A Novel Load Balancing Algorithm Based on Utility Functions and Fuzzy Logic in Heterogeneous Wireless Networks

Sheng Jie, Yang Zheng, Tang Liang-rui, Hao Jian-hong
School of Electric and Electronic Engineering
North China Electric Power University
Beijing, China

***Abstract*—Aiming at working out the problem of existing load balancing algorithms that failed to consider the requirements of different types of services sufficiently in heterogeneous wireless networks, a new load balance scheduling algorithm was proposed in this paper. Firstly, three utility functions were designed to reflect three implementation-specific aspects, including the effective bandwidth, real-time calls blocking rate and the average non-real-time calls transmission time as per the needs of different types of services combined with the actual performance of each available network in heterogeneous wireless networks. Further, a fuzzy logic algorithm was adopted to fulfill the fuzzy decision of three utility functions above. Finally, we adjusted the load of each network according to the fuzzy decision value to achieve equilibrium. Simulation results showed that the proposed load balancing algorithm performs much better than reference algorithms with respect to the system blocking probability, average transmission time and network load rate equalization.**

Keywords-heterogeneous wireless networks; load balancing; utility functions; fuzzy logic

I. INTRODUCTION

The interworking between heterogeneous third-generation cellular networks and wireless local area networks (WLANs) is one promising evolution approach to fourth-generation wireless networks, which can exploit the complementary advantages of the cellular network and WLANs [1,2]. Resource management for the 4G-oriented integrated network is an important open issue that deserves more research efforts[3].Load balancing is an effective way to improve the network operation quality, which on the one hand can be used to alleviate the problem of unequal distribution of the heterogeneous network resources, and on the other hand can enhance system capacity and service quality as well as improve the utilization of increasingly strained radio resources, that has become a hot research spot[4].

Currently, a large number of scholars in and abroad have done a lot of researches in load balancing for heterogeneous networks, and there are different load balancing methods. Literature [4] presented a policy-based resource management framework that used the access control and vertical handover strategies respectively to handle the service requests of new calls and handoff calls, so as to maintain the dynamic balance of the load. However, the algorithm did not conduct a detailed analysis for library strategy design. The sojourn time based load balancing algorithm proposed in [5] reduced the blocking rate of the heterogeneous system to some extent and improved resource utilization at the expense of increasing the switching rate. Hyukmin Son [6] provided the concept of soft load balancing and studied the optimal load split ratio under a specific network topology that is not universal. The MLB algorithm mentioned in [7] took into account the QoS requirements of different services and link-layer retransmission mechanism in order to choose the network with the smallest ratio of average service resources consumption and available network resources to access. These algorithms could not distinguish the QoS requirements for different types of services as well as the priority of various services for different types of networks.

The main contribution of this paper is proposing a more comprehensive solution to the load balancing problem over integrated cellular and WLAN systems. We take into account the needs of different services combined with the actual performance of each available network in heterogeneous environment and develop proper utility functions and fuzzy logic based load balancing algorithm. First of all, three utility functions are designed to reflect three implementation-specific aspects of network performance, including the effective bandwidth, real-time (RT) calls blocking rate and the average non-real-time (NRT) calls transmission time. Further, a fuzzy logic system is adopted to fulfill the fuzzy decision values of three utility functions. Finally, we adjust the load of each network according to the fuzzy decision value to achieve equilibrium.

The rest of this paper is organized as follows. The system model and the general idea of the algorithm is introduced in section 2,the utility functions and fuzzy logic based load balancing algorithm is proposed in section 3, simulation results and performance are analyzed in section 4, conclusions and further work are given in the final section.

II. SYSTEM MODEL AND GENERAL IDEA

To make the algorithm universal, we adopt the stratified semi-centralized system architecture of heterogeneous wireless networks. The system consists of several cellular cells and

978-1-4673-0024-7/10/$26.00 ©2012 IEEE

WLAN hotspots, the base station (BS) of cellular and wireless access point (AP) of WLAN are connected to the upper mobile control and administration center (CAC). Thus, the load balancing information will not be broadcasted in the whole network that reduces the overhead considerably. When any of the candidate networks' traffic load condition changes, the BS or AP will report the fresh load information to CAC. Therefore, the implementation of load balancing requires a mutual cooperation among BS, AP and CAC. The control process of this utility functions and fuzzy logic based load balancing approach is divided into three parts, including monitoring, analysis and decision making as well as implementation, as shown in Figure 1.

