Operators Intervention Strategies with New MCDM for Load Balancing in Heterogeneous Mobile Networks

Bassem El Zant
Department of Electricity and Mechanics, Saint-Joseph University, Faculty of Engineering – ESIB
Campus of Sciences and Technology, Mar Roukos, Mkalles, Beirut, Lebanon
bassem.elzant@usj.edu.lb

Abstract. Wireless networks are designed to operate independently. With the many existing inconveniences of 3G networks, operators are persuaded to switch to heterogeneous access networks. This will mainly reduce the costs that operators should pay in order to move to 4G; and thus make use of the existing networks infrastructures, to come up with heterogeneous environment where multiple wireless technologies coexist. In this paper, three strategies along with the operators’ intervention are programmed: binary, binary with return to initial condition, and fuzzy. The reason behind this is to select the best strategy (least drop probability, least delay, best performance parameters, etc), which is the fuzzy one; whereby regular filling of the different networks is provided, and thus network fluctuation is well illustrated. As for the worst strategy, it is the binary with return to initial condition; whereby operators will have less resources to exploit, and so less throughput and higher delay.

Keywords: Multi-criteria decision method (MCDM), Heterogeneous Mobile Networks, Satisfaction-based decision method, Binary and Fuzzy logic.

1. Introduction

Wireless networks are designed to operate independently. With the explosion of traffic, telecom operators are confronted with the problem of mobile infrastructure saturation. The use of new technologies (Wireless MAN-Advanced, development of WiMAX), and LTE-Advanced) enables the operators to integrate different radio technologies already deployed; such as pooling of resources of WiFi, mobile WiMAX, the LTE and HSPA +. Besides, taking into consideration the 3 G disadvantages (whereby the deployment has proven to be costly), and the higher costs to be paid in order to migrate to 4G networks or to increase the density of the existing 3G networks, operators are increasingly convinced by deploying heterogeneous access networks. In the heterogeneous networks, multiple wireless technologies coexist, and the radio resource management is coordinated. In such networks, mobile users can connect to different radio access technologies. To optimize the system performance, network operators aim to balance loads, as much as possible, in its various radio access networks.

In heterogeneous wireless networks, the main challenge is to keep connections among the different networks like WiFi, WiMax, WLAN etc. The 4th generation of wireless (NGWN/4 G) networks is expected to present heterogeneity in terms of wireless technologies and services. The mainly advantage of the mobile networks 3G (UMTS and 1xEV-/DV) is their global coverage. However, the weaknesses of 3G lie in their bandwidth capacity and operating costs. On the other hand, the WLAN technology such as IEEE 802.11 offers higher bandwidth with low operating costs, although it covers a relatively short range. In addition, technological advances in the evolution of mobile devices made possible the support of different Radio Access Technologies. This raised much interest for integration and interoperability of 3 G wireless networks and wireless local networks, to take advantage of their respective potentials. In this paper, we will start by defining the heterogeneous mobile networks and their benefits before presenting different intervention strategies.

2. Heterogeneous Mobile Networks

Heterogeneous networks integrate different radio technologies that have been already deployed to combat the problem of network saturation, and to share various resources of the operators. Two methods could be used to access the heterogeneous networks: the loose coupling, and the tight coupling. Besides, there are two entities responsible for the management of the radio resources: the RRM for the local processing of resources, and the CRRM for common resource management.

3. Problem

Network selection is made by the user as per the provided criteria: the throughput and the cost. However, to optimize the overall performance, networks operators should intervene. They play on guarantees of QoS (the offered throughput), and economic incentives (the cost). This will guide the final network selection of the user the way system performance will be the best. To alleviate networks, the operators offer a lower throughput or a higher cost. However, to attract new arrivals, they provide higher throughput, or lower cost. Different strategies are described to change the throughput and the cost.

