On the Suitability of Load Balancing Principles in Heterogeneous Wireless Access Networks

X. Gelabert, J. Pérez-Romero, O. Sallent, R. Agustí

Departament de Teoria del Senyal i Comunicacions, Universitat Politècnica de Catalunya (UPC)
c/ Jordi Girona, 1-3, 08034, Barcelona, Spain.

[xavier.gelabert, jorperez, sallent, ramon]@tsc.upc.edu, http://www.gcr.tsc.upc.edu/

Abstract— **Initial RAT selection is a key Common RRM strategy, where users requiring service are to be efficiently allocated in the existing and available RATs. Although load balancing is a possible guiding allocation principle, sometimes it may not be convenient. This paper evaluates and compares a load balancing and a service-class RAT selection policy in order to discuss the suitability of the former in situations where different service-class mixings are present. Results indicate a tight dependency between this appropriateness and the mixing of demanding services.**

Key words: Common Radio Resource Management, Radio Access Technology selection, load balancing, heterogeneous wireless access networks, GERAN, UTRAN.

1. INTRODUCTION

Future mobile networks are thought to consist of a flexible and open architecture comprising a large variety of different radio access technologies (RAT) that constitute a so-called heterogeneous network. Bearing this in mind, the concept of Common Radio Resource Management (CRRM) arises in order to efficiently manage the common pool of radio resources that are available in each one of the existing RATs [1] [2].

One of the tasks the CRRM entity should carry out is the RAT selection, either at the session initiation or during the session lifetime (i.e. in case of vertical handover). Users demanding service are to be efficiently allocated to the existing and available RATs according to a certain policy denoted as π. The CRRM may then respond to these policies, which take into account both technical and/or economical aspects, such as demanding service class, load at each RAT, revenue, etc.

Load balancing is a possible guiding principle for resource allocation in which a policy π^{load} will distribute the load among all resources as evenly as possible. However, in some situations a load balancing policy may not be desirable, at least not as a unique policy to be applied. Indeed, at one stage, an operator may be more interested in allocating users according to a service policy, $\pi^{service}$, e.g. because it increases the revenue, rather than performing a load balancing assignation. However, these service policies may drive to a very uneven situation. Consider for example that voice users are assigned to RAT_1 and interactive users to RAT_2. If the number of voice users increases while the interactive users remain constant, this may lead to the blocking of voice users in RAT_1, while RAT_2 could accommodate those users perfectly. Moreover, the performance of a certain service may differ from one RAT to another and consequently the eligibility of the RAT should also consider this fact.

Our work intends to evaluate and compare two different RAT selection policies, namely, a service class policy (π_S) and a load balancing policy (π_L), in order to discuss the suitability of load balancing in this multi-access/multiservice network.

This paper is organized as follows: Section 2 reviews some of the related work found in the literature. Section 3 presents the service-class and load balancing policies. The system models and simulation parameters are shown in section 4. In section 5 some results are presented followed by conclusions in section 6.

2. RELATED WORK

CRRM has been identified as an important issue by the 3GPP, which defines some recommendations and architectures for CRRM operation, refer to [3] and [4].

The term load balancing appears in the literature in a wide variety of contexts but profusely in the area of distributed computing where, e.g., jobs or tasks are to be assigned to a set of processors [5]. In the context of wireless access networks, load balancing may refer to the allocation of users requesting a given service to a certain cell, carrier frequency, radio access technology, etc. This allocation may be at a call/session establishment, i.e. initial RAT selection, or within an ongoing call (i.e. during vertical handover). Note that, in most cases, this assignment of mobile terminals imposes a more complex set of constraints than the case of assigning tasks to processors due to inherent

This work is partially funded by the IST-EVEREST project and by the COSMOS grant (ref. TEC2004-00518, Spanish Ministry of Science and Education and European Regional Development Fund).

properties of the wireless link, such as time-variant channel conditions, limited assignment of terminals to cells, RATs, frequencies, etc.

Load balancing algorithms have been considered to improve the performance among cells in single-RAT wireless cellular networks [6]. In this particular case, the algorithm operates when the coverage areas of different base stations overlap. Thus, whenever a mobile station can attach to more than one base station, the new call can be directed to the base station with greatest number of available channels, i.e. the least loaded base station.

