Performance Evaluation of Multimedia Services in Cellular Networks

Péter Fazekas, Sándor Imre and Miklós Telek
Department of Telecommunications, Budapest University of Technology and Economics
Pázmány P. sétány 1/D, H-1117, Budapest, Hungary, E-mail:{fazekasp}@hit.bme.hu

Abstract—**In this paper a new analytical method is presented for analyzing cellular networks with connections that change their transmission rate during the session. The method is based on a Markovian model of a single base station and an approximate recursive formula following from the Markov model. This formula is applicable for calculating call level system parameters rapidly with reasonable error. Two simple admission control strategy is also presented and investigated. The accuracy of the proposed approximate method is verified by computer simulations.**

Keywords— **multimedia services, traffic analysis,Phase Type distributions, Kaufman-Roberts formula**

I. Introduction

Present days we face the enormous development of high-speed communication networks, especially the rapid growth of capacity and availability of wireless networks. Future cellular systems will provide enough bandwidth for the customers to use multimedia network applications (video telephony, web browsing, data applications, etc.) besides voice connections.

Along with the evolution from second generation cellular networks towards the third generation, the development of wireless data networks (like wireless ATM testbeds [1]-[4] or Hiperlan 2 [5]) guarantees the technical bases of using multimedia communications in wireless networks.

Unlike voice calls in conventional cellular networks, multimedia connections generally do not generate traffic at constant rate. Since the radio channel is spare and expensive, cellular networks should not allocate capacity for a connection according to its peak rate. Rather the medium access method of multimedia cellular systems will allocate radio capacity according to the connection's instantaneous transmission rate.

The simplest method of providing variable amount of capacity for a connection is to use TDMA/TDD access method. In this case a scheduler is responsible for distributing the time slots of the TDMA frame among the connections. A connection may use different number of time slots in different frames, according to its instantaneous need. In a CDMA environment the versatility of the allocated capacity may be fulfilled by means of the variation of the length of the spreading code.

For network planning and rough network dimensioning purposes analytical models are required that handle multimedia connections. Numerous papers analyze the performance parameters of medium access methods that provide variable amount of capacity [6]-[10], however a general model is still needed to investigate the effect of multimedia connections in cellular networks. Although several papers are devoted to analyze cellular networks with multiple service classes (e.g. [11]-[17] and references therein), these models do not consider connections that require variable amount of capacity during the session. Generally the variable nature of multimedia traffic is considered with the use of effective bandwidth, that is constant during the connection.

This work was supported by the Inter-University Centre for Telecommunications and Informatics (ETIK) and OTKA grant No. F-032268 and T-30685

In analytical models of cellular systems generally three random time variables are used to describe user behavior. The session length characterizes the time of a connection. It is very common to model this time with exponential distribution. However, when considering multimedia connections this assumption does not always hold. The dwell time is used to describe user mobility. This is also often modeled with exponential distribution, although this is reasonable only under very special circumstances. The channel holding time describes the time a customer spends communicating in a radio cell. This time is usually derived from the dwell time and the session length.

In this paper we present a modeling method that handles the versatility of the generated traffic of a connection. Moreover, the dwell time and the session length of customers is modeled with a very general type of distribution. The method is based on a Markovian model of a base station and an approximate recursive formula is used to calculate system parameters.

The paper is organized as follows. In Section II the modeling assumptions are presented, including the user describing times, the traffic model of a connection and the base station model. Section III is devoted to the deduction of the service time distribution. In Section IV and V the system describing Markov chain and the approximate formula is presented. This is followed by numerical results and conclusions.

II. Analytical modeling

The aim of the method proposed in this paper is to determine connection level system parameters of a single base station. These parameters are the probability of call blocking and handover failure along with the channel utilization.

Two types of connections may appear at the examined base station. In the paper we use the term new connection meaning a connection that is initiated within the coverage area of the base station. Handover calls are already established connections that arrive to the base station from the vicinity of the radio cell. We suppose that always a large number of idle customers roam within the radio cell, and they initiate new connections independently with small probability. Therefore the number of new connection attempts during a time interval is approximated with a Poisson distributed random variable, e.g. the arrival of new connections forms a Poisson process. Because of similar reasons the arrival of handover attempts is also modeled with a Poisson

process. The rates of the new calls and handover attempts are denoted with λ_N and λ_H respectively. Obtaining these rates is out of the scope of this paper, but the rates may either be results of measurements or calculated according to the method proposed in [15]-[17].

We suppose that K types of customers may appear in the system. Users of different types has different traffic characteristics (session length and generated traffic), in this case they belong to different service classes. In addition, mobiles may differ by means of their mobility. This means that customers belonging to the same service class may belong to different user types, if their dwell times are different.

A. Dwell time and session length

The dwell time of a customer describes its mobility. The dwell time begins at the instant when the mobile enters the coverage area of the examined base station and ends when the mobile leaves the cell, regardless it is communicating or not. We introduce the notion of residual dwell time, which is the remaining of the dwell time after the mobile starts transmission at the given cell. From the definition of the residual dwell time it is clear that the residual dwell time of a mobile that initiates handover into the cell is equivalent with its dwell time since it starts transmission in the cell at the same instant when it enters the cell. The residual dwell time of a mobile that initiates new call is different, since it sets up the connection somewhere inside the cell. Authors dealing with mobility models therefore often define two different dwell time distributions for handover and for new calls (e.g. [18] and [19] and references therein). Here we suppose that the residual dwell times are given by distribution, or by statistical data.

The session length begins when the mobile starts transmitting (somewhere in the network, not necessarily in the examined cell) and ends when the mobile terminates the call. However, in our approach the session length is not exponentially distributed, therefore its distribution is not memoryless. If a communicating mobile arrives to the cell some time has already elapsed from the beginning of its connection. For further analysis we are interested in the residual time of the session. The residual session length is the time interval between the admission to the cell (as handover or new connection) and the connection termination. It follows from the definition that the residual session length of new calls is equivalent with the session length. The determination of the residual session length is out of the scope of this paper, however we suppose that it is given for all connection types by distribution or by statistical data.

The channel holding time is the time a mobile spends communicating in the cell. If the mobile terminates its call before leaving the cell, its channel holding time is equal to its residual session length. If the mobile leaves the cell before terminating the session, the channel holding time is equal to the residual session length. Namely the channel holding time is the minimum of the residual dwell time and the residual session length.

