A Survey for Load Balancing in Mobile WiMAX Networks

Camellia Askarian[1] and Hamid Beigy[2]

[1]School of Engineering and Science, Sharif University of Technology-International Campus, Kish Island, Iran
askarian.camellia@gmail.com
[2]Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
begiy@sharif.edu

ABSTRACT

Mobile Worldwide Interoperability for Microwave Access (Mobile WiMAX) is a wireless access technology based on IEEE 802.16e standard. Mobile WiMAX supports mobility, high utilization of radio resources, beneficial Quality of Service (QoS) framework, and flexible centralized scheduling, which are offered by connection oriented MAC layer. Due to the architecture of MAC layer in mobile WiMAX, admission control, packet scheduling, handover, and load balancing mechanisms can be implemented in the MAC layer by vendors. In this paper, we explain two main features in MAC layer of Mobile WiMAX networks; handover and load balancing. Additionally, we investigate load balancing algorithms suggested in wireless and WiMAX networks to find out appropriate way of load balancing in mobile WiMAX networks.

KEYWORDS

IEEE 802.16e, Mobile WiMAX Networks, Load Balancing, Quality of Service, Handover.

1. INTRODUCTION

Mobile Worldwide Interoperability for Microwave Access (WiMAX) is an IEEE broadband wireless access technology, aims to provide mobility enhancement for Mobile Stations at the vehicular speed for long distances, in a mobile environment [1]. Fixed WiMAX, is based on the IEEE 802.16-2004 standard [2] and supports only restricted coverage and roaming. IEEE 802.16e edited*physical*(PHY) and *medium access control* (MAC) layers contents to improve the system performance.
WiMAX compromises numerous advantages, such as improved performance and robustness, end-to-end IP-based networks, secure mobility and broadband speeds for voice, data, and video, support for fixed and mobile systems, efficient and adaptive coding and modulation techniques, scalable channel sizes, sub-channelization schemes, multiple-input-multiple-output antenna systems, and quality of service [3]. Mobile WiMAXintroduces proper mobility to the WiMAX system. Comparing 802.16 standard and other 802 families, 802.16 leads a very high utilization of radio resources and a good Quality of Service (QoS) framework provided by connection oriented MAC and responsive centralized scheduling.

The main goal of this standard is to provide mobility development support for Mobile users moving at the vehicular speed. The IEEE 802.16e introduces many changes to PHY and MAC layer protocols due to mobility support it required addressing new issues that were not necessary in 802.16-2004, such as handover.

DOI : 10.5121/acij.2012.3213

Since MAC layer in IEEE 802.16e supports handover and load balancing features which are basic elements of this paper, this layer will be explained.

As the Mobile WiMAX is a kind of a multi-service wireless network, a major challenge is the optimal allocation and utilization of the available radio link between users and resources. Therefore, resource utilization and QoS within the radio link of one *base station* (BS) have been taken into the consideration and are still the object of many research projects. As WiMAX progresses to include mobility, a cellular infrastructure and overlapping cells, system wide *radio resource management* (RRM) and QoS seem to be attracting.

Based upon the mentioned reasons, we aim to introduce the best solution for this problem. According to our research, it seems applying load balancing algorithms which concentrate on controlling load of the system and keeping it in the balance state by triggering handover process can be a proper solution. Though, the main approach of this survey is load balancing algorithms with handover.

In mobile WiMAX networks, the aim of load balancing process is to control and manage the system in the way that users can achieve guaranteed QoS. In order to satisfy this goal, mobile users should switch from current *serving base station* (SBS), which is highly loaded, to lightly loaded BSs.Based on the key features of IEEE 802.16e, WiMAX forum network architecture and also handover procedure in WiMAX networks, this paper focuses on the process of conducting load balancing with handover in mobile WiMAX networks and load balancing algorithms in wireless and WiMAX networks.

The rest of this paper is organized as follows. The process of load balancing, load definition and load measurement will be explained in section 2. In addition to these concepts, load balancing methods will be classified and described in more details. The process of handover in mobile WiMAX networks will be introduced in the section 3. Subsequently, in sections 4 and 5 a literary review of the load balancing algorithms in wireless networks and WiMAX networks will be given. Finally, in the last section, the survey is summarized.

2. Load Balancing Process

Usually, load balancing can be described as the process of dividing and distributing jobs among more than one server, accordingly more jobs can be served and the entire system can perform more efficiently. Load balancing has been frequently operated in computer systems for load sharing, but has also been used in telecommunications [4].

Generally, load balancing can be done in a static or dynamic way. Static load balancing is independent of situation of the system whereas in dynamic load balancing, decisions are made regarding to the current loading status of the system and availability of resources.

Load definition, load measurement and load balancing mechanism are essential elements in load balancing process [5]. In what follows, the definition of load, the process of load evaluation, and also load balancing mechanisms in mobile WiMAX networks will be presented.

2.1 Load Definition

The role of load metric in load balancing algorithms is to estimate whether a system is balanced or not. Hence, it should be able to describe the load situation of the system accurately regarding to the utilization of shared resources. The shared resources in the radio link of an OFDMA system, such as WiMAX networks are time, frequency, and power. The consumption of these resources depends on the transmission power and the modulation and coding scheme (MCS) [4].

Load metric has been defined in several ways based on the features of the networks. The most common metrics which were used in traditional cellular networks are the number of calls and the probability of call blocking, while packet loss, throughput and delay were used in wireless networks such as wireless LAN (WLAN).

In [6] the load metric defined by considering several factors such as the number of users currently connected to an access point and the mean *received signal strength indicator* (RSSI) value of these users. The proposed load metric in [7] is based on the RSSI value and the bandwidth that can be gained when a new user connects to an access point. However, these factors are not sufficient to predict the probability of collisions and available systems' bandwidth. The authors in [7] proposed bandwidth as the main metric, which is the percentage of time the access point is busy transferring data during some time interval.

As in the mobile WiMAX networks, *mobile stations* (MS) can have many service flows and each of them may have different features, evaluating the number of users or connections cannot be accurate enough. Additionally, throughput does not consider the type of applied MCS into account, thus throughput cannot recognize when the maximum resource utilization has been reached. Also, delay, the rate of packet loss, and rate of call blocking only represent implicit information of the load situation of the system and can be useful only for decision making [4].

