Regular Article

Efficient Load Balancing Algorithm over Heterogeneous Wireless Packet Networks

Quoc-Thinh Nguyen-Vuong[1], **Nazim Agoulmine**[2]

[1]France Telecom, Arcueil, France
[2]University of Evry Val d'Essonne, Evry, France

Correspondence: Quoc-Thinh Nguyen-Vuong, thinh.nguyen@polytechnique.org

Manuscript communication: received 25 January 2010, accepted 16 October 2010

Abstract– **The paper aims at improving the load balancing algorithm in wireless packet cellular networks and particularly in coordinated heterogeneous wireless packet networks. Our main contributions are two-fold. First, we introduce a new approach to compute the network load metric based on the radio link quality and scheduling information, which can be applied to any wireless packet system. This load metric hides the radio resources heterogeneity of different access technologies from the load balancer. Secondly, we propose a new practical load balancing algorithm which provides a more efficient way to manage the scarce radio resources. The proposed approach has been compared with different existing schemes and the results show the superiority of the proposed solution.**

Keywords– **Load balancing, heterogeneous networks, handover, radio resource management, utility.**

1 Introduction

Current trends in wireless network evolution indicate a desire to integrate different wireless access technologies to offer an *always best connected* environment for mobile users. Along with the rapid growth in demand for high data rate and high Quality of Service (QoS) multimedia communications as well as the scarcity of radio resources, an efficient Radio Resource Management (RRM) scheme is highly required. An operator can deploy different technologies or interwork with other technologies owned by other operators to enable the global roaming capability through a coordinated heterogeneous access network environment. An advanced Common RRM (CRRM) is a motivation for interworking among these networks, and also a challenge to overcome.

The interworking between different Radio Access Technologies (RAT) can be distinguished into open, loose and tight couplings [1, 2]. The stronger the coupling is, the more efficient the resources can be commonly utilized. In this work, we consider the tight coupling case where different access technologies are being deployed by a single operator or by cooperative operators. Available radio resources of coupled networks will be jointly managed. We hence adopt the CRRM architecture introduced by the 3rd Generation Partnership Project (3GPP) [3] and further used in [4–11]. CRRM is defined as a platform to gather information from the Base Stations (BS) of different RATs, and to control the resource allocation of all BSs to optimize the overall system performance.

Generally, the load balancing plays an important role in the CRRM. The load balancing algorithm consists of accepting or denying a new incomming user request and forcing users connected to a heavily loaded BS to hand over to a lightly loaded one. To do so, we need to define load thresholds for the admission control and the handover enforcement. The latter is mainly due to the load balancing and not to the user mobility.

Our contribution is to introduce a new approach to quantify the load in wireless packet networks and a novel load balancing algorithm. The remainder of the paper is organized as follows. After an overview of existing load balancing algorithms in Section 2, a new load metric and a new load balance index are proposed in Section 3. Section 4 focuses on our proposed load balancing algorithm. Section 5 is devoted to show the performance evaluation of our approach compared to other reference strategies. Conclusions are drawn in the last section.

2 Related Work

In the joint RRM research area, most of previous work mainly focused on identifying the functionalities of the CRRM architectural components, and designing the protocols for control exchanges between these components [3–5, 12]. Besides, the resource allocation scheme which aims at quantifying the amount of resources allocated to each user in such a way to maximize the operator's revenue or the user's satisfaction has also been increasingly studied [13–15]. However, the load balancing between different BSs and different RATs has not been sufficiently considered. Although the load balancing is much related to the resource allocation, they are two separable aspects. The load balancing can be considered on the one hand as an objective of the resource allocation scheme and on the other hand as a

1859-378X–2011-0107 © 2011 REV

constraint for the resource allocation optimization. In this work, we only focus on the load balancing issue.

An adaptive threshold for load balancing based handover enforcement initiation was introduced in [16]. Although this approach makes it possible to detect the need of initiating a handover, the suitable target access network is not addressed. Another solution for RRM algorithm based on fuzzy logic and reinforcement learning was presented in [7, 8]. However, the admission control is just a primary step in the load balancing process as it only deals with the incomming calls. Even if an efficient admission control algorithm [7, 8, 17] is used, overload situations might still occur, e.g., due to the mobility of high-rate packet data users or the inherent fluctuation of the transmission channel.

All the load balancing solutions have been based on a fundamental resource unit notion, called "*load*". The load metric represents the occupation ratio of a BS. The load of a cellular network is usually computed through the received power and the interference level [18] whereas the load of a Wireless Local Area Network (WLAN) is simply computed through the number of users connected to an access point [7, 8]. The load can be computed in different manners for different systems. As a result, the same load value for two different systems does not mean the same load situation. As such a comparison is the basis of any cross-system load balancing solution, having a same semantic of the load metric is mandatory. The existing load computation methods, which are based on the interference [18] or the throughput [19], do not allow the load variation anticipation prior to the situation where a user moves into/out of a cell. The estimation of future interference or throughput values is really challenging. Accordingly, we will not be able to make the right decision to achieve an efficient resource balancing.

