M3S – MULTIMEDIA MOBILITY MANAGEMENT AND LOAD BALANCING IN WIRELESS BROADCAST NETWORKS

Bogdan Ciubotaru, Gabriel-Miro Muntean

Performance Engineering Laboratory, RINCE
School of Electronic Engineering
Dublin City University
Dublin, Ireland
{bogdan, munteang}@eeng.dcu.ie

Keywords: Wireless broadcast networks, load balance, mobility management.

Abstract

Wireless local-area networks represent a viable broadband Internet access solution for enterprise, residential and public areas. Due to its short range radio, multiple wireless access points are necessary to cover a certain physical area. The tendency of mobile users to group in certain areas of interest and of mobile devices to connect to the access point with the highest signal strength determines the overall network load to be highly unbalanced. To overcome this issues which drastically affects user bandwidth share as well as the efficiency of network resource utilization, special load balancing techniques has to be employed. This paper presents the Multimedia Mobility Management System (M3S), a quality oriented mobility management framework which aims at maximizing user perceived multimedia quality by efficiently distributing the traffic load over all the communication resources available. Simulation based testing results are presented, outlining the performance of M3S against other load balancing techniques which rely on mobile device re-association with least congested access points.

1 Introduction

Using wireless technologies as the last link in the Internet access networks have uncountable advantages including user mobility and ubiquitous service providing. Unlike the wired communication technologies where each link uses a dedicated transport medium, wireless mobile devices have to share the common radio spectrum which is a scarce and consequently highly valuable resource.

Wireless local-area networks (WLAN) defined by IEEE 802.11 standard represent a viable solution for wireless broadband access for enterprise, residential and public areas. Due to its short range radio, multiple wireless access points are necessary to cover a certain area. As different studies has shown [1, 3] mobile users tend to group in certain areas of interest (hot-spots) determining an uneven distribution of traffic load among network access points (AP). Traffic load unbalance in hot-spots is usually determined by the fact that mobile devices scan the wireless channels to find the nearby APs and connect to the AP with the strongest signal (RSSI).

Selecting the AP only based on signal strength correlated with mobile users inclination to group in certain areas will determine the AP closest to the hot-spot to be highly loaded while the other APs to be lightly or totally unloaded. This load unbalance affects drastically the bandwidth share that the user is allocated as well as the optimum network resource utilization.

The network load unbalance can be reduced by using special user association control algorithms [4]. Selecting the AP to which a mobile device will associate when entering the network, based on AP load metrics, can significantly improve the effect on other users’ bandwidth share and consequently user perceived quality-of-service (QoS).

As user QoS demands are highly dynamic related to both time and location, simply by controlling the users admission to the network does not optimize the network resource allocation on a long term basis. To overcome traffic load variations due to user mobility and dynamic demands, load balancing algorithms needs to be implemented. The main goals of association control and load balancing techniques is to guarantee a minimum user bandwidth share and QoS level while efficiently exploiting network resources.

In this paper the Multimedia Mobility Management System (M3S) is presented. M3S is a mobility management framework for multimedia applications aiming at improving user perceived multimedia quality while using wireless networks for content delivery. M3S efficiently distributes the overall multimedia traffic over multiple simultaneous connections, each connection using a different AP to exchange data packets. Load balancing is performed using M3S’s core handover management component, Smooth Adaptive Soft-Handover Algorithm (SASHA) which is a quality oriented handover management system. SASHA gradually transfers the load from one connection
(communication channel) to another, based on a Quality of Multimedia Streaming (QMS) metric. QMS is a comprehensive quality metric used to assess the capacity of each communication channel to provide a certain level of quality for multimedia content delivery in the context of an overall QoS demand.

By using multiple simultaneous connections M3S can provide better user bandwidth share and QoS provisioning then a load balancing solution which rely on re-association with a less loaded AP when the current AP is overloaded.

The remainder of this paper is as follows. Section 2 presents the related work regarding load balancing techniques. M3S architecture is presented in Section 3 while the simulation-based testing environment and performance evaluations are discussed in Section 4 and 5. In Section 6 conclusions and future work are presented.

2 Related Work

Load balancing in wireless access networks have already received significant attention, several solutions being proposed in the literature.

Various wireless network equipment manufacturers have already incorporated load balancing features in their products [12, 13]. In these implementations the wireless APs broadcast load information within the beacon messages informing the potential candidates for association about their level of congestion. Based on this information wireless mobile devices can chose to associate with the least congested AP then with the closest one, which presents the highest RSSI.

