Beiträge aus der Informationstechnik

Mobile Nachrichtenübertragung
Nr. 82

Henrik Klessig

Advanced Network and Mobile Data Traffic Models and their Application to Cellular Network Optimization

VOGT

Dresden 2016

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

Bibliographic Information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Zugl.: Dresden, Techn. Univ., Diss., 2016

Die vorliegende Arbeit stimmt mit dem Original der Dissertation „Advanced Network and Mobile Data Traffic Models and their Application to Cellular Network Optimization“ von Henrik Klessig überein.

Gesetzt vom Autor

ISBN 978-3-95947-006-3

Jörg Vogt Verlag
Niederwaldstr. 36
01277 Dresden
Germany

Phone: +49-(0)351-31403921
Telefax: +49-(0)351-31403918
e-mail: info@vogtverlag.de
Internet : www.vogtverlag.de

Technische Universität Dresden

Advanced Network and Mobile Data Traffic Models and their Application to Cellular Network Optimization

Henrik Klessig

von der Fakultät Elektrotechnik und Informationstechnik der Technischen Universität Dresden

zur Erlangung des akademischen Grades

Doktoringenieur

(Dr.-Ing.)

genehmigte Dissertation

Vorsitzender:	Prof. Dr.-Ing. habil. Gerald Gerlach
Gutachter:	Prof. Dr.-Ing. Dr. h.c. Gerhard Fettweis
	Prof. Dr. Di Yuan
Tag der Einreichung:	21. Juni 2016
Tag der Verteidigung:	23. September 2016

To the memory of my father.

To Amilia.

Abstract

Without doubt, network operators continuously face an inevitable increase in mobile data traffic demand, the largest share of which will be video streaming. Additionally, spatio-temporal traffic fluctuations cause local capacity bottlenecks in cellular networks, which are strongly connected with rising interference levels and worsening quality of experience. Network densification and self-organization (SON) capabilities are considered to counteract such problems in upcoming 5G networks. However, a large number of wireless access nodes and future smart IoT and MTC devices render network management increasingly complex. Flow-level modeling techniques help avoid enormous simulation effort and enhance SON solutions.

Flow-level models for interference-coupled cellular data networks are generalized by the aspect of admission control in this thesis. A comprehensive performance evaluation framework with respect to various key performance indicators is thereby established. The metrics presented include the distribution of video startup delays and the probability of video buffer starvations to address the increasing popularity of mobile video services. The traffic demand distribution across the network is identified as the main extrinsic factor that affects the quality of experience perceived by single mobile users. In particular, the amount of traffic served by neighboring base stations can have a similarly strong impact as increased traffic load in a cell.

Motivated by this, a detailed analysis of the spatial traffic distribution is carried out based on call traces measured across a 3G network. It becomes apparent that the traffic density can be modeled according to a log-normal distribution. An efficient method to generate large sets of log-normally distributed and spatially correlated traffic maps is introduced, which facilitates an effective design and a reliable evaluation of self-organizing network algorithms if it is applied along with the extended flow-level model. The application is illustrated by means of two specific examples. In one example, advanced models for phased-array antennas are combined with the flow-level model and an algorithm is proposed, which jointly adjusts the spherical directions of multiple sub-beams per antenna for a set of base stations in the network. It is shown that the beamforming algorithm outperforms state-of-the-art antenna tilt optimizers in terms of cell edge throughput by a factor of three on average. However, beamforming performance is sensitive to the traffic hot spot size and their traffic demand intensity. In particular, the algorithm robustness decreases as the traffic standard deviation and the correlation distance increase.

Kurzfassung

Netzbetreiber sehen sich zunehmend mit der Erhöhung des mobilen Datenverkehrs, wovon Video-Streaming den größten Anteil haben wird, konfrontiert. Zudem verursachen räumlich-zeitliche Verkehrsschwankungen in Mobilfunknetzen lokale Kapazitätsengpässe, die mit steigender Interzell-Interferenz und schlechter Nutzerqualität einhergehen. Eine weitere Verdichtung der Netze und Fähigkeiten zu ihrer Selbstorganisation werden als wichtige Aspekte der zukünftigen fünften Mobilfunkgeneration angesehen, um solchen Problemen zu entgegnen. Allerdings werden Netzwerkmanagementlösungen durch die steigende Anzahl von Basisstationen und mobilen Endgeräten, z. B. für IoT- und MTC-Anwendungen, immer komplexer. Flow-Level-Modelle können helfen, großen Simulationsaufwand bei der Netzwerkplanung und -selbstoptimierung zu vermeiden und Lösungen dafür zu verbessern.

In dieser Arbeit wird ein Flow-Level-Modell für Interferenz-gekoppelte, zellulare Mobilfunknetze durch den Aspekt der Zugangskontrolle verallgemeinert, sodass sich ein umfangreiches Werkzeug zur Bewertung solcher Netze hinsichtlich vieler Performanzmetriken ergibt. Solche Metriken umfassen beispielsweise die Verzögerungen der Videowiedergabe und die Wahrscheinlichkeit, dass der Videopuffer am Endgerät leer läuft. Die Datenverkehrsnachfrage wird als größter extrinsischer Einflussfaktor hinsichtlich der durch den Nutzer empfundenen Netzqualität identifiziert.

