Optimal Mobility-aware Admission Control in Content Delivery Networks

Novella Bartolini[*]
Università di Roma “La Sapienza”
Via Salaria 113, 00198 Roma, Italy
novella@dsi.uniroma1.it

Emiliano Casalicchio, Salvatore Tucci[†]
Università di Roma ”Tor Vergata”
Via del Politecnico, 00133 Roma, Italy
{casalicchio,tucci}@ing.uniroma2.it

Abstract

This paper addresses the problem of mobility management in Content Delivery Networks (CDN). We introduce a CDN architecture where admission control is performed at mobility aware access routers. We formulate a Markov Modulated Poisson Decision Process for access control that captures the bursty nature of data and packetized traffic together with the heterogeneity of multimedia services. The optimization of performance parameters, like the blocking probabilities and the overall utilization, is conducted and the structural properties of the optimal solutions are also studied. Heuristics are proposed to encompass the computational difficulties of the optimal solution when several classes of multimedia traffic are considered.

1. Introduction

Content Delivery Networks (CDN)[9] are based on mechanisms of content replica server placement, measurement and selection and request redirection. Mobile terminals connecting to the CDN via a wireless interface are likely to change their point of access several times during a session. Handover procedures must be performed at the access routers, with user and session profile management, to ensure that the mobile users will continuously benefit of the increased performance of a CDN architecture, selecting the replica server at each CDN access point. Though handovers at internet access points occur at lower frequency than handovers at the wireless interface, they can significantly degrade the performance of long session services like audio and video streaming, while they are sufficiently tolerated in web services where the probability that handovers occur during user think times is considerably high.

We consider an all-IP scenario where mobile access is enabled by Mobility Aware Routers (MAR) [6]. The control area of a MAR extends over the coverage area of one or more wireless access points (see figure 1). MARs provide replica server selection mechanisms, profile management, context transfer [5] and admission control. When a mobile client crosses the overlapping region between two adjacent MAR service areas, the ongoing service management must be transferred from the first router to the one associated with the next coverage area, performing a content aware handover procedure at the internet access level.

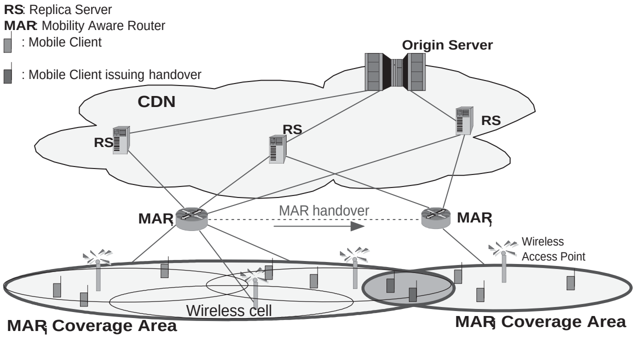


Figure 1. Content Delivery Network architecture enhanced with mobile access capability.

Focus of this paper is the performance evaluation of access control policies for MARs. The problem of admission control is well known in the field of wireless and wired networks [2, 7]. To the best of our knowledge there is no preceding work on admission control for mobile access to CDN. We propose a Markov Modulated Poisson Process

[*] The work of N. Bartolini has been funded by the WEB-MINDS project supported by the Italian MIUR under the FIRB program.

[†] The work of E.Casalicchio and S. Tucci has been funded by the PERF project supported by the Italian MIUR under the FIRB program.

(MMPP) enhanced with decision support to formulate admission policies that optimize performance parameters such as the handover failures and new requests blocking probabilities, or the overall bandwidth utilization. The proposed model captures the bursty nature of data and packetized traffic, typical of multimedia network applications accessible through CDNs. It also considers a proper differentiation of service classes and prioritization of handover over new requests. Through the proposed model, optimal policies can be computed, once basic statistic parameters defining the traffic of requests are known, by means of commonly used methods of operations research like linear programming or value iteration[4, 8]. A structural analysis of optimal policies is given. Some heuristics are also proposed to avoid excessive computational burden when several classes of multimedia services are considered. Performance comparisons between the heuristics and the optimal policies are also given.