Given the residual session lengths and residual dwell times, we suppose that a Phase Type distribution (PH)[20] is fitted to the two distributions (or according to the statistics of the two times). Then these fitted PHs are used in the analysis as the residual dwell time and the residual session length.

A PH random time is a mixture of n exponentially distributed phases. Upon the beginning of the process the initial phase is chosen according to the initial probability vector. After completing a phase the process jumps to another phase or terminates. A probability matrix determines the next phase or the termination of the process after a phase is completed.

In other terms, a PH time is the time a finite state Markov chain reaches an absorbing state. The distribution is determined by the initial probability vector $\underline{t}$ and the infinitesimal generator matrix $\boldsymbol{T}$ and a column vector $\underline{T}^0$. Matrix $\boldsymbol{T}$ contains the rates among the non-absorbing states of the Markov chain and the column vector $\underline{T}^0$ contain the rates from each state into the absorbing state.

Studies show ([21]-[23]) that most distributions can be accurately approximated by an appropriately chosen Phase Type distribution. Moreover, if measurement data is available about a random variable with unknown distribution, a properly chosen PH can be fitted according the data series and this PH approximates the variable's unknown distribution accurately. The use of PH distributions to describe the dwell time is not unknown in the literature. The exponential distribution is itself the simplest (one phase) PH, therefore our model includes the cases when the dwell time or the session length has exponential distribution. In [16] and [17] sum of hyperexponential (SOHYP) distribution was used as dwell time, that is also a PH distribution.

The use of PH distributions in analytical modeling has the advantage that - although the PH itself is not memoryless - its phases are exponentially distributed. At the expense of larger state space well known queuing theory methods can be applied when using PH distributions.

B. Generated traffic

A customer's generated data during a connection is described by a finite state Markov chain. For a type k connection it is characterized by the initial probability vector $\underline{q}^{(k)}$ and rate matrix $\boldsymbol{Q}^{(k)}$. Each state of the Markov chain is assigned with a transmission rate these rates are the elements of the vector $\underline{c}^{(k)}$. Thus the customer can transmit with a given set of possible transmission rates. The customer begins its transmission with a rate that is determined by the initial probability vector of the underlying Markov chain. The transmission continues with this rate for an exponentially distributed time, then the transmission rate changes according to the rate matrix of the Markov chain. This model does not capture all properties of multimedia traffic (for instance LRD properties of video transmission) but it is widely used in the literature to describe several services [24]-[26].

If a new call is initiated, its first transmission rate is determined by the initial probability vector of the Markov chain. However, handover connections have been active before admission to the base station. Since the changes of the transmission rate of a traffic is generally very fast compared to the dwell time and the connection length, we assume that the underlying Markov chain reaches its equilibrium until the instant of handover. Thus we suppose that the initial distribution of the traffic describing Markov chain for a type k handover connection is the steady state distribution of the Markov chain, obtained from the

well known equation:

$$0 = \underline{q}^{(k),H} \boldsymbol{Q}^{(k)},$$

where the subscript H denotes handover connection.

C. Base station operation

The base station is characterized by the transmission capacity of all the radio channels it controls. This capacity is denoted by C_0 and it is expressed in the same units as the transmission rates of connections. Generally, the failure of a handover attempt is less tolerable than restricting a new connection. Moreover, we suppose that some service types do not tolerate blocking. Therefore we assume that the base station does not allow the use of the total capacity for all connection types, namely it preserves some capacity for several connection types and for handover calls. We denote the maximum available capacity for type k handover and type k new calls with $C_{k,H}$ and $C_{k,N}$, respectively. If the total occupied capacity at the base station is C_{oc} and the vector $\underline{c}^{(k)}$ contains the possible transmission rates of type k connections, a handover call arriving with transmission rate $c_i^{(k)}$ is admitted if $C_{oc} + c_i^{(k)} \leq C_{k,H}$. The same applies for new calls as well.

We investigate two simple admission control policies to handle the event when a type k connection arrives with a transmission rate $c_i^{(k)}$ so that $C_{oc} + c_i^{(k)} > C_{k,H}$ if it is a handover call, or in case of new call $C_{oc} + c_i^{(k)} > C_{k,N}$. This is the event when a connection attempt cannot be admitted because of its too high instantaneous transmission rate. The two policies are:

- policy 1: the connection is immediately blocked. In case of handover connection or blocking sensitive connection type this is not tolerable.
- policy 2: the connection is forced to reduce its transmission rate. If $c_j^{(k)}$ is the highest transmission rate so that $C_{oc} + c_j^{(k)} \leq C_{k,H}$ or $C_{oc} + c_j^{(k)} \leq C_{k,N}$, the connection begins its transmission with rate $c_j^{(k)}$. The connection is only blocked when $C_{oc} + c_{min}^{(k)} > C_{k,H}$ or $C_{oc} + c_{min}^{(k)} > C_{k,N}$, where $c_{min}^{(k)}$ is the lowest possible transmission rate. Forcing a connection to transmit with a lower rate at packet level may cause increased queuing delay or even packet losses due to the overload of transmission buffers. Thus the connection is not refused but it suffers some degradation of QoS parameters.

For different connection types and for handover and new connections the base station may apply different admission control policy. The applied policy as well with the available amount of capacity must be carefully set, according to the sensitivity and negotiated QoS of different connection types.

According to our model, mobiles change their transmission rate during a session. This results in the change of the amount of occupied capacity. If a connection raises its transmission rate so that the occupied capacity would be greater than the allowed maximum, the base station simply restrict this change and the mobile continues transmission with the previous rate.

III. Service time distribution

Our aim is to create an queuing model of the system presented in the previous section. To formulate this model, a service time distribution is needed, that has the following properties:

- its distribution is the same as the distribution of the channel holding time,
- it contains the instantaneous transmission rate of the connections.

