Load Balancing in Heterogeneous Networks with LTE

Pedro Ganço and Luís M. Correia

Abstract—The main objective of this work was to find a method for balancing the load among the GSM, UMTS and LTE systems and for that purpose one defined metrics for load computation based on the number of fixed radio resources each system has, and then analysed in a time-based network simulator a model that moves traffic to balance the load. One also studied the impact of varying the user density, radio configuration, number of base stations and other relevant scenarios, and then designed a scenario that enables the performance analysis of the model developed for load balance. The model allowed to increase the performance in heterogeneous scenarios in overloading situations. The main results show some improvement in the overall QoS parameters, while increasing the average load for all RATs and increasing the average number of users connected per second. The main results in terms of QoS are the reduction of drop rate calls from 2.3% to 1.2% and the reduction of blocking probability from 15.2% to 3.4% measured for the same scenario. The model developed for this work may be used as a general framework for load balancing in heterogeneous networks.

Index Terms — Heterogeneous Networks, Load Balancing, Radio Resource Management, LTE.

I. INTRODUCTION

ROUGHLY every 10 years a great technology shift occurs in the wireless communications world introducing new systems and ways of communication. In Portugal, mobile devices penetration now exceeds more than one device per person, and the trend is to find many radio technologies working simultaneously in such devices like GSM/GPRS, UMTS/HSDPA, LTE, WiFi and Bluetooth. This wireless world is a heterogeneous world and operators exploit at their best the functionalities any given technology provides. GSM with its wide spread coverage and many years of settlement provides all the basic needs for reliable voice calls in almost any part of the country. UMTS has been going on for more than 10 years now, and is a reasonable system in terms of coverage, serving current traffic demands, although the appearance of smartphones and new cloud services are pushing demand to higher bitrates, only feasible with new technologies, like LTE.

With all this radio access technologies coexisting in the same space and at the same time, usually owned by the same telecom operator, naturally, the trend is to integrate the communication among all the systems in the most efficient way, transparently to the end user. This work exploits this need for networks’ efficiency through load balancing in heterogeneous scenarios with GSM, UMTS and LTE. Henceforth, its main objective consisted of researching a technique to measure load in a heterogeneous scenario and to develop a model capable of balancing load among GSM, UMTS and LTE. 3G and 4G systems will reinforce their market penetration, hence, it is crucial for mobile operators to balance the load among systems, and efficiently use their network resources. According to [1], the trend to offload traffic among different systems will increase. This is due to the increasing mobile capabilities for connection simultaneously to more than one radio technology.

In face of the presented exploding traffic challenges and heterogeneous environment, where many different systems co-exist, it is the operator’s best interest to be able to balance the load among existing systems.

The scope of this work is to provide a model that enables load balancing among GSM, UMTS and LTE, taking quality of service into account. The focus is only on the mentioned three systems: GSM (2G), UMTS (3G) and LTE (4G), since they represent a huge portion of the total mobile users among the 3 generations deployed around the world.

This work aims to bring some insights on ways to measure load in a heterogeneous radio environment with GSM, UMTS and LTE, as well as to provide some practical ways of sharing load among these different RATs (Radio Access Technology) for the purpose of service quality improvement.

The paper is organized as follows. Section II contains the essential aspects of GSM, UMTS and LTE, as well as information on load balance and its state of the art. The models developed in this work can be found in Section III, and in Section IV the scenarios to test the model and results drawn from such models are presented. Finally, conclusions drawn from this work are presented in Section V, along with a few remarks on future work.

II. BASIC ASPECTS OF LOAD BALANCING

Load metrics differ from system to system, and depend on the number of users in a given cell, services being provided, quality of service and many other different parameters. Capacity per system is a very important variable in load balancing. In GSM, capacity depends mainly on the number RF Transceivers (TRX) installed in the Base Station (BS), since each user is assigned its own timeslot (at full-rate). As for UMTS, capacity strongly depends on the interference level among users, since all UEs (User Equipment) are operating at the same frequency due to CDMA. In an LTE cell, the total number of users depends on the total bandwidth available, the average bandwidth of the Resource Blocks (RB) allocated per
user, and capacity gains driven by users positioning and diversity.

In order to balance the load in a heterogeneous network, one can employ many approaches, although, in the scope of this work, the focus is on balancing the load by periodic Handovers (HO) of load and by Call Admission Control (CAC). The former is responsible by applying methodologies of Vertical Handover (VHO) or Horizontal Handover (HHO) between GSM, UMTS and LTE in order to harmonise the networks current load sharing, and the latter by the analogy for accepting or denying every new incoming call/session request and HO of ongoing calls.