3 Load Metric & Balancing Index

3.1 Load Metric Definition

Some High Speed Packet Access (HSPA) network operators have recently encountered a network saturation by so many Iphone users. Along with the increase of multimedia and data-intensive applications, the future fourth-generation networks will promisingly experience an extremely high load situation. In this paper, we present only the cross-system downlink load balancing. However, the solution is still valid for uplink load balancing. Traditionally, the load metric corresponding to the resource occupation ratio varies from 0 to 1. As the circuit-switched cellular network has been progressively migrated towards all-IP packet network, here we only consider the load balancing for wireless packet networks. In wireless packet networks, the channel access is dynamically assigned to mobile users by a scheduler running in the BS (see Figure 1). The scheduler decides which packets are transmitted to their corresponding destinations at an instant (depending on the required QoS of each user and radio link quality between the user and the BS). Contrary

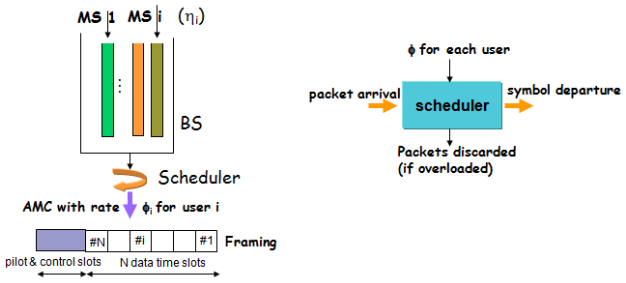

Figure 1. Scheduler in a base station.

to a fixed resource allocation in circuit networks, the resource allocation in packet networks is much more dynamic. An overload situation will cause a delay or packet loss to some specific connections, but not necessarily an outage of connections. It is thus interesting to be able to estimate the overload degree. The way to balance the load in packet networks is thus different from circuit networks.

The packet scheduling is an active research area. Generally, based on the transmission channel estimation, the BS will adapt the modulation and coding scheme to transmit packets in such a way to maximize the throughput and minimize the packet error rate. Recently, the QoS priority has been also taken into account in the packet scheduling [20]. Compared to the load balancing, a global strategy involving all the BSs in the system, the packet scheduling is just a local strategy at each individual BS. We can see that if the total requesting resource (i.e., packet arrival rate mapped with the modulation and coding rate) is higher than the capacity of the BS (i.e., symbol departure rate at the physical layer), some users will not get their required QoS. In other words, the BS is overloaded.

We define load ρ as the ratio of the required resources to the total resources. If the amount of the required resources of all users connected to a BS is greater than or equal to its total resources, this BS is considered as overloaded. In differentiated QoS wireless networks, the objective of the scheduler is to guarantee the QoS required by the non-best-effort users. Hence, the required resources information used for load computation is the guaranteed bit rate corresponding to the running application of each user. Alternatively, the required resources of a communication is its arrival rate at the BS. As a First-In-First-Out buffer is implemented at the BS for each connection, the packet arrival rate can be simply retrieved. In the following, for simplicity, each communication is assumed to have a guaranteed bit rate $\eta(Kbps)$.

At the physical layer, multiple transmission modes comprising of a pair of modulation scheme and Forward Error Control (FEC), as in IEEE802.11/16, 3GPP and 3GPP2 standards, are available to each user. Given the modulation and coding rate of $\phi(bits/symbol)$, the packet of N_p bit is mapped to a block of N_p/ϕ symbols after modulated and coded. Hence, the required resources of a call can be expressed as $\frac{\eta}{\phi}$ $(Ksymbol/s)$.

The total resources of a BS can be referred to as the

number of data symbols that the BS can transmit in downlink during one second, i.e., data symbol rate R_s. For example, in HSDPA system, the channel multiplexing is in time domain where each Transmission Time Interval (TTI) consisting of three slots (or $2ms$) can carry 480 data symbols. Within each TTI, a maximum of 15 parallel codes can be assigned to one user or shared between several ones. Hence, the total resources become $15 \times 480 symbols/(2ms) = 3.6 Msymbols/s$. In an OFDM system like Worldwide Interoperability for Microwave Access (WiMAX), or 3GPP Long-Term Evolution (LTE), the resources consist of OFDM symbols in the time domain and sub-carriers in the frequency domain. The downlink data symbol rate is equal to *(number of downlink OFDM symbols)×(number of data sub-carriers)/(frame duration)*. Meanwhile, in the direct-sequence CDMA system like Universal Mobile Telecommunications Systems (UMTS) or CDMA2000, the symbol rate depends on the spreading factor $S(chips/symbol)$ of the used code. As the chip rate $R_c = R_s \times S$ $(chip/s)$ is a fixed value, we choose it as the total resources parameter.

Now let M denote the number of currently connected users at a BS of total resource R_c. Each user i is characterized by a required guaranteed bit rate η_i, a modulation and coding rate ϕ_i and an associated spreading factor S_i. If speading factor does not exist, we set $S_i = 1$. The load of a BS is given as:

$$\rho = \frac{1}{R_c} \sum_{i=1}^{M} \frac{\eta_i S_i}{\phi_i}. \quad (1)$$

This load metric definition takes into account not only the user's required resource but also the radio link quality between the user and the BS. If the link quality is so poor to guarantee the connection or the user is outside the corresponding BS's radio coverage, the corresponding modulation and coding rate ϕ will be set to 0. If the BS accepts this user request, its load becomes infinity. Thus, the load balancing algorithm will refuse the connection and/or force the user to handover to another neighboring access network. Using this definition, the resources heterogeneity among different access systems will be hidden from the load balancing. In other words, the load balancing scheme is based only on the load values of different access nodes regardless of underlying technologies and underlying scheduling schemes. The load balancing over heterogeneous networks is somewhat similar to that over a homogeneous network.

