Article:
Vasilakis, Vasileios, Moscholios, Ioannis and Logothetis, Michael (2018) Quality of Service Differentiation in Heterogeneous CDMA Networks:A Mathematical Modelling Approach. Wireless Networks. pp. 1279-1295. ISSN: 1022-0038

https://doi.org/10.1007/s11276-016-1411-z

Noname manuscript No.
(will be inserted by the editor)

Quality of Service Differentiation in Heterogeneous CDMA Networks: A Mathematical Modelling Approach

Vassilios G. Vassilakis · Ioannis D. Moscholios · Michael D. Logothetis

the date of receipt and acceptance should be inserted later

Abstract Next-generation cellular networks are expected to enable the coexistence of macro and small cells, and to support differentiated quality-of-service (QoS) of mobile applications. Under such conditions in the cell, due to a wide range of supported services and high dependencies on efficient vertical and horizontal handovers, appropriate management of handover traffic is very crucial. Furthermore, new emerging technologies, such as cloud radio access networks (C-RAN) and self-organizing networks (SON), provide good im-

V. G. Vassilakis

School of Computing & Engineering, University of West London, London, United Kingdom

Tel.: +44-7466-863403

E-mail: vasileios.vasilakis@uwl.ac.uk

I. D. Moscholios

Dept. of Informatics & Telecommunications, University of Peloponnese, Tripolis, Greece

E-mail: idm@uop.gr

M. D. Logothetis

WCL, Dept. of Electrical & Computer Engineering, University of Patras, Patras, Greece

E-mail: mlogo@upatras.gr

plementation and deployment opportunities for novel functions and services. We design a multi-threshold teletraffic model for heterogeneous code division multiple access (CDMA) networks that enable QoS differentiation of handover traffic when elastic and adaptive services are present. Facilitated by this model, it is possible to calculate important performance metrics for handover and new calls, such as call blocking probabilities, throughput, and radio resource utilization. This can be achieved by modelling the cellular CDMA system as a continuous-time Markov chain. After that, the determination of state probabilities in the cellular system can be performed via a recursive and efficient formula. We present the applicability framework for our proposed approach, that takes into account advances in C-RAN and SON technologies. We also evaluate the accuracy of our model using simulations and find it very satisfactory. Furthermore, experiments on commodity hardware show algorithm running times in the order of few hundreds of milliseconds, which makes it highly applicable for accurate cellular network dimensioning and radio resource management.

Keywords quality of service · handover · cdma · cloud radio access network

1 Introduction

Future generation cellular networks are expected to support services that require varying degrees of quality-of-service (QoS), but at the same time enabling a wide range of real-time and data applications [1]. Furthermore, the appearance of highly promising technologies, such as the cloud radio access networks (C-RAN) and self-organizing networks (SON), can provide higher efficiency and greater scalability through the use of software-defined networking (SDN), network function virtualisation (NFV), and data center processing capabilities [2, 3]. On the other hand, the design and deployment of future cellular networks is further complicated due to the heterogeneous nature of various coexisting communication

technologies and protocols, especially when trying to satisfy tight energy consumption constraints [4–6]. In such environments, providing acceptable QoS under the conditions of high user mobility is not a straightforward task. Call handovers from one small cell to another or between macro and small cells can severely degrade the QoS of existing users if no proper call admission control (CAC) is performed [7, 8].

Furthermore, future fifth generation (5G) cellular networks [9,10] are expected to simultaneously utilize multiple and different channel access methods, such as code division multiple access (CDMA) [11,12] and frequency division multiple access (FDMA), in a multi-tier fashion [13]. Recent advances in multicarrier CDMA (MC-CDMA) systems indicate good robustness against multipath propagation channels and high performance in terms of data transmission rates [11]. On the other hand, the radio resource planning and management in CDMA-based systems is a difficult task due to the multiple access interference (MAI), which is caused by both inter- and intra-cell mobile traffic [14,15]. This is especially true in a heterogeneous wireless environment with dense small cell deployments and an increasing number of end devices with dual connectivity [16].

In this work, we propose a teletraffic model that: a) takes into account the aforementioned requirements of next-generation cellular networks; b) enables derivation of important performance metrics, both network-centric and user-centric; c) is easily implementable in real world, with short running times; and d) requires minimal storage and computing resources. Especially in the case of emerging wireless multimedia systems and video streaming over wireless networks, fast estimations of video quality in real time is of major importance [17,18].

To model a wide range of current and future applications, we distinguish the three following generic classes: *fixed traffic*, *elastic traffic*, and *adaptive traffic* [19]. Fixed traffic refers to calls that demand a fixed amount of resources and service time (also referred to

as holding time). This resource demand can not be reduced and the requesting user either gets it all or nothing. A typical example is a fixed-rate real-time video streaming service. Elastic and adaptive traffic refers to calls that have more than one contingency resource requirements. The demanded amount of resources depends on the resource availability in the system, but also on the varying radio conditions in the cell. The holding time of an elastic call is inversely proportional to the amount of resources that the particular call has been allocated. A typical example is the file transfer service, where the transfer completion time depends on the available data rate. Finally, the holding time of an adaptive call remains fixed, irrespective of the amount of allocated resources. A typical example is the adaptive real-time video streaming service, where the video quality can be reduced in response to the data rate reduction, but the completion/holding time remains fixed.

Furthermore, for each generic traffic type, we incorporate different QoS levels. Each QoS level dynamically captures service requirements such as bandwidth, signal-to-noise ratio (SNR), and service time. Our proposed mathematical model takes also into account call handovers, dynamic CAC, and MAI, and results in analytical expressions for system state probabilities and other important system parameters. Next, we develop a time- and space-efficient algorithm for the calculation of state probabilities. Experiments on commodity hardware show algorithm running times in the order of hundreds of milliseconds, which makes it applicable in highly dynamic heterogeneous environments and even in cases of fast moving users. This time- and space-efficiency has been achieved by introducing a number of approximations. However, as our evaluation shows, the impact of the approximations is negligible and the accuracy of produced results is very good. Furthermore, by configuring some parameters, it is possible to trade-off efficiency for accuracy. Having determined state probabilities, call blocking probabilities (CBP), radio resource utilization (RRU), and service throughput, can be easily calculated.

The remainder of the paper is organised in the following way. Section 2 presents the literature review. It also gives an overview of our proposed model and states the contribution of this work. In Section 3, we review two existing models, proposed for wired, connection-oriented networks. In Section 4, we propose our new mathematical model for heterogeneous CDMA-based networks. Subsection 4.1 describes our Markov chain modelling approach, provides a simple example that illustrates the basic concepts, and specifies the adopted approximations that enable the derivation of recursive and efficient formulas. A practical algorithm for determining the system state probabilities is proposed in Subsection 4.2. Subsection 4.3 calculates important performance metrics, such as CBP, RRU, and service throughput. In Section 5, we present the applicability framework for our proposed model. In particular, the applicability in C-RAN using SON, SDN, and NFV technologies is discussed. Section 6, investigates three case studies aiming at evaluating our proposed approach. In particular, the analytical results that are derived via the approximate model are compared with simulation results and the accuracy is very good. For completeness, we also compare our model with other proposed models in the literature. Section 7 concludes the paper and gives future work directions.

2 Literature Review and Model Overview

Below we discuss the relevant teletraffic models proposed in the literature. We also provide a descriptive overview of our proposed approach, leaving formal definitions to the sections that follow. Finally, we state the contribution of this work, compared to other published works.

The well-known Erlang multi-rate loss model (EMLM) has been extensively used for performance modelling and analysis of multi-rate loss systems under the presence of Pois-

son traffic [20], [21] (the EMLM is presented in Subsection 3.1). The EMLM results in a recurrent formula, known as Kaufman-Roberts (K-R) recursion, that allows the CBP calculation when the complete resource sharing (CRS) policy [22] is used. Following its first inception several decades ago, numerous modifications of the EMLM have been proposed for both wireline and cellular networks. These works are discussed below.

In [23], calls that are blocked due to unavailability of resources, may retry multiple times, requesting for less resources. In the model of [24], arrived calls may have multiple contingency resource demands and the appropriate amount of resources that is allocated to each call depends on the total amount of occupied resources in the system and on a set of predefined thresholds, common to all services. The connection-dependent threshold model (CDTM) proposed in [25] can be seen as a generalization of the aforementioned retry and threshold models. In particular, the CDTM allows the parameterization of individual thresholds that can be defined on a per service basis (the CDTM is presented in Subsection 3.2). The aforementioned models are applicable to connection-oriented networks and are not suitable for the modelling of cellular networks with sophisticated radio resource management (RRM) schemes. In [26], the CDTM is extended to allow call bandwidth compression/expansion. Later, the model of [26] was investigated under the bandwidth reservation (BR) policy in [27], where some bandwidth is reserved for certain service-classes in order to achieve CBP equalization.

In this work we concentrate on the uplink of heterogeneous CDMA systems. The cell is modelled as a multirate loss system with a given amount of radio resources. The amount of resources is not fixed and depends on the network conditions, the activity of accepted calls (i.e., whether the call is on a transmission or a silent mode), and other factors. Hence, we talk about the *soft capacity*. The demanded amount of resources of a particular service can be derived from a number of service parameters such as the SNR, data transmission

rate, and call activity [28]. Arriving calls are accepted into the system according to the CRS policy. That is, a call is accepted if and only if there is sufficient amount of resources in the system. In particular, a CAC mechanism performs an estimation of the required resources by measuring the increase of radio interference (both intra- and inter-cell) as a result of call's acceptance into the system. CAC and other RRM functions can be implemented either as distributed or as centralized SON functions, as discussed in Section 5. Due to MAI of CDMA-based systems, if a call is accepted, the SNR of other calls in the system is reduced. Hence, if according to estimations, the SNR of other calls is going to drop below a tolerable level (dependent on the QoS of each call), the arriving call must not be accepted. In other words, a call must not be accepted if it will cause an increase of the interference above a certain level. Taking into account the aforementioned peculiarities of cellular CDMA modelling, the EMLM has been enhanced in [29] (referred to as the W-EMLM) considering only the fixed traffic generated by new calls. Later, the model of [29] was extended to take into account handover traffic as well [30].

