Mobility-aware QoS assurance in software-defined radio access networks: an analytical study

Vassilakis, Vassilios, Moscholios, Ioannis, Bontozoglou, Andreas and Logothetis, Michael (2015) Mobility-aware QoS assurance in software-defined radio access networks: an analytical study. In: 1st IEEE Conference on Network Softwarization (NetSoft), 13-17 Apr 2015, London, UK.

http://dx.doi.org/10.1109/NETSOFT.2015.7116192

Mobility-Aware QoS Assurance in Software-Defined Radio Access Networks: An Analytical Study

Vassilios G. Vassilakis*, Ioannis D. Moscholios†, Andreas Bontozoglou‡, Michael D. Logothetis§
* Computer Laboratory, University of Cambridge, Cambridge, U.K.
† Dept. of Informatics & Telecommunications, University of Peloponnese, Tripolis, Greece
‡ School of Computer Science and Electronic Engineering, University of Essex, Colchester, U.K.
§ Dept. of Electrical & Computer Engineering, University of Patras, Patras, Greece

Abstract—**Software-defined networking (SDN) has gained a tremendous attention in the recent years, both in academia and industry. This revolutionary networking paradigm is an attempt to bring the advances in computer science and software engineering into the information and communications technology (ICT) domain. The aim of these efforts is to pave the way for completely programmable networks and control-data plane separation. Recent studies on feasibility and applicability of SDN concepts in cellular networks show very promising results and this trend will most likely continue in near future. In this work, we study the benefits of SDN on the radio resource management (RRM) of future-generation cellular networks. Our considered cellular network architecture is in line with the recently proposed Long-Term Evolution (LTE) Release 12 concepts, such as control-data plane split, heterogeneous networks (HetNets) environment, and network densification through deployment of small cells. In particular, the aim of our RRM scheme is to enable the macro base station (BS) to efficiently allocate radio resources for small cell BSs in order to assure quality-of-service (QoS) of moving users/vehicles during handoffs. We develop an approximate, but very time- and space-efficient algorithm for radio resource allocation within a HetNet. Experiments on commodity hardware show algorithm running times in the order of a few seconds, thus making it suitable even in cases of fast moving users/vehicles. We also confirm a good accuracy of our proposed algorithm by means of computer simulations.**

Keywords*—*Radio access network; software-defined networking; quality of service; radio resource management; handoff.

I. Introduction

Software-defined networking (SDN) has recently attracted lots of research efforts and has gained a tremendous attention from both academic and industry communities [1]. It can be described as a revolutionary networking paradigm attempting to incorporate the solutions developed within the computer science and software engineering domains into the information and communications technology (ICT) domain. The ultimate goal is to enable completely programmable networks, which can be achieved by decoupling the control and data planes. The expected benefits of SDN include, but not limited to the following: a) decreased CAPEX and OPEX for network operators, by reducing the cost of hardware, automating services and management, and reducing power consumption; b) flexibility in terms of deployment and operation of new infrastructure, services, and applications; c) faster innovation cycles due to creation of enhanced services/applications and new business models.

Due to the aforementioned reasons, there are many SDN standardization efforts by various industry associations and standards bodies, such as the Open Network Foundation (ONF), IETF's Forwarding and Control Element Separation (ForCES) Working Group, IRTF's SDN Research Group (SD-NRG), ETSI's Industry Specification Group for Network Functions Virtualization (ISG NFV), etc [2]. The ONF's OpenFlow protocol [3], [4] is considered as the first and most adopted SDN standard today. There have also been successful efforts on OpenFlow and SDN real-world deployments [5]. Following these trends, a number of SDN research activities in the cellular networks domain have recently started [6], [7]. It is also argued that SDN is going to play a major role and will become an integral part of future 5th generation (5G) cellular systems [8], [9]. Some preliminary studies on feasibility and applicability of SDN concepts in cellular networks show quite positive and promising results, both in the radio access network (RAN) [10], [11] and in the mobile core network [12].

In this work, we study the benefits of SDN on the radio resource management (RRM) of future-generation RAN. That is, our considered cellular network architecture is in line with the current 3GPP LTE Advanced (LTE-A) standard and the broadly accepted 5G concepts, such as user/control plane split, heterogeneous networks (HetNets) environment as a combination of macro cells and small cells, co-existence of multiple channel access schemes, etc. In particular, the aim of our RRM scheme is to enable the macro base station (macro-BS) to efficiently allocate radio resources for small cell BS (sc-BS) in order to assure quality-of-service (QoS) of moving users/vehicles during handoffs. To this end, we describe the call arrivals and departures as a continuous-time Markov chain (CTMC) and derive an efficient recursive formula for the calculation of state probabilities. Based on that, the required number of radio resources per small cell can be determined. We also model the handoffs between small cells and determine the required capacity in terms of radio resources of the target small cell. In order to achieve good time- and space-efficiency, we propose an approximate algorithm for cell capacity calculation. Experiments on commodity hardware show running algorithm times in the order of a few seconds. This property makes the proposed recursive algorithm suitable for dynamic radio resource allocation during handoffs, even in cases of fast moving users/vehicles. We also confirm the good accuracy of our proposed algorithm by means of computer simulations.

