Modeling and Analysis of Broadband Cellular Networks with Multimedia Connections

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
Tel:(36) 1 463-3227, Fax:(36) 1 463-3263
E-mail: fazekasp@hit.bme.hu

Abstract

As mobile networking is moving towards the direction of providing high-speed multimedia services, the presence of connections that do not generate traffic at constant rate is becoming reality in wireless environment. For preliminary network design purposes analytical models are necessary that examine the performance or wireless net-works with such connections. In this paper an analytical method is presented to calculate call-level system parameters of cellular networks with multimedia connections. The method is based on an approximate recursive calculation. The accuracy of the approximation is verified by computer simulations. The derivation of the distributions of user describing time variables and two simple admission control policies are also presented and investigated.

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

Introduction

One of the major directions of today's telecommunication research and industry is the integration of high speed multimedia transmission with mobile access. This movement is based on the exponentially growing number of mobile telephone subscribers and internet users, along with the rapidly increasing computational capacity of portable devices.

The evolution of second generation mobile systems (GSM, IS95, etc.) towards the third generation (UMTS, IMT2000 family) has the aim of satisfying the increasing demand of using integrated network services with wireless mobile access. As 3G systems are going to begin operation, packet switched data applications and other multimedia services will spread instead of simple voice connections. However, third generation systems will not provide as high transmission capacity as wired LANs today.

Besides the development of 3G systems enormous research is being carried in the field of wireless LANs and wireless systems that provide several tens of Mbps transmission rate. Currently the aim of these systems is not to provide large scale coverage but to assure high speed mobile access in certain locations (university campuses, offices, etc.). The IEEE 802.11 wireless LAN has the capacity of 5 Mbps, Hiperlan/2 [5] networks can carry up to 56 Mbps at the physical layer. During the 90's several wireless ATM testbeds were built [1]-[4]. These have transmission rates from 2 Mbps up to 20 Mbps.

As we have seen the networking technology of the near future allows the use of any type of application (web browsing, video telephony, file transfer, etc.) in wireless environment. However, multimedia applications often do not generate traffic at constant rate. To utilize the radio channel efficiently it is not worth allocating capacity for a connection according to its peak transmission rate, rather a medium access method is necessary that is able to provide resources according to the connection's instantaneous requirement.

At first sight such a medium access method may not seem trivial, however such protocols are widely proposed in the literature. [6]-[10]. The most plausible approach is to apply TDMA/TDD MAC protocols. The TDMA channel is reasonable to serve packet communications with fixed packet lengths. The TDD approach allows the protocol to adaptively change the uplink/downlink capacity according to traffic parameters, by simply changing the length of the uplink/downlink phase of the MAC frame. Every TDMA wireless MAC frame should contain a contention phase (ALOHA type), where new connections can be set-up. The variability of a connection's used amount of capacity is fulfilled by means of the variable number of time slots it may use in different time frames. A scheduler is placed in the base station that allocates time slots to users, according to their instantaneous need. Determination of the number of needed time slots may be performed at the mobile, than this case this number is submitted to the base station via either a control channel or in-band or via the contention period. If the base station determines the number of assigned time slots to a user it may use previously negotiated information about the user or the customer may submit state information (buffer saturation, instantaneous bit-rate, etc.) according to which the scheduler distributes the time slots among users.

We may conclude that in broadband cellular networks connections may appear that occupy variable amount of radio resources during a connection. Although numerous papers are devoted to the evaluation of several performance parameters of cellular networks with multiple service classes [11]-[17], these models do not consider the versatility of a connection's required capacity. In this paper we present an analytical modeling method that takes into account that a connection's required capacity is changing during the session. The model presented here does not assume any particular networking technology, therefore is useful for investigating Wireless ATM networks as well. The existence of VBR and ABR/UBR service classes of ATM also necessitates such a model, where the amount of radio capacity assigned to a user is changing during the connection's lifetime.

In analytical models of cellular networks generally three random time variables are used to characterize user behavior. The dwell time is the time a mobile spends in a cell while it is roaming throughout the area. This time is also often referred as cell residence time or sojourn time. It is common to model the dwell time as a random variable with negative exponential distribution, although this assumption applies only under special circumstances. Other authors suggested different dwell time distributions. In their often referred paper [15] Hong and Rappaport proposed a dwell time distribution dependent on the speed of the mobile and the radius of the cell. In [16] and [17] sum of hyperexponential distributions

(SOHYP) was used to model the dwell time. Lin et al. used general distributions as cell residence time distributions. In [19] and [20] the authors shown that the cell residence time is properly described by generalized gamma distribution. In this paper we allow the dwell time to have arbitrary distribution, or the knowledge of the residence time distribution is not necessary at all, rather measurement data on this time is sufficient.

The second time variable to describe customer behavior is the connection duration. It is common in teletraffic modeling theory to assume that this time follows a negative exponential distribution. However, this assumption only holds for voice connections, but other connection types do not necessarily have this property. In this study we assume that the call holding time also has arbitrary distribution, or only measurement data is available on this time.

The third user describing time is the channel holding time. This is the time a customer spends communicating in the cell. This is usually derived from the dwell time and the call holding time.

This paper is organized as follows. In Section 1 the modeling environment is described, along with the mobile describing time variables, the traffic model of the connections and the base station model. Section 2 contains the derivation of the service time distribution for the abstract model of the modeled system. In Section 3 the Markov model of a single base station is investigated. We present the approximate method for calculating system parameters in Section 4. This is followed by numerical results and conclusions.

1 Characterization of the modeling environment

The model proposed in this paper concentrates on a single base station. The aim is to determine call level system parameters: new call blocking probabilities, handover failure probabilities and channel utilization.

A connection can admit to the base station as a new call within the coverage area of the base station or as a handover attempt from the neighborhood of the cell. In the rest of the paper the first is referred as new call, the second type is the handover call. We suppose that the number of users roaming in the cell and in the vicinity of the cell is large and they initiate new call or handover independently and with low probabilities, therefore the incoming flow of new calls and handover calls form two independent Poisson processes with rates of 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 there are K several user types present in the system. Customers belonging to different types may belong to different service classes, in this case the characteristics of their generated traffic and their session lengths are different. Users of different types may also differ by means of their dwell times, i.e. their mobility behavior. Thus customers from the same service class may belong to different user types.

1.1 User describing times

A customer of type k is described by its random session length τ_L^k and its dwell time τ_D^k. We suppose that the customer's connection length is independent of its mobility behavior, therefore τ_L^k and τ_D^k are independent random variables.

In this paper we suppose that τ_L^k and τ_D^k have arbitrary distributions or given by statistical data. Then we suppose that a Phase Type distribution (PH)[21] is fitted to the two distributions. Then these fitted PHs are used in the analysis as the dwell time and the 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 rate matrix of the Markov chain has the following structure:

$$\begin{vmatrix} \boldsymbol{T} & \underline{\mathrm{T}}^0 \\ 0...0 & 0 \end{vmatrix} \tag{1}$$

The distribution is determined by the initial probability vector $\underline{t}$ and the infinitesimal generator matrix $\boldsymbol{T}$. 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.

The distribution function and the pdf of a PH with the above parameters are:

$$F(x) = 1 - \underline{t}e^{\boldsymbol{T}x}\underline{h}, \quad f(x) = \underline{t}e^{\boldsymbol{T}x}\underline{T}^0, \tag{2}$$

where $\underline{h}$ is a column vector with 1 in its each position.

Studies show ([22]-[24]) that most distributions can be accurately approximated by an appropriately chosen Phase Type distribution. Moreover, if measurement data is available for 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.

To give insight in the accuracy of the PH approximation we fitted a PH with 12 phases by the method presented in [24]. The original distribution was the generalized gamma distribution, with pdf:

$$f(t; a, b, c) = \frac{c}{b^{ac}\Gamma(a)} t^{ac-1} e^{-(\frac{t}{b})^c},$$

where $\Gamma(a)$ is the gamma function defined as $\Gamma(a) = \int_0^\infty y^{a-1}e^{-y}dy$. In [19] and [20] the authors have shown that the dwell time can be properly modeled by this distribution. Figure 1 and 2 show the distribution functions and the densities of the generalized gamma distribution and the fitted PH distribution. The results prove that the fitted PH approximates the original distribution with reasonable error.

The use of PH distributions has two advantages. One is that it can approximate any other distribution, therefore in analytical models PH distributions can substitute other distributions that appear in any problem. The reason why the use of a fitted PH is favored instead of the original distribution is that the PH consists of exponential phases. Although the PH itself does not have the memoryless property of the exponential distribution, its phases does. Thus standard queuing theory techniques can be used for examining systems with PH service times, at the expense of an extended state space, since each phase of the PH has to be tracked during the analysis.

Thus we assume that after a PH fitting procedure the session length distribution of a type k connection is given as a PH with parameters $\underline{l}^{(k)}$, $\boldsymbol{L}^{(k)}$ and $\underline{L}^{(k),0}$, the dwell time is also given as a PH distribution with parameters $\underline{d}^{(k)}$, $\boldsymbol{D}^{(k)}$ and $\underline{D}^{(k),0}$.

To determine the channel holding time of a connection the dwell time and the call holding time are not sufficient, rather we introduce two new time variables.

