Adaptive Neuro-Fuzzy Inference System for Dynamic Load Balancing in 3GPP LTE

Aderemi A. Atayero and Matthew K. Luka Department of Electrical & Information Engineering Covenant University Ota, Nigeria

Abstract—ANFIS is applicable in modeling of key parameters when investigating the performance and functionality of wireless networks. The need to save both capital and operational expenditure in the management of wireless networks cannot be over-emphasized. Automation of network operations is a veritable means of achieving the necessary reduction in CAPEX and OPEX. To this end, next generations networks such WiMAX and 3GPP LTE and LTE-Advanced provide support for selfoptimization, self-configuration and self-healing to minimize human-to-system interaction and hence reap the attendant benefits of automation. One of the most important optimization tasks is load balancing as it affects network operation right from planning through the lifespan of the network. Several methods for load balancing have been proposed. While some of them have a very buoyant theoretical basis, they are not practically implementable at the current state of technology. Furthermore, most of the techniques proposed employ iterative algorithm, which in itself is not computationally efficient. This paper proposes the use of soft computing, precisely adaptive neurofuzzy inference system for dynamic QoS-aware load balancing in **3GPP LTE.** Three key performance indicators (i.e. number of satisfied user, virtual load and fairness distribution index) are used to adjust hysteresis task of load balancing.

Keywords-ANFIS; 3GPP; LTE; Neural Network; Fuzzy Logic; Load balancing; Virtual load.

I. INTRODUCTION

The third generation project (3GPP) Long Term Evolution (LTE) has the core objective of meeting the increasing performance needs of mobile broadband. Some of the key features of LTE include: high spectral efficiency, very low latency, support of variable bandwidth, simple protocol architecture, and support for Self-Organizing Networks (SON) operation. SON operation was introduced to improve overall system performance through efficient operations and maintenance. Load balancing belongs to SON's self-optimizing functions, which are engineered towards reducing overall operational expenditures (OPEX) by minimizing workload for site survey, analysis of network performance and other operational and maintenance tasks that require human intervention. Generally, self-optimization involves the use of User Equipment (UE) and evolved Node B (eNB) measurements and performance measurements for network auto-tuning. The objective of load balancing is to ensure an equitable distribution of cell load among cells or to transfer part of the traffic from congested cells with the aim of improving

the overall system capacity and network performance indices [1], [2].

The process of load balancing begins with detecting network load imbalance by periodically exchanging information between neighbouring eNBs (over the X2 interface) to compare the cells load. To realize an efficient intra LTE based load balancing, the load information must consist of both radio resource usage, which corresponds to the uplink and downlink physical resource block (PRB) usage as well as generic measurements representing non-radio-related resource usage. The non-radio-related load parameters include: Transport Network Load (TNL) indicator, Hardware (HW) load indicator, and available capacity for load balancing as a percentage of total cell capacity. For inter-RAT (radio access technology) load information must include another parameter known as Cell Capacity Class Value (CCCV), which is a relative capacity indicator. An algorithm to distribute the loads towards neighboring cell(s) with minimum number of cell reselection or handover is then implemented to achieve load balancing.

Several algorithms have been envisaged. In [3], a load balancing algorithm aimed at finding the optimum handover (HO) offset value between the overloaded cell and a possible target cell was implemented. Another paradigm to load balancing for LTE networks was investigated in [4]. The approach is based on a network formulation of heterogeneous services with different quality of service requirements. In [5], Wang et al. used a network utility-based load-balancing framework to develop an algorithm called Heaviest-First Load Balancing (HFLB). Another approach postulated in [6] involves the integration of another self-optimization function handover parameter optimization to offset handover problems associated with load balancing. All of the aforementioned methods and algorithms are however based on iterative processes, which are computationally expensive. This is a serious limitation to a generalized load-balancing scheme.

In addition, since load balancing using handover is a computationally demanding task, it is desirable to divide and allocate resources between users who have data to transmit. If the desired load balancing is not achieved, then a handover is enforced. Moreover, to realize a generic load balancing, both radio resource usage and non-radio resource parameters must be incorporated. These challenges point to the need for the development of a robust, computationally less expensive and as a consequence cost effect approach. In this research work, an Adaptive Neuro-fuzzy Inference System (ANFIS) is proposed for the implementation of dynamic load balancing in LTE.