For multi-RAT wireless access networks the allocation problem is extended in a way that resources may be assigned in different RATs. Literature has covered this topic in the last years in a scarce number of papers, with special focus on the effects of load balancing in inter-RAT handover procedures.

In particular, in [7], the effect of tuning the load-based handover (HO) thresholds depending on the load of inter-system/inter-layer/inter-frequency cells is studied. In order to avoid unnecessary HOs and HO signaling, a minimum load threshold ensures that no load balancing activities are carried out below that value. However, to reduce the HO attempts and HO failure rates, adjustable thresholds using neighbor load information are suggested and evaluated.

In [8], a force-based load balancing approach is proposed for initial RAT selection and vertical HO decision making. So-called *forces* model the diametric aspects of gain and cost of a decision. This decision is based on the load in the target and the source cell, the QoS difference between the radio links, the time elapsed from the last HO and the HO overhead.

Nevertheless, abovementioned references either compare results obtained in the combined UMTS/GSM systems with the disjoint systems and observe the so-called *trunking gain*, [8], or just consider a single load balancing approach with changes on the algorithm parameters [7]. Therefore, the suitability of applying load balancing techniques in a multi-RAT scenario as opposed to applying other techniques in the same scenario is not addressed.

In this paper we deal with the initial RAT selection problem for new incoming users requesting service in either of the available RATs. In order to retain the effect of load balancing in initial RAT selection, vertical HO is not considered, although will be investigated in future work. Access selection in heterogeneous networks has been also covered in the literature in a number of papers, please refer to [9] for further details.

3. LOAD BALANCING AND RAT SELECTION PROCEDURES

A basic RAT selection policy can be defined as a function that selects a suitable RAT according to some input parameters, in our case, the service class and the load in each RAN. The performance of the RAT selection policies are evaluated considering UTRAN and GERAN access technologies supporting a mixing of voice and interactive (www) users. Let the following policies be considered hereon:

- *Service policy (π_S)*: At first, this policy attempts to assign voice users to GERAN and www users to UTRAN. If no resources are available in GERAN, voice users try admission in UTRAN. Similarly, rejected www users in UTRAN will attempt admission in GERAN. If no resources are available in any of the RATs, the user gets blocked.
- *Load balancing policy (π_L)*: Upon call/session arrival, this policy adaptively selects the RAT with the minimum load metric, as described in the following, provided that there are available resources in this RAT. Otherwise, the user gets blocked.

An influential run-time parameter in a load balancing decision-making procedure is the *load metric*. In our study the following metrics are considered:

In UTRAN, the well-known load factor expressions [10] are used in their window-averaged form defined as

$$\eta_T(i) = \frac{\sum_{j=1}^{T} \eta(i-j)}{T} \quad (1)$$

where T is the window size for averaging given in number of UTRAN frames.

The uplink load factor in the ith frame is estimated as

$$\eta_{UL} = 1 - \frac{P_N}{I_{total}} \quad (2)$$

with P_N the background thermal noise and I_{total} the total received wideband power.

The downlink load factor in the ith frame is

$$\eta_{DL} = \frac{P_{total}}{P_{\max}} \quad (3)$$

where P_{total} is the total downlink transmission power and $P_{\max}$ is the maximum Node-B transmission power.

As for GERAN, a useful way to measure the data load is to measure the average amount of Time SLots (TSL) utilized by GSM/EDGE services [11]. A window-averaged timeslot utilization factor is defined

$$TSL_{utilisation,T} = \frac{\sum_{j=1}^{T} TSL_{utilisation}(i-j)}{T} \quad (4)$$

with the timeslot utilization factor in the ith frame being

$$TSL_{utilisation} = \frac{Used\ TSL\ in\ previous\ frame}{Available\ TSL\ for\ GSM + EGPRS} \quad (5)$$

where T is the window size for averaging given in number of EGPRS frames. The above expressions are particularized for both the uplink and downlink.