The major resource measurement component in mobile WiMAX is one slot. This is a proper and accurate indicator of resource utilization since it describes the resources not just in terms of throughput, but also regarding to the applied MCS and thus, channel conditions are also considered [4]. It has also been an accepted choice of load metric in the WiMAX forum network architecture [8].

2.2 Load Measurements

In load balancing algorithms a criterion is needed which can estimate load situation of the system. As it was stated in [5] another important component in load balancing algorithms is the load measurement. This element is used to evaluate the load status of the system. Load measurement can be used as an awareness of starting the calculation of load metric[9]. Based on the resulted value, the system is managed by initiating a load balancing procedure.

Due to the fact that varied traffic generated by MSs in mobile WiMAX networks cause in different values of throughput and resource utilization, realizing the load state of the system by comparing throughput or even resource utilization of BSs is not feasible. Hence, the load state of the whole system should be described with one value.In order to achieve this value, the following balance index has been introduced in [10]:

$$\beta = \frac{\left(\sum_{i=1}^{n} U_i\right)^2}{n\left(\sum_{i=1}^{n} U_i^2\right)}, \qquad (1)$$

where index β provides a value between 0 and 1, where 1 specifies that the system is balanced and 0 represents an unbalanced state. The value n indicates the number of BSs, and U_i is the resource utilization of BS_i. To evaluate how balanced the system is, the average load of the entire system should be calculated in the following way as given in[11]:

$$L = \frac{\sum_{i=1}^{n} U_i}{n}. \qquad (2)$$

The gained value of this formula should be compared with the individual resource utilization $U_i$of each base station. As it is mostly probable, the uplink and downlink subframe division in mobile WiMAX are static resource utilization of a BS will be defined as follows [4]:

$$U_i = max\left(U_{DL,i}(A), U_{UL,i}(A)\right),(3)$$

where $U_{DL,i}(A)$ and $U_{UL,i}(A)$ represent the resource utilizations of downlink and uplink subframes with a given association matrix A, respectively. The association matrix A illustrates serving base station of each MSs and its elements can gain two values 0,1. Value 1 for entry (i, j)means MS $j$is connected to BS i and value0 indicates that MS$j$is not connected to the BSi. The set of possible $a_{i,j} = 1$ values is limited to the BSs covering the overlapped area where the MS resides.

H. Velayos et al.for the first time, have proposed formulas 1 and 2 as a load balance index in wireless LAN [11], while T. Casey et al. have developed this algorithm and applied it in mobile WiMAX networks to distinguish load status of the network. According to the gained results in [4], it can be concluded that, using these formulas helps to specify load state of the system more accurately.

2.3 Load Balancing Mechanisms

Load balancing schemes that aim to solve the problem in overloaded cells can generally be classified into resource allocation and load distribution schemes [12], which contain different methods illustrated in the Figure 1. The idea behind balancing the system load with resource allocation is to bringthe resources (unoccupied frequencies) to where most of the users are located[13], this category includes fixed and dynamic channel allocation schemes. In load balancing based on the load distribution the goal is to direct the trafficto where the resources are. This category also includes two main methods; MS and BS initiated handover.These subtitles are further described.

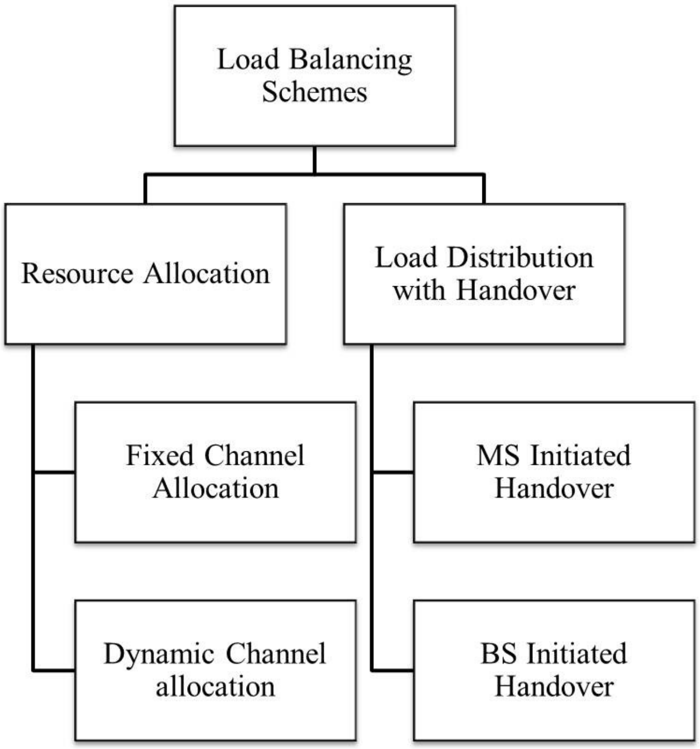

Figure 1. Classification of load balancing mechanisms

2.3.1 Resource Allocation Schemes

The scheme of load balancing of the system by using resource allocation is based on carrying the unused resources to the area where most of users are placed. In this approach, a centralized component assigns additional or free resources to overloaded cells. Resource allocation can be categorized into two main classes, *fixed channelallocation* (FCA) and *dynamic channel allocation* (DCA).

In an FCA class, a fixed number of channels are allocated to each base station. Though, this scheme does not use the channel sufficiently because of the variability of the traffic. DCA as an enhancement to the FCA can adapt itself with changes in traffic and adjusts frequency assignments relevant to the traffic load.

Most of the major researches in this domain have proposed *channel borrowing algorithms* (CBR) which states in the second category, dynamic channel allocation. The main principle of CB algorithms is using remained resources of cells with lower rate of traffic [14]. Although Mobile WiMAX provides a flexible way to allocate frequency resources in DCA manner between BS, it won’t be applied at least in the early stages of WiMAX deployment. However, FCA was preferred based on its simple mechanism.

2.3.2 Load Distribution Schemes

In order to balance the load of the system through the load distribution scheme the offered traffic should be directed to where more resources are available. The main way of operating load distribution is to use handover-based algorithms. Handover process would be conducted in two ways: MS and BS initiated handovers.

The first method for load distribution is *load balancing based MS initiated handover.* In this process, the load balancing logic locates in MSs, hence MSs in the overloaded cells configures the load situation and chooses the least congested access points [15]. This approach has been implemented in WLAN terminals and it can be also used in mobile WiMAX networks, based on the available resource information broadcasted in the *MOB-NBR_ADV* message.