4. Existing Methods and Related Work

Different methods exist in the literature to improve the quality of services and try to provide solution to the above problems.
In [1] we can see how the handover technique is used to redirect the mobile user’s service network from current network to a new network or one Base Station (BS) to another one or from one Access Point (AP) to another one using the same or different technologies in order to reduce the processing delay in the overlapping area. Handover network type [11] has horizontal and vertical handover. The homogenous wireless network performs horizontal handover, if there are two BSs using the same access technology. This type of mechanism use signal strength measurements for surrounding BSs to trigger and to perform the handover decision. [2] describes the concept of being always best connected describe the user experience and business relationship in an ABC environment and outline different aspects of an ABC solution that broad the technology and business base of 3G. It shows how in heterogeneous wireless networks environment, Always Best Connected (ABC) requires dynamic selection of the best network and access technologies when multiple options are available simultaneously. The Mobile Station (MS) or BS should be equipped with multiple network interfaces to reach different wireless network. The authors in [4] have done a comparison among SAW, Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), Grey Relational Analysis (GRA) and Multiplicative Exponent Weighting (MEW) for vertical handoff decision. The vertical handoff decision algorithm for heterogeneous wireless network has been discussed in [3]. The author formulated the problem as Markov decision process. And the vertical handoff decision is formulated as fuzzy multiple attribute decision making (MADM) in [5]. A vertical handoff decision scheme DVHD uses the MADM method to avoid the processing delay in [6], when the goal in [7] of the authors was to reduce the overload and the processing delay in the mobile terminal. A novel vertical handoff decision scheme to avoid the processing delay and power consumption has been proposed in this paper. A novel distributed vertical handoff decision scheme using the SAW method with a distributed manner to avoid the drawbacks has been suggested in [9] when the authors in [9] define using the emerging IEEE 802.21 standard a Media Independent Handover (MIH) functions as transport service in order to offer a vertical handoff decision with a minimum of processing delay. A four-step integrated strategy for MADM-based network selection has been proposed in [10]. In [12] a comparative analysis of MADM methods including SAW, MEW, TOPSIS, ELECTRE, VIKOR, GRA, and WMC is illustrated with a numerical simulation, showing their performance for different applications such as: voice and data connections, in a 4G wireless system.

5. KEY PERFORMANCE INDICATORS (KPIs)

Different KPIs are defined to assess the user satisfaction and operators gain. To assess the satisfaction of the user, we take into consideration the applications’ types. So, for the streaming and inelastic applications, we consider the throughput, the delay and the drop probability. And for the elastic applications, we consider the throughput and performance test \(d/ d_c\), where \(d\) is the effective flow received per the user and \(d_c\) represents the flow of comfort. It then sets the user’s satisfaction by:

\[
S = w_{QoS}S_{QoS} + w_{cost}S_{cost}
\]

Where \(w_{QoS}\) and \(w_{cost}\) represent the respective weights assigned to the QoS criteria and monetary cost, and \(S_{QoS}\) and \(S_{cost}\) the respective satisfaction of the user. The functions of satisfaction are identical to those defined for the calculation of the best alternative and satisfaction will be measured at the actual rate received by the user. To assess the operator gain we calculate the average revenue per user for consumption per Kbyte.

6. OPERATOR STRATEGIES

The operator strategies are implicitly embedded in the system’s information. This will influence the decision process. The operator will inform the user of the minimum and maximum throughput of each network access with the corresponding monetary cost and the user is who choose the most suitable access technologies among different ones. We can say that the operator will affect and guide the user when making the final decision. In our work, the monetary cost will be fixed when the QoS incentives will vary dynamically, and vice versa to optimize short and long term goals. In both cases the monetary cost and the different incentives for QoS will reflect the operator strategies and contribute in the final decision of the user. The operator will hide the total capacity of the system and will only provide to the users information about the incentives of QoS and presents the different guaranteed throughputs offered with the cost. On the other hand, when the operator is ready to reserve, in a RAT j, a band for the session of the service class i, excellent throughput will be suggested, and then the class i will attract new coming sessions to the RAT j. So to avoid new sessions, the operator can offer excellent throughput or low cost, which will push the user to pass to one RAT. In conclusion, the dynamic incentives of QoS variation or cost variation will allow the operator to more or less attract new users to a class in a specific RAT and then the operator will contribute in the final decision of the user.

7. NETWORK SELECTION PROCESS

Before opening a new session or to make a handover, the mobile must evaluate different alternatives and select the best network and the best class of service. For this multi-criteria decision, we should consider the QoS requirements, radio conditions, the cost and the preferences of the user. The different decisions attributes are as following: the minimum rate guarantees \(d_{min}\), the maximum rate \(d_{max}\) and cost \(C\). Figure 1 shows the hierarchical representation of the criteria.

During the first stage, the general criteria must take into consideration the user’s preferences. The user may prefer
paying more for a better quality of service or he prefers to save money and get low quality. The user therefore assigns weights to QoS incentives and the cost according to his/her preference. However the secondary criteria depend on the type of application, for example the minimum throughput is critical for a CBR application (constant bit rate) when this did not make sense for a Best Effort application. As following distinct types of applications:

a) Inelastic applications: work with constant throughput applications, therefore the maximum throughput will not be considered and has no importance so his weight will be null. d<sub>min</sub> and C will be taken into consideration in this case.

b) Streaming applications: used for real time services with variable throughput (ex: Video service using Mpeg4). Three parameters will be considered for those applications: d<sub>max</sub>, d<sub>min</sub> and C.

c) Elastic applications: used for data transfer services such as the transfer of files, email, and web traffic services. For these applications just d<sub>max</sub> and C will be considered, when d<sub>max</sub> will be ignored because these applications don’t require guarantees in QoS.