1.2 Residual dwell time and session length

Let $\tau^k_{L,R}$ denote the residual lifetime of the session length of a type k connection after it enters the cell. This time variable begins at the instant of admission to the base station and lasts until the connection is terminated in the cell or in another cell after several more handovers. Obviously, for new connections the residual session length is equal to the session length, i.e. for new calls $\tau^k_{L,R} = \tau^k_L$. Since handover calls were initiated in another cell, some time has already elapsed between the call initiation and the instant of handover. If the session length has a negative exponential distribution, the residual session length of handover connections also has the same distribution, due to the memoryless property of the exponential distribution. Here the session length is not exponentially distributed, therefore the distribution of the residual session length of handover connections is not the same as the distribution of the session length. The determination of the residual session length of handover connections is completely out of the scope of this paper, however an analytical method is currently being elaborated by the authors to determine the distribution of the residual session length with the knowledge of the session length itself and the dwell time distribution. Here we assume that the residual session length distribution is known by primarily approximate user analysis or statistics and a PH fitting is done to achieve the PH distributed residual session length.

The residual dwell time of a type k connection ($\tau^k_{D,R}$) is the time that lasts after the connection admits to the base station until the mobile leaves the cell, regardless it has terminated its call or not. It is obvious from the definition that for handover calls the residual dwell time is equal to the dwell time, i.e $\tau^k_{D,R} = \tau^k_D$. If the dwell time is modeled with exponential distribution, the residual dwell time of new calls also has the same exponential distribution, due to the memoryless property of the exponential distribution. Here we assume PH dwell time distributions, therefore this equation generally does not hold.

Authors of papers devoted to the derivation of the dwell time distribution usually do not use the notion residual dwell time distribution, rather generally two different dwell time distributions are investigated for new calls and for handover calls. Regardless the notion, two dwell times are necessary to describe new connections and handover connections of the same type. In this case two PH fitting procedures are performed to obtain the two PH dwell times.

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.

If $\tau_{D,R}^k$ and $\tau_{L,R}^k$ is given, the channel holding time τ_C^k is determined from the above two times. If the mobile leaves the cell by handing over to another base station, its channel holding time is the residual dwell time, if it terminates its call before handover, the channel holding time is the residual session length, i.e. the channel holding time is the shorter time of the above two:

$$\tau_C^k = min(\tau_{D,R}^k, \tau_{L,R}^k). \quad (3)$$

1.3 Traffic model

A customer's generated data during a connection is characterized 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. 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 multimedia transmission) but it is widely used in the literature to describe several services

[25]-[27].

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.

1.4 Base station model

The base station is modeled as a channel pool of C_0 units of capacity. This capacity is expressed in the same units as the transmission rates of the connections. We assume that different connection types tolerate call blocking differently. Moreover, generally blocking on-going handover calls is less tolerable than the blocking of new calls. Therefore we suppose that the base station preserves some capacity for blocking sensitive connections and for handover connections. Namely there is a maximum available capacity $C_{k,H}$ and $C_{k,N}$ for type k handover and new connections. A type k connection can be admitted if the total occupied capacity after admission is less then or equal to $C_{k,H}$ or $C_{k,N}$.

Let C_{oc} denote the amount of occupied capacity at the base station and $\underline{c}^{(k)}$ denote the vector containing the possible transmission rates of a type k connection. 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.

The applied admission control policy may be different for several connection types and for handover and new connections. 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.

Because the connections change their transmission rate during the session, the amount of occupied capacity at the base station may change without the arrival or termination of a connection. In the case when a connection tries to switch to a higher transmission rate but there is not enough idle capacity, we assume that the session is not terminated but the transmission continues with the same rate.

2 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, i.e. $min(\tau_S^k, \tau_{D,R}^k)$ for new calls and $min(\tau_{L,R}^k, \tau_D^k)$ for handover connections,

- it contains the instantaneous transmission rate of the connections.

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_S^k. Then N_D^k group is formed, each group containing all the N_S^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 $i, j = 1, \dots N_S^k, n, m = 1, \dots, N_D^k$.

An example of composing the channel holding time distribution is shown in Figure 3. In this example both the residual dwell time and the residual session length has a two phase acyclic PH (often referred as Coxian) distribution.

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:

$$\boldsymbol{T}^{(k)} = \boldsymbol{D}^{(\boldsymbol{k})} \oplus \boldsymbol{L}^{(k)}, \quad \underline{T}^{(k),0} = \underline{D}^{(k),0} \oplus \underline{L}^{(k),0},$$
$$\underline{t}^{(k),N} = \underline{d}_R^{(k)} \otimes \underline{l}^{(k)}, \tag{4}$$
$$\underline{t}^{(k),H} = \underline{d}^{(k)} \otimes \underline{l}_R^{(k)}$$

where $\oplus$ and $\otimes$ denotes the Kronecker sum and product, respectively. It is shown in the Appendix that a PH composed in the described manner, with the parameters of (4) has the distribution of the minimum of the two PHs it was composed from. It is obvious that the channel holding time has $N_T^{(k)} = N_D^{(k)} \cdot N_S^{(k)}$ 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. An example of constructing the service time distribution is shown in Figure 4. Here, for the sake of simplicity the channel holding time has only two phases and the number of possible transmission rates is only two.

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

$$\begin{aligned} \boldsymbol{S}^{(k)} = \boldsymbol{Q}^{(k)} \oplus \boldsymbol{T}^{(k)}, \quad \underline{S}^{(k),0} = \underline{h}_{N_Q^{(k)}} \otimes \underline{T}^{(k),0}, \\ \underline{s}^{(k),N} = \underline{q}^{(k),N} \otimes \underline{t}^{(k)}, \\ \underline{s}^{(k),H} = \underline{q}^{(k),H} \otimes \underline{t}^{(k),H} \end{aligned} \tag{5}$$

where $\boldsymbol{S}^{(k)}$, $\underline{S}^{(k),0}$ and $\underline{s}^{(k),N}$ are the parameters of the PH service time distribution for new calls, $N_Q^{(k)}$ denotes the number of transmission rates of a type k connection.

In the Appendix we show, that the PH distribution composed from the channel holding time and the traffic describing Markov chain 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 $(m-1) \cdot N_L^{(k)} * N_D^{(k)} + (j-1) \cdot N_L^{(k)} + i$ 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.

In (4) and (5) the superscript H and N differentiates between the initial probability vectors of the channel holding time and the service time of handover and new calls. Obviously if two different dwell times are given for handover and new calls, the rate matrices of the channel holding time and the service time are also different.

Because handover calls and new calls of type k are different by means of their residual session length and residual dwell time, 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.

3 The driving process

Given the incoming process is 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 $\underline{n} = [\underline{n}^{1,N}, \ldots, \underline{n}^{K,N}, \underline{n}^{1,H}, \ldots, \underline{n}^{K,H}]$, 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

$$\begin{gathered} r_1^{(k)} = r_2^{(k)} = ... = r_{N_L^{(k)} N_D^{(k)}} = c_1^{(k)}; \\ r_{N_L^{(k)} N_D^{(k)}+1}^{(k)} = r_{N_L^{(k)} N_D^{(k)}+2}^{(k)} = ... = r_{2N_L^{(k)} N_D^{(k)}}^{(k)} = c_2^{(k)} \\ ; ...; \\ r_{(N_Q^{(k)}-1)N_L^{(k)} N_D^{(k)}+1} = ... = r_{N_Q^{(k)} N_L^{(k)} N_D^{(k)}} = c_{N_Q^{(k)}}^{(k)} . \end{gathered} \tag{6}$$

The valid states of the system are those, where

$$\begin{gathered} \underline{n}^{k,N} \cdot \underline{r}^{(k)} \le C_{k,N}, \quad \underline{n}^{k,H} \cdot \underline{r}^{(k)} \le C_{k,H} \quad k = 1...K \\ \textstyle\sum_{k=1}^{K} \underline{n}^{k,N} \cdot \underline{r}^{(k)} + \sum_{k=1}^{K} \underline{n}^{k,H} \cdot \underline{r}^{(k)} \le C_0. \end{gathered} \tag{7}$$

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.

Assuming the phase type structure defined by $\underline{s}^{k,N}$ and $\boldsymbol{S}^{(k)}$ is not redundant the Markov chain describing the variation of system states is finite and irreducible, hence its steady state distribution exists.

3.1 State transitions

To calculate performance parameters of the presented system, the steady state distribution of the resulting Markov process is needed. To calculate the steady state distribution, the knowledge of the possible state transitions and the rates of these transitions is necessary.

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 $N_D^{(k)} \times N_L^{(k)} \times N_Q^{(k)}$ 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:

$$\lambda_N \cdot \alpha_k \cdot \sum_{j: r_j^{(k)} \geq C_{k,N} - C_{oc}} s_j^{(k),N} \;,\; \lambda_H \cdot \alpha_k \cdot \sum_{j: r_j^{(k)} \geq C_{k,H} - C_{oc}} s_j^{(k),H} \qquad (8)$$

  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, $m = 1 ... C_0$.

4 Calculating channel occupancy probabilities

Here we propose an approximate method to calculate channel occupancy probabilities. This method has negligible computational 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:

$$\lambda_N \alpha_k s_i^{(k),N} p(\underline{n}') + \sum_{j,j\neq i} p(\underline{n}' + \underline{e_j}) S_{ji}^{(k)} \cdot (j+1) =$$

$$p(\underline{n}' + \underline{e_i})(i+1) \cdot (S_i^{(k),0} + \sum_{j,j\neq i} S_{ij}^{(k)}) \qquad . \tag{9}$$

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:

$$-\lambda_N \alpha_k \underline{s}^{(k),N} p(\underline{n}') = [(n'_1 + 1) p(\underline{n}' + \underline{e_1}) \cdots (n_i + 1) p(\underline{n}' + \underline{e_i}) \cdots (n'_P + 1) p(\underline{n}' + \underline{e_P})] \boldsymbol{S}^{(k)} \quad ,$$

where for the sake of simplicity, the number of phases of the service time ($N_D^{(k)} \cdot N_L^{(k)} \cdot N_Q^{(k)}$) is denoted by P.