The above mentioned models result in recurrent formulas. Among other notable works, [31] proposes a model for elastic traffic with fixed transmission rate slow down factors of in-service calls. This work has been extended in [32] in order to have state-dependent rather than fixed slow down factors. In particular, in-service elastic calls may change their occupied resources, but having different resource requirements upon arrival is not allowed. In [33], an analytical model for dynamic streaming systems is proposed. The wireless channel is modelled as a continuous time Markov process and a set of differential equations is constructed to characterize the buffer starvation probability. The proposed model enables determination of QoS metrics for dynamic and adaptive streaming services. For the downlink of CDMA systems, a number of efficient analytical models have been already proposed [34–37]. There-

fore, in this work, we focus on the uplink direction. Our aim, among others, is to explicitly incorporate the handover traffic into the analytical model.

The proposed model is called Wireless Handover Connection-Dependent Threshold Model (WH-CDTM). As our analysis in Section 4 shows, the steady state probabilities in the WH-CDTM do not have a product form solution (PFS) [38, 39]. Therefore, we introduce appropriate approximations and obtain an equivalent reversible Markov chain, for which a PFS exists. We also derive an approximate expression for the calculation of state probabilities that enables the calculation of CBP for both new and handover traffic.

Compared to our previous paper [30], the contribution of this work is as follows: a) we enhance the mathematical model with elastic and adaptive traffic types; b) we introduce different QoS levels for each traffic type; c) we propose a practical and easy implementable algorithm for the determination of system state probabilities; d) in addition to CBP, we derive expressions for other important performance metrics, such as RRU and service throughput; e) having benefited by more available space, we present a more detailed description of the mathematical model, providing more diagrams and examples, and much more detailed calculations; f) we provide the applicability framework for our model using C-RAN and SON technologies; g) the evaluation section has been substantially extended; h) we provide both analytical and simulative comparison of our model with other existing models in the literature.

3 Background

3.1 Overview of the Erlang Multi-rate Loss Model (EMLM)

Consider a system that has R discrete resources. Calls arrive to the system according to a Poisson process. Assume that each call belongs to one of S independent services. The arrival

rate of service $s \in S$ calls is denoted by ar_s. A call of a service $s \in S$ demands r_s resources from the system. If the number of demanded resources is available in the system, the call occupies them for a generally distributed holding time with mean t_s. On the other hand, if the demanded resources are not available, the call is blocked and leaves the system without further affecting it. This effectively implements the CRS policy, mentioned in Section 2. The total number of occupied resources in the system is denoted by r and can be determined as follows:

$$r = \sum_{s=1}^{S} c_s r_s \tag{1}$$

where c_s is the number of service s calls in the system at any given moment. In the following, r is also referred to as the *system state*.

From the above it is clear that r takes values between 0 and R (inclusive). When new calls are accepted into the system, r increases, whereas when calls depart from the system and release the previously occupied resources, r decreases. The probability that the system state is $r \in [0, ..., R]$ is denoted by $P(r)$.

It has been proven that the following local balance exists in the EMLM [20]:

$$ar_s P(r - r_s) = \frac{ac_s(r)}{t_s} P(r) \tag{2}$$

where $ac_s(r)$ is the average number of service s calls in state r.

The above equation essentially says that when the system is in equilibrium, transitions between adjacent states ($r - r_s$ and r in this case) occur at equal rates. Eq. (2) is often re-written in the following way

$$tl_s P(r - r_s) = ac_s(r) P(r) \tag{3}$$

where tl_s is the offered traffic-load of service s, defined as $tl_s = ar_s t_s$.

In many cases it is desirable to know the percentage of the total resources that is occupied by a particular service. This is captured by the *resource share* of a service s, defined as

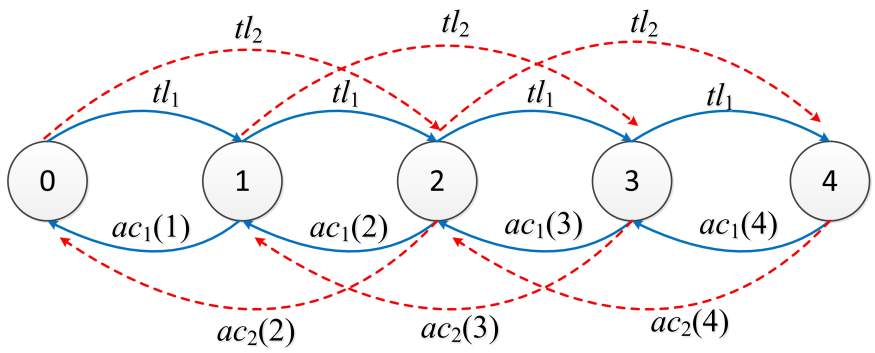

Fig. 1: State transition diagram in the EMLM.

$rs_s = \frac{c_s r_s}{r}$ (instantaneous). The average resource share, $ars_s(r)$, of service s in state r can be calculated with the aid of (3) as follows:

$$ars_s(r) = \frac{ac_s(r)r_s}{r} = \frac{tl_s P(r-r_s)r_s}{P(r)r} \tag{4}$$

The aforementioned concepts can be easier understood using a simple example. Consider a system with $R = 4$ resources and $S = 2$ services. Resource demands per call are $r_1 = 1$ and $r_2 = 2$ for the 1st and 2nd service, respectively. Figure 1 shows the state transition diagram for this system. We observe that $r = 4$ is a *blocking state* for both services, since no available resources are left in this state. We also observe that $r = 3$ is a blocking state for the 2nd service, but not for the 1st. This is due to the fact that 1 resource is available (and 3 resources are occupied) in this state.

In the general case, a new service s call (that demands r_s resources) is accepted in the system if and only if $r + r_s \leq R$. Hence, $r = R - r_s + 1, ..., R - 1, R$ are blocking states for service s, whereas the remaining states are non-blocking states. The CBP of a particular service s can be calculated by summing the state probabilities of all blocking states:

$$CBP_s = \sum_{j=R-r_s+1}^{R} P(r) \tag{5}$$

The state probabilities, $P(r)$ ($r = 1, ..., R$), that are required in the above equations, can be recursively determined by the well-known K-R formula [20], [21]:

$$rP(r) = \sum_{s=1}^{S} tl_s r_s P(r - r_s) \tag{6}$$

where $P(r) = 0$ for $r < 0$ and for $r > R$, using the normalization $\sum_{r=0}^{R} P(r) = 1$.

3.2 Overview of the Connection-Dependent Threshold Model (CDTM)

The EMLM, described in Section 3.1, supports only fixed traffic. That is, each call demands a fixed number of resources. Below we describe the CDTM, which is an extension of the EMLM that supports elastic traffic [25].

Consider a system with R discrete resources and S independent services. Calls arrive to the system according to a Poisson process with the arrival rate of ar_s for service $s \in S$. A call of service s has $D(s)$ contingency resource demands, denoted as $r_{s,d}$, $d \in [1, ..., D(s)]$. By convention $r_{s,d}$ is a strictly increasing function with respect to d for every s. The choice of a particular demand depends on the system state r (defined as in the EMLM) and the set of resource thresholds of the particular service. The thresholds of service s are denoted as $TH_{s,d}$, $d \in [1, ..., D(s)]$ and are used as follows. If at the time of a call arrival $r \leq TH_{s,1}$, then the call demands $r_{s,1}$ resources. If $TH_{s,d-1} < r \leq TH_{s,d}$ ($d \in [2, ..., D(s)]$), then the call demands $r_{s,d}$ resources. Finally, if $r > TH_{D(s)} = R$, then the call is blocked and lost. Note that by convention R represents the highest threshold $TH_{D(s)}$.

As mentioned in Section 1, one of the main characteristics of elastic traffic is that the holding time of a call is inversely proportional to the amount of resources the call is given. Hence, each of the aforementioned contingency resource demands has a corresponding holding time, denoted as $t_{s,d}$, $d \in [1, ..., D(s)]$. Note that $t_{s,d}$ is a strictly decreasing function with respect to d for every s.

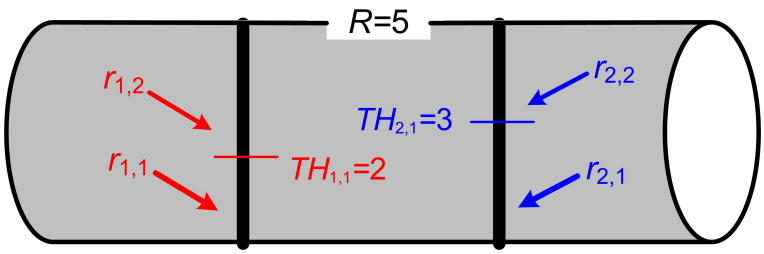

Fig. 2: Principles of the CDTM.

The aforementioned concepts are explained below with the aid of a simple example. Consider a system with $R = 5$ resources and $S = 2$ services. Each service has two contingency demands. That is $D(1) = D(2) = 2$. Consequently, (apart from the trivial threshold $TH_{s,2} = R = 5$) there is one threshold per service. Let us assume that $TH_{1,1} = 2$ and $TH_{2,1} = 3$. Let us also assume the following contingency demands: $r_{1,1} = 3$, $r_{1,2} = 1$, $r_{2,1} = 3$, and $r_{2,2} = 2$. This basic CDTM concept has also been illustrated in Fig. 2 and in Fig. 3 we show the state transition diagram for the 1st service.

By $ac_{1,1}(r)$ and $ac_{1,2}(r)$, we denote the mean number calls of the 1st service in state r, with resource demands $r_{1,1}$ and $r_{1,2}$, respectively. We observe that the 1st service has one blocking state, $r = 5$. Similarly, the 2nd service (although not shown in Fig. 3) has two blocking states, $r = 4$ and $r = 5$. We also observe that the corresponding Markov chain is irreversible. This is obvious from the fact that while there are some transitions from higher states to lower, there are no corresponding transitions from a lower state to higher in all cases. This means that the CDTM system does not have a PFS and we will have to resort to approximations in order to derive an efficient and recurrent formula for state probabilities, $P(r)$.