In general, upon call/session arrival, the cell selection procedure selects the base station with best received signal strength, in the case of GERAN, and

best E_c/I_0 in the case of UTRAN. When using the load balancing algorithm, the network selects two target base stations, one for each supported RAT. These base stations are chosen to be the ones with best signal strength and best E_c/I_0 for GERAN and UTRAN respectively. For the selected base stations, load metrics are measured and users allocated according to the defined policy.

From here on, and for brevity purposes, the term *load* accompanied by the corresponding RAT name will be used when referring to the load factor in UTRAN and the timeslot utilization factor in GERAN.

4. SIMULATION SETUP

A scenario with UTRAN and GERAN access technologies is considered. We assume a 2.25x2.25 km^2 area with 7 collocated omnidirectional cells for GERAN and UTRAN. Sites are separated a distance of 1 km. It is assumed for GERAN that the 7 cells represent a cluster where each cell works with different carrier frequencies. Three carriers per cell, belonging to the 1800 MHz, are used in GERAN, while a single carrier is considered in UTRAN. The urban macrocell model is assumed with shadowing deviation of 10 dB [12].

The parameters of the UE and BS in UTRAN and GERAN cells are summarized in Table 1.

Table 1. UTRAN and GERAN BS and UE parameters.

UTRAN BS parameters	
Max. transmitted power	43 dBm
Thermal noise	-104 dBm
Common Control Channels Power	33 dBm
Max. DL power per user	41 dBm
UTRAN UE parameters	
Max. transmitted power	21 dBm
Min. transmitted power	-44 dBm
Thermal noise	-100 dBm
DL Orthogonality factor	0.4
GERAN BS parameters	
DL transmitted power	43 dBm
Thermal noise	-117 dBm
Number of carriers	3
EGPRS slots	All slots reversible except slot 0 of first carrier
GERAN UE parameters	
Max. transmitted power	33 dBm
Min. transmitted power	0 dBm
Thermal noise	-113 dBm
Multislot class	2 UL, 3 DL, 4 UL+DL

A service-class mix of voice and interactive users is considered with mobility over the scenario at 3 km/h. Voice calls are generated according to a Poisson process with an average call rate of 10 calls/h/user and exponentially distributed call duration with an average 180 s. In UTRAN, the RAB for voice users is the 12.2 kb/s speech one defined in [13], considering a dedicated channel (DCH) with spreading factor 64 in the uplink and 128 in the downlink. As for GERAN, voice users are allocated to a TCH-FS (traffic channel with full-rate speech), i.e. one time slot in each GSM frame.

Interactive users follow the www browsing model given in [14], with 5 pages per session, an average reading time between pages of 30s, an average of 25 objects (packets) per page, and inter-arrival packet time 0.125s for the UL and 0.0228s for the DL. The average packet size is 366 bytes. This leads to an average bit rate during activity periods of 24 kb/s in the uplink and 128 kb/s in the downlink. A session rate of 24 sessions/h/user is assumed. WWW browsing service is provided in UTRAN by means of DCH making use of transport channel type switching to RACH/FACH during inactivity periods. The considered RAB assumes a maximum bit rate in the uplink of 64 kb/s (corresponding to a minimum spreading factor of 16) and in the downlink of 128 kb/s (with a spreading factor of 16). The RAB characteristics are given in [13]. In turn, the www service in GERAN is provided through a PDCH (Packet Data Channel) with round robin scheduling. A link adaptation mechanism selects the highest modulation and coding scheme (MCS) that ensures the CIR requirements. The considered maximum allowed MCS in our study is MCS-7, corresponding to a bit rate of 44.8 kb/s per time slot. Then, assuming that the multislot class allows up to 2 uplink slots and 3 downlink slots (see Table 1), the maximum bit rate is 89.6 kb/s in the uplink and 134.4 kb/s in the downlink. Consequently, in terms of maximum bit rate, similar values are considered for both UTRAN and GERAN, thus enabling consistent comparisons.

A summary of the main RRM parameters residing at the local RRM entities in both UTRAN and GERAN are given in Table 2.

Table 2. RRM parameters.