This paper is organized as follows. In Section II, we briefly review the relevant works. In Section III, we describe

978-1-4799-7899-1/15/$31.00 ©2015 IEEE

our considered model for RAN in a HetNet. In Section IV, we analyse the model using a CTMC and derive recursive equations for state probabilities and call blocking probabilities. In Section V, we perform simulation experiments to verify a good accuracy of our proposed algorithm. Finally, we conclude in Section VI.

II. Relevant Works

Below we briefly review a number of recent works on cellular software-defined RAN. These can be broadly classified into works on LTE RAN [10], [11], [13] and on HetNet RAN [14], [15]. Gudipati *et al.* [10] propose the SoftRAN, a software-defined control plane of LTE RAN. The control is performed by a virtual macro-BS, that acts as a centralized controller, and by a number of distributed radio elements. The radio elements perform local control decisions within a cell, whereas the centralized controller performs the inter-cell control tasks. However, SoftRAN mostly considers macro-BSs and, therefore, is not suitable for more advanced HetNet scenarios. Li *et al.* [11] propose an SDN solution for both LTE RAN and mobile core network. The control functions are abstracted from user devices and from forwarding elements. Also, a local agent is introduced to make fast control decisions for time-critical services. Bansal *et al.* [13] propose the OpenRadio, a programmable data plane that aims at supporting a continuous wireless network evolution. They introduce a software abstraction layer and a set of Application Programming Interfaces (APIs), that are both declarative and modular. OpenRadio can support different wireless protocol stacks, such as LTE, WiFi, and WiMax. Yang *et al.* [14] propose the OpenRAN, a software-defined HetNet RAN architecture based on cloud computing. According to dynamic radio network requirements, a centralized controller establishes a virtual BS from a pool of radio resources. On the other hand, the baseband processing utilizes a pool of cloud computing resources. Shrivastava *et al.* [15] propose an SDN-based framework for LTE-A HetNets. They study the problem of elastic radio resource sharing in a multi-operator environment of Frequency Division Duplex (FDD) macro-BS and Time Division Duplex (TDD) sc-BS. The proposed solution achieves reduced application layer delay, as confirmed by system-level simulation experiments.

We also briefly review some analytical models proposed for calculating cellular network capacity and for performing RRM [16], [17], [18], [19]. These models are based on the Kaufman-Robers (K-R) algorithm [20], [21] to achieve time- and space-efficient calculations. Staehle *et al.* [16] extend the K-R algorithm for the calculation of state probabilities and eventually the call blocking probabilities in the uplink of a Code-Division Multiple Access (CDMA) system. The call arrival process is assumed to be Poisson and calls have fixed bandwidth requirements. The derived approximate algorithm shows very good accuracy and short running times, as confirmed by simulations. This work has been extended in order for several important features to be included. Moscholios *et al.* [17] incorporated batched Poisson traffic, which is more peaked and bursty than Poisson traffic. The applied call admission control (CAC) is based on the partial batch blocking discipline. That is, one or more calls of an arriving batch are discarded, if the available cell resources are not enough. Another important extension is made by Vassilakis *et al.* [18], in order to incorporate elastic traffic and thus permitting handoff calls to reduce their bandwidth requirements, so that blocking is avoided when the cell load is above a predefined threshold. Hanczewski *et al.* [19] concentrated on soft handoff, and modelled the so-called *active set* of cells that participate in call handoff (having the best signal-strength-to-noise ratio). The proposed computational model is of a relatively low complexity. It makes use of a k-cast connections system; a handoff-call is blocked (may not be lost, because of the presence of hard handoff), when at least 1 connection out of several (i.e., the number of cells in the active set) is not available.

III. HetNet RAN Model

A. Basic Definitions and Assumptions

In this section, we describe our considered model for a cellular network. We assume a HetNet with one macro cell, controlled by a macro-BS and covering S small cells, each controlled by a sc-BS. In our considered scenario, the macro-BS is equipped with an SDN controller that may dynamically allocate radio resources to small cells. We consider K independent service-classes, with each representing a QoS level for mobile users (MUs).

In our model, an MU may be in one of the following two operational modes.

- ON-mode: MU is busy having a call in progress.
- OFF-mode: MU is idle and ready to receive a call.

The total cell radio resource consists of C_{total} radio resource units (RRUs). The definition of RRU for various channel access schemes is presented in the next subsection. The centralized controller dynamically allocates C_s RRUs for each small cell $s(s = 1, ..., S)$. For example, this could be done as in SoftRAN [10]. MUs that are in ON-mode, consume a number of RRUs of their corresponding small cell, as will be explained in the following subsection.