To approximate reversibility we assume the following:

1. The number of calls with resource demand $r_{s,1}$ is negligible in states $r > TH_{s,1}$.

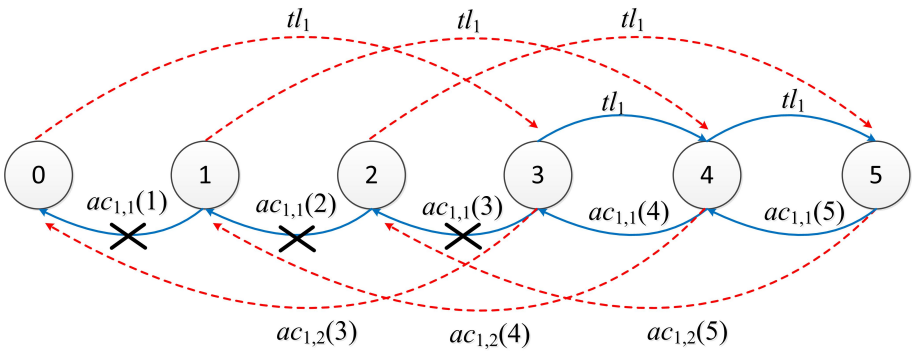

Fig. 3: State transition diagram in the CDTM (1st service).

2. The number of calls with $r_{s,d}$ for $1 < d < D(s)$ is negligible in states $r \leq TH_{s,d-1}$ and $r > TH_{s,d}$.
3. The number of calls with $r_{s,D(s)}$ is negligible in states $r < TH_{s,D(s)-1}$.

The purpose of the above approximations is to remove the "redundant" transitions so that the remaining Markov chain becomes reversible.

In the system of Fig. 2 after the introduced approximations, the number of calls with resource demand $r_{2,1}$ in states $r = 1$, 2, and 3 is considered negligible. Hence, as indicated with X's, the transitions $3 \rightarrow 2$, $2 \rightarrow 1$, and $1 \rightarrow 0$ are removed. In a similar way, the transitions $5 \rightarrow 3$ for calls with $r_{2,1}$ are removed as well.

In addition to the above approximations, we also assume that local balance (eq. (7)) exists between adjacent system states (Fig. 4). This essentially means that the transition rates from lower states to higher are equal to the corresponding transition rates from higher states to lower.

$$ar_s\delta_{s,d}(r - r_{s,d})P(r - r_{s,d}) = t_{s,d}ac_{s,d}(r)\delta_{r,d}(r)P(r) \tag{7}$$

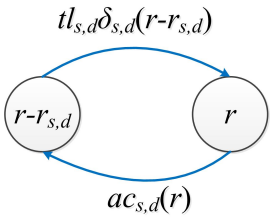

Fig. 4: Local balance in the CDTM.

where the parameters $\delta_{s,d}(r)$ are given by:

$$\delta_{s,1}(r) = \begin{cases} 1, & \text{if } r \leq TH_{s,1} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

$$\delta_{s,d}(r) = \begin{cases} 1, & \text{if } TH_{s,d-1} < r \leq TH_{s,d} \\ 0, & \text{otherwise} \end{cases} \quad (d = 2, ..., D(s) - 1) \tag{9}$$

$$\delta_{s,D(s)}(r) = \begin{cases} 1, & \text{if } r > TH_{s,D(s)-1} \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

Consequently, dividing both sides of (7) by $rP(r)$ and substituting $ar_s t_{s,d} = tl_{s,d}$, the resource share of service s with demand $r_{s,d}$ in state r can be calculated by:

$$ars_{s,d}(r) = \frac{tl_{s,d}\delta_{s,d}(r)P(r - r_{s,d})}{rP(r)} \tag{11}$$

Following the aforementioned approximations and calculations, the state probabilities can be calculated by the recursion below [25]:

$$rP(r) = \sum_{s=1}^{S} \sum_{d=1}^{D(s)} tl_{s,d} r_{s,d} \delta_{s,d}(r) P(r - r_{s,d}) \tag{12}$$

for $r = 1, ..., R$ and $P(r) = 0$ for $r < 0$, using the normalization $\sum_{r=0}^{R} P(r) = 1$.

A new call of service s can be accepted to the system only if its smallest resource demand, $r_{s,D(s)}$, is available. This means that the blocking states are $r = R - r_{s,D(s)} + 1, ..., R$. Hence, the CBP of a given service s can be calculated by:

$$CBP_s = \sum_{r=R-r_{s,D(s)}+1}^{R} P(r) \tag{13}$$

4 The proposed Wireless Handover Connection-Dependent Threshold Model (WH-CDTM)

4.1 Markov chain modelling

Our aim is to model a CDMA system accommodating S independent services.

The system offers $D(s)$ different QoS levels to each service $s \in S$. In the following, a service s call of QoS level d, $d \in [1, ..., D(s)]$, is referred to as a service s, d call. We also distinguish between *new* and *handover* calls.

To characterize a *new* service s, d call, we use:

- $DR^N_{s,d}$: data rate.
- $t^N_{s,d}$: holding time.
- $SNR^N_{s,d}$: signal-to-noise ratio.

Similarly, to characterize a *handover* service s, d call, we use:

- $DR^H_{s,d}$: data rate.
- $t^H_{s,d}$: holding time.
- $SNR^H_{s,d}$: signal-to-noise ratio.

The aforementioned parameters can be used to define the *load factor* of a service s, d call, as follows:

$$LF^t_{s,d} = \frac{SNR^t_{s,d} DR^t_{s,d}}{W + SNR^t_{s,d} DR^t_{s,d}} \tag{14}$$

where W is the CDMA chip rate and $t \in \{N, H\}$.

In order to determine the resource demand of a service s, d call, we discretize its load factor:

$$r_{s,d}^{t} = \lfloor \frac{LF_{s,d}}{g} \rfloor \tag{15}$$

where g is the discretization unit. The selection of g can be subject to optimization. A small value of g will produce larger discretization error, while a larger value will produce a larger state space. In our experiments, in Section 6, we use $g = 0.001$.

We classify the services as follows:

- *fixed traffic*: when $D(s) = 1$, meaning that a single QoS level is supported.
- *elastic traffic*: when $D(s) > 1$ and the holding time depends on the QoS level.
- *adaptive traffic*: when $D(s) > 1$ and the holding time is fixed and is independent of the QoS level.

Calls of each service s arrive to the system according to a Poisson process with mean ar_s^t. The traffic-load of service s, d is defined as $tl_{s,d}^t = ar_s^t t_{s,d}^t$.

One of the main characteristics of CDMA-based communication is that all calls utilize the same frequency band and their signals are distinguished via different codes. Since in practice the codes are non-orthogonal, signals generated by each call are perceived as noise by other in-service calls.

In general, the noise/interference in CDMA systems can be classified into:

- I_{intra}: The intra-cell interference that is generated by the calls of the same cell.
- I_{inter}: The inter-cell interference that is generated by the calls of adjacent cells.
- T_{noise}: The thermal noise generated at the receiver.

A typical way of implementing CAC is by estimating the *noise rise* via the following formula [28]:

$$NR = \frac{I_{intra} + I_{inter} + T_{noise}}{T_{noise}} \tag{16}$$

A pre-defined upper bound for the noise rise, NR_{max}, is used for call blocking/admission decisions.

Another and more convenient quantity for CAC modelling is the *cell load*, defined by:

$$CL = \frac{I_{intra} + I_{inter}}{I_{intra} + I_{inter} + T_{noise}} \tag{17}$$

Note that while NR can take arbitrarily high values, CL has a theoretical maximum value of 1. In practice, however, a typical value is $CL_{max} = 0.8$ [29].

By manipulating (16) and (17) we can express CL in terms of NR:

$$CL = \frac{NR - 1}{NR} \tag{18}$$

The cell load consists of the *intra-cell load*, CL_{intra}, that is generated within the cell, and the *inter-cell load*, CL_{inter}, that is generated in the adjacent cells.

The calculation of CL_{intra} is straightforward and is based on the load factors of accepted calls:

$$CL_{intra} = \sum_{t \in \{N,H\}} \sum_{s=1}^{S} \sum_{d=1}^{D(s)} ac_{s,d}^{t} LF_{s,d}^{t} \tag{19}$$

On the other hand, CL_{inter} can not be easily determined because load information from adjacent cells is required. For this reason and similarly to other works (e.g., [30]) we model CL_{inter} as a log-normal random variable.

Having defined CL, the CAC mechanism can be based on the following conditions:

$$CL + LF_{s,d}^{N} \leq CL_{max}^{N} \tag{20}$$

$$CL + LF_{s,d}^{H} \leq CL_{max}^{H} \tag{21}$$

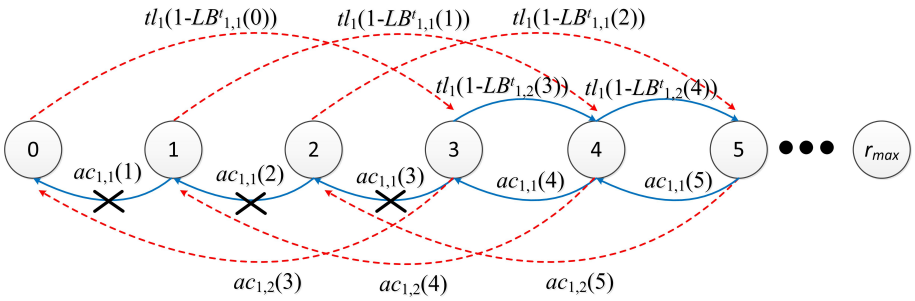

Fig. 5: State Transition Diagram in the WH-CDTM.

where CL^N_{max} and CL^H_{max} are the CAC thresholds for *new* and *handoff* calls, respectively.

Below we present an example to explain the WH-CDTM. Consider the simple example of Section 3.2 but with the following difference. Each state r is a *conditionally blocking* state. In particular, the probability that a service s, d call is blocked in state r is referred to as the *local blocking factor* (LBF) and denoted as $LB^t_{s,d}(r)$. Note that, the CDTM can be seen as a special case of the WH-CDTM with LBFs of various states being either 0 or 1.

Figure 5 depicts the state transition diagram (STD) for the 1st service. It can be observed that transitions to higher states occur at a reduced rate (due to the factor $1 - LB^t_{s,d}(r)$) compared to the the example of Fig. 3. The highest reachable state, r_{max}, is the one where $LB^t_{s,d}(r) \rightarrow 1$. This enables us to model the *soft capacity* feature of CDMA systems.

We also observe that the corresponding Markov chain is irreversible. This is obvious from the fact that while there are some transitions from higher states to lower, there are no corresponding transitions from lower states to higher. This means that the WH-CDTM system does not have a PFS and we will have to resort to approximations in order to derive an efficient and recurrent formula for state probabilities, $P(r)$.