II. SYSTEM MODELING AND LOAD METRIC DETERMINATION

A. Introduction

The proposed system consists of a five-layer ANFIS that takes three inputs viz.: a) SINR – the Signal to Interference Noise Ratio of the users; b) the virtual load of a cell and c) the load distribution index of the entire network. The output of the ANFIS system is a quality of service (QoS) indicator that is used to decide either scheduling or handover, in order to achieve load balancing. The Network model is based on a 3GPP downlink multi-cell network serving users with homogenous QoS requirement. Specifically, constant bit error rate (CBR) users are taken into account. Other QoS requirements can be easily added. The SINR is used as a metric measuring the link quality of the link model [7]. Performance analysis is hinged on two factors, namely: fairness distribution of the virtual load and the link Block Error Ratio (LBER).

B. Link Model

The post-equalization symbol SINR was determined from three parts of the link measurement model. These constituent models include: (i) shadow fading, (ii) macroscopic pathloss and (iii) small scale fading (for Multiple-Input-Multiple Output). The propagation pathloss due to distance and antenna gain can be modeled by the macroscopic pathloss between an eNodeB sector and a UE. The pathloss can be noted as L_{mp,T_i, U_j} where T_i is the ith transmitter (denoted as 0 for the attached eNodeB and 1, ..., N for the interfering eNodeBs. U_j is the jth UE which is located at an (x, y) position. The pathloss was generated using a distance dependent pathloss of 128.1 + 37.6log₁₀(R[Km]) [8] and a $\theta_{3dB} = 65^{\circ}/15$ dBi antenna [9].

Shadow fading occurs due to obstacles in the propagation path between the eNodeB and UE. Shadow fading can be seen as the changes in the geographical properties of the terrain associated with the mean pathloss derived from the macroscopic pathloss model. It is often approximated by a lognormal distribution of standard deviation 10 dB and mean 0 dB. A UE moving in the Region of Interest (ROI) will experience a slowly changing pathloss due to the shadow fading of the attached eNodeB being correlated with the shadow fading of the interfering eNodeBs. Shadow fading can be denoted by L_{sf,T_i, U_j} . The large scales fading (shadow fading and pathloss) are position dependent and time-invariant.

Small-scale fading results primarily due to the presence of reflectors and signal scatter agents that cause multiple versions of the transmitted signal to arrive at receiver. The small scale fading is modeled as a time dependent process for different transmission modes. One of the MIMO transmission modes is the Open Loop Spatial Multiplexing. The MIMO OSLM channel can be modeled to obtain the per-layer SINR. This transmission mode consists of a precoding for Spatial Multiplexing (SM) with large-delay Cyclic Delay Diversity (CDD) [10]. The OLSM MIMO precoding is defined by:

$$\begin{bmatrix} y_{(0)}(i) \\ \vdots \\ y_{(N_t-1)}(i) \end{bmatrix} = W(i)D(i)U \tag{1}$$

Where, N_t – Number of transmit antennas; v – Number of layers (a layer is a mapping of symbols to the transmit antenna); W(i) = $N_t \times v$ – precoding matrix; D, U – $v \times v$ diagonal matrixes introducing the CDD.

For the MIMO OLSM, the SINR for the UE can be expressed as:

$$SINR_{c,u} = \frac{\alpha_i L_{sf,0,U} L_{pl,0,U} P_1}{\beta_i P_1 + \gamma_i \sigma^2 + \sum_{1}^{N_{int}} \theta_{i,1} L_{sf,T_i,U_j} L_{pl,T_i,U_j} P_1}$$
(2)

Where α_i and β_i model the channel estimation errors; $P_1 = P_{tx}/v$ represents the homogenously distributed transmit power; γ_i models a simple Zero Forcing (ZF) receiver noise enhancement σ^2 is the uncorrelated receiver noise and θ models the interference. $L_{sf,T_i,\,U_j}$ and $L_{pl,T_i,\,U_j}P_1$ stand for the shadow fading and pathloss between the UE, u and its attached eNodeB c (for $T_i=0$) and its interference (for $T_i=1,...,N_t)$ respectively.