To evaluate the different alternatives of a session and select the best choice, a method for multi-criteria (RMD) decision will be adopted. Indeed, it defines a utility function which depends on the standard weight of the values of the various criteria. Several utility functions have been proposed, the simple additive weighting (SAW) and the multiplicative exponent weighting (MEW), but these methods do not consider the current needs of the session. For example, when a user using CBR application and is willing to pay, the best alternative will be severed without taking into account the throughput required for the session, and thus an overqualified alternative will be severed, while there is another less expensive alternative meeting the requirements of the session.

8. SELECTION OF NETWORK ARCHITECTURE

We suggest a hybrid approach taking into account the preferences of the operator and the user. Indeed, the policies of the operator are implicitly integrated in the system information. This information will be sent to all mobile terminals. The mobile terminal will decode the information, assess different options and then choose the best network. The operator offers three classes of services (Premium, Regular, and Economic). Financial and QoS considerations will determine the selection of the best alternative. An alternative is defined by a combination of available access technology (WiMax, WiFi, 3G ...) and classes of services such as Premium, Regular, and Economic. When a new session will be opened or where a hand over, the mobile terminal will receive information from the system, decodes the incentives of the operator, evaluates the different alternatives and then chooses the best network with the service class. The information of the available alternatives includes:

*The monetary cost:* we suggested a cost according to the QoS. As there are different classes of services (Premium, Regular, and Economic) in different networks, the network operators provide different flows for each. Indeed, a Premium session will be more priority, will get a better quality of service and will be more expensive than the Regular, and the Economic. Mobile users will be charged according to their priority (class of service). The prices will change dynamically in real time according to the conditions of the networks, the radio resources used efficiently, and the performance of the system will be improved. As it’s more complex to implement a dynamic price, we consider a static price with the different classes of service which will be fixed and does not change with the network load conditions. The flat pricing strategies are not used because they will cause a waste of resources and will force light users to subsidize heavy users and prevent the deployment of internet service quality. In addition, the use of flat price will result in the congestion of a computer network and will degrade the performance of the system.

Accordingly, as we will provide a guaranteed QoS, we propose a model based on the used volume, and then the sessions will be billed according to the amount of traffic that they transmit. The price of traffic per unit will depend on the radio resource access and the class service. In conclusion, the monetary cost is for each unit of traffic and in our case is defined per Kbyte.

*QoS incentives:* different QoS parameters to consider are based on the application’s requests. They specify the minimum number (n<sub>min</sub>) and the maximum (n<sub>max</sub>) units of radio resource localized for a session. These sub-parameters depend on radio resources and the service class. The total traffic for a specific RAT is hidden. Different n<sub>min</sub> and n<sub>max</sub> are generated for the different classes of services reflecting the strategies of the operator. These sub-parameters don’t necessarily reflect the conditions of the network but rather the operator’s wishes to serve the different sessions of different classes. Based on the SNR report, the mobile terminals will adopt the modulation type and the FEC for the encoding. This is why the number of bits per RRU and the minimum and maximum flows depend on radio’s conditions. Indeed, during the evaluation of different alternatives, the mobile terminal combines its radio conditions (which differed from rat to another) and with the reported QoS Sub-parameters then determines the minimum and maximum expected flow.
for OFDM-based technologies the minimum flow expected will be:

\[
d_{\min} = \frac{d_{\min} \times N_u \times K \times R_c}{g_f}
\]  

(2)

With \(d_{\min}\): the minimum guaranteed of OFDM reserved symbol, \(N_u\): the number of carrier used for data transmission, \(K\): the number of bits per symbol module (vary with the modulation), \(R_c\): FEC report, and \(g_f\): scheduling interval.

