Journal of ELECTRICAL ENGINEERING, VOL. 58, NO. 3, 2007, 152–160

CLUSTERING AND NEURAL MODELING FOR PERFORMANCE EVALUATION OF MOBILE COMMUNICATION NETWORKS

Dimitar Radev * — **Izabella Lokshina** **

In this paper, we present a core network model of universal mobile telecommunication system with calls that belong to one of four service classes and arrive randomly. Arriving calls are granted service based on specific service class, required maximum and minimum bandwidth, and available network resources. Performance of priority-based dynamic capacity allocation, suitable for the wireless ATM system is analyzed. Scheduling of the ATM cell transmission in each time division multiple access frame for the uplink is based on a priority scheme. Blocking probability and throughput parameters for bandwidth sharing policy are considered, and partial overlap link is implemented. The clustering procedure for the performance analysis of the mobile communication networks and the blocking probability and throughput measurements are introduced as Markov reward models enhanced with vector quantification and neural modeling. The optimal link occupancy probability distribution is determined using neural network that was trained on the base of Kohonen rules. Simulation and numerical results are shown.

K e y w o r d s: mobile communication networks, Markov reward models, clustering, and neural modeling

1 INTRODUCTION

We can identify a feature of the mobile communication network that the access part is totally independent from the core network as one of the key architectural aspects of the third-generation mobile communication networks. This autonomy becomes a reason for generating different services and developing different techniques to evaluate parameters of their performance.

All multi-service networks, providing guaranteed blocking probability and guaranteed throughput level, as for example, core of next generation UMTS/IMT (universal mobile telecommunication system), have the access components independent from the core network. UMTS is a third-generation broadband, packet-based transmission of text, digitized voice, video, and multimedia at data rates up to 2 Mbps (megabits per second) that offers a consistent set of services to mobile computer and phone users no matter where they are located in the world. Once UMTS is fully available geographically, computer and phone users can be constantly attached to the Internet as they travel and, as they roam, have the same set of capabilities no matter where they travel to, getting access through a combination of terrestrial wireless and satellite transmissions. The access network provides a core-network-technology-independent access platform to all core networks and network services from the mobile terminals. UMTS core network can be GSM-based (global system for mobile communications), providing the access to ISDN/PSTN (integrated service digital network/public switched telephone network) networks and services. Another part of the UMTS core network can be based on the GPRS (general packet radio services) network, providing the packet-switched access to the Internet or other IP-based packet networks. In order to provide the access to various core networks, the standardization of the interface between the access and core networks is required; that arrangement allows the access network with various radio-access-technology-dependent and mobility functions to be totally removed from the core network.

UMTS/IMT networks can be classified based on four service classes: conversational, streaming, interactive and background. The conversational class provides high quality access to a range of different services, including high bit rate services. This class is suitable for the demanding user who wishes to receive bandwidth assurance similar to that of the CBR (constant bit rate) class in an ATM. The streaming class is designed to carry high bandwidth with VBR (variable bit rate) services, such as medium- or high- quality video or teleconferencing service. The interactive class supports less demanding services, typically supported by today's best effort IP networks, including file transfer, web browsing, or telnet applications. The holding time of interactive class calls typically depends on the throughput. The background class is of the best effort type, meaning that background calls receive whatever bandwidth is "left over" with the calls of the higher priority service classes. Examples of this service class include e-mail and low quality file transfers. With respect to their holding time, the background class is similar to the interactive class. The fact, that it is relatively simple to obtain the mean and variance of the distribution, but still difficult to obtain the distribution itself, becomes a reason for approximations applied to determine a discrete version of a normal density function for link occupancy distribution (Racz *et al*, 2001), [1]. ATM networks meet

* University of Rousse 8 Studentska Street 7017 Rousse, Bulgaria E-mail: dradev@abv.bg; ** SUNY Oneonta 324B Netzer Administration Bldg. Oneonta, NY 13820, USA, E-mail: lokshiiv@oneonta.edu

ISSN 1335-3632

flexible GoS (blocking probability) and QoS (throughput) multimedia service requirements through statistical multiplexing the fixed-size cells of different traffic types. Recently these capabilities were extended to the wireless networks. ATM Forum has been developing a set of functional specifications for WATM (wireless ATM), including MATM (mobile ATM) for mobility support within an ATM network, and the radio access layer for the ATM-based wireless access. Mobile ATM protocol extensions include the handoff control, the location management for the mobile terminals, the routing aspects of the mobile connections, the traffic/QoS control for the mobile connections, and the wireless network management. The radio access layer specifications include the wireless control, the data link layer, the medium access control, and the physical layer (Xhafa and Tonguz, 2004), [2]. Different MAC protocols (medium access control) have been proposed to support multimedia services, such as PRMA (packet reservation multiple access) and its variants, the distributed queue request update multiple access, the dynamic slot assignment, DTDMA (dynamic time division multiple access), *etc.* The main objective is to achieve an efficient utilization of the radio channel through an appropriate scheduling of a variety of traffic classes with different burstiness characteristics and GoS/QoS requirements, since the available bandwidth that has to support multiple users is limited. Based on these three different priorities for different types of the ATM traffic, the reservation mechanism over the MAC frame and the retransmission of the error cells can be provided (Fantacci, 2000), [3].

In this work we consider wireless ATM network in the dynamic multi-service UMTS/IMT environment. We attempted to evaluate WATM parameters corresponding to the link occupancy distribution with use of Markov reward model, enhanced by the vector quantification and neural modelling.