3.2 Load Balancing Index

One of the key elements in the load balancing is the balance index used to measure the balance of resources in a system. Such an index was first introduced in [21] and recently used in [19]. It is defined as:

$$\xi_1 = \frac{(\sum_i \rho_i)^2}{K \sum_i \rho_i^2}, \quad (2)$$

where K is the number of neighbouring BSs over which the load can be distributed. In fact, ξ_1 is a correlation factor between the load vector $[\rho_1, \ldots, \rho_K]$ and the vector $[1, \ldots, 1]$. If all BSs have the same load level, then $\xi_1 = 1$. The load balancing target is to maximize ξ_1. However, this balance index exposes serious limitations. Consider a scenario where a new user at the overlapped zone of three BSs as depicted in Figure 2(a) wants to initiate a communication. Given that $\{\rho_A = 0.8, \rho_B = 0.4, \rho_C = 0.3\}$ are the current load of BS A, B and C and $\{\Delta\rho_A = 0.1, \Delta\rho_B = 0.2, \Delta\rho_C = 0.7\}$ are the added load if the new user attaches to BS A, B and C, respectively. By using objective function ξ_1, the new user will attach to BS C as it results in the highest balance index $\xi_1 = 0.89$. The BS C becomes overloaded ($\rho_C = 1$). And we can see that if the user connects to either BS B or BS C, this will not generate the overloaded situation.

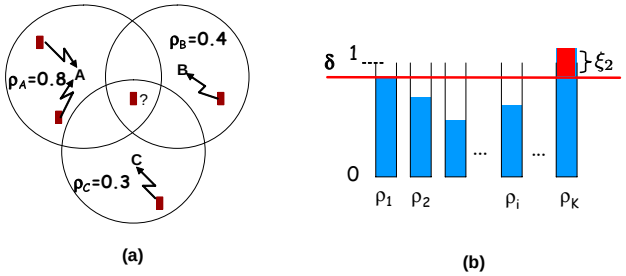

Figure 2. (a) Problem of using ξ_1; (b) Load balancing index ξ_2 computation.

As a user will generate different added loads when connecting to different access nodes, it becomes difficult to maintain all BSs at the same load value. Also, in a heavily loaded system, the balancing objective ξ_1 tries to evenly distribute the load to all BSs, which leads to a situation where all BSs will be overloaded. It may be better to degrade the QoS of only several users instead of all users. When the load between the BSs has not been balanced yet but all the BSs are not in the imminent overloaded situation, it is not necessary to maximize ξ_1 by forcing the users to attach to another BS. To resolve the overload situation in the exemplary scenario, one may suggest adding a constraint like $\rho_i < 1 \ \forall i$ while trying to maximize ξ_1 to accommodate the revealed limitation. It seems to be a good solution in a lightly loaded system. But, this constraint is never satisfied in a heavily loaded system. Accordingly, the objective of load balancing algorithm is to minimize the effect of overload situation and not to avoid the overload situation (because it is not always guaranteed in a finite capacity system).

In order to improve the above limitations, our load balancing mechanism objective is to avoid the overload if possible or to reduce overloading situation in access networks. The idea is to detect imminent overload situations and start to redistribute the load from heavily loaded access networks to lightly loaded ones. A system is considered as load-balanced if all BSs have a load below a specific threshold $0 < \delta < 1$. It is motivated by the avoidance of unnecessary load balancing operations that wastes the resource and causes undesired handovers. Usually, in the load control strategy, operators reserve an amount of resources $(1 - \delta)$, known also as *guard channel* ratio, for handing over users as well as

for system redundancy. The choice of threshold δ can be inspired by the research on guard channel optimization in [22] and we do not address such a choice in this work. Accordingly, we propose a new balance index ξ_2:

$$\xi_2 = \sum_{i=1}^{K} \max(\rho_i - \delta, 0). \quad (3)$$

If there exists $\rho_i > \delta$, then $\xi_2 > 0$. The greater index ξ_2, the closer to an overload state the network is. Note however that $\xi_2 > 0$ does not mean an overload situation (i.e., since ξ_2 may be greater than 0 but $\rho_i < 1$ for all i). The objective of the load balancing is now to minimize ξ_2. In the previous scenario, the overload situation does not occur while using ξ_2 as an objective function since $\xi_2(C)$ (that is the value of ξ_2 while network C is selected) is clearly greater than $\max\{\xi_2(A), \xi_2(B)\}$ for any chosen δ.

4 Load Balancing Algorithm

4.1 Optimal Algorithm Formulation

We provide here a formulation of an optimal load balancing algorithm. Assuming that at a given instant our system consists of M currently connected users and K BSs of different access technologies. Let us denote $\mathcal{W} = (w_{ij})$, $i = 1, \ldots, M$, $j = 1, \ldots, K$ as a generated load matrix where w_{ij} is the load generated at BS j while user i attaches to it. If user i is not in the radio coverage of BS j, then $w_{ij} = \infty$. The balancing algorithm will be triggered upon the imminent overload situation. Results of the algorithm should come out with an assignment $\sigma = (\sigma_{ij})$, where $\sigma_{ij} = 1$ if user i is decided to attach to BS j and $\sigma_{ij} = 0$ otherwise. The optimal assignment $\sigma^\star$ is given as

$$\sigma^\star = \arg\min_{\sigma} \sum_{j=1}^{K} \max(\rho_j - \delta, 0), \quad (4)$$

where $\rho_j = \sum_{i=1}^{M} w_{ij}\sigma_{ij}$, subject to the following conditions: $\sigma_{ij} = 0$ if $w_{ij} = \infty$ and only one element σ_{ij} in each row i of matrix σ is non-zero. We assume that if user MS i is in coverage of a particular BS then MS i will be allocated the resource (there exists i such that $\sigma_{ij} = 1$). In other words,

$$\exists j : w_{ij} \neq \infty \Rightarrow \sum_{j=1}^{K} \sigma_{ij} 1_{\{w_{ij} \neq \infty\}} = 1. \quad (5)$$

