CALL ADMISSION CONTROL IN WIRELESS MULTIMEDIA NETWORKS

Novella Bartolini[1], Imrich Chlamtac[2]

[1] Dipartimento di Informatica, Università di Roma "La Sapienza", Roma, Italy
novella@dsi.uniroma1.it
[2] Center for Advanced Telecommunications Systems and Services (CATSS)
University of Texas at Dallas, Dallas, Texas, U.S.A.
chlamtac@utdallas.edu

Abstract - This paper addresses the call admission control problem for the multimedia services that characterize the third generation of wireless networks. In the proposed model each cell has to serve a variety of classes of requests that differ in their traffic parameters, bandwidth requirements and in the priorities while ensuring proper quality of service levels to all of them. A Semi Markov Process is used to model multi-class multimedia systems with heterogeneous traffic behavior, allowing for call transitions among classes. It is shown that the derived optimal policy establishes state-related threshold values for the admission policy of handoff and new calls in the different classes, while minimizing the blocking probabilities of all the classes and prioritizing the handoff requests. It is proven that in restrictive cases the optimal policy has the shape of a Multi-Threshold Priority policy, while in general situations the optimal policy has a more complex shape.

Keywords - Cellular Networks, Semi Markov Decision Processes, Handoff, Call Admission Control, Randomized Policies.

I. Introduction

The advent of the third generation of wireless multimedia services, brought about the need to adapt the existing mobile cellular networks to make them carry the various classes of multimedia traffic, voice, video, images, web documents, data or a combination thereof.

The user needs to have ubiquitous access to a wide variety of services while roaming throughout the area covered by the wireless network. In these systems the base station of each cell in the network has to serve several streams of requests, corresponding to multimedia services. The wireless multimedia network has to be able to support multiple classes of traffic with different Quality of Service (QoS) requirements, i.e., different *number of channels* needed, *holding time of the connection*, *cell residence time* and *priority*. Furthermore, the system will have to support dynamic transitions of the calls among classes, that may be required at run time, according to the users' needs, corresponding to bandwidth adaptation requirements and to type of service changes required at run-time.

In order to support multimedia applications, the system capacity has to be increased significantly. A common way of achieving this is by employing microcells and picocells [1]. The reduction in cell size leads to an increased spectrum-utilization efficiency, but also results in more frequent handoffs and makes connection level QoS more difficult to achieve.

Since it is impractical to completely eliminate handoff drops, the next best alternative is to provide a *probabilistic* guarantee on quality. The system can reduce the chances of unsuccessful handoff by assigning higher priority to handoff requests than that assigned to initial access requests belonging to the same class, because from a user's perspective, the forced termination of an ongoing call is less desirable than getting a busy signal due to the block of an initial access attempt. At the same time the system must guarantee a proper prioritization among requests of different classes. In fact, emergency, or very high priority new requests, can have higher priority than handoff requests of a low priority class. For this reason the choice of the Call Admission Control (CAC) policy plays a major role impact in multimedia wireless networks.

The purpose of this paper is to propose a model to study and optimize the QoS parameters that characterize the traffic in a multi-class environment, by means of access control policies.

Several works have been proposed to analyze the problem of CAC in a multi-class environment. Some of them [5], [7] only considered two classes while comparing different reservation strategies. In [13] a call admission policy named Hybrid Cutoff Priority Policy (HCPP) is considered while evaluating the performance of a model that does not allow for class transitions, while in [11], [12] SMDP models are proposed to be used at runtime to improve the QoS by means of CAC policies, without analyzing the structure of the optimal solution and without considering the possibility of having class transitions of calls, except in [11], where a limited bandwidth requirement adaptation is considered to solve the problem of the highly varying resource availability.

In this paper we propose a model that takes into account an heterogeneous, multi-class environment, that allows for call transitions among classes, responding to multimedia traffic requirements of 3G wireless systems. This model is exploited to evaluate the behavior of a wide class of policies and is optimized by means of stochastic analysis with SMDPs. Linear programming methods permit to discard non stationary and randomized policies from the search for the optimum. The structural analysis of the optimal long-run cost function is instead realized through dynamic programming methods that allow the structural analysis of the optimal CAC policy.