Introducing the vector:

$$\underline{F}^{(k),N} = (\frac{(n'_1 + 1) p(\underline{n}' + \underline{e_1})}{p(\underline{n}')}, ..., \frac{(n'_P + 1) p(\underline{n}' + \underline{e_P})}{p(\underline{n}')}),$$

we have

$$\underline{F}^{(k),N} = -\lambda_N \cdot \alpha_k \cdot \underline{s}^{(k),N} \cdot (\boldsymbol{S}^{(k)})^{-1}. \tag{10}$$

The vector defined by (10) 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:

$$\boldsymbol{S}_{ij}^{(k)}(x) = 0 \ , \ \ s_j^{(k),N}(x) = 0 \ , \ \ \forall j : r_j^{(k)} > C_{k,N} - x. \tag{11}$$

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:

$$s_i^{(k),N}(x) = \sum_{j:\, r_j^{(k)} > C_{k,N} - x} s_j^{(k),N} + s_i^{(k),N}, \tag{12}$$

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

Since the factor of (10) is calculated from the parameters of the service time distribution, it depends on the occupied capacity as well, therefore (10) gets the general form of

$$\underline{F}^{(k),N}(x) = -\lambda_N \cdot \alpha_k \cdot \underline{s}^{(k),N}(x)(\boldsymbol{S}^{(k)}(x))^{-1}. \tag{13}$$

Kaufman [28] and Roberts [29] 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 (13) 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 5 the results have reasonable error.

Following the pattern proposed by Kaufman and Roberts we define $\tilde{p}(m)$ and $p(m)$, 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 $m > 0$

$$\tilde{p}(m) = \sum_{k=1}^{K} \sum_i \tilde{p}(m - r_i^{(k)}) \frac{r_i^{(k)}}{m} F_i^{(k),N}(C_0 - m + r_i^{(k)}) + \tilde{p}(m - r_i^{(k)}) \frac{r_i^{(k)}}{m} F_i^{(k),H}(m - r_i^{(k)})$$

and

$$p(m) = \tilde{p}(m) \frac{1}{\sum_{m=0}^{C_0} \tilde{p}(m)} \ . \tag{14}$$

4.1 Performance parameters

If the channel occupancy probabilities are given as (14) and (14), 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:

$$p_B^{(k),N} = \alpha_k \cdot \sum_{i=1}^{N_Q^{(k)}} \sum_{m=C_{k,N}-c_i^{(k)}}^{C_0} p(m) \ . \tag{15}$$

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:

$$\hat{p}_B^{(k),N} = \alpha_k \cdot \sum_{m=C_{k,N}-c_{min}^{(k)}}^{C_0} p(m). \tag{16}$$

The channel utilization is simply given as:

$$\varrho = \sum_{m=0}^{C_0} m \cdot p(m). \tag{17}$$

5 Numerical results

5.1 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 VBR nature of the connections 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 5 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 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.

5.2 Effect of the reserved capacity and the base station policy

The accuracy of the approximation depends on the available capacity as well. Therefore with the same arrival rate different accuracy can be achieved by varying the amount of reserved capacity for handover calls.

Figure 6 and Figure 7 show the effect of the amount of reserved capacity on the utilization of the radio channel under heavy and lighter load conditions. In the first case the total arrival rate was 17 calls per minute, in the second case 12. Trivially the utilization increases as the amount of capacity is decreased. In case of high arrival rate this increment of the utilization is more significant, since the probability of having larger capacity is higher and with small reserved capacity new calls can occupy nearly all the channel. In case of light load the utilization increases very slightly, this is because the probability of occupying large capacity is low therefore the amount of reserved capacity does not effect the admission of new calls. It is clear from both graphs that the base station policy of handling connections that arrive with too big capacity demand does not effect the utilization very much, even in the case of heavy load the biggest difference between the two utilizations is about 0.015. This also holds for the accuracy of the approximate method: the results obtained by simulation and the approximate formula diverges with only 0.01, that is a very reasonable error considering channel utilization. In case of light load the curves of utilizations are hard to distinguish, nor the base station policy nor the calculation of the utilization (approximate or simulation) affect the utilization very much.

Figure 8 and 9 show the blocking probabilities of new and handover calls for the two base station policies. On the legends of the figures *HO* refers to handover calls and *NEW* refers to connections initiated within the cell. It is clear that in both cases the blocking probability of new calls decreases as the amount of reserved capacity is decreased, while the blocking probability of handover connections increases. Applying policy 2 results in a reasonable decline in the blocking probabilities. Since as we have seen the policy does not effect the utilization remarkably, using the second policy increases the overall system performance. It is also clear that with not very high arrival rate it makes no sense reserving two big amount of capacity for handover connections since this results in a steep rise of the new call blocking probability.

5.3 Effect of the reserved capacity and the base station policy with three service classes

The following results were achieved in a system with 20 Mbps channel capacity (this is the capacity of the experimental WATM network called WAND [1]). Because the accuracy of the proposed approximate method has been verified, these results were calculated with the approximate method only. Three types of connections were supposed: voice calls with 32 kbps transmission speed, blocking insensitive multimedia connections with three possible transmission speeds (64 kbps lowest speed, 256 kbps average and 512 kbps peak rate) and blocking sensitive multimedia calls (128 kbps lowest, 256 kbps average and 386 kbps peak rate). The distinction blocking sensitive and insensitive means that by varying the amount of available capacity for a connection type and the applied admission control policy, we would like to reduce the blocking probability for sensitive sessions and also we do not tolerate the growth in the blocking probability of handover calls of any type.

To examine the effect of available capacity for different user types we lowered the amount of available capacity for insensitive new calls, while the available capacity for new voice calls was 19.6 Mbps. For all connection types policy 1, i.e. immediate blocking was applied in the case when a call arrives with too big instantaneous capacity demand. Figure 10 shows the blocking probabilities of all connection types as the amount of available capacity for insensitive new calls decreases. The available capacity for the blocking insensitive connection type is given as the proportion of the total capacity. On this figure type 1 refers to blocking insensitive, type 2 to blocking sensitive calls. As we decrease the available capacity for insensitive new calls its blocking probability dramatically increases, however it results in a slight decline of the blocking probabilities of all other types. Figure 11 shows the blocking probabilities in the same situation, the only difference is that for sensitive handover calls policy 2 is applied. As it is clear from the graph applying policy 2 results in a 0.01 decline of the blocking probability for this type.

Figure 12 shows the channel utilization for the above two scenarios. Since the reduction of the available capacity for insensitive new calls mean that fewer connections are admitted to the base station, the utilizations slightly decrease. However this shrinkage is less then 0.007 that is negligible. If policy 2 is applied for sensitive handover connections the utilization is better (since fewer calls are dropped), but this achievement is again not significant (about 0.3 percent).

6 Conclusions and future works

In this paper we have investigated a modeling method to analyze a single base station with multiclass multimedia connections. We proposed a method to determine the residual session length of handover connections, if the session length is modeled with PH distribution, that do not have the memoryless property. We also proposed a method to calculate the dwell time distribution of new connections in the cell. 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 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.

Appendix

Channel holding time

The Kronecker product of matrix $\boldsymbol{A}$ and matrix $\boldsymbol{B}$ is defined as the hypermatrix

$$\boldsymbol{A} \otimes \boldsymbol{B} = \{A_{ij} \cdot \boldsymbol{B}\},$$

the Kronecker sum is defined as

$$\boldsymbol{A} \oplus \boldsymbol{B} = \boldsymbol{A} \otimes \boldsymbol{I_m} + \boldsymbol{I_n} \otimes \boldsymbol{B},$$

where n and m are the dimensions of the quadratic matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ respectively, $\boldsymbol{I_m}$ and $\boldsymbol{I_n}$ are the m and n dimensional identity matrices. The Kronecker sum and product have the following properties:

- $(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = \boldsymbol{AC} \otimes \boldsymbol{BD}$,
- $e^{\boldsymbol{A} \oplus \boldsymbol{B}} = e^{\boldsymbol{A}} \otimes e^{\boldsymbol{B}}$.

Theorem: If a PH with descriptors $\boldsymbol{T}$ and $\underline{t}$ is constructed from a PH with $\boldsymbol{D}$, $\underline{d}$ and another PH with $\boldsymbol{L}$, $\underline{l}$ according to the method described in Section 2, its distribution is the minimum of the two.

Proof: The distribution of the minimum of two independent random variables with distribution functions $F(t)$ and $G(t)$ is $1 - (1 - F(t))(1 - G(t))$. Substituting the distributions of the two PH (2), we get:

$$1 - \underline{d}e^{\boldsymbol{D}t}\underline{h}_D \cdot \underline{l}e^{\boldsymbol{L}t}\underline{h}_L.$$

Let us introduce the notations $\underline{a}(t) = \underline{d}e^{\boldsymbol{D}t}$, $\underline{b}(t) = \underline{l}e^{\boldsymbol{L}t}$. Then this minimum distribution is given as:

$$1 - (\sum_{i=1}^{D} a_i(t))(\sum_{i=1}^{L} b_i(t)).$$

The distribution of the constructed PH is $1 - (\underline{d} \otimes \underline{l})e^{(\boldsymbol{D} \oplus \boldsymbol{L})t}(\underline{h}_{D \cdot L})$, according to (4). Using again the properties of the Kronecker operations we get:

$$1 - (\underline{d} \otimes \underline{l})(e^{\boldsymbol{D}t} \otimes e^{\boldsymbol{L}t})h_{DL} = 1 - (\underline{d}e^{\boldsymbol{D}t} \otimes (\underline{l}e^{\boldsymbol{L}t}\underline{h}_{DL}.$$

Substituting $\underline{a}(t)$, $\underline{b}(t)$ and the definition of the Kronecker product, we get:

$$1 - \sum_{i=1}^{D} a_i(t) \cdot (\sum_{j=1}^{L} b_j(t)) = 1 - (\sum_{i=1}^{D} a_i(t))(\sum_{i=1}^{L} b_i(t))$$

q.e.d.