One may note that the constraint on binary integer variables σ_{ij} makes our optimization problem non-convex, and therefore far more difficult to solve. In the worst case where any user can connect to any BS, by using potentially exhaustive search, we need to compute the values of ξ_2 for K^M possibilities of σ to find out $\sigma^\star$. Such optimal algorithm is thus impractical for implementation since it requires an exponential computation time, especially in a large wireless network with thousands of users and BSs. Also, such an assignment may lead to a reallocation of resources for all users which implies a significant amount of handover and overheads.

4.2 Proposed Load Balancing Algorithm

Our aim is to design a feasible and suboptimal solution for load balancing while minimizing the resource rearrangement and the computation effort. When a user initiates a connection, the end-user device selects a suitable access network among available ones using the network selection mechanism. The load value of each access node may be used in the network selection evaluation if the user has access to this information. The user will be able to not select the heavily loaded access node. Besides, the access node may refuse the user's connection request based on its admission control policy if it is heavily loaded. Despite the use of an admission control, the overload of an access node still happens due to the transmission channel fluctuation, the mobility or the application data rate changes. To handle the load balancing, on-going calls will be transferred from an access network to another. The two main targets of our proposed algorithm are the admission control and the network-initiated handover.

4.2.1 Admission control: The admission control is employed to admit or reject a new originating or handing over communication in order to avoid overload situations. A connection request to a specific BS will be accepted if the BS's load, including the contribution of the incoming communication, is below an admission threshold δ_{AC}. Otherwise, the new incoming communication will be redirected to the least loaded overlapped access network. If all BSs in the coverage area could not accommodate the new communication, the connection request is rejected. If the incoming communication is a handing over one, the admission threshold is greater than the one used for a new originating communication. It is generally preferable to refuse the new calls rather than to drop the on-going calls. That explains also why we choose a load balancing threshold $\delta < 1$. In our solution, we choose to always accept the handing-over users.

It is noteworthy that a number of previous publications [7, 8, 17] have considered the admission control as a means to achieve load balancing. However, the admission control is just a first step in the load balancing process as it only deals with incoming communications and it does not treat the load fluctuation of on-going ones. Moreover, trying to redirect an originating communication to a less loaded access system (redirect from one technology to another) may not be possible if the communication is initiated from a single-mode terminal. In this case, it may be better to accommodate the originating single-mode user and to force a multi-mode user to make a vertical handover to a coordinated access system. That motivates the need to use handover enforcement to effectively distribute the load over the heterogeneous systems.

4.2.2 Handover enforcement: In addition to the admission control, it is essential to have a mechanism to detect and handle imminent overload situations. Such mechanism is known as the handover enforcement since its main role is to select *suitable* users in a heavily loaded access network and force them to handover

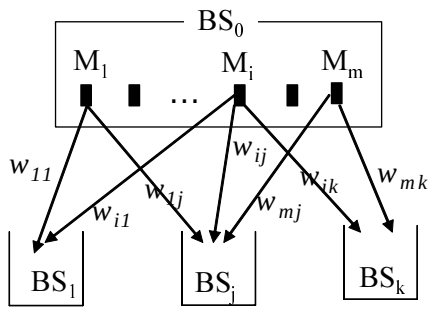

Figure 3. Illustration of load balancing algorithm.

to *suitable* lightly loaded overlapped ones. The main output of the handover enforcement is to determine a set of pairs, *suitable* user and *suitable* target access network, for the handover execution.

Let $\mathcal{M} = \{M_1, M_2, \ldots, M_m\}$ denote a set of mobile users currently connected to a heavily loaded BS_0 that needs to be unloaded. The set of neighboring BSs overlapped with BS_0 is denoted by $\mathcal{B} = \{BS_1, BS_2 \ldots, BS_k\}$ and the current load of each of the neighboring BSs is $\wp = \{\rho_0, \rho_1, ..., \rho_k\}$. The load balancing scenario is illustrated in Figure 3. While the load of BS_0 is still greater than δ and the load balance index ξ_2 can still be decreased, we move a user M_I to a BS_J in such a way that the new arrangement minimizes ξ_2. This operation consists of one move, one user from BS_0 to BS_J, at a time. We recognize that in some situations two consecutive moves, one user from BS_0 to BS_J and another user from BS_J to BS_K, can help to reduce the overall load balance index, which could not be achieved by the one-move operation. The two-move operation requires more computation effort. Accordingly, we propose a handover enforcement including of two steps:

In the first step, we identify a move (I, J) of suitable user M_I from an overloaded BS_0 and suitable BS_J for load balancing handover. (I, J) is given by

$$(I, J) = \arg\min_{(i,j)} \xi_2(0, i, j), \tag{6}$$

where

$$\xi_2(0, i, j) = \max(\rho_0 - w_{i0} - \delta, 0) + \max(\rho_j + w_{ij} - \delta, 0) + \sum_{l \neq \{0,j\}} \max(\rho_l - \delta, 0). \tag{7}$$

Here, w_{ij} is the load contribution of user M_i at BS_j while M_i connects to BS_j. Also, $w_{ij} = \infty$ if M_i is not in the radio coverage of BS_j.