In Section 2 we introduce a Markov Modulated Poisson Decision model and related optimization problems and we analyze the structural properties of the optimal admission control policy. In Section 3 some heuristics are proposed. In Section 4 performance comparisons among optimal policies and heuristics are given, while Section 5 concludes the paper.

2. A Markov modulated decision process for session based access control

The access to CDN services occurs in the form of a *session* consisting of many individual phases characterized by different resource utilization and traffic parameters [3]. We consider a decision model with different classes representing the possible phases composing a session. We also consider a class differentiation to account for different types of services. Class transitions are also introduced to model phases in the behavior of the considered traffic due to its burstiness and bandwidth adaptiveness. It is without loss of generality that in our model we consider classes both as different type of services and as different phases of the same type of service. Class i denotes either a particular type of service, or a single session phase of a Markovian modulation of the whole service duration in different exponential phases. A class transition is also used to model the "on/off" period of bandwidth consumption of packetized traffic.

Each MAR can assign up to C resource units and manages N classes with further differentiation between new and handover requests. Each request of class i has a class dependent bandwidth requirement of b_i bandwidth units. Each class is characterized by exponential arrivals λ_i^{n} and λ_i^{h}. An ongoing session is Markov modulated, among different phases of resource consumption and idle times, creating a MMPP of services. We indicate with μ_i^{n} and μ_i^{h} the average termination rates of class/phase i requests. The state of the Markov Modulated Poisson Decision Process we are introducing is represented through a $2N$-dimensional vector $\mathbf{x} \triangleq (x_1^{\mathsf{n}}, x_2^{\mathsf{n}}, \ldots, x_n^{\mathsf{n}}, \ldots, x_1^{\mathsf{h}}, \ldots, x_n^{\mathsf{h}})$ in which the random variables x_i^{n} and x_i^{h} represent the number of new and handover requests respectively, being served in class i. The state space is defined as $\Lambda = \{\mathbf{x} : \sum_{i=1}^{N} b_i \cdot (x_i^{\mathsf{n}} + x_i^{\mathsf{h}}) \leq C; x_i^{\mathsf{n}}, x_i^{\mathsf{h}} \geq 0\}$. When the system is in state $\mathbf{x}$, an accept/reject decision must be taken for each type of request. Thus, the action space can be denoted by $\mathcal{B} = \{(a_1^{\mathsf{n}}, a_2^{\mathsf{n}}, \ldots, a_N^{\mathsf{n}}, a_1^{\mathsf{h}}, \ldots, a_N^{\mathsf{h}}) : a_i^{\mathsf{n}}, a_j^{\mathsf{h}} \in \{0, 1\}, i, j = 1, \ldots, N\}$ where the indicators a_i^{n} and a_j^{h} denote the acceptance (1) or the denial of service (0) of class-i new requests or handover respectively.

The set of rates which characterizes the process is bounded above by the maximum outgoing rate Γ, hence, the process is uniformizable and discrete analysis methods can be applied to its embedded Markov chain. Two formulations of the objective function are used in this paper to optimize the blocking probabilities of handover and new requests, and the overall utilization, while keeping the handover failure probability constrained. We indicate with $\mathcal{S}$ the finite set of all feasible couples of vectors of the kind (*state, decision*). Let $\tilde{p}_{\mathbf{sj}}^{\mathbf{a}}$ denote the uniformized transition probability from state $\mathbf{s}$ to state $\mathbf{j}$ when the decision $\mathbf{a}$ is taken. We define $\pi_{\mathbf{s},\mathbf{a}}$ as the probability that the system is in state $\mathbf{s}$ and contemporaneously takes the decision $\mathbf{a}$.