We also consider two different types of MUs: *new* and *handoff*. The first type refers to MUs that are trying to establish an initial connection. The second type refers to MUs that move from one cell to another while a call is on progress. The SDN controller facilitates the handoffs between small cells by allocating appropriate radio resources. If the required radio resources are not available at the time of arrival, the MU is blocked. The CAC policy, used at the macro-BS, is expected to guarantee that the *handoff-call* blocking probability will be significantly lower than the *new-call* blocking probability.

Each service-class k $(k = 1, ..., K)$ has a finite number of *new* MUs, denoted by $U_{k,N}$, and *handoff* MUs, denoted by $U_{k,H}$. In the following, we will use the indices N and H to refer to *new* and *handoff* MUs, respectively. The traffic generated by a service-class k *new* MU is denoted by $a_{k,N}$, whereas the traffic generated by a service-class k *handoff* MU is denoted by $a_{k,H}$. To model the transitions of MUs between ON and OFF modes, we define the *activity factor*, v_k, of service-class k as the ratio of the duration of ON periods over the total call duration.

B. Defining the Radio Resource Unit

In order to describe our model as a CTMC, we need to appropriately define the RRU. However, this definition depends on the channel access method that is used in the network. In case of Frequency Division Multiple Access (FDMA)/ Time Division Multiple Access (TDMA)-based schemes, the RRU definition is straightforward. For example, in case of LTE Orthogonal FDMA (OFDMA), one RRU may be equivalent to one resource block (RB). That is, if the LTE channel bandwidth is 9 MHz and one subcarrier is 15 kHz, then there are in total 600 subcarriers. In is known that one OFDMA RB consists of 12 subcarriers (during one time slot of duration 0.5 ms), which corresponds to 50 RBs per channel.

However, in case of CDMA-based systems, due to their soft capacity and inter-cell interference, the RRU definition is slightly more complicated and is shown below. As it is known, the capacity of CDMA-based systems is limited by the multiple access interference (MAI). That is, if a new MU is admitted in the system, this will cause interference to all other existing MUs. To model this characteristic, the *cell load*, n, which represents the total occupied cell radio resource, has been introduced [16]. The *cell load* consists of the *intra-cell load*, n_{intra}, (caused my MUs of the same cell) and the *inter-cell load*, n_{inter}, (caused my MUs of neighbouring cells).

The bit rate of an MU depends on its service-class k and on the cell load, n, at the moment of call arrival. In particular, we define two rate thresholds: an upper threshold, $n^U_{k,t}$, and a lower threshold, $n^L_{k,t}$ for service-class k, type $t (t \in \{N, H\})$ MUs. If the cell load is below the lower threshold, then the MU requests its *peak (P)* bit rate, $R^P_{k,t}$. If the cell load is above the lower threshold but below the upper threshold, then the MU requests its *medium (M)* bit rate, $R^M_{k,t}$. Finally, if the cell load is above the upper threshold, then the MU requests its *low (L)* bit rate, $R^L_{k,t}$.

When a *new* or *handoff* MU is accepted in the cell, the cell load, n, is increased by the so-called *load factor*, $LF^r_{k,t} (r \in \{P, M, L\})$, which is defined as follows:

$$LF^r_{k,t} = \frac{R^r_{k,t} SNR_k}{W + R^r_{k,t} SNR_k} \tag{1}$$

where, SNR_k is the required signal-to-noise ratio (SNR) for service-class k and W is the chip rate. A typical value in W-CDMA systems is $W = 3.84$ Mcps.

Now, to define the cell load, n, load factors, $LF^r_{k,t}$, and load thresholds, $n^U_{k,t}$, $n^L_{k,t}$ in terms of RRUs, we use a *basic discretization unit*, g, as follows:

$$j = \lfloor \frac{n}{g} \rfloor, b^r_{k,t} = \lfloor \frac{L^r_{k,t}}{g} \rfloor, J^U_{k,t} = \lfloor \frac{n^U_{k,t}}{g} \rfloor, J^L_{k,t} = \lfloor \frac{n^L_{k,t}}{g} \rfloor \tag{2}$$

where the floor function $\lfloor x \rfloor$ gives the largest integer less than or equal to x. A typical range of values used for g is [0.001 - 0.005] [18].

IV. Calculating Blocking Probabilities

A. Call Admission Control

When a *new* MU starts a call or a *handoff* MU arrives to a small cell, it requires $b^r_{k,t}$ RRUs. If these RRUs are available in the small cell, the MU is accepted. Otherwise, the SDN controller may allocate additional RRUs to this cell, thus increasing its capacity, C_s. This is possible only if there are some available RRUs in the macro cell, out of C_{total}. If not, the MU is blocked.

We consider the following CAC policy for the acceptance of a service-class k, type t, rate r call:

$$j_{intra} + j_{inter} + LF^r_{k,t} \leq j_{max,t} \tag{3}$$

where $j_{intra} = \lfloor \frac{n_{intra}}{g} \rfloor$, $j_{inter} = \lfloor \frac{n_{inter}}{g} \rfloor$, and $j_{max,t}$ is the CAC threshold for type t MUs.