UTRAN RRM parameters	
Admission method UL	Based on load factor
Admission method DL	Based on transmitted power
UL admission threshold (η_{max})	1.0
DL admission threshold (P_{max})	42 dBm
Active Set size	1
Replacement hysteresis	3 dB
Time to trigger handover	0.64 s
GERAN RRM parameters	
Link adaptation period	1s.
BS_CV_MAX	15
GPRS_MS_TXPWR_MAX_CCH	43 dBm
GPRS_RESELECT_OFFSET	-2 dB
GPRS_RXLEV_ACCESS_MIN	-105 dBm
Max. number of TBFs per slot	UL: 8, DL:32
L_RXLEV_UL_H	-100 dBm
L_RXLEV_DL_H	-100 dBm
MS_RANGE_MAX	35 km
P5	3
P8	3

Considered QoS parameters set the BLER target at 1% and 10% for voice and interactive users respectively in both RATs. Dropping occurs in UTRAN when BLER is 1dB below target during 20 s. In GERAN, dropping happens when BLER is 5dB below target during 20 s. or when 10 consecutive unsuccessful HO occur.

As for the load balancing parameters, load metrics are considered in the uplink direction, i.e. those described by equations (2) and (4) considering the uplink timeslot utilization factor in the last case. Load averaging windows are chosen to be of length 1 second.

5. RESULTS

Given that policy π_S allocates users according to the demanded service-type, we can foresee that the traffic mix will impact the performance of this policy. Therefore, in order to evaluate the suitability of policy π_L, we consider two representative service mixes, SM1 and SM2, which are chosen so that different stress conditions are noted in GERAN when policy π_S is applied. In SM1 the number of interactive users is fixed while voice users increase. On the contrary, in SM2 the number of voice users is fixed while the number of interactive users increases.

Results are shown in the uplink direction although the same trend was observed in the downlink.

5.1. Service Mix 1

Figure 1 shows the average cell load of the central base station in both RATs for SM1 when policies π_S and π_L are used. Note that for π_S policy, voice users are directed to GERAN while not full; otherwise, requests are transferred to UTRAN. Load balancing policy π_L behaves as expected, maintaining cell load levels in both RATs at approximately the same level.

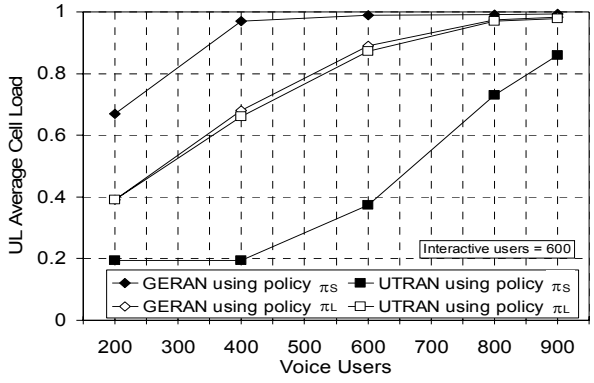

Figure 1. Average UL load for policies π_S and π_L.

Figure 2 illustrates the total aggregated throughput for SM1 when using policies π_S and π_L. Results show an improvement of total aggregated throughput with policy π_L. Due to π_S policy, users in GERAN bear higher load conditions (see Figure 1), which in turn causes dropping to increase. Therefore, throughput contribution of these users, mostly voice users allocated by π_S, diminishes. Recall that no inter-RAT handovers are considered in this study.

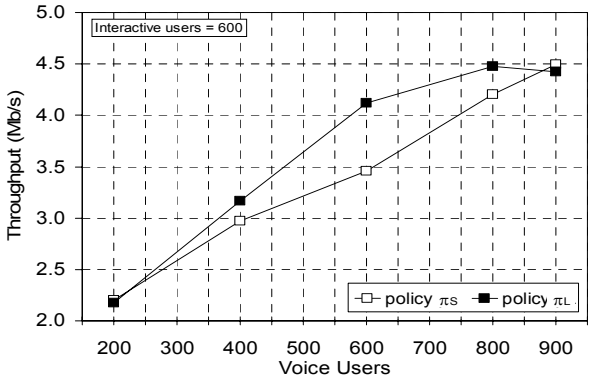

Figure 2. Total UL aggregated throughput with SM1.

