Admission Control Scheme for Proxy Mobile IPv6 Networks

Nika Naghavi, Vasilis Friderikos, Toktam Mahmoodi, Hamid Aghvami
Centre for Telecommunications Research, King's College London
{nika.naghvai, vasilis.friderikos, toktam.mahmoodi, hamid.aghvami}@kcl.ac.uk

***Abstract*—This paper's central aim is to address the issue of resource management at the Local Mobility Anchors (LMA) in the Proxy Mobile IPv6 (PMIPv6) networks. A class-based admission control is proposed to improve the bottleneck effect caused by triangular routing in PMIPv6, where resource units are rationed amongst different classes of traffic according to their QoS requirements. The PMIPv6 network is modeled as an M/M/m/m tandem queuing network with two types (classes) of arrival process and an analytical model is presented. Performance of our proposed admission control scheme is evaluated through simulation and results are compared to the case where no distinction in terms of resource unit allocation between classes of traffic was considered.**

I. Introduction

Next generation of mobile networks are expected to support a variety of mobile devices such as smart phones, Personal Digital Assistants (PDAs), and laptop computers over full IP based mobile networks. As a result the need to have high-speed Internet access and an intelligent mobility management support along with QoS mechanisms is ever increasing. In order to provide an efficient mobility support various macro and micro mobility management protocols have been proposed. One of the most well known standards for mobility support in IPv6 based communication networks is Mobile IPv6 [1]. Such a macro mobility solution has some key shortcomings such as handover latency, packet loss and signalling overhead [2]. Micro-mobility protocols such as Cellular IP [3], fast handovers for mobile IPv6 [4], and hierarchical mobile IPv6 [5] and others [6] have been proposed to overcome some of the shortcomings of the plain Mobile IPv6. All of these protocols are based on the idea of implementing a mobility agent that hides the local mobility of the Mobile Node (MN) so that Binding Updates do not have to be forwarded to the Home Agent (HA) every time the MN change its point of attachment.

Proxy Mobile IPv6 (PMIPv6) is a network based mobility management protocol that enables IP mobility without involving the MN [7]. Its advantages over the existing protocols range from minimising the handover latencies, to reducing the overheads such as signalling over the wireless link and non-complex deployment. According to the PMIPv6, the MN does not participate in any of the mobility-related signalling and the serving network is in charge of the MN's mobility management. The two main logical entities that provide a network-based mobility management support in PMIPv6 are Local Mobility Anchor (LMA) and Mobility Access Gateway (MAG). The LMA is similar to the HA in MIPv6, in a sense that it has a binding cache memory for all the registered MNs and manages the MNs binding updates. The MAG is usually run on the Access Router (AR) of the MN. The MAG's main duty is to perform mobility-related signalling as well as detecting and performing handovers on behalf of all the MNs that are attached to its access link [2].

In PMIPv6, all traffic originated from or sent to an MN, regardless of their QoS requirements, has to go through the bidirectional tunnel between the MAG and the LMA [2]. It is possible to bypass the LMA for localized routing, if a MAG has traffic that is in one of its access links and it is destined to another one of the same MAG's access links i.e. both the MN and Correspondent Node (CN) are located within the same MAG. In this paper we focus on the scenarios where the MN and the CN are not located within the same MAG and the traffic has to go through the bidirectional tunnel between the MAG and the LMA to be sent via the LMA. In such scenario, managing the resources at the LMA becomes highly important. Thus, we mainly focus on this issue and propose an admission control scheme that uses a class-based approach and treats traffic according to their Quality of Service (QoS) requirements.

With the advances in communication technology, mobile devices have become one of the most important areas in Internet growth. The mobile users access a wide range of multimedia rich traffic, which demands for new protocols and management schemes in order to be able to provide differentiated service according to their requested QoS. The primarily assumption in the all IP networks is that n classes of traffic are present, but aggregating flows into a few number of classes according to their QoS requirements is much simpler traffic management task than providing QoS for each of the n individual classes of traffic. In this respect, we assume two classes of real-time and non real-time traffic, and design our admission control based on these two categories. A similar distinction between different classes of traffic was made in [8] and [9].