A given MCS (Modulation Coding Scheme) requires a certain SINR (measured at the receiver of the UE) to operate with an acceptably low BER (Bit Error Rate) in the output data. An MCS with a higher throughput needs a higher SINR to operate [11]. We assume that the best modulation coding scheme (MCS) is used for a given SINR and the highest data rate R(SINR) is achievable, this can be represented by Shannon formula as shown below:

$$R(SINR_{u}) = \log_{2}(1 + SINR_{u})$$
(3)

For better approximation to realistic MCS, the mapping function is scaled by an attenuation factor (of say 0.75) and is bounded by the minimum required SINR of -6.5dB and a maximum bitrate of 4.8 bps/Hz.

C. Load Metric

The amount of Physical Resource Blocks (PRBs) required by user u can be expressed as:

$$N_u = \frac{D_u}{R(SINR)_u \cdot BW} \tag{4}$$

Where D_u – is required data rate; BW – is the transmission bandwidth of one resource block (180 kHz for LTE).

The load of cell c is thus expressed as the ratio of the sum of the required resources of all users connected to cell c to the total number of resources N_t :

$$\rho_c = min\left(\frac{\sum_{u:X(u)=c} N_u}{N_t}, 1\right)$$
(5)

If we chose the number of unsatisfied users as assessment and simulation metric, then we can focus on the CBR traffic rather than the network throughput. In this case, the UEs either get exactly the CBR or they totally unsatisfied. Equation (5) implies that the cell load parameter should not exceed 1 for all users to be satisfied. This can be extended to give a general indication of how overloaded (or otherwise) a cell is, by defining a virtual load given by:

$$\widehat{\rho_C} = \frac{\sum_{u:X(u)=c} N_u}{N_t} \tag{6}$$

Where $\widehat{\rho_c} \leq 1$ implies that all users in the cell are satisfied, $\widehat{\rho_c} = U$ means 1/U of the users are satisfied. The total number of unsatisfied users in the whole network (with a total number of M_c users in cell c) is given by:

$$z = \sum_{c} max \left(0, M_{c} \cdot \left(1 - \frac{1}{\widehat{\rho_{c}}} \right) \right)$$
(7)

For performance analysis, the use of a fairness distribution proposed in [12] is employed. Thus, the load distribution index measuring the degree of load balancing of the entire network is given as:

$$\mu(t) = \frac{[\sum_{c} \rho_{c}(t)]^{2}}{|N| \cdot [\sum_{c} \rho_{c}(t)]^{2}}$$
(8)

Where |N| is the number of cells in the network (used for simulation) and t is the simulation time. The load balance index $\mu(t)$ takes the value in the interval $\left[\frac{1}{|N|},1\right]$. A larger μ indicates a more balanced load distribution among the cells. Thus, the load distribution index is 1 when the load is completely balanced. The aim of load balancing (for CBR users) is to maximize is to maximize $\mu(t)$ at each time t.

In order to improve the load balancing performance among adjacent cells, it is necessary to find the optimum target cell. This can be achieved by adopting a two-layer inquiry scheme proposed in [13]. The source eNB (the cell requiring load balancing) request load state and environment state from all neighbouring eNBs (first layer cells). The load state is the load of the first layer cell and the environment state is the average load of the first layer cell's adjacent cells excluding the one to be adjusted (denoted as the second layer cells). The overall state of the first layer cell i is obtained by a weighted combination of the load state (LS_i) and environment state (ES_i) in one figure as follows:

$$OS_i = \alpha LS_i + (1 - \alpha)ES_i \tag{9}$$

Where the environmental state is given by:

$$ES_{i} = (\rho_{1i} + \rho_{2i} + \dots + \rho_{ni})/n = \frac{\sum_{j=1}^{n} \rho_{j}}{n}$$
(10)

 $LS_i = \rho_i$, the load of first layer cell i, and α is a parameter that indicates the relative contribution of LS_i and ES_i to OS_i .