Figure 3 shows the average weighted packet delay of interactive users for both policies. This delay is defined by means of the contributed throughput in each RAT:

$$WD_{www} = d_{www}^{UTRAN} \cdot \alpha + d_{www}^{GERAN} \cdot (1-\alpha) \qquad (6)$$

with d_{www}^{j} the average packet delay for interactive users in each RAT; and α, $(1-\alpha)$ the fractions of interactive throughput served through UTRAN and GERAN respectively. This performance measure enables a fair comparison between both RATs due to the uneven traffic distribution introduced by policy π_S.

Results indicate that interactive users being allocated with policy π_S undergo lower average packet delays, which benefit the perceived QoS of those users.

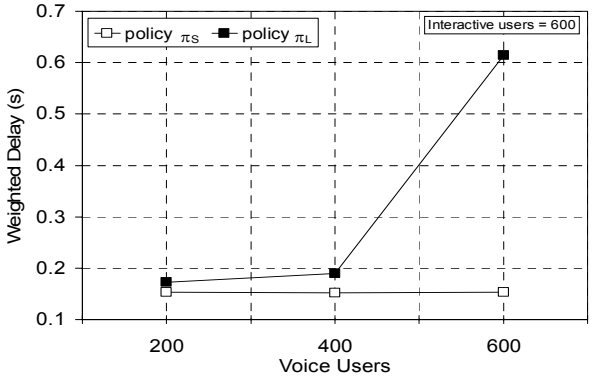

Figure 3. Average weighted packet delay for interactive users.

It has been shown that, for SM1, load balancing policy tradeoffs the overall performance of the system, in terms of total aggregated throughput, with the performance reduction of interactive users. Interactive users that are forced to GERAN by means of load balancing procedures may exhibit degradation in terms of average delay packet delay.

5.2. Service Mix 2

The average cell load for SM2 is depicted for the abovementioned policies in Figure 4. In this case, the stress is not set on GERAN, which can manage its share

of 200 users. As for UTRAN, a moderate load increase is observed, although easily handled. On the other hand, policy π_L exhibits the expected behavior in terms of similar load levels in both RATs.

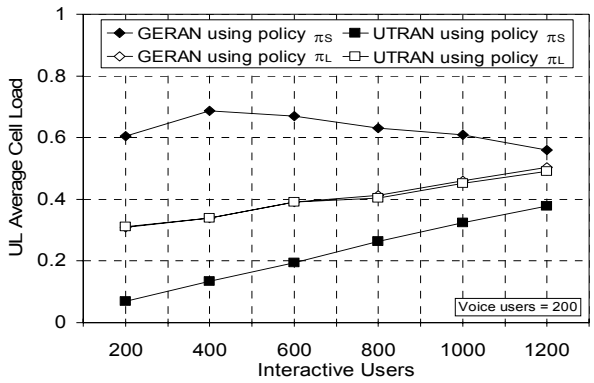

Figure 4. Average UL load for policies π_S and π_L.

Figure 5 shows the total uplink aggregated throughput for policies π_S and π_L. For this service mixing, policy π_L does not show a visible improvement with respect to the service class policy π_S. Average load curves (Figure 4) indicate that load levels are kept low, compared to SM1, and therefore no severe dropping occurs.

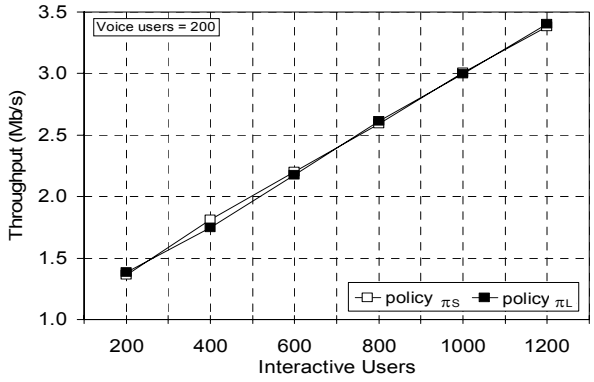

Figure 5. Total UL aggregated throughput with SM2.