The admission control proposed in this paper gives higher priority to the real-time traffic by allocating more resources. We model the PMIPv6 as a tandem queuing network where each node operates as an independent M/M/m/m queue. Capacity of each node is assumed to be divided into m resource units, and the admission control scheme works as follows: real-time traffic can benefit from simultaneous use of several resource units and non real-time traffic can occupy single resource unit. Simulation results demonstrate considerable benefits in using our proposed scheme in terms of reduction in the total blocking probability as well as blocking probability per class of traffic. The main novelty of this paper is division of the available resources to multiple resource units and design of the admission policies associated with that. Furthermore, this work is the first to model the PMIPv6 as an M/M/m/m tandem queuing network.

The rest of this paper is organized as follows. In Section II the problem description, along with network model, and the state transition diagram of our queuing model is detailed. Section III, our proposed admission control scheme is briefed. Simulation results and analysis are provided in Section IV, and finally the paper is concluded in Section V.

II. System Model

This section sets out a formal definition of the analytical model used in the performance evaluation section. In this paper, PMIPv6 network is modeled as a tandem queuing network. Capacity of each node is assumed to be divided into m resource units which is equivalent to m servers and each node operates as an independent M/M/m/m queue or Erlang Loss system [10] (Chapter 3, Section 3.4.3), with two types of arrival process. In today's internet most of the session are background traffic or in other words are non real-time [11][12], in this respect in our work we consider categorizing the traffic into two classes: I) Non real-time traffic and II) Real-time traffic. Sessions from both classes of traffic have to traverse through the LMA node and if both classes of traffic are treated the same, it may result in QoS disruption. The proposed admission control scheme in this paper, takes a class-based approach when it comes to resource allocation and operates as follows: assuming that the bandwidth required by the real-time traffic is much larger than the bandwidth required by the non real-time traffic, each real-time session will benefit from the simultaneous use of several resource units whilst each non real-time session can occupy one of the m resource units at a time.

The amount of bandwidth required by the real-time and non real-time traffic is a and b Kbps respectively. Let's assume that the capacity of all nodes in the network is equal to each other and is denoted by C_T. Assuming that $a \gg b$, then it can be stated that the total bandwidth in each node is equal to m servers or resource units, m being equal to η as shown in equation 2. In the proposed model in this paper, admitting a non real-time traffic requires one of the η resource units and admitting a real-time traffic requires the simultaneous use of several resource units. Number of required resource units by the non real-time and real-time traffic are shown by m_b and m_a respectively, where $m_b = 1$ and $m_a = a/b$. Considering one class of traffic at a time and assuming that there is no arrival from the other class, it can be stated the number of real-time traffic that can be admitted at each node is equal to γ and the number of non real-time traffic that can be admitted at each node is equal to η such that:

$$\gamma = C_T/a \tag{1}$$

$$\eta = C_T/b \tag{2}$$

One of the primarily assumptions in this paper is that there is an identical service time offered in each node of the network with higher service time for the real-time traffic, this assumption restores the independence of inter-arrival times and packet lengths; hence the Kleinrock independence approximation is valid and nodes in the network can be modeled as independent M/M/m/m queuing systems [10] (Chapter 3).

Another valid assumption is the offline route discovery, i.e. admitting the sessions at the gateway is performed with the prior knowledge of the bottlenecks in the network. Dijkstra algorithm is used to find the set of shortest paths in the network. It is essential to mention that the proposed admission control scheme is performed on the downlink traffic, and we consider routing all sessions through the LMA node whilst using our proposed admission control scheme, hence the major bottleneck in the network is the LMA itself.