The residual dwell time of a type k customer is given by its rate matrix $\boldsymbol{D}^{(k)}$, the vector $\underline{D}^{(k),0}$ and the initial probability vector $\underline{d}^{(k)}$. Similarly, the descriptors of the residual session length of a type k connection are $\boldsymbol{L}^{(k)}$, $\underline{L}^{(k),0}$ and $\underline{l}^{(k)}$. The channel holding time of a type k call also has a PH distribution and can be composed from the residual dwell time and the residual session length as follows (note that the residual dwell time of handover connections is the dwell time and the residual session length of new calls is the session length). Let the number of phases of the residual dwell time be denoted by $N_D^{(k)}$, that of the residual session length is $N_L^{(k)}$. Then $N_D^{(k)}$ group is formed, each group containing all the $N_L^{(k)}$ phases of the session length. Among the phases of a group the rates are the rates of the residual session length distribution, i.e. taken from the matrix $\boldsymbol{L}^{(k)}$. Between he appropriate phases of different groups the rates are the same as the rates of the residual session length. This means, that the rate between phase i of group n and phase j of group m (that corresponds to phase n and m of the PH residual dwell time) is:

- $\boldsymbol{L}_{ij}^{(k)}$ if $n = m, \quad i \neq j$,
- $\boldsymbol{D}_{nm}^{(k)}$ if $i = j, \quad n \neq m$,
- 0 if $n \neq m, \quad i \neq j$,

for $i, j = 1, ... \, N_L^{(k)}, n, m = 1, ... \, , N_D^{(k)}$.

According to this composition of the channel holding time, its parameters $\underline{t}^{(k)}$, $\boldsymbol{T}^{(k)}$ and $\underline{T}^{(k),0}$ can be computed as:

$$\begin{gathered} \boldsymbol{T}^{(k)} = \boldsymbol{D}^{(k)} \oplus \boldsymbol{L}^{(k)}, \quad \underline{T}^{(k),0} = \underline{D}^{(k),0} \oplus \underline{L}^{(k),0}, \\ \underline{t}^{(k)} = \underline{d}^{(k)} \otimes \underline{l}^{(k)}, \end{gathered} \tag{1}$$

where $\oplus$ and $\otimes$ denotes the Kronecker sum and product, respectively. It is easy to prove that a PH composed in the described manner, with the parameters of (1) has the distribution of the minimum of the two PHs it was composed from. It is obvious that the channel holding time has $N_T^{(k)} = N_D^{(k)} \cdot N_L^{(k)}$ phases.

To include the connection's instantaneous bandwidth requirement in the service time distribution, similar procedure has to be done with the channel holding time and the traffic describing Markov chain that was performed to create the channel holding time distribution. The role of residual session length is replaced by the channel holding time and the role of residual dwell time is performed with the states of the traffic describing Markov chain.

It follows from the construction of the service time distribution that its descriptors are given as:

$$\begin{gathered} \boldsymbol{S}^{(k)} = \boldsymbol{Q}^{(k)} \oplus \boldsymbol{T}^{(k)}, \quad \underline{S}^{(k),0} = \underline{h}_{N_Q^{(k)}} \otimes \underline{T}^{(k),0}, \\ \underline{s}^{(k)} = \underline{q}^{(k)} \otimes \underline{t}^{(k)}, \end{gathered} \tag{2}$$

where $\boldsymbol{S}^{(k)}$, $\underline{S}^{(k),0}$ and $\underline{s}^{(k)}$ are the parameters of the PH service time distribution and $N_Q^{(k)}$ denotes the number of transmission rates of a type k connection. It is straightforward to show that the PH distribution composed in the described manner has the same distribution as the channel holding time. The instantaneous transmission rate can be also tracked, since if a connection

is receiving the $(m-1)\cdot N_L^{(k)}\cdot N_D^{(k)}+(j-1)\cdot N_L^{(k)}+i$ th phase of its service time, its transmission rate is $c_m^{(k)}$ (and it is in the ith phase of the residual session length and in the jth phase of the residual dwell time). The service time has $N_L^{(k)}\cdot N_D^{(k)}\cdot N_Q^{(k)}$ phases, thus as it is described in the next section causes that the resulting Markov process has multiple dimensions.

Because handover calls and new calls of type k are different by means of their residual session length and residual dwell time along with the initial probability vector of the traffic describing Markov chain, the channel holding time and the service time is also different for handover and new calls of the same type. Thus for each connection type two service times are required, this is totally $2K$ service times. We do not allow the change of service type of a connection during the session, thus the $2K$ service times can be handled independently. For the rest of the paper we distinguish the service times of handover and new calls, this is denoted by the index H and N.

IV. The driving process

Given the incoming process Poisson and the service time is PH and customers change their amount of occupied capacity according to the phase of their service time, formally we have to solve the multiclass M/PH/C_0 queue with phase dependent capacity requirements.

The state of the resulting Markov process is the vector $\underline{n}=[\underline{n}^{(1),N},\dots,\underline{n}^{(K),N},\underline{n}^{(1),H},\dots,\underline{n}^{(K),H}]$, where the ith element of vector $\underline{n}^{(k),N}$, $n_i^{(k),N}$ denotes the number of type k customers arrived as new call receiving the ith phase of the type k new call service time.

Let the vector $\underline{r}^{(k)}$ contain the transmission rate of a type k user that is in the ith phase of its service time $r_i^{(k)}$ in its ith position. It is clear form the composition of the service time that

$$\begin{aligned} r_1^{(k)} &= r_2^{(k)} = \dots = r_{N_L^{(k)}N_D^{(k)}} = c_1^{(k)};\\ r^{(k)}_{N_L^{(k)}N_D^{(k)}+1} &= r^{(k)}_{N_L^{(k)}N_D^{(k)}+2} = \dots = r^{(k)}_{2N_L^{(k)}N_D^{(k)}} = c_2^{(k)}\\ &;\dots;\\ r_{(N_Q^{(k)}-1)N_L^{(k)}N_D^{(k)}+1} &= \dots = r_{N_Q^{(k)}N_L^{(k)}N_D^{(k)}} = c^{(k)}_{N_Q^{(k)}}\ . \end{aligned} \tag{3}$$

The valid states of the system are those, where

$$\begin{aligned} \underline{n}^{(k),N}\cdot\underline{r}^{(k)}\le C_{k,N},\quad \underline{n}^{(k),H}\cdot\underline{r}^{(k)}\le C_{k,H}\quad k=1\dots K\\ \textstyle\sum_{k=1}^{K}\underline{n}^{(k),N}\cdot\underline{r}^{(k)}+\sum_{k=1}^{K}\underline{n}^{(k),H}\cdot\underline{r}^{(k)}\le C_0. \end{aligned} \tag{4}$$

This simply means that the amount of occupied for each type handover and new calls can not exceed the maximum available capacity for that type and the total amount of occupied capacity cannot exceed the base station capacity.