The second and the most significant load distribution method that should be considered is *load distribution based BS initiated handover.* In this scheme, the load balancing algorithm resides in BSs and the congested SBS forces the MS to handover to a less congested TBS. This method is suitable for mobile WiMAX, as it enables stronger control for BSs also providing better QoS in the whole network [4], [16], [12], [17] .

3. HANDOVER IN MOBILE WIMAX

As the main way of load balancing is distributing load of the network with the handover process, overviewing handover process is necessary. One of the most important features of mobile WiMAX through mobility domain is handover. Handover means changing the serving BS by the MS if another BS is available to achieve higher throughput in the network. Furthermore, handovers are the essential component of system wide resource utilization and QoS. In 802.16e, handoff process may be initiated based on two reasons. The first one is caused by fading of the signal and interference level within the current cell or sector, while the second one is based on the fact that another cell or BS can offer a higher level of QoS for the MS, so MS will transfer to a neighbor cell or BS [18].

3.1 Types of Handover in Mobile WiMAX

This paper has classified handover types from four points of view which is a novel classification of handover in wireless and WiMAX networks. Types of handover can be categorized as*Technology*, *Structural*, *Initiation*, and *Execution mechanisms*aspects.From the technological point of view, it will be clarified that the handover algorithm will be executed between two different technologies or among same technologies. In the structural aspect, we will investigate handover is triggered among different BSs or among different channels in one BS. Initiation aspect shows who is responsible for directing handover, MS or BS and execution aspect clarifies type of handover based on handover types which were defined in the Standard.

- ***Technological Aspect:*** Handovers are also achievable among different technologies. It means that, handover can be occurred between Mobile WiMAX and other wireless technologies, or vice versa. It can be categorized into two classes, *horizontal* and *vertical handovers*. The former occurs when handover is within a single technology and the latter is the handover within different technologies such as wireless LAN and mobile WiMAX[19], [20],[21].

- ***Structural Aspect:*** Handover does not always mean changing the BS, it can also happen within the same BS but among various channels in the same BS, which names intra-cell handover [22] while in inter-cell handover MSs transfer from one BS to another BS [16].

- ***Initiation Aspect:*** The deterioration of radio signal causing in handing over the MS connection to a suitable target BS (TBS). In this case, MS is responsible for conducting handover, so it is named as MS initiated handover.In the case of handover trigger due to an unbalanced distribution of traffic, usually BSs detect the unbalance situation in the system and BS initiates handover. So, it can be called as *BS InitiatedHandover*[13].

- ***Execution Mechanisms Aspect:*** Mobile WiMAX provides three types of handover from the execution aspect as follows. Hard handover (HHO) is mandatory however fast base station switching and macro diversity handover are optional[18].

  1. *Hard Handover (HHO):* Thisprocedure changes the BS using a “Brake-Before-Make” way, which means that the MS disconnects from the serving BS before connecting to the TBS and has no connection during two phases; Synchronization to the TBS and Network re-entry.

  2. *Fast Base Station Switching handover (FBSS):* In FBSS and MDHO, the MS is connected to one or more BSs during the handover execution. In other words during handover execution phases 3 and 4, which will be discussed in the following section, are repeated for many BSs. Thus, it will become quite heavy procedure.

     In this mechanism the MS maintains a list of active BSs that it has established a connection to them. The MS is able to receive and transmit every frame from any BS within this ”Active set” and therefore no handover interruption should occur[23]**.** An important point that arises here is that in the FBSS all the BSs in the diversity set receive the data addressed to the MS, but only one of them transmits the data over the air interface whereas the others eventually drop the received packets.

  3. *Macro Diversity Handover (MDHO):* In this type of handover, the MS saves a list of BSs proficient to the MDHO on its coverage area. This group is known as diversity set. One BS always exists in the diversity set that is expressed as an Anchor BS. The procedure of MDHO is begun by the MS when it decides to receive or transmit from multiple BSs simultaneously. With a MDHO handover the MS is able to communicate with all of the BSs enabling diversity combining to be used to reach the optimum signal quality in both downlink and uplink [24].

The rest of this section discusses about basic Handover procedure proposed by IEEE 802.16e. This type of handover initiates by MS. During a handover procedure the following steps will be observed: (1) cell reselection, (2) HO decision and initiation, (3) HO cancellation, (4) synchronization to target BS down-link, (5) use of scanning and association results, (6) ranging, (7) termination with the serving BS, (8) dropping during HO, and (9) network entry/re-entry [3]. These steps can be summarized into four main phases which will be explained in more details:

1. Cell re-selection,
2. Handover decision and initiation (followed by resource reservation and admission control by the Target BS)
3. Synchronization to the Target BS downlink (interruption in the connection)
4. Network re-entry (including ranging and interruption in the connection).

1. *Cell Reselection*

This phase denotes the process of an MS scanning or associating with one or more BSs to determine their availability and efficiency as an HO target. So, MS should extract a set of potential TBSs that it can handover to. The first set of these TBSs will be gained from a message names *network topology advertisement message* (*MOB NBR-ADV*) which is illustrated in the Figure 3.

The actual cell re-selection is initiated when the MS starts to scan potential TBSs in order to find out if they can provide sufficient signal strength and quality. The MS will apply for permission of scanning from the Serving BS through a *MOB SCN-REQ* message. The SBS responds with a *MOB SCN-RSP* message specifying a period of time when it will buffer the traffic sent to the MS, permittingthe MS to scan the TBS. Hence, such scanning can threaten the QoS of delay sensitive service flows[25].

2. *Handover decision and initiation*

Based on the MS initiated handover, while the signal strength and quality received by the MS weakened and passed a pre-defined threshold, the MS will trigger the scanning procedure. In this phase, MS will decide to handover to a suitable TBS based on the achieved measurement of signal strength and quality of the TBS.

There are several algorithms that could be used for decision making in order to initiate the handover, and the most famous one is relative hysteresis algorithmwhich reduces the ”Ping-Pong” effect [26], [27]. Ping-Pong effect occurs due to quick reaction to rapidly changing channel conditions. If the condition for a handover is satisfied, the MS will trigger the handover by sending a *MOB MSHO-REQ* message. In this message the MS will include a list of TBSs it would prefer to handover to.