Upon arrival of a new session at the gateway, the class of the session is distinguished and the decision as to admit or block the session at the LMA is made accordingly. The process involves checking the number of available resource units at the LMA, and the number of required resource units depending on the class of the traffic. The same process is then repeated for every node in tandem till the session reaches its destination. If all the η resource units are busy at any of the nodes on the path, the session gets blocked.

A. *Network Model*

A PMIPv6 network is considered as a connectivity graph $G(V;E)$, where V is a set of routers and E is the set of links in the network. K is the set of LMA routers in the PIMPv6 network, M is the set of MAGs and g is the gateway router such that $M \in V - \{K, g\}$. Set of all of the shortest paths from the gateway to each MAG that contains the LMA is defined as P. Let $P_{g,k}^{LMA}$ and $P_{k,m}^{MAG}$ be the set of all paths from gateway g to the LMA $k \in K$ and from LMA $k \in K$ to MAGs $m \in M$ respectively. As mentioned earlier, in this research paper we are only interested in routing through the paths that contain an LMA router. The session arrivals for both classes of traffic $(n = 2)$ is assumed to follow the Poisson distribution with an average arrival rate λ_n and exponentially distributed service time with mean $1/\mu_n$. Service time offered in all of the nodes in the network is identical and the service time offered to the real-time traffic is higher than the service time of non real-time traffic; i.e. if $\mu_1 = \mu$ then $\mu_2 = 5\mu$. Our topology is shown in Figure 1.

B. *State Transition Diagram*

In this section the state transition diagram for one node is analyzed. State (i, j) represents the presence of i non real-time and j real-time requests in the system. Let λ_1 and λ_2 be the average call arrival, and μ and 5μ the mean service time of non real-time and real-time traffic respectively. Consider an LMA node within a PMIPv6

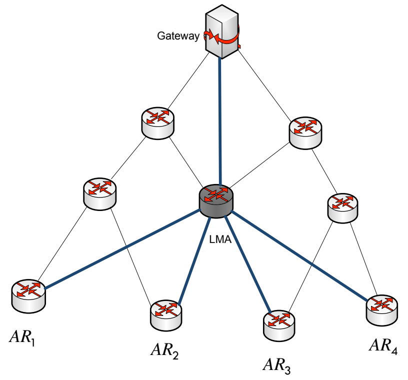
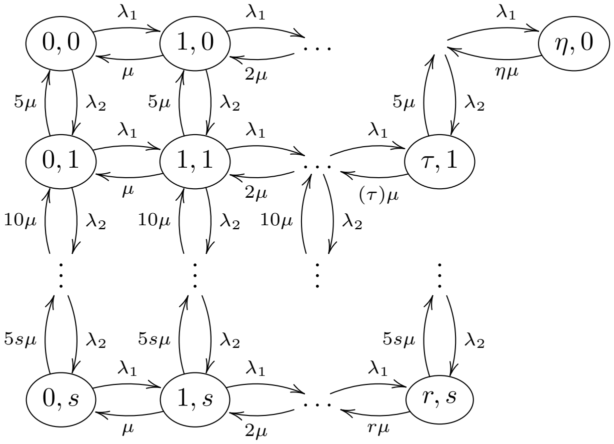


Fig. 1. Topology

network with capacity C_T, this can be analyzed by a 2-dimensional Markov chain [13] (Chapter 3), as follows:



Let $P(i,j)$ be the steady-state probability of system being in state (i,j). The probability of system being in an equilibrium state can be found by solving the general balance equation (flow in = flow out) for an internal state:

$$\begin{aligned}\lambda_1 p(i-1,j)+\lambda_2 p(i,j-1)+(i+1)\mu p(i+1,j)\\ +(j+1)5\mu p(i,j+1)=j5\mu p(i,j)+i\mu p(i,j)\\ +\lambda_2 p(i,j)+\lambda_1 p(i,j)\end{aligned} \tag{3}$$

in conjunction with the method suggested in [13] (Chapter 3), this is shown below :

$$P(i,j)=\frac{\frac{1}{i!}(\frac{\lambda_1}{\mu})^i\frac{1}{j!}(\frac{\lambda_2}{5\mu})^j}{\sum_{j=0}^{\eta}\frac{1}{j!}(\frac{\lambda_2}{5\mu})^j\ \sum_{i=0}^{\eta-jm_a}\ \frac{1}{i!}(\frac{\lambda_1}{\mu})^i} \tag{4}$$