Service time distribution

Theorem: If a PH with descriptors $\boldsymbol{S}$, $\underline{s}$ is constructed from the PH of $\boldsymbol{T}$, $\underline{t}$ with pdf $f(t)$ and a Markov chain with $\boldsymbol{Q}$, $\underline{q}$ according to the method described in Section 2, its pdf is also $f(t)$, i.e. the two distributions are the same.

Proof: the pdf of the constructed PH according to (5) is:

$$g(t) = (\underline{q} \otimes \underline{t})e^{(\boldsymbol{Q} \oplus \boldsymbol{T})t}(\underline{h}_M \otimes \underline{T}^0).$$

Using the second and the first property of the Kronecker operations this is:

$$g(t) = (\underline{q} \otimes \underline{t})(e^{\boldsymbol{Q}t} \otimes e^{\boldsymbol{T}t})(\underline{h}_M \otimes \underline{T}^0) =$$

$$(\underline{q}e^{\boldsymbol{Q}t}) \otimes (\underline{t}e^{\boldsymbol{T}t})(\underline{h}_M \otimes T^0) =$$

$$(\underline{q}e^{\boldsymbol{Q}t})\underline{h}_M \otimes (\underline{t}e^{\boldsymbol{T}t}T^0).$$

The first argument of this Kronecker product is always one since $\underline{q}e^{\boldsymbol{Q}t}$ is a probability vector, therefore the expression reduces to $1 \otimes \underline{t}e^{\boldsymbol{T}t}T^0 = \underline{t}e^{\boldsymbol{T}t}T^0 = f(t)$, q.e.d.

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

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

[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] Yi-Bing Lin; Mohan, S.; Noerpel, A., Queueing channel assignment strategies for PCS hand-off and initial access, *Universal Personal Communications*, 1994. Record., 1994 Third Annual International Conference on , 1994 Page(s): 365 -369