There are plenty of different ways suggested in the literature to implement load balancing via VHO. In [2], Ferrús et al. suggest the use of parameters such as service type, network conditions, operators policies, user preferences and signal level for a VHO decision. In [3], the authors suggest using a combination of signal level, service delay sensitivity, service financial cost, and mobile conditions to trigger VHO. Other authors, such as [4] and [5], give much importance to signal quality by triggering the VHO due to signal level, signal to interference ratio and BER (Bit Error Ratio) measurements. More focused on energy efficiency, in [6] the selection of Radio Access Technology (RAT) is the one leading to lower energy consumption to increase Mobile Terminal (MT) battery life while respecting Quality-of-Service (QoS); when considering energy efficiency from the operators’ viewpoint, in [7] a wide set of saving techniques are proposed to reduce networks infrastructure power consumption.

In the current work, the triggering point for a HO in a heterogeneous networks is based on the idea that, once a given load threshold is reached, some of the traffic is handed over to a legacy system to free radio resources.

Chiu et al. [8] introduced for the first time a balance index to measure the balance of resources in a system, given by:

$$\xi_1 = \frac{\left(\sum_i \rho_i \right)^2}{n \sum_i \rho_i^2}$$

(II.1)

where $n$ is the number of neighbouring BSs over which the load can be distributed and $\rho_i$ represents the load of BS $i$.

This balance index quantifies the balance among neighbouring BSs and equals 1 when all BSs have the same load and tends to $1/n$ when the load is severely unbalanced, therefore, its target is to be maximised. To define a BS load-state and define such a threshold parameter, $\delta$, for a VHO decision, the definition of average load is given by:

$$\bar{\rho} = \frac{\sum_i \rho_i}{n}$$

(II.2)

where the BS load state is then 1) under-loaded: when the BS load is below $\bar{\rho}$; 2) balanced: when the BS exceeds the load average by less than $\delta$; and 3) overloaded: when the BS load exceeds $\bar{\rho}$ by $\delta$.

This so called “load” metric represents the occupation ratio of a BS, and can be calculated in different manners depending on the system, therefore, the load value of two different networks may not represent the same load situation. Some common methods for load computation are based on interference, [9], or throughput, [10], though, Nguyen-Vuong et al. [11] identified some limitations to this balance index and load metrics since a given user may generate different added loads depending on the BS it is connected to. Concerning such limitations, [11] proposes a new algorithm to deal with LB in heterogeneous packet networks. First, a general load metric has been defined as the ratio of the required resources to the total ones, to hide the resources heterogeneity among different networks (see Section III for different systems load computation), thereafter, a new balance index was defined as:

$$\xi_2 = \sum_{i} \max(\rho_i - \delta, 0)$$

(II.3)

The objective of the new LB algorithm is now to minimise $\xi_2$ until it is equal to zero or there is no further improvements to be made. The LB algorithm proposed is employed to prevent overload situations (by admitting or rejecting new communications) and employed also in managing ongoing communications (by forcing HO in imminent overload situations). A connection request to a specific BS is only accepted if the BS’s load, including the contribution of the incoming communication, is below an admission threshold, otherwise, it is redirected to the least loaded overlapped access network.

The connection is rejected if there is no BS available for connection in the coverage area, unless it is a HO. Handovers are always accepted and this HO enforcement is done in a two-step process: in the first step, a move ($I, J$) is identified to move a mobile user $M_i$ from an overloaded $BS_0$ to a suitable $BS_j$. If the network is still overloaded, the algorithm runs for a two-move operation moving a user $M_i$ from $BS_0$ to $BS_j$ and then move a user $M_l$ from $BS_j$ to $BS_k$ until it can no longer improve the balance index $\xi_2$. Fig. 1 represents the algorithm explained above, where $wij$ is the load contribution of user $M_i$ at $BS_j$ while $M_i$ connects to $BS_j$.

This work takes all the presented state of the art into consideration in Section III, although one may outline the suggested algorithm of Nguyen-Vuong et al. [11] in the LB for VHO and HHO, whereas for the CAC, one based the ideas on [12] and evaluated the performance of such developed models in Section III in heterogeneous environments containing all GSM, UMTS and LTE.

III. MODELS AND ALGORITHMS

A. Propagation, Traffic Generation and User Mobility Models

The propagation model considered in this work is applied by taking urban scenarios as a reference, since it is in such
scenarios that LTE is being mainly deployed, therefore, one considers the COST 231 Walfish-Ikegami Model for the path loss calculation for all RATs. It is important to note that GSM operates at the 900 MHz band, UMTS at 2100 MHz and LTE at 2600 MHz, and since the validity range of the COST 231 Walfish-Ikegami Model goes from 800 MHz to 2000 MHz the implementation of the model introduces some propagation measurement errors. The buildings’ height, streets’ width and all the remaining parameters of the model are within their validity range and further characterised in the reference scenario on Section IV.