In the state diagram, horizontal arrows to the right and left correspond to the arrival and departure of non real-time traffic into the system. It can be concluded that a non real-time traffic is blocked if all the η servers are busy and this occurs when the system is at the rightmost of any row. Using Equation (4) and summing the probabilities of the $s+1$ states, the probability of a non real-time traffic being blocked or denied access to the LMA can be computed as follows:

$$P_{b_{non-rt}}=\frac{\sum_{j=0}^{s}\frac{1}{j!}(\frac{\lambda_2}{5\mu})^j\frac{1}{(\eta-jm_a)!}\ (\frac{\lambda_1}{\mu})^{\eta-jm_a}}{\sum_{j=0}^{s}\frac{1}{j!}(\frac{\lambda_2}{5\mu})^j\ \sum_{i=0}^{\eta-jm_a}\ \frac{1}{i!}(\frac{\lambda_1}{\mu})^i} \tag{5}$$

The vertical arrows up and down, represent arrival and departure of the real-time traffic into the node. A real-time traffic is blocked if at least $\tau+1$ servers are busy, where $\tau=\eta-m_a$, in other words when less than m_a resource units are available. In order to calculate the blocking probability of real-time traffic $P_{b_{rt}}$, $\eta+1$ probabilities are summed and the probability that a session from the real-time class is blocked can be computed as:

$$\begin{aligned}P_{b_{rt}}=\frac{\sum_{j=0}^{s-1}\frac{1}{j!}(\frac{\lambda_2}{5\mu})^j\ \sum_{i=\eta-(j+1)m_a+1}^{\eta-jm_a}\frac{1}{i!}(\frac{\lambda_1}{\mu})^i}{\sum_{j=0}^{s}\frac{1}{j!}(\frac{\lambda_2}{5\mu})^j\sum_{i=0}^{\eta-jm_a}\ \frac{1}{i!}(\frac{\lambda_1}{\mu})^i}+\\ \frac{\frac{1}{s!}(\frac{\lambda_2}{5\mu})^s\sum_{i=0}^{r}\frac{1}{i!}(\frac{\lambda_1}{\mu})^i}{\sum_{j=0}^{s}\frac{1}{j!}(\frac{\lambda_2}{5\mu})^j\sum_{i=0}^{\eta-jm_a}\frac{1}{i!}(\frac{\lambda_1}{\mu})^i}\end{aligned} \tag{6}$$

III. Admission Control Scheme

The state diagram illustrated in the previous section represents one node in the network, while the network is modeled as z independent M/M/m/m queues. This is illustrated in figure 2, in which each node can be analyzed by the same state diagram illustrated in section II-B.

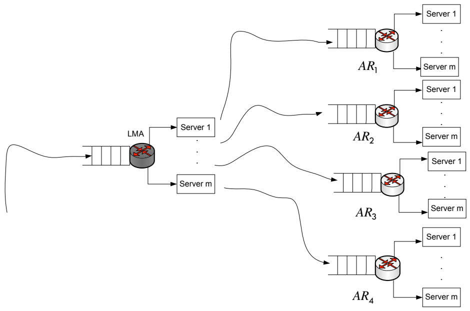


Fig. 2. Independent M/M/m/m queues

Two of the primarily assumptions of this research paper is that the capacity of each node is divided into η resource units and all the traffic regardless of their class are routed through the LMA node or nodes (i.e no routing optimization scheme is incorporated). Looking at the system from the downlink and considering Figure 1 and 2, it can be concluded that the only bottleneck in the system is the LMA node and each session has to pass through two independent queues in tandem before it reaches its destination. This means that a session travels one of these four routes: LMA-AR$_1$, LMA-AR$_2$, LMA-AR$_3$, or LMA-AR$_4$. The admission control scheme for the downlink works as explained below, the same principle stands for the reverse direction (uplink):

- Upon arrival of a new session x at the gateway, the class and destination of the session are distinguished. Destination of session x can be one of the four ARs, i.e. $D_x\in\{AR_1,AR_2,AR_3,AR_4\}$ and class of session x can be either non real-time or real-time.