A. State transitions

State transitions can happen because of the following events: a handover or new call arrival, a customer leaves the system by handover or by connection termination, a customer changes its phase of service time. To simplify notations, in the following we do not use the whole state vector $\underline{n}$, rather its sub-vector $\underline{n}^{(k),N}$ or $\underline{n}^{(k),H}$ that changes as the result of state transition. The rates of state transitions are dependent on the applied admission control policy as well.

The arrival rate is denoted by λ_N and λ_H for new and handover calls and let the vector $\underline{\alpha}$ contain in its kth position the probability that an arriving connection is of type k. If for type k policy 1 is applied, the state transition rates are the following:

- state transition due to a new call arrival:

this event results in a state transition from state $\underline{n}^{(k),N}$ to state $\underline{n}^{(k),N}+\underline{e_i}$ at rate $\lambda_N\cdot\alpha_k\cdot s_i^{(k),N}$, where e_i is the $N_D^{(k),N}\times N_L^{(k),N}\times N_Q^{(k)}$ dimensional vector filled with 0s and one 1 at its ith position;

- state transition due to a handover arrival:

this result in a state change from state $\underline{n}^{(k),H}$ to state $\underline{n}^{(k),H}+\underline{e_i}$ at rate $\lambda_H\cdot\alpha_k\cdot s_i^{(k),H}$,

- state transition due to a call termination (by handover out of the cell or by connection termination):

this event results in a state transition from state $\underline{n}^{(k),N}$ to state $\underline{n}^{(k),N}-\underline{e_i}$ at rate $n_i^{(k),N}\cdot S_i^{(k),0}$;

- state transition due to a phase change from phase i to j:

this event results in a state transition from state $\underline{n}^{(k),N}$ to $\underline{n}^{(k),N}+\underline{e_j}-\underline{e_i}$ at rate $n_i^{(k),N}\cdot S_{ij}^{(k)}$.

To classify the state transition probabilities when policy 2 is applied, we again use the notion C_{oc} to denote the total amount of occupied capacity at the base station. The state transitions are:

- a new call arrival result in a transition from state $\underline{n}^{(k),N}$ to state $\underline{n}^{(k),N}+\underline{e_i}$, a handover call results in a state transition from state $\underline{n}^{(k),H}$ to state $\underline{n}^{(k),H}+\underline{e_i}$ with the rates of:

$$\lambda_N\cdot\alpha_k\cdot\sum_{j:r_j^{(k)}\ge C_{k,N}-C_{oc}} s_j^{(k),N},\ \lambda_H\cdot\alpha_k\cdot\sum_{j:r_j^{(k)}\ge C_{k,H}-C_{oc}} s_j^{(k),H} \tag{5}$$

respectively.

By observing the difference between the rates applying the two admission control policies it is clear that the second policy allows more connections to be admitted in case of overloaded system.

The steady state distribution of the process can be achieved by enumerating the states and creating the transition rate matrix of the system accordingly. Obtaining the steady state distribution then means solving a set of linear equations. However, due to the multiple dimensions of the process the state space may become very large. Even in case of the simplest model (two phase dwell time and session length, two possible transmission rates, single connection type) the number of states can exceed 10^6. In this case the storage and computational capacity of today's computers is not enough, thus solving a set of several million linear equations is impossible.

Fortunately, to obtain blocking probabilities and channel utilization we do not need the steady state distribution explicitly. It is enough if the probabilities of having m units of capacity occupied is known, $m=1\dots C_0$.

V. Approximate calculation of channel occupancy probabilities

Here we propose an approximate method to calculate channel occupancy probabilities. This method has negligible computa-

tional complexity and provide results with reasonable error.

If we consider the base station to have infinite capacity and all the connection types can occupy any amount of capacity, the resulting process has a product form solution. This means that in equilibrium the probability of a state can be calculated as the product of the probability of one of its neighbors and a multiplying factor. This factor has the same form in the whole state space, therefore the probability of each state can be calculated easily by a recursive formula. Moreover, if the connections could not change their transmission rates (constant bit-rate sources) the system then also would have a product form solution.

A Markov chain has a product form solution if and only if local balance equations hold throughout the whole state space. While global balance equations mean that the rates out of a state hold balance with the rates into that state, local balance means that transition rates crossing a given "surface" at the state space hold balance. The multiplying factors of the product form solution is also calculated from the local balance equations.

We observed, that in our problem at the majority of the state space the local balance equations hold. The local balance equations change in those states that represent an amount of total occupied capacity that is too big, so some connections can not be admitted or the raise of transmission speed is not possible. We refer to these states as the states of a blocking sub-space. In these blocking sub-spaces local balance equations also hold, but have different form comparing to the equations of the non-blocking space. Therefore we conclude that the form of local balance equations depends on the total occupied capacity of the base station, which is a key observation for further analysis. This means that the multiplying factors that appear in the product form solution also depend on the occupied capacity.

By examining base stations with infinite capacity we realized that at the non-blocking sub-space the following state transitions hold local balance:

- transitions that result in an increment of the number of users receiving a particular phase of the service time: arrival or phase change of a customer,
- transition that result in the diminution of the number of customers receiving the same phase: one customer leaves the base station by handover or by call termination, or phase change.