Dies dient als Motivation, eine detaillierte Analyse der räumlichen Verkehrsverteilung in einem 3G-Netzwerk basierend auf gemessenen Call Traces durchzuführen. Es stellt sich heraus, dass, unter anderem, die Verkehrsdichte entsprechend einer Lognormal-Verteilung modelliert werden kann. Darauf aufbauend wird eine effiziente Methode entwickelt, um lognormal-verteilte und räumlich korrelierte Verkehrskarten zu generieren, welche, zusammen mit dem Flow-Level-Modell, ein effektives Design und eine zuverlässige Evaluierung von Netzoptimierungsalgorithmen ermöglicht. Dies wird anhand zweier Beispiele erläutert. In einem Beispiel wird ein Modell für Phased-Array-Antennen mit dem Flow-Level-Modell kombiniert und ein Algorithmus vorgeschlagen, welcher die Richtungen mehrerer Teilstrahlen je Antenne für eine Gruppe von Basisstationen optimiert. Es wird gezeigt, dass dieser Beamforming-Algorithmus herkömmlichen Antennenneigungswinkel-Optimierern hinsichtlich des Durchsatzes am Zellrand im Durchschnitt um den Faktor drei überlegen ist. Jedoch nimmt die Robustheit mit der Standardabweichung und dem räumlichen Korrelationsabstand des Verkehrs ab.

Acknowledgment

Art is I; science is we.

Claude Bernard

I believe that art and science are not necessarily mutually exclusive, as science demands innovative and creative thinking as well especially in engineering research. But as Claude Bernard correctly points out, scientific work is not necessarily produced by only one individual. Science rather produces results that demand collaboration and a consensus among many participating individuals. Consensus is mostly needed, when it comes to publishing scientific results, and it can often be achieved, only when each individual collaboratively makes his or her small contribution to the overall work – be it by joining discussions, by direct involvement in producing results, or by helping develop professional skills. Unsurprisingly, creating a Ph.D. thesis is not much different, so I owe a debt of gratitude to all who helped.

My deepest gratitude goes to my advisor, Gerhard Fettweis, for believing in my skills and giving me the chance to pursue this adventurous journey four years ago. I am very thankful for his support whenever it was needed and his guidance that helped me sharpen my professional skills. Also, I want to thank my second reviewer, Prof. Di Yuan, for his willingness to review the thesis, his interest in my work, and all the valuable comments. I owe special thanks to my colleagues and friends Albrecht, David, Philipp, Maciej, Meryem, Norman, Sascha, and Vinay, who were always willing to join discussion and to work collaboratively, sometimes during the night to meet seemingly unrealistic deadlines. They made the last years very enjoyable, both on a professional and a private level. My former students David, Felix, Hagen, Henning, Jens, Lucas, and Michael deserve many thanks for their collaborative work as well, which taught me the competences of target-oriented personal interaction and successful leadership.

I warmly thank my parents and my grandma for reminding me of all my accomplishments, which kept me motivated, my parents in law for being a wonderful part of my family, and Olla for her extraordinarily strong love and her endless patience during the last couple of years.

Summer 2016 - Henrik Klessig

Contents

1 Introduction

The subsequent paragraphs aim at highlighting the need for appropriate analytical network and traffic models in light of current developments of the fifth generation mobile technology (5G). In particular, we stress the importance of a mapping between technical metrics and quality of experience-related (QoE) ones in the context of the limited-capacity problem due to the tremendous increase of traffic demand. An outline of the thesis and literature recommendations follow thereafter.

1.1 Motivation

Without doubt, mobile data traffic demand is going to increase considerably during the next decade. In particular, according to [Cis] the traffic volume generated by mobile devices will increase by a factor of eight between 2015 and 2020. As of 2020 smart phones will account for four-fifths of the demand and 4G (Fourth Generation Mobile Communications Systems) traffic will hold a the largest share of the total mobile data traffic demand. In addition, mobile video traffic demand will increase by a factor of 11 between 2015 and 2020 ultimately accounting for about 75 % of the entire demand in 2020.

Cellular networks are and will remain capacity-limited as a result of the ever-increasing traffic demand. Limited capacity generally relates to highly loaded and, in the context of modern cellular frequency-reuse-one networks, to interference-limited systems rather than noise-limited systems. In addition, because mobile data traffic is likely to fluctuate over time and space [RBK14], limited capacity and inter-cell interference are anything but static phenomena, neither in space nor in time. As a consequence thereof and with the goal of improving the quality of the network experienced by the users, a joint treatment of dynamic inter-cell interference and spatio-temporal characteristics of the traffic demand is required to be able to allocate sufficient capacity to high traffic locations.

However, sole technical metrics, such as the signal-to-interference-and-noise ratio (SINR), are not meaningful to describe the mobile users' quality of *experience*. In fact, the relation between technical and QoE-related key performance indicators (KPIs) is rather complex. For example, limited capacity and an increased interference level decreases the user throughput. For mobile video users, this translates to worse experience of buffered video streaming services (such as *YouTube* [Had+11]), which is, among others, characterized through a rapid increase of the so-called video startup delay (the time between initiating the transfer of the video data and the actual playback). It has been found in [KS12] that video users start abandoning the video service already if the video startup delay exceeds two seconds with an

additional increase of 5.8 % of the rate of abandoning for each additional second of the delay. Nevertheless, it is not straightforward to establish a direct relation between interference levels and video startup delays or the rate of abandoning mobile video services.

5G cellular networks are expected to counteract the limited-capacity problem in the future mainly through two paradigms: (1) densification in space, for example, by ultra-dense small cell network deployments and/or massive MIMO (multiple-input multiple output) techniques, and (2) densification in frequency, for instance, with novel waveform designs and/or millimeter wave (mmWave) communications, [And+14; Bhu+14]. Spatio-temporal traffic dynamics additionally call for flexible and adaptive network optimization algorithms that comprise self-planning, self-optimization, and self-healing capabilities. Such algorithms are commonly referred to as self-organizing network (SON) algorithms. For instance, a SON use case related to the limited-capacity problem is the capacity and coverage optimization use case, see [Scu+08; Sch+08; 3GP14] among others.

Theoretical network modeling is of great benefit for an accurate evaluation and prediction of relevant QoE-related KPIs especially in the context of 5G developments, such as the massive deployment of wireless access points. Based on analytical models, network design and optimization are more effective, more efficient, and less complex. Furthermore, the use of realistic spatial data traffic distributions ensures a reliable evaluation of KPI statistics and an effective design of optimization algorithms with realistic input.