Regarding the models for traffic generation, one takes advantage of some of the work developed by Serrador [12], who explains the models used for traffic generation for voice services, web browsing, video calls, e-mail, file sharing and music streaming services. On top of those, the present work contributes to the improvement of the existing simulator by adding new services, namely, Machine-to-Machine (M2M) communication, whose traffic models were adapted from the services already implemented, in order to add eHealth, smart meters, surveillance and domotics services. These new services were mapped onto the traffic models developed by [12] according to their service classes and characteristics.

In order to reproduce MTs’ behaviour, one implemented the Random Walk Mobility Model developed by [Serr12], which gives memory-less mobility patterns, as each step is calculated without any information from the previous one. At regular time intervals, both the direction (uniform distributed from 0º to 360º) and the speed of the MTs are updated. MTs’ speed is calculated using the Radio Resource Allocation algorithm implemented in the simulator. The speed of the MTs is then used to calculate the new position of the MT in each time interval, which is updated at each time step.

### B. Load Metrics for Heterogeneous RATs

The great challenge in balancing the load among heterogeneous RATs is in finding a common load measurement for all RATs that take the system’s differences in allocating resources into account. Consider for instance one overloaded UMTS BS transferring some load to a neighbouring LTE one. The UMTS BS’s release of MTs, via VHO, implying, e.g., a 20% reduction of its load would not necessarily imply an equivalent 20% increase on the LTE BS’s load, since the radio resource allocation in UMTS and LTE differ, as seen in Section II. Therefore, one must first define what load is and afterwards define a common metric to exchange load from users and services among BSs, regardless of their RAT.

The maximum number of users in a BS is a good starting point to obtain a threshold for the maximum number of users allocated to a given BS, but its value is hard to compare among different RATs, since it may depend on the interference levels, SNR and other highly variable (and hard to measure) parameters. Load could also be defined in terms of the BS’s maximum throughput, power budget and many other ways, as shown in Section II, but for the purpose of performing load balance in a heterogeneous RAN, one has taken the approach of finding a common metric for load focusing only on the radio resource allocation.

Henceforth, to compute the load of a BS, one defines a generic Radio Resource Unit (RRU) that represents the amount of radio resources a BS has to connect to MTs. The load of a BS, regardless of its RAT, can now be defined as follows:

\[
\rho[\%] = \left(1 - \frac{N_{RRU_{a}}}{N_{RRU_{t}}}\right)100 \quad (III.1)
\]

where \(N_{RRU_{a}}\) is the number of RRU available for traffic and \(N_{RRU_{t}}\) is the total number of RRU.

The concept of RRU hides the RATs differences and in order to quantify the total number of RRUs a given BS has, one must understand the fixed amount of radio resources any technology has to communicate and the way resources are consumed by users.

In a TDMA/FDMA system, like GSM, where one considers only voice, load consists mainly on the number of users one can accommodate for phone call services or SMS. The allocation of resources is done in periods of 4.61 ms of Timeslots (TS), where any given radio channel at full-rate transmission has capacity for 8 different users. GSM900’s DL band ranges from 935 MHz to 960 MHz, therefore a separation of 200 kHz per channel gives us a theoretical maximum of 125 frequency channels, although, one should consider that more than one telecom operator uses the spectrum and that not all radio channels can be used just for traffic purposes.

A more realistic approach to reality may consider the attribution of 10 radio channels for traffic purposes allocated to a single telecom operator, each carrying 8 TS, imposing a maximum of 80 radio resources per Time Transmission Interval (TTI) in a GSM BS, obtained by the following computation of total radio resources in GSM:

\[
N_{RRU_{GSM}} = N_{TS} N_{fc} \quad (III.2)
\]

where \(N_{TS}\) is the number of users allocated to TSs during a TTI (Time Transmission Interval) in a radio channel (which equals to 8 for a full-rate transmission) and \(N_{fc}\) is the number of radio channels available for traffic.

The radio resources (timeslots) capacity can be taken from [13], where, for example, for voice calls, the data rate is 22.8 kbps.

Regarding UMTS’s BSs, for the purpose of this work one will only consider the more recent radio access technology release, i.e., HSDPA. In this system, channel multiplexing is done in the time domain, where each TTI with 2 ms of duration can carry 480 symbols. Within each TTI, a maximum of 15 parallel codes per carrier can be assigned to one user (or shared among several) for traffic usage. This variable number of codes per carrier assigned to a user imposes the total number of RRU in UMTS:

\[
N_{RRU_{UTMS}} = N_{co} \times N_{ca} \quad (III.3)
\]

where \(N_{co}\) is the number of codes and \(N_{ca}\) is the number of carriers (typically, there are 3 carriers per BS).