2 CORE NETWORK AND WATM ARCHITECTURE

In this section we discuss different aspects of the dynamic UMTS/ITM access networks and focus at advantages that can be obtained through the wireless ATM system as switching and multiplexing technology. The main driving force for the UMTS/IMT standards is the user demand for new features and capabilities that require both increased coverage and bandwidth over next generation networks. These new features and capabilities include:

- The efficient support of deterministic bit rate and statistical bit rate services.
- The seamless wideband Internet/intranet access.
- The multimedia communication capabilities.
- The capacity and capability to serve to the whole population with global and seamless radio coverage.
- The radio resources flexibility for multiple networks and traffic types.
- The radio bearer capabilities up to 2 Mb/s (with wide coverage).
- The low cost of services and terminals.
- The flexible introduction of new services and technical capabilities.

The summary of existing and forthcoming technologies in terms of coverage vs. bandwidth is shown at Figure 1. The limited range of existing GSM services offers the data rates up to 9.6 kb/s only. Within the GSM framework, the GPRS enables the higher-bit-rate services with the data rates up to 170 kb/s that can be allocated according to the actual user demand. The UTMS/IMT standards are expected to offer full coverage at the data rates from 384 kb/s up to 2Mb/s in the close range communications. Wireless ATM, wireless LAN and wireless IP technologies are expected to further increase the service rate of UTMS/IMT up to 20 Mb/s in indoor environments, such as offices and public buildings.

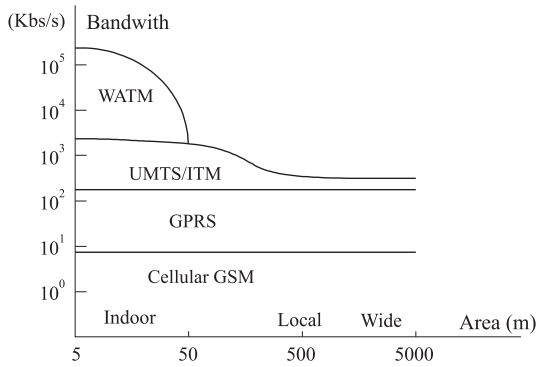

Fig. 1. Bandwidth shearing of the wireless system

UTMS/IMT cannot provide narrow and wideband services with the data rates above 2 Mb/s for global wide-area coverage. However, we propose the scenario that provides even higher bit rate services with local area technologies, such as wireless ATM and wireless IP. In order to access these high bit rate services the users need to have the dual port terminals that allow the access to both WATM/WIP and WCDMA within the UTMS/IMT concept.The hub-based configuration is demonstrated in Figure 2, in which N WATM terminals can send and receive traffic to/from the hub that can be connected to a high-speed ATM backbone network via an ATM switch or multiplexer.

The hub acts also as a scheduler. The MAC layers are used to control the cell transmission over the radio channel. The MAC protocol uses a dynamic TDMA (time-division multiple-access) scheme for the uplink from the user terminals to the hub. Since the downlink transmission (from the hub to the user terminals) operates in a broadcast mode, it can also apply a TDM mode (time-division multiplexing). The WATM implementation can be designed in two different ways. It can be based entirely on the ATM structure with additional error control and resource control through the MAC scheduling, in order to provide QoS guarantees at the ATM connection

end points as a single ATM network. Otherwise, it can be based on networking, which processes the AAL PDU (protocol data units) in an optimized manner that suitable for wireless environment (Eneroth *et al*, 1999), [4]. The configuration of the WATM protocol suite, which transports the ATM cells transparently across the wireless medium with the wireless MAC, is shown in Figure 3. The user terminal can communicate to another one within the same cell/cluster as in the case of wireless LAN (local area network), or to the external user terminal through the fixed broadband ATM networks. The topology of each WATM cell/cluster can be one of the following two types: the broadcast-based ad hoc WATM or the hub-based WATM.

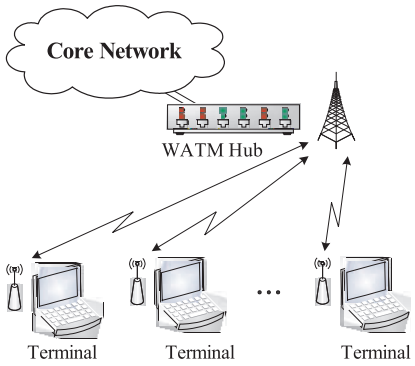

Fig. 2. Example of the hub-based cell/cluster operating in a TDMA/TDD mode

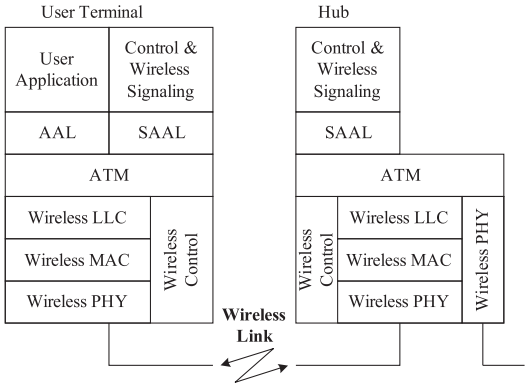

Fig. 3. ATM-based protocol suite of user terminal and hub

In the broadcast-based ad hoc WATM configuration, a set of N WATM user terminals can share the bandwidth/frame in a demand assignment TDMA mode. Traffic bursts sent by a terminal contain the ATM cells with the virtual circuit identifiers or virtual path identifiers. Using the wireless MAC address, the appropriate WATM terminal can receive the WATM cells. The operation of radio link at the physical layer uses only one frequency band as in the dynamic TDMA with TDD MAC protocols. The user terminals in the cell or cluster can nominate one of them to be a main scheduler, and another one to be a standby scheduler. At the beginning of each frame, the scheduler can send the frame reference and frame control information indicating the allocation of both the request and traffic slots, as is shown in Figure 4. The request slots can be accessed by the user terminals in a fixed-assignment or slotted-Aloha mode. They are used by the user terminals to send their requests for the traffic slots