The requirements of such network and traffic models are ultimately characterized by the following four main features:

1. Scalability to extremely large and complex wireless networks that consist of hundreds of nodes to address the spatial base station (BS) densification in 5G networks,
2. An accurate evaluation *and* a reliable prediction of KPI statistics with low computational effort for the integration with SON algorithms,
3. A paradigm shift from rather technical metrics to user-specific QoE metrics, such as video streaming-related KPIs, and
4. The ability to capture the effects of spatio-temporally fluctuating data traffic demand and hence dynamic inter-cell interference.

The thesis at hand provides a holistic, flexible, yet accurate performance evaluation framework, which addresses all of the aforementioned aspects. The framework is based on flow-level models and the notion of so-called *elastic data flows*. This work also illustrates how this framework shall be used when designing (SON-)algorithms by means of ascertained examples and in combination with realistic spatial data traffic distributions.

1.2 Overview of the Thesis

This thesis is mainly divided into four parts (Chapters 2 to 5).

Chapter 2 reviews relevant prior art with respect to flow-level modeling in wireless networks, spatial traffic modeling and model-based network optimization, and highlights shortcomings thereof. This chapter also details the extensions and improvements compared to prior art made in the Chapters 3 to 5.

The extension of flow-level models for interference-coupled wireless data networks by admission control is presented in **Chapter 3**. This extension allows the low-complexity computation of a large set of QoE-related KPIs, which include the video startup-delay distribution and the video buffer starvation probability. The resulting performance evaluation framework is generic in the sense that it can be applied to any cellular deployment (base station positions, base station type and mix, transmit power, antenna type, etc.) and to any arbitrary data traffic demand distribution. The accuracy of the model is illustrated by the comparison with discrete event simulations in a small hexagonal setup. The focus is on the impact of increased traffic load in neighboring cells on network performance.

Because the spatial traffic distribution is the key input for the framework described above, measured spatial traffic data is analyzed in **Chapter 4**. Based on the statistical results obtained through the analysis of the traffic data, a low-complexity method is proposed to efficiently generate large sets of random spatial data traffic maps with the same statistical characteristics. This method is explicitly helpful in terms of comprehensive performance evaluation of SON algorithms.

The application of both, the flow-level model from Chapter 3 and the spatial traffic model from Chapter 4, is illustrated in **Chapter 5** by means of two example SON algorithms. They are *data offloading* using cell range expansion (CRE) and further enhanced inter-cell interference coordination (FeICIC), and *traffic-adaptive beamforming* using phased-array antennas. In addition to the accuracy of the flow-level model shown in Chapter 3, we further highlight its usefulness by showing that it can be the basis for (multi-) objective optimization. More specifically, the idea of data offloading in this thesis is to shift as much traffic as possible to more power-efficient small cells to enhance the total network energy efficiency, a metric that can be derived easily from flow-level KPIs presented in Chapter 3. The goal of the coordinated beamforming algorithm is to reveal the optimization potential of performing cell load balancing and interference reduction with adjusting the directions of multiple sub-beams per base station. The optimization potential and the algorithm robustness are quantified using various traffic distributions, which are provided by the spatial traffic model introduced in Chapter 4.

Conclusions and recommendations for future work are provided in **Chapter 6**.

1.3 Literature Recommendations

The following chapters cover a number of diverse topics from different fields related to wireless communications. For that reason, a list of basic literature and further reading is provided below.

For the understanding of **Chapters 2 and 3**, the reader is expected to be familiar with the basics of probability theory, stochastic processes, and queuing theory, which can be studied using the books by Robert B. Ash [Ash70] and R. G. Gallager [Gal14]. We also recommend the books by L. Kleinrock for further reading about queuing theory [Kle75; Kle76]. The problem of performance evaluation of interference-coupled cellular networks is well introduced in the research paper by T. Bonald *et al.* [Bon+04] and in the article by I. Siomina and D. Yuan [SY12a]. The book by G. S. Fishman [LJ02] gives information about simulation techniques, especially discrete event simulation, which is used for the validation of our analytical results.

Chapter 4 requires some knowledge about surface interpolation techniques and random field theory, which can be gained from the books by G. Farin [Far02] and E. Vanmarcke [Van10].

Network optimization techniques and SON are treated in **Chapter 5**. An overview about SON can be found in the book by S. Hämäläinen *et al.* [HSS12]. Some details about the particular optimization problems using data offloading in heterogeneous networks and three-dimensional beamforming can be found in the articles by A. Aijaz *et al.* [AAA13] and by H. Halbauer *et al.* [Hal+13], respectively.

2 An Overview of Flow-Level & Spatial Traffic Models

Before we review state-of-the-art flow-level and spatial traffic models, we characterize flow-level modeling in the context of wireless network performance evaluation. In general, performance evaluation can be carried out through simulations (mainly categorized into dynamic system level simulations, discrete event simulations, and Monte Carlo simulations), or through the use of analytical network models and tools (main approaches are flow-level modeling and stochastic geometry), see Table 2.1.