That is, the MU is accepted if and only if, after the acceptance, the total number of occupied RRUs is not going to exceed the threshold $j_{max,t}$. The thresholds $j_{max,t}$ are constrained by the total system capacity, C_{total}:

$$j_{max,t} \leq C_{total}, \forall t \tag{4}$$

Note that the *rate thresholds* $J^U_{k,t}$ and $J^L_{k,t}$ of (2) are used to select the appropriate bit rate for MUs, whereas the *CAC threshold* $j_{max,t}$ of (3) is used to decide whether to accept or block an MU.

Due to the CAC of (3), some of the arriving calls in the cell may be blocked. The probability that an MU of service-class k, type t, and rate r is blocked when arriving at an instant with j_{intra} is called *local blocking probability* (LBP) and defined as:

$$\beta^r_{k,t}(j_{intra}) = Pr[j_{intra} + j_{inter} + LF^r_{k,t} > j_{max,t}] \tag{5}$$

By performing some calculations, we get:

$$\beta^r_{k,t}(j_{intra}) = 1 - CDF_{j_{inter}}(j_{max,t} - j_{intra} - LF^r_{k,t}) \tag{6}$$

for $LF^r_{k,t} \leq j_{max,t} - j_{intra}$, where $CDF_{j_{inter}}()$ is the cumulative distribution function of j_{inter}.

B. Tutorial Example

Below we explain the aforementioned concepts with the aid of a simple example. Consider a small cell, s, and $K = 2$ service-classes. The total system capacity is $C_{total} = 10$ RRUs, whereas the current allocated capacity of this particular small sell is $C_s = 6$ RRUs. The first service-class has only a peak data rate of $R^P_{1,t} = 64$ Kbps, whereas the second has a peak data rate of $R^P_{2,t} = 256$ Kbps, medium data rate of $R^M_{2,t} = 192$ Kbps, and low data rate of $R^L_{2,t} = 128$ Kbps. If one RRU is equivalent to a 64 Kbps data rate, then $b^P_{1,t} = 1$. The first service-class has no rate thresholds and, therefore, $b^P_{1,t} = b^M_{1,t} = b^L_{1,t}$. The second service-class has one rate threshold: $J^U_{2,N} = 6$ for *new* MUs and two rate thresholds, $J^U_{2,H} = 6$, $J^L_{2,H} = 3$ for *handoff* MUs. Therefore, $b^P_{2,N} = b^M_{2,N} = 4 > b^L_{1,N} = 2$ and $b^P_{2,H} = 4 > b^M_{2,H} = 3 > b^L_{2,H} = 2$.

Consider three MUs in the small cell. The first MU is *new* and belongs to the first service-class. The second MU is also *new* but belongs to the second service-class. The third MU is *handoff* and belongs to the second service-class. At a given moment, some MUs may be in ON mode, whereas other will be in OFF mode. The parameter j, considered as the *system state*, represents the number of RRUs occupied in the cell, assuming that all MUs are in ON mode. Therefore, in this example $j = b^P_{1,N} + b^P_{2,N} + b^L_{2,H} = 1 + 4 + 2 = 7$. Since $j > C_s$, the macro-BS will allocate (if available) $j - C_s = 1$

RRU to the small cell. The parameter j_{intra}, defined as the *resource occupancy*, represents the actual number of occupied RRUs (i.e., takes into account only MUs in ON mode). In this example, $j_{intra} = 0$ if all MUs are in OFF mode, $j_{intra} = 1$ if the first MU is in ON mode and two other MUs are in OFF mode, $j_{intra} = 4$ if the second MU is in ON mode and two other MUs are in OFF mode, etc.

It is clear from the above that j depends on the number of MUs and their service-classes, whereas j_{intra} also depends on the MU's mode. Also note that $0 \leq j_{intra} \leq j$. When all MUs are in OFF mode, we have $j_{intra} = 0$; when all MUs are in ON mode, we have $j_{intra} = j$.

C. Continuous-Time Markov Chain Modeling

In this subsection, we describe the process of arrivals and departures of MUs as a CTMC. This will enable us to derive an efficient formula for the calculation of system state probabilities. Based on them, we will then be able to determine the *new-call* and *handoff-call* blocking probabilities. To understand the CTMC modeling approach, consider a cellular system of capacity $C_{total} = 3$ RRUs and $K = 2$ service-classes. In Fig. 1 the four possible states $j = 0, 1, 2, 3$, are represented by circles and the transitions between states are represented by arrows. The resource requirements of the first service-class are $b^P_{1,N} = 3$, $b^L_{1,N} = 2$, $b^P_{1,H} = 2$, and $b^L_{1,H} = 1$, for peak-rate *new* MUs, low-rate *new* MUs, peak-rate *handoff* MUs, and low-rate *handoff* MUs, respectively. Similarly, the corresponding resource requirements of the second service-class are $b^P_{2,N} = 2$, $b^L_{2,N} = 1$, $b^P_{2,H} = 2$, and $b^L_{2,H} = 1$. Also assume that the upper thresholds are $J^U_{1,t} = 1$ and $J^U_{2,t} = 2$ for the first and the second service-class (both *new* and *handoff*) MUs, respectively. For simplicity assume that there are no lower thresholds $J^L_{k,t}$.

