU Dresden, Germany, and is currently working on system level models and high resilience in 5G networks.



Albrecht J. Fehske Dr. Albrecht Fehske received his diploma as well as PhD in electrical engineering with highest honors from Technical University of Dresden. From 2010 to 2014 Albrecht was leading a team at TU-Dresden and worked on variety of research projects with tier one vendors and operators. He co-authored more than 40 papers in the areas of mobile network energy efficiency, self-organizing networks, queuing theory, and mathematical optimization. Among others, he was awarded the 2014 Fred W. Ellersick prize of the IEEE Communications Society. In 2014 Albrecht co-founded Airrays, a startup company working on large scale antenna arrays for 5G mobile communications, where he is currently active.

Gerhard P. Fettweis Prof. Dr.-Ing. Gerhard Fettweis received his Dipl.-Ing. and doctoral degrees under the guidance of Prof. H. Meyr from Aachen University of Technology (RWTH) in Aachen, Germany, in 1986 and 1990 resp. From 1990 to 1991 he was a Visiting Scientist at the IBM Almaden Research Center in San Jose, CA. From 1991 to 1994 he was responsible for signal processor development at TCSI Inc., Berkeley, CA. Since 09/1994, he holds the Vodafone Chair at TUD. Prof. Fettweis has been an elected member of the IEEE Solid State Circuits Society Board (Administrative Committee) since 1999 and an IEEE Fellow since 2009. He was elected as IEEE Distinguished Lecturer for Solid State Circuit Society (2009 to 2010) and for the Vehicular Technology Society (2011 to 2013). He also serves on several supervisory boards, and on advisory committees of companies and research institutes. In 2009 he initiated the Leading Edge Cluster *Cool Silicon* and led it until 2010. Since 2011 he leads the Collaborative Research Center (CRC) *Highly Adaptive Energyefficient Computing (HAEC)* at TUD as its speaker. As chairman of TUDs focus area Information Technology and Microelectronics, he coordinates the proposal for a Cluster of Excellence within the German Excellence Initiative *Center for Advancing Electronics Dresden (cfAED)* addressing different paths to advance electronics beyond current/upcoming limitations of CMOS technology.