To approximate reversibility we assume the following:

1. The number of calls with resource demand $r_{s,1}$ is negligible in states $r > TH_{s,1}$.

2. The number of calls with $r_{s,d}$ for $1 < d < D(s)$ is negligible in states $r \leq TH_{s,d-1}$ and $r > TH_{s,d}$.
3. The number of calls with $r_{s,D(s)}$ is negligible in states $r < TH_{s,D(s)-1}$.

The purpose of the above approximations is to remove the "redundant" transitions so that the remaining Markov chain becomes reversible.

In the system of Fig. 5, after the introduced approximations, the number of calls with resource demand $r_{2,1}$ in states $r = 1$, 2, and 3 is considered negligible. Hence, as indicated with X's, the transitions $3 \rightarrow 2$, $2 \rightarrow 1$, and $1 \rightarrow 0$ are removed. In a similar way, the transitions $5 \rightarrow 3$ for calls with $r_{2,0}$ are removed as well.

Having defined the necessary approximations and before the calculation of state probabilities, we need to introduce the notion of call activity and define another important metric named *resource occupancy*.

In CDMA systems, a call during its lifetime can be modeled as a series of *active* and *passive* periods. During the active periods, the call transmits data and occupies radio resources. On the other hand, during the passive periods, the call does not transmit and does not occupy any resources. The probability that a call of service s is active is called *activity factor* and denoted as a_s. The system state r refers to the total number of occupied resources in the system assuming that all users are active. Hence r is essentially an upper bound for the actual number of occupied resources denoted by c. In particular, $0 \leq c \leq r$ with $c = 0$ if every call is passive and $c = r$ if every call is active at a given moment.

The probability that c resources are occupied in state r is named *resource occupancy* and can be determined via the following recursion:

$$RO(c|r) = \sum_{s=1}^{S} \sum_{d=1}^{D(s)} \sum_{t \in \{N,H\}} ars_{s,d}^{t}(r)[a_s RO(c - r_{s,d}^{t} | r - r_{s,d}^{t}) + (1 - a_s) RO(c | r - r_{s,d}^{t})] \tag{22}$$

for $r = 1, ..., r_{max}$ with $RO(0|0) = 1$ and $RO(c|r) = 0$ for $c > r$.

The derivation of (22) is explained as follows. To reach the state $(c|r)$ there are two potential previous states: $(c - r^t_{s,d}|r - r^t_{s,d})$ and $(c|r - r^t_{s,d})$. If the system was previously in state $(c - r^t_{s,d}|r - r^t_{s,d})$ then after the arrival of an active call (which happens with probability a_s), both c and r are increased by $r^t_{s,d}$. If the system was previously in state $(c - |r - r^t_{s,d})$ then after the arrival of a passive call (which happens with probability $1 - a_s$), only r is increased by $r^t_{s,d}$. Hence, in both cases the system will reach state $(c|r)$.

Having determined the resource occupancy of every state, we can now calculate the LBFs as follows:

$$LB^t_{s,d}(r) = \sum_{c=0}^{r} LBP^t_{s,d}(c) RO(c|r) \tag{23}$$

where $LBP^t_{s,d}(c)$ is the *local blocking probability* and can be determined via (24) and (25), below [30].

$$LBP^t_{s,d}(c) = \lfloor \frac{1 - CDF_{CL}(x)}{g} \rfloor \tag{24}$$

with $CDF_{CL}(x)$ the cumulative distribution function of CL given by:

$$CDF_{CL}(x) = \big(1 + erf(\frac{\ln x - \mathrm{E}[CL]}{\mathrm{VAR}[CL]\sqrt{2}})\big)/2 \tag{25}$$

where $\mathrm{E}[CL]$ and $\mathrm{VAR}[CL]$ are the expected value and the variance of CL, respectively, and $erf(\cdot)$ is the well-known *error function*.

The resource share in the WH-CDTM can be calculated similarly to (11) by incorporating LBFs:

$$ars^t_{s,d}(r) = \frac{ac^t_{s,d}(r) r^t_{s,d}}{r} \tag{26}$$

The parameters $ac^t_{s,d}(j)$ of (26), are obtained from:

$$ac^t_{s,d}(r) = \frac{tl^t_{s,d} P(r - r^t_{s,d})(1 - LB^t_{s,d}(r - r^t_{s,d}))}{P(j)} \tag{27}$$

Finally, the probability of each state can be calculated as in (12) but with the incorporation of LBFs:

$$P(r) = \frac{1}{r} \sum_{s=1}^{S} \sum_{d=1}^{D(s)} \sum_{t \in \{N,H\}} \left[(tl_{s,d}^{t}(1 - LB_{s,d}^{t}(r - r_{s,d}^{t}) r_{s,d}^{t} \delta_{s,d}(r) P(j - r_{s,d}^{t})) \right] \quad (28)$$

for $r = 1, ..., r_{max}$ and $P(r) = 0$ for $r < 0$, with $\lim_{r_{max} \to \infty} \sum_{r=0}^{r_{max}} = 1$. As it is shown below in Subsection 4.2, j_{max} is the state in which the local blockings $LB_{k,l}^{t}(j)$ are practically equal to 1.

4.2 Recursive Algorithm for the Calculation of State Probabilities

Below we present our proposed algorithm for the calculation of state probabilities. The algorithm is based on the analysis presented in the previous subsection.

Input

1: $S, D(s), TH_{s,d}, tl_{s,d}^{t}, CL_{max}^{t}$

Precalculation

2: determine each $r_{s,d}^{t}$ from (14) and (15)

Initialization

3: $\hat{P}(0) \leftarrow 1$

4: $ac_{s,d}^{t}(0) \leftarrow 0$

5: $ars_{s,d}^{t}(0) \leftarrow 0$

6: $RO(0|0) \leftarrow 1$

7: $LB_{s,d}^{t}(0) \leftarrow 0$

8: $r \leftarrow 0$

9: $\epsilon \leftarrow 10^{-4}$

10: **While** $|1 - LB^t_{s,d}(r)| < \epsilon$ **do**

11: $r \leftarrow r + 1$

12: determine $\delta^t_{s,d}(r)$ from (8) and (9)

13: determine $\hat{P}(r)$ from (28); **print** $\hat{P}(r)$

14: determine $ac^t_{s,d}(r)$ from (27)

15: determine $ars^t_{s,d}(r)$ from (26)

16: **For** $c \in [1, ..., r]$ **do**

17: determine $RO(c|r)$ from (22)

18: determine $LBP^t_{s,d}(c)$ from (24) and (25)

19: **End for**

20: determine $LB^t_{s,d}(r)$ from (23)

21: **End while**

22: $r_{max} \leftarrow j$

The algorithm calculates the so-called un-normalized state probabilities, denoted by $\hat{P}(r)$. It assigns an arbitrary (un-normalized) probability, $\hat{P}(0) = 1$, to state $r = 0$ and, subsequently calculates all the probabilities in the while loop (lines 10-21). We observe that the algorithm runs until $LB^t_{s,d}(r) \approx 1$. Which essentially means that higher states are unreachable due to local blockings having a blocking probability of almost 1. The algorithm's running time and its accuracy depends on the selected parameter ϵ. The smaller is ϵ the better is the accuracy (i.e., $LB^t_{s,d}(r_{max})$ is closer to 1). The bigger is ϵ the shorter is the algorithm's running time. Our experiments show that for practical purposes $\epsilon = 10^{-4}$ is a good choice.

Having determined $\hat{P}(j)$'s, the state probabilities are calculated as follows:

$$P(r) = \frac{\hat{P}(r)}{\sum_{r=1}^{r_{max}} \hat{P}(r)}, \text{ for } r = 1, ...r_{max} \tag{29}$$

Below, we briefly discuss the algorithm's steps. It starts by reading input, and precalculating and initializing various parameters (lines 1-9). Then it enters the while loop, where in each iteration the state j is increased by 1. The determination of parameters δ (line 12) is straightforward, as it is based on known quantities. Next, the calculation of $\hat{P}(r)$ (line 13) is based on state probabilities and LBFs of previous states, which have been already calculated in previous steps. The calculation of $ac^t_{s,d}(r)$ requires knowledge of current and previous sate probabilities, which have been calculated before, and of local blockings from previous states, which are also known at this step. The average resource share, $ars^t_{s,d}(r)$, is determined using $ac^t_{s,d}(r)$ from the previous step. Similarly, recursive calculations of $RO(c|r)$ (lines 16-19) require $ars^t_{s,d}(r)$, calculated in previous step and $RO(c|r)$'s of previous states. Finally, the local blockings, $LB^t_{s,d}(r)$, can also be determined (line 20), since all other required parameters of current and previous states are already known.

As it can be observed, the computational complexity, in terms of required mathematical operations, of the proposed algorithm is very low. Furthermore, our experiments on commodity hardware show algorithm running times in the order of few hundreds of milliseconds, which makes it highly applicable for cell dimensioning and dynamic radio resource allocation (RRA), even under challenging conditions.

4.3 Performance Metrics

In this subsection, we derive analytical expressions for a number of important performance metrics. In particular, we determine the CBP, RRU, and service throughput. All of them use as a basis the state probabilities calculated in Subsection 4.2.

To determine the CBP of service s we add all the state probabilities multiplied by the corresponding LBFs:

$$CBP_s^t = \sum_{r=1}^{r_{max}} \sum_{d=1}^{D(s)} P(r)\delta_{s,d}(r)LB_{s,d}^t(r) \tag{30}$$

Recall, that the system state r essentially corresponds to the amount of occupied radio resources in the cell at a given time. Having previously determined the state probabilities, we can now determine the average RRU, U, which essentially corresponds to the average system state r and is given by:

$$U = \sum_{r=1}^{r_{max}} rP(r) \tag{31}$$

Finally, the throughput, T_s^t, of service s calls is determined as follows:

$$T_s^t = \frac{\sum_{r=1}^{r_{max}} \sum_{d=1}^{D(s)} r_{s,d}^t ac_{s,d}^t(r)\delta_{s,d}(r)P(r)}{\sum_{r=1}^{r_{max}} \sum_{d=1}^{D(s)} ac_{s,d}^t(r)\delta_{s,d}(r)P(r)} \tag{32}$$

The numerator represents the average resource consumption per call of a given service-class. It takes into account the mean number of calls of a particular service and different QoS levels. Hence, the numerator is calculated by adding, for all states, the resource requirements $r_{s,d}^t$ multiplied with the average number of calls, $ac_{s,d}^t(r)$, per state r. The denominator represents the average number of calls of a particular service across all system states. Hence, T_s^t, represents the average amount of radio resources occupied by a call.