If the system is still overloaded, we search a possible two-move operation to reduce the overload situation: move a user M_I of BS_0 to BS_J and then move a user M_L of BS_J to BS_K. (I, J, L, K) is given by

$$(I, J, L, K) = \arg\min_{(i,j,l,k)} [\xi_2(0, i, j) + \xi_2(j, l, k)], \tag{8}$$

where

$$\xi_2(j, l, k) = \max(\rho_j - w_{lj} - \delta, 0) + \max(\rho_k + w_{lk} - \delta, 0) + \sum_{r \neq \{j,k\}} \max(\rho_r - \delta, 0). \tag{9}$$

Instead of balancing the resources of the overall system as described in the optimal algorithm, our proposed solution aims at redistributing locally the load of a heavily loaded BS around its neighbouring overlapped BSs. In turn, the neighbouring BS will redistribute its load to its own neighbouring BSs and so on. By doing so, the load of the overall system will be then balanced. In fact, the handover enforcement will be triggered when the load of a specific BS is greater than δ. The algorithm execution is continued until $\xi_2 = 0$ or we cannot find a handover to improve index ξ_2. In our proposition we only consider one-move and two-move operations during the handover enforcement since considering more than two consecutive moves is not realistic in on-line system due to its computation time.

5 Performance Evaluation

In this section, we first show the effectiveness of our new load balance index ξ_2 which is used as an objective function in our proposed load balancing scheme. Next, the performance of our proposed solution is compared with the optimal solution and a reference scheme. The chosen reference solution employs an *advanced* admission control [7, 8, 17], in which a new incoming communication will be redirected to the least loaded BS. The smallest load value includes the load of the new incoming communication.

5.1 Simulation and Performance Metric

We consider a simulation scenario in which users start and stop dynamically their communication sessions. Each communication is associated with a guaranteed bit rate η which is randomly generated in the interval $\eta \in [200, 3000] Kbps$. Assume that a user has only one communication session at a time and the duration of each communication follows an exponential distribution with a selected averaged value of 5 minutes. A user has the possibility to connect to a random number of BSs. As we focus on the load balancing operation, the simulation of the physical and MAC layers is not necessary in order to observe the load balancing performance. Therefore, the radio link quality between a user and its reachable BSs (i.e., the modulation and coding rates) is randomly selected at the beginning of each communication session. The modulation and coding rate ϕ varies from 0 (i.e., radio link is very poor for the connection or user is outside the BS's radio coverage) to 4 $bit/symbol$. The capacity of each BS is randomly selected in the interval $[1, 10]$ $Msymbol/s$.

The performance is evaluated by means of a user satisfaction degree. When MS i is connected to BS j, the achievable throughput of MS i is given by (inspired by [23])

$$T_i = \frac{\eta_i}{\rho_j} g(\gamma_{ij}) = \frac{\eta_i}{\rho_j} \frac{L}{B} [1 - 0.5 \exp(-v\gamma_{ij})]^B, \tag{10}$$

where γ_{ij} is the SNR of the radio link between MS i and BS j, B is the block size, L is the number of

data bit within the block size B and v is the specified constant depending on the considered technology. In fact, $g(\gamma_{ij})$ is the probability that the radio frame of size B is transmitted without errors. And $\frac{\eta_i}{\rho_j}$ represents the achievable data rate if user MS i connects to BS j. For sake of simplicity, we assume that there is no errors on radio transmission, i.e., $g(\gamma_{ij}) = 1$. The achievable throughput of user MS i, connected to BS j, is thus equal to $T_i = \frac{\eta_i}{\rho_j}$.

To compute the user satisfaction, we use the modified Sigmoid utility function proposed in [24]. Based on the achievable throughput, the user satisfaction degree is given as

$$u_i(T_i) = \begin{cases} 1, & T_i > \eta_i \\ \frac{(\frac{T_i - \eta_i^{min}}{0.5\eta_i - \eta_i^{min}})^\zeta}{1+(\frac{T_i - \eta_i^{min}}{0.5\eta_i - \eta_i^{min}})^\zeta}, & \eta_i \geq T_i \geq \eta_i^{min} \\ 0, & \text{otherwise} \end{cases}, \quad (11)$$

where η_i^{min} is the minimum acceptable bandwidth threshold of MS i. The parameter ζ is the tuned steepness parameter that follows $\zeta \geq 2$. In fact, we assume that a user will be completely satisfied ($u_i = 1$) if his achievable throughput is greater or equal to what he asks for (i.e., $T_i \geq \eta_i$). And he will be half satisfied ($u_i = 0.5$) if he gets only a half amount of throughput that he asks for (i.e., $T_i = 0.5\eta_i$). In this simulation part, we assume that $\eta_i^{min} = 0$ and $\zeta = 3$.