Tab. 2.1. Overview of System Models and Simulation Approaches

	System level simulation	Discrete event simulation	Flow-level models	Monte Carlo simulation	Stochastic geometry
Emulate dynamics	yes	yes	yes	limited	no
Computational efficiency	– – [1]	–	++	–	++
Scalability	– –	–	++	+	++
Flexibility (specific scenarios)	++	+	+	+	– –
Accuracy	++	+	+	–	–
Applicability to SON	– –	– –	+	–	–

[1] – – very low, – low, + high, ++ very high

Discrete event simulations and flow-level models form one group where the *dynamic* behavior of network elements and users is emulated. Since the evaluation of the dynamic behavior is costly in terms of computational effort, the aforementioned tools are associated with a less elaborate consideration of link-level aspects. In contrast, Monte Carlo simulations and stochastic geometry attempt to characterize the *static* network behavior by taking into account *snapshots* of the network's state, which are characterized by, for example, the number and positions of users in Monte Carlo simulations. Or they aim at obtaining KPI statistics of random deployments and/or user distributions by, for example, modeling their positions as Poisson Point Processes in stochastic geometric approaches. Dynamic system level simulations form another group, in which specific architectural or algorithmic aspects are emulated in minute detail. Simulation approaches usually exhibit a remarkable computational effort since a lot of events (system level simulation,

discrete event simulation) or user drops (Monte Carlo simulation) are necessary to obtain entire KPI statistics reliably. The computational effort is also the reason why simulations poorly scale with the number of wireless nodes, that are base stations (BSs) and user terminals. Low scalability is especially the case for system level simulations, in which the entire Open Systems Interconnection (OSI) protocol stack (or many parts of it) is emulated. However, the flexibility of simulation approaches is high with respect to environmental conditions, such as user or traffic distributions, or with respect to architectural aspects, such as complex algorithms or BS deployments. This flexibility facilitates an accurate performance evaluation if the statistical validity is ensured. Stochastic geometry appears to be an attractive solution to evaluate the performance of arbitrarily large networks analytically and therefore with low computational effort. Nonetheless, the flexibility of this approach is limited. For instance, finding analytical expressions for relevant KPIs considering arbitrary user or BS *distributions* may become very cumbersome. An analysis is often possible only in very specific cases, such as uniform distributions of nodes.

Flow-level models can be seen as the analytical counterpart of discrete event simulations. They are considered to be a powerful tool to describe the dynamics of Internet traffic [Rob01] in wired and wireless systems since 2001. Flow-level modeling is based on queuing theory and aims at deriving analytical expressions for a wide range of KPIs, such as the so-called flow sojourn time or the mean number of concurrent flows per server. Flow-level models combine the benefits of mathematical solutions, low computational effort and a scalability with the number of network elements, and the advantages of discrete event simulations, flexibility and accuracy. Moreover, arbitrary BS deployments and spatial traffic or user distributions (see Section 2.2) can be considered, which constitutes a fundamental advantage over approaches using stochastic geometry. Flexibility, accuracy, and computational efficiency make flow-level models very attractive for the integration with SON algorithms, see Section 2.3. In what follows, we further sub-divide flow-level models into categories, namely *unbounded versus bounded* and *isolated versus interference-coupled systems*. Interference-coupled systems are of particular importance to modern cellular network performance evaluation.

2.1 Flow-Level Models for Cellular Networks

Since the mathematical analysis of performance at the Internet Protocol (IP) packet layer or user session layer is rather intricate, the notion of so-called *data flows* [Rob01; Fre+01] has become established for modeling and quantifying the performance of packet-switched networks in the last fifteen years. A data flow is a set of IP packets that belong to specific objects, such as a web page, an Email, a video stream, etc. A user usually initiates multiple data flows during one session. As a result thereof, a session is characterized by a number of consecutive data flows

and a so-called thinking time between the flows. The scientific field of flow-level models is very broad in general and very diverse in terms of its applications. For instance, the research community distinguishes between *elastic* data flows, the data rate of which adapts to the available bandwidth or channel conditions, and *live streaming* flows with strict and fixed bandwidth requirements. Here we focus on flow-level models applied in a wireless network context, more specifically adapted to cellular data networks, and consider elastic data traffic only, which includes *buffered* streaming as well.

The tool set for characterizing the performance at flow level is queuing theory. In the wireless network context, each BS represents *one* server in a queuing system. Radio resources, for example, resources in a time-frequency grid, are shared among concurrent elastic data flows at the server. The resulting flow-level model accounts for the dynamics of incoming data flows as well as the dynamics of the service process, or more specifically, the presence of other data flows competing for the same resources in a cell. The dynamics are formally modeled by a stationary continuous-time random process $\{X_{i,t} : t \in \mathbb{R}_+\}$, which describes the number of concurrently active flows served by BS i at each time instant $t \in \mathbb{R}_+$, where $\mathbb{R}_+$ denotes the set of positive real numbers. We write X_i for this process and $X_i(t)$ for the value it takes at time t in the remainder of this thesis.

Unbounded Isolated Systems

It is a well-known result that the steady-state probabilities π_i of the random process $X_i(t) \in \mathbb{N}_+$ ($\mathbb{N}_+$ is the set of positive integers) become

$$\pi_i(x_i) := \Pr[X_i(t) = x_i] = (1 - \rho_i)\,\rho_i^{x_i} \quad \text{for } \rho_i < 1 \tag{2.1}$$

in an unbounded system with processor sharing (PS) service discipline [Kle76] and Poissonian arrivals. The load ρ_i of the i^{th} BS in Eq. (2.1) is defined as the ratio of an arrival rate λ_i of flows and a service rate μ_i, that is $\rho_i = \lambda_i/\mu_i$. The probability operator is denoted as $\Pr[\cdot]$. According to Kendall's notation [Ken53], such a system is referred to as an M/M/1/∞-PS system if in addition to exponentially distributed inter-arrival times (Poissonian arrivals) the service time is exponentially distributed as well. The process $X_i(t)$ is a so-called continuous-time Markov process if the system possesses the aforementioned properties. It is characterized by the *memorylessness* property, which means that the probability of a state transition is only dependent on the current state but not on the preceding states. There exist also other service disciplines in queuing theory, such as First-In First-Out (FIFO) or Last-In First-Out (LIFO). However, the PS discipline is more appropriate in the wireless context, where multiple mobiles compete for same resources and where the mobiles are, in principle, served concurrently. In particular, the egalitarian processor sharing (EPS) discipline is the queuing-theoretical counterpart of the

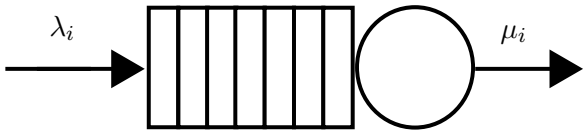

Fig. 2.1. Standard queuing system with data flow arrival rate λ_i (input) and service rate μ_i (output).

popular Round Robin scheduler, which is often implemented in practical systems [Kle76] and mostly assumed for wireless BSs.