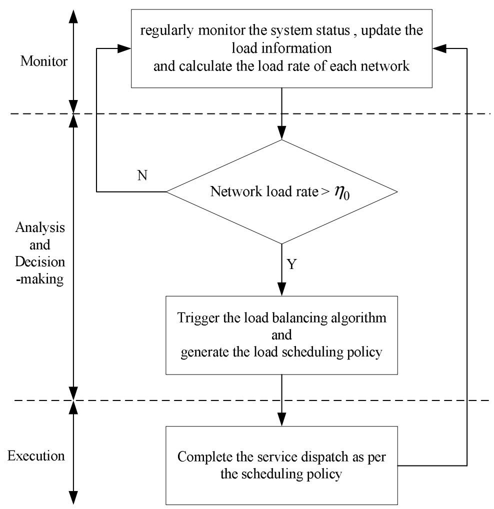

Figure 1. The proposed load balancing framework for heterogeneous wireless networks

In the first place, the CAC regularly monitors and updates the load information that periodically reported by the BS and the AP to calculate the load rate of each network. In this paper, we take the network bandwidth utilization rate as the assessment of the load parameters, i.e., the ratio of the occupied effective bandwidth and the total available bandwidth of the network. Secondary, if the network load rate continuously greater than the load threshold (η_0) for a certain period of time, then the load balancing is triggered. Otherwise, the CAC continues to monitor the system status. Note that the value of η_0 should be slightly less than the maximum of network available bandwidth resources in order to avoid the blocking of handoff calls that caused by the lack of resources. Further, the load dispatch strategy is implemented as per the load balancing algorithm. Then, the CAC executes the scheduling of services as per the dispatch strategy, and the load balancing process is completed. Finally, the CAC goes back to monitor the system load information.

III. Proposed Load Balancing Algorithm

A. Network performance indicators

In this paper for simplification we considered that call setup to each network follows Poisson distribution with the mean arrival rate of λ_{RT} and λ_{NRT}, which are corresponding with RT and NRT services respectively. Call duration is exponentially distributed with mean of $1/\mu$. Each individual network is supposed to provide N_{RT} and N_{NRT} connects for RT and NRT services. Based on the Erlang distribution model, the blocking rate of RT services can be obtained as follows:

$$P_{RT} = \frac{(\lambda_{RT} / \mu)^{N_{RT}} / N_{RT}!}{\sum_{i=0}^{N_{RT}} (\lambda_{RT} / \mu)^i / i!} \quad (1)$$

Let f_{NRT} denote the average data file size, $C_{NRT}(i)$ is the residual bandwidth of current network for NRT service, then the NRT service load-bearing strength ρ_{NRT} is defined as follows:

$$\rho_{NRT}(i) = \frac{\lambda_{NRT} * f_{NRT}}{C_{NRT}(i)} \quad (2)$$

T_{NRT} represents mean transfer time for NRT service [8].

$$T_{NRT} = \frac{(\rho_{NRT})^{N_{NRT}+1}(N_{NRT} * \rho_{NRT} - N_{NRT} - 1) + \rho_{NRT}}{\lambda_{NRT}(1-(\rho_{NRT})^{N_{NRT}})(1-\rho_{NRT})} \quad (3)$$

B. Utility functions

There are different standards for performance evaluating of the QoS between real-time and non-real-time calls. In general, real-time services are more concerned about the blocking rate, including the new call blocking and handover call dropping rate, rather than real-time business that focusing on the length of average transmission time [9]. Hence, we explore three specific performance parameters, i.e., the effective available bandwidth, RT service blocking rate and NRT services mean transmission time, to measure the load conditions and business performance of each network. More over, we design three different utility functions to evaluate the three parameters respectively.

The utility function of effective available bandwidth for net i can be obtained readily through a direct application of Eq.(4).

$$U_{BW}(i) = C_{avl}(i) / C_{total} \quad (4)$$

where $C_{avl}(i)$ is indicative of effective available bandwidth for net i while C_{total} represents the total bandwidth of the network.

When it comes to the RT service, the most important principle is to ensure that the blocking rate of business within in the scope of the requirements. The utility function of RT service blocking rate ($U_{RT}(i)$) of net i is given as:

$$U_{RT}(i) = \lg P_{RT}(i) / \lg P_{RT\_tag} \quad (5)$$

where $P_{RT}(i)$ denotes the blocking rate of net i for RT service , and P_{RT_tag} is the maximum value of blocking rate that RT service can tolerate. As shown in Eq.(5), the smaller that the $P_{RT}(i)$ is, the lager $U_{RT}(i)$ will to be.