If an MS initiated network controlled handover is conducted, the SBS will first send a message through the backbone network to the TBS investigating if they have sufficient amount of resources to admit the MS. The TBSs will respond by signifying the type of QoS they can offer.

The set of TBSs that have adequate resources will be transmitted to the MS in a *MOB BSHO-RSP* message. The MS will indicate to the TBS that it has selected to handover to through a *MOB MSHO-IND*. The MS can cancel the handover, in case that the TBS signal quality has fallen, by asserting the BS ID of the SBS in the *MOB MSHO-IND* message.

3. Synchronization to Target BS Downlink

Since in the Hard Handover execution the MS connection will be disrupted, the main goal of phases 3 and 4 in terms of QoS is to perform handover quickly and reliably. In this phase, on the way to connect with the target BS, the MS synchronizes to a downlink transmissions of the target BS and acquires downlink (DL) and uplink (UL) transmission parameters. If the MS had previously received an *MOB_NBR-ADV* message containing target *base station ID*(BSID), physical frequency, *uplink* and *downlink channel descriptor*(DCDand UCD). Otherwise, the MS synchronizes with the target BS by scanning the potential channels of DL frequency band till it finds a valid DL signal[8].

4. Network Re-entry

The network re-entry is executed in a similar way as the initial network entry. The process of re-entry through handoff can be improved and done in advanced. The reauthorization of MS and key exchange is performed and the MS registers with the target BS, which is intended for agreement of some capabilities. Now, the MS has re-entered the network of the target BS and the service flows can be created again by continuing their normal operation. At last, the previous serving BS can be released. The target BS can acquire information from the serving BS through backbone connection, or even from other network entities.

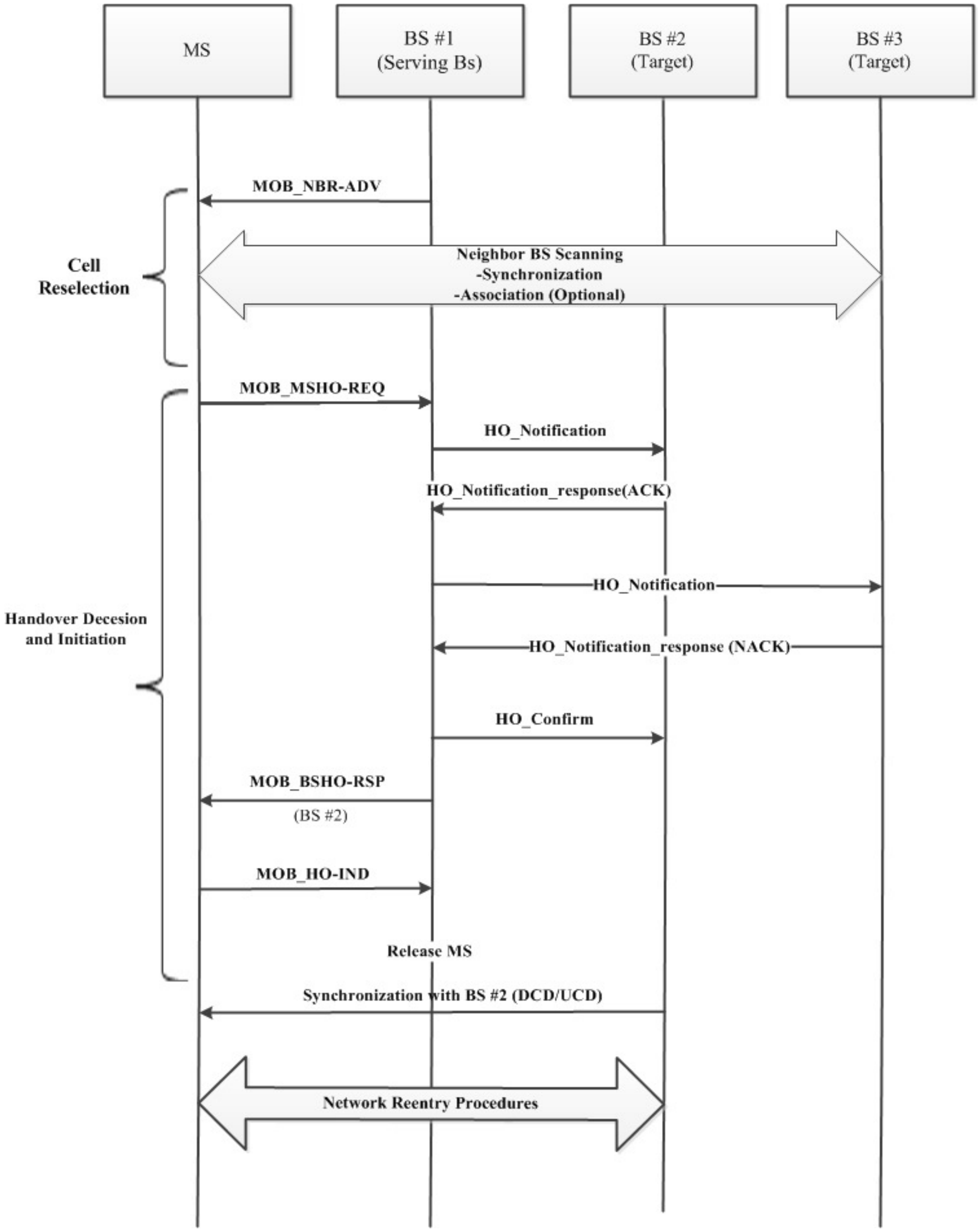

Figure 2. Handover Operation [3]

4. Load Balancing Algorithms in Wireless Networks

In cellular systems, hot-spot cells problem arises when available wireless resources at some location are not sufficient to support the users' requirements. The hot-spot cell can potentially cause blocked and dropped calls and can lead also to degradation of the service quality. Thus, several algorithms were proposed in this domain to handle the overloading cells.

4.1 Resource Allocation Scheme

The Mobile Assisted Connection Admission (MACA) algorithm [28] is a channel allocation algorithm based on channel borrowing scheme to balance the load of cellular networks. In this scheme, some special channels are utilized to straighten associated mobile units from different cells. Thus, a mobile unit, which is in a hot-spot cell unable to connect to its current BS. However, it may be able to connect to its neighboring lightly-loaded BSs through a two-hop link.