In this paper, we will use the averaged user satisfaction over all users in the system, given by (12), as the performance metric to compare different load-balancing algorithms.

$$U = \frac{1}{M}\sum_{i=1}^{M} u_i(T_i). \quad (12)$$

5.2 Validation of the Load Balance Index ξ_2

We employ indexes ξ_1 and ξ_2 as load-balancing objective functions. Another strategy consisting in minimizing the total load of all BSs is also examined. The performance of the three strategies is illustrated in Figure 4. In this simulation, the number of BSs in the system is fixed at 20. The value of threshold δ here is selected as $\delta = 0.95$. Note further that when we change the number of BSs or users in the system, the whole system configuration (e.g., ϕ, BS's capacity, η) is modified. The comparison of the user satisfaction or balance index between different network configurations (number of BS and number of user) is not relevant. Note however that we keep the same initial network configuration to test the different load balancing algorithms.

From Figure 4, the averaged user satisfaction is decreased when the number of users increases. As the number of BS in the system is fixed, a large number of users results in a high-load system. The user satisfaction is thus declined. The averaged user satisfaction is also averaged over many simulation repetitions. We observe that the ξ_2-based strategy gives the best performance compared to the two other strategies in

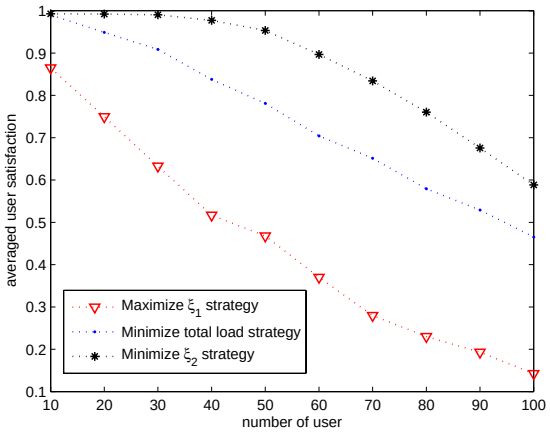

Figure 4. Averaged user satisfaction vs. load balancing objective function strategies.

any simulated network-load contexts. When the overall network load increases, the ξ_1-based strategy exposes clearly its limitation. For example, when the number of users is 60, the averaged user satisfaction given by the ξ_2-based strategy is around 0.9 while that given by ξ_1-based strategy is lower than 0.4. The ξ_1 strategy is not suitable since an equalization of all BSs' load is sometimes wasteful and does not lead to a good system performance. The performance of ξ_1 strategy is even worse than the total load minimization strategy. The later strategy does not result in an efficient resource utilization either because minimizing the total load does not mean a minimization of the system overload level. The simulation result affirms the efficiency of the ξ_2-based load balancing strategy.

5.3 Performance of the Proposed Load Balancing Strategy

We compare the performance of our proposed scheme with the impractical optimal solution. As the optimal solution requires a great computation time, the number of users arriving at a time is limited to 15 and a small number of BSs are considered. However, each user requires a high bit rate η ($400 \leq \eta \leq 3000$) to introduce an important load in the system. According to Figure 5 and Figure 6, it is clear that our proposed algorithm performs very well compared to the optimal one. Indeed, the balance indexes ξ_2 given by our solution are almost the same as those of the optimal one. We can see that the averaged user satisfaction of our proposed solution is slightly smaller than that of the optimal one in some situations (e.g., when the number of BSs is equal to 4 or 6). Most of the cases, the balance index and averaged user satisfaction provided by the our solution are identical to the optimal one.

The simulation shows clearly that our proposed algorithm provides a very close result compared to optimal but impractical one. Remind that our handover enforcement is based on two search steps: one-move and two-move operation searching. In order to investigate the advantage offered by the two-move operation

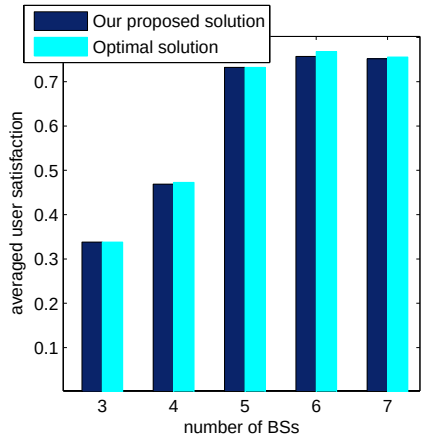

Figure 5. Averaged user satisfaction between our solution and the optimal one.

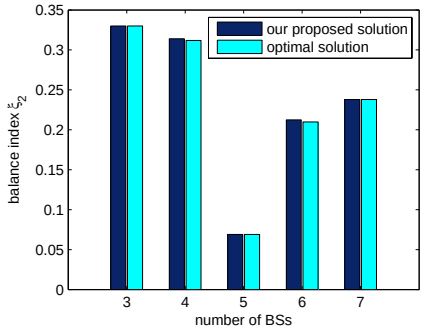

Figure 6. Balance index ξ_2 between our solution and the optimal one.

searching, we compare the balance index ξ_2 obtained from the algorithm using only one-move searching and the one using both one-move and two-move searching. In this simulation, the number of BS is fixed to 20. The balance index ξ_2 is presented in Figure 7. The result is averaged over 50 simulation repetitions for each chosen number of users. The balance index ξ_2 obtained by our proposed algorithm is smaller than the one using only one-move operation searching. It means also that the averaged user satisfaction of the proposed algorithm is better than the one-move algorithm. In fact, in the two-

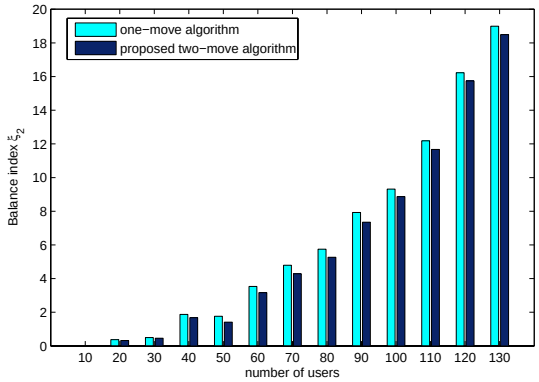

Figure 7. Load balancing using one-move handover enforcement vs. the one using two-move handover enforcement.