Anytime a given MT is using a service in HSDPA, the number of RRUs this service consumes can be calculated as follows:

\[
N_{RRU_{UTMS}} = \frac{B_{srv}}{N_{sym} \times M_{code}} \quad (III.4)
\]

Where \(B_{srv}\) is the service bitrate, \(N_{sym}\) is the number of symbols transmitted in a second (480 symbols in a TTI,
correspond to 240 000 symbols in a second), and \( M_{\text{code}} \) is the modulation and coding rate in use (e.g., a modulation of 16-QAM, and unitary coding rate, generates 4 bits per symbol).

In an OFDM system such as LTE, RBs are sent every 0.5 ms and its total number depends on the bandwidth, e.g., for a bandwidth of 20 MHz one has 100 RB. Note however that the TTI is 1 ms, hence, the available RBs per 1 ms is 200. The amount of RBs can be used as the measurement for the total number of RRU in LTE, \( N_{\text{RRU,LTE}} \).

An LTE BS generates 7 OFDM symbols every 0.5 ms in the time-domain and 12 sub-carriers (1 RB) in the frequency domain, therefore, the required resources a given service demands in LTE is given by:

\[
N_{\text{RRU},\text{LTE}} = \left[ \frac{B_{\text{srv}}[\text{bps}]}{N_{\text{RB}} \times N_{\text{sym}}[\text{symb/s}] \times M_{\text{code}}[\text{bits/symb}]} \right] \tag{III.5}
\]

where \( N_{\text{RB}} \) is the number of RB.

Not all RRUs are used for traffic. Some are used for signalling, e.g., therefore to compute the number of RRU available for traffic, \( N_{\text{RRU}a} \), one should consider that for UMTS, a maximum of 15 out of 16 codes can be assigned for traffic. The available RRU can be obtained by:

\[
N_{\text{RRU}a,\text{UMTS}} = (N_{\text{CO}} - 1) \times N_{\text{c}}
\]

\[
= \sum_{i}^{N_{\text{USR}}} \left[ \frac{B_{\text{srv,i}}[\text{bps}]}{N_{\text{sym}}[\text{symb/s}] \times M_{\text{code},i}[\text{bits/symb}]} \right] \tag{III.6}
\]

where \( N_{\text{USR}} \) is the number of users connected to the BS, \( B_{\text{srv,i}} \) is the bitrate of the i-th user connected to the BS, and \( M_{\text{code},i} \) is the modulation coding rate of the i-th user connected to the BS. The summation parcel consists of adding the number of RRU's required from all users connected to the UMTS BS. As for GSM and LTE, one will reserve 10% of the total resources for signalling; therefore the available RRU can be defined by:

\[
N_{\text{RRU}a} = N_{\text{RRU}} \times 0.9 - \sum_{i}^{N_{\text{USR}}} N_{\text{RRU},i} \tag{III.7}
\]

where \( N_{\text{RRU},i} \) is the number of required resources by the i-th user connected to the BS.

The load of any BS can now finally be computed by (III.1), replacing the total number of RRRUs, \( N_{\text{RRU}} \), and the number of available RRUs, \( N_{\text{RRU}a} \), accordingly to its RAT.

The load metrics developed in this section make it possible to measure the impact of transferring load from one BS to another, via VHO, taking the differences of the RATs into consideration. Based on this framework, one can develop a load balance algorithm to be implemented in a heterogeneous radio access network.

C. Load Balance Algorithm

In order to implement a load balancing in a Heterogeneous Network (HN) environment, one can take many different approaches as mentioned previously. In this work, load balance is studied in order to optimise network capacity and allocate the maximum number of users while respecting their QoS requirements. Hence, due to the dynamic load state of a heterogeneous network, there are two main moments when it may be required to intervene: when a BS reaches its load threshold and when a new call/session request arrives. For the sake of simplicity, from now on, the term “call” will be used as any connection request from a MT to a BS, whether it refers to a voice call in CS, or a session request in PS. Furthermore, “load” is the one defined in (III.1).

The algorithm receives as input information all the BSs statuses in terms of load and coverage, and the different services priority list for every RAN, previously predefined. As a result, the new incoming call or HO request, if any solution is available, is routed to one of the available RANs: GSM, UMTS or LTE.

When the LB algorithm is triggered by a new call, it runs the CAC process; if it comes from a HO request from an overloaded BS it runs the Forced Handover (FHO) routine. The former process, described in Fig. 2, starts by sorting the reachable BSs by the MT originating the call using the Service Priority List mentioned above. Then, it checks if any BS may accept the connection without exceeding its threshold load, otherwise it runs the FHO Routine to make any BS available.