Since no transition is possible between phases of service time distributions of different connection types, the above description of local balance equations is formulated for a type k connection that was initiated within the cell. The number of customers receiving different phases of the service time is described by $\underline{n}^{(k),N}$, but for sake of simplicity in this derivation we denote this with $\underline{n}'$. The local balance equations has the form of:

$$\begin{gathered}\lambda_N \alpha_k s_i^{(k),N} p(\underline{n}')+ \\ \textstyle\sum_{j,j\neq i} p(\underline{n}'+\underline{e_j}) S_{ji}^{(k)} \cdot (j+1) = \\ p(\underline{n}'+\underline{e_i})(i+1)\cdot (S_i^{(k),0} + \textstyle\sum_{j,j\neq i} S_{ij}^{(k)}) \quad . \end{gathered} \tag{6}$$

Using the properties of Markov chains $S_i^{(k),0} + \sum_{j,j\neq i} S_{ij}^{(k)} = -S_{ii}^{(k)}$ and writing the equations into vector form, we get:

$$\begin{gathered}-\lambda_N \alpha_k \underline{s}^{(k),N} p(\underline{n}') = \\ [(n_1'+1)p(\underline{n}'+\underline{e_1})\cdots(n_i+1)p(\underline{n}'+\underline{e_i})\cdots \\ (n_P'+1)p(\underline{n}'+\underline{e_P})]\boldsymbol{S}^{(k)} \quad , \end{gathered} \tag{7}$$

where for the sake of simplicity, the number of phases of the service time is denoted by P.

Introducing the vector:

$$\underline{F}^{(k),N} = (\frac{(n_1'+1)p(\underline{n}'+\underline{e_1})}{p(\underline{n}')}, \ldots, \frac{(n_P'+1)p(\underline{n}'+\underline{e_P})}{p(\underline{n}')}),$$

we have

$$\underline{F}^{(k),N} = -\lambda_N \cdot \alpha_k \cdot \underline{s}^{(k),N} \cdot (\boldsymbol{S}^{(k)})^{-1}. \tag{8}$$

The vector defined by (8) would play the role of the multiplying factor in the product form solution if the base station had infinite capacity.

As we described, in blocking sub-spaces the available capacity is small, so not every transmission rate changes can be completed, or some connections cannot be admitted. This means that blocking sub-spaces has less dimensions then the non-blocking space, because the number of customers receiving certain phases can not increase (in these phases customers transmit with too high rate), but customers may arrive into other phases that represent lower transmission rates. In other terms, we can conclude that in blocking sub-spaces the service time changes: those entries of the initial probability vector that correspond to the "unreachable" phases of the service time are set to zero, and also those elements of the rate matrix that represent transition to one of the unavailable phases are also set to zero.

Formulating the change of the service time of a type k new call, when policy 1 is applied and x units of capacity is occupied we get:

$$\boldsymbol{S}_{ij}^{(k)}(x) = 0 \ , \ s_j^{(k),N}(x) = 0 \ , \ \forall j : r_j^{(k)} > C_{k,N} - x. \tag{9}$$

The diagonal elements of the rate matrix must be updated so that $\sum_j \boldsymbol{S}_{ij}^{(k)}(x) + S_i^{(k),0} = 0$. Since blocking sub-spaces are viewed as having less dimensions if those rows and columns of the rate matrix and the entries of the initial probability vector that correspond to unreachable phases are eliminated, the reduced initial probability vector and rate matrix describes the process in that sub-space properly. For handover connections the rate matrix and the initial probability vector is changed analogously.

If the second admission control policy is applied, the rate matrix changes the same way as for policy 1. If $r_i^{(k)}$ denotes the highest transmission rate such that $r_i^{(k)} \leq C_{k,N} - x$ the initial probability vector changes as:

$$s_i^{(k),N}(x) = \sum_{j:\ r_j^{(k)} > C_{k,N}-x} s_j^{(k),N} + s_i^{(k),N}, \tag{10}$$

where $s_j^{(k),N}$ means the jth element of the original initial probability vector.

Since the factor of (8) is calculated from the parameters of the service time distribution, it depends on the occupied capacity as well, therefore (8) gets the general form of

$$\underline{F}^{(k),N}(x) = -\lambda_N \cdot \alpha_k \cdot \underline{s}^{(k),N}(x)(\boldsymbol{S}^{(k)}(x))^{-1}. \tag{11}$$

Kaufman [27] and Roberts [28] proposed a recursive formula to compute channel occupancy distribution in a shared channel. They considered connections with constant capacity requirements. Their method provides exact values in case of a product form Markov chain. The method is based on a one dimensional mapping of a multi-dimensional state space.

As we described, in our problem the local balance holds in the majority of the state space and modified local balance equations hold in blocking sub-spaces as well. If the system is not heavily loaded, the probability mass of the blocking sub-spaces is negligible compared to that of the non-blocking space. Moreover, depending on the parameters of the service time distribution the multiplying factor (11) of blocking sub-spaces does not differ very much from the multiplying factor of the non-blocking space. Considering these we introduce a modified version of the Kaufman-Roberts formula to calculate channel occupancy probabilities. Although this formula does not provide exact solution, as we describe it in Section VI the results have reasonable error.

Following the pattern proposed by Kaufman and Roberts we define $\tilde{p}(m)$ and $v(m)$, the relative and the normalized probability of that m amount of capacity is occupied in equilibrium. $\tilde{p}(m)$ is computed as $\tilde{p}(m) = 0$ for $m < 0$, $\tilde{p}(0) = 1$, and for $m > 0$

$$\tilde{p}(m) = \sum_{k=1}^{K} \sum_i \tilde{p}(m - r_i^{(k)}) \frac{r_i^{(k)}}{m} F_i^{k,N}(m - r_i^{(k)}) + \tilde{p}(m - r_i^{(k)}) \frac{r_i^{(k)}}{m} F_i^{k,H}(m - r_i^{(k)}) \quad (12)$$

and

$$p(m) = \tilde{p}(m) \frac{1}{\sum_{m=0}^{C_0} \tilde{p}(m)} . \quad (13)$$

A. Performance parameters

If the channel occupancy probabilities are given as (12) and (13), the performance parameters of the system are calculated as follows.

The call blocking probability in case of applying policy one for a type k call initiated in the cell is:

$$p_B^{(k),N} = \sum_{i=1}^{N_Q^{(k)}} q_i^{(k)} \cdot \sum_{m=C_{k,N}-c_i^{(k)}}^{C_0} p(m) . \quad (14)$$

The same measure for handover calls is calculated analogously.