The state transition diagram of the CTMC for the first service-class is shown in Fig. 1. Transitions from lower to higher states occur due to call arrivals, whereas transitions from higher to lower states occur due to call departures.

We use the following notation:

- $\lambda_{k,N}(j)$: effective transition rate from state (j) to state $(j + b_{k,N})$ due to *new* MUs of service-class k
- $\lambda_{k,H}(j)$: effective transition rate from state (j) to state $(j + b^P_{k,H})$ if $j < J^U_{k,H}$, or to state $(j + b^L_{k,H})$ if $j \geq J^U_{k,H}$, due to *handoff* MUs of service-class k
- $\mu_{k,t}$: service rate of service-class k, type t MUs
- $M^r_{k,t}(j)$: mean number of service-class k, type t, rate r MUs in state j

In Fig. 1 we observe that some transitions, for example from (0) to (3), have the corresponding transitions in the reverse direction, i.e. from (3) to (0). For cases like that, we can assume the following *local balance* between to corresponding states:

$$\lambda_{k,t}(j)q(j) = \mu_{k,t} M^r_{k,t}(j + b^r_{k,t})q(j + b^r_{k,t}) \tag{7}$$

for $j = 0, ..., C_{total} - b^r_{k,t}$

We also observe that some transitions to lower states, for example from (2) to (1), have no corresponding transitions to higher states. For cases like that, in order to approximate the reversibility, we will assume that these transitions to lower sates do not exist. That is, it will be assumed that the corresponding transition rates are negligible (i.e., $M^L_{1,H}(2) \approx 0$). In general, we assume the following:

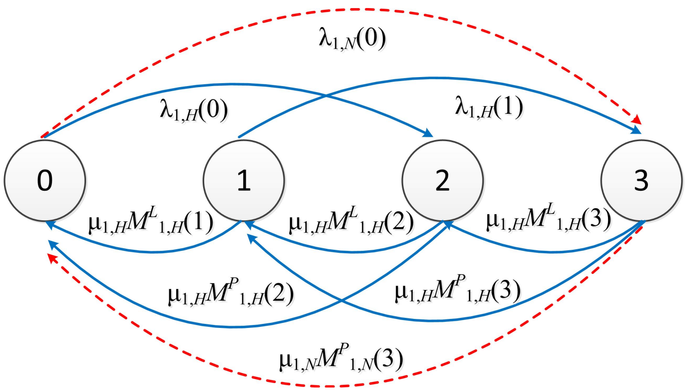

Fig. 1. State transition diagram for the first service-class.

$$\begin{aligned} &M^L_{k,t}(j) = 0 \text{ if } j < J^U_{k,t} \\ &M^M_{k,t}(j) = 0 \text{ if } j < J^L_{k,t} \text{ or } j \geq J^U_{k,t} \\ &M^P_{k,t}(j) = 0 \text{ if } j \geq J^L_{k,t} \end{aligned} \tag{8}$$

It is clear that some approximation errors are introduced through the assumptions (7) and (8). However, in Section V, we show that the impact of these approximations on the accuracy of the derived results is not significant.

In Fig. 1, the states $j > C_{total} - b^P_{1,H} = 3 - 2 = 1$ are *blocking states* for *handoff* MUs of the first service-class, because the available RRUs in these states are less than required. As a consequence, transitions from states $j = 2, 3$ to higher states are not possible for *handoff* MUs. Similarly, there are three blocking states, $j = 1, 2, 3$, for *new* MUs of the first service-class.

The transition rates $\lambda_{k,t}(j)$ depend on the number of MUs in OFF mode, $U_{k,t} - M^r_{k,t}(j)$, the arrival rate of MUs, $\lambda_{k,t}$, and the blocking probability in state j, denoted as *state blocking factor*, $SBF^r_{k,t}(j)$:

$$\lambda_{k,t}(j) = (U_{k,t} - M^r_{k,t}(j))\lambda_{k,t}(1 - SBF_{k,t}(j)) \tag{9}$$

The SBFs are calculated in the following subsection.

D. Calculating the State Blocking Factors

In this subsection, we derive formulas for the calculation of $SBF^r_{k,t}(j)$. First, we define the *resource share*, $RS^r_{k,t}(j)$, of service-class k, type t, rate r MUs in a state j as follows:

$$RS^r_{k,t}(j) = \frac{M^r_{k,t}(j) b^r_{k,t}}{j} \tag{10}$$

The resource share represents the ratio of RRUs allocated to a particular MU category over the total number of allocated RRUs in a given state j.