To achieve maximum efficiency, load metrics represent a very important component of association control and load balancing techniques. Different metrics are proposed in the literature [10], typically relying on the number of mobile devices currently associated with an AP, the RSSI of users currently connected to the AP or the bandwidth a new user can get if it is associated with a certain AP.

Network load balancing techniques can be developed as centralized or distributed algorithms. Some of the proposed solutions can be used in both centralized and distributed systems.

Balachandran et al. present a centralized load balancing solution in [2]. To access the network, mobile users have to submit to the admission control server a service level specification request which contains the minimum and maximum bandwidth required. Based on the overall network load information the central coordination server can grant access to the currently associated AP or can advise the mobile device to switch the radio channel and associate to another, least congested AP. To improve network resource allocation on a larger scale the central coordination server may advise the mobile user to change its location within the range of other APs which are very lightly loaded and can provide a better QoS. Another centralized solution is presented in [4].

The proposed algorithm runs on a network operation centre (NOC) and decides on the optimum user-AP association to improve resource fare-shearing. Due to user mobility and network dynamics, NOC periodically reassesses the network load and recalculates the optimal user-AP association.

The main drawbacks of a centralized approach to association control and load balancing is the reduced scalability, single point of failure represented by the centralized coordination unit and also high maintenance costs involve by the central unit.

A completely distributed load balancing solution is presented in [11]. In this approach, agents running at each wireless AP, monitor the congestion level of its host. These agents communicate over the wired backbone infrastructure, used to connect the APs to the core network, broadcasting and receiving load information to and from other agents residing on neighbour APs. Based on this information each agent can determine locally the distribution of network resources. Each agent will estimate the load level of its AP, which can be under-loaded, balanced and overloaded. Under-loaded APs will accept new mobile stations while overloaded APs will determine some of the already connected mobile devices to handoff to other least loaded APs. Distributed load balancing based on “cell-breathing” technique is discussed in [9].

The main drawback of distributed load balancing techniques is the autonomous decision making performed by each network device which can involve unpredictable behaviour for the network, leading to suboptimal resource allocation or even service quality degradation.

The work in [6] presents a load balancing algorithm which can be used both as centralized and distributed solution. This algorithm considers time and channel allocation in an interactive manner aiming at optimizing network resource allocation in the context of heterogeneous fairness and service requirements.

3 M3S System Architecture

Most of the load balancing techniques relies on determining the optimal user-AP association and routing the whole traffic through the selected AP. In certain areas when the network is highly loaded the infrastructure may not be able to accommodate a new user with its required bandwidth share only by associating it with a single AP. In this situation a multi-connection load balancing technique has to be employed. Such a solution is mostly appropriate for multimedia applications where multi-stream transport solutions are already proposed [8].

3.1 Multimedia Mobility Management System

Multimedia Mobility Management System is a quality oriented mobility management framework for multimedia applications. M3S aims at maximizing user perceive multimedia quality by distributing efficiently the traffic load over multiple simultaneous connections (communication channels). Figure 1 generally describes the principle of using multiple simultaneous connections to deliver multimedia content to mobile users. In the context of M3S multiple
simultaneous connections refers only to the wireless links, assuming that the APs are connected to the core network via wired high-speed backbone infrastructure which is capable of providing the required bandwidth and QoS. Also it is assumed that the mobile device is capable of parallel communication over multiple channels. This can be achieved by using multiple interfaces or possibly the more flexible concept of software radio [5].

As presented in Figure 1 M3S consist of a client side module and a server side module. Figure 2 shows a more detailed architectural representation of M3S. The client-side module scans the wireless medium determining the available APs which have a sufficient RSSI level to sustain a communication channel to the server. The client opens a connection with the server mostly for each of the wireless resource available (AP). The traffic allocation to these connections is performed by the server based on the Quality of Multimedia Streaming metric which is assessed for each individual connection and periodically updated to monitor and react to user and network dynamics.

