Hidden Markov Model based Resource Estimation and Call Admissions in Mobile Cellular Networks

Vinay K. Bharadwaj *

ECE Department
Rice University, Houston
Texas, 77030 USA
vickyvin@rice.edu

Abhay Karandikar

Department of Electrical Engineering
Indian Institute of Technology–Bombay
Powai, Mumbai - 400076, India
karandi@ee.iitb.ernet.in

ABSTRACT

A new call admission algorithm with better performance guarantees is proposed with emphasis on decoupling of the given cell from communication with its neighbours. The algorithm is formulated with the view that the bandwidth requirements of anticipated incoming and outgoing handoff calls can be trained. Accordingly, a Hidden Markov Model (HMM) based resource estimation and call admission scheme is suggested. A novel way of maintaining the imposed performance metrics is outlined. Simulation results validate the performance of the algorithm against the existing ones.

1. INTRODUCTION

During the last few years, there has been an explosion of research in cellular mobile communications. The future generation of mobile networks are expected to support multimedia traffic like voice, video and data with varying bandwidth requirements. Since the bandwidth available in a mobile network is limited, it should be utilized in the most efficient manner. As an active mobile call moves from one cell to another, the required resources are allocated in the destination cell. This phenomenon is called Handoff. If sufficient resources are not available in the destination cell, then the handoff call may get terminated. A base station in a cell may, therefore, be required to reserve resources for the anticipated handoff traffic from other cells. This may sometimes mean denying access to a new call originated within the cell, even if resources are available.

Since reserving resources for handoff traffic would increase the probability of new calls getting blocked, there is always a trade–off between the two Quality of Service (QoS) metrics, *ie*, probability of call blocking and probability of handoff dropping. Hence, one of the most important aspects in the design of a call admission algorithm in a cellular network is to strike a balance between the bandwidth utilization and the probability of handoff call termination.

*This work was performed while the author was at IIT–Bombay

One common call admission scheme proposed in the literature is the guard channel scheme [1, 2]. The guard channel scheme gives higher priority to handoff calls. In this scheme, a portion of the bandwidth (or a set of channels) is always reserved for handoff. Instead of reserving a fixed portion of the bandwidth, a distributed call admission scheme proposed in [3] takes into account the state of neighbouring cells. This is done by base station exchanging state information with the adjacent cells periodically [3].

In [4], the amount of bandwidth to be reserved is dynamically varied using a concept called "Shadow Cluster". Shadow Cluster Concept can be viewed as a message system where mobile terminals inform the base stations in their neighbourhood about their requirements, position and movement parameters. With this information, base stations predict future demands, reserve resources accordingly, and admit only those mobile terminals which can be supported adequately.

The fundamental premise of the shadow cluster concept is that every mobile terminal with an active wireless connection exerts an influence upon the cells in the vicinity of its current location and along its direction of travel. As an active mobile terminal travels to other cells, the region of influence moves, following the active mobile terminal to its new location. The base stations currently being influenced are said to form a shadow cluster. The shadow is strongest near the active mobile terminal and fades away depending on factors such as the distance to the mobile terminal, current call holding time and priority, bandwidth resources being used, and the mobile terminal's trajectory and velocity.

In a cellular network with shadow cluster implemented on it, every base station must inform its neighbours about the future location probabilities of the active mobile terminals currently under its control. For each active mobile terminal, only those base stations which belong to that particular mobile terminal's shadow cluster are informed about these probabilities. For a detailed analysis involving the calculation of these mobile probabilities and the associated call admission control algorithm, the reader is directed to [4]. The main drawback of this algorithm, is that it requires complicated communication between base stations of neighbouring cells. Additionally, an estimate of the call holding time is required to compute the number of cells that the mobile will be visiting in near future.

In [5, 6], the cellular network is assumed to be the lower layer of wireline network. An architecture and an admission policy has been proposed using “Virtual Connection Tree” concept for minimizing call controller invocation overheads due to high rate of handoffs. In [7, 8], the author has formulated the states of the cells in terms of mobile platforms active within the cell and then resolved the state transition dynamics in terms of call arrival and departure processes.