Next, we define the *resource occupancy*, $RO(j_{intra}|j)$, as the conditional probability that j_{intra} RRUs are allocated in state j. The calculation of $RO(j_{intra}|j)$ is performed as follows:

$$RO(j_{intra}|j) = \sum_{k=1}^{K}\sum_{t}\sum_{r} RS^{r}_{k,t}(j)[v_k RO(j_{intra}- b^{r}_{k,t}|j - b^{r}_{k,t}) + (1-v_k)RO((j_{intra}|j - b^{r}_{k,t})] \tag{11}$$

for $j = 1, ..., j_{max}$ and $j_{intra} \leq j$, with $RO(0|0) = 1$ and $RO(j_{intra}|j) = 0$ for $j_{intra} > j$. Due to space limitations, we do not present the full derivations of (11).

Having determined the resource occupancy, we now can calculate the state blocking factors as follows:

$$SBF^{r}_{k,t}(j) = \sum_{j_{intra}=0}^{j} \beta^{r}_{k,t}(j_{intra})RO(j_{intra}|j) \tag{12}$$

E. State Probabilities and Call Blocking Probabilities

Having determined the SBFs in the previous subsection, we can now solve the local balance equations of (7) and derive the state probabilities:

$$q(j) = \frac{1}{j}\sum_{k=1}^{K}\sum_{t}\sum_{r}[(U_{k,t} - M^{r}_{k,t}(j) + 1)a_{k,t}(1- SBF^{r}_{k,t}(j - b^{r}_{k,t})b^{r}_{k,t}\delta^{r}_{k,t}(j)q(j - b^{r}_{k,t}))] \tag{13}$$

for $j = 1, ..., C_{total}$ and $q(j) = 0$ for $j < 0$, where $\sum_{j=0}^{C_{total}} = 1$.

The parameters $\delta^{r}_{k,t}(j)$ are used in (13) to impose the approximations of (8) and are defined as follows:

$$\delta^{P}_{k,t}(j) = \begin{cases} 1, & \text{if } j < J^{L}_{k,t} \\ 0, & \text{otherwise} \end{cases}$$
$$\delta^{M}_{k,t}(j) = \begin{cases} 1, & \text{if } J^{L}_{k,t} \leq j < J^{U}_{k,t} \\ 0, & \text{otherwise} \end{cases} \tag{14}$$
$$\delta^{L}_{k,t}(j) = \begin{cases} 1, & \text{if } j \geq J^{U}_{k,t} \\ 0, & \text{otherwise} \end{cases}$$

Finally, we can calculate the call blocking probabilities of both *new* and *handoff* MUs as follows:

$$B_{k,t} = \sum_{j=0}^{J^{L}_{k,t}-1} q(j)SBF^{P}_{k,t}(j) + \sum_{j=J^{L}_{k,t}}^{J^{U}_{k,t}-1} q(j)SBF^{M}_{k,t}(j) + \sum_{j=J^{U}_{k,t}}^{C_{total}} q(j)SBF^{L}_{k,t}(j) \tag{15}$$

These are calculated by adding all the state probabilities multiplied by the corresponding state blocking factors for *peak*, *medium*, and *low* data rates.

V. Evaluation

In order to evaluate the accuracy of our proposed analytical model, we simulate the HetNet scenario using SIMSCRIPT III [22]. User arrivals and departures are simulated according to the description of Section III. When there are not enough available radio resources in the small cell, macro-BS allocates them, as long as there are unused resources in the macro cell or other small cells. If no sufficient resources are available, an arriving MU is blocked. The simulator records such blocking events to produce in the end, call blocking probabilities for each service-class, for both *new* and *handoff* MUs. We also analytically calculate blocking probabilities as described in Section IV. The calculation of blocking probabilities is based on (15). These calculations, performed on commodity hardware, can be produced within a few seconds and also require very little memory, due to the recursive calculation of (13). This enables macro-BS to perform fine-grained RRM and resource allocation to small cells. Recall that this efficiency comes at the cost of approximations (7) and (8). However, by comparing analytical and simulation results, we show that the approximation errors are not significant.

TABLE I. Offered traffic (erl)

	1st service-class		2nd service-class	
	new	***handoff***	***new***	***handoff***
traffic point	$a_{1,N}$	$a_{1,H}$	$a_{2,N}$	$a_{2,H}$
1	0.10	0.02	0.05	0.01
2	0.20	0.04	0.10	0.02
3	0.30	0.06	0.15	0.03
4	0.40	0.08	0.20	0.04
5	0.50	0.10	0.25	0.05
6	0.60	0.12	0.30	0.06
7	0.70	0.14	0.35	0.07
8	0.80	0.16	0.40	0.08