To make it more simple and homogeneous the QoS incentives for different radio technologies, we will express the QoS sub-parameters in terms of minimum and maximum guaranteed flow \(d_{\min}\) and \(d_{\max}\) (instead of \(n_{\min}\) and \(n_{\max}\)). To evaluate the different alternatives, the mobile terminal will determine the expected flow that represent the result of multiplication of minimum flow (resp. maximum) of the class of service in the alternative with the gain of modulation \(g_M\) and \(g_c\) coding gain.

\[
d_{\text{user}}^{\text{min or max}} = d_{\text{min or max}} \times g_M \times g_c
\]

(3)

9. New Multi-Criteria Decision Method

As shown in the previous paragraph, there are different existing methods used to decide the selection of an alternative. The method used in this paper is a multi-criteria decision method (MCDM). It defines the alternative as a combination between a network and one of the classes of service Premium, Regular or Economic. The alternatives will be evaluated according to their monetary cost, minimum flow that they guarantee and the maximum rate they offer based on the satisfaction of the user. We define then a function of satisfaction for each type of session (inelastic, streaming and elastic) and user profile. The HWAN (or NGWAN) allow the efficient use of available radio resources and then they can serve more customers that will generate more profit. Mobile users can connect simultaneously or not, the different access technologies that meet their needs in terms of QoS or cost. As cited before, our method is a hybrid method (shared between network and users). We define an environment that integrates the operator’s objectives and the user’s preferences. So our method of decision is based on the user’s satisfactions. Our method selects the best alternative based on the expected user satisfaction. The utility function is defined as the weighted sum of partial satisfaction functions. The function of partial satisfaction \(s_c(p)\) depends on the decision criterion \(c\) and the profile of the user \((p)\). There are two types of users: those who are willing to pay for best performance and those who prefer to save. As mentioned before, there are three different types of applications, so we will have six users’ profiles:

a) Profile 1: The user Initializes an inelastic session and is willing to pay.

b) Profile 2: The user Initializes a streaming session and is willing to pay.

c) Profile 3: The user Initializes an elastic session and is willing to pay.

d) Profile 4: The user Initializes an inelastic session and prefers to save money.

e) Profile 5: The user Initializes a streaming session and prefers to save money.

f) Profile 6: The user Initializes an elastic session and prefers to save money.

The function of satisfaction expected \(S(a_i)\) for the alternative \(a_i\) is given by:

\[
S(a_i) = w_{d_{\min}, p} \times S_{d_{\min}, p} + w_{d_{\max}, p} \times S_{d_{\max}, p} + w_{\text{cost}, p} \times S_{\text{cost}, p}
\]

(4)

Where \((w_{d_{\min}, p}, w_{d_{\max}, p}, w_{\text{cost}, p})\) represents the static weight vector of profile \(p\), and \(S_{d_{\min}, p}, S_{d_{\max}, p}, S_{\text{cost}, p}\) represent the functions of partial satisfaction of profile \(p\).

The function of satisfaction of flow depends on the QoS needs of the session. Inelastic applications are characterized by a fixed flow, \(R_f\), and the QoS requirements for these applications are strict and inflexible and therefore the function of satisfaction of the minimum flow ensures is defined by:

\[
S_{d_{\min, p}} = \begin{cases} 
0 & \text{if } d_{\min}(a_i) < R_f \\
1 & \text{if } d_{\min}(a_i) \geq R_f 
\end{cases}
\]

(5)

Figure 2. Form of the function of satisfaction for an inelastic session

Where \(d_{\min}(a_i)\) represents the minimum rate guaranteed by the alternative \(a_i\). In this case the function of maximum satisfaction will not be considered.

Streaming sessions require a minimum rate but also a maximum flow as it is real-time applications. Their function of satisfaction is in the form of sigmoid and is defined by:

\[
S_{d_{\min, p}} = 1 - \exp \left( -\frac{\alpha (d_{\min}(a_i) - R_{av})}{R_{av}} \right) 
\]

(6)

\[
S_{d_{\max, p}} = 1 - \exp \left( -\frac{\alpha (d_{\max}(a_i) - R_{av})}{R_{av}} \right) 
\]

(7)

Where \(\alpha, \beta\) are positive constants that determine the shape of the sigmoid, \(d_{\max}(a_i)\) and \(d_{\min}(a_i)\) are the maximum and minimum flow guaranteed by the alternative \(a_i\) and average flow \(R_{av}\).
We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord “Format” pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.

Figure 3. Form of the function of satisfaction for the streaming sessions.

For elastic sessions, the function of satisfaction is a concave function defined by:

\[ S_{d_{\text{max},p}} = 1 - \exp\left( -\frac{d_{\text{max}}(a_i)}{R_c} \right) \] (8)

Where \( R_c \) is the flow of comfort.

Figure 4. Form of the function of satisfaction of the elastic sessions.