In this paper, we formulate the problem of resource estimation and call admission in cellular networks using Hidden Markov Model (HMM). Section 2 explains the motivation for training based call admissions followed by the algorithm in Section 3. We give a brief overview of HMM, then discuss our scheme and a way to compute the decision threshold. The simulation results of the algorithm are discussed in Section 4 followed by conclusions in Section 5

2. MOTIVATION FOR TRAINING

We assume a cellular network system where the incoming handoffs from neigbouring cells have different bandwidth requirements. As assumed in [4], the bandwidth requirements of a mobile call can be expressed in terms of bandwidth units bu. We consider that a base staion has a total of C bandwidth units. Based on an analysis of the drawbacks of existing solutions briefly discussed above, we emphasize on the following criteria when formulating the new call admission algorithm.

- Decouple the interaction between neighbouring base stations.
- Exploit the uniqueness of geographic location of every cell.
- Should be easily implementable in every base station.

We concentrate our attention on a particular cell called the reference cell. A call admission algorithm employed by the reference base station should take into consideration the network state of neighbouring cells and estimate the bandwidth units that would be required by future handoff traffic into the reference cell.

The central concept is that our reference base station is expected to encounter similar handoff traffic patterns as it operates at a fixed location and in a reasonably fixed environment. The base station can, therefore, learn future bandwidth requirements of handoff traffic from past measurements. Our adaptive learning algorithm is based on Hidden Markov Model (HMM) that has been used in pattern recognition and image processing applications.

3. FORMULATION OF RESOURCE ESTIMATION AND CALL ADMISSION

3.1. A Brief Outline of HMM

HMMs are widely used in image training and pattern recognition problems. It is a generic tool, which when trained to encode a characteristic observation sequence in such a way that if an observation sequence having many characteristics similar to the one encountered earlier is given, it should be able to identify it.

The idea is to use HMM to train the incoming and outgoing handoff traffic patterns in a cell. Once the HMM parameters are trained to these sequences, then during the admission of a new call, future handoff traffic requirements can be estimated using the trained HMM model.

A Hidden Markov model (HMM) is a Markov model in which the state transitions are not directly visible but are observed through the feature vectors. These feature vectors are assumed to have a probability distribution within each state. An HMM can, therefore, be characterized by the following parameters :
N= number of states in the HMM;
M= total number of observation symbols;
$V = \{v_1, \dots, v_M\}$ the discrete set of possible observation symbols
T = length of observation sequence, *ie*, the number of observed symbols;
i_t denotes the state of the HMM at time t
$\pi = \{\pi_i\}$, $\pi_i = Pr(i_1 = i)$ is the probability of being in state i at the beginning of the state sequence;
$A = \{a_{ij}\}$ where $a_{ij} = Pr(i_{t+1} = j \mid i_t = i)$ is the probability of being in state j at time $t+1$ given that the HMM is in state i at time t ;
$B = \{b_j(v_k)\}$ where $b_j(v_k) = Pr(v_k \text{ at } t \mid i_t = j)$ the probability of observing the symbol v_k at time t given that HMM is in state j at time t;
O_t = observation symbol vector observed at instant t.

Given an HMM $\Lambda = (A, B, \pi)$ the following questions are pertinent :

- *Probability of an Observation Sequence:* Given the model $\Lambda = (A, B, \pi)$, how do we compute $P(O|\Lambda)$, the probability of occurence of the observation sequence $O = O_1, O_2, \dots, O_T$.
- *Finding an optimal state sequence:* Given model Λ, how is the optimal state sequence I = $i_1, i_2, \cdots, i_T$ chosen so that the probability $P(O, I \mid \Lambda)$, where $O = O_1, O_2 \cdots, O_T$, is maximized? The Viterbi algorithm [9] is used to find this optimal state sequence.
- *Training the HMM:* Tuning the model parameters of Λ so that $P(O, I \mid \Lambda)$ is maximized constitutes the training phase of the model. This is done using the Baum-Welch Re-estimation Formulae or Segmental k-Means Algorithm [9].