We consider the following experimental setup. The first service-class has $U_{1,N} = 50$ *new* MUs (per small cell) and $U_{1,H} = 10$ *handoff* MUs (coming from neighbouring small cells). It has one rate threshold, $n^{U}_{1,N} = 0.5$ for *new* MUs and two rate thresholds: upper, $n^{U}_{1,H} = 0.5$, and lower, $n^{L}_{1,H} = 0.3$ for *handoff* MUs. The corresponding bit rates (in Kbps) for the first service class are: $R^{P}_{1,N} = 384$, $R^{L}_{1,N} = 144$, $R^{P}_{1,H} = 384$, $R^{M}_{1,H} = 144$, $R^{L}_{1,H} = 64$. The second service-class has $U_{2,N} = 25$ *new* and $U_{2,H} = 5$ *handoff* MUs. It has no rate thresholds for *new* MUs and one rate threshold $n^{U}_{2,H} = 0.5$ for *handoff* MUs. The corresponding bit rates (in Kbps) for the second service class are: $R^{P}_{2,N} = 256$, $R^{P}_{2,H} = 256$, $R^{L}_{2,H} = 122$. The activity factor is selected to be $v_1 = 1$ and $v_2 = 0.67$ for the first and the second service-class, respectively.

We consider a lognormally distributed *inter-cell interference* with coefficient of variation $CV[I_{inter}] = 1$ and with mean $E[I_{inter}] = 10^{-18}$ mW. The power spectral density of the thermal noise is assumed to be $N_0 = -174$ dBm/Hz. In the analytical model for discretization we use $g = 0.001$. The offered traffic load (in erlangs) per MU of the two service-classes is shown in Table I.

Figures 2 and 3 show analytical and simulation blocking probabilities for both *new* and *handoff* calls. The presented simulation results are mean values of 10 runs with 95% confidence interval. The resultant reliability ranges of our measurements are very small and, therefore, we present only mean values. We observe that the analytical results are very close to simulation results. This confirms that the introduced approximations do not significantly impact the accuracy of the proposed analytical model. We also observe that the blocking probabilities of *handoff* MUs are lower than the blocking probabilities of *new* MUs. This is in line with adopted resource

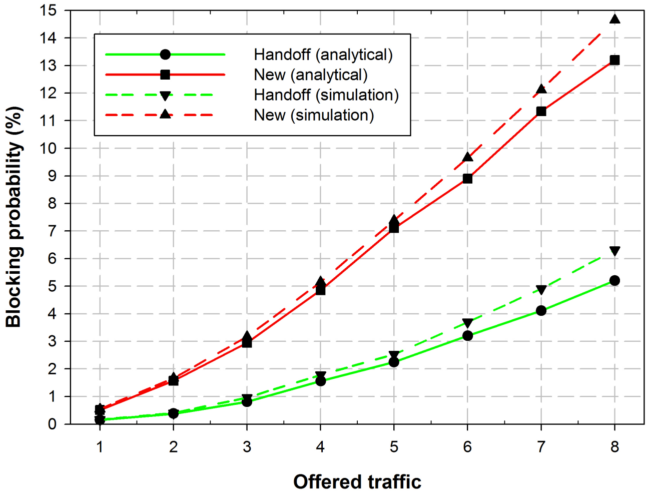

Fig. 2. Call blocking probabilities vs offered traffic-load (1st service-class).

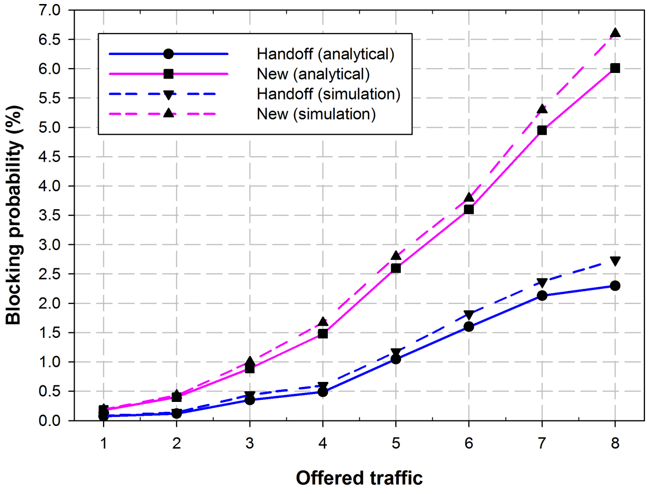

Fig. 3. Call blocking probabilities vs offered traffic-load (2nd service-class).

allocation strategy, which gives higher service priority to *handoff* MUs. We have also considered other experimental scenarios, with different bit rates, rate thresholds, and activity factors, and a larger number of service-classes. In all cases the accuracy of the proposed analytical model is similar to what is shown in Figs. 2 and 3. Due to space limitations, these results are not presented.

VI. Conclusion

In this paper, we have presented a novel analytical framework for software-defined radio access network. We take into account multiple quality-of-service classes, user mobility, and finite number of users in the cell. User arrivals and departures are described by a continuous-time Markov chain and time- and space-efficient recursive formulas are derived. This allows for a software-driven control at the macro base station for efficient radio resource management in a heterogeneous network. Our simulation results show that the accuracy of the proposed approximations is very satisfactory.