It is shown that the optimal policy establishes state-related threshold values for handoff and new calls in the different classes. It is also shown that under certain assumptions the optimal policy has the shape of what we call *Multi-priority Threshold Policy* (MTP) that, to the best of our knowledge, has not been proposed before. MTP is an adaptation of a resource utilization policy proposed in [4] under the name of *Upper Limit* (UL) policy to solve the problem of admitting multi-rate traffic in ATM wireless networks without taking into account the handoff prioritization problem. Under MTP two threshold values are defined for each class, T_i^{New} and T_i^{Hoff}. A class-i call is accepted as long as enough bandwidth is available and the total number of busy channels is such that the acceptance of another class-i call does not cause the occupancy level to exceed the threshold T_i^{New} or T_i^{Hoff} depending on if the call is new or if it is a handoff call respectively. The handoff calls of the highest priority stream of request are always accepted.

The definition of a threshold value even for handoff calls allows for the case in which a low priority handoff call could be denied service in the hope to make place for high priority calls, even if newly originated (e.g., emergency new calls).

The results of our analysis have an immediate practical application as the optimal policy can easily be computed once basic statistic parameters defining the traffic of requests are known, by solving an optimization problem by means of very commonly used methods of operations research [8], [10], [14], [15].

II. A Multi-class Semi Markov Decision Model for CAC

We consider a decision model that allows for multiple class calls that differ in their QoS requirements and traffic parameters, allowing for call transitions among classes.

Each cell in the system has to give service to N classes of call requests that we assume to be generated according to a Poisson distribution with mean arrival rates λ_i^{New} $0 \le i \le N$ for new calls, λ_i^{Hoff}, $0 \le i \le H$ for handoff calls, where the condition $H \le N$ takes into account the possibility that some classes of traffic do not have handoffs. There are C channels available in each cell assuming a Fixed Channel Allocation (FCA) scheme. Each new or handoff call of class i has a class dependent bandwidth requirement that is measured in number of unit channels and represented with b_i.

The unencumbered communication time and the *cell residence time* of a class-i call are assumed to be exponentially distributed with mean rate μ_i^c and μ_i^e respectively. We use $\mu_i \triangleq \mu_i^c + \mu_i^e$, $\forall i \in \{1, 2, \dots, N\}$.

We now introduce the notation that will be used throughout the section.
The state of the SMDP is represented through an N-dimensional vector $\mathbf{x} \triangleq (x_1, x_2, \dots, x_N)$ in which the variable x_i represents the number of ongoing class-i calls. The state space Λ is defined in the following way

$$\Lambda \triangleq \left\{ \mathbf{x} : \sum_{i=1}^{N} b_i \cdot x_i \le C; x_i \ge 0 \right\}.$$

When the system is in state $\mathbf{x}$, an accept/reject decision must be made for each type of possible arrival, i.e., an origination of a class-i new call, or the arrival of a class-i handoff call. Thus, the action space $\mathcal{B}$ can be expressed by

$$\mathcal{B} = \Big\{ (a_1^{New}, a_2^{New}, \dots, a_N^{New}, a_1^{Hoff}, \dots, a_N^{Hoff}) : \\ a_i^{New}, a_j^{Hoff} \in \{0,1\}, i, j = 1, \dots, N \Big\}.$$

where the indicators a_i^{New} and a_j^{Hoff} denote the acceptance (1) or the denial of service (0) of class-i new calls or handoff respectively.

A call may transit from one class to another satisfying bandwidth adaptation requirements due to the traffic conditions, or according to the user's needs to change type of service at runtime. This could happen when a user asks for new additional services while having an ongoing connection. For example a user with an ongoing voice call could ask for additional bandwidth to set a video-conference without interrupting his connection. In this case the call would transfer from the voice class to the video class.