If we denote the minimum possible capacity requirement of a type k call with $c_{min}^{(k)}$, the call blocking probability for a type k call initiated in the cell applying the second service policy has the form of:

$$\hat{p}_B^{(k),N} = \sum_{m=C_{k,N}-c_{min}^{(k)}}^{C_0} p(m). \quad (15)$$

The channel utilization is simply given as:

$$\varrho = \sum_{m=0}^{C_0} m \cdot p(m). \quad (16)$$

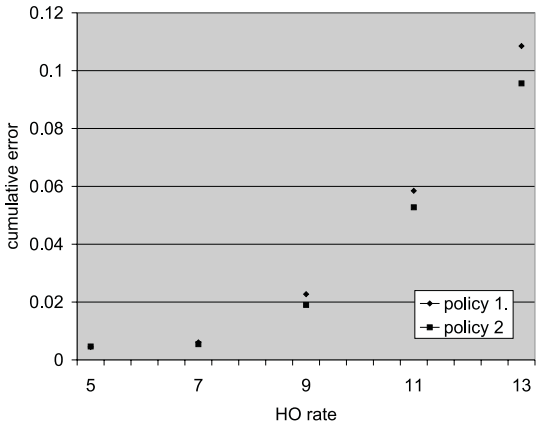

Fig. 1. Cumulative error of the approximation versus arrival rate

VI. Numerical results

A. Accuracy of the proposed approximation

As we described it in the previous section, we intuitively feel that our approximation method based on the Kaufman-Roberts formula gives more accurate results if the system under examination "nearly" has a product form, i.e. if the majority of the probability mass is distributed in the non-blocking part of the state space. This means that if the base station under consideration is slightly loaded, the approximation gives nearly accurate results. Under heavy load conditions the blocking states has higher probabilities, therefore the approximation becomes less accurate. To give insight of this dependency on the load, we examined a base station with $C_0 = 100$ units of capacity and single customer type. The versatility of the generated traffic was fulfilled by means of three possible transmission rates: $1, 2$ and 4 units. The mean dwell time and mean residual call holding time of handover connections were 5.88 and 2.93 minutes respectively, that of connections initiated in the cell were 4.537 and 5 minutes. An amount of 10 units of capacity was reserved for handover connections. In such a system a total arrival rate of 15 calls/minute resulted in a highly overloaded system. Figure 1 shows the accuracy of our approximation compared with simulation as the incoming rate of calls initiated in the cell is constantly 4 per minute and the arrival rate of handover calls rises. The accuracy is plotted for the two base station policies. The measure of accuracy is the cumulative error that is calculated as: $\sum_{m=0}^{C_0} |p_{sim}(m) - p_{app}(m)|$,where $p_{sim}(m)$ and $p_{app}(m)$ are the probability of having m capacity occupied obtained by simulation and the approximate method, respectively.

It is clear from the figure that as the load of the base station increases, the error of the approximation increases dramatically, although even in case of high arrival rate the cumulative error is only about 0.1! Under light load conditions the error of the approximation affects only the third or fourth decimal value of the occupancy probabilities. Thus we have seen, that the approximate method can achieve slight inaccuracy and it needs little computational complexity. In our case the running time of the simulation was 500 times longer then that of the approximate method.