References

[1] B. Nunes, M. Mendonca, X. Nguyen, K. Obraczka, T. Turletti, "A survey of software-defined networking: Past, present, and future of programmable networks," IEEE Comm. Surv. & Tut., vol. 16, no. 3, pp. 1-18, 2014.

[2] J. M. Halpern, "Standards collisions around SDN," IEEE Commun. Mag., vol. 52, no. 12, pp. 10-15, 2014.

[3] ONF, "OpenFlow Switch Specification," v. 1.4.0, Oct. 14, 2013.

[4] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, J. Turner, "OpenFlow: enabling innovation in campus networks," ACM SIGCOMM Computer Communication Review, vol. 38, no. 2, pp. 69-74, 2008.

[5] M. Kobayashi, S. Seetharaman, G. Parulkar, G. Appenzeller, J. Little, J. Van Reijendam, P. Weissmann, N. McKeown, "Maturing of OpenFlow and software-defined networking through deployments," Computer Networks, no. 61, pp. 151-175, 2014.

[6] M. Yang, Y. Li, D. Jin, L. Zeng, X. Wu, A. Vasilakos, "Software-defined and virtualized future mobile and wireless networks: A survey," Mobile Networks and Applications, 2014 (in press).

[7] M. Tomovic, M. Pejanovic-Djurisic, I. Radusinovic, "SDN based mobile networks: Concepts and benefits," Wireless Personal Communications, vol. 78, no. 3, pp. 1629-1644, 2014.

[8] P. Demestichas, A. Georgakopoulos, D. Karvounas, K. Tsagkaris, V. Stavroulaki, J. Lu, C. Xiong, J. Yao, "5G on the horizon: Key challenges for the radio-access network," IEEE Vehicular Technology Magazine, vol. 8, no. 3, pp. 47-53, 2013.

[9] P. K. Agyapong, M. Iwamura, D. Staehle, W. Kiess, A. Benjebbour, "Design considerations for a 5G network architecture," IEEE Commun. Mag., vol. 52, no. 11, pp. 65-75, 2014.

[10] A. Gudipati, D. Perry, L. E. Li, S. Katti, "SoftRAN: Software defined radio access network," Proc. 2nd ACM SIGCOMM workshop on Hot Topics in Software Defined Networking, pp. 25-30, 2013.

[11] L. E. Li, Z. M. Mao, J. Rexford, "Toward software-defined cellular networks," Proc. IEEE European Workshop on Software Defined Networking (EWSDN), pp. 7-12, 2012.

[12] X. Jin, L. E. Li, L. Vanbever, J. Rexford, "Softcell: Scalable and flexible cellular core network architecture," Proc. 9th ACM Conf. on Emerging Networking Experiments and Technologies, pp. 163-174, 2013.

[13] M. Bansal, J. Mehlman, S. Katti, P. Levis, "OpenRadio: a programmable wireless dataplane," Proc. 1st ACM Workshop on Hot Topics in Software Defined Networks, pp. 109-114, 2012.

[14] M. Yang, Y. Li, D. Jin, L. Su, S. Ma, L. Zeng, "OpenRAN: a software-defined ran architecture via virtualization," Proc. ACM SIGCOMM, pp. 549-550, 2013.

[15] R. Shrivastava, S. Costanzo, K. Samdanis, D. Xenakis, D. Grace, L. Merakos, "An SDN-based framework for elastic resource sharing in integrated FDD/TDD LTE-A HetNets," Proc. 3rd IEEE International Conference on Cloud Networking (CloudNet), pp. 126-131, 2014.

[16] D. Staehle, A. Mäder, "An analytic approximation of the uplink capacity in a UMTS network with heterogeneous traffic," Proc. ITC-18, Berlin, Germany, Sept. 2003.

[17] I. D. Moscholios, G. A. Kallos, V. G. Vassilakis, M. D. Logothetis, "Congestion probabilities in CDMA-based networks supporting batched Poisson input traffic," Wireless Personal Communications, vol. 79, no. 2, pp. 1163-1186, Nov. 2014.

[18] V. G. Vassilakis, I. D. Moscholios, J. S. Vardakas, M. D. Logothetis, "Handoff modeling in cellular CDMA with finite sources and state-dependent bandwidth requirements," Proc. IEEE CAMAD-2014, Athens, Greece, Dec. 2014.

[19] S. Hanczewski, M. Stasiak, P. Zwierzykowski, "A new model of the soft handover mechanism in the UMTS network," Proc. IEEE/IET 9th CSNDSP, Manchester, U.K., July 2014.

[20] J. Kaufman, "Blocking in a shared resource environment," IEEE Trans. Commun., vol. 29, no. 10, pp. 1474-1481, 1981.

[21] J. W. Roberts, "A service system with heterogeneous user requirements," In: G. Pujolle (Ed.), Performance of Data Communications Systems and Their Applications, North Holland, Amsterdam, pp. 423-431, 1981.

[22] Simscript III, http://www.simscript.com [March 2015].