For a detailed analysis, the reader is directed to [9, 10]

3.2. HMM based Call Admission Scheme

We consider a discete time system. The time axis is assumed to be slotted. The calls are generated or handed off only at the end of a slot. Since the forced termination of an ongoing call is more disturbing to a subscriber than the inability to initiate a new call, it is therefore reasonable to assume that the cost of a handoff traffic rejection is more than the cost of rejecting a new call. Under this assumption, it has been proved in [12] that an optimal policy should always accept a handoff request if the bandwidth requirements can be met. Accordingly, we also assume that handoff calls are always admitted into the cell if the bandwidth units requested by an handoff call are available at the time of handoff. Hence, the control policy is to decide whether to accept or reject a new call.

Let C be the total bandwidth units available in our reference cell. Let a be the bandwidth units requested by a new call at the current time interval. (Let us assume it to be $t = 0$.) Let C_t be the bandwidth units in usage at t^{th} time slot. Let us assume that a new call is generated at time $t = 0$. If the bandwidth requested by the new call is not available, then the call is blocked. Otherwise, we determine whether the probability that the bandwidth will be available throughout the call dwell time is appropriately high or not. In other words, we admit a new call if the following condition is satisfied.

$$Pr[C_t + a \leq C] > Th_t \qquad \forall t = 1, 2 \dots \eta \tag{1}$$

Where η is the call dwell time in slots and Th_t is the value of threshold that can be determined from the required QoS guarantees. The call dwell time can be estimated from the mobile's position, velocity and cell size. By admitting a new call, if the bandwidth usage exceeds the total available bandwidth at any time, then some calls will be forcely terminated. The value of threshold Th_t is thus related to the probability of forced termination of a call. The determination of the threshold parameter will be discussed in the subsequent section.

Let H_t^i denote the amount of bandwidth units requested in slot 't' by incoming handoff traffic. Similarly, let H_t^o denote the amount of bandwidth units released in slot 't' by outgoing handoff traffic plus call termination traffic. It can then be shown [13] that

$$\begin{aligned} Pr[C_t + a \leq C] &= Pr[C_t \leq C - a] && (2)\\ &= \sum_{r=0}^{C-a} \sum_{l=0}^{C} Pr[C_{k-1} = l] Pr[Z_t = r - l] && (3)\\ Pr[Z_t = b] &= \sum_{l=0}^{C} Pr[H_t^o = l] Pr[H_t^i = b + l] && (4) \end{aligned}$$

where $Z_t = H_t^i - H_t^o$. Note that the value of C_0 is known at time $t = 0$. We are now left with the problem of determining decision probabilities or in otherwords $Pr[Z_t]$. These probabilities can be determined using HMM based training.

Let our HMM be denoted by $\Lambda = (A, B, \pi)$ where A is the state transition matrix, B is the observation symbol probability matrix and π is the initial state distribution vector [9, 10]. This HMM learn the difference sequence $Z = Z_1, Z_2, \dots,$. Training can be done using Baum-Welch Re-estimation formulae or Segmental K-means Algorithm [9, 10]. Once the HMM is trained, the above probabilities can be computed as follows.

$$Pr[Z_t = l|\Lambda] = \sum_{\forall i_1 i_2 \dots i_t} a_{i_0 i_1} a_{i_1 i_2} \dots a_{i_{t-1} i_t} b_{i_t}(Z_t) \tag{5}$$

where a_{ij} and $b_{i_t}(Z_t^i)$ are the transition probabilities and observation symbol probabilities for HMM model Λ.

The state i_0 can be computed by determining the optimum state sequence $I_- = i_{-T} \dots i_{-1}, i_0$ using the Viterbi Algorithm [9, 10] so that $Pr(Z_-, I_-|\Lambda)$ is maximized.

3.3. Determination of Decision Threshold

In this section, we determine the threshold Th. As stated above in our scheme, a new call is blocked if $P_t = Pr[C_t + a \leq C] < Th_t$, for any $t = 1, 2, \dots, \eta$. Let *pb* denote the probability of a new call blocking. Then we have

$$pb = 1 - P_0 P_1 . P_2 \dots P\eta \tag{6}$$

For maintaining a specified value of *pb*, this equation can be used to determine the thresholds Th_t's in Equation 1 in each time interval.