Fig. 2.1 depicts the common illustration of a queuing system with an arrival intensity λ_i and an average service rate μ_i, both given in data flows per unit time. The maximum number of concurrent flows being served by the system described above is infinite. This means that the system is unbounded and that the number of active data flows can grow indefinitely, in particular if $\lambda_i \geqslant \mu_i$. In this specific case, the system is said to be *unstable*, which is the most common drawback of such systems if they are operated at high load. In general, steady-state probabilities $\pi_i(x_i)$ cannot be computed and the derivation of KPIs fails for unstable systems.

There has been extensive work on modeling of and investigating on flow-level dynamics of such unbounded systems in a wireless communications context. For example, the authors in [BBP04] and [Bon05] study the impact of user mobility and opportunistic scheduling on the performance of wireless systems, respectively. They found, for example, that user mobility generally improves the cellular performance at flow level. Furthermore, capacity gains of frequency reuse schemes in OFDMA networks have been analyzed in [BH09]. A very interesting, yet mathematically challenging, approach can be found in [CA13], where flow-level models are combined with stochastic geometry to derive the distribution of the load within the network.

Bounded Isolated Systems

Bounded Markovian systems, such as the M/M/1/K_i-EPS system, restrict the maximum number of concurrently active flows through admission control to a positive integer K_i, such that $X_i(t) \in \{0, \ldots, K_i\}$. If the system is full, that is $X_i(t) = K_i$, any other arriving flow is not admitted service and is blocked. Therefore, bounded systems are stable in the sense that the number of flows cannot grow indefinitely. The generating function of the steady-state probabilities of the M/M/1/K_i-EPS queue is given as

$$\pi_i(x_i) := \Pr\left[X_i(t) = x_i\right] = \begin{cases} \dfrac{(1-\rho_i)\,\rho_i^{x_i}}{1-\rho_i^{K_i+1}}, & \text{for } \lambda_i \neq \mu_i, \\ \dfrac{1}{K_i+1}, & \text{for } \lambda_i = \mu_i. \end{cases} \tag{2.2}$$

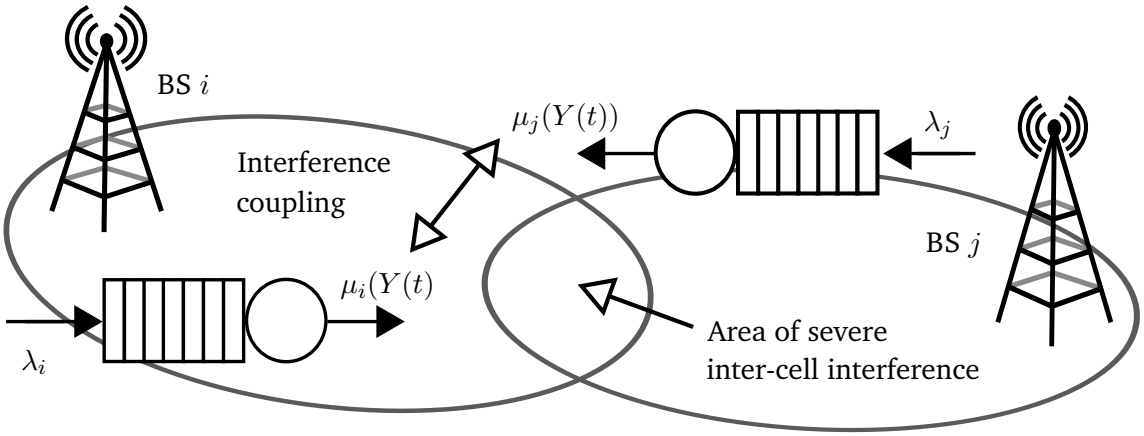

Fig. 2.2. Interference-coupling between interfering BSs. The activity of BS i affects the service rate μ_j of BS j and vice versa. This causes a mutual coupling among all BSs through the interference process $Y(t)$.

Bounded systems have been used to study the integration of elastic and live-streaming traffic in [BH07], and to derive the startup delay distribution and buffer starvation probabilities of buffered streaming services in [Xu+13].

Unbounded, Interference-Coupled Systems

Dynamic inter-cell interference introduced by one BS affects the performance of neighboring BSs in modern frequency-reuse-one wireless networks. Let $\{Y_t : t \in \mathbb{R}_+\}$, in short $Y(t)$, denote a stationary, continuous-time random vector process, the i^{th} element of which quantifies the interference level generated by BS i. The number of BSs considered is denoted as N. Since the dynamics of interference directly translate to a variation of the data rates experienced by individual users in neighboring cells and therefore to a variation of their mean service rate, we write $\mu_i\left(Y(t)\right)$ for the queue's service rate. The severe impact of inter-cell interference, which depends on the activities of the BSs, evokes a *mutual coupling* of the flow-level service processes in all BSs, see Fig. 2.2.

A first attempt to jointly characterize the complex interactions of flow-level dynamics, $X_i(t)$, and inter-cell interference dynamics, $Y(t)$, has been made in [Bon+04]. The authors derive second-degree approximations, which are upper and lower performance bounds of the mean number of active flows served by a BS. The simple assumption is that interfering BSs provide maximum or minimum data rates by experiencing either no or full interference, respectively. This assumption has two drawbacks. Only the performance of one cell under consideration is approximated and its own impact on the performance on surrounding BS is neglected.