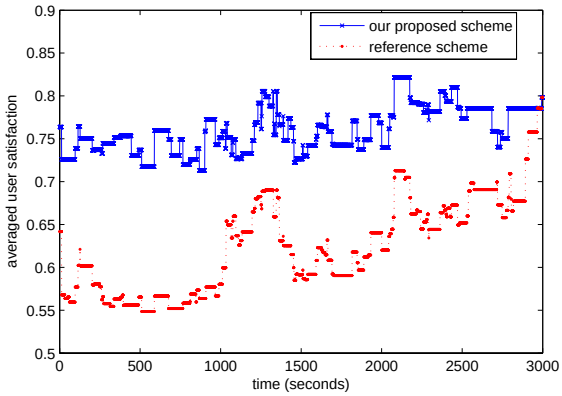

Figure 8. Performance comparison between our solution and the reference one.

move operation searching, we consider the handover enforcement of the user connected to a non-overloaded BS. This move unfreezes a enough resource amount on this BS for a possible incoming enforced handover user. One may note that we can also improve the load balancing by considering the three-move operation searching. However, the proposed algorithm, taken into account the two-move operation, provides already a very close to optimal result. A possible improvement from the three-move operation searching is not much significant compared to its computation time.

We compare now our proposed scheme with the one using advanced admission control algorithm [7, 8, 17] in which a new incoming communication will be redirected to the least loaded BS. We start two separate simulations using the same initial load-balanced system, the same user arrival process and the same running application scenario. The number of BSs is set to 10 and the number of users is set to 30. The load variation of the system is due to the communications start/stop. The averaged user satisfaction of the two systems, the one managed by our proposed load balancing scheme and the one managed by the reference scheme, is observed at every instant of the simulation duration and is depicted in Figure 8. We observe that the averaged user satisfaction degree in the system managed by our proposed scheme is much higher than in the system managed by the reference one. In fact, our proposed scheme uses a simple admission control compared to the advanced admission control of the reference one. The key of our scheme is based on the handover enforcement process that handles imminent overload situations. The results show clearly the effectiveness of our solution which is furthermore feasible for implementation in both homogeneous and coordinated heterogeneous networks.

6 Conclusion

In this paper, we have proposed a new load metric which makes it possible to formulate the load balancing as a classic optimization problem. This novel load metric for wireless packet networks is based on the

packet scheduling and the radio link quality information. Thank to this new metric, the heterogeneity of different access technologies can be removed. It also facilitates the load balancing operations since it allows load variation anticipation. We introduced a new load balancing index to measure the overload degree of a system. This balancing index leads to minimize the overload degree of a system instead of equalizing the load among the access nodes within a system. We designed a load balancing scheme which consists of an admission control and a handover enforcement. The proposed handover enforcement based on one-move and two-move iterative search is one of the feasible suboptimal solutions to the problem. The solution can be used in on-line system because it does not require much computation time and because it operates in a distributed way instead of a usual centralized way. It was shown that our proposed approach outperforms the existing approaches. We also showed that the performance of the proposed scheme is very close to the optimal but unimplementable solution. In the future work, we plan to investigate the joint load balancing and resource allocation optimization in heterogeneous networks.

References

[1] S.-L. Tsao and C. Lin, "Design and evaluation of UMTS-WLAN interworking stategies," in *Proc. 56th IEEE Vehicular Technology Conference (VTC)*, 2002, pp. 777–781.

[2] L. Gras, Q.-T. Nguyen-Vuong, Y. Ghamri-Doudane, N. Agoulmine, M. Kassar, B. Kervella, and G. Pujolle, "Terminal mobility in mobile and wireless realm: Tight, loose vs. very loose coupling," in *Proc. 1st Int. Workshop on Seamless Service Mobility (SSMO)*, Marrakech, 2007.

[3] 3GPP, "Improvement of RRM across RNS and RNS/BSS (Release 5)," 3rd Generation Partnership Project (3GPP), Tech. Rep. TR25.881 v5.0.0, December 2001.

[4] N. Vulic, S. Groot, and I. Niemegeers, "Common radio resource management for WLAN-UMTS integration at radio access level," in *Proc. 14th IST Mobile and Wireless Communications Summit*, Germany, June 2005.

[5] J. Perez-Romero *et al.*, "Common radio resource management: Functional models and implementation requirements," in *Proc. 16th IEEE Int. Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, vol. 3, 2005, pp. 2067–2071.

[6] R. Ferrus, A. Gelonch, O. Sallent, and J. Perez-Romero, "Vertical handover support in coordinated heterogeneous radio access networks," in *Proc. 14th IST Mobile and Wireless Communications Summit*, Germany, 2005.