For NRT service, more attention should be paid to the mean transfer time of data packets. The utility function attribute of NRT service mean transfer time is now calculated as Eq.(6), which is an index of reference of EXP algorithm[10].

$$U_{NRT}(i) = \exp(\frac{T_{NRT\_avg} - T_{NRT}(i)}{1+\sqrt{T_{NRT\_avg} / T_{NRT\_tag}}}) \quad (6)$$

where T_{NRT_avg} is the average value of NRT service mean transfer time of networks in the heterogeneous system, $T_{NRT}(i)$ is NRT service mean transfer time of net *i*, whereas T_{NRT_tag} represents maximum value of transmission time that NRT service can tolerate. Eq.(6) is an illustration of the fact that when $T_{NRT}(i)$ is smaller than T_{NRT_avg} ,the value of $U_{NRT}(i)$ will obtain an exponential growth.

C. Fuzzy logic based network performance evaluation

In the first place, choose the values of three utility functions to fully evaluate the network performance. Define three fuzzy subsets, and take them as the system input variables, as well as define the membership functions and fuzzy linguistic variables of each input variable. The input fuzzy linguistic variables are shown in Table1, which also indicates the quality of performance for each level. The triangle and trapezoidal functions are used as the membership functions of input variables. In the second place, define a fuzzy subset, which is named by *Fuzzy_out*, as the system output variable. Select the triangle membership functions, and define three fuzzy linguistic variables (L, M, H), which represent the comprehensive performance (CP),as shown in Table 1.The next step is defining the fuzzy evaluation rules of the fuzzy logic system. Since all of the parameters have three fuzzy linguistic variables, there are totally 27 inference rules. As the length of the article is limited, we do not discuss the input and output variables membership functions and fuzzy inference rules in detail.

TABLE I. FUZZY SET ABBREVIATIONS USED IN RULE BASES

	Input			Output
	U_{BW}	U_{RT}	U_{NRT}	*CP*
L	Few Available bandwidth	High Blocking rate	Long Transmission time	Bad
M	Medium Available bandwidth	Medium Blocking rate	Medium Transmission time	Medium
H	Adequate Available bandwidth	Low Blocking rate	Short Transmission time	Good

The three steps above create a complete fuzzy system based on utility functions. According to the value of each utility function, using the 27 rules of the system, we can obtain 27 *Fuzzy_out* values. Then, the values of L, M and H are able to be obtained by the minimum value method, according to their membership functions. The crisp value of *Fuzzy_out* is determined by using the following formula:

$$Fuzzy\_out = \frac{\sum M_i \times W_i}{\sum M_i} \quad (7)$$

where M_i is the degree of membership in output singleton *i*, W_i is the weight value for the output singleton *i*.

D. Process of Load Balancing Algorithm

The process of the proposed load balancing algorithm is designed in the following.

Step1: CAC checks whether a candidate network is over load by its load rate, and if there is, the load balancing process will be triggered.

Step2: Obtain the load information of individual networks in the heterogeneous system, including the numbers of RT and NRT services that are serving in each network. Record the information as the initial state of the heterogeneous system.

Step3: Calculate the utility function values for the parameters of the effective available bandwidth, RT service blocking rate and NRT services mean transmission time.

Step4: Obtain the *Fuzzy_out* values of candidate networks by using utility functions and the fuzzy logic system.

Step5: It is the implementation of load scheduling algorithm that selects a suitable service as per the type of the overloaded load network to the one with the maximum *Fuzzy_out*. Take WLAN as an example, we encourage the NRT service to access WLAN. Therefore, when the type of overloaded network is WLAN, handoff preference will be given to RT service. By contrast, when it comes to Cellular, NRT service will be selected with priority.

Step6: Update the business distribution of each network and calculate the mean square deviation (*M*) of network load rates. If *M* is larger than ε, which is far less than 1, then go back to step 2 and continue to the next round of load scheduling. If *M* is less than ε, then the distribution of services will be put out as the balanced load distribution, and the process of load scheduling is completed.

The load scheduling strategy can be received by the comparison of the eventual RT and NRT services distribution with the initial state.

IV. PERFORMANCE EVALUATION

A. Simulation model

To validate the effectiveness of the load balancing scheme, in the following we create a simulation model. We assume that there are three adjacent cellular network cells in the heterogeneous wireless system and six WLAN hotspots at the edge region of the intersection area of three cells where exists overlapped coverage. In this way, there are totally nine available networks for service accessing in this model. Call setup to each network follows Poisson distribution and randomly distribute in the nine networks. The mobility of users during the load balancing process is out of consideration in this paper. The system parameters given in Table 2 are used in the following performance evaluation.