To optimize the blocking probabilities of new and handover requests, we assign penalty values H_i and L_i for the denial of access to handover and new class-i requests respectively. In this first case, the optimization problem has the following linear programming formulation (1), where $\pi_{\mathbf{sa}} \geq 0:$

$$\begin{array}{ll}\text{Min} \sum_{(\mathbf{s},\mathbf{a}) \in \mathcal{S}} \frac{\pi_{\mathbf{s},\mathbf{a}}}{\Gamma} \sum_{i=1}^{N} [\lambda_i^{\mathsf{n}} L_i (1 - a_i^{\mathsf{n}}) + \lambda_i^{\mathsf{h}} H_i (1 - a_i^{\mathsf{h}})] & \\ \sum_{(\mathbf{s},\mathbf{a}) \in \mathcal{S}} \pi_{\mathbf{sa}} = 1 & \\ \sum_{\mathbf{a} \in \mathcal{B}_{\mathbf{j}}} \pi_{\mathbf{ja}} = \sum_{(\mathbf{s},\mathbf{a}) \in \mathcal{S}} \tilde{p}_{\mathbf{sj}}^{\mathbf{a}} \pi_{\mathbf{sa}}, & \mathbf{j} \in \Lambda. \end{array} \tag{1}$$

In a second case we optimize the overall utilization, with a constraint on the total blocking probability of handover requests to be less than $P_{\mathsf{h}}^{\mathsf{MAX}}$. The linear programming problem is the following (2), where $\pi_{\mathbf{sa}} \geq 0:$

$$\begin{array}{ll}\text{Max} \sum_{(\mathbf{sa}) \in \mathcal{S}} \pi_{\mathbf{s},\mathbf{a}} \cdot \sum_{i=1}^{N} s_i \cdot b_i & \\ \sum_{(\mathbf{s},\mathbf{a}) \in \mathcal{S}} \pi_{\mathbf{sa}} = 1 & \\ \sum_{\mathbf{a} \in \mathcal{B}_{\mathbf{j}}} \pi_{\mathbf{ja}} = \sum_{(\mathbf{s},\mathbf{a}) \in \mathcal{S}} \tilde{p}_{\mathbf{sj}}^{\mathbf{a}} \pi_{\mathbf{sa}} & \\ \sum_{(\mathbf{s},\mathbf{a}) \in \mathcal{S}} (1 - \prod_{i=1}^{N} a_i^{\mathsf{h}}) \pi_{\mathbf{sa}} \leq P_{\mathsf{h}}^{\mathsf{MAX}} & \mathbf{j} \in \Lambda. \end{array} \tag{2}$$

The solution of (1) and (2) leads to two optimal policies to which we will refer with the names Opt and CMU respectively. In [1] we show that in the general case described in this paper the optimal solution has a complex shape that establishes state-related decisions for the admittance of handover and new requests in the different classes.

The decision processes here formulated allow the search for the optimal policy in a wide general class that includes also randomized policies [4]. The existence of a deterministic optimal solution can be proven through the analysis of the set of constraints of problem (1). In analogous way it can also be proven that problem (2) has an optimal solution that is deterministic in all states but one. An N-dimensional property of monotony and convexity of the optimal cost function of problem (1) in the number of ongoing services of each class, can be proven by means of dynamic programming methods (by induction on the number of steps k) [1]. From this property we can conclude that the optimal solution of problem (1), though threshold based, cannot be reduced to a fixed threshold or to a multi-threshold policy. The optimal state-related threshold value for each class depends on the occupancy level of the other classes (see [1] for details).

More complex is the structural analysis of the optimal solution of problem (2) showing a randomized decision in at least one state.

3. Access Control Heuristics

The optimization method introduced in Section 2 can be used off-line to calculate the optimal policy once some workload parameters are known. The proposed optimization method becomes prohibitive when there is fast variability of traffic parameters because the optimal policy cannot be computed at real time in case of high number of service classes, due to the high dimensionality of the problem.

Under certain assumptions, that is when the different classes have the same bandwidth requirements and the same exit and completion rates, the optimal policy has the shape of a multi-threshold priority policy (MTP) [1].

Under MTP two threshold values are defined for each class, $T_i^{\texttt{new}}$ and $T_i^{\texttt{hoff}}$, where $T_i^{\texttt{new}} \leq T_i^{\texttt{hoff}} \leq C$. A class-$i$ request 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 request does not cause the occupancy level to exceed the threshold $T_i^{\texttt{new}}$ or $T_i^{\texttt{hoff}}$ depending on the origination of the request. Under MTP, the handoff requests of the highest priority class are always accepted. The definition of a threshold value even for handoff requests allows for the case in which a low priority handoff request can be denied service in the hope of making place for higher priority requests, even if newly generated. Experimental results have shown that the best choices of the thresholds are in relation to the parameters μ_i, b_i and to the priorities of each class. We therefore define the following heuristic to tune the threshold values: $T_i^{\texttt{new}} = A \cdot \frac{L_i \mu_i}{b_i} + B$ and $T_i^{\texttt{hoff}} = A \cdot \frac{H_i \mu_i}{b_i} + B$, where A and B are constant values to be calculated as follows.