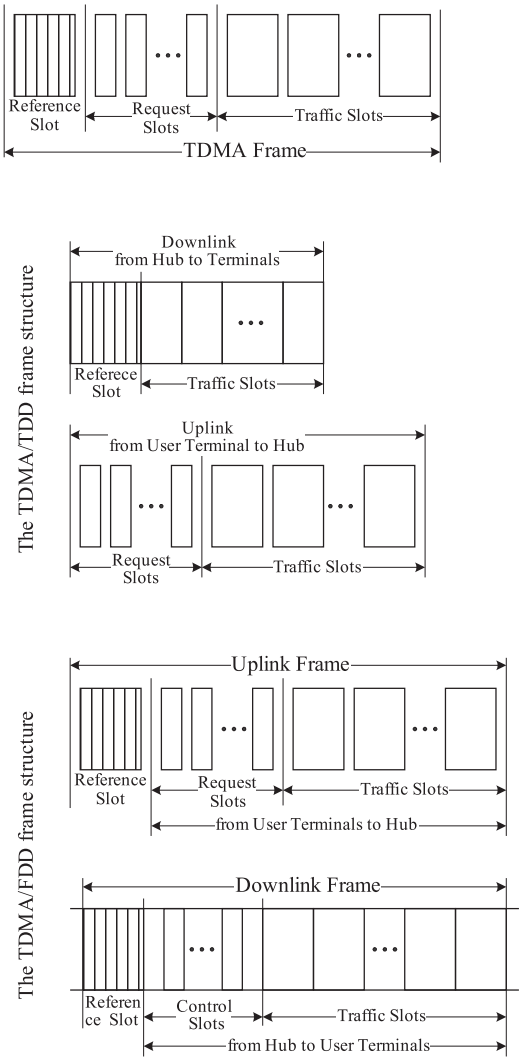

Fig. 4. The frame structure in a broadcast bandwidth sharing TDMA mode

Following the allocation of the scheduler, the user terminals send the traffic in the assigned traffic slots. The standby scheduler will take over the control and allocation, if the main scheduler fails. Each WATM terminal can transmit during the reserved slots and if not transmitting, can store traffic bursts received from the broadcast medium. Both uplink and downlink channels can share the same frequency slot as shown in Figure 4. In this arrangement, both the user terminal and the hub operate in a TDD mode. On the other hand, the uplink and downlink channels can occupy a pair of frequency slots, and then the terminals and the hub operate in a full-duplex FDD (frequency-division diplexing) mode. Uplink traffic time-slots are allocated to the user terminals on demand

by the scheduler (the hub) using reservation and/or contention in association with the UPC mechanism (usage parameter control), so that the negotiated traffic contract of declared ATM connections can be maintained (Ganesh Babu *et al*, 2001), [5].

Each user terminal generally supports five categories of the multimedia traffic: CBR, rt-VBR (real-time), nrt-VBR (non-real-time), ABR and UBR. The capacity allocation for the CBR traffic is straightforward. Since the CBR traffic uses fixed capacity during the entire connection, it affects the system only by reducing the total capacity with the amount of static bandwidth allocated to the CBR connections. Other traffic categories, except UBR, are subject to the CAC (call admission control) according to the traffic contracts. The UBR traffic is given no capacity commitments and therefore, is served with the remaining capacity, if available. The ABR traffic can be controlled through the feedback, depending on the congestion in the network. On the other hand, rt-VBR and nrt-VBR are uncontrollable within the committed traffic profile. Therefore, we consider the QoS commitments to be satisfied for the uncontrollable re-VBR and nrt-VBR traffic in the proposed scheduling scheme applicable to both the TDMA/TDD and TDMA/TDM/FDD configurations.

As shown in Figure 4, the uplink TDMA frame contains a frame marker to denote the frame beginning, several signalling time slots for the capacity request ad internal control signalling, and a number of traffic time slots, since each no overlapping traffic time slot can accommodate one ATM cell. In each frame the user terminal can send its capacity request in a designed signalling time slot that contains the number of the slots needed for rt-VBR and nrt-VBR traffic arrived in the previous frame, to the base station. Its request can also include other relevant parameters. The scheduler first stores all requests received from different user terminals in a request table. Then it calculates the capacity allocated to the rt-VBR traffic. If available capacity remains, it continues the capacity allocation to the nrt-VBR traffic. After both the rt-VBR and nrt-VBR requests are satisfied, and if there is still capacity available, the scheduler can provide the user terminals in a round-robin manner for transmission of the ABR and UBR traffic. The scheduler prepares and broadcasts the time-slot assignment to all user terminals, effective in the next frame. In the time-slot assignment, the scheduler does not mention explicitly the number of the rt-VBR and nrt-VBR cells that the user terminals can transmit. Instead, it only indicates the time slots allocated to a given user terminal. The user terminal keeps the rt-VBR traffic in a buffer for one frame and the nrt-VBR cells in a FIFO queue, since the nrt-VBR traffic can tolerate the delay variations. It will transmit the real-time cells first. If the number of the rt-VBR cells exceeds the allocated capacity, the excess cells can be lost as the maximum CTD (cell transfer delay) of the rt-VBR traffic must be met. Otherwise, it will continue sending the nrt-VBR cells stored in its data FIFO queue in the remaining allocated time slots, and the nrt-VBR cell can be lost only when overflow occurs in the queue.