The transition from a class i to another class j occurs with negative exponential distribution with mean rate π_{ij}. These transitions take place between states characterized by a different number of busy channels (in bandwidth units) but with the same total number of ongoing calls as shown in Figure 1.
If the system is in state $\mathbf{x} \in \Lambda$ and the action $\mathbf{a} \in \mathcal{B}$ is chosen, then the dwell time of the state $\mathbf{x}$ is $\tau(\mathbf{x}, \mathbf{a})$ where

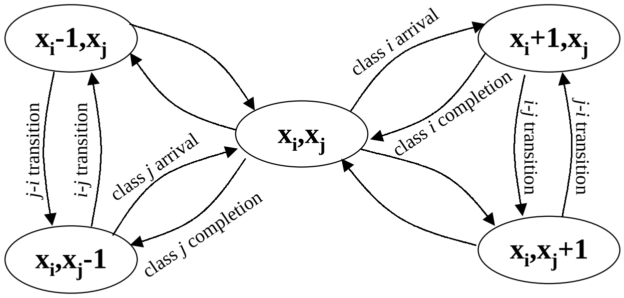

Fig. 1
A 2-class model with class transitions.

$$\tau(\mathbf{x},\mathbf{a}) = 1/\left[\sum_{i=1}^{N}(\lambda_i^{New}a_i^{New} + \lambda_i^{Hoff}a_i^{Hoff} + \right. \\ \left. +x_i\mu_i + \sum_{j=1}^{N} x_i\pi_{ij})\right] \quad (1)$$

where $\pi_{ii} \triangleq 0 \forall i \in \{1, \cdots, N\}$.

The system is forced to pay a penalty H_i if a handoff call of class i is refused. If the service is denied to an initial attempt of access in class i, the system will have to pay a lower penalty $L_i \leq H_i$. It may happen that $H_i < L_j$ for some indexes $i \neq j$, meaning that the priority of class-j calls is higher than the priority of class i enough that even a class-i handoff call may be interrupted in the hope of making place for new calls in class j. This can happen in presence of very high priority traffic, like emergency calls.

Equation 1 shows the dependency of the dwell time of the states of the process both on the decision $\mathbf{a}$ and on the state $\mathbf{x}$. The set of rate that characterizes the process is bounded above by the maximum outgoing rate. Hence we conclude that the process is uniformizable. Adding dummy transitions from states to themselves, a uniform Poisson process can be constructed which governs the epochs at which transitions take place.

The uniform rate will be an upper bound on the total outgoing rate from each state. The following definition of Γ fits our needs.

$$\Gamma = \sum_{i=1}^{N}[C\mu_i + \lambda_i^{Hoff} + \lambda_i^{New} + C\sum_{j=1}^{N}\pi_{ij}]. \quad (2)$$

The transition probabilities of the uniformized process are then

Class-i arrival: $\mathbf{y} = \mathbf{x} + \mathbf{e_i}$

$$\tilde{p}_{\mathbf{xy}}^{\mathbf{a}} = (\lambda_i^{new} \cdot a_i^{new} + \lambda_i^{hoff} \cdot a_i^{hoff})/\Gamma \quad \forall i \in \{1, 2, \ldots, N\} \\ \text{where} \quad a_i^{new} = a_i^{hoff} = 0 \quad \text{if} \quad \mathbf{y} \notin \Lambda. \quad (3)$$

Class-i departure: $\mathbf{y} = \mathbf{x} - \mathbf{e}_i$

$$\tilde{p}_{\mathbf{xy}}^{\mathbf{a}} = x_i \cdot \mu_i/\Gamma \quad \forall i \in \{1, 2, \ldots, N\}. \quad (4)$$

Class $i - j$ transition: $\mathbf{y} = \mathbf{x} - \mathbf{e_i} + \mathbf{e_j}$

$$\tilde{p}_{\mathbf{xy}}^{\mathbf{a}} = x_i \cdot \pi_{ij} \cdot a_{ij}^{trans}/\Gamma \quad \forall i \in \{1, 2, \ldots, N\}, \\ \text{where} \quad a_{ij}^{trans} = 0 \quad \text{if} \quad \mathbf{y} \notin \Lambda \quad \text{and} \quad a_{ij}^{trans} = 1 \\ \text{otherwise}. \quad (5)$$