- The decision as to admit or block the session at the LMA is made on the basis of the number of available resource units and number of required resource units.
- After the session is admitted at the LMA, it occupies one or m_a of the η resource units depending on its class of traffic during the session's life-time (till the end of its service time).
- If enough resource units are available in the second or last node, the session will be forwarded to the second node and remains in one or m_a of the η servers of the second/last node till its service time is finished. Otherwise, the session gets blocked.

Algorithm 1 describes the decision process of the proposed class-based admission control.

Algorithm 1 Class-based Admission Control Scheme

```
flow $x$ arrives at the gateway
find destination MAG $m \in M$ and class of flow $x$
**if** enough resources available at LMA $k \in K$ **then**
    flow $x$ is admitted at the LMA $k$
    **if** flow $x$ is real-time **then**
        **repeat**
            **rt resource allocation**
            **if** enough resources are available at the next node on $P^{MAG}$ towards MAG $m$ **then**
                **forward**
            **else**
                **block**
            **end if**
        **until** MAG $m$ is reached
    **else**
        **repeat**
            **non-rt resource allocation**
            **if** enough resources are available at the next node on $P^{MAG}$ towards MAG $m$ **then**
                **forward**
            **else**
                **block**
            **end if**
        **until** MAG $m$ is reached
    **end if**
**else**
    **block**
**end if**
**rt resource allocation**: assign $m_a$ resource units
**non-rt resource allocation**: assign $m_b$ resource units
**forward**: flow $x$ is forwarded to the next node on $P^{MAG}$ towards MAG $m$
**block**: flow $x$ is blocked
```

IV. Performance Evaluation

The proposed solutions in this paper are further investigated by means of MATLAB simulations. The main focus is on the results of simulations carried out in the MATLAB but also results of the mathematical analysis are presented to verify the simulation environment. We use a uniform random generator to generate our flow or session arrivals, bandwidth of each real-time and non real-time session is considered to be $a = 500$ Kbps and $b = 100$ Kbps respectively. It is also assumed that $m_a = 5$ and $m_b = 1$, which means each real-time traffic will access the five resource units simultaneously whereas each non real-time traffic only occupies one resource unit at a time.

A. Blocking Probability of One Node

In this section only one node is considered and the results of the analytical model developed in section II-B are compared to the simulation results. The capacity of the LMA node is assumed to be $C_T = 5$ Mbps, and η is equal to 50. Simulation results illustrated in Figure 3 are achieved by running 100 rounds of simulation using different seeds. Figure 3 shows the simulation and analytical results on the same plot, while their difference is less than 5 percent. This gives us confidence to further extend the simulation environment to a tandem queuing network. The high blocking probability of real-time traffic in the range of 0.5 and 1, is due to only 30% of this low total arrival rate being real-time.

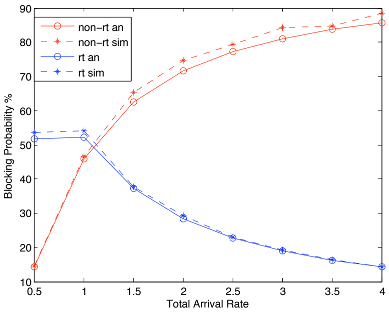


Fig. 3. Blocking Probability vs arrival rate, where arrival rate is total arrival rate of real-time and non real-time traffic, with 70% of the total arrival rate being non real-time at all the points

B. Blocking Probability (Total and per class)

To investigate the impact of our proposed admission control scheme on the total and per class blocking probability, we consider the topology shown in Figure 1. In the examined topology, there are one gateway node, four AR nodes (MAGs) and five intermediate nodes including the LMA node that provide the backhaul routing. Here after η is equal to 200. We assume the P shortest paths from the gateway to each of the four ARs are computed based on the Dijkstra algorithm [10] (Chapter 5, Section 5.2.3).