Due to the increasing complexity of wireless networks and the increasing number of nodes per unit area, studying the coupling among interfering BSs has gained momentum again in the last four years. A detailed analytical study on the cell load coupling through a time-averaged interference approximation has been carried out

in [SY12a]. Studying the framework of interference function calculus in [Yat95], the authors in [Cav+14] concluded the framework's potential usefulness to efficiently characterize the load distribution across BSs and other performance metrics, which can be derived from the BS loads. However, one drawback of using interference function calculus in the context of cellular network performance evaluation is the assumption of time-averaged interference. Since users are served in a best-effort manner in most OFDMA-based network technologies, such as Wifi or Long Term Evolution (LTE), interference power experienced by users follows an on/off scheme rather than a time-average characteristic. Let for further derivations

$$X(t) := (X_1(t), \ldots, X_N(t)) \tag{2.3}$$

denote the random vector process, which collects all BS states and describes the state of the entire network. The authors in [FF12] make the following assumption in order to capture the aforementioned on/off characteristic of the process $Y(t)$

Assumption 1 (Best effort service)**.** If there is at least one active data flow, the corresponding BS is said to be *active*. An active BS allocates all radio resources and therefore transmits with full power, as long as all flows have been served.

Assumption 1 ultimately results in the following relation between the random process $X(t)$ and the interference process $Y(t)$

$$Y(t) = \operatorname{sgn}(X(t)). \tag{2.4}$$

Since the service rate μ_i is a function of the interference process $Y(t)$ and the process $Y(t)$ varies on the same time scale as the flow dynamics $X_i(t)$, we have to – in contrast to the time-averaged interference approach – consider the temporal dynamics of both processes jointly. A two-dimensional state transition diagram for two interference-coupled BSs is illustrated in Fig. 2.3. As can be seen, the transition rates $\mu_{(\cdot)}$ from higher to lower states depend on the activity of other BSs, that means on the fact that they serve at least one flow or not. Moreover, as users experience different data rates depending on their location within the cell, they represent different *classes* of flows. Therefore, the problem at hand relates to performance evaluation of *multi-class processor sharing queuing systems with mutually interference-modulated service rates*. The derivation of the steady-state probabilities

$$\pi(x) := \Pr[X(t) = x] \tag{2.5}$$

of such queues appears to be extremely difficult due to the *combination* of the following three facts:

1. The modulation of the service rate of some cell under consideration, that is the variation of the external interference and hence the data rates within the cell,

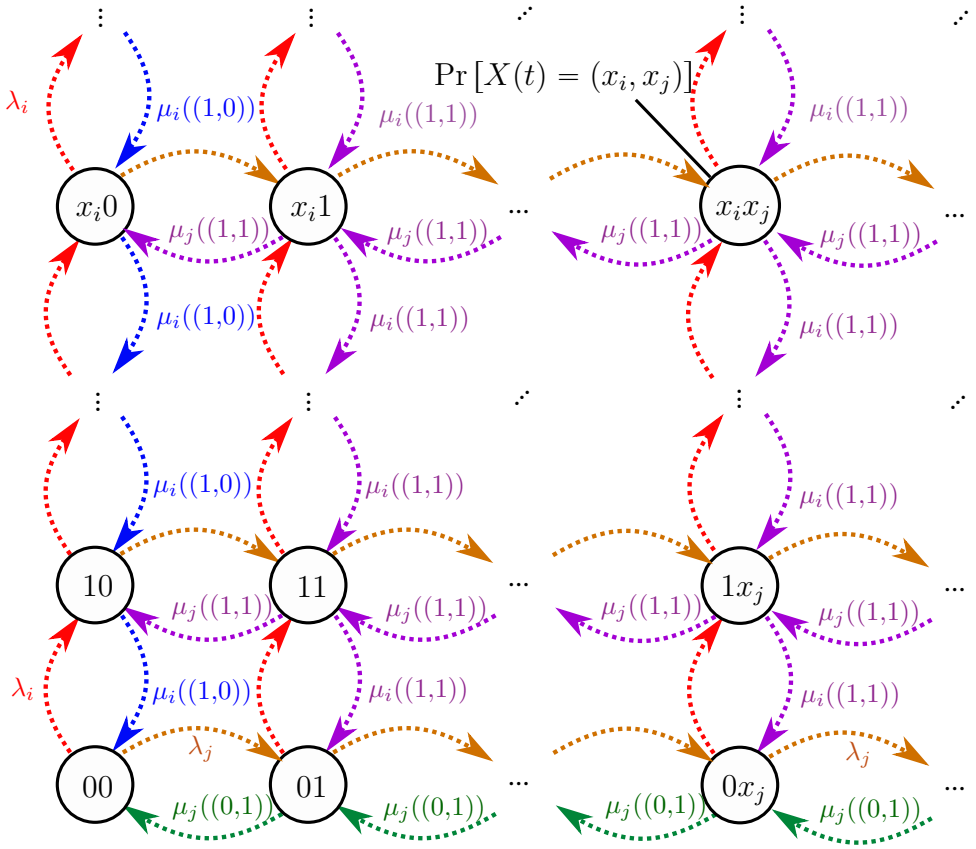

Fig. 2.3. State transition diagram for two interference-coupled BSs. The service rates μ_i are state-dependent, which means that they depend on the interference scenario $Y(t) = \operatorname{sgn}(X(t))$.

happens on the same time scale as the flow dynamics (arrivals and departures of flows) within this cell. More severely, the service rate usually changes *during* the service of data flows.

2. The modulation of the service rate is experienced by *all* concurrently active data flows due to the assumption of the PS service discipline. More importantly, the modulation is experienced *differently* by the flows because the impact of inter-cell interference on the data rates is stronger at the cell edge compared to the cell center (multi-class queuing system).
3. The service rates are *mutually modulated* through relation (2.4).