[19] 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

[20] 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

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

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

[23] 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

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

[25] 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

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

[27] 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

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

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

Acknowledgements

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

Figure 1. Distribution function of the generalized gamma and the fitted PH

Figure 2. Pdf of the generalized gamma and the fitted PH

Figure 3. Construction of the channel holding time distribution

Figure 4. Construction of the service time distribution

Figure 5. Cumulative error of the approximation versus arrival rate

Figure 6. Utilization of heavily loaded base station

Figure 7. Utilization of lightly loaded base station

Figure 8. Blocking probabilities under heavy load conditions

Figure 9. Blocking probabilities under light load conditions

Figure 10. Blocking probabilities with all types policy 1

Figure 11. Channel utilization

Figure 12. Blocking probabilities with type 2 HO policy 2

Figure 13. Channel utilization

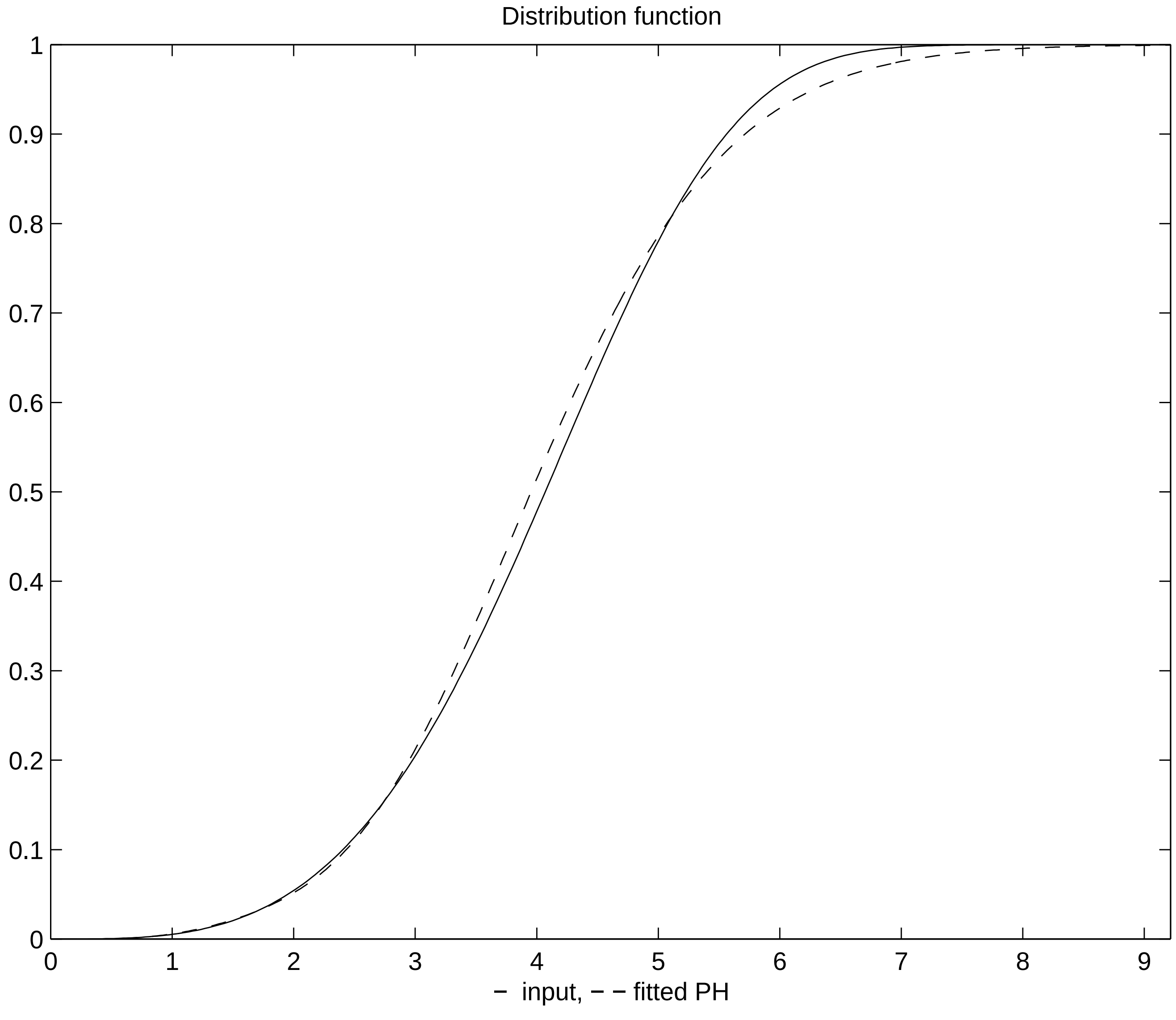

Figure 1: Distribution function of the generalized gamma and the fitted PH

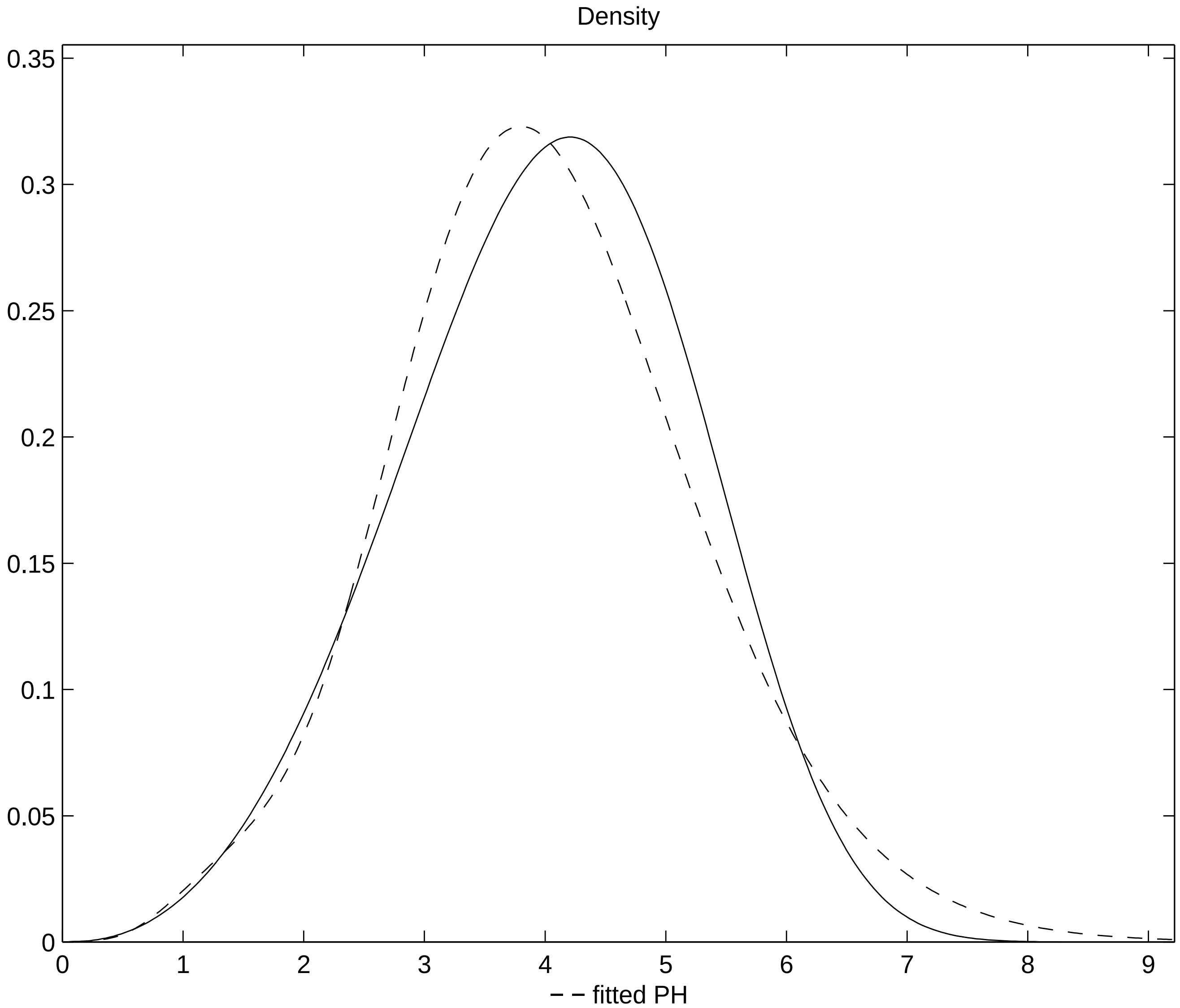

Figure 2: Pdf of the generalized gamma and the fitted PH

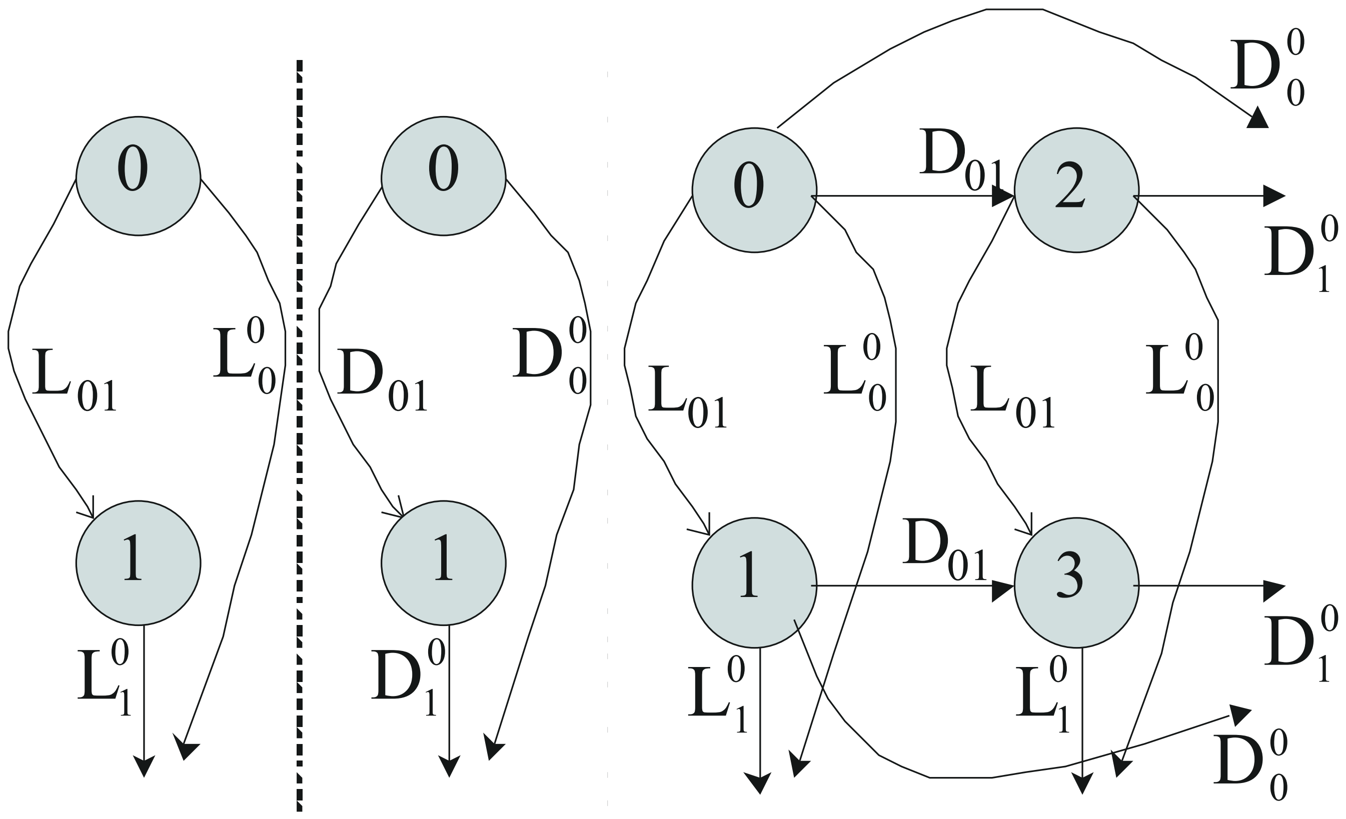

Figure 3: Construction of the channel holding time distribution

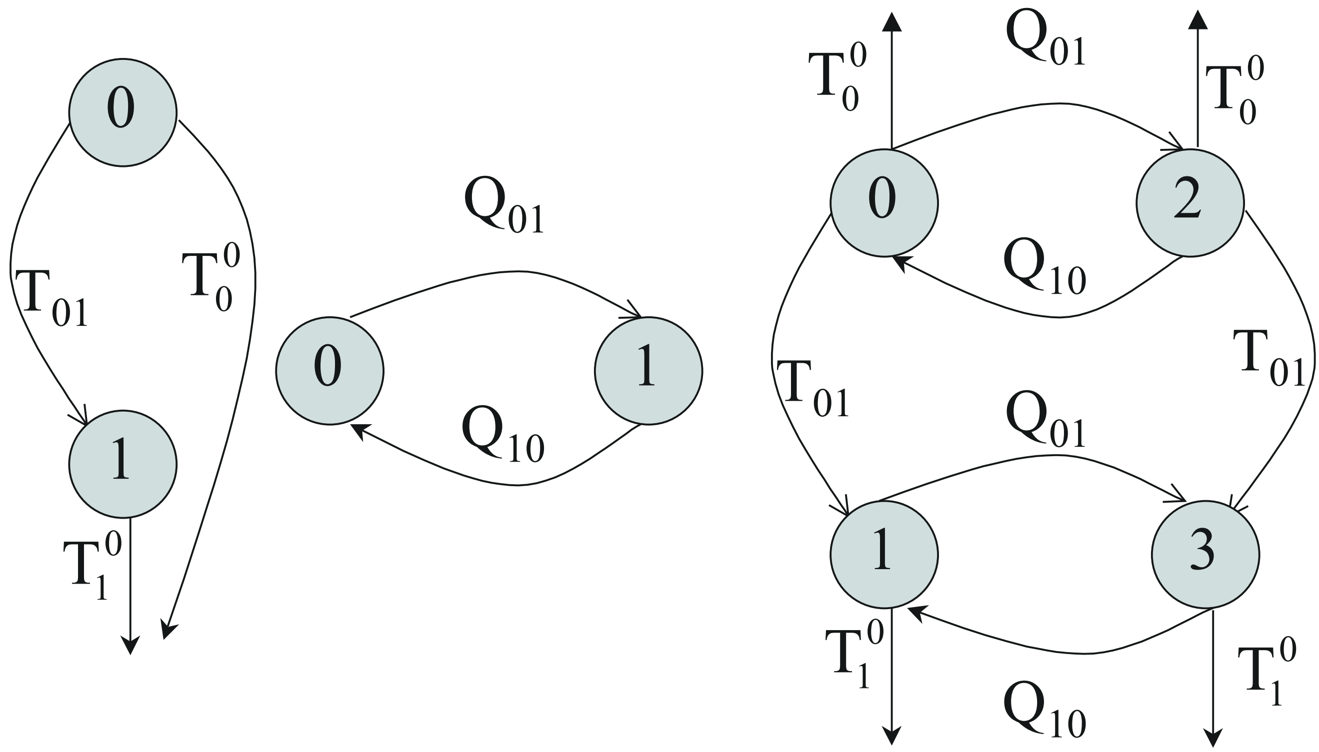

Figure 4: Construction of the service time distribution

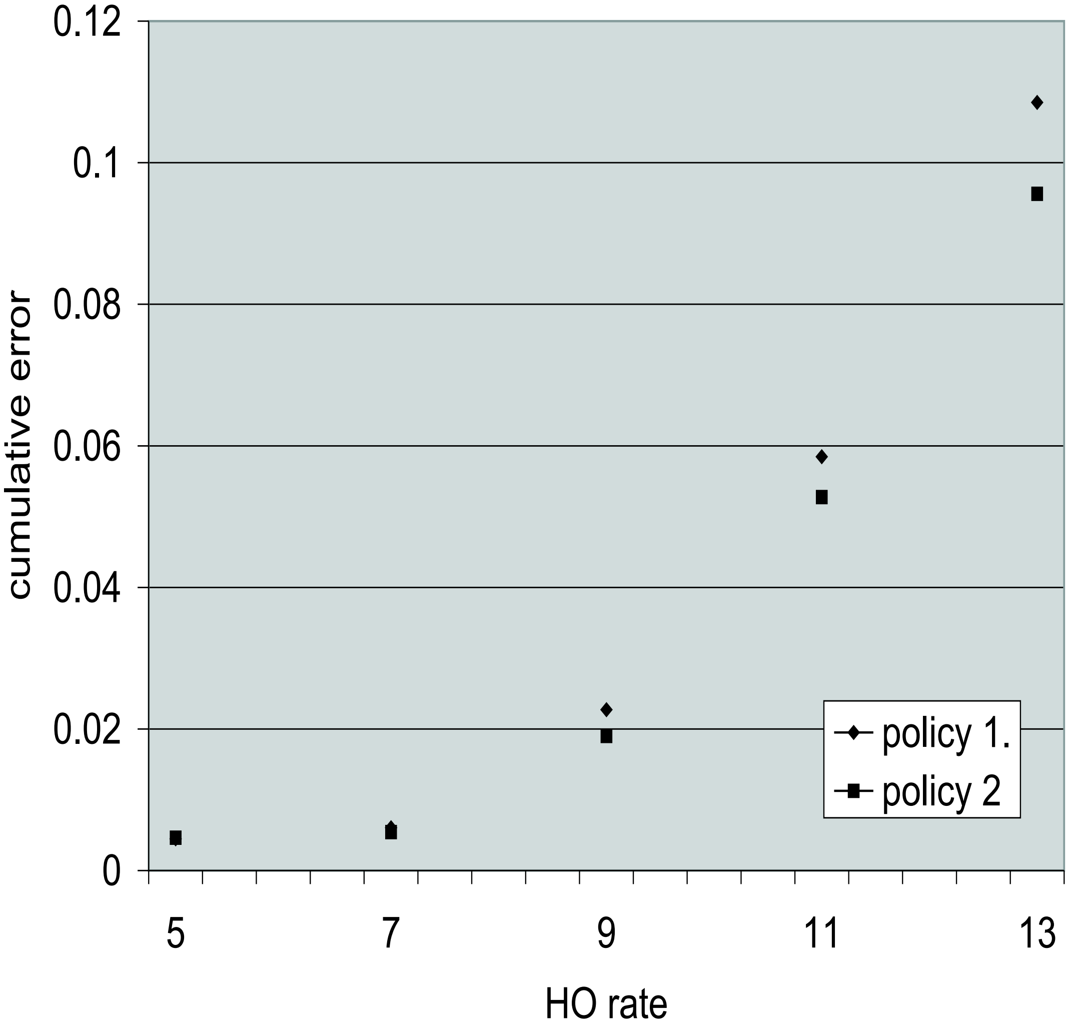

Figure 5: Cumulative error of the approximation versus arrival rate