3 BANDWIDTH SHARING MODEL WITH MARKOV CHAIN

The dynamic UMTS/IMT networks measurements and performance analysis requires that the call-level models are available which take into account the blocking probability and throughput trade-off in a dynamic multi-service environment, *ie*, where calls arrive and depart randomly in time. The most important network performance characteristics are the link occupancy distribution and the end-to-end blocking, which could be defined as the probability of a cell blocked of all the routes from its origination node to its destination node. Generally, the calculation of the link occupancy distribution in those networks is very difficult; and practically it is often based on the decomposition into series of fixed routing problems with link-by-link decompositions. Even considering only a single link, the computation of occupancy distribution is not trivial and differs significantly from the traditional circuit-switching networks, in which the modeling often is based on the Poisson process. In order to analyze overflow traffic, *eg* in alternative routing, it is necessary to consider less tractable arrival processes like renewal, or MMPP (Markov modulated Poisson process). Then the assumption is that the holding time distribution is exponential. On the other hand, the class of state-dependent Poisson processes has been used for the arrival process, where the computation of the link occupancy distribution, and thereby the link blocking measurements, are rather complicated. Designed for the multi-service environment modeling, a CTMC (continuous time Markov chain) is used. Since the mobile core network exercises the admission control on the link-by-link basis, a single transmission link is defined as Markov reward model (Radev *et al*, 2005), [5]. Usually the issue of bandwidth sharing should be considered in the context of dynamically arriving and departing flows, which naturally calls for the application of classical multi-rate loss models. These models have proved to be useful in the measuring and performance evaluation of circuit switched similar to ATM networks; but their direct application in multi-service networks like the internet is non-trivial because of several reasons. At first, it is not possible to associate a constant bandwidth with elastic services. The bandwidth occupied by the elastic flow depends on the current load on the link and on the scheduling and rate control algorithms applied in the network nodes. Then, the notion of blocking, when applied to elastic flows, needs to be reconsidered because an arriving elastic flow might get into service even if at the arrival instant there is no, or a very small, bandwidth available. At last, for many services the actual residency time of the elastic flows depend on the throughput that the flow receives. For instance, a file transfer protocol (ftp) session would last longer if

its throughput decreased. Since we cannot directly use the reservation-based multi-rate models, we try to find meaningful extensions to allow the inclusion of both QoS - assured and elastic traffic into a common framework.

Let us consider the single transmission link of capacity C. Calls arriving to the link belong to one of the following three traffic classes, as is shown in Figure 5. Non-adaptive stream or rigid traffic class flows are characterized by their peak bandwidth requirement b_1, flow arrival rate λ_1, and departure μ_1. This class also is known as the conversational service class. Adaptive stream class flows are characterized by their peak bandwidth requirement b_2, minimum bandwidth requirement b_2^{min}, and flow arrival rate λ_2 and departure rate μ_2. This class also is known as the streaming class. Although the bandwidth occurring by adaptive flows may fluctuate as a function of the link load, their actual holding time is not influenced by the received throughput throughout their residency in the system. This is the case, for instance, of an adaptive video code; which because of throughput degradation, decreases the quality of the video images and thereby occupies less bandwidth. Elastic class flows are characterized by their peak bandwidth requirement b_3, minimum bandwidth requirement b_3^{min}, flow arrival rate λ_3 and their ideal departure rate μ_3. This class also is known as the interactive class. The ideal departure rate is experienced when the peak bandwidth is available. The actual instantaneous departure rate is proportional to the bandwidth of the flows. This class can be further classified into two subclasses. If the minimum accepted bandwidth is 0, then this class is the model of the best effort traffic class. If the minimum accepted bandwidth is greater than zero, then this class corresponds to the "better-than-best-effort" traffic class. A typical example of this class is the ftp. Three types of flows with throughput guarantees arrive according to independent Poisson processes, the holding time for the conversational and streaming class calls are exponentially distributed, and the holding time of the interactive class calls are determined with single queuing models. We are interested in the performance of two simple strategies of bandwidth sharing that provide GoS/QoS bound for the three service classes.