Looking at the system from the downlink and referring to Figure 1 and 2, it can be stated that each flow arrival at the gateway has to go through 2 nodes till it reaches the MN at one of the serving ARs (or MAGs) and that the LMA node is in tandem with the other four ARs.

In the simulation scenarios, the total arrival rate is $\lambda_1 + \lambda_2$. Statistics show that large part of traffic in the Internet is associated to the peer to peer download [12], while previous report by Cisco mentioned 70% of traffic is associated to the peer to peer download [11]. Therefore in our simulation scenario 70% of the total arrivals are non real-time traffic. Similar to the previous section, each result is produced based on 100 rounds of simulations using different seeds. Firstly the total blocking probability for each arrivals rate is computed. This result is then compared to the method where there is no distinction between the real-time and non real-time traffic in terms of number of resource units used i.e $m_b = m_a = 5$, and sessions from both classes of traffic require the simultaneous use of five resource units at a time. This is shown in Figure 4, it can be gathered that the proposed admission control

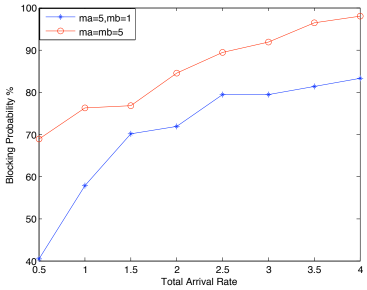


Fig. 4. Total Blocking Probability vs arrival rate,where arrival rate is total arrival rate of real-time and non real-time traffic, with 70 percent of the total arrival rate being non real-time at all the points

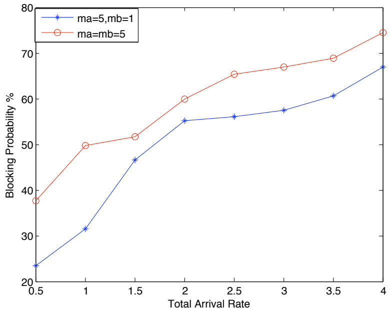


Fig. 5. Blocking Probability of non real-time traffic vs arrival rate, where arrival rate is total arrival rate of real-time and non real-time traffic, with 70 percent of the total arrival rate being non real-time at all the points

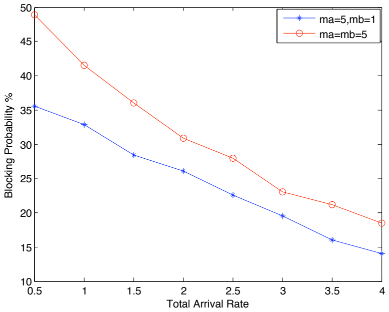


Fig. 6. Blocking Probability of real-time traffic vs arrival rate, where arrival rate is total arrival rate of real-time and non real-time traffic, with 70 percent of the total arrival rate being non real-time at all the points

scheme results in a significant decrease in total blocking probability comparing to the case where sessions from both classes of traffic are treated the same. Moreover, as the session arrival rate increases and as a result the congestion level in the network becomes higher, the gap between the blocking probability of the proposed admission control and the case where $m_b = m_a = 5$ becomes larger and our proposed admission control scheme outperforms the conventional method considerably. Secondly, blocking probability per class of traffic using our proposed admission control is attained. This is then compared with the scenario where $m_b = m_a = 5$. Results in Figure 5 demonstrate that by using our proposed admission control, blocking probability of non real-time traffic has lowered. This was expected as in our proposed scheme each non real-time traffic requires one resource unit at a time at each node, whereas with no distinction between real-time and non real-time traffic each non real flow requires the simultaneous use of $m_a = 5$ resource units at each node.