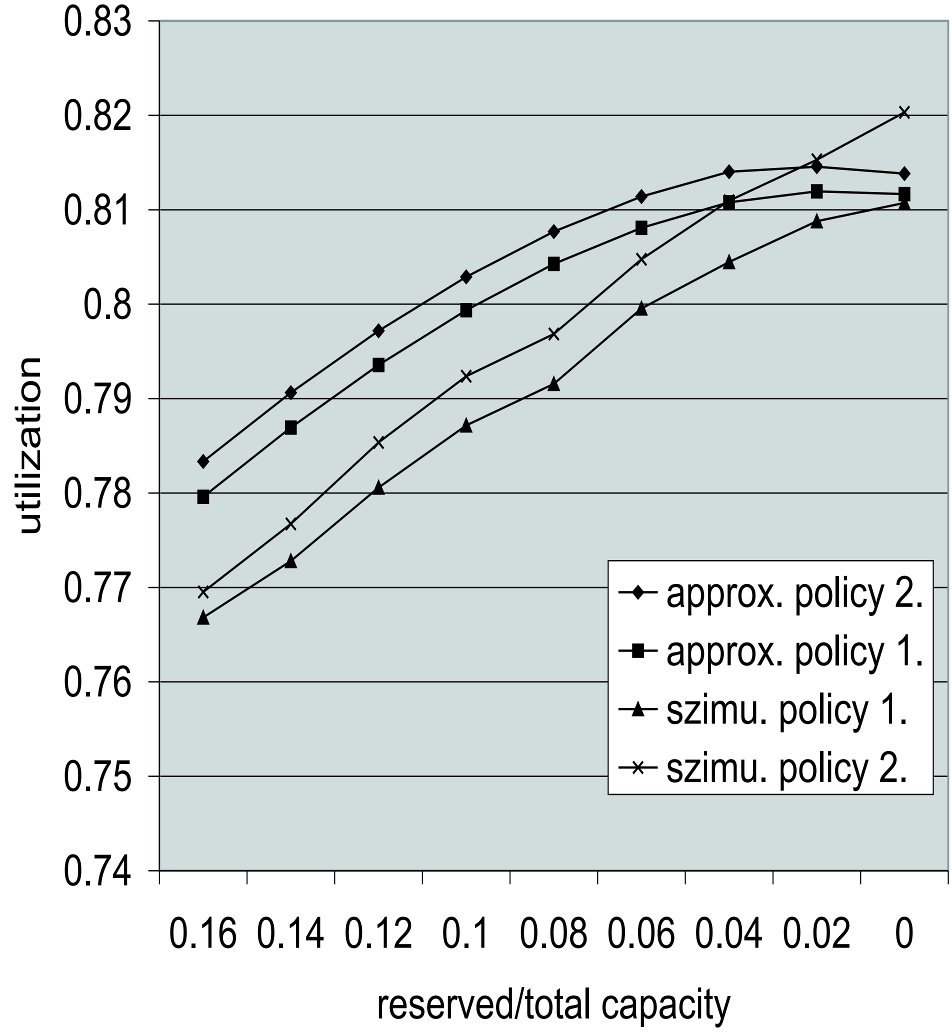

Figure 6: Utilization of heavily loaded base station

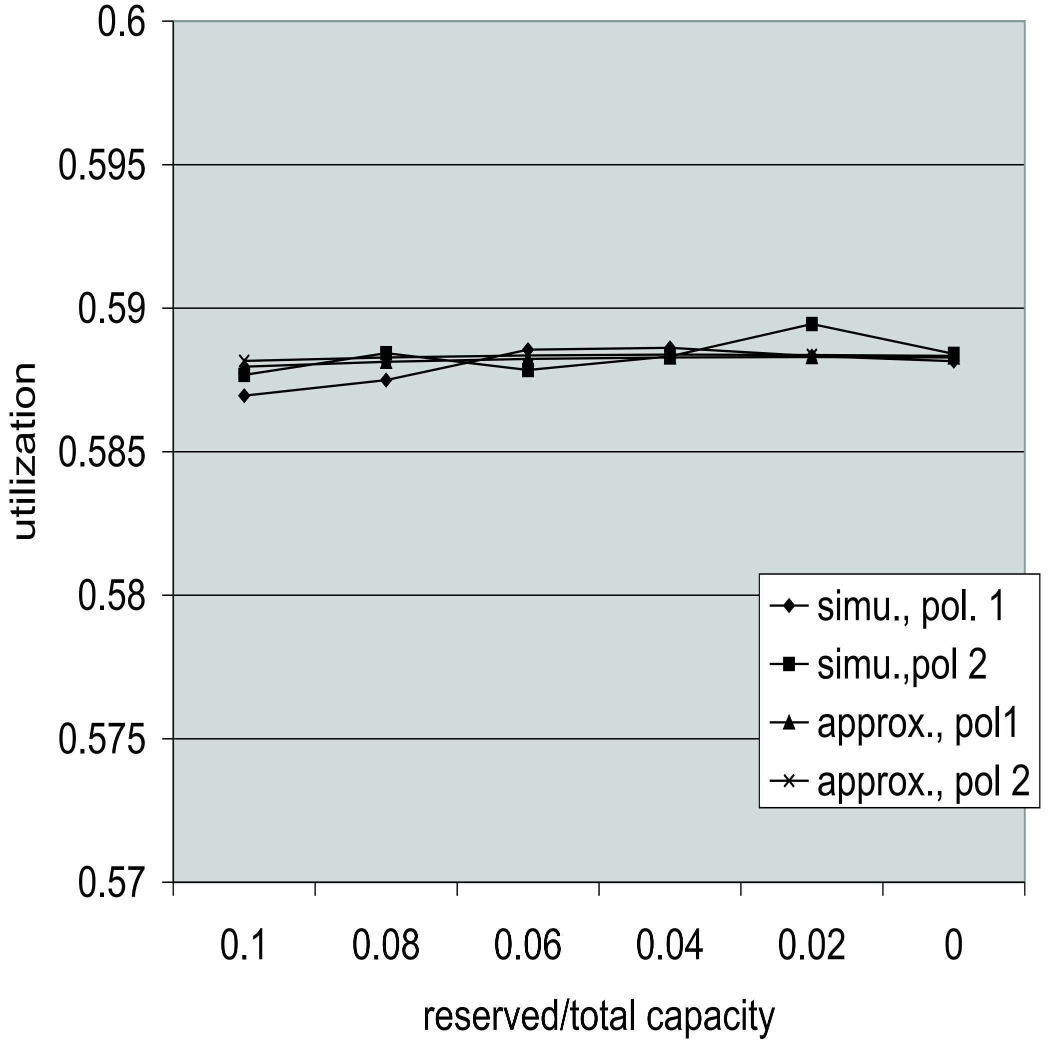

Figure 7: Utilization of lightly loaded base station

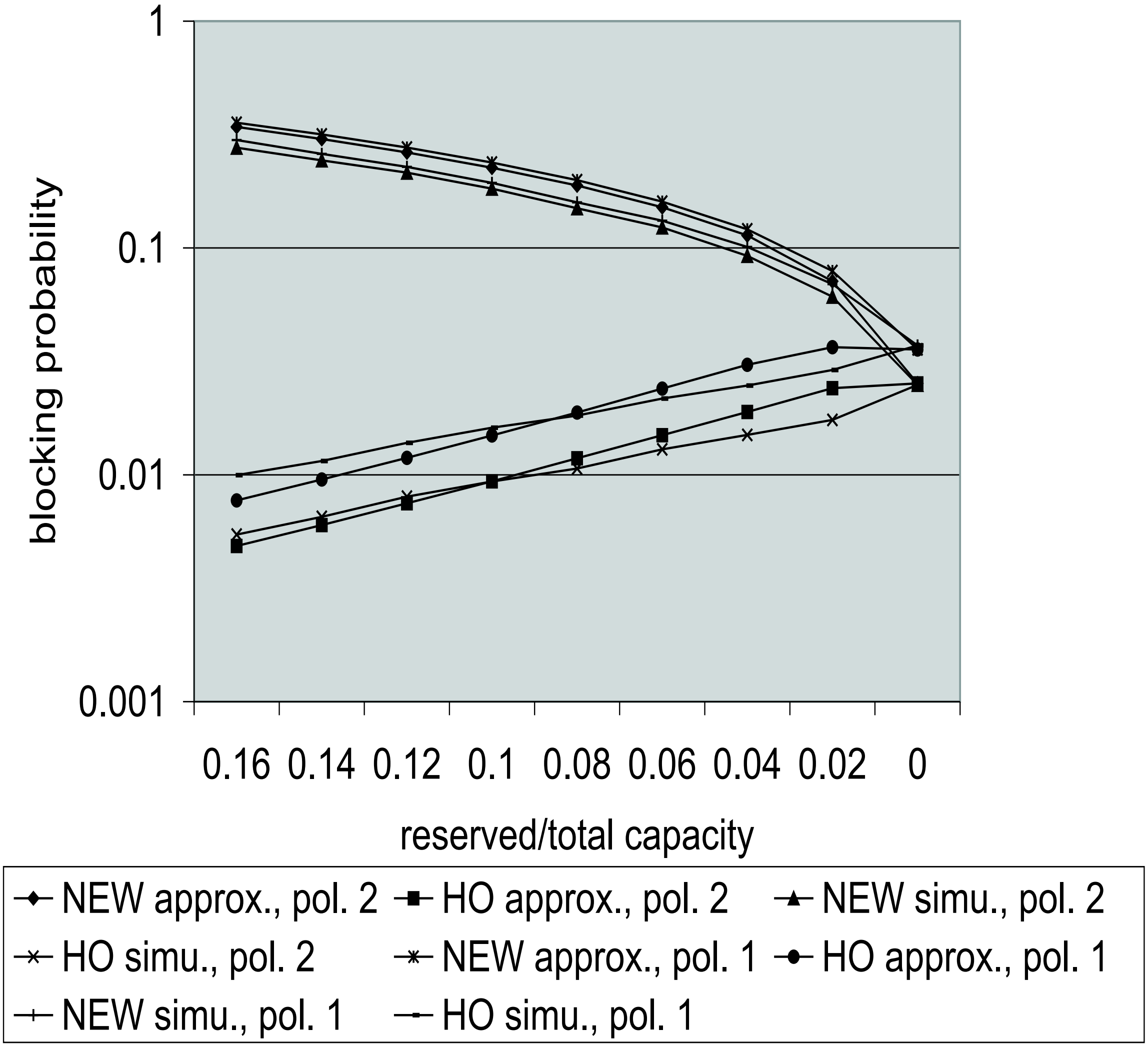

Figure 8: Blocking probabilities under heavy load conditions

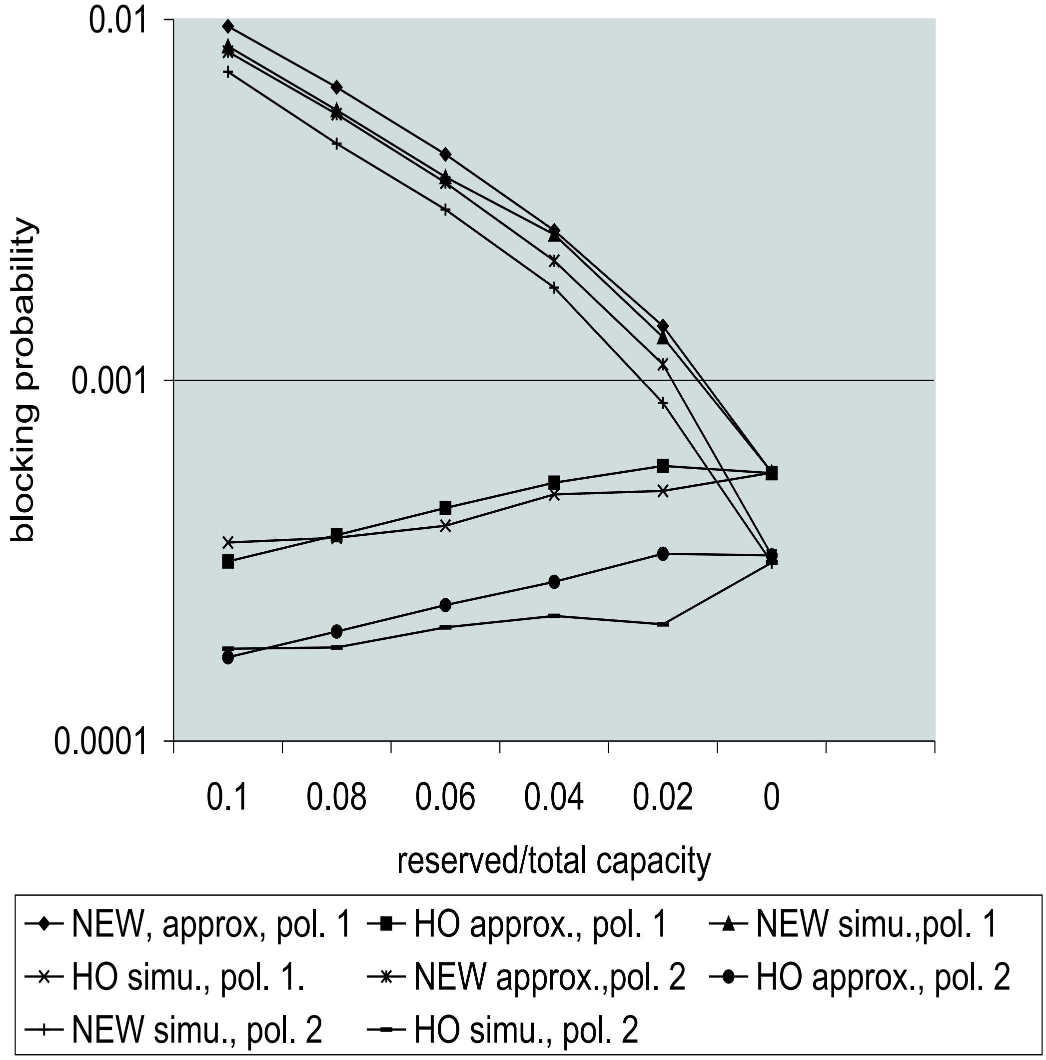

Figure 9: Blocking probabilities under light load conditions

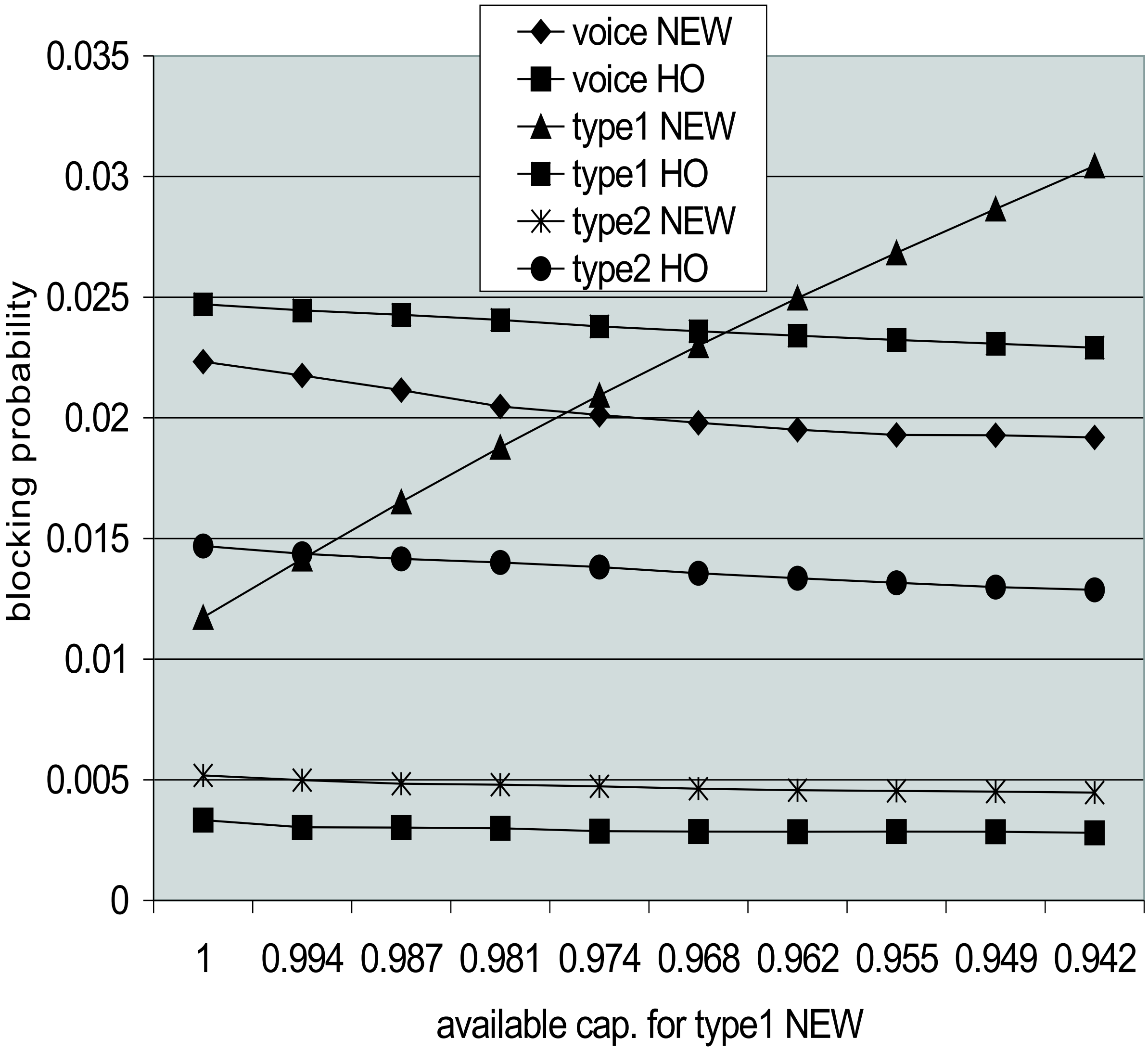

Figure 10: Blocking probabilities with all types policy 1

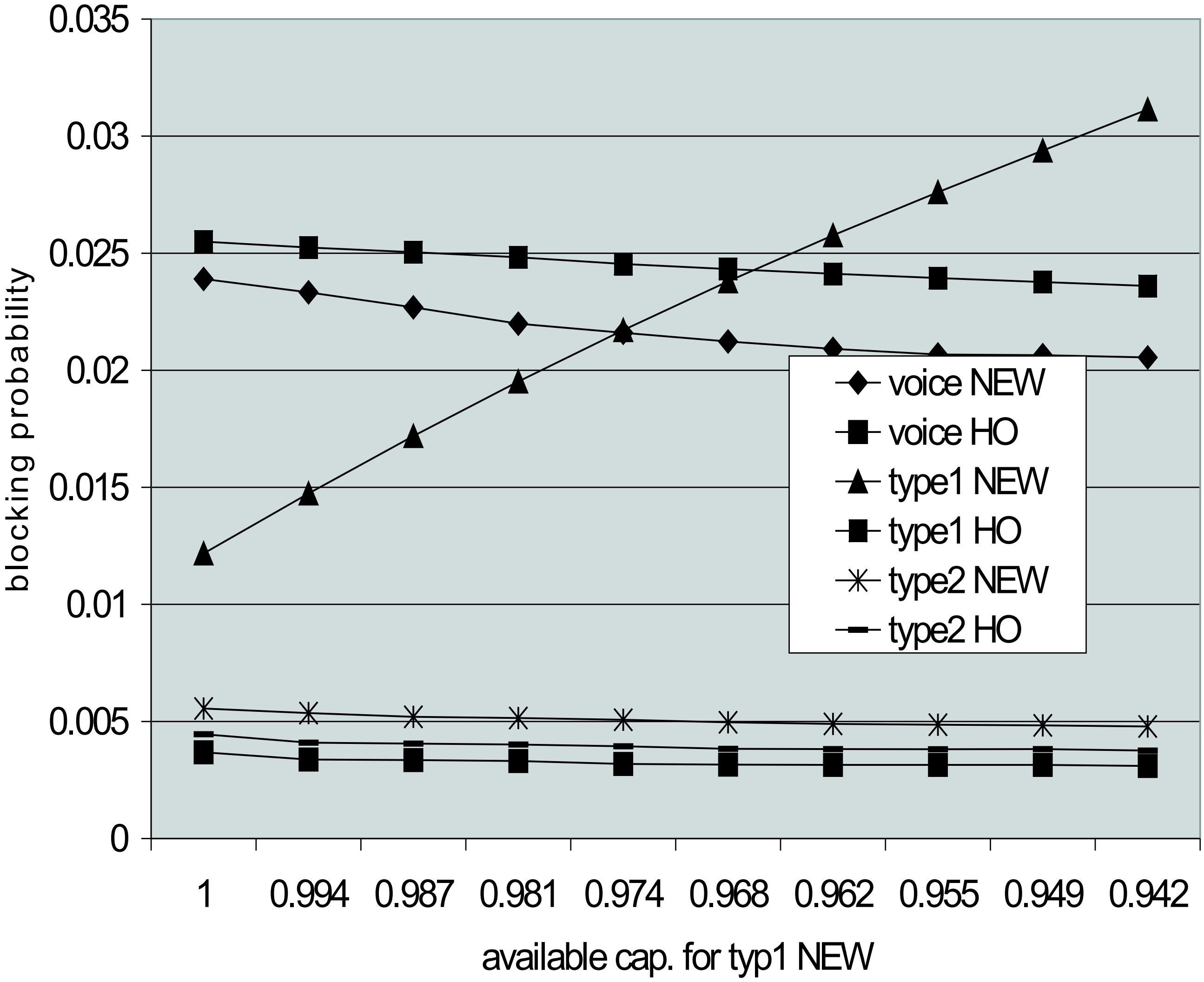

Figure 11: Blocking probabilities with type 2 HO policy 2

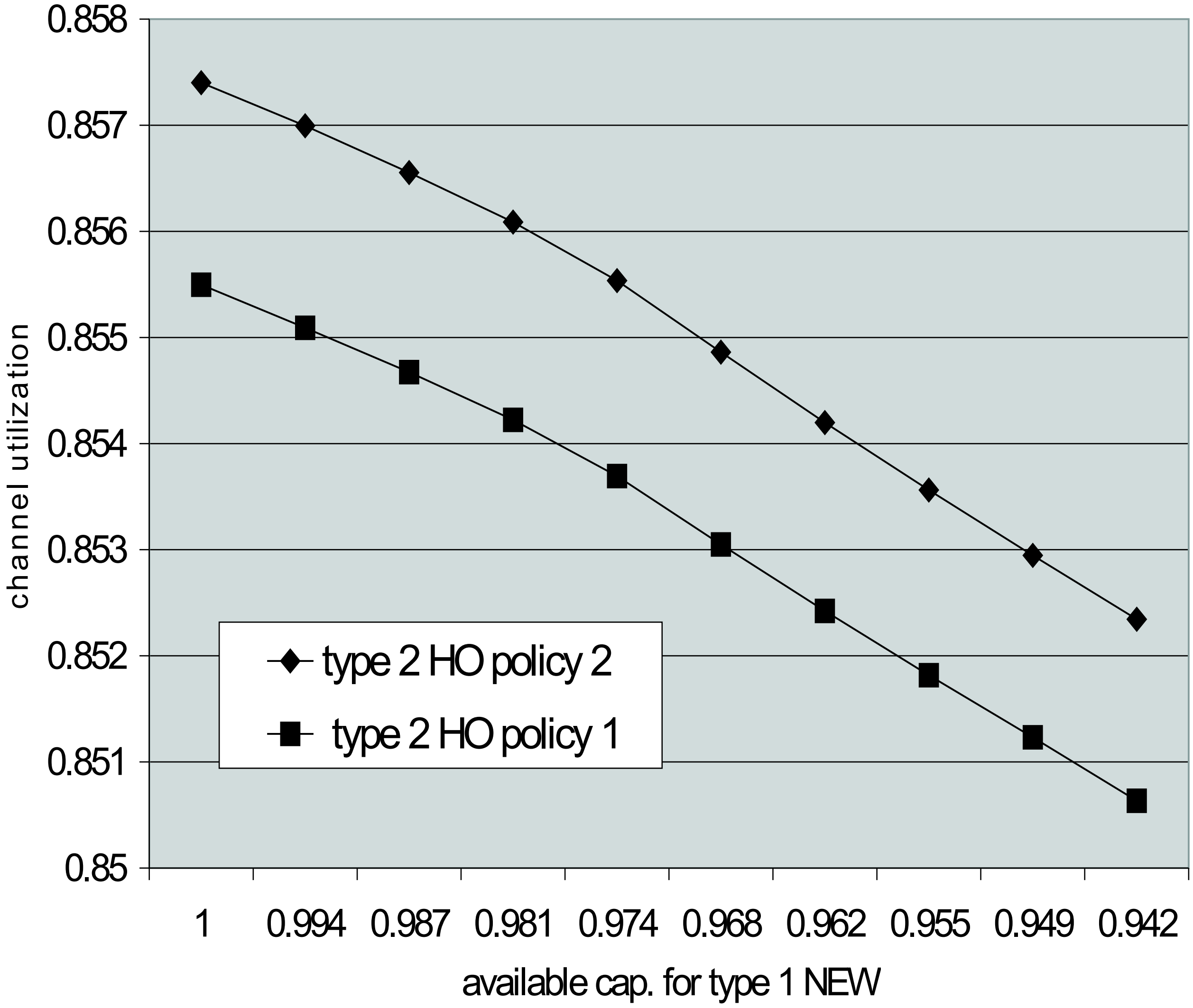

Figure 12: Channel utilization