The CAC algorithm ends by establishing the new connection or blocking the call/delaying it in case of unavailable resources.

![Fig. 2. Call Admission Control Algorithm](image)

A BS under the coverage of an MT is able to accept a new call only on two conditions. First, in order to receive a new call, the targeted BS has to calculate the required number of RRU's incoming service demands (represented by index “inc”) and verify if they do not surpass the number of available RRUs, as stated in:

\[
N_{\text{RRU},\text{inc}} < N_{\text{RRU}} \tag{III.8}
\]

Thereafter, if the number of required RRUs is lower than the available ones, the new load calculated for the targeted BS to receive the new call, \( \rho' \), must not exceed a given load threshold, \( \rho_{\text{thr}} \):

\[
\rho' = \left(1 - \frac{N_{\text{RRU}} - N_{\text{RRU},\text{inc}}}{N_{\text{RRU}}} \right) \times 100 < \rho_{\text{thr}} \tag{III.9}
\]

If both conditions are true, the BS is able to receive the call. When the LB Algorithm is triggered by a HO request, it means that one or more BSs have reached their load thresholds. In this case, the LB Algorithm uses the FHO
Routine shown in Fig. 3. The forced HO algorithm starts by ordering the BSs with the most overloaded ones on top of the list, unless the routine has been used by the CAC algorithm, where it gives priority to the BSs reachable by the MT originating the new call. Then, it follows a procedure similar to the one described in [11], where the network, by changing MTs allocation (either by HHO or VHO), tries to ease the load of the overloaded BSs. The re-allocation of MTs stops when there are no more BSs an MT can be allocated to or when there are no more overloaded BSs.

Fig. 3. Forced Handover Algorithm

D. Practical Simulation Aspects

The Common Radio Resource Management (CRRM) Simulator used to obtain the results of the present thesis is a system level, time-based simulator with a 10 ms resolution, developed by Serrador [12] over the Microsoft Visual Studio 2005 platform. The tool implements all fundamental systems functionalities, like power control, link control, basic channel code management, radio bearer service, load control, access control, propagation estimation and interference estimation and generation. Furthermore, the simulator is also capable of generating users and heterogeneous traffic services. The original version of this simulator only considered two different types of RATs, i.e., UMTS (both releases 99 and 5) and WiFi, hence, many modifications were required for the purpose of this thesis. The WiFi module was disabled during simulation time and two new dedicated modules were developed to add GSM and LTE, as represented in the CRRM Simulator block diagram in Fig. 4. The CRRM Algorithm and Policies Engine (mainly regarding HO methods and load status computation) were also re-designed according to the algorithm and models defined in the present work. The CRRM Simulator can be divided into three main blocks, identified by the green, yellow and blue areas of Fig. 4. The green block represents all the input information needed to run the simulator for a given scenario; the yellow block contains the RRM functionalities of the program and it is where the new models and algorithms were implemented; the blue block is where the simulator calculates and then writes the outputs into a file.

The development of the GSM and LTE modules required a huge effort, since they required a structural change in the original simulator. Although, most of the radio features could be adapted from the existing RATs, they were spread in many parts of the simulator. The simulator consists mainly of 53 text files written in C++, 55 header files and some more auxiliary ones. All of the three blocks mentioned above had to be changed, but the main developments are highlighted in grey in Fig. 4. The models previously described were translated into code at the CRRM Algorithms and Policies Engine block and the new GSM and LTE modules were written from scratch, adapting the code from the existing UMTS RAT and also by adapting the implementations done by Venes’s contributions to the simulator [14]. Venes has improved the simulator to add WiMAX. All RATs implement a COST231 Walfish-Ikegami propagation model and noise and interference estimation. Many minor routines had to be implemented or adapted from the original simulator, e.g., the one represented in Fig. 5 to check if there are enough RRU modules in a BS whenever a user is trying to connect.

Due to space constraints, the expanded description of the simulator and its parameters is left for consult in the full thesis [15].

IV. ANALYSIS OF RESULTS

A. Reference Scenario

The Reference Scenario (REF) to evaluate the load balance performance of the models previously described in Section III is an urban environment supporting multiple RATs, i.e., OFDMA (LTE), CDMA (UMTS), and FDMA/TDMA (GSM). It was designed in such a way that it allows the simulation of a variety of different (sometimes unrelated) parameters and yet
being possible to compare the load impact among those scenarios. Therefore, REF considers only 3 co-located BSs, 1 per RAN, Fig. 6. This approach avoids VHOs that would mask the real impact on the load of the BSs and also focus the traffic within the coverage area around the site.