 OS_i gives a comprehensive load information of the first layer cell, thereby indicating whether the eNodeB can be a target cell. Taking the value of $\alpha = 0.2$ equation (9) can be expressed as:

$$OS_i = (0.2 \times \rho_i) + 0.8 \times \left(\frac{\sum_{j=1}^n \rho_j}{n}\right) \tag{11}$$

III. ADAPTIVE NEURO-FUZZY INFERENCE SYSTEM

Adaptive Neuro-Fuzzy Inference System (ANFIS) otherwise referred to as Adaptive Network-based Fuzzy inference System was proposed in [14]. ANFIS is a blend of Fuzzy Logic (FL) and Artificial Neural Network (ANN) that captures the strengths and offsets the limitations of both techniques for building Inference Systems (IS) with improved results and enhanced intelligence. Fuzzy logic is associated with the theory of fuzzy set, which relates to classes of objects with rough boundaries in which membership is a matter of degree. It is an extensive multivalued logical system that departs in concept and substance from the traditional multivalued logical systems. Much of fuzzy logic may be viewed as a platform for computing with words rather than numbers. The use of words for computing is closer to human intuition and exploits the tolerance for imprecision, thereby lowering the cost of the solution [15]. However, there are no known appropriate or well-established methods of defining rules and membership functions based on human knowledge and experience. Artificial Neural Networks are made up of simple processing elements operating concurrently. These elements model the biological nervous system, with the network functions predominantly determined by the connections between the elements. Neural Networks have the ability to learn from data by adjusting the values of the connections (weights) between the elements. Merging these two artificial intelligence paradigms together offers the learning power of neural networks and the knowledge representation of fuzzy logic for making inferences from observations.

A. Basic ANFIS Architecture

The ANFIS architecture described here is based on type 3 fuzzy inference system (other popular types are the type 1 and type 2). In the type 3 inference system, the Takagi and Sugeno's (TKS) if-then rules are used [16]. The output of each rule is obtained by adding a constant term to the linear combination of the input variables. Final output is then computed by taking the weighted average of each rule's output. The type 3 ANFIS architecture with two inputs (x and y) and one output, z, is shown in figure1.



Figure 1. Type 3 ANFIS Architecture

Assuming the rule base contains two first order TKS ifthen rules as follows:

Rule 1: if x is
$$A_1$$
 and y is B_1 , then $z_1 = p_1 x + q_1 y + r_1$
Rule 2: if x is A_2 and y is B_2 , then $z_2 = p_2 x + q_2 y + r_2$

The ANFIS structure is the functional equivalent of a supervised, feed-forward neural network with one input layer, three hidden layers and one output layer, whose functionality are as described below:

Layer 1: Every node in this layer is an adaptive layer that generates the membership grades of the input vectors. Usually, a bell-shaped (Gaussian) function with maximum equal to 1 and minimum equal to 0 is used for implementing the node function:

$$O_i^1 = \mu_{A_i}(x) = \frac{1}{1 + |(x - c_i)/a_i|^{2b_i}}$$
(12)

Where O_i^1 is the output of the ith node in the first layer, $\mu_{A_i}(x)$ is the membership function of input x in the linguistic variable A_i . The parameter set $\{a_i, b_i, c_i\}$ are responsible for are responsible for defining the shapes of the membership functions. These parameters are called premise parameters.

Layer 2: Each mode in this layer determines the firing strength of a rule by multiplying the membership functions associated with the rules. The nodes in this layer are fixed in nature. The firing strength of a particular rule (the output of a node) is given by:

$$w = O_i^2 = \mu_{A_i}(x) \cdot \mu_{B_i}(y), i = 1, 2, \dots$$
(13)

Any other T-norm operator that performs fuzzy AND operation can be used in this layer.

Layer 3: This layer consists of fixed nodes that are used to compute the ratio of the ith rule's firing strength to the total of all firing strengths:

$$\overline{w} = O_i^3 = \frac{w_i}{w_1 + w_2}, i = 1, 2, ...$$
 (14)

The outputs of this layer are otherwise known as normalized firing strength for convenience.

Layer 4: This is an adaptive layer with node function given by:

$$\overline{w}_i z_i = O_i^4 = \overline{w}_i (p_i x + q_i y + r_i) \tag{15}$$

This layer essentially computes the contribution of each rule to the overall output. It is defuzzification layer and provides output values resulting from the inference of rules. The parameters in this layer $\{p_i, q_i, r_i\}$ are known as consequent parameters.