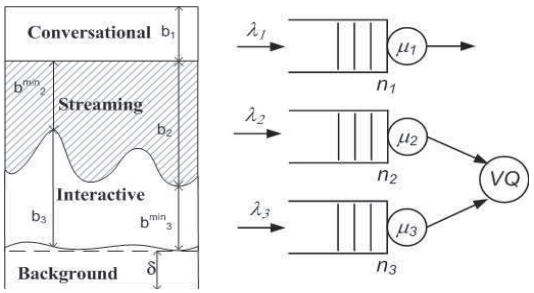

Fig. 5. Model of a single transmission link

The throughput of the conversational calls in the rigid traffic is simple queue n_1 with a constant peak bandwidth requirement b_1. The residency time of the adaptive traffic depends not only on the amount of data they want to transmit, but also on the bandwidth they receive during their holding times. The amount of data transmitted through an elastic traffic depends on the received bandwidth. The partial overlap link (POL) of the other two classes can be presented as two parallel single queues. Based on description of the service classes, we clearly need to simulate the service classes with only throughput guarantee (*ie*, conversational, streaming and interactive), and to compute the bandwidth leftover for the background class. Then, we present the stochastic stationary process as a network of three queues in the state space. The input parameters of the system are the set of arrival rates $(\lambda_1, \lambda_2, \lambda_3)$ and departure rates (μ_1, μ_2, μ_3), the bandwidths (b_1, b_2, b_3), and throughput constraints $\tilde{\theta}^{min}$, $\hat{\theta}^{min}$. The state model is uniquely characterized by the triple $n_1, n_2, n_3)$, where n_1 is the number of states in the conversational flows, n_2 is the number of states in the streaming flows, and n_3 is the number of states in the interactive flows. In order to obtain the performance measurement, the CTMC's generator matrix $\boldsymbol{Q}$ and the BSP are defined as bandwidth sharing policy, so that the link capacity C is divided into two parts: a common part and a part reserved only for the streaming and interactive flows. Let assume that C denotes the portion of the link capacity that is dedicated to the first three service classes, and δ denotes the portion of the link capacity that is dedicated to the background class. There are only two possibilities for δ. When $\delta = 0$ we have Policy 1, which is a complete sharing. When $\delta > 0$ we have Policy 2, which is a complete portioning. There are the following possibilities that could be true for both policies. If there is enough bandwidth for all flows to get their respective peak bandwidth demands, then class-2 and class-3 flows would occupy b_2 and b_3 bandwidth units respectively. If there is a need in bandwidth compression, *ie* $n_1 b_1 + n_2 b_2 + n_3 b_3 > C - \delta$, then the bandwidth compression of the flows would be divided in equal parts between both classes as long as minimum rate constraint is met for both classes. If there is a need for further bandwidth compression, but either one or both classes does not tolerate further bandwidth decrease at the time of a new flow arrival, then the service class that tolerates further compression would decrease the bandwidth occupied by its flows equally, as long as the minimum bandwidth constraint would be kept for this traffic class. The basic assumptions of the above rules demonstrate that both the streaming and interactive flows are greedy. It means that they always occupy the maximum possible bandwidth on the link, which is the smaller of their peak bandwidth requirements (b_2 and b_3, respectively), and the equal share of the bandwidth left for them by the conversational flows. All streaming and interactive in-progress flows share proportionally the available bandwidth among themselves, *ie* the newly arrived flow and the in-progress flows will be squeezed to the same compression values. After that, if a newly arriving flow decreases the flow bandwidth below minimal accepted value,

and is not admitted to the system, then it is blocked and lost. By this reason, it is important to develop bandwidth sharing models for the streaming and interactive flows. We propose the models, based on Markov chain, in which only transitions between neighboring states are allowed. Non-zero transition rates for the feasible states are described according to (1).

$$\begin{aligned} &q(n_1,n_2,n_3 \to n_1+1,n_2,n_3) = \lambda_1 \\ &q(n_1,n_2,n_3 \to n_1,n_2+1,n_3) = \lambda_2 \\ &q(n_1,n_2,n_3 \to n_1,n_2,n_3+1) = \lambda_3 \\ &q(n_1,n_2,n_3 \to n_1-1,n_2,n_3) = n_1\mu_1 \\ &q(n_1,n_2,n_3 \to n_1,n_2-1,n_3) = n_2\mu_2 \\ &q(n_1,n_2,n_3 \to n_1,n_2,n_3-1) = n_3\rho_3(n_1,n_2,n_3)\mu_3 \end{aligned} \tag{1}$$

The first three equations represent the state transitions due to call arrivals, while the second three equations represent the transitions due to call departures. The $n_3\rho_3(n_1,n_2,n_3mu_3$ quantity denotes the total bandwidth of the interactive flows when the system is in state (n_1,n_2,n_3), and the compression of the interactive flow is denoted as ρ_3. The BSP policy is completely determined by specifying the following output parameters: the capacity of the common part, and the maximum number of the streaming and interactive flows, N_2 and N_3. At this point, the set of feasible states of the CTMC model S is determined according to (2).

$$S = (N_2+1)(N_3+1) \tag{2}$$

The stochastic system clustering presents a steady state analysis for the states j of state space S, $j \in S$; in every discrete time epoch. The states are arranged on a grid with as many dimensions as the number of queues. Each of the axes is representing the number of customers in one of the queues.

Our goal is to determine the occupancy and arrival distributions, the partial overlap, and the call blocking probabilities of the feasible states of the queuing system. In order to evaluate density function and performance of the BSP allocation strategy, the maximal cuts and the partitions of the graph should be calculated applying Markov chain model (Kulkarni, 1995), [7]. The clustering procedure is developed with use of the VQ (vector quantification) and neural modeling. Next, we consider a small telecommunication system with a link capacity $C = 25$, and for two-dimensional presentation $n_1 = 1$ is kept fixed, *ie* , the available bandwidth for the streaming and interactive flows is 24 bandwidth units. The peak bandwidth requirements for these flows are $b_2 = 7$ and $b_3 = 5$, respectively. The flows are characterized with minimum accepted bandwidth, which we set to $b_{2min} = 4.2$ and $b_{3min} = 2.8$. The cut-off parameters are $N_2 = 4$ and $N_3 = 2$, respectively. The system has 15 feasible states as shown at Figure 6, out of which there are 5 states - (1,4,0); (1,3,1); (1,4,1); (1,3,2); (1,4,2); where at least one of the flows is compressed below the peak bandwidth specified by b_2 and b_3.