The values of the highest threshold must be set to C because otherwise the policy would cause some resources being left idle. The value of the lowest threshold impacts on the amount of reserved resources for high priority classes and for handoff, therefore must be set low as much as we think resources should be reserved. It must also be set positive to guarantee that the lowest priority new requests are not always discarded. We fix the values of the highest and lowest thresholds, $\texttt{max}_{i=1..N} T_i^{\texttt{hoff}} = C$ and $\texttt{min}_{i=1..N} T_i^{\texttt{new}} = \frac{C}{k}$.

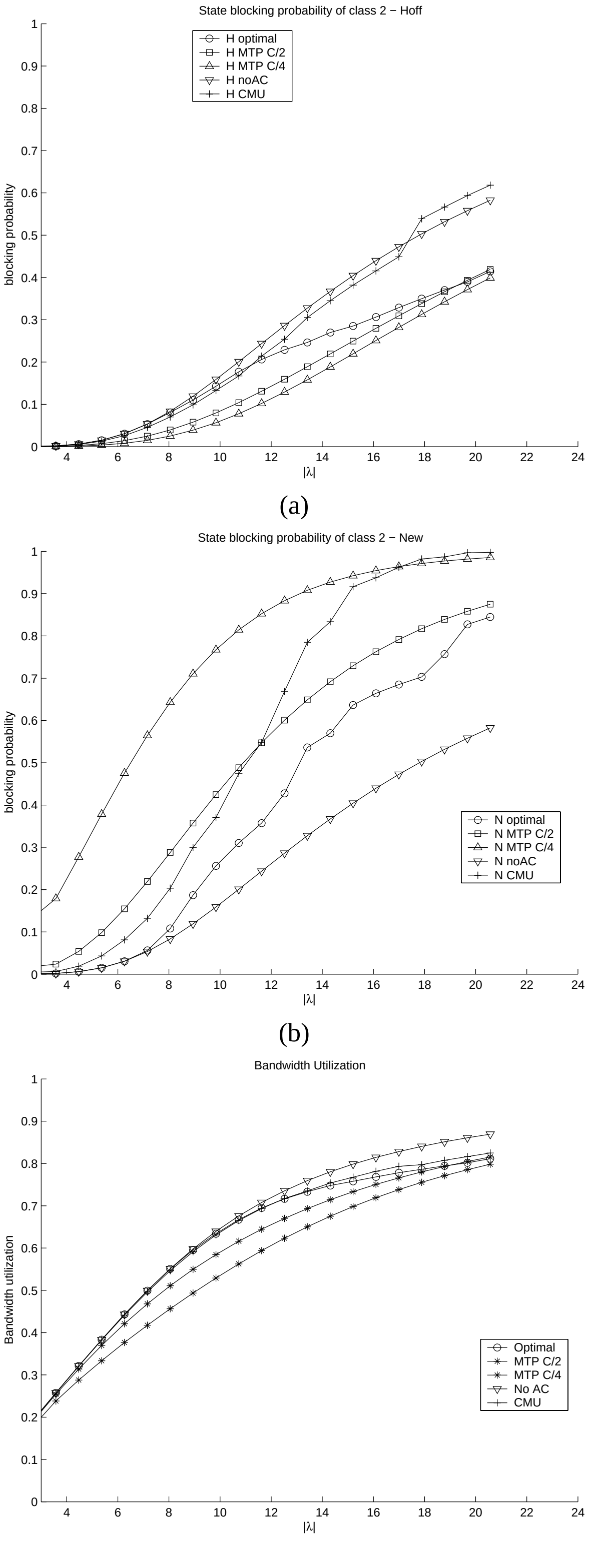
Experimental results show that, in most cases, $k = N$ and $k = 2 * N$ are acceptable values. We refer to MTP-C/2 for the heuristic policy in which $k = N$ and to MTP-C/4 for the heuristic where $k = 2 * N$. The heuristics values for the constants A and B can be computed setting the highest and lowest thresholds among all the classes. Once A and B are computed, the two threshold values for each class can be obtained for MTP sizing.