5 Applicability Framework

In this section we present the applicability framework for our proposed model. Initially, we introduce our considered C-RAN architecture that has been enhanced with the concepts of SDN and NFV. Next, we briefly introduce the SON technology. Finally, we describe how our approach could be applied to enable RRM by utilizing the hybrid SON technology in C-RAN.

5.1 The Considered C-RAN Architecture

Our considered network architecture is presented in Fig. 6. Three main parts are distinguished: a pool of remote radio heads (RRHs), a pool of baseband units (BBUs), and the evolved packet core (EPC). RRHs are connected to BBUs via the common public radio interface (CPRI) with a high-capacity fronthaul using microwave E-band, millimeter wave, or optical fiber. BBUs form a centralized pool of data center resources and denoted as C-BBU. C-BBU is connected to the EPC via the backhaul connection.

To further benefit from the advances in the areas of NFV, we consider virtualized BBU resources (V-BBU) [40] where the BBU functionality and services have been abstracted from the underlying infrastructure and virtualized in the form of virtual network functions (VNFs). To realize the virtualization, a virtual machine monitor (VMM) is used to manage the execution of BBUs. The possibility to run the control programs on general purpose computing/storage resources [41,42], as facilitated by NFV, enables the deployment of very flexible control functions for different mobile users (MUs), as required.

To benefit from the advances in the areas of SDN, an SDN controller (SDN-C) has been placed on top of the VMM. The SDN-C is responsible for routing decisions and configures the packet forwarding elements to forward packets to/from MUs. The applicability of SDN to mobile networks is intended to bring a systematic abstraction and modularity of the functions within the RAN, enabling a hierarchical control architecture in which the high control layer controls lower layers through defining behaviors without the need to know their specific implementation [43–45].

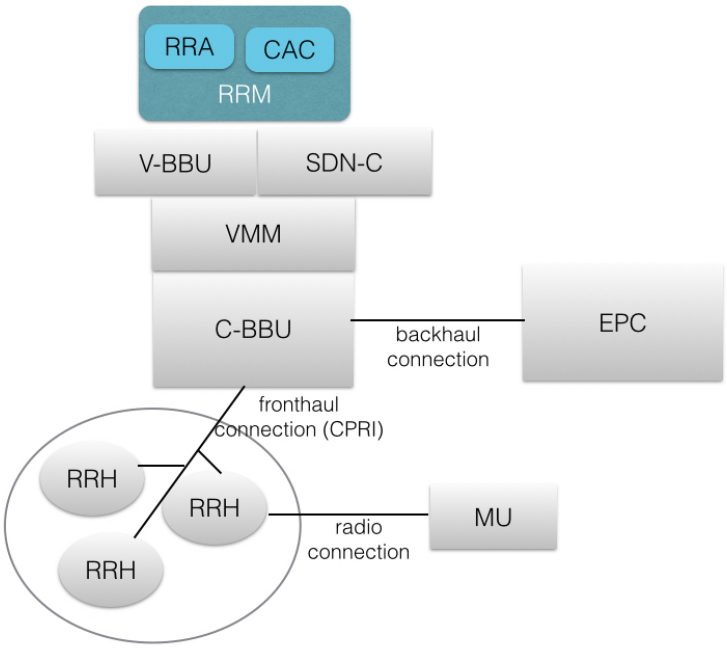

Fig. 6: Reference architecture.

5.2 Self-Organizing Network

SON refers to a set of features and capabilities for automating the operation of a network so that operating costs can be reduced and human errors minimized [46]. The incorporation of SON features in cellular networks can support and/or replace common activities, such as manual planning, deployment, optimization, and maintenance activities. These features can make network operations simpler and faster by enabling more autonomous and automated processes.

SON functions can be categorized as follows: self-planning, self-optimization, and self-healing. Our proposed WH-CDTM model mainly targets the self-optimization objective, but can also greatly facilitate the self-planning objective. The goal of self-optimization is as follows. Once the network is in operational state, the self-optimization includes the set of

processes intended to improve or maintain the network performance in terms of coverage, capacity, and QoS by tuning the different network settings [47, 48]. SON functions might automatically tune global operational settings of the base station (BS) (e.g., maximum transmit power and channel bandwidth) as well as specific parameters corresponding to the RRM functions (e.g., CAC thresholds and handover offsets).

5.3 Realizing a RRM function using hybrid SON

In this subsection we provide specific information on how our proposed approach could be realized using the SON technology in C-RAN. Let us consider the hybrid SON (hSON) where a part of the SON functionalities are centralized (cSON) at the EPC level, while others are distributed (dSON) at the RAN level (Fig. 7). The cSON sends configuration parameters to the dSON, whereas the dSON replies with performance measurements and alarms. The cSON determines the configuration parameters based on a number of performance-related objectives. In the case of the WH-CDTM, these objectives specify upper bounds for CBP per service, a target RRU, and a target throughput per service, as defined in (30), (31), and (32), respectively. The dSON is configured to report to the cSON at regular time intervals various relevant measurements, such as the SNR, $SNR^t_{s,d}$, that is used in (14) for determining the service load factor, $LF^t_{s,d}$. The cSON, upon receiving these measurements, will execute the recursive algorithm of Subsection 5.2 to determine the new state probabilities. The latter will then be used to identify whether any of the objectives has been violated (e.g., the CBP of a particular service, as determined in (30), is above the predefined level). The dSON is also configured to send an alarm message when the *observed* measurements for a performance-related objective, such as the CBP or the RRU, are outside the acceptable values. In response to the alarm message, the cSON will execute the recursive algorithm

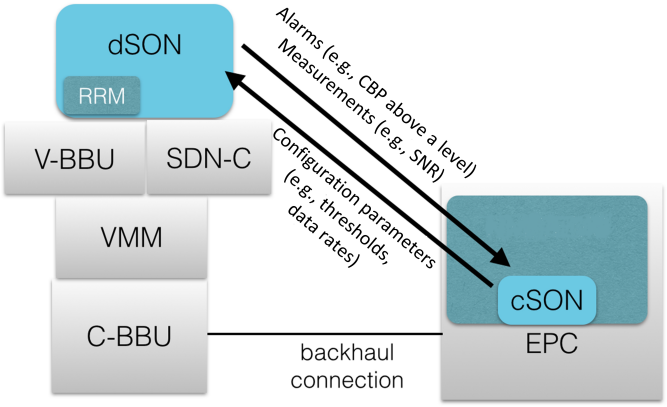

Fig. 7: Realizing a RRM function using hybrid SON in C-RAN.

of Subsection 5.2 for different sets of input parameters until the desired outcome has been reached. The updated configuration parameters that will be sent to the dSON could be for example the thresholds $TH_{s,d}$ or the data rates $DR^t_{s,d}$. Note that our derived algorithm has been particularly optimized for time-constrained operations and can operate on very short timescales (in the order of seconds).

In order to enable sharing of virtualized BBU resources among MUs, based on the architectural model of Fig. 6, the RRM function (e.g., CAC, RRA, and handover) must be implemented as a VNF. For this to be achieved, appropriate open control interfaces must be established [41]. Then, the RAN customization model can be facilitated by instantiating different VNFs of a RRM function even on a per multiple operator basis. That is, in scenarios where multiple (real or virtual) mobile network operators (MNOs) are sharing the same physical RAN infrastructure. For the realization of the NFV-based implementation, the current management architecture can be extended to incorporate the management of virtualized networks. In the context of 3GPP this is addressed in [49].

6 Numerical Examples

To evaluate the applicability and the accuracy of our proposed model, we consider three use cases. The 1st use case concerns two services and different levels of inter-cell interference. The 2nd use case concerns three services and a fixed level of inter-cell interference. The 3rd use case concerns two services with handover traffic and a fixed level of inter-cell interference.

We present analytical and simulation results for the WH-CDTM. The simulation tool used in our experiments is SIMSCRIPT III [50]. To produce simulation results we perform each experiment 6 times and calculate the mean value with 95% confidence interval. For comparison with other proposed models, we also present the analytical CBP results of the W-EMLM [29].

Finally, we have tested the speed of analytical calculations on commodity hardware. In all cases, we get running times in the order of hundreds of milliseconds. This property makes the derived recursive algorithm highly applicable for accurate cellular network dimensioning and RRM.

6.1 Use Case 1

Consider a cellular CDMA system that accommodates two services with the following parameters (shown in Table 1):

- *1st service*: adaptive video streaming with two contingency data rates $DR_{1,1} = 12.2$ Mbps and $DR_{1,2} = 6.2$ Mbps. The selection of the rate is performed according to threshold $TH_{1,1} = 0.7$. The activity factor for this service is $a_1 = 0.67$ and the SNR is $SNR_1 = 5$ dB.

– *2nd service*: elastic data transfer with two contingency data rates of $DR_{2,1} = 64$ Mbps and $DR_{2,2} = 32$ Mbps, with threshold $TH_{2,1} = 0.6$. The activity factor for this service is $a_2 = 0.8$ and the SNR is $SNR_2 = 4$ dB.

We generate traffic for both services based on 8 different traffic sets, as shown in Table 2. Each value of the table represents the total offered traffic-load of a particular service in Erlangs. We consider thermal noise $T_{noise} = -174$ dBm/Hz and two levels of inter-cell interference: $E[I_{inter}] = 3 \times 10^{-18}$ mW and $E[I_{inter}] = 5 \times 10^{-18}$ mW.

In Figs. 8 and 9 we present our experimental CBP results for both the WH-CDTM and the W-EMLM versus the offered traffic. Figure 8 shows the comparative results for the 1st service, whereas Fig. 9 the results for the 2nd service. For the WH-CDTM, we present both analytical and simulation results. The fact that analytical and simulation results are very close to each other, indicates that introduced approximation errors in the WH-CDTM are negligible and the accuracy of the analytical model is very satisfactory. This is especially true when the traffic-load is small or moderate. Also, even when increasing the inter-cell interference, the model's accuracy remains satisfactory.