Finally, the average weighted packet delay exhibited by interactive users in both RATs is depicted in Figure 6. Similar to the case of SM1, a degradation of delay performance is noted by forcing load balancing among RATs when interactive users are actually best served in UTRAN. Note that the degradation in terms of packet delay is less severe than for SM1, in part because GERAN can now manage interactive users better.

6. CONCLUSIONS

This paper has focused on the impact of load balancing in initial RAT selection procedures compared to service-class based policies by means of simulations. Results revealed a tight dependency between the suitability of load balancing RAT selection and service-class mixing. It has been shown that even though the overall throughput may increase with load balancing policies, this at the expense of interactive traffic performance. Nevertheless, other service type mixings showed no type of throughput improvement at all.

Future work includes extending the concepts presented here considering vertical handover algorithms.

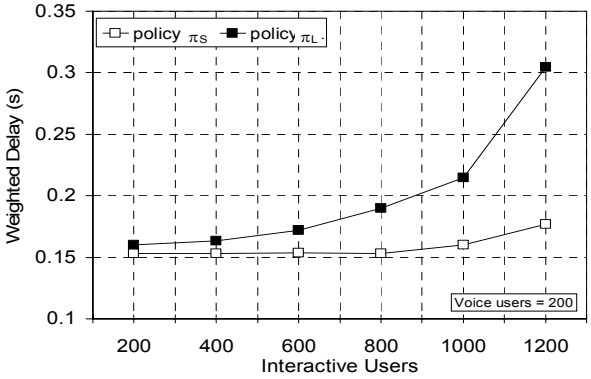

Figure 6. Average weighted packet delay for interactive users.

REFERENCES

[1] A. Tölli, P. Hakalin, H. Holma, "Performance Evaluation of Common Radio Resource Management (CRRM)", ICC2002, NY, USA.

[2] J. Pérez-Romero, O. Sallent, R. Agustí, M.A. Díaz-Guerra, *Radio Resource Management strategies in UMTS*, John Wiley & Sons, 2005.

[3] 3GPP TR 25.881 v5.0.0 "Improvement of RRM across RNS and RNS/BSS".

[4] 3GPP TR 25.891 v0.3.0 "Improvement of RRM across RNS and RNS/BSS (Post Rel-5) (Release 6)".

[5] G. Cybenko, "Dynamic Load Balancing for Distributed Memory Multiprocessors", IEEE Journal on Parallel and Distributed Computing, Vol. 7, Oct. 1989, pp. 279-301.

[6] T. Chu, S. Rappaport, "Overlapping Coverage with Reuse partitioning in Cellular Communication systems", *IEEE Trans. on Vehicular Technology*, vol. 46, no. 1, pp.41-54, Feb 1997.

[7] A. Tölli, P. Hakalin, "Adaptive load balancing between multiple cell layers", VTC 2002-Fall. 2002 IEEE 56th Vol. 3, 24-28 Sept. 2002 pp.1691 – 1695.

[8] A. Pillekeit, F. Derakhshan, E. Jugl, A. Mitschele-Thiel, "Force-based load balancing in co-located UMTS/GSM networks", VTC 2004-Fall. 2004 IEEE 60th Vol. 6, 26-29 Sept. 2004 pp. 4402 – 4406.

[9] J. Pérez-Romero, O. Sallent, R. Agustí, "Policy-based Initial RAT Selection algorithms in Heterogeneous Networks" accepted at MWCN '05. Marrakech-Morocco 19-21 Sept. 2005.

[10] H. Holma, A. Toskala (Eds.) *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications*, John Wiley & Sons, 2002.

[11] T. Halonen, J. Romero, J. Melero (Eds.) *GSM, GPRS and EDGE Performance, Evolution Towards 3G/UMTS*, John Wiley & Sons, 2002.

[12] 3GPP TR 25.942, v5.1.0 "RF System Scenarios, Release 5. (2002–06)".

[13] 3GPP TS 34.108 "Common Test Environments for User Equipment (UE); conformance testing".

[14] UMTS 30.03 v3.2.0 TR 101 112 "Selection procedures for the choice of radio transmission technologies of the UMTS", ETSI, April, 1998.