Satisfaction of the monetary cost function:

This function depends on the user’s tolerance, it is modeled by:

\[ S_{\text{cost},p} = \exp\left( \frac{\text{cost}(a_i)}{\lambda_p} \right) \] (9)

Where \( \text{cost}(a_i) \) is the monetary cost of the alternative \( a_i \) and \( \lambda_p \) a positive constant that depends on the profile \( p \). More the user is tolerated in terms of cost, more \( \lambda_p \) is greater (\( \lambda_1 = \lambda_2 = \lambda_3 \) and \( \lambda_4 = \lambda_5 = \lambda_6 \)).

Figure 5. Form of the function of satisfaction of the monetary cost.

10. SIMULATION

For the simulation, three similar Radio Access Technologies (R1, R2 and R3) are considered. They all have the same service classes (Same Cost and QoS incentives for all classes: Premium, Regular and Economic). Our goal behind this simulation is to let the user make a choice in terms of Cost and QoS incentives; as he won’t be concerned with the technology behind this.

As a result, there will be possibility that there is no load balancing: R1 can be filled first, then R2 and R3. To avoid this case, it is necessary that the operators intervene in the network. The intervention will occur at special moments that will be determined later for load balancing.

10.1. Discrete Events System

A discrete event system is a system described by discrete state variables, i.e. changes occur on the occurrence of a set of states.

We have different types of possible events during the lifetime of the system; thus, we must describe the operating logic between events (determine state changes for each event and the events that result). In our system, we define three main events: session Arrival (A), session departure (D) and the end of a frame (FT).

Figure 6: Discrete Events System

If the event is an arrival we must expect the departure, mark the session as active, and expect the new arrival. If it’s a departure, we must free the resources and mark the session as terminated. And if it is an end of frame, there must resources allocation.

Matlab has been used to make the simulation, as well as the programming of the intervention of the operator under different strategies. The operator can intervene either by adjusting the Cost or the QoS incentives.

The following approaches have been simulated for the adjustment of QoS incentives:
10.1.1 Binary logic
The QoS incentives are decreased in each threshold.

\[
D = \begin{cases} 
D_{\text{init}} & \text{if Network Occupancy} < S_1 \\
D_{\text{init}} \cdot X & \text{if } S_1 < \text{Network Occupancy} < S_2 \\
D_{\text{init}} \cdot Y & \text{if Network Occupancy} > S_2 
\end{cases}
\]

10.1.2 Binary logic with return to initial conditions
It’s similar to the binary logic. But in this scenario, the value of the QoS incentives are returned to their initial values after decreasing the QoS incentives in the three networks in each threshold.

\[
D = \begin{cases} 
D_{\text{init}} & \text{if Network Occupancy} < S_1 \\
(D_{\text{init}} - D_f) \cdot (d_{\text{min}} + \text{Total} - S_1) / (S_2 - S_1) & \text{if } S_1 < \text{Network Occupancy} < S_2 \\
D_f & \text{if Network Occupancy} > S_2 
\end{cases}
\]

All the above scenarios could be programmed by increasing costs instead of decreasing the QoS incentives.

11 Results

- QoS (binary logic)
- QoS (binary logic + R)
- QoS (Fuzzy logic)

As per the above figures, it is obvious that the fuzzy logic intervention offers the best performances for both the inelastic and the partially elastic sessions. These performances are shown by the delay (Figure 8, 10), and by the drop probability (Figure 7, 9). The main reason is that the different networks are filled in a regularly and continuous way with the fuzzy logic. This will give the sessions a higher amount of resources to exploit in each network. Therefore, sessions will have a higher throughput and thus a lower delay and drop probability. On the other hand, it’s clear that the binary logic with return to initial conditions provides the worst performance. Whereby, the operators are having fewer resources to exploit by offering higher throughputs with initial conditions.

As per the binary logic, it has approximately the same results as the fuzzy logic. The fuzzy logic could be considered as binary logic with multiple thresholds.
Different operators’ intervention strategies were presented. Each strategy is based on a specific logic: binary, binary with return to initial conditions, and fuzzy logic. After the implementation of these strategies, it was obvious that fuzzy logic presents the best performances; whereby the lowest delays, drop probability and the highest performance parameters are met. Besides, the operators benefit from a higher financial gain. The results also show that the operators should reduce the QoS incentives.

For future work, these methods could be implemented with SON (Self Organization Networks), which is a new technology presented in the LTE standard for auto-configuration, auto-optimization and auto-exploitation of cellular networks equipments for mobile telephony.

13 Acknowledgment

I would like to thank my classmate Rita Maria Salameh and Mr. Melhem El Helou for their great support.

14 REFERENCES