Queuing systems with modulated service rates are rarely considered in literature and only a few analytical results for very specific systems exist, such as for M/MM/1-FCFS (MM: Markov-modulated; FCFS: first come first served) queues, see for instance [ZG99; MG05]. To this end, the authors in [FF12; FF13] propose a state aggregation method, in which subsets of the states are subsumed to state aggregates to approximate the performance of interference-coupled wireless networks with *unbounded* state space. Indeed, the resulting joint consideration of the processes $X(t)$ and $Y(t)$ yields far more accurate results than assuming time-averaged interference [FF13].

Bounded Interference-Coupled Systems

So far, bounded interference-coupled systems have not been analyzed. One exception is [MK10], in which the authors propose a conservative cell load approximation and use the Erlang-B loss formula to account for stabilization of the network and computing blocking probabilities. Although the Erlang-B loss formula and related M/M/c/c models (c is the number of dedicated channels, which are used by c data flows at maximum) hold for computing blocking probabilities of voice calls in telephone networks, such as 2G wireless networks, they are not applicable to 3G or 4G systems, in which Internet data traffic is usually transmitted through BSs according to best-effort service disciplines, such as PS.

Impact of Flow Dynamics on Buffered Streaming Traffic Performance

Theoretical modeling of video performance in wireless networks has gained a lot of interest in the past four years due to the increasing popularity of ubiquitous mobile video services. For instance, the authors in [Xu+12b] model the impact of fast fading and scheduler strategies on the distribution of playback intervals. They conclude that QoE optimization should be carried out at the BS instead by individual users (or through their applications). In another publication [Xu+12a], the authors model the video playout buffer as a M/M/1 queue on packet level to compute the exact distribution of video buffer starvation events. A first attempt to model the impact of flow-level dynamics on startup delay distributions and buffer starvation probabilities is made in [Xu+13]. It has been found that flow-level dynamics predominate effects induced by fast fading and variable bitrate streaming in bounded isolated M/M/1/K_i-EPS wireless systems.

2.2 Spatial Data Traffic Models

There exist a lot of studies on the analysis of mobile data traffic statistics already. Many of them are based on geographical and demographical factors, see [TTG98] among others, or theoretical approaches using point process theory, such as the one in [MSY14]. However, the correlation of such data with actual traffic demands is not guaranteed. Other studies, as the one presented in [Nan+13], focus on the analysis of the temporal characteristic of data volumes downloaded through individual BSs and extrapolate daily or weekly traffic profiles. Such profiles are very helpful for network-wide performance evaluation with respect to, for example, network energy efficiency. Furthermore, detailed temporal analysis of High-Speed Downlink Packet Access (HSDPA) traffic in [Lan+12] shows that the number of connected users and their throughput follow a characteristic daily curve and that the arrival of users can indeed be modeled as a Poisson process. The finding that the user arrival is Poissonian is another motivation for using the flow-level models, in which we use the terms *data flow* and *mobile terminal* or *user* interchangeably.

Spatial analysis of mobile traffic based on live-network measurements is carried out in the publications [MRM11] and [Pau+11], in which the load distribution across cells of a nation-wide GSM (Global System for Mobile Communications) network and a nation-wide 3G network have been investigated, respectively. Interestingly, the authors in [MRM11] found that the 2G cell load distributions can be modeled by a mixture of log-normal distributions. In addition, the 2G area traffic *density* can be approximated by log-normal mixture distributions [Lee+14] as well.

Based on the aforementioned insights on more realistic traffic distributions, the authors in [LZN13; Lee+14] propose a method to efficiently generate log-normally distributed spatial data traffic maps with the help of three parameters: the mean and the variance of the traffic as well as a proxy parameter to control the spatial correlation. The method relies on the exponentiation of two-dimensional Gaussian random fields that are, according to the central limit theorem, generated through the summation of cosines with uniformly distributed angular frequencies and phase shifts. However, it appears to be cumbersome to control the spatial correlation via the proxy parameter if the generation of traffic maps with a specific correlation distance is envisaged. Furthermore, it is important to note that the spatial traffic analysis mentioned above is based on aggregated traffic volumes measured in the BSs. The common procedure is to distribute the traffic volumes uniformly within the Voronoi cells generated by the BS locations. The result of this procedure is that the spatial resolution of the maps generated is ultimately limited to the area of macro cells.

2.3 Model-Based Network Optimization

In general, SON can be categorized into heuristic short-time scale (milliseconds to seconds) algorithms, so-called SON use cases, or model-based medium- to long-term (minutes to days) optimization algorithms. Heuristic approaches often focus on the improvement of a small subset of KPIs or on the solution of specific problems, such as reducing the number of call drops. Since different SON use cases may operate on the same network parameters, for instance the BS transmit power, they have to be coordinated thus leading to sub-optimal results. Algorithms that are based on flow-level models, such as the one presented in our publication [Feh+13a], usually operate on time-scales of several minutes to a few hours. The reasons are that they are based on average spatial traffic distributions and that they rely on the computation of long-term KPIs, such as the average cell load or throughput statistics. The advantages of flow-level based network optimization are the accurate description of KPIs as a function of network parameters, and therefore a more reliable and effective network optimization. We briefly shed light on some related existing work below since one aspect of this thesis is the application of flow-level models to SON-algorithms.

Flow Level-Based Network Optimization

Since obtaining KPIs using flow-level modeling relies on the computation of cell loads within the coverage areas of BSs, such models are often applied to cell size optimization [CAA12], cell load balancing [Kim+10; SY12b], or power allocation and range assignment [YLY14] algorithms. The goal is to steer the BS loads to enhance the fairness among users or other metrics. More advanced analytical adaptations of the models allow for the design and analysis of more complex algorithms, such as inter-cell interference coordination [CAA13], intra-site CoMP (Coordinated Multi-Point) [KBE13], or coordinated beamforming [KBE15]. Due to the flexibility of flow-level models, they can also be applied in multi-parameter and multi-objective SON algorithms. For instance, joint user association and antenna down-tilt optimization algorithms can be found in our publications [Kle+12; Kle+13; Feh+14; Feh+13b]. Furthermore, joint bandwidth allocation and small cell switching methods have been developed in [Bar+13].