During any time slot, if the bandwidth usage in the cell exceeds the total bandwidth available, some of the calls will be forcibly terminated. The probability of forced termination of a call is given by

$$pf = Pr[C_t + a > C] \qquad \forall k = 1, 2, \dots, \eta \tag{7}$$

If a minimum probability of termination of existing calls pf_{min} is to be maintained by the network, then it may impose a stronger threshold Th_t for the probabilities in Equation 1. Hence for a new call to be admitted, both the thresholds imposed by pb and pf_{min} have to be satisfied in all time slots within its residence time, else the new call will be blocked.

4. PERFORMANCE EVALUATION OF ALGORITHM

We have evaluated the performance of our algorithm by simulations. The simulation parameters are similar to that of [4] and are stated as follows.

- A series of 10 cells of width 1 km spanning 10 km of a highway is considered. New calls originate in each of the cells with an exponential inter-arrival time. So the new call arrivals are Poissonian.
- New call can be of 3 traffic types: Video (3 bu), Audio (2 bu) and Voice (1 bu)
- New Call holding time for all traffic types is the same.
- The mobile from which new calls originate can have speeds 70, 90 or 100 km/h. The probability of each speed is 1/3 with a uniform distribution. The mobile can move in either direction with a probability of 1/2, with the random deviates picked from a uniform distribution.
- Each cell has a capacity of $C = 30\ bu$.
- Time axis is slotted in intervals of 10s.
- Traffic is simulated for 480 hours. The first 240 hours are used for training of HMM parameters and the next 240 hours are used for testing the call admission algorithm.

The training is done on the difference sequence $Z = Z_1, Z_2 \ldots$, where $Z_t = H_t^i - H_t^o$. A left-right HMM model is used. In a left-right HMM model, the transition probabilities $a_{ij} = 0$ for $j > i + 1$ and $j < i$ (for $i = 1, 2, \ldots, N$).

We have compared the performance of our algorithm with the guard channel or constant resource reservation algorithm. We consider three algorithms–one with no resevation of bandwidth (CRR0), second with 10% reservation of bandwidth (CRR1) and the third with 20% reservation of bandwidth. Let c_pf0, c_pf1 and c_pf2 denote the probability of forced termination and c_pb0, c_pb1 and c_pb2 denote the probability of call blocking for the three algorithms respectively. Similarly, let c_bwu0, c_bwu1 and c_bwu2 denote the average bandwidth utilizations for these algorithms. The constant resource reservation algorithm–CRR0 with zero percent reservation is equivalent to random admissions of the calls irrespective of whether they are new or handoff calls.

4.1. Simulation Results

The simulation results in Figure 1 illustrate the variation of new call load in terms of average connection requests per second. We assume a threshold on decision probabilities Th to be equal to 0.97. The number of HMM states N taken for simulation is 5. The results show that the forced termination probability of CRR1 and CRR2 is lower than that of our algorithm but this is at the cost of very high blocking probability (50-60%) when compared to our algorithm ($<$ 25%). The cost incurred by the HMM based scheme for maintaining lower forced termination probability and blocking probability is a slight decrease in the average bandwidth utilization. And that too, such a trace is observed only at higher new call arrival rates. CRR0 performs better in blocking probability and bandwidth utilization but its forced termination probability is much higher.

The simulation results in Figure 2 indicate that as the threshold on the decision probabilities becomes more stringent, the forced termination probability decreases with an increase in the blocking probability. In our simulations, the number of new calls is roughly around half the number of handoff calls. This simulation has these numbers around 1466 and 2364 respectively. At a threshold of 0.2, 329 new calls are blocked and 898 handoff calls are dropped. At a threshold of 0.99, 571 new calls are blocked and 709 handoff calls are dropped. Summing up, a small fraction of around 50 calls are denied access when the threshold increases. So for a slight sacrifice in the bandwidth utilization, the forced termination probability– a QoS metric is brought down appreciably.