In this paper, we evaluate and compare the algorithm's performance with two other reference schemes for load balancing. In the first reference scheme, two types of network works independently without using load balancing. The second reference scheme is MLB algorithm [7] that considers different QoS requirements and the link-layer retransmission mechanism in order to choose the network with the smallest ratio of average service resources consumption to access. Computer simulation is used to evaluate their performances due to difficulties in analytical modeling.

TABLE II. SYSTEM PARAMETERS FOR PERFORMANCE EVALUATION

Parameter	Value	
Network types	WLAN	Cell
Network ID	1-6	7-9
Total effective bandwidth(Mbps)	5.5	2
Average arriving rate of RT call(Calls/s)	0.1-1	
Average arriving rate of NRT call(Calls/s)	1-10	
Bandwidth request of RT call(Kbps)	30	
Bandwidth request of NRT call(Kbps)	64	
Average call duration(s)	100	
Load balancing trigger threshold	0.9	
Mean square deviation of load rate	0.03	

B. *Simulation results*

The main performance evaluation of load balancing algorithm is the RT service blocking probability, average transmission time of NRT service and the load rate of networks in the system. The simulation results are compared and analyzed in the following.

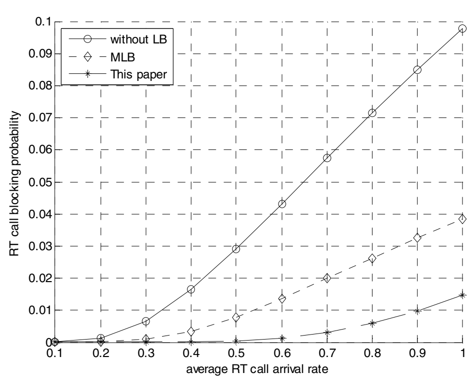

Figure 2. RT call blocking probability

Figure 2 shows the RT call blocking probabilities of the three schemes in comparison. It can be seen that the blocking rate of three kinds of simulation methods increases correspondingly with the increasing of average RT call arrival rate. After the introduction of load balancing, RT service blocking probability of the overall heterogeneous system is significantly decreased. The RT call blocking rate of the proposed load balancing algorithm is as much as 50 percent lower than that of the second reference scheme. The significant performance improvement comes from the fact that the proposed scheme is based on integrated utilities of RT service and NRT service, focusing on the RT business transferred out from overloaded WLAN network as well as the NRT business transferred out from overloaded cellular network, which has given full play to the two types of networks in the provision of the provided services performance. In addition, within a load balance scheduling period, the amount of the transferred load will be greater than that of the reference algorithms which comes out the result that the efficient available resource of overloaded area will increase greatly.

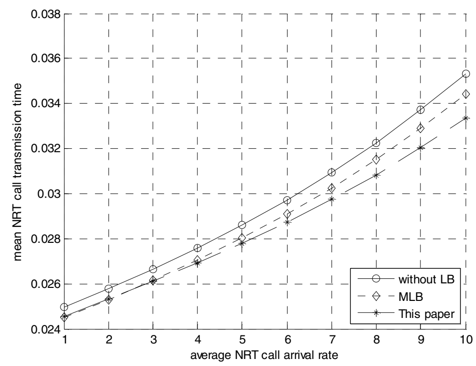

Figure 3. NRT call mean transfer time of each candidate network

Figure 3 shows comparison of the mean NRT call transfer time of the three schemes. The mean NRT call transfer time of MLB algorithm is almost 5 percent lower than that of the first reference scheme. A thorough examination of the fact reveals that it is unreasonable to make the load scheduling only depend on the load rate, which may cause a slight improvement of the network performance, such as the transmission delay of data packet. In contrast, the average NRT call transmission time of the proposed algorithm is almost 10 percent lower than that of the scheme without load balancing. This performance improvement comes from the fact that the algorithm takes into account the network load combined with the blocking rate and other parameters to make an integrate assessment that brings about a much more reasonable load scheduling strategy. The scheduling result still shows that the utilization of the proposed algorithm not only substantially reduced the mean NRT call transfer time of the overloaded load networks, but also greatly improved the load rate equalization of each network in the heterogeneous system.