The comparison of the WH-CDTM with the W-EMLM in Figs. 8 and 9, reveals that the WH-CDTM can achieve lower CBP compared to the W-EMLM. The difference is bigger in the cases of high traffic. This is because when the offered traffic is high and the population of calls in the system increases, the W-EMLM is not able to reduce the fixed amount of resources occupied by in-service calls. On the other hand, the WH-CDTM by utilizing the rate thresholds, is able to accommodate more calls with reduced resources. Another observation is that the WH-CDTM significantly outperforms the W-EMLM when the inter-cell interference is low.

Table 1: Use case 1: Service parameters.

	1st service	**2nd service**
Description	Adaptive video streaming	Elastic data transfer
Data rate (Mbps)	$DR_{1,1} = 12.2$ $DR_{1,2} = 6.2$	$DR_{2,1} = 64$ $DR_{2,2} = 32$
Rate threshold	$TH_{1,1} = 0.7$	$TH_{2,1} = 0.6$
Activity	$a_1 = 0.67$	$a_2 = 0.8$
SNR (dB)	$SNR_1 = 5$	$SNR_2 = 4$

6.2 Use Case 2

Consider a cellular CDMA system that accommodates three services with the following parameters (shown in Table 3):

- *1st service*: adaptive video streaming with two contingency data rates $DR_{1,1} = 12.2$ Mbps and $DR_{1,2} = 8.4$ Mbps. The selection of the rate is performed according to threshold $TH_{1,1} = 0.7$. The activity factor for this service is $a_1 = 0.5$ and the SNR is $SNR_1 = 5$ dB.
- *2nd service*: elastic data transfer with two contingency data rates of $DR_{2,1} = 64$ Mbps and $DR_{2,2} = 32$ Mbps, with threshold $TH_{2,1} = 0.6$. The activity factor for this service is $a_2 = 1.0$ and the SNR is $SNR_2 = 4$ dB.
- *3rd service*: adaptive video streaming with three contingency data rates $DR_{3,1} = 144$ Mbps, $DR_{3,2} = 128$ Mbps, and $DR_{3,2} = 112$ Mbps. The selection of the rate is

Table 2: Use case 1: Offered traffic (erl).

	1st service	**2nd service**
1	5	2.5
2	10	4.0
3	15	5.5
4	20	7.0
5	25	8.5
6	30	10.0
7	35	11.5
8	40	13.0

performed according thresholds $TH_{3,1} = 0.4$ and $TH_{3,2} = 0.6$. The activity factor for this service is $a_3 = 0.3$ and the SNR is $SNR_3 = 3$ dB.

We generate traffic for the three services based on 8 different traffic sets, as shown in Table 4. Each value of the table represents the total offered traffic-load of a particular service in Erlangs. We consider thermal noise $T_{noise} = -174$ dBm/Hz and inter-cell interference $E[I_{inter}] = 3 \times 10^{-18}$ mW.

In Figs. 10 and 11 we present our experimental CBP results for the WH-CDTM and the W-EMLM versus the offered traffic. Figure 10 shows the comparative results for the 1st and the 3rd services, whereas Fig. 11 the results for the 2nd service. For the WH-CDTM, we present both analytical and simulation results. We observe that the introduced approximation

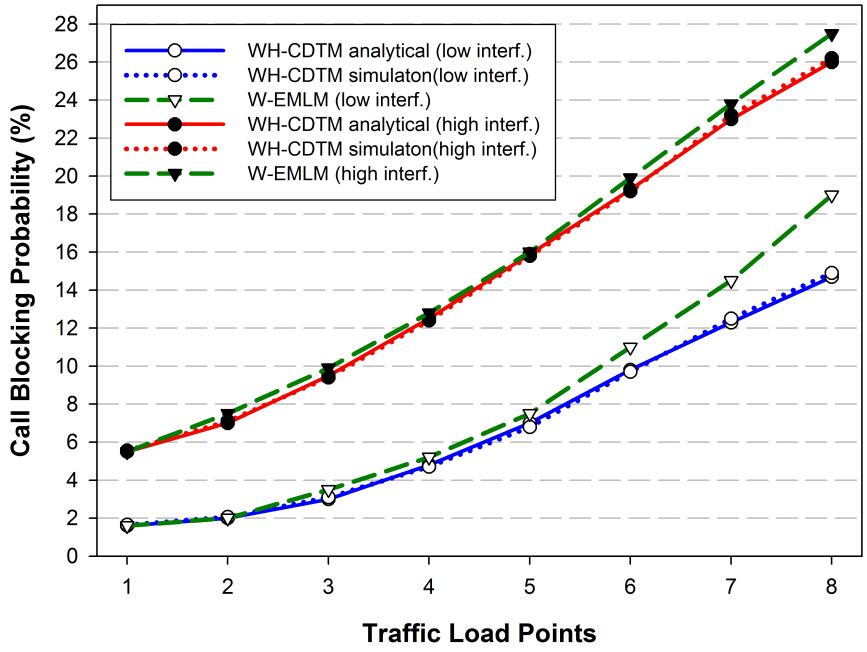

Fig. 8: Call blocking probabilities for the WH-CDTM and the W-EMLM: use case 1 (1st service).

errors are negligible and the accuracy of the analytical model is very satisfactory in all cases. Also, the comparison of Figs. 10 and 11, reveals that the WH-CDTM can achieve lower CBP compared to the W-EMLM. We observe that the difference is bigger in the cases of high traffic.

6.3 Use case 3

Consider a cellular CDMA system that accommodates two services and handover traffic with the following parameters (shown in Table 5):

- *1st service*: fixed-rate live video streaming with data rate $DR_{1,1} = 144$ Mbps. The activity factor for this service is $a_1 = 0.67$ and the SNR is $SNR_1 = 3$ dB.

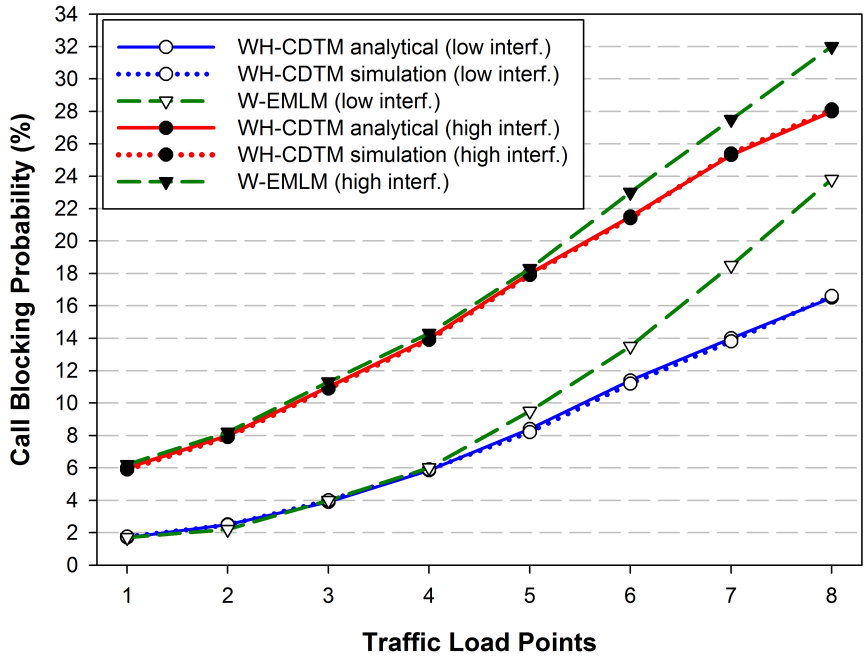

Fig. 9: Call blocking probabilities for the WH-CDTM and the W-EMLM: use case 1 (2nd service).

– *2nd service*: elastic data transfer with two contingency data rates of $DR_{2,1} = 384$ Mbps and $DR_{2,2} = 320$ Mbps, with threshold $TH_{2,1} = 0.6$. The activity factor for this service is $a_2 = 1.0$ and the SNR is $SNR_2 = 4$ dB.

We generate traffic for both services based on 6 different traffic sets, as shown in Table 6. Each value of the table represents the total offered traffic-load of a particular service in Erlangs. We consider thermal noise $T_{noise}= -174$ dBm/Hz and inter-cell interference $E[I_{inter}] = 2 \times 10^{-18}$ mW. The CAC thresholds for *new* and *handover* calls are $CL^{N}_{max} = 0.75$ and $CL^{H}_{max} = 0.8$, respectively.

Figures 12 and 13 present our experimental CBP results for the two services, for both *new* and *handover* traffic, respectively. We observe that the analytical and simulation results

Table 3: Use case 2: Service parameters.

	1st service	**2nd service**	**3rd service**
Description	Adaptive video streaming	Elastic data transfer	Adaptive video streaming
Data rate (Mbps)	$DR_{1,1} = 12.2$ $DR_{1,2} = 8.4$	$DR_{2,1} = 64$ $DR_{2,2} = 32$	$DR_{3,1} = 144$ $DR_{3,2} = 128$ $DR_{3,3} = 112$
Rate threshold	$TH_{1,1} = 0.7$	$TH_{2,1} = 0.6$	$TH_{3,1} = 0.4$ $TH_{3,2} = 0.6$
Activity	$a_1 = 0.5$	$a_2 = 1.0$	$a_3 = 0.3$
SNR (dB)	$SNR_1 = 5$	$SNR_2 = 4$	$SNR_3 = 3$

are very close to each other. This shows that accuracy of the analytical model is very satisfactory. We also observe that, due to higher CAC thresholds, the CBP for handover traffic is lower compared to the CBP of new traffic.

7 Conclusion and Future Work

In this paper, we present a novel teletraffic model for heterogeneous CDMA-based cellular systems. Different QoS requirements as well as the handover traffic have been explicitly incorporated into the model. The call arrival process has been modelled as a Poisson distribution and a complete radio resource sharing policy is assumed. Handover calls, having relatively low CAC threshold, receive higher priority compared to new calls. The cellular system has been described as a continuous-time Markov chain and provides an efficient

Table 4: Use case 2: Offered traffic (erl).