Dummy transitions from each state to itself: $\mathbf{y} = \mathbf{x}$

$$\tilde{p}_{\mathbf{xy}}^{\mathbf{a}} = \frac{1}{\Gamma} \cdot [\Gamma - \sum_{i=1}^{N}(\lambda_i^{new}a_i^{new} + \lambda_i^{hoff}a_i^{hoff} + x_i\mu_i + \\ + \sum_{j=1}^{N} x_i a_{ij}^{trans}\pi_{ij})] \\ \forall i \in \{1, 2, \ldots, N\}. \quad (6)$$

Otherwise: the transitions between other states that are not considered in this list have probability 0.
The uniformized cost function is:

$$\tilde{r}(\mathbf{s},\mathbf{a}) = \frac{1}{\Gamma}\sum_{i=1}^{N}[\lambda_i^{New}(1-a_i^{New}) \cdot L_i + \lambda_i^{Hoff}(1-a_i^{Hoff}) \cdot H_i]. \quad (7)$$

Studying the properties of the described uniformized process, it can be confirmed, without loss of generality, that the decision model can be restricted to include the only processes with no transient states and with only one communicating class, i.e., to the only *unichain* processes.

We indicate with $\mathcal{S}$ the finite set of all feasible couples of vectors of the kind (*state, decision*). The unichain property, together with the finiteness of $\mathcal{S}$ implies the existence of a unique stationary state probability distribution which is independent of the initial state of the process. The existence of a stationary policy allows us to conclude that an optimal solution can be expressed through a decision variable $x_{\mathbf{sa}}$ that represents the probability for the system to be in state $\mathbf{s}$ and contemporaneously to take the decision $\mathbf{a}$.

The Linear Programming (LP) formulation associated with our SMDP for the minimization of the cost paid in a long-run execution is given by:

$$
\begin{aligned}
&\text{Minimize} \\
&\textstyle\sum_{(\mathbf{s},\mathbf{a})\in\mathcal{S}} \tilde{r}(\mathbf{s},\mathbf{a})\cdot x_{\mathbf{s},\mathbf{a}} \\
&\text{constrained to} \\
&x_{\mathbf{sa}} \geq 0 \qquad\qquad (\mathbf{s},\mathbf{a})\in\mathcal{S} \\
&\textstyle\sum_{(\mathbf{s},\mathbf{a})\in\mathcal{S}} x_{\mathbf{sa}} = 1 \\
&\textstyle\sum_{\mathbf{a}\in\mathcal{B}} x_{\mathbf{ja}} = \sum_{(\mathbf{s},\mathbf{a})\in\mathcal{S}} \tilde{p}^{\mathbf{a}}_{\mathbf{sj}} x_{\mathbf{sa}} \qquad \mathbf{j}\in\Lambda.
\end{aligned} \tag{8}
$$

Once known the parameters that characterize the traffic behavior in the considered cell, the linear programming problem given by Equation 8 can be solved by means of the well known methods of the operations research [9], thus obtaining the optimal policy for CAC. This problem can also be solved at run-time, and it can be used to calculate the steady state probabilities that are necessary to obtain the values of the main QoS parameters.

III. Optimization of the Multi-Class Model

We now want to obtain general properties of the optimal CAC policy. For this purpose we keep on analyzing the optimization problem.

The decision process formulated in Section II allows the search for the optimal policy in a wide general class that also includes the so called *fractional* or *randomized* policies. Stationary randomized policies are characterized by state-related non-deterministic decisions, i.e. in each state the decision is made with a certain fixed probability that defines the considered policy.

Nevertheless fractional policies can be excluded from our study since, by analyzing the shape of the restraint matrix of the linear programming formulation given by Equation 8, we can prove the existence of an optimal deterministic solution.