The services distribution among the three systems deployed obeys a criteria for RAN selection. On Table I it is represented this criteria where “1” means the preferred system for a given service to be steered into, and “3” the least one. Taking also into consideration that the simulator works on 10 ms time basis, it has been necessary to normalize the available radio resources for each RAT accordingly. On Table II one can see the normalization used to obtain the number of RRUs per 10 ms. More details about the REF scenario characterization can be found at [15], whose results obtained through simulation can be found at Table III.

### B. Load Performance Analysis

In order to test the performance of the load balance algorithm in a heterogeneous network, many scenarios have been simulated. The present section studies the impact of user and network parameters to draw conclusions on the load balance effectiveness in improving the performance of the system.

1) Impact of Users Density

The impact of the number of users on the load was tested for REF just varying the users’ density per km$^2$. The simulation plotted in Fig. 7 and Fig. 8 shows a slight increase of load for both UMTS and LTE. The reason for such low load increases is due to more than half of the traffic being for voice calls, that is being steered to GSM BSs, and because, for the remaining packet services, LTE does not suffer a great impact on resource allocation with a 10 MHz bandwidth where RRUs for LTE are abundant for the current traffic demands. For the GSM BS, on the other hand, one can observe a high correlation between users’ density and load state. For instance, the increased variation from 3 000 users/km$^2$ to 6 000 users/km$^2$ represents a 25% increase in the load for GSM. It is also notably that for REF, an average of 34 users connected to the GSM BS are consuming 52% of the BS’s total resources.

The main conclusions drawn from the analysis of the variation of the number of users is that it hardly affects the availability of UMTS and LTE, due to the amount of RRUs these two systems have for traffic. Thereafter, it is necessary also to study the impact of changing the radio configuration parameters of LTE to be possible to analyse the impact on load. The following section studies this in further detail.
2) **Impact of Radio Parameters**

In order to test the impact of LTE radio parameters, one has simulated the same heterogeneous reference scenario with a single GSM, UMTS and LTE BSs, varying just LTE’s bandwidth for 5 MHz, 10 MHz, 15 MHz and 20 MHz. The results show little or no impact on the average number of connected users in all 3 RATs. Once the user density is defined, the number of connected users per system is always the same, apart from a small variation due to randomness of the simulator traffic/users generation. The bandwidth variation of the LTE BS only impacts on LTE BS’s load as plotted in Fig. 10 from the results in Table IV. Based on the information provided in Section III, one can preview the number of RRU s available in LTE for each bandwidth scenario and consequently understand the great impact it creates by changing LTE’s bandwidth. Bandwidth allocation is directly related to the number of RBs generated in LTE leading to the total RRU s in Table V, recalling that the TTI in LTE is 1 ms and that the simulator time resolution is 10 times greater.

Considering (III.5), knowing that in a normal CP there are 7 OFDM symbols per RB, and assuming a constant modulation coding rate of 4 bits per symbol, the number of required resources any service demands is given by:

$$N_{RRU_{LTE}} = \left\lceil \frac{B_{srv} \text{ [bits per 10ms]}}{N_{RB_{[per 10ms]}} \times 28 \text{ [bits]}} \right\rceil$$  \hspace{1cm} (IV.1)

Before applying the equation, it is necessary to consider that the total number of RB, $N_{RB}$, should be multiplied by 0.90 to reserve 10% of resources for signalling purposes. The number of required resources by each service has a larger impact on the load state of the LTE BS the lower the bandwidth is. For instance, for the 1.4 MHz bandwidth scenario, it is only possible to serve 15 users simultaneously in a file sharing service.

Other radio parameters, such as the number of radio channels allocated for communications in GSM or the number of codes given in UMTS, were kept unchanged during simulation, but the results also show how such slight changes in the radio parameters cause drastic changes in the load state of BSs.

### TABLE IV

<table>
<thead>
<tr>
<th>Bandwidth [MHz]</th>
<th>Simulation Output</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>Average</td>
<td></td>
</tr>
<tr>
<td>Users per second</td>
<td>11.8</td>
</tr>
<tr>
<td>Average Load, $\rho$ [%]</td>
<td>26.0</td>
</tr>
</tbody>
</table>

### TABLE V

<table>
<thead>
<tr>
<th>Bandwidth [MHz]</th>
<th>RRU s variation with bandwidth.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RB [per 0.5ms]</td>
</tr>
<tr>
<td>-----------------</td>
<td>---------------</td>
</tr>
<tr>
<td>1.4</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>10</td>
<td>50</td>
</tr>
<tr>
<td>15</td>
<td>75</td>
</tr>
<tr>
<td>20</td>
<td>100</td>
</tr>
</tbody>
</table>

**Fig. 7** Impact of users’ density on the average number of connected users.

**Fig. 8** Impact of users’ density on average load.

**Fig. 9** Impact of users’ density on blocking probability.

**Fig. 10** Impact of LTE bandwidth variation on the BS’s load.
3) Variation of Number of Base Stations

For this case, two situations were studied: the impact of adding another site with all 3 RATs collocated (named “2 sites”) and the scenario where only one LTE BS is added to REF (“1 site + LTE”). The new scenarios were designed in such a way that BSs are all within the same coverage area to infer on the impact of load distribution among the different RATs.