Layer 5: There is only one fixed node in this layer. It computes the overall output as the summation of contribution from each rule:

$$\sum_{i} \overline{w}_{i} z_{i} = O_{i}^{5} = \sum_{i} \frac{w_{i} z_{i}}{\sum_{i} z_{i}}$$
(16)

B. Hybrid Learning Algorithm

The objective of learning is to tune all the adjustable parameters to make the ANFIS output match the desired data. In order to improve the training efficiency, a combination of learning algorithms is adopted to adjust the parameters of the input and output membership functions. The consequent parameters are optimized using the least square method with the antecedent parameters fixed. After updating the consequent parameters, the gradient descent method using back-propagation training algorithm is used to fine-tune the premise parameters. Assuming the premise parameters are held fixed, then the overall output of the ANFIS will be a linear combination of the consequent outputs given by:

$$z = \overline{w}_{1}z_{1} + \overline{w}_{2}z_{2}$$

= $\overline{w}_{1}(p_{1}x + q_{1}y + r_{1}) + \overline{w}_{2}(p_{2}x + q_{2}y + r_{2})$
= $(\overline{w}_{1}x)p_{1} + (\overline{w}_{1}y)q_{1} + (\overline{w}_{1})r_{1} + (\overline{w}_{2}x)p_{2} + (\overline{w}_{2}y)q_{2}$
+ r_{2} (17)

IV. DESIGN OF LOAD BALANCING INFERENCE SCHEME

In the first stage, the crisp variables, the virtual load of the source cell, the load fairness distribution index and number of unsatisfied users are converted into fuzzy (linguistic) variables in the fuzzification process. The fuzzification maps the three input variables to fuzzy labels of the fuzzy sets. Each linguistic variable has a corresponding membership function. A sigmoidal membership function (precisely, the product of two sigmoidal function) was used in this work. As there are three inputs and 4 fuzzified variables, the inference system has a set of 64 rules (figure 2).



Figure 2. Rule Viewer for the inference system

The neural network training helps select the appropriate rule to be fired.

Next, the rules are de-fuzzified to produce quantifiable results. Defuzzification can be achieved using several techniques such as maximum methods, center of gravity



Figure 3. ANFIS structure for Proposed Dynamic Load Balancing

method, center of singleton method etc. The center of gravity method is adopted for this work. The defuzzified output is further is then used to schedule resources or handover users to achieve a dynamic load balancing. The structure of the Model used is depicted in figure 3.

The model consists of 158 nodes, 64 rules, 256 linear parameters and 48 nonlinear parameters. The total number of parameter is very important in deciding the number of training data pairs required. In order to realize a good generalization capability, it is recommended to have the number of training data points to be many times larger than the number of parameters being evaluated [15]. 1326 input/output pairs of training data was used for training. Thus, the ratio between the data points and parameters is about four times (1326/304).

For parameter optimization, hybrid training (which combines least mean squares and back-propagation) was used. To ascertain how well the training data models the load balancing system, model validation was incorporated. Model validation involves presenting input/output data sets on which the inference system not trained to the inference system to check the degree to which the inference system model predicts the corresponding data set outputs values. For this work, model validation was accomplished using a checking data set of 1326 input/output pairs. The checking data helps prevent the potential of model over-fitting of the data. This is accomplished by selecting model parameters that correspond to the minimum

checking data model error. Fig.4 show training data tested on the checking data.



Figure 4. Checking data tested on training data

V. RESULTS AND DISCUSSION

The ANFIS system uses the hysteresis value for a QoS aware dynamic load balancing. The inference system increases the hysteresis as the virtual load of the cell increases. The virtual load has an overriding effect over the fairness distribution index in determining the result of the inference. When these two factors are the predominant input metrics, the relationship is illustrated in figure 5.



Figure 5. Hysterisis as a function of Fairness Index and Virtual Load

As the number of satisfied users increases, hysteresis value decreases (figure 6). Conversely, when the number of unsatisfied users in the network increases, the hysteresis value also increases to trigger. This results in sustaining or triggering load-balancing process.



Figure 6. Effect of using virtual load and number of satisfied users for ANFIS load balancing.

Figure 7 shows the effect of fairness index and the number of satisfied users in determining the value. The result reveals that the numbers of satisfied users have a more domineering effect over fairness index in determining load balancing.