To investigate the shape of the optimal policy we recur to Derman's theorem [6] according to which every policy that is optimal under a discounted cost criterion with a discount factor α close to 1 is also optimal under the average cost criterion. Thence we restrict our study to the optimization of a discounted cost function over an infinite horizon $W^\alpha(\mathbf{s})$, that has the following formulation obtained by means of dynamic programming methods, made possible by the certainty of the existence of an optimal deterministic solution.

$$
\begin{aligned}
W^\alpha(\mathbf{s}) = \frac{1}{\eta+\Gamma}\Bigg\{ &\sum_{i=1}^{N} s_i\mu_i W^\alpha(\mathbf{s}-\mathbf{e_i}) + \\
&+\sum_{i=1}^{N}(C-s_i)\mu_i W^\alpha(\mathbf{s}) + \sum_{i=1}^{N}(\lambda_i^{new}+\lambda_i^{hoff})W^\alpha(\mathbf{s}) + \\
&+\sum_{i=1}^{N}\lambda_i^{new}\min_{\mathbf{a}\in\mathcal{B}}\{a_i^{new}\left[W^\alpha(\mathbf{s}+\mathbf{e_i})-W^\alpha(\mathbf{s})\right]+ \\
&+(1-a_i^{new})\hat{L}_i\} + \\
&+\sum_{i=1}^{N}\lambda_i^{hoff}\min_{\mathbf{a}\in\mathcal{B}}\left\{a_i^{hoff}\left[W^\alpha(\mathbf{s}+\mathbf{e_i})-W^\alpha(\mathbf{s})\right]+\right. \\
&\left.+(1-a_i^{hoff})\hat{H}_i\right\} + \\
&+\sum_{i=1}^{N}\sum_{j=1}^{N}\left[\pi_{ij}s_iW^\alpha(\mathbf{s}-\mathbf{e_i}+\mathbf{e_j})+(C-s_i)\pi_{ij}W^\alpha(\mathbf{s})\right]\Bigg\}
\end{aligned} \tag{9}
$$

where $\hat{L}_i \triangleq L_i/\alpha$ and $\hat{H}_i \triangleq H_i/\alpha$.

Then the optimal policy chooses the best action to take in each state with the following rule:

- High priority customers of class i are accepted only if $\Delta_i^\alpha(\mathbf{s}) \triangleq W^\alpha(\mathbf{s}+\mathbf{e_i}) - W^\alpha(\mathbf{s}) \leq \hat{H}_i$.
- Low priority customers of class i are accepted only if $\Delta_i^\alpha(\mathbf{s}) \triangleq W^\alpha(\mathbf{s}+\mathbf{e_i}) - W^\alpha(\mathbf{s}) \leq \hat{L}_i$.

An N-dimensional property of monotony and convexity in the number of ongoing calls in each class can be proven by means of dynamic programming methods.

Since $\Delta_i^\alpha(\mathbf{s})$ is non-negative and non-decreasing in the number of ongoing calls in class i, thence we can find integer values $t_{L_i}(\mathbf{s})$ and $t_{H_i}(\mathbf{s})$ such that

$$
t_{L_i}(\mathbf{s}) = \left[\arg\min\left\{\Delta_i^\alpha(\mathbf{s}) > \hat{L}_i\right\}\right]_i,
$$

and

$$
t_{H_i}(\mathbf{s}) = \left[\arg\min\left\{\Delta_i^\alpha(\mathbf{s}) > \hat{H}_i\right\}\right]_i.
$$

Therefore, the optimal policy regarding the decision to accept or refuse to serve requests belonging to any one of the classes, cannot be easily reduced to a fixed thresholds policy like MTP or HCPP, the threshold value for each class strictly depends on the level of occupancy of the other classes. Each threshold has to be evaluated for the possible levels of occupancy, for example by means of the LP method seen in Section II concerning the optimization problem of Equation 8.

Since there is an optimal stationary policy, if the traffic parameters do not change significantly at runtime, the optimal policy can be evaluated once for all, by means of the linear programming formulation of this optimization problem. The computational complexity of this optimal solution is proportional to the size of the restraint matrix of the problem, but if an average traffic behavior can be identified, the solution can be computed once for all the time the esteem of the traffic parameters is sufficiently accurate.