[7] R. Agusti, O. Sallent, J. Pérez-Romero, and L. Giupponi, "A fuzzy-neural based approach for joint radio resource management in a beyond 3G framework," in *Proc. 1st Int. Conference on Quality of Service in Heterogeneous Wired/Wireless Networks (QSHINE)*, vol. 4, 2004, pp. 216–224.

[8] L. Giupponi, J. Agusti, J. Perez-Romero, and O. Sallent, "Joint radio resource management algorithm for multi-RAT networks," in *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, St Louis, MO, 2005, pp. 3851–3855.

[9] B. B. Chen and M. C. Chan, "Resource management in heterogeneous wireless networks with overlapping coverage," in *Proc. 1st Int. Conference on Communication System Software and Middleware (COMSWARE)*, India, 2006, pp. 1–10.

[10] A. Hasib and A. Fapojuwo, "Analysis of common radio resource management scheme for end-to-end QoS support in multiservice heterogeneous wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 4, pp. 2426–2439, 2008.

[11] C. Shin and J. Cho, "A preliminary study on common radio resource management in heterogeneous wireless networks," in *Proc. 3rd Int. Conference on Ubiquitous Information Management and Communication (ICUMIC)*, 2009, pp. 278–282.

[12] P. Stuckmann *et al.*, "The EUREKA gandalf project: Monitoring and self-tuning techniques for heterogeneous radio access networks," in *Proc. 61st IEEE Vehicular Technology Conference (VTC)*, vol. 4, 2005.

[13] W. Shen and Q. Zeng, "Resource allocation schemes in integrated heterogeneous wireless and mobile networks," *Journal of Networks*, vol. 2, no. 5, 2007.

[14] L. Giupponi, R. Agusti, J. Perez-Romero, and O. Sallent, "Improved revenue and radio resource usage through inter-operator joint radio resource management," in *Proc. IEEE Int. Conference on Communications (ICC)*, 2007, pp. 5793–5800.

[15] D. Niyato and E. Hossain, "A noncooperative game-theoretic framework for radio resource management in 4G heterogeneous wireless access networks," *IEEE Transactions on Mobile Computing*, vol. 7, no. 3, pp. 332–345, 2008.

[16] Y. Zhang, K. Zhang, Y. Ji, and P. Zhang, "Adaptive threshold joint load control in an end-to-end reconfigurable system," in *Proc. 15th IST Mobile and Wireless Communications Summit*, Greece, 2006.

[17] G. Bianchi and I. Tinnirello, "Improving load balancing mechanisms in wireless packet networks," in *Proc. IEEE Int. Conference on Communications (ICC)*, 2002, pp. 891–895.

[18] U. Bernhard, E. Jugl, J. Mueckenheim, H. Pampel, and M. Soellner, "Intelligent management of radio resources in UMTS access networks," *The Bell Labs Technical Journal*, vol. 7, pp. 109–126, 2003.

[19] H. Velayos, V. Aleo, and G. Karlsson, "Load balancing in overlapping wireless LAN cells," in *Proc. IEEE Int. Conference on Communications (ICC)*, 2004, pp. 3833–3836.

[20] Q. Liu, X. Wang, and G. Giannakis, "A cross-layer scheduling algorithm with QoS support in wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 55, no. 3, pp. 839–847, 2006.

[21] D. Chiu and R. Jain, "Analysis of the increase and decrease algorithms for congestion avoidance in computer networks," *Computer Networks and ISDN systems*, vol. 17, no. 1, pp. 1–14, 1989.

[22] R. Ramjee, R. Nagarajan, and D. F. Towsley, "On optimal call admission control in cellular networks," in *Proc. IEEE Int. Conference on Computer Communications (INFOCOM)*, 1996, pp. 43–50.

[23] X. Zhang, E. Zhou, R. Zhu, S. Liu, and W. Wang, "Adaptive multiuser radio resource allocation for OFDMA systems," in *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, vol. 6, 2005.

[24] Q. Nguyen-Vuong, Y. Ghamri-Doudane, and N. Agoulmine, "On utility models for access network selection in wireless heterogeneous networks," in *Proc. IEEE/IFIP Network Operations and Management Symposium (NOMS)*, Brazil, April 2008.

Quoc-Thinh Nguyen-Vuong is currently a project manager at France Telecom for IMS-based value-added service solutions. Prior to joining France Telecom, he spent almost three years as a senior consultant at Davidson Consulting. He had expertise in Alcatel-Lucent's 3G HSxPA technology introduction and deployment for various large mobile operators like Vodafone, AT&T, China Unicom, Orange France. He received his masterŠs degree from University of Paris VI and his PhD from the University of Evry (France) in computer networks respectively in 2005 and 2008. He received the engineering diplomas of the Ecole Polytechnique (France) and Telecom ParisTech (France) in 2005. His research interests include the resource management in wireless network, mobility management, fixed and mobile service convergence, all-IP mobile network management and next generation network.

Nazim Agoulmine is currently full professor at the University of Evry, France. He is the head of the Networks and Multimedia Systems Research Group (LRSM). He received his master's degree and his PhD in computer science respectively in 1989 and 1992 from the University of Paris XI. Prior to joining the University of Evry, France, he was an associate professor at the University of Versailles (France) for 8 years, associated professor at the University of Quebec at Montreal (Canada) and senior scientist at GMD-Fokus (Germany). His research interests include network and service management, autonomic communications, multimedia communication systems, fixed and mobile networks integration, Quality of Service control.