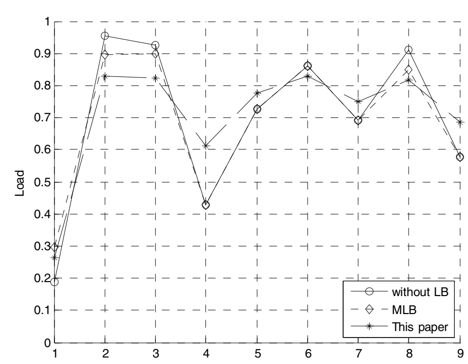

Figure 4. Load rate of each candidate network

Figure 4 depicts the load rate changes of different algorithms after load balancing process. As show in the figure, the load rate of the overloaded load networks that over the threshold η_0 are all reduced dramatically after load balance scheduling. The load rate distribution of the proposed scheme is much more uniform than that of the other two schemes. The changes and the fluctuations of provided algorithm tend to be smooth and steady which comes into being a better network performance than that of the other ones.

In conclusion, the proposed utility functions and fuzzy logic based load balancing algorithm is much better than the reference algorithms in improving the quality and the performance of the service for different types of business in heterogeneous system as well as equating the traffic load of the networks. On the one hand, the generation and implementation of the scheduling strategy can avoid the influence of handoff mechanism on the load balancing behavior. On the other hand, adjustable distinguish degree of different types of service can be realized by using the calculation of controllable network utility functions, which can flexibly avoid the overloaded load phenomenon caused by the inadequate network selection for specific services.

V. CONCLUSIONS AND FURTHER WORK

Heterogeneous wireless networks provide chances to use different types of networks to enhance availability of system by balancing loads among the candidate networks. In this paper we have discussed a load balancing architecture under heterogeneous wireless networks. A utility functions and fuzzy logic based load balancing algorithm is proposed and the system indicators in the aspects of blocking probability, average transmission time and network load rate are analyzed. The simulation results of the algorithm show that the RT service blocking probability is significantly reduced and the average NRT service transmission time of the overloaded network is effectively decreased. The average NRT service transmission time and network load rate tends to be equal at the same time. Significant performance improvement is observed in comparison with two other reference schemes. For the further work, we will develop an effective analytical model for the integrated networks' load balancing problem and investigate how to determine the system parameters so as to maximize the system performance.

ACKNOWLEDGMENT

The work was supported by grants from the Major National Science and Technology Special Project (2010ZX03006-005-001) and the Fundamental Research Funds for the Central Universities (11QX47).

REFERENCES

[1] D. Aleksandar, M. Juan, W. Yongbin,*et al*., "A survey on 3GPP heterogeneous networks," IEEE Wireless Communications, Vol. 18, pp. 10-21, June 2011.

[2] S. Wenxiao, F. Shaoshuai, W. Nan,*et al*., "Fuzzy neural network based access selection algorithm in heterogeneous wireless networks," Journal on Communications, Vol. 31, pp. 151-156, September 2010.

[3] R. Skehill, M. Barry, W. Kent, *et al*., "The common RRM approach to admission control for converged heterogeneous wireless networks," IEEE Wireless Communications, Vol. 14, pp. 48-56, April 2007.

[4] S. Wei , Z. Weihua, C. Yu, "Load balancing for cellular/WLAN integrated networks," IEEE Network, Vol. 21, pp. 27-33, January 2007.

[5] Z. Guangxi,N. Guoqin,W. Hongya, "Research on sojourn-time-based dynamic load balancing algorithm in heterogeneous hierarchical wireless networks," Journal on Communications, Vol. 27, pp. 29-36,41, April 2006.

[6] H. Son, S. Lee, K. Soo-Chang, *et al*., "Soft load balancing over heterogeneous wireless networks," IEEE Transactions on Vehicular Technology, Vol. 57, pp. 2632-2638, July 2008.

[7] S. Zhou,L. Tiejun, Z. Qianqian,*et al*., "Radio access selection algorithm in heterogeneous wireless packet networks," Journal of System Simulation, Vol. 20, pp. 2762-2769, May 2008.

[8] S. Wei, J.,Hai Z. Weihua,*et al*. "all Admission Control for Integrated Voice/Data Services in Cellular/WLAN Interworking," in *Proc.*IEEE ICC, June 2006, Vol. 12, pp. 5480-5485.

[9] B. Hongyan, *Joint resources management technology in Heterogeneous wireless networks*. Shang Hai : Shanghai Jiaotong Univ. Press, 2008, pp. 42-47.

[10] Shakkottai S,Stolyar A L. "Scheduling algorithms for a mixture of real-time and non-real-time data in HDR," Bell Laboratories Lucent Technology, 2000.