As it was mentioned before, this scheme is similar to channel borrowing method, proposed in [29].Das et al. have proposed thatwhile the cell does not contain enough resources, it will borrow one or more channels from its neighbouring cells provisionally. As this algorithm is based on CB scheme, it can adjust itselfwith variable traffic.

4.2 Load Distribution Scheme

Adaptive Cell Sizing (ACS) scheme in is an illustrative algorithm, which manages the transmitting power of BS in CDMA cellular system[30]. In this paper handover based algorithm which use functionalities of handover mechanism were recommended in order to distribute load of hot-spot cells.

In [11], a load-balancing scheme for overlapping wireless LAN cells is presented. Load balancing logic is implemented in agents running in each access point. Agents broadcast the local load level periodically via the Ethernet backbone and determine whether the access point is overloaded, balanced or under-loaded by comparing it with received reports. The access point's throughput is used as a load metric. Overloaded access points force the handoff of some stations to balance the load. Only under-loaded access points accept roaming stations to minimize the number of handovers and avoiding "Pig-Pong" effect. This algorithm is an inter-cell, horizontal, and BS initiated handover algorithm from structural, technological, and initiation points of views, respectively.

In [12], 4G mobile networks which are expected to support various multimedia services over IP network were considered. While 4G mobile networks use OFDM as the PHY layer technology, it can support hard handoff mechanism from execution point of view. So, a handover-based algorithm was proposed in this paper. This algorithm aimsto distribute the load of the system in order to manage the hot-spot cells. Most of load balancing algorithms concentrate on balancing the traffic load of the hotspot cell without deliberating the load status of the neighbouring cells while this paper has considered this issue.

This algorithm initiates the hard handover process, considering the load state of the target cell in addition to the load situation of the serving cell, and users received signal strength of the user. Based on this consideration, handover initiation time changes dynamically. The handover will be started earlier than scheduling time, if the received signal strength is less than an offered threshold and the target cell is in an acceptable load state. However, the algorithm does not conduct the handovers earlier when the target cell is overloaded; instead it uses conventional handover algorithms.

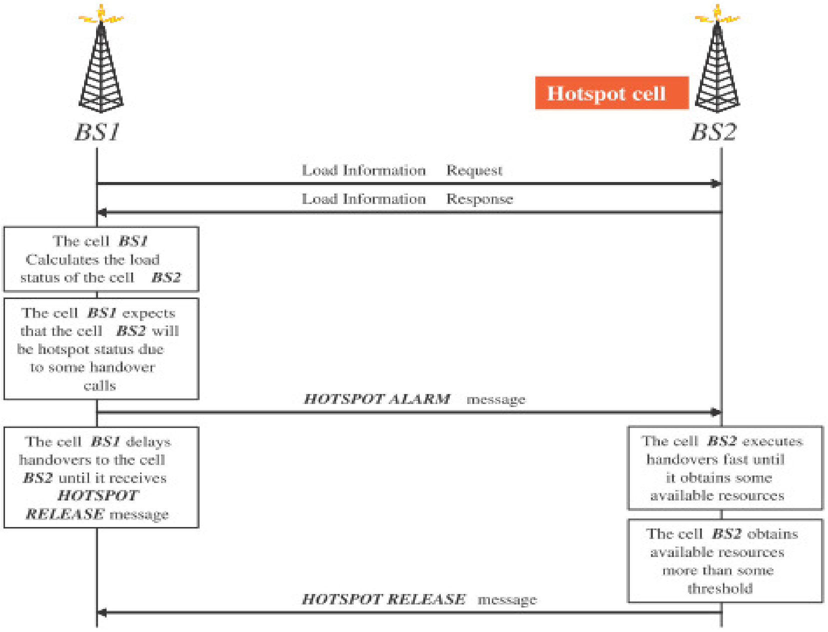
BS1
Hotspot cell
BS2
Load Information Request
Load Information Response
The cell BS1 Calculates the load status of the cell BS2
The cell BS1 expects that the cell BS2 will be hotspot status due to some handover calls
HOTSPOT ALARM message
The cell BS1 delays handovers to the cell BS2 until it receives HOTSPOT RELEASE message
The cell BS2 executes handovers fast until it obtains some available resources
The cell BS2 obtains available resources more than some threshold
HOTSPOT RELEASE message

Figure 3. Process of adaptive handover time scheme[14]

Figure 3 shows the process of adaptive handover time scheme. When the current serving cell, BS1 receives the report from a mobile station which includes information about MS signal strength and signal strength is below the specific threshold, it means the MS Initiated handover is necessary. Though, BS1 demands the load information of the target cell, BS2. If BS1 clarifies that BS2 will become the status of hotspot due to increased load caused by handover calls, BS1 directs the hotspot alarm message, *HOT-SPOT ALARM* to BS2.Then BS1 postpones all handovers caused by users who belong to the overlapping area of BS1 and BS2. BS2 identifies that its status will be hotspot due to additional load affected by handover calls. Then, BS2 decreases its load by executing all potential handovers earlier than what is scheduled in conventional handover mechanism.

During the fast handover execution, BS2 can achieve available resources and prepare to handover calls happened in the near future. If BS2 reaches to the light loaded status due to fast handover execution and acquires the sufficient number of available resources, it transmits *HOT-SPOT RELEASE* message to BS1. This algorithm proposed a horizontal, inter-cell, and BS-initiated handover.

5. Load Balancing Algorithms in WiMAX Networks

As traffic and the MCS in traditional networks is homogeneous and fixed, load balancing with resource allocation methods, such as channel borrowing is simple to manage. However, in mobile WiMAX, fluctuating traffic characteristics, MCS variations and distributed access network structure make such resource allocation very challenging. So far, no load balancing algorithm based on resource allocation was proposed.

Based on the reasons mentioned, this section will take a look at the research conducted for load balancing with load distribution in mobile WiMAX networks. As a result of the previous researches it can be concluded that load distribution will most probably be the way load balancing is performed at least in the early stages of mobile WiMAX. Therefore, the aim in the case of congestion is to direct connections to where free resources exist.

5.1 Load Distribution Algorithms

In time of revising related work, there have been a few papers discussing load balancing with handovers in IEEE 802.16e based systems such as algorithms in [16]and [31], up to now. As far as the authors are aware, the load balancing issue was only investigated from the viewpoint of the Mobile WiMAX system, its architecture and functions [4], [17].