5. CONCLUSIONS

In this paper, we have addressed the problem of resource estimation and call admission within a training based framework. The emphasis has been one decoupling of the given cell from communication with its neighbours. We have formulated the algorithm with the view that the presence of the surrounding cells which is reflected in the amount of bandwidth units of anticipated incoming and outgoing handoff calls can be trained. We have evaluated the performance of the algorithm by simulations. For simulation purposes, we have considered the case of rectangular cells laid across a highway. The algorithm performs better than the constant reservation algorithms in terms of maintaining lower forced termination and blocking probabilities. We have inroduced a new measure called Decision Probabilities that is found to have a good control over the QoS metrics. The simulation results also indicate that as the threshold on the decision probabilities increases, the forced termination probability is brought down appreciably for a slight sacrifice in bandwidth utilization.

REFERENCES

[1] Carlos Oliveira, Jaime Bae Kim and Tatsuya Suda, "Quality-of-Service Guarantee in High-Speed Multimedia Wireless Networks", *ICC'96*, pp 728-734.

[2] Roch Guerin, "Queueing-Blocking System with Two Arrival Streams and Guard Channels", *IEEE Trans. Comm.*, Vol. 36, Feb. 1988, pp.153-163.

[3] M. Naghshineh and Mischa Schwartz, "Distributed Call Admission Control in Mobile/Wireless Networks" *IEEE Journal Select. Areas Comm.*, vol. 14, May 1996, pp 711-717.

[4] D. A. Levine, I. F. Akyildiz and M. Naghshineh, "A Resource Estimation and Call Admission Algorithm for Wireless Multimedia Networks Using the Shadow Cluster Concept" *IEEE/ACM Trans. Networking*, Vol. 5, Feb. 1997, pp 1-12.

[5] A.S. Acampora and M. Naghshineh, "An Architecture and Mehodology for Mobile Executed Handoff in Cellular-ATM Networks", *IEEE Journal Select. Areas Comm.*, Oct. 1994, pp 1365-1375.

[6] A.S. Acampora and M. Nagshineh, "Control and QoS provisioning in High Speed Microcellular Networks", *IEEE Pers. Comm.*, second quarter, 1994.

[7] S.S. Rappaport, "The Multiple-Call Handoff Problem in High Capacity Cellular Communication Systems", *IEEE Trans. Vehicular Tech.*, Aug. 1991, pp.546.

[8] S.S Rappaport, "Blocking, Handoff and Traffic Performance for Cellular Communication Systems with Mixed Platforms" *IEE Proc-I*, Oct. 1993, pp.389.

[9] Rakesh Dugad and U.B.Desai, "A Tutorial on Hidden Markov Models" *Tech. Rep., SPANN Lab, IIT Bombay*, May 1996.

[10] L.R. Rabiner and B.H. Juang, "An introduction to Hidden Markov Models", *IEEE ASSP Mag.*, June 1986, pp4-16.

[11] Julian Keilson and Oliver C. Ibe, "Cutoff Priority Scheduling in Mobile Cellular Communication Systems", *IEEE Trans. Comm.*, Vol. 43, Feb./Mar./April 1996, pp.1038-1045.

[12] Manjari Asawa, "Optimal Admissions in Cellular Networks with Handoffs", *ICC'96*, pp.1143-1147.

[13] Vinay K. Bharadwaj, "A Framework for Call Admissions in Mobile Cellular Networks", *MTech Thesis, IIT-Bombay*, June 1998.