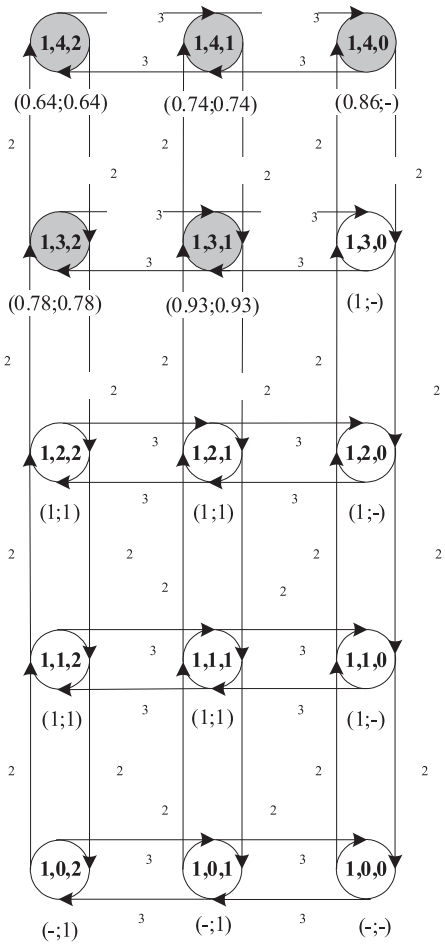

Fig. 6. Feasible states for a link capacity C=25

4 CLUSTERING PROCEDURE WITH KOHONEN NEURAL NETWORKS

We can associate the clustering problem with a discrete distribution with independent margins, where the random values $X_1^{(2)},...,X_k^{(2)}$ and $X_1^{(3)},...,X_k^{(3)}$ are distributed into M classes $S_1,\dots S_M$. Then again, the number of streaming n_2 and interactive n_3 flows can be presented as a couple of time series $\{x^{(2)}\}$ and $\{x^{(3)}\}$, for which we can determine the probability function using the VQ and Kohonen learning rules. Kohonen network algorithm (Radev *et al*, 2004) [8] provides a transformation of the input space into the patches with corresponding code vectors. It has an additional feature that the centers are arranged in a low dimensional structure (rectangular grid). We introduce vector quantification to the learning neural model in order to determine the probability density for each feasible steady state of Markov chain (Hofmann and Buhmann, 1998), [9]. Equal width of class boundaries and equal number of target values in each class can be applied as criteria for clustering (Radev and

Radeva, 2003), [10]. Rectangular boundaries between separated classes are orthogonal to coordinate axes. The set of input/target couples, for which the probability function of the input space, can be defined as (3), and the neural network can be trained according to (4), where $\boldsymbol{x}_j$ is the couple of two $N-$dimensional input vectors, and $\boldsymbol{C}_j$ is the M-dimensional vector describing the conditions of target classes.

$$\{\boldsymbol{x}_1, \boldsymbol{C}_1\}, \{\boldsymbol{x}_1, \boldsymbol{C}_1\}, \ldots, \{\boldsymbol{x}_j, \boldsymbol{C}_j\}, \ldots, \{\boldsymbol{x}_N, \boldsymbol{C}_N\} \quad (3)$$

$$\begin{aligned} \boldsymbol{x}_j &= \{X_j^{(2)}, X_j^{(3)}\}, \quad \{j = 1, ..., N\} \\ \boldsymbol{C}_j &= \{S_1, S_2, \ldots S_k, \ldots S_M\}, \quad \{k = 1, ...M\} \end{aligned} \quad (4)$$

The hidden neurons of the first layer initialize the weight matrix Wkj, and compete, and then identify the neuron-winner, which has the minimum Euclid distance dk. It obtains output value equal to 1, and defines the corresponding target class, as is shown in (5).

$$\begin{aligned} &S_k = 1, \quad \text{for} \quad d_k^{\min}, \quad d_k = \sqrt{\sum_{j=1}^{N} (X_j - W_{kj})^2} \\ &S_k = 0, \qquad \text{otherwise} \end{aligned} \quad (5)$$

At the linear layer, the neuron-winner has a negative feedback corresponding to the rest of neurons, and a strong positive feedback relative to itself, and this is used in the learning process. The coefficients of the neurons in each consequent epoch q of the training process are altered according to Kohonen learning rule, as is introduced in (6).

$$W_{kj}(q) = W_{kj}(q-1) \pm \xi\big(X_j(q) - W_{kj}(q-1)\big) 0 < \xi \leq 1 \quad (6)$$

The coefficient depends on the training epoch number q, and can be in advance adjusted in interval [0, 1] (standard 0,1). The sign of the training coefficient ξ is positive for the neuron-winner, and negative for the neighboring neurons. As a result, the area of neighboring neurons for the neuron-winner is transformed in the training process decreasing Euclid distances. We can see that the neural network is determined with two hidden neurons in each cluster class at the competitive layer. At the same time, consequent adjustment of the target class on the horizontal and vertical axes produces rectangular zones at the linear layer. Each class is described via a rectangular class face, which is proportional to the approximated probability density function.

5 LEARNING VECTOR QUANTIFICATION AND OPTIMAL OCCUPANCY PROBABILITY DISTRIBUTION

We recommend using the LVQ (learning vector quantification) in the neural model in order to determine the value of the occupancy probability density function for each discrete feasible state of two-dimensional CTMC set. The probability density function of the input vector space corresponds to the values of the streaming and interactive time series, and it is indicated in (7).

$$f(x_1^{(2)}, ..., x_k^{(2)}, x_1^{(3)}, ..., x_k^{(3)}) = \frac{\partial^{n\times 2}}{\partial x_1^{(2)}, ..., x_k^{(2)}, x_1^{(3)}, ..., x_k^{(3)}} \boldsymbol{F}\left(x_1^{(2)}, ..., x_k^{(2)}, x_1^{(3)}, ..., x_k^{(3)}\right) \quad (7)$$