In [16] an inter-frequency reuse handover among *frequency assignments* (FA) is suggested to solve unbalanced load distribution in 802.16e systems. In these systems multiple FAs can be assigned to each BS, it would be possible to transmit MSs from overloaded FA to another in order to balance the load distribution among FAs.As this handover occurs among Fas in the same BS, it is an intra-cell handover algorithm from structural point of view. This algorithm has used *MOB_NBR-ADV* message to indicate target FA instead of using De/Re-register command. The MS in overloaded FA can gain information about neighbouring BSs through this message and distinguish if its current BS obtains multiple FAs or not. Afterward the MS selects the most suitable and available FA based on its channel and load condition.

In the basic handover process, MSs in overloading FAs or BSs need to scan neighbouring BSs and gain *received signal strength indicator*(RSSI) and*signal to interference plus noise ratio*(SINR) information of neighbouring FAs and BSs, to find the appropriate target BS or FA for handover. However, as this kind of handover is in the same BS, MS just requires the load and channel condition of FAs in its serving BS. Due to this reason, scanning process is not necessary during inter-FA handover process.In this scheme load balancing is not controlled by the network, it is thus not that appropriate in terms of Mobile WiMAX network architecture. Also, as the load definition takes the percentage of dropping packet into consideration it is not a reasonable criterion for mobile WiMAX networks. This is a horizontal, intra-cell handover algorithm from technological and structural points of view.

In [22] an FA load balancing scheme is introduced in WiBRo systems. WiBRo is defined as a subset of the IEEE 802.16 standard and is developed based on the mobile WiMAX. The proposed algorithm uses scheduling and dual handover procedures where MS and BS initiated handovers are used adaptively. MS initiated handover is used to assign MSs to suitable FAs and uses an admission control algorithm to check availability of resources.BS initiated handover is performed when average throughput of non-real time service flows is above threshold. Through BS initiated handover procedure, BS decides to transfer MSs to other FAs cause of lack of resources. In the first step of this procedure, a BS compares resource consumption in other FAs, and then selects the one with the lowest resourceconsumption as a target FA. Finally, it sends *MOB_BSHO-REQ* message to a MS in order to inform it about transferring to other FA.The goal of this algorithm was decreasing rate of call dropping by restricting resource utilization in non-real time MSs. However, since this scheme has only considered throughput as the load parameter, it cannot be a proper method to satisfy QoS of service flows. This algorithm is designed based on horizontal and intra-cell approaches in technological and structural aspects, respectively.

In [4] an algorithm was proposed for Mobile WiMAX networks which can balance the load of the system with directed handovers. In this algorithm the load balancing logic resides in BSs, therefore BSs control the load situation of the network. In this method, BSs force MSs residing in overlapping areas to switch their connection from overloaded BSs to underloaded ones. The algorithm has used a feature specified in the WiMAX forum network architecture which is spare capacity report.As BSs control load of the system, they should communicate with each other and retrieve information about the resource utilization of their neighbour BSs. Therefore, this algorithm implemented *spare capacity reporting* (SCR) procedure including *Spare_Capacity_Req* and *Spare_Capacity_Rep* messages which are being sent between BSs in one cluster.

The *load balancing cycle* (LBC) has been defined as the load evaluation of the network and directed handover initiation are performing through these cycles. LBC starts by receiving SCR from peer BSs, and based on achieved information the average load of the system can be computed and compared with the load of each BSs. The scheme has defined three possible loading statuses: *underloaded, balanced, overloaded* BSs. The BS is underloaded and balanced if its load is below or equivalent to the average load of the system, respectively. The BS is overloaded if its load is exceeds a threshold which is calculated by considering the hysteresis margin.After identification of load state of the BS, overloaded BSs start initiating directed handovers of MSs in overlapping areas. Rescue handovers can be accepted in by BSs in all of states, where directed handover will be accepted only in underloaded BSs. The LBC will be finished by handing over the last MS or reaching to the balanced state. This algorithm can prevent “Ping-Pong” effect and utilizes resources efficiently. This algorithm is an horizontal handover from technological point of view, and has used inter-cell method in structural aspect.

A novel handover algorithm to balance the load of the layers in a multi-reuse scenario in Mobile WiMAX systems is proposed in[17]. This algorithm aims to balance the resources between layers by moving MSs from one layer to another one based on their QoS requirements and channel condition. They have divided bandwidth among three layers in one BS, and each layer owns $B/3$ of the total bandwidth. This paper has considered both rescue handover and directed handover in terms of *Default* and *Load Balance* handover, respectively. The algorithm checks possibility of Default handover based on the RSS and SINR criteria in the first step and performs it where it is necessary, otherwise it will check the possibility of Load Balance handover. The process of this step will start by calculating the load of the layer which can be the layer average throughput, number of occupied slots, and the number of MSs in per Layer. The layer with the most fluctuation of the load will be selected as an unbalance layer.

The second step is generating a list of unbalanced pairs based on the gained information of layers. The next step is ranking of MSs in unbalance layer using cost function. Then, the MS with the highest cost function value will be handed over from inner layer to the outer layer. This process will last till the last MS handed over to the target BS. Finally, the layers that took part of the initial evaluation are removed from all other layer pairs that they take part. This algorithm could manage load balancing situation of the system and keeping a satisfactory QoS for the VoIP calls, and also could avoid “Ping-Pong” effect. As the proposed algorithm checks rescue handover and then checks load situation of the network. However, the system should tolerate overload status more than reasonable period, as the algorithm should check the probability of the rescue handover at first and then performs load balancing algorithm. Due to this reason, the delay of directed handover will be increased. This is an horizontal, intra-cell handover algorithm. In this algorithm both MS and BS initiated handovers are performed, but load balance handover is a BS initiated handover.

Another interesting load balancing algorithm was presented in [31] for IEEE 802.16e based networks, to distribute load of the system based on directed handover. The scheme attempts to find the optimal MS-BS association set to balance the utilization of common resources in the whole system. The algorithm goes through every possible association combination, trying to minimize the maximum resource utilization of slot and power resources in each BS. It has the ability to distribute the load very electively, but comes with some problems.