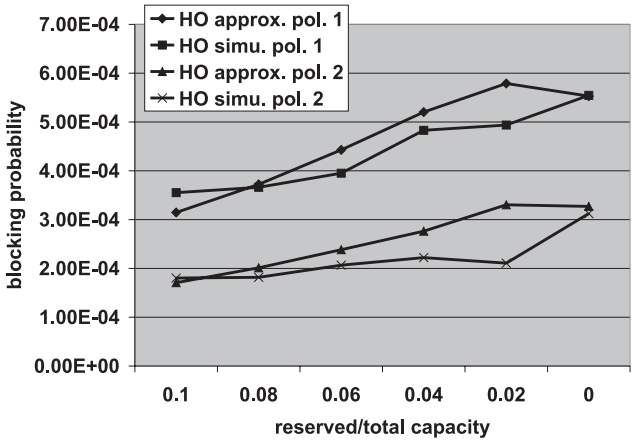

Fig. 2. Handover blocking probabilities

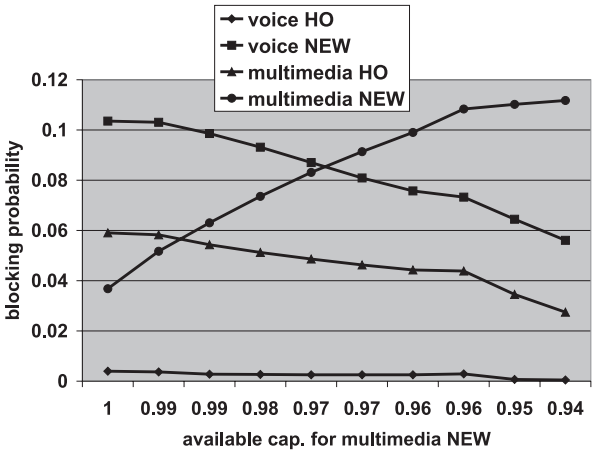

Fig. 4. Blocking probabilities with all types policy 1

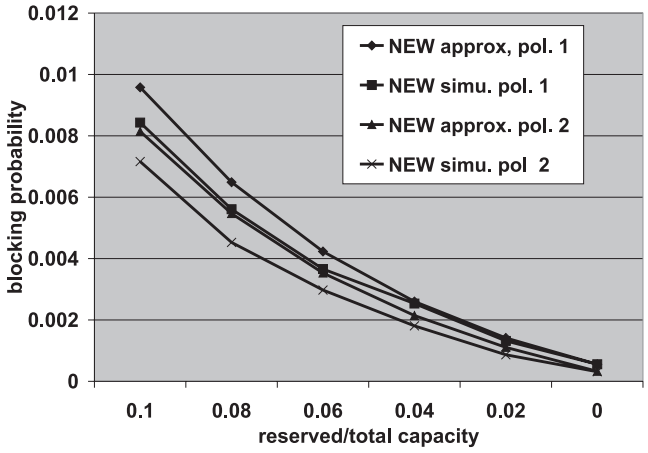

Fig. 3. New call blocking probabilities

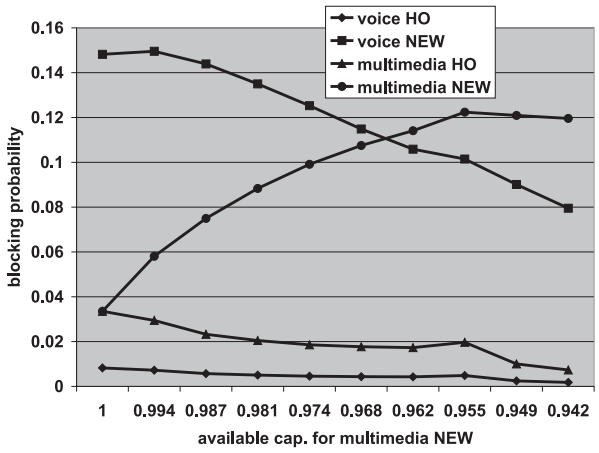

Fig. 5. Blocking probabilities with multimedia HO policy 2

B. Blocking probability

The blocking probabilities were investigated in two scenarios. The first one was described shortly in the previous subsection. In the second scenario a base station was considered with 20 Mbps transmission capacity. This is equal to the capacity of the wireless ATM testbed called WAND [1]. Two connection types were present in the system: voice calls with constant 32 kbps rate and multimedia connections. The latter had 128 kbps minimum, 256 kbps average and 1024 kbps peak transmission rate. The dwell time of the customers was modeled with generalized gamma distribution, according to [18] and [19]. Phase type distributions with six phases were fitted to the dwell time distributions with the method described in [23].

The first scenario was examined by the approximate method and computer simulations as well. Figure 2 and Figure 3 shows the blocking probabilities of handover and new connections as the amount of capacity preserved for handover calls decreases, for both admission control policies. On the legends of the figures *HO* refers to handover calls and *NEW* refers to connections initiated within the cell. Clearly, the blocking probability of new calls decreases as we decrease the capacity preserved for handover calls. On the other hand, the handover blocking probability increases. The decline of the new call blocking probability is more significant than the rise of the handover call blocking probability.By examining the two graphs we may conclude that results achieved by the approximate method are reasonably accurate. Applying policy 2 results in a reasonable decline in the blocking probabilities.

Figure 4 and 5 show the results of the second examined scenario. Since the accuracy of our recursive method was verified earlier this scenario was investigated only with the new approximate method. In this case we supposed that new voice calls are rejected if the total occupied capacity is greater than 19.2 Mbps. Figure 4 show the blocking probabilities when immediate call blocking (policy 1) is applied in the network and the maximum available capacity for multimedia new connections is decreasing. The available capacity of multimedia new connections is given as the proportion of the total capacity. The decline of the maximum available capacity results in a steep rise of the blocking probability of multimedia new calls along with a significant decline of the blocking probability of all other types.

Figure 5 shows the results when for multimedia handover connections policy 2 is applied. The effect of this policy is clear. It significantly reduces the blocking probability of multimedia handover calls, but slightly increases the blocking probability of

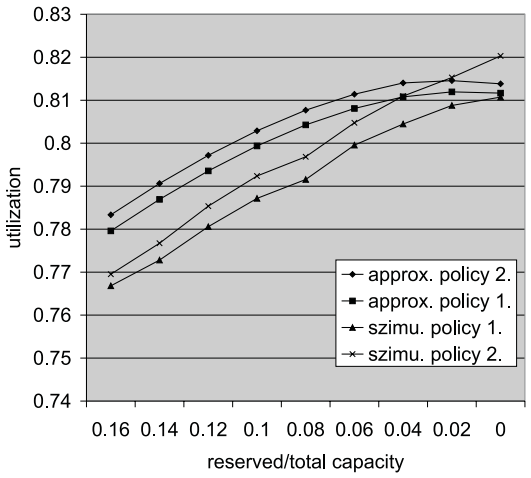

Fig. 6. Utilization of heavily loaded base station

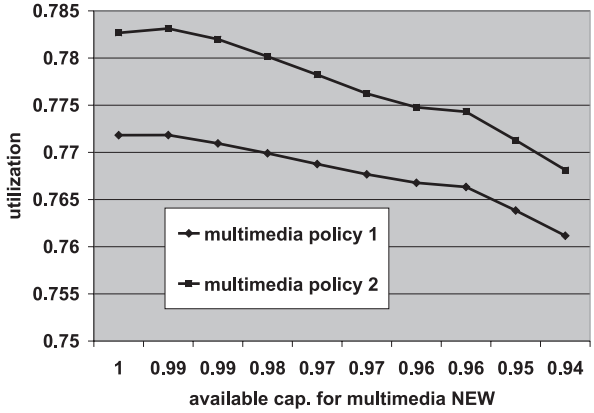

Fig. 7. Channel utilization

other types. This is because more multimedia connections can admit to the base station thus the probability of occupying more capacity is bigger.

C. Channel utilization

The channel utilization of the first scenario is plotted in Figure 6. If the system is lightly loaded the utilization barely changes as the capacity preserved for handover decreases, neither the applied policy affects the utilization significantly. Therefore the utilization is shown in case of heavy load. By observing Figure 6 we may conclude that applying policy 2 slightly raises the utilization. The utilization is growing if the amount of preserved capacity for handover connections is declining. Since in this case the base station is heavily loaded the approximate method is less accurate, however the difference between the utilization achieved by simulation and by the approximate method is about 0.01 which is negligible.

Figure 7 shows the channel utilization of the second scenario. Since the reduction of the available capacity for multimedia new calls mean that fewer connections are admitted to the base station, the utilization slightly decreases. However this shrinkage is about 0.015 that is negligible. If policy 2 is applied for multimedia handover connections the utilization is better (since fewer calls are dropped), but this achievement is again not significant.

VII. Conclusions and future works

In this paper we have investigated a modeling method to analyze a single base station with multiclass multimedia connections. The proposed model is very general and scalable due to the flexibility that comes with the use of PH distributions. We also proposed and investigated two simple admission control policies. The analytical calculation presented here is based on the recursive Kaufman-Roberts formula. Although the results obtained using our method are not accurate, we have proven that the error of the approximation is negligible, especially under light load circumstances. We investigated the effect of the applied admission control policy and reserved bandwidth. The applied policy does not affect the channel utilization very much, but applying policy 2 reduces call blocking probabilities. In multiclass environment reducing the available capacity for a connection type rapidly increases its blocking probability, but only slightly decreases that of other types.