We consider an approach to link occupancy neural modeling that consists of four basic phases: the PVQ (preliminary vector quantification), the weight centers determined with the bipartition, the class boundaries optimization, and the LVQ. We introduce a preliminary separation of the input space values into different classes in the phase of the PVQ, and determine the rectangular boundaries of the separated classes that are orthogonal to the coordinate axes. The linear layer fits the boundaries of the target classes in such a manner that we have the same number of the target values in each target class. A small telecommunication system with the link capacity $C = 25$ and partial overlap was considered in this paper. The streaming and interactive flows were presented as separate M/G/1 queuing systems. For each queue, 500 stochastic variables were generated in interval (1,100) for inter-arrival times with Poisson distribution, and the scaling factors λ_2 and λ_3;; and for service times with Erlang distribution, and the scaling factors μ_2 and μ_3; and the positive discrete values for time series $\{x^{(2)}\}$ and $\{x^{(3)}\}$ were obtained. The process generating the stochastic variables started with the equal seeds for each couple in the time series, and the synchronization helped with the variance reduction. We considered 15 target classes with an equal number of the target values as the PVQ. A smaller class face corresponded to a superior probability density link occupancy function as demonstrated in Figure 7. We optimized the PVQ applying the bipartition algorithm (Radev and Radeva, 2003),[10] which searched for the new boundaries of the target classes. The separation of the input space helped to determine the weight centers for the obtained classes. We separated the input values into $M-1$ couples of clusters. The bipartition algorithm separated the input space $\Re$ into two clusters $\Re_1$ and $\Re_2$, while $\Re_1$ coincided with class S_1. The weight center W_1 for cluster $\Re_1$ was defined with Kohonen neural network learning and training. The space $\Re_2$ was separated into two new clusters $\Re_{21}$ and $\Re_{22}$, while $\Re_{21}$ coincided with class S_2, and the weight center W_2 was defined. With use of the described procedure, the weight centers W_k, $(W_k^{(2)}, W_k^{(3)})$ of all classes were defined. The weight centers obtained after 100 training epochs are indicated in Figure 8 as the black points. Then we used the midpoint of Manhattan distances between the weight centers of the neighboring classes as the criteria of the optimal class boundaries. The number of the target values could be different within the new optimal class boundaries. We determined Manhattan distances between the neighboring weight centers according to (8).

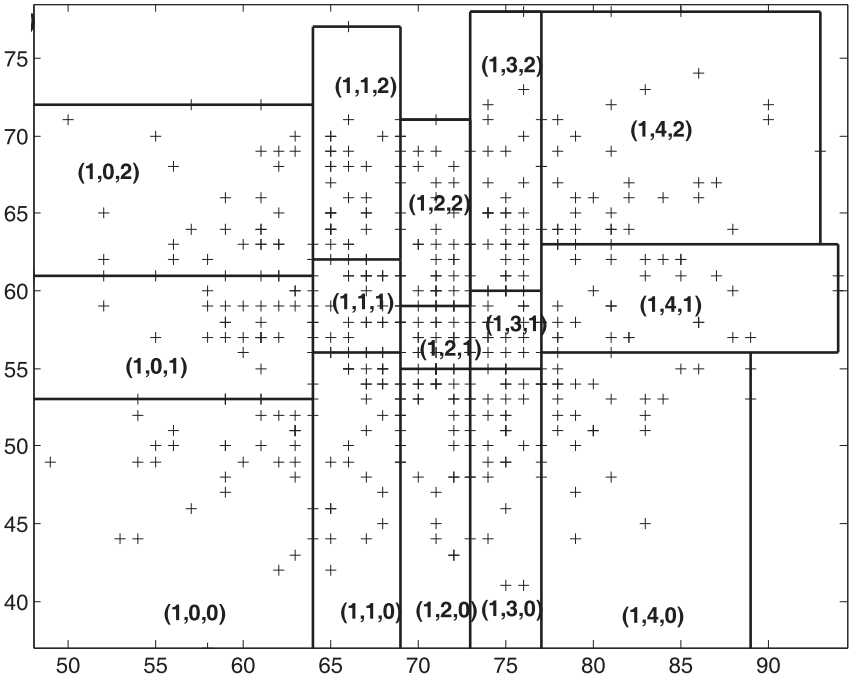

Fig. 7. Preliminary VQ with equal number of target values in each class

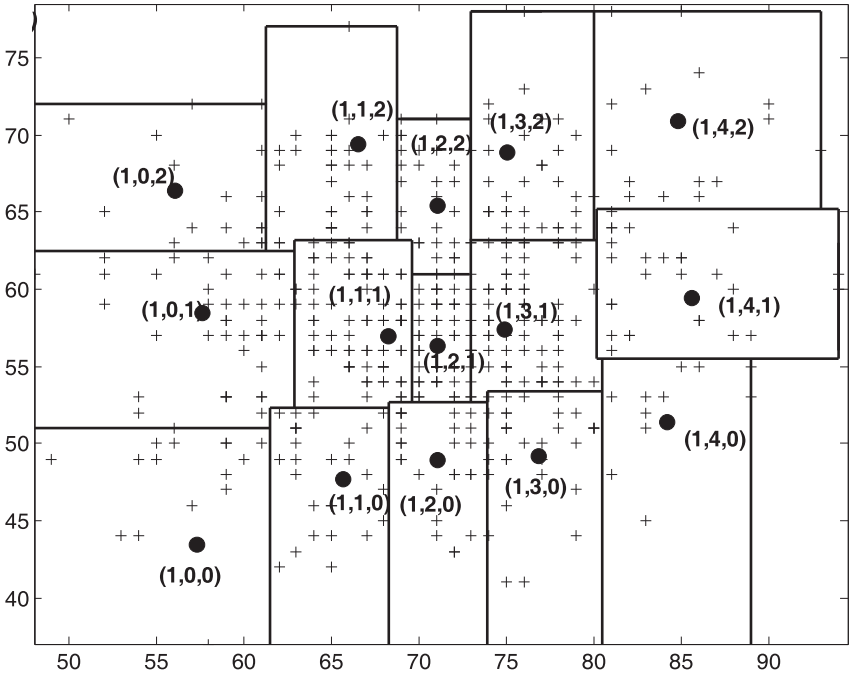

Fig. 8. Optimal boundaries between classes

$$D_{k,k+1} = \left\| W_k^{(2)} - W_{k+1}^{(2)} \right\| + \left\| W_k^{(3)} - W_{k+1}^{(3)} \right\| \quad (8)$$