To provide the information required by the server to assess QMS, the client monitors QoS parameters on each connection and also harvest information related to user preferences, wireless interfaces power consumption and data transfer service costs. As depicted in Figure 2, IEEE 802.21 Media Independent Handover (MIH) may be used to harvest network related information.

$$QMS^s = w_1 * QoS_{s, grade} + w_2 * QoE_{s, grade} + w_3 * Cost_{s, grade} + w_4 * Peff_{s, grade} + \sum_{i} w_i * UPref_{i, grade}$$ (1)

For maximum efficiency and flexibility weights are associated with each component. Weights normalization is required, so the condition from equation (2) has to be respected.

$$\sum_{i} w_i = 1$$ (2)

3.3 Smooth Adaptive Soft-Handover Algorithm

SASHA is a novel quality oriented handover management and is the core component of M3S. SASHA performs handover from one AP to other by gradually transferring the load from one connection to the other based on QMS scores. Figure 3 presents an example of load distribution between two APs performed using SASHA in the situation of a mobile device moving within the overlapping area of the two APs. In stage 1 the mobile node is closer to AP1, all the traffic being routed through AP1 due to better QMS scores of AP1 connection determined by the AP2 link fading. In stage 2 the mobile node is positioned in between the two APs, QMS scores being similar for both links. In this situation the traffic is distributed evenly among the two connections. In stage 3 the mobile node...
node moves closer to AP2 determining AP1 link to fade and consequently the QMS scores of AP1 to drop. In this situation the whole traffic will be routed through AP2.

For simplicity, the example in Figure 3 was presented in three distinct stages. In a real scenario the QMS scores will be continuously updated following the dynamics of the network determining the traffic to be continuously balanced between the available communication links. The same traffic distribution algorithm, as described above in case of node mobility, can be applied when an AP became congested, requiring the load to be partially or totally transferred to a new AP.

Figure 4 describes the load distribution algorithm used by SASHA. This algorithm is distributed over the M3S client and server modules and consists on several phases. The first phase is to scan the wireless medium and determining the available networks. In the second phase some of the networks are selected, based on user preferences, power consumption or cost, and the corresponding connections with the server are established. These two phases are performed by the client module. The next two phases, phase 3 and 4 are performed by the server. In phase 3 the QMS scores are computed for each individual connection. In phase 4 the traffic is distributed over the existing connections according to the QMS scores. Each connection will transport a share of the overall traffic load, proportional with its QMS score.

To provide the user with at least the minimum level of QoS, which in this case represents bandwidth, the total throughput provided by the existing connections is checked against a threshold negotiated with the user. If the total bandwidth is close to the threshold or in the worst case below the threshold phase 1 and phase 2 of the algorithm are initiated at the client module.

4 Simulation-based Testing Environment

M3S performance was evaluated by simulations conducted using NS-2 Network Simulator [14] enhanced by Marco Fiore’s realistic radio patch [15]. Figure 5 presents the simulation scenario which is based on two access points connected to a common router through wired infrastructure. The router is further connected to a multimedia streaming server. The mobile device is positioned between the two APs within their radio coverage overlapping area. As the performance of M3S is assessed from the point of view of load balancing due to network congestion and not user mobility the mobile device is considered to have a fixed position.

The background traffic is generated by four wireless nodes. Two of the traffic generator nodes are within AP1 coverage area and are associated with it and the other two are with AP2 coverage area and consequently are associated with this AP. The constant bit rate background traffic used for M3S performance evaluation is presented in Figure 6. Four load levels are considered for each AP. Totally unloaded (0Mbps), lightly loaded when the background traffic is 2.5 Mbps, fully loaded when 3.5 Mbps background traffic is generated and overloaded when 4.5 Mbps background traffic is considered. As it can be seen in Figure 6 the sequence of light and high traffic load periods are complementary to avoid the situation...
when both APs are overloaded, making minimum QoS provisioning impossible.

To implement M3S on NS-2 the wireless node was changed to allow multiple simultaneous connections over different wireless channels. M3S was deployed in an application which is capable of sending a constant bit rate multimedia content (1.5 Mbps). A re-association based load balancing technique, referred in this paper as RLB, is implemented by switching the traffic from one AP to the other according to the induced level of congestion. The RLB implementation is optimistic as the time required to assess AP load is not considered. For the situation when no load balancing technique is used the traffic is constantly routed through one of the APs.

5 Performance Evaluation and Result Analysis

The performance of the load balancing technique based on M3S was evaluated using Throughput, Loss as QoS metrics and PSNR as a user perceived quality metric. The simulation-based testing results for M3S load balancing, re-association load balancing (RLB) and the situation when no load balancing is used (NO LB) are presented in Figure 7, 8 and 9.