4. Numerical results

In this section we provide performance comparisons among the proposed heuristics (MTP-C/2, MTP-C/4) and the optimal policies (OPT, CMU) with a case without admission control (noAC). Three performance metrics are considered: bandwidth utilization, handoff failure and new request blocking probability for each class. We consider only two service classes for simplicity, with a resource availability of $C = 12$ units and a request mix with $\mathbf{b} = (1, 2)$ and $\mu = (2, 5)$.

Figures 2(a) and 2(b) show the blocking probability of new and handoff requests for class 2 requests (results regarding other classes are omitted). These figures point out how the proposed heuristics approximate the behavior of the optimal policies. MTP-C/4 shows a lower handoff failure probability at the expense of an increase in the new request blocking probability when compared with MTP-C/2. MTP-C/4 in fact reserves more resources to handoff requests and therefore causes a resource waste. MTP-C/2 represents a tradeoff between blocking probability of new and handoff requests. Though noAC does never show very high blocking probability, it doesn't provide any kind of prioritization of classes and therefore is only used as term of comparison. CMU shows a bad performance in terms of new request blocking probability and handoff failures. It indeed tends to lead the system to saturation and therefore to a high probability of being in a blocking state for all classes.

Comparing the admission control policies in terms of resource utilization (see figure 2(c)), the best performance is



(a)

(b)

(c)

Figure 2. (a) Blocking probability of handoff requests and (b) new requests. (c) Overall bandwidth utilization.

obviously achieved in absence of admission control with noAC. CMU approximates this behavior until the system reaches very high load conditions. MTP approaches the behavior of the optimal policies in terms of resource utilization, when the system load increases.

5. Conclusion

In this paper we addressed the problem of mobile access control to CDN networks. We introduced a reference architecture to show that access control must be performed by CDN access routers in a mobility aware mode, guaranteeing the agreed QoS and context transfer while the users move. The necessity of handoff management at access points is pointed out. A Markov Modulated Poisson Decision Process which captures the bursty nature of the packetized traffic that reaches the access routers is introduced and optimized according to different objective functions, like the handoff failure probability, the new request blocking probability and the overall utilization. To investigate the behavior of the optimal policy, a structural analysis of the average cost function is conducted. Some heuristics are also proposed to overcome the computational difficulties when many types of services are considered.

References

[1] N. Bartolini. Admission control in multimedia heterogeneous traffic environment. *RR I-01/02, University of Rome*.

[2] N. Bartolini and I. Chlamtac. Improving call admission control procedures by using handoff rate information. *Journal of Wireless Communications and Mobile Computing*, 1(3), 2001.

[3] H. Chen and P. Mohapatra. Session-based overload control in qos-aware web servers. *Proc. of INFOCOM 2002, New York, NY*.

[4] D.P.Heyman and M.J.Sobel. *Stochastic Models in Operations Research*. McGraw-Hill, 1984.

[5] O. H. L. et al. *RFC3374: Context Transfer Problem Statement*. http://www.faqs.org/ftp/rfc/pdf/rfc3374.txt.pdf.

[6] Z. Jiang, K. Leung, B. Kim, and P. Henry. Seamless mobility management based on proxy servers. *Proc. of WCNC 2002, Orlando, FL, April*, 2002.

[7] T. Kwon, Y. Choi, C. Bisdikian, and M. Nagshineh. Qos provisioning in wireless/mobile multimedia networks using an adaptive framework. *ACM/Baltzer Wireless Networks 9*, 2003.

[8] H. C. Tijms. *Stochastic modelels. An algorithmic approach*. John Wiley & Sons, 1994.

[9] D. C. Verma. *Content Distribution Networks*. John Wiley & Sons, Inc., 2002.