The boundaries of the target classes were shifted parallel to the coordinate axis for all the classes until the calculated midpoint of Manhattan distances was going to be reached. Then we recalculated the numbers of the target values within the new rectangular classes as shown in Figure 8. We used the LVQ to obtain the approximated values of the occupancy probability distribution in the target classes for input time series in this phase. The accuracy of obtained results depended on the number of training epochs, and it increased with growth of

the number of the learned target values. The efficiency of the recommended clustering procedure is illustrated in Table 1, in which the occupancy probability density function was calculated for 15 feasible states, and the influence of the PVQ in shifting the class boundaries is also demonstrated. The LVQ probability density function is decreased for the both classes with the minimum and maximum values. For example, the minimum density distribution for class (1,0,0) is decreased from 1.406 to 1.189 and the maximum density distribution for class (1,2,1) is decreased from 28.104 to 23.74, respectively.

Cutbacks of LVQ density function and low values of density distribution in the classes with partial overlap are results of corrections of the link occupancy allocated strategy.

Table 1. Density Distribution, %

Class	PVQ	LVQ
(1,0,0)	1.405	1.189
(1,1,0)	4.818	5.241
(1,2,0)	7.026	6.674
(1,3,0)	8.029	5.435
(1,4,0)	2.342	1.702
(1,0,1)	3.513	3.160
(1,1,1)	11.242	14.202
(1,2,1)	28.104	23.740
(1,3,1)	16.863	14.943
(1,4,1)	3.011	2.759
(1,0,2)	1.916	2.327
(1,1,2)	4.497	3.998
(1,2,2)	1.044	9.011
(1,3,2)	4.684	4.375
(1,4,2)	1.506	1.248

6 CONCLUSIOND

In this paper, we presented the dynamic multi-service UMTS/IMT core network model with calls that belong to one of four service classes and arrive randomly, and provide its performance analysis. The arriving calls were granted service based on the specific service class, required maximum and minimum bandwidth, and available network resources. Performance analysis was based on priority-based dynamic capacity allocation that is suitable for the wireless ATM system. The scheduling of ATM cell transmission in each uplink TDMA frame was provided with the priority scheme. Blocking probability and throughput parameters for bandwidth sharing policy were considered, and partial overlap link was implemented.

The clustering procedure with Kohonen neural network, described as two-dimensional Markov reward model was developed and applied. Recommended approach included four phases of the neural modeling: the preliminary vector quantification, the weight centers determined with use of the bipartition, the class boundaries optimization, and the learning vector quantification. Developed model provided the opportunity to determine the optimal probability occupancy density function for the feasible steady states of the Markov chain. The efficiency of the probability occupancy density function used for the quick and easy performance analysis of the partial overlap allocation scheme was shown based on the obtained numerical results.

References

[1] RACZ, S.—TELEK, M.—FODOR, G.: Call level performance analysis of 3rd generation mobile core network, Proceedings of the IEEE International Conference on Communications, ICC, Helsinki, Finland **2** (2001), 456–461.

[2] XHAFA, A. E.—TONGUZ, O.: Dynamic priority queueing of handover calls in wireless networks: an analytical framework, IEEE Journal on Selected Areas in Communications **22** No. 5 (2004), 604–916.

[3] FANTACCI, R.: Performance evaluation of prioritized handoff schemes in mobile cellular networks, IEEE Transaction on Vehicular Technology **49** No. 2/ (2000), 485–493.

[4] ENEROTH, G.—FODOR, G.—LEIJONHUFVUD, G.—RACZ, A—SZABO, I.: Applying ATM/AAL2 as switching technology in third-generation mobile access networks, IEEE Communication Magazine **38** No. 6 (1999), 112–122.

[5] GANESHBABU, T.—LE-NGOC, T.—HAYES, F.: Performance of a priority-based dynamic capacity allocation scheme for wireless ATM systems, IEEE Journal on Selected Areas in Communications **19** No. 2 (2001), 355–369.

[6] RADEV, D.—LOKSHINA, I.—RADEVA, S.: Clustering procedure for performance analysis of mobile communication networks, Proceedings of the EUROSIS-ETI 2005: 3rd Industrial Simulation Conference (ISC'05) at Fraunhofer IPK, Berlin, Germany,, 83–87.

[7] KULKARNI, V. G.: Modeling and Analysis of Stochastic Systems, Chapman Hall, 1995.

[8] RADEV, D.—LOKSHINA, I.—RADEVA, S.—INSINGA, R.: Neural Modeling of Link Occupancy Distribution for Broadband Telecommunication Transmission, Neural Modeling of Link Occupancy Distribution for Broadband Telecommunication Transmission, Proc. of the DSI 2004: 35-th Annual Meeting of the Decision Sciences Institute, Boston, USA (2004), 3001-3009.

[9] HOFMANN, T.—BUHMANN, J.: Competitive Learning Algorithms for Robust Vector Quantization, IEEE Transaction on Signal Processing **46** No. 6 (1998), 1665-1675.

[10] RADEV, D.—RADEVA, S.: Artificial Intelligence Modeling of Stochastic Processes in Digital Communication Networks, Journal of Electrical Engineering **54** No. 9-10, 255–259.

Received 1 July 2006

Dimitar Radev PhD is an Associate Professor at Department of Communication Technique and Technologies at University of Rousse, Bulgaria. His research interests are connected with Teletraffic Theory, Simulation and Modeling of Communication Networks, and Performance Analysis of Queuing Systems.

Izabella Lokshina, PhD is an Associate Professor of Management Information Systems and Academic Coordinator of MMI at Division of Economics and Business at SUNY Oneonta, USA. Her research interests include Fuzzy Systems and Neural Networks, and System Modeling.