Figure 7. Effect of using fairness index and number of satisfied users for ANFIS load balancing

VI. CONCLUSION

In summary, we have presented in this paper three key performance indicators for consideration in LTE dynamic load balancing i.e. the number of satisfied (dissatisfied) users, the fairness index and the virtual load of the source are three key performance indicators that can be used for dynamic load balancing in LTE and by extension to all SONs. This becomes especially important in the consideration of different network architectures [17]. As seen from the results, the number of satisfied (dissatisfied) users plays a more dominant role as the key performance indicator (KPI) especially where QoS is a major consideration. The virtual load of the cell is the next most important key performance indicator for fine-tuning the load balancing decision. Although the Fairness index did not reflect well in comparison to the other KPIs, it is also important especially where the load balancing in the network can be achieved by a more even (fairer) distribution of resource to users. In other words, the fairness index can be used as the KPI for deciding scheduling, while the virtual load and number of satisfied/unsatisfied users can be employed for handovers when and where necessary, thereby achieving a dynamic QoS-aware load balancing.

REFERENCES

- 3GPP TS 36.300 version, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2", Technical Specification, version 10.4.0 Release 10, June 2011
- [2] S. Sesia, I. Toufik and M. Baker, "LTE The UMTS Long Term Evolution: From Theory to Practice", 1st edition, John Wiley & Sons, Ltd., West Sussex, UK, 2009
- [3] A. Lobinger et al., "Load Balancing in Downlink LTE Self-Optimizing Networks", IEEE 71st VTC 2010, Taipei, Taiwan, June 2010.
- [4] H. Wang et al, "Dynamic Load Balancing in 3GPP LTE Multi-Cell Networks with Heterogenous services", ICST Conference, Beijing, 2010
- [5] H. Wang, "Dynamic Load Balancing and Throughput Optimization in 3GPP LTE Networks", IWCMC 2010, Caen, France, July, 2010
- [6] 3GPP TS 36.201 V9.1.0 (2010-03), "LTE Physical Layer: General Description."
- [7] WINNER Technical Report, "Assessment of advanced beamforming and MIMO technologies," WINNER, Tech. Rep. IST-2003-507581, 2005
- [8] Technical Specification Group RAN, "E-UTRA; LTE RF system scenarios," 3rd Generation Partnership Project (3GPP), Tech. Rep. TS 36.942, 2008-2009.
- [9] Technical Specification Group RAN, "Physical layer aspects for E-UTRA," 3rd Generation Partnership Project (3GPP), Tech. Rep. TS 25.814, 2006.
- [10] Technical Specification Group RAN, "E-UTRA; physical channels and modulation," 3rd Generation Partnership Project (3GPP), Tech. Rep. TS 36.211 Version 8.7.0, May 2009.
- [11] 3GPP, 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (EUTRA), "Radio Frequency (RF) system scenarios" (Release 8), Technical Report TR 36.942, available at http://www.3gpp.org.
- [12] R. Jain, D.M Chiu and W. Hawe, "A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Systems", Technical Report, Digital Equipment Corporation, DEC-TR-301, 1984.
- [13] L. Zhang, Y. Liu, M. Zhang, S. Jia, and X. Duan, " A Two-layer Mobility Load Balancing in LTE Self-Organization Networks" IEEE Internal Conference on Communication Technology, Beijing, China, 2011, pg. 925 – 929.
- [14] Jyh-Shing Roger Jang, "ANFIS: Adaptive Network-Based Fuzzy Inference System", IEEE trans., on Systems, Man and Cybernetics, Vol. 23, No. 3, May-June, 1993, pg. 665-685.
- [15] Mathwork Inc., "Fuzzy Logic ToolboxTM User Guide ", Ver., 2.2.14, Sep., 2011. Available from www.mathworks.com
- [16] T. Takagi and M. Sugeno, "Fuzzy Identification of Systems and its application to modelling and control," IEEE trans. on Systems, Man and Cybernetics, Vol. 15, pp.116-132, 1985
- [17] A.A. Atayero, M.K. Luka, M.K. Orya, J.O. Iruemi "3GPP Long Term Evolution: Architecture, Protocols and Interfaces", International Journal of Information and Communication Technology Research (IJICTR), Vol.1, №7, pp. 306–310, Nov. 2011.