The results, plotted in Fig. 11 and Fig. 12 suggest a general increase in the number of connected users and a general decrease in the load state of the BSs. This would be expected, since the addition of a BS provides more coverage area, increasing the number of users, and more availability of resources for traffic cutting the average load of a given RAT. Nonetheless, some results differ from this description. When one single LTE BS is added, for instance, there is a decrease in the number of connected users in GSM and UMTS. This result is rather interesting in terms of easing other systems and virtual sharing the load even without applying any load balancing model.

With an increased LTE coverage, more users try to connect to this RAT (the REF’s priority table for packet services steer most of the traffic to LTE), thereby leading to less users connecting to GSM, and consequently leaving more free resources in UMTS to accept voice and video calls, easing in the GSM load state.

C. Load Balance Model Performance

The analysis from previous sections suggests that a slight variation of radio resources has a large impact on the average system load. One can also infer that a great number of users will easily saturate the network’s capacity, hence, in order to evaluate the load balance developed models, the scenario summarised in Table VI was first simulated without load balance routine active in order to draw conclusions. The load balancing testing scenario consists of the collocation of the 3 RATs in two sites within an urban scenario of 10,000 users/km². The scenarios analysed so far suggest that the BSs hardly reach half of its load capacity with the current setup, so the radio resources had to be reduced. For this purpose, the number of carriers in UMTS has been reduced from 3 to 2, cutting the available RRUs for traffic from 225 to 145; in LTE, instead of considering a 10 MHz bandwidth allocation, now one considers only 5 MHz, reducing the number of RRUs available for traffic to only 450, instead of 900. The results from previous sections also highlight the large consumption of resources in GSM due to the service penetration of voice calls, so, in order to reduce the user’s allocation in GSM to better analyse the model, the service penetration was also redefined. The obtained results, although increasing the expected load state per BS, do not reach the defined load threshold of 70%, therefore showing little or no improvement in the overall performance of the load balance model developed. Nonetheless, the blocking probability performance parameter is reduced from 0.6% to null. This result shows some improvement with the LB model activated, since the CAC algorithm designed is now capable of finding more BS within reach (regardless of its RAT) increasing the odds of successful connections before blocking an incoming call.

Nonetheless, the blocking probability improvement is not that interesting, considering that the scenario without LB active registers a blocking probability below the reference QoS value of 1%, so in order to test the efficiency of the model developed, one must reach the load threshold to infer the behaviour of the BSs in the situation of resources scarcity. Such scenario could be designed by increasing the user density around the site, recreating, for example, a music festival event in a real life network, but such scenario would demand a great amount of time and a huge computational effort. In order to avoid such scenario, the high load demand has been virtually induced in all BSs by changing their load threshold to 25%. The results are shown in Figs. 13, 14, 15, 16 and 17., whose scenarios description for LB and “NoLB” refer to the scenario described above and the scenario “Threshold 25% LB” and “Threshold 25% NoLB” refer to the same scenario just considering a 25% threshold value.

The results for a threshold of 25% show that, without LB, some voice users, after the GSM BSs have reached their load threshold, try to connect to UMTS, increasing both the average number of users connected to UMTS and increasing the UMTS BSs load state, until they also reach the load threshold.

LTE BSs are not affected in this scenario by the load threshold in case there is no LB algorithm running. On the other hand, one observes that, with LB, LTE increases its number of connected users, almost reaching the threshold. As for UMTS, the activation of the LB routine saturates all of its RRUs changing the load state from 20% to 25%, increasing the average number of connected users by 8. Moreover, without LB, GSM blocks most of the calls in the admission control, which is why the blocking probability is 15.2%, against the LB scenario where this parameter is 3.4%. Regarding the delay parameter, all 4 scenarios register a value much lower than 30ms, which is much lower than the expected one, since the overload conditions predict higher delays in the transfer of packets. Regarding the drop rate, LB presents an effective
improvement in the overall RAN reducing it from 2.3% to 1.2%.
From these results, one may conclude that the model only has a notorious impact for an overloaded network. Only for the case where a 25% load threshold has been established, the model starts to react to the situation moving MTs from one BS to another, taking advantage of HHO and VHO to better distribute the traffic.