The algorithm is complex in the case that it does not only balance system load but also tries to decrease the load which will increase the number of directed handovers and hence might threaten QoS fulfillment and causing “Ping-Pong” effect. The issue of when directed handovers should be directed regards to fluctuating traffic and how much unbalance should be allowed is not considered in this algorithm. Furthermore, implementing this algorithm in a distributed manner in the BSs, causes in a high amount of signalling overhead. This algorithm would be more suitable in a centralized approach, but it decreases scalability. Many of the ideas presented in the paper, could be used for load balancing in the later stages of Mobile WiMAX deployment but for now, it seems too complex.

In [32]a new distributed uplink packet scheduling algorithm in WiMAX networks is proposed. This algorithm estimates available resources for each connection in the uplink and aims to provide required resources for connections according to the characteristics and QoS requirements of each connection. In the case that the uplink cannot fulfill the required resource of connections, it means that it is overloaded and handover should be initiated. Though one or some user terminals in the overlapping cells will be selected and transferred to the neighbouring under-loaded cells.However, for the first time, from point of view of the uplink packet scheduling, the proposed algorithm in [32] has used a handover priority function that chooses the best user terminal from user terminals in the overlapping areas according to their scheduling services and also based on their received bandwidth requests. It can be concluded that, this algorithm used a load balancing with handover scheme in order to schedule packets. The simulation results show that this packet scheduling algorithm can increase the overall throughput of the network.

As it was mentioned before, WiMAX has a large coverage area but limited data rate for each end user. However WLAN uses free frequency and in the case that the network is not congested it can provide higher data rate for end users. As a result, WiMAX and WLAN networks are complementary in terms of service characteristics. Due to these reasons, authors in [20] have considered the WiMAX/WLAN heterogeneous network and studied load balancing for multi-service load balancing in an overlay heterogeneous network. They have examined system performance of an overlay heterogeneous wireless network where elastic applications share network capacity with prioritized streaming applications. In this paper it was assumed that all the traffic first arrives to the WiMAX network and streaming applications are given strict pre-emptive priority over elastic applications in WiMAX.

As one possible way for interaction between WiMAX and WLAN is to distribute the load among both of them by considering service requirements of applications and network service characteristics, a load balancing model has been proposed in [20]. They have modelled their system in the way that all streaming applications are served in WiMAX while elastic applications are served in the remained capacity of WiMAX and in the whole capacity of WLAN. A real time load balancing algorithm is proposed in this paper begins while the system state changes. The proposed algorithm tried to improve the performance of elastic applications by performing vertical handoff on them. Through this handover mechanism elastic applications will be dispatched to the WLAN and are served there.

This algorithm includes two phases, *handover check* and *handover candidate selection*. Handover check will be conducted based on predicting ending time of elastic applications. In the second phase, all new arrival and also old elastic applications can be selected based on their remaining size. Two cases are considered for conducting vertical handover; in the first case, elastic application which is connected to WiMAX will be dispatched to the WLAN if its performance is not acceptable, the other method is the direction of vertical handover upon arriving an elastic application to the WiMAX network, in this case the application will be dispatched to WLAN.

In [33] a QoS aware load balancing algorithm is proposed. Load has been evaluated based on characteristics of traffic classes. In the overload detection phase, while *load balancing cycle* (LBC) starts, each BS compares elements in its matrix with the corresponding element in the threshold matrix. Mobile WiMAX introduces five classes of QoS called UGS, rtPS, ertPS, nrtPS, and BE which forms columns. Furthermore, rows consist of the most effective performance metrics so-called Downlink Throughput (Th_{DL}), Uplink Throughput (Th_{UL}), Delay and Jitter.

Two more matrices are generated according to *Thresholds* and *Hysteresis Margins* while the number of columns and rows of these matrices are equal to the previous ones'. Matrices of MSs and BSs will be updated periodically and they will be compared with the *Threshold* matrix at the end of each period. While an element of BS matrix exceeds the corresponding element of threshold matrices, it will be concluded overload situation will take place. In this case, overloaded BS should initiate handover by considering the class of QoS which caused overload state. Handover mostly will occur because of non-real time connections; otherwise the hysteresis margin should be taken into account. When real time connections exceed both threshold and the attachment of threshold and hysteresis margin matrices, handover is necessary for real time service flows, too.

This algorithm is designed in such a way that it can have best treatment based on the number of MSs in overlapping areas. If the number of MSs is less than a pre-defined value an optimized method will be performed, otherwise a less complicated handover will be done. In the optimized handover method all of the situations that MSs can be switched to TBS have been considered. The matrix of selected MSs should be attached to the matrix of TBS besides detach from the matrix of serving base station (SBS). Variance of these computations should be calculated separately.

As it was mentioned in the previous chapters, Mobile WiMAX networks include fluctuating traffic characteristics and distributed access network architecture, so the best method to balance the load of the system is load distribution with handover. Since mobile WiMAX specifies five classes of QoS and performance parameters provided by each service classes is different, the load should be defined based on these QoS classes and associated performance parameters. Any other load definition would not be efficient enough. So, through load balancing algorithms load definition should be taken into account in addition to the mechanism of handover.

6. CONCLUSION

In this survey, for the first time, we have provided a comprehensive survey of load balancing algorithms in wireless networks. To achieve this goal, a number of concepts were explained. In general, load balancing can be conducted in a static or dynamic manner. Static load balancing is independent of the state of the system where as in dynamic load balancing, decisions are made based on the current loading situation and availability of resources. Load balancing can also be done in a distributed or centralized way. Based on mobile WiMAX network's specifications, such as fluctuating traffic characteristics, variation of MCS and distributed access network structure, load distribution with BS initiated handover scheme will be the best way to balance load of the system. Based on background research, it can be concluded that BS initiated handover scheme will be the most appropriate way to balance load of mobile WiMAX networks. Also, as the load definition plays an important role in the load balancing algorithms it should be defined in the more accurate and inclusive manner.

REFERENCES

[1] "IEEE Xplore - IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems," 2004. [Online].Available:http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1350465.[Accessed: 09-Dec-2011].

[2] IEEE Computer Society.;IEEE Microwave Theory and Techniques Society.;IEEE Standards Board.;Institute of Electrical and Electronics Engineers., *IEEE standard for local and metropolitan area networks. amendment 1□: management information base*. New York□:: Institute of Electrical and Electronics Engineers,, 2005.

[3] S. Ahson, *WiMAX standards and security*. Boca Raton Fla.: CRC Press, 2008.