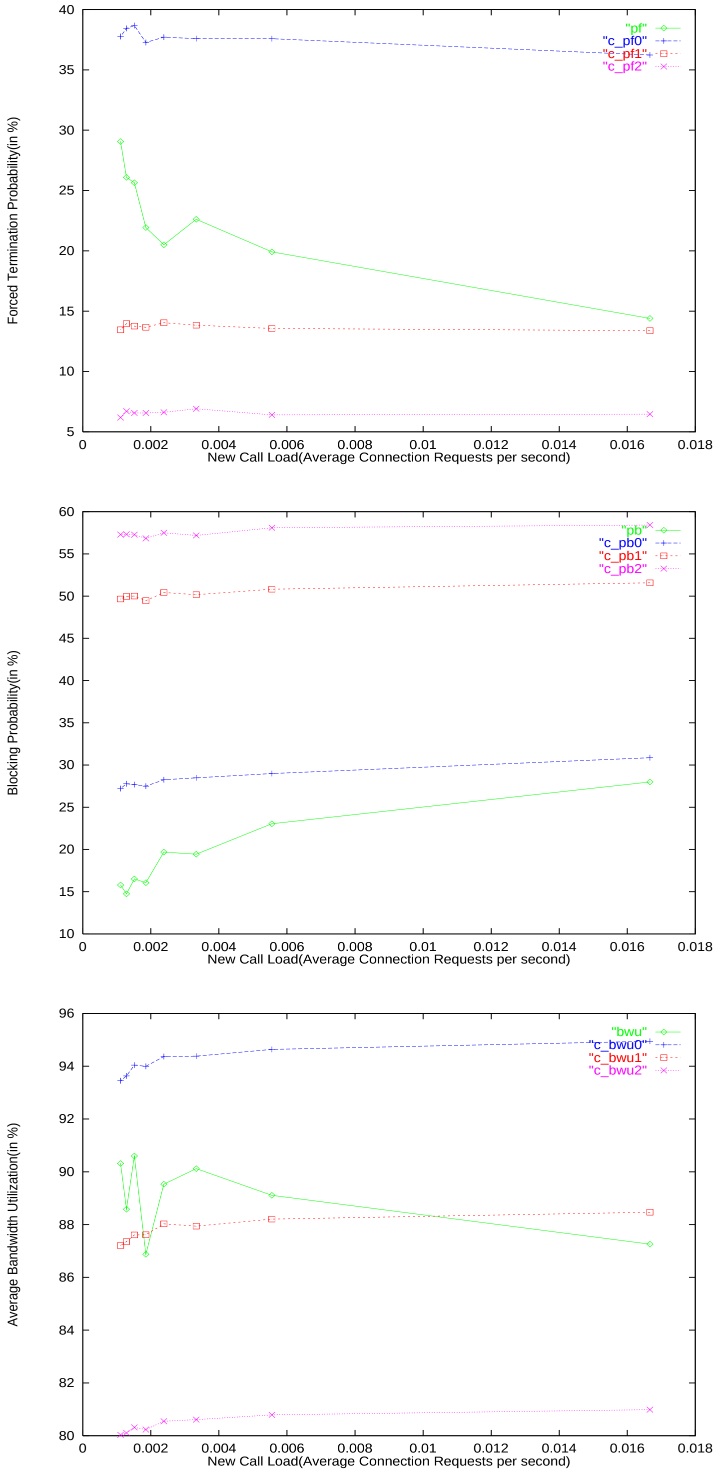

Figure 1: Simulation 1: Effect of New Call Arrival Rate (Mean Call Duration = 30s, Th = 0.97, N = 5, C = 30 bu)

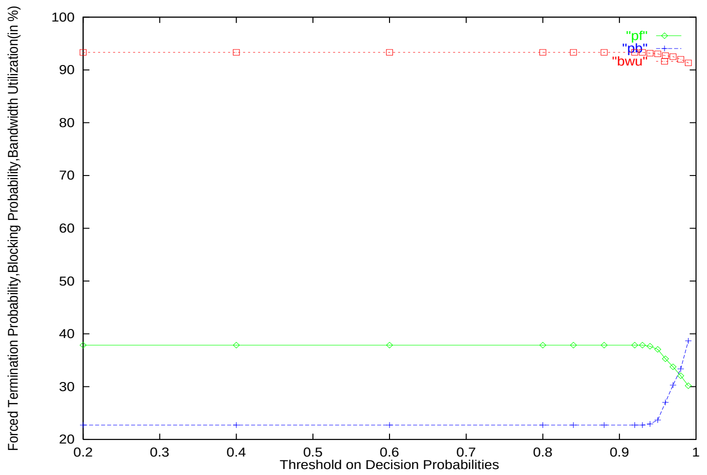

Figure 2: Simulation 2: Effect of Threshold (Mean New Call Interval = 600s, Mean Call Duration = 60s, C = 30 bu)