	1st service	**2nd service**	**3rd service**
1	2	1	0.75
2	6	2	1.0
3	10	3	1.25
4	14	4	1.5
5	18	5	1.75
6	22	6	2.0
7	26	7	2.25
8	30	8	2.5

expression for state probabilities. Next, important performance metrics, such as call blocking probabilities, radio resources utilization, and service throughput, can be determined. We present an applicability framework for C-RAN, which can exploit our proposed approach using SDN, NFV, and SON technologies. We evaluate the accuracy of our model using simulations and find it very satisfactory. Finally, experiments on commodity hardware show algorithm running times in the order of few hundreds of milliseconds. This property makes our algorithm highly applicable for accurate cellular network dimensioning and radio resource management.

As a future work we intend to extend the proposed model to include both the Poisson and the batched Poisson traffic types. When the connection requests arrive in batches, a batch can be either fully or partially accepted in the cell, depending on the availability of radio

Table 5: Use case 3: Service parameters.

	1st service	**2nd service**
Description	Live video streaming	Elastic data transfer
Data rate (Mbps)	$DR_{1,1} = 144$	$DR_{2,1} = 384$ $DR_{2,2} = 320$
Rate threshold	—	$TH_{2,1} = 0.6$
Activity	$a_1 = 0.67$	$a_2 = 1.0$
SNR (dB)	$SNR_1 = 3$	$SNR_2 = 4$

Table 6: Use case 3: Offered traffic (erl).

	1st service (new)	**1st service (handover)**	**2nd service (new)**	**2nd service (handover)**
1	1.0	0.1	0.2	0.05
2	1.25	0.2	0.3	0.1
3	1.5	0.3	0.4	0.15
4	1.75	0.4	0.5	0.2
5	2.0	0.5	0.6	0.25
6	2.25	0.6	0.7	0.3

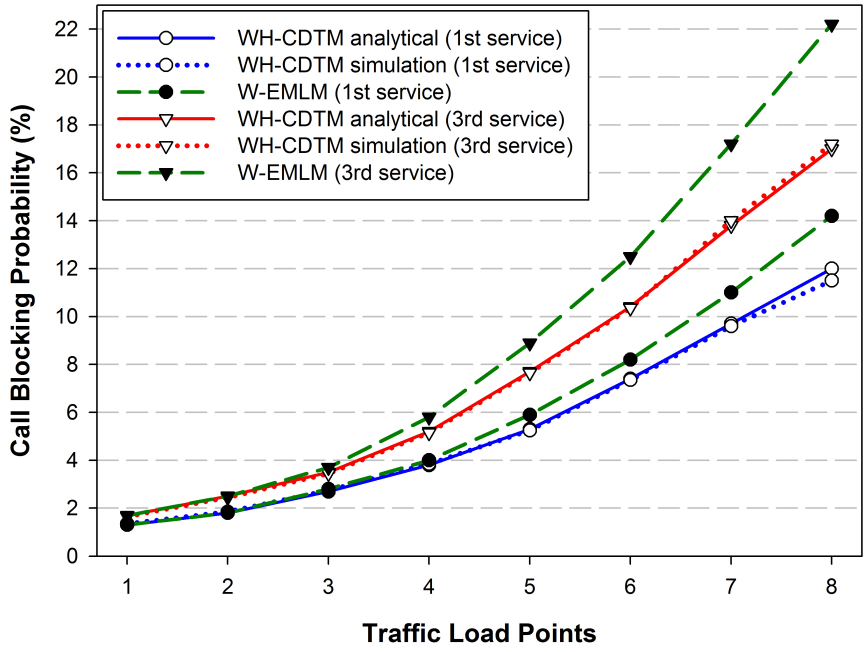

Fig. 10: Call blocking probabilities for the WH-CDTM and the W-EMLM: use case 2 (1st and 3rd services).

resources. In the second case, some calls of the batch will be serviced and the rest will be blocked. Other possible extensions of our model are the incorporations of different resource sharing policies, such as the bandwidth reservation (BR) and the multiple fractional channel reservation (MFCR) policies. The BR policy introduces a service priority to benefit high-speed calls and can be used to achieve CBP equalization among calls of different services [51, 52]. On the other hand, the MFCR policy enables a fine-grained QoS assessment by allowing the reservation of real (not integer) number of channels [53, 54].

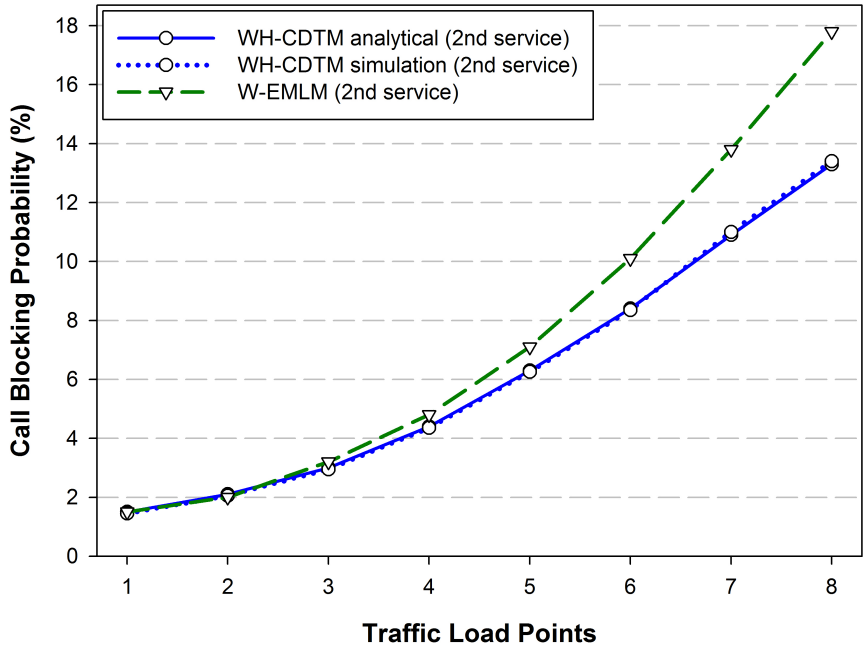

Fig. 11: Call blocking probabilities for the WH-CDTM and the W-EMLM: use case 2 (2nd service).

References

1. Shuminoski, T. and Janevski, T. (2015). 5G mobile terminals with advanced QoS-based user-centric aggregation (AQUA) for heterogeneous wireless and mobile networks. Wireless Networks 22(5), 1553-1570.
2. Ericsson (2015). Cloud RAN: The benefits of virtualisation, centralisation, and coordination. Ericsson White Paper.
3. Fujitsu (2014). The benefits of cloud-RAN architecture in mobile network expansion. Fujitsu White Paper.
4. Kyriazis, G. and Rouskas, A. (2016). Design and operation of energy efficient heterogeneous mobile networks. Wireless Networks 22(6), 2013-2028.
5. Galinina, O., Andreev, S., Turlikov, A., and Koucheryavy, Y. (2014). Optimizing energy efficiency of a multi-radio mobile device in heterogeneous beyond-4G networks. Performance Evaluation 78, 18-41.
6. Peng, C., Lee, S.-B., Lu, S., and Luo, H. (2014). GreenBSN: Enabling energy-proportional cellular base station networks. IEEE Transactions on Mobile Computing 13(11), 2537-2551.

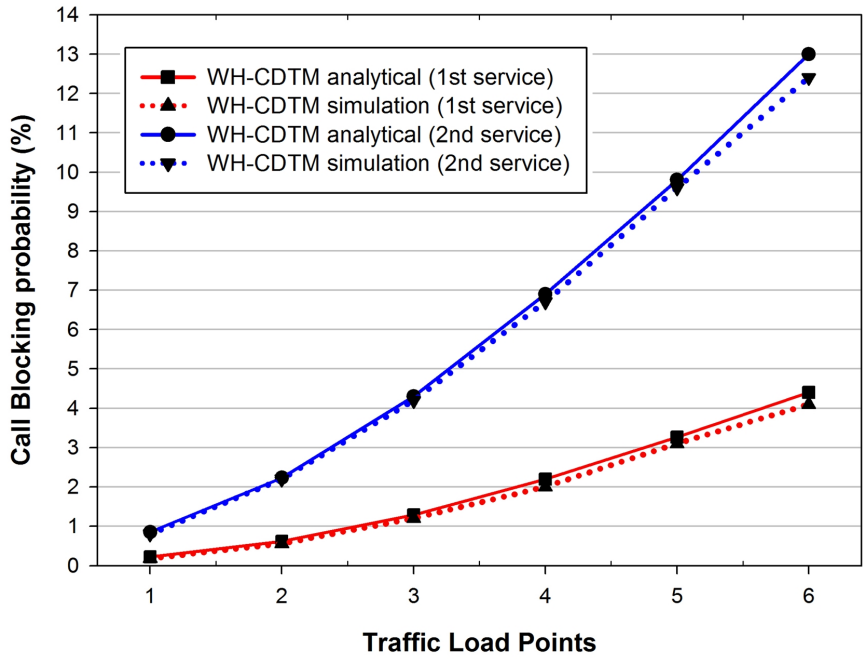

Fig. 12: Call blocking probabilities for the WH-CDTM: use case 3 (new traffic).

7. Abdulova, V. and Aybay, I. (2015). Performance evaluation of call admission control schemes with new call reattempts in wireless cellular networks. Wireless Personal Communications 84(4), 1-21.

8. Hwang, H.Y., Lee, H., Roh, B., and Kim, S. (2016). Joint resource allocation, routing and CAC for uplink OFDMA networks with cooperative relaying. Wireless Networks 22(5), 1493-1503.

9. Chavez-Santiago, R., Szydelko, M., Kliks A., Foukalas, F., Haddad, Y., Nolan, K. E., Kelly, M. Y., Masonta, M. T., and Balasingham, I. (2015). 5G: The convergence of wireless communications. Wireless Personal Communications 83(3), 1-26.

10. Demestichas, P., Georgakopoulos, A., Karvounas, D., Tsagkaris, K., Stavroulaki, V., Lu, J., Xiong, C., and Yao, J. (2013). 5G on the horizon: Key challenges for the radio-access network. IEEE Vehicular Technology Magazine 8(3), 47-53.

11. Al-Junaid, A.F. and Al-Kamali, F.S. (2016). Efficient wireless transmission scheme based on the recent DST-MC-CDMA. Wireless Networks 22(3), 813-824.

12. Wang, Z., Fan, S., and Rui, Y. (2014). CDMA-FMT: A novel multiple access scheme for 5G wireless communications. In: IEEE 19th International Conference on Digital Signal Processing (DSP), 898-902.