Using Spatial Traffic Models for Network Optimization Performance Evaluation

Despite the availability of methods for parameterizable generation of spatial data traffic distributions, literature generally lacks investigations on the performance of optimization algorithms considering a wide range of spatial traffic scenarios. Performance evaluation is mostly restricted either to uniform user distributions, to arbitrarily chosen scenarios, or to very specific cases, such as a single traffic map obtained in a certain city or region.

2.4 Contributions and Outline of this Thesis

This section briefly discusses the limitations and shortcomings of the aforementioned state-of-the-art solutions, from which the contributions of the thesis at hand are derived. Related aspects that are not discussed in this work are listed subsequently.

2.4.1 Limitations of Prior Art and Contributions Derived

The limitations of prior art and corresponding contributions are as follows:

A - Analysis of the Interference-Coupling of Unbounded EPS Systems

So far, admission control in conjunction with a detailed investigation of the effect of inter-cell interference dynamics and PS service discipline has not been considered yet. In fact, neglecting admission control is not realistic, since operators usually attempt to *guarantee* a minimum of quality of service through admission control for a certain percentage of the mobile users. This is especially important for (video) streaming applications with minimum bandwidth requirements [DPR04].

Another disadvantage of existing models for unbounded systems is that such systems are unstable if the arrival rate in a cell exceeds the average service rate. Then performance evaluation is impossible.

Contribution: The flow-level model from [FF12; FF13] is generalized in this theses by the introduction of **admission control** at the BSs. One is thereby able to consider interference-coupled cellular networks, the BSs of which are represented as servers in **bounded** and/or **unbounded** queuing systems. A holistic performance evaluation framework is presented, which includes the aforementioned extensions and addresses a **wide range of important KPIs**. The derivations of the KPIs and the numerical analysis have been published in parts in the research papers [KFF14; Kle+16a].

B - Buffered Streaming Performance in a Single Cell

To our best knowledge, [Xu+13] is, so far, the only attempt to model and analyze the impact of flow-level dynamics on buffered video streaming KPIs. Nevertheless, the assumptions on the underlying network model are rather simplistic. Firstly, the authors assume isolated cells, in which the performance is not affected by inter-cell interference. Secondly, they assume homogeneous data rates for all concurrent flows within the cell (*single-class* PS). Both assumptions limit the general validity of the model in a wireless context considerably.

Contribution: The modeling approach in [Xu+13] is taken and generalized by considering the impact of inter-cell interference coupling and multi-class PS on streaming flow performance in the thesis at hand. In particular, **approximations of the startup delay distribution** and **video playout buffer starvation probabilities** are provided in a **multi-cellular context**, that is for different locations, across a cell, and also within the entire network. The derivation of the startup delay distribution has been published in [KF15a; KF15b].

The contributions to the aspects A and B are presented in **Chapter 3**.

C - Spatial Traffic Analysis based on Measured Data Volumes in BSs

So far, spatial traffic distributions have been modeled based on measured 2G and 3G data volumes transferred through macro BSs. A common approach is to distribute the traffic demand evenly within the BSs' Voronoi cells. This inherently limits the spatial resolution of the data traffic distribution to the size of *macro* cells. This is particularly unfavorable for the investigation of the performance of networks with much smaller cells, such as *pico* or *femto* cells, because traffic hot spots that could be easily covered by small cells are hidden through the poor resolution. The work in [LZN13; Lee+14] presents an efficient algorithm to generate spatial log-normally

distributed data traffic demand maps. However, the authors' model relies on the inaccurate approach using Voronoi cells of macro BSs as described above. Another drawback is that a correlation distance cannot be considered directly as input variable, which makes the generation of maps with a specific correlation distance cumbersome.

Contribution: In order to analyze and model spatial traffic distributions more realistically and more accurately, measurements from so-called call traces from a live 3G network, which are geo-located through different techniques, are taken. The measurements allow for a **higher spatial resolution** of traffic maps, which is mainly independent of the cell size. Data processing steps are provided to enable an **analysis of hot spot characteristics** and **statistical traffic distributions** based on the data measured. Furthermore, a novel method is presented for the **efficient generation of spatially correlated data traffic demand maps**. In contrast to the approach in [LZN13; Lee+14], this method considers the **correlation distance as direct input**. The analysis and modeling approaches are part of **Chapter 4** and have been published in [Kle+14; KSF15].

D - Optimization Algorithm Performance for Limited Spatial Traffic Scenarios

Although many network optimization algorithms rely on realistic flow-level models or dynamic system level simulations, the importance of spatial traffic distributions is often neglected. More often than not, simple user or traffic distributions are assumed for the evaluation of network optimization algorithms, such as a uniform distribution. Statements on the algorithms' applicability to certain environments or about its robustness to changes in spatial traffic demand are very limited as a result of this assumption.

Contribution: The **flexibility of the generalized flow-level model (A)** is illustrated by adapting certain modeling aspects to fit two specific algorithmic approaches in **Chapter 5**, namely: **data offloading** using so-called further enhanced inter-cell interference coordination (FeICIC) and cell range expansion (CRE), and **traffic-adaptive beamforming** with phased-array antennas. Furthermore, the spatial traffic model from C is applied to the beamforming approach to ensure a **performance evaluation with realistic traffic distributions under a wide range of settings**, such as the mean traffic, hot spot intensity, and hot spot size. The optimization algorithms as well as recommendations to apply the spatial traffic model have been published in [KGF14b; KGF14a; KF14; KSF15].

Chapter 6 concludes this thesis and provides recommendations for future work.

All the aforementioned extensions and contributions in their entirety form a holistic performance evaluation framework, which meets the requirements listed in