Fig. 14 shows how the LB routine increases the average load in the network for all RATs, while increasing the number of users being served and providing a slight improvement in the analysed QoS parameters.

### TABLE VI
Parameters changed from REF to test the load balance model.

<table>
<thead>
<tr>
<th>Simulator Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>User density [users/km²]</td>
<td>10 000</td>
</tr>
<tr>
<td>Number of carriers in UMTS</td>
<td>2</td>
</tr>
<tr>
<td>LTE’s bandwidth [MHz]</td>
<td>5</td>
</tr>
<tr>
<td>Voice</td>
<td>35.0</td>
</tr>
<tr>
<td>WWW</td>
<td>15.0</td>
</tr>
<tr>
<td>Video</td>
<td>3.6</td>
</tr>
<tr>
<td>E-mail</td>
<td>15.3</td>
</tr>
<tr>
<td>Streaming</td>
<td>18.0</td>
</tr>
<tr>
<td>FTP</td>
<td>8.1</td>
</tr>
<tr>
<td>Smart Met.</td>
<td>1.7</td>
</tr>
<tr>
<td>eHealth</td>
<td>0.9</td>
</tr>
<tr>
<td>Domotics</td>
<td>0.4</td>
</tr>
<tr>
<td>Surveillance</td>
<td>2.0</td>
</tr>
</tbody>
</table>

**V. CONCLUSIONS**

The approach to achieve the objective of this work can be split into two parts: 1) common load metrics framework; 2) models and algorithms development to better distribute load among the three RATs. At first, one had to find a common framework to compute load, regardless of the system, i.e., a given service may be overloading a UMTS base station, but its load may be better allocated to an LTE one, due to the different radio resources allocation scheme. Hence, some radio resources units were established to measure the impact of services in the base stations’ load. The simulation tool, SimCell, was developed upon previous versions from Serrador [12] and Venes [14]. A great effort was put in the simulator development to add two new systems (GSM and LTE), as well as making an overall overview on the main functionalities of the simulator to cope with these new systems. In the verge of the Internet of Things era, it would be unwise not to consider the impact of some new services already starting to be deployed; therefore, M2M services were also implemented and added to the existing ones. The simulator is now capable of generating traffic for GSM, UMTS and LTE where users
can seamlessly perform vertical handovers. The simulator has traffic models for 10 types of services, namely, voice calls, web browsing, video calls, e-mail, music streaming, file sharing, smart meters, eHealth, domotics and surveillance. Due to the scope of the thesis, the modules regarding WiFi and WiMAX from previous versions were not further developed, in order to comply with same load balancing routines created for GSM, UMTS and LTE, and were left out from the simulator’s development. Nonetheless, it will be possible, without much effort, to upgrade the simulator to be fully operational for GSM/GPRS, UMTS/HSDPA, LTE, WiFi, and WiMAX.

The analysis has shown the influence of many parameters in load within a heterogeneous scenario. Results show that the increase of the number of users in the scenario has a linear influence in the load state of base stations. From simulation, one concludes that changing LTE’s bandwidth has the biggest impact on the load state. By reducing the bandwidth from 10 MHz to 5 MHz, the load in the LTE base station increases by 12.3%; the theoretical analysis supports this idea, showing how a low bandwidth allocation for LTE leads to higher load states and fewer users allowed for connection. The results show, at first, little or no impact in the usage of the load balancing model, due to the abundance of RRUs in the scenario: the developed model is useful only in overloading situations.

In order to increase load in the scenario, one could increase the density of users and traffic demands, which would require a huge computational effort. Due to time constraints, such heavy load scenarios could not be tested, thereby, one was forced to lower the load threshold for all base stations to infer on the load balance algorithm. After decreasing the load threshold from 70% to 25% the LB algorithm has shown some improvement in the overall QoS parameters analysed. The average load for all RATs has increased, while increasing the average number of users connected per second. The main results in terms of QoS are the reduction of drop rate calls from 2.3% to 1.2% and the reduction of blocking probability from 15.2% to 3.4% measured for the same scenario. Note that such load threshold would not be realist in a real life network for the scenario tested, but it is a valid way to test the algorithm performance and to recreate a heavy load scenario like a music festival event.

It is important to take into account that, despite the effort to recreate a real wireless communication system, the world is far too complex, and many simplifications had to be introduced. The simulator performs most of the functionalities a radio management systems demands, but it does not strictly follow the existing communication protocols.

Despite all simplifications, the obtained results provide some insights and basic notions on how to deal with the load balancing challenge.

As suggestions for future work regarding the topic, it might be interesting to add some cost benefits on implementing this solution. The first version of Serrador’s simulator considered such cost functions and it would have been an interesting subject to continue, this time, applied with the newly developed LTE and GSM modules.

REFERENCES