From Equation 9 we can see that if μ_i does not depend on the class, that is $\mu \triangleq \mu_i$, $\forall i \in \{1, 2, \ldots, N\}$, and if the bandwidth requirements are the same in every class, that is $b \triangleq b_i$, $\forall i \in \{1, 2, \ldots, N\}$, then the optimal discounted cost function over an infinite horizon $W^\alpha(\mathbf{s})$ does not depend on the particular state $\mathbf{s}$, but on its total occupancy level only and the policy MTP can be proven to be optimal in this particular case.

Thence MTP is not optimal even without class transitions, unless with some assumptions, i.e. the different classes, even though with different priorities, have the same bandwidth requirements and the same cell residence and call holding times, i.e., MTP is optimal only for non multimedia traffic. HCPP is optimal in an even more restrictive case, that is only if the classes of requests have also the same priority, i.e. with only one class of traffic, thus obtaining the well known CPP [2].

IV. Conclusions

In this paper we conducted an analysis and optimization of QoS of wireless networks by means of innovative models created with the methodology of SMDPs. SMDPs, in a very general shape, are proven to be an excellent method of evaluation of QoS parameters and CAC policies, creating a parametric model that makes us able to compare the behavior of several admission policies, while optimizing an objective function in the shape of an average cost over an infinite horizon. SMDPs can be used to optimize blocking probabilities of new and handoff calls of several classes of multimedia requests. SMDPs reveal that the optimal policy establishes state-related threshold values for handoff and new calls in the different classes. It was proven that only if the classes have the same bandwidth requirements and the same service time, i.e. for prioritized non-multimedia traffic, the optimal policy has the shape of MTP.

References

[1] A. S. Acampora, M Naghshineh, "An architecture and Methodology for Mobile-Executed Handoff in Cellular ATM Networks", *IEEE Journal on Selected Areas in Communications*, Vol.12, No. 8, October 1994, pp.1365-1374.

[2] N. Bartolini, "Handoff and Optimal Channel Assignment in Wireless Networks" Mobile Networks and Applications, ACM/Kluwer Academic Publishers, Vol.6, No.6, November 2001.

[3] A. T. Bharucha-Reid, *Elements of the theory of Markov processes and their applications.* McGraw-Hill, 1960.

[4] S. K. Biswas, B. Sengupta, "Call Admissibility for Multirate Traffic in Wireless ATM Networks", in *Proceedings of IEEE INFOCOM '97*, pp. 650-658.

[5] J. Chen, M. Schwartz, "Two-tier resource allocation for a multimedia micro-cellular mobile system", IBM Research, Hawthorne, New York, submitted for publication.

[6] C. Derman, *Finite state Markovian decision processes.* Academic Press, 1970.

[7] B. Epstein, M. Schwartz, "Reservation Strategies for Multimedia Traffic in a Wireless Environment" in *IEEE 45th Vehicular Technology Conference*, Chicago IL, July 1995.

[8] K. Hastings, *Introduction to the mathematics of operations research*, Dekker, 1989.

[9] D. P. Heyman, M. J. Sobel, *Stochastic models in operations research.* Vol. 1 and Vol.2, McGraw-Hill, 1984.

[10] J. Keilson, *Markov chain models. Rarity and exponentiality.* Springer-Verlag, New York, 1979.

[11] S. Kim, T. Kwon, Y. Choi, "Call Admission Control for Prioritized Adaptive Multimedia Services in Wireless/Mobile Networks", in IEEE *VTC 2000 Spring*, Tokyo, 2000.

[12] T. Kwong, Y. Choi, M. Nagshineh, *Optimal Distributed Call Admission Control for Multimedia Services in Mobile Cellular Network* in Proceedings of 5th Inl. Workshop on Mobile Multimedia Communivations MoMuC'98, October 12-14, 1998 Berlin.

[13] B. Li, C. Lin, S. T. Chanson, "Analysis of a Hybrid Cutoff Priority Scheme for Multiple Classes of Traffic in Multimedia Wireless Networks", *Wireless Networks*, ACM Baltzer Science Publishers, Vol. 4, No. 4, 1998, pp.279-290.

[14] S. Ross, *Applied probability models with optimization applications*, Holden-Day, 1970.

[15] H. C. Tijms, *Stochastic modeling and analysis. A computational approach*, John Wiley & Sons, 1986.