The blocking probability of real-time class of traffic is computed, results of which are compared to the case where $m_b = m_a = 5$ in Figure 6. Our admission control scheme results in reduction of blocking probability of real-time traffic, especially under high congestion in the network. As the total session arrival rate increases, the number of arrivals from the non real-time class of traffic increases 70% more than the real-time class of traffic. As a result, $P_{b_{nrt}}$ increases and $P_{b_{rt}}$ decreases as the total arrival rate increases. Overall it can be concluded that by treating the two classes of traffic differently we have managed to achieve a great reduction in both total and per class blocking probabilities.

V. Conclusions

This research paper has considered modeling the nodes within the PMIPv6 network as independent two-dimensional M/M/m/m queues. We propose an admission control scheme where the real-time and non real-time traffic are treated differently. Each real-time traffic requires simultaneous use of $m_a > 1$ resource units, whereas each non real-time traffic is admitted at one of the m resource units. Our proposed admission control scheme is briefly explained and finally the results are illustrated and discussed in detail. Simulation results demonstrate promising benefits in terms of reducing the total blocking probability as well as blocking probability per class of traffic.

An interesting extension to our proposed work is investigating the scenarios where more bottleneck nodes are present in the network, and also topologies where more nodes are in tandem per path. Moreover, using an optimized routing strategy whilst applying this admission control scheme maybe an attractive dimension to delve into.

References

[1] D. Johnson, C. Perkins, and J. Arkko, “Mobility Support in IPv6,” *RFC 3775*, june 2004.

[2] K.-S. Kong, W. Lee, Y.-H. Han, M.-K. Shin, and H. You, “Mobility Management for All-IP Mobile Networks: Mobile IPv6 vs. Proxy Mobile IPv6,” *IEEE Wirel. Commun.*, vol. 15, no. 2, pp. 36–45, april 2008.

[3] A. Campbell, J. Gomez, S. Kim, A. Valko, C.-Y. Wan, and Z. Turanyi, “Design, Implementation, and Evaluation of Cellular IP,” *IEEE Personal Commun.*, vol. 7, no. 4, pp. 42–49, auguest 2000.

[4] R. Koodli, “Fast Handovers for Mobile IPv6,” *RFC 4068*, july 2005.

[5] H. Soliman, C. Castelluccia, K. El Malki, and L. Bellier, “Hierarchical Mobile IPv6 Mobility Management (HMIPv6),” *RFC 4140*, august 2005.

[6] P. Reinbold and O. Bonaventure, “IP Micro-Mobility Protocols,” *IEEE Commun. Surveys Tutorials*, vol. 5, no. 1, pp. 40–57, quarter 2003.

[7] S. Gundavelli, K. Leung, V. Devarapalli, K. Chowdhur, and B. Patil, “Proxy Mobile IPv6,” *IETF RFC 5213*, august 2008.

[8] A. D. Pragad, V. Friderikos, P. Pangalos, and A. H. Aghvami, “QoS Aware Dynamic Route Optimization for Proxy Mobile IPv6 Networks,” *Wirel. Commun. Mob. Comput.*, vol. 11, no. 4, pp. 508–521, april 2011.

[9] A. Pragad, P. Pangalos, V. Friderikos, and A. Aghvami, “Dynamic qos aware route optimization for networks with mobility agents,” in *IEEE Consumer Commun. and Net. Conf. (CCNC)*, january 2009, pp. 1 –5.

[10] D. P. Bertsekas and R. Gallager, *Data Networks (2nd Ed.)*. Prentice Hall, 1992.

[11] *Managing Peer-To-Peer Traffic With Cisco Service Control Technology*, Cisco White Paper, 2005.

[12] *Cisco Visual Networking Index: Forecast and Methodology, 2011-2016*, Cisco White Paper, 2012.

[13] L. Kleinrock and R. Gail, *Queueing Systems: Problems and Solutions*. John Wiley and Sons, 1996.