[4] T. Casey, N. Veselinovic, and R. Jantti, "Base Station Controlled Load Balancing with Handovers in Mobile WiMAX," 2008, pp. 1–5.

[5] K. Wu, "Load balancing of elastic data streams in cellular networks - Google Scholar," Helsinki University of Technology.

[6] T.-C. Tsai and C.-F. Lien, "IEEE 802.11 hot spot load balance and QoS-maintained seamless roaming," *tsai2003ieee*, 2003.

[7] A. J.Nicholson, Y. Chawathe, M. Y.Chen, B. D. Noble, and david Wetherall, "Improved Access Point Selection," *MobiSys '06 Proceedings of the 4th international conference on Mobile systems, applications and services*, 2006, pp. 233–245.

[8] "WiMAX Forum Network Architecture (Stage 2: Architecture Tenets, Reference Model and Reference Points- Release 1.1.0)." 03-Feb-2009.

[9] J. bum Ryou, "Adaptive Load Balancing Metric for WLANs," PHD Thesis, Oregon State University, 2011.

[10] D.-M. Chiu and R. Jain, "Analysis of the increase and decrease algorithms for congestion avoidance in computer networks," *Computer Networks and ISDN Systems*, 1989, vol. 17, pp. 1–14.

[11] H. Velayos, V. Aleo, and G. Karlsson, "Load balancing in overlapping wireless LAN cells," 2004, vol. 7, pp. 3833–3836.

[12] D. Kim, N. Kim, and H. Yoon, "Adaptive handoff algorithms for dynamic traffic load distribution in 4G mobile networks," 2005, pp. 1269–1274.

[13] T. Casey, "Base Station Controlled Load Balancing withHandovers in Mobile WiMAX," Master's Thesis, HELSINKI UNIVERSITY OF TECHNOLOGY, 2008.

[14] S. K. Das, S. K. Sen, and R. Jayaram, "A dynamic load balancing strategy for channel assignment using selective borrowing in cellular mobile environment," *Wirel. Netw.*, 1997, vol. 3, pp. 333–347.

[15] D. Kim, M. Sawhney, and H. Yoon, "An effective traffic management scheme using adaptive handover time in next-generation cellular networks,"*Int. J. Netw. Manag.*, 2007, vol. 17, pp. 139–154.

[16] S. H. Lee and Y. Han, "A Novel Inter-FA Handover Scheme for Load Balancing in IEEE 802.16e System," *Vehicular Technology Conference. VTC2007-Spring. IEEE 65th*, 2007, pp. 763–767.

[17] J. J. Bazzo, A. M. Cavalcante, M. J. de Sousa, L. Kuru, and J. Moilanen, "Load Balance for Multi-Layer Reuse Scenarios on Mobile WiMAX System," 2010, pp. 1–5.

[18] A. Mäkeläinen, "Analysis of Handoff Performance in Mobile WiMAX Networks," Master's Thesis, Helsinki University of Technology, 2007.

[19] Y. Liu and C. Zhou, "A Vertical Handoff Decision Algorithm (VHDA) and a Call Admission Control (CAC) policy in integrated network between WiMax and UMTS," 2007, pp. 1063–1068.

[20] J. Xu, Y. Jiang, and A. Perkis, "Multi-service load balancing in a heterogeneous network," 2011, pp. 1–6.

[21] S. R. Poorahmadi and H. Beigy, "IMPROVING HANDOVER LATENCY BY USING CROSS-LAYER DIRECT COMMUNICATION MODEL IN IEEE 802.16E BROADBAND WIRELESS ACCESS NETWORKS," 2010, pp. 133–137.

[22] K. P. Moon, J. Park, J. Kim, H. Choo, and M. Y. Chung, "Dual Handover Procedures for FA Load Balancing in WiBro Systems," 2008, pp. 54–59.

[23] Z. Becvar and J. Zelenka, "Handovers in the Mobile WiMAX," *Research in Telecommunication technology*, 2006, vol. 1, pp. 147–150.

[24] S. K. Ray, K. Pawlikowski, and H. Sirisena, "Handover in Mobile WiMAX Networks: The State of Art and Research Issues," *IEEE Communications Surveys & Tutorials*, 2010, vol. 12, pp. 376–399.

[25] Z. Yan, L. Huang, and C.-C. J. Kuo, "Seamless high-velocity handover support in mobile WiMAX networks," 2008, pp. 1680–1684.

[26] J.-H. Yeh, J.-C. Chen, and P. Agrawal, "Fast Intra-Network and Cross-Layer Handover (FINCH) for WiMAX and Mobile Internet," *IEEE Transactions on Mobile Computing*, 2009, vol. 8, pp. 558–574.

[27] H.-M. Sun, S.-Y. Chang, Y.-H. Lin, and S.-Y. Chiou, “Efficient Authentication Schemes for Handover in Mobile WiMAX,” 2008, pp. 235–240.

[28] X. WU, B. MUKHERJEE, S.-H. Gary CHAN, and B. BHARGAVA, “Assuring Communications by Balancing Cell Load in Cellular Network,” 2003, vol. E86-B.

[29] S. K. Das, S. K. Sen, and R. Jayaram, “A Structured Channel Borrowing Scheme for Dynamic Load Balancing in Cellular Networks,” *International Conference on Distributed Computing Systems*, 1997, pp. 1216–1228.

[30] X. H. Chen, “Adaptive traffic-load shedding and its capacity gain in CDMA cellular systems,”*IEE Proceedings - Communications*, 1995, vol. 142, p. 186.

[31] S. Moiseev, S. Filin, M. Kondakov, A. Garmonov, A. Savinkov, Yun Sang Park, Do Hyon Yim, Jae Ho Lee, Seok Ho Cheon, and Ki Lae Han, “Load-Balancing QoS-Guaranteed Handover in the IEEE 802.16e OFDMA Network,” *IEEE Global Telecommunications Conference*, 2006, pp. 1–5.

[32] S. Nazari and H. Beigy, “A new distributed uplink packet scheduling algorithm in WiMAX newtorks,” *Future Computer and Communication (ICFCC), 2010 2nd International Conference on*, Wuhan, 2010, vol. 2, pp. V2–232–V2–236.

[33] C. Askarian Amiri, “Load Balancing with Handover in Mobile WiMAX Networks,” Sharif University of Technology-International Campus, Kish Island, Iran, 2012.