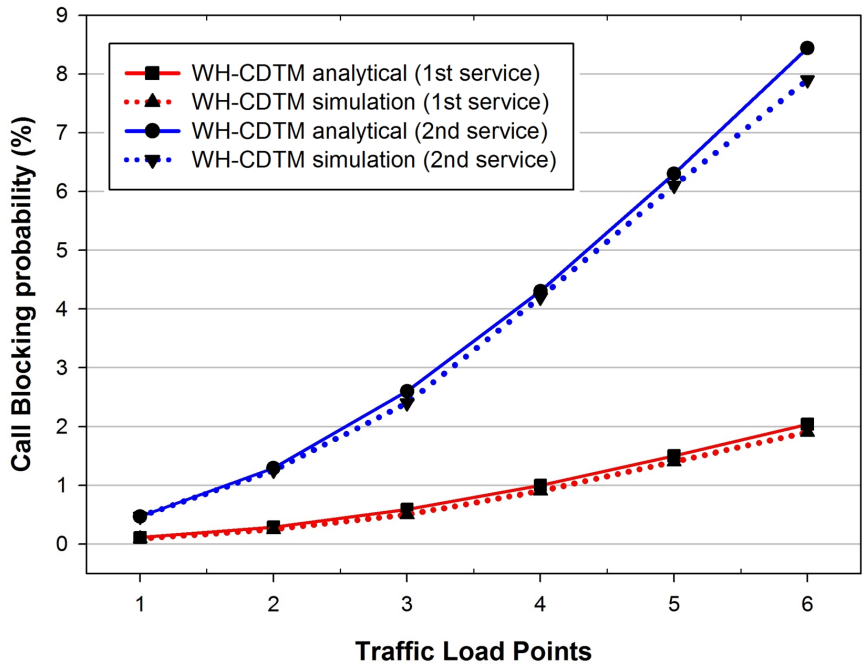

Fig. 13: Call blocking probabilities for the WH-CDTM: use case 3 (handover traffic).

13. Hossain, E., Rasti, M., Tabassum, H., and Abdelnasser, A. (2014). Evolution toward 5G multi-tier cellular wireless networks: An interference management perspective. IEEE Wireless Communications 21(3), 118-127.

14. Ravindrababu, J., Rao, K., and Rao, R. (2014). Interference and complexity reduction in multi-stage multi-user detection in DS-CDMA. Wireless Personal Communications 79(2), 1385-1400.

15. Mahadevappa, R. H. and Proakis, J. G. (2002). Mitigating multiple access interference and intersymbol interference in uncoded CDMA systems with chip-level interleaving. IEEE Transactions on Wireless Communications 1(4), 781-792.

16. Wang, H., Rosa, C., and Pedersen, K.I. (2016). Dual connectivity for LTE-advanced heterogeneous networks. Wireless Networks, 22(4), 1315-1328.

17. Aguiar, E. et al. (2014). A real-time video quality estimator for emerging wireless multimedia systems. Wireless Networks 20(7), 1759-1776.

18. Su, G.M., Su, X., Bai, Y., Wang, M., Vasilakos, A.V., and Wang, H. (2015). QoE in video streaming over wireless networks: Perspectives and research challenges. Wireless Networks 22(5), 1571-1593.

19. Vassilakis, V. G., Moscholios, I. D., and Logothetis, M. D. (2008). Call-level performance modelling of elastic and adaptive service-classes with finite population. IEICE Transactions on Communications 91(1), 151-163.

20. Kaufman, J. S. (1981). Blocking in a shared resource environment. IEEE Transactions on Communications 29(10), 1474-1481.

21. Roberts, J. W. A service system with heterogeneous user requirements. Performance of Data Communications Systems and Their Applications, Amsterdam, The Netherlands: North-Holland 29(10), 423-431.

22. Yang, S.-T. and Ephremides, A. (1996). On the optimality of complete sharing policies of resource allocation. In: IEEE Conference on Decision and Control, 299-300.

23. Kaufman, J. S. (1992). Blocking in a completely shared resource environment with state dependent resource and residency requirements. In: Eleventh Annual Joint Conference of the IEEE Computer and Communications Societies INFOCOM'92., 2224-2232.

24. Kaufman, J. S. (1992). Blocking with retrials in a completely shared resource environment. Performance Evaluation 15(2), 99-116.

25. Moscholios, I. D., Logothetis, M. D., and Kokkinakis, G. K. (2002). Connection-dependent threshold model: A generalization of the Erlang multiple rate loss model. Performance Evaluation 48(1), 177-200.

26. Vassilakis, V. G., Moscholios, I. D., and Logothetis, M. D. (2007). Call-level performance modelling of elastic and adaptive service-classes. In: IEEE International Conference on Communications (ICC), 183-189.

27. Vassilakis, V. G., Moscholios, I. D., and Logothetis, M. D. (2012). The extended connection-dependent threshold model for call-level performance analysis of multi-rate loss systems under the bandwidth reservation policy. International Journal of Communication Systems 25(7), 849-873.

28. Holma, H. and Toskala, A.(2006). WCDMA for UMTS, Chichester: John Wiley & Sons.

29. Staehle, D. and Mäder, A. (2003). An analytic approximation of the uplink capacity in a UMTS network with heterogeneous traffic. Teletraffic Science and Engineering 5, 81-90.

30. Vassilakis, V. G., Moscholios, I. D., Vardakas, J. S., and Logothetis, M. D. (2014). Handoff modeling in cellular CDMA with finite sources and state-dependent bandwidth requirements. In: IEEE 19th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), 26-30.

31. Altman, E. (2002). Capacity of multi-service cellular networks with transmission-rate control: A queueing analysis. In: ACM 8th Annual International Conference on Mobile Computing and Networking, 205-214.

32. Fodor, G. and Telek, M. (2005). A recursive formula to calculate the steady state of CDMA networks. In: ITC-19, Beijing, China, 1285-1294.

33. Xu, Y., Zhou, Y., and Chiu., D. M. (2014). Analytical QoE models for bit-rate switching in dynamic adaptive streaming systems. IEEE Transactions on Mobile Computing 13(12), 2734-2748.

34. Mäder, A. and Staehle, D. (2004). Analytic modeling of the WCDMA downlink capacity in multi-service environments. In: 16th ITC Specialist Seminar, 229-238.

35. Daskalopoulos, I. B., Vassilakis, V. G., and Logothetis, M. D. (2009). Thorough analysis of downlink capacity in a WCDMA cell. In: Mobile Lightweight Wireless Systems, Springer Berlin Heidelberg, 1-14.

36. Sangeetha, M., Bhaskar, V., and Cyriac, A. R. (2014). Performance analysis of downlink W-CDMA systems in Weibull and Lognormal fading channels using chaotic codes. Wireless Personal Communications 74(2), 259-283.

37. Papadaki, K. and Friderikos, V. (2012). Multi-rate control policies for elastic traffic in CDMA networks. Performance Evaluation 69(10), 510-523.

38. Aein, J. M. (1978). A multi-user-class, blocked-calls-cleared, demand access model. IEEE Transactions on Communications 26(3), 378-385.

39. Ross, K. W. (1995). Multiservice loss models for broadband telecommunication networks. Springer Science & Business Media.

40. Checko, A., Christiansen, H., Yan, Y., Scolari, L., Kardaras, G., Berger, M., and Dittmann, L. (2015). Cloud RAN for mobile networks - a technology overview. IEEE Commun. Surveys & Tutorials 17(1), 405-26.

41. ETSI GS NFV-MAN 001 (V1.1.1) (2014). Network Function Virtualisation (NFV); Management and Orchestration.

42. Akyildiz, I.F., Lin, S.C., and Wang, P. (2015). Wireless software-defined networks (W-SDNs) and network function virtualization (NFV) for 5G cellular systems: An overview and qualitative evaluation. Computer Networks 93, 66-79.

43. Chen, T., Matinmikko, M., Chen, X., Zhou, X., and Ahokangas, P. (2015). Software defined mobile networks: concept, survey, and research directions. IEEE Communications Magazine 53(11), 126-133.

44. Haleplidis, E., Salim, J.H., Denazis, S., and Koufopavlou, O. (2015). Towards a network abstraction model for SDN. Journal of Network and Systems Management 23(2), 309-327.

45. Vassilakis, V.G., Moscholios, I.D., Alzahrani, B.A., and Logothetis, M.D. (2016). A software-defined architecture for next-generation cellular networks. In: IEEE International Conference on Communications (ICC), 1-6.

46. 3GPP TR 32.500 v12.1.0 (2014). Self-Organizing Networks (SON); Concepts and requirements (Release 12).

47. Ramiro, J. and Hamied, K. eds. (2011). Self-organizing networks (SON): Self-planning, self-optimization and self-healing for GSM, UMTS and LTE, Wiley.

48. Sallent, O., Pérez-Romero, J., Ferrús, R., and Agustí, R. (2016). Small cells as a service: From capacity provisioning to full customisation. In: European Conference on Networks and Communications (EuCNC), Athens, Greece.

49. 3GPP TR 32.842 v13.1.0 (2015). Telecommunication management; Study on network management of virtualized networks (Release 13).

50. SIMSCRIPT III, http://www.simscript.com [last accessed Nov. 2016]

51. Glabowski, M., Kaliszan, A., and Stasiak, M. (2008). Asymmetric convolution algorithm for blocking probability calculation in full-availability group with bandwidth reservation. IET Circuits, Devices & Systems 2(1), 87-94.

52. Moscholios, I.D., Logothetis, M.D., Vardakas, J.S., and Boucouvalas, A.C. (2015). Congestion probabilities of elastic and adaptive calls in Erlang-Engset multirate loss models under the threshold and bandwidth reservation policies. Computer Networks 92, 1-23.

53. Cruz-Pérez, F.A., Vázquez-Ávila, J.L., and Ortigoza-Guerrero, L. (2004). Recurrent formulas for the multiple fractional channel reservation strategy in multi-service mobile cellular networks. IEEE Communications Letters 8(10), 629-631.

54. Moscholios, I.D., Vassilakis, V.G. and Logothetis, M.D. (2016). Call blocking probabilities for poisson traffic under the multiple fractional channel reservation policy. In: 10th IEEE/IET International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP), 1-5.