As presented in Figure 6 three network loading situations were considered. In the first situation, one of the APs is overloaded (4.5 Mbps background traffic) while the other one is totally unloaded (0 Mbps background traffic). In the second case one AP is overloaded (4.5 Mbps background traffic) while the other is lightly loaded (2.5 Mbps background traffic). In the last case both APs are fully loaded (3.5 Mbps background traffic). These loading conditions are sequential in time and repetitive. As outlined in Figure 7, M3S performs well in all these situations presenting insignificant loss rates excepting the period when both APs are fully loaded (100s – 150s) when a 0.08 Mbps loss rate is encountered. From the point of view of PSNR the average values are constantly above 51 db with a maximum of 70 db.

The re-association-based load balancing technique performs well when at least one AP is lightly loaded or unloaded. As it can be seen in Figure 8, the throughput is maximum (1.5 Mbps) for all network load situations except the one when both AP are fully loaded (100s – 150s) when the throughput drops to 1 Mbps. The same behaviour can be observer by analyzing loss and PSNR in Figure 8.

When no load balancing technique is used, as presented in Figure 9, the QoS drops dramatically when the AP is overloaded as the mobile device remains connected to this AP. When the background traffic decreases, the AP becomes lightly loaded or totally unloaded determining a QoS increase to the maximum as it can be seen in Figure 9 (150s – 250s).

In Table 1 the average values of Throughput, Loss and PSNR are presented for three individual load situations (phases). In Phase 1 AP1 has 4.5 Mbps background traffic (overloaded) and AP2 has 0 Mbps background traffic (unloaded). In Phase 2, AP1 has 4.5 Mbps background traffic (overloaded) and AP2 gets 2.5 Mbps background traffic (lightly). In Phase 3 both APs have fully loading background traffic (3.5 Mbps).

As it can be seen in the table the performance of M3S load
balancing and re-association (RLB) load balancing are similar when at least one AP is capable of providing the required QoS. The performance of M3S load balancing improves with 51% in terms of throughput and 157% in terms of PSNR compared to load balancing techniques based on user-AP re-association.

Future works will assess the performance of SASHA with more background traffic patterns including variable bit-rate traffic. Various network loading scenarios will be used as well as different network topologies. Heterogeneous wireless environments will also be considered.

Acknowledgements

The support of Microsoft Research and Irish Research Council for Science, Engineering and Technology is gratefully acknowledged.

References


Table 1: Throughput, Loss and PSNR for three distinct load phases.

<table>
<thead>
<tr>
<th>Phase</th>
<th>Phase</th>
<th>Throughput (Mbps)</th>
<th>Loss (Mbps)</th>
<th>PSNR (db)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SASHA</td>
<td>1</td>
<td>1.50</td>
<td>0.00</td>
<td>57.6</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.50</td>
<td>0.00</td>
<td>56.8</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.46</td>
<td>0.08</td>
<td>51.0</td>
</tr>
<tr>
<td>RLB</td>
<td>1</td>
<td>1.50</td>
<td>0.00</td>
<td>55.3</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.50</td>
<td>0.00</td>
<td>55.5</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.99</td>
<td>0.51</td>
<td>19.8</td>
</tr>
<tr>
<td>NO LB</td>
<td>1</td>
<td>0.69</td>
<td>0.80</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.10</td>
<td>0.39</td>
<td>23.55</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.29</td>
<td>0.20</td>
<td>33.16</td>
</tr>
</tbody>
</table>

Figure 9 Throughput, Loss and PSNR obtained without load balancing

6 Conclusions and Future Work

Load balancing represents a key element in providing high quality wireless broadband access which is mostly required by the users accessing multimedia content over the Internet. This paper presents the Multimedia Mobility Management System (M3S) which performs load balancing using the novel quality oriented handover management system, Smooth Adaptive Soft-Handover Algorithm (SASHA).

M3S improves user bandwidth share and the overall network resource allocation by distributing the application traffic over several simultaneous connections exploiting all the communication resources available. As SASHA distributes shares of the overall application traffic on several connections using different APs the user bandwidth share improves with 51% and the user perceived quality improves by 157% compared to load balancing techniques based on user-AP re-association.

Future works will assess the performance of SASHA with more background traffic patterns including variable bit-rate traffic. Various network loading scenarios will be used as well as different network topologies. Heterogeneous wireless environments will also be considered.