The proposed method here is also suitable for investigating the performance of several dynamic capacity allocation methods, when the available capacity for different user types is not constant but depend on several parameters of the connection types. Currently a method is being elaborated to obtain the optimal value of the amount of available capacities and applied admission control policies to optimize the overall system performance.

References

[1] Mikkonen, J.; Aldis, J.; Awater, G.; Lunn, A.; Hutchinson D.; The Magic WAND - Functional Overview. *IEEE Journal on Selected Areas in Communications*, Vol. 16, No. 6, Aug. 1998, pp. 953-972

[2] Raychaudhuri, D. *et. al.* WATMnet: A Prototype Wireless ATM System for Multimedia Personal Communication. *IEEE Journal on Selected Areas in Communications*, Vol. 15, No. 1, Jan. 1997, pp. 83-95

[3] Eng, K.Y. *et.al* A wireless broadband ad-hoc ATM local-area network *Wireless Networks* 1995 pp.161-173

[4] http://www.cselt.it/sonah/AWACS

[5] http://www.hiperlan2.com

[6] Sriram, K. Performance of ATM and variable length packet access in broadband HFC and wireless networks. *Universal Personal Communications*, ICUPC '98 IEEE 1998, pp. 495-501

[7] Lixin, Wang; Hamdi, M. HAMAC: an adaptive channel access protocol for multimedia wireless networks. *Computer Communications and Networks*, 1998. Proceedings. 7th International Conference on pp. 404 -411

[8] Passas, N.; Paskalis, S.; Vali, D.; Merakos, L. Quality-of-Service Oriented Medium Access Control for Wireless ATM Networks. *IEEE Communications Magazine*, Nov. 1997, pp. 42-50

[9] Zhang, Z.; Habib, I.; Saadawi, T. A bandwidth reservation multiple access protocol for wireless ATM local networks, *Military Communications Conference, 1996. MILCOM '96*, Conference Proceedings, IEEE, Volume: 3, 1996 Page(s): 954 -958

[10] Horikawa, H.; Inoue, M.; Hatori, M.; Mizumachi, M. Prioritized wireless access protocols for real-time VBR traffic. *Personal, Indoor and Mobile Radio Communications*, 1996. PIMRC'96., Seventh IEEE International Symposium on , Vol. 3, pp. 918 -922

[11] M. Ajmone Marsan, S. Marano, C. Mastroianni, M. Meo, Performance Analysis of Cellular Mobile Communication Networks Supporting Multimedia Services, in *Proceedings of MASCOTS'98*, July 19-24,1998, Montreal, Canada

[12] S. Greiner, G. Bolch, K. Begain, A Generalized Analysis Technique for Queuing Networks with Mixed Priority Strategy and Class Switching. *Computer Communications*, vol. 21, pp 819-832, 1998.

[13] M. Ajmone Marsan, S. Marano, C. Mastroianni and M. Meo, Performance analysis of cellular mobile communication networks supporting multimedia services, *Mobile Networks and Applications*, Volume 5 , Issue 3 (2000) Pages 167-177

[14] J. G. Markoulidakis, G. L. Lyberopoulos and M. E. Anagnostou, Traffic model for third generation cellular mobile telecommunication systems *Wireless Networks*, Volume 4, Issue 5 (1998) Pages 389-400

[15] D. Hong and S.S. Rappaport, Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures, *IEEE Transactions on Vehicular Technology*, 35(3) August, 1986 Pages 77-92

[16] Orlik, P.V.; Rappaport, S.S. Traffic Performance and Mobility Modeling of Cellular Communications with Mixed Platforms and Highly Variable Mobilities. *Proceedings of the IEEE*, Vol. 86, No. 7, July 1998, pp. 1464-1479

[17] Orlik, P.V.; Rappaport, S.S. A model for teletraffic performance and channel holding time characterization in wireless cellular communication with general session and dwell time distributions *Selected Areas in Communications*, IEEE Journal on , vol 16, Issue: 5, June 1998, pp. 788 -803

[18] Zonoozi, M.; Dassanayake, P. User Mobility Modeling and Characterization of Mobility Patterns. *IEEE Journal on Selected Areas in Communications*, Vol. 15, No. 7, Sep. 1997, pp. 1239-1252

[19] Zonoozi, M.M.; Dassanayake, P.; Faulkner, M. Mobility modelling and channel holding time distribution in cellular mobile communication systems. *Global Telecommunications Conference*, 1995. GLOBECOM '95., IEEE , Vol. 1 , 1995 pp. 12 -16

[20] M. Neuts. Probability distributions of Phase Type. In *Liber Amicorum Prof. Emeritus H. Florin*, pages 173–206. University of Louvain, 1975.

[21] A. Bobbio and M. Telek, "A benchmark for PH estimation algorithms: results for Acyclic-PH," *Stochastic Models*, vol. 10, pp. 661–677, 1994.

[22] A. Horváth and M. Telek, Approximating heavy tailed behaviour with phase type distributions, *3rd International Conference on Matrix-Analitic Methods in Stochastic models*, MAM3, Leuven, Belgium, *Notable Publications Inc.*, 2000, pp. 191-214

[23] S. Asmussen, O. Nerman & M. Olsson, Fitting phase-type distribution via the EM algorithm, *Scand. J. Statist*. 23, 419-441 (1996)

[24] Sen, P.; Maglaris, B.; Rikli, N.-E.; Anastassiou, D. Models for packet switching of variable-bit-rate video sources. *IEEE Journal on Selected Areas in Communications*, Vol. 7 No. 5, June 1989 pp. 865 -869

[25] Special Issue on Packet Speech and Packet Video *IEEE Journal on Selected Areas in Communications*, Vol. 7, No. 5, 1989

[26] Elwalid, A. I.; Mitra, D. Effective Bandwith of General Markovian traffic Sources and Admission Control of High Speed Networks *IEEE Transaction on Networks*, Vol. 1, No. 3, July 1993 pp. 329-343

[27] Kaufman, J. Blocking in a Shared Resource Environment *IEEE Transactions on Communications*, Vol.Com-29, No. 10, Oct. 1981 pp. 1474-1481

[28] Roberts, K. *Performance of Data Communication Systems and their Applications.* North-Holland-Elsevier Science Publishers, 1981