TCP Level Investigation of Parallel Transmission over Heterogeneous Wireless Networks

Alessandro Bazzi and Gianni Pasolini
IEIIT-BO/CNR, DEIS-University of Bologna,
V.le Risorgimento 2, 40136 Bologna, Italy.
Email: alessandro.bazzi@unibo.it, gianni.pasolini@unibo.it

Abstract—Communication networks beyond 3G will see the interconnection of heterogeneous radio access networks in order to always provide the best quality of service in the most efficient way to users with multi-mode terminals. This scenario poses a number of critical issues, which have to be faced in order to get the best from the integrated network.

In this paper we investigate the issue of parallel transmission over multiple radio access networks: in particular we propose simple but effective radio resource management algorithms, whose performance is investigated by means of a simulative approach in UMTS-IEEE802.11a and IEEE802.16e-IEEE802.11a heterogeneous networks (adopted as case studies).

An insight on TCP level critical aspects, strictly related to the parallel transmission over multiple access networks, is also provided.

I. INTRODUCTION

It is a common opinion among researchers that communication networks beyond 3G will see the interconnection of heterogeneous radio access networks in order to always provide the best quality of service in the most efficient way. The realization of such a scenario will allow, in fact, not only the pursuit of the “Always Best Connected” paradigm, but also an increase of the efficiency in the networks usage by fully exploiting the peculiarities, in terms of capacity, cost, coverage and support of users’ mobility, of the different radio access technologies (RATs) that could be deployed in the same coverage area.

Several steps have already been taken in the direction of RATs integration: protocols to connect wireless local area networks (WLAN) and 3G cellular networks are currently under standardization (see for example [1] and [2]), and user terminals able to operate with more than one communication technology are already a reality.

Nonetheless, this scenario poses a number of critical issues, which are mainly related to the architecture of future heterogeneous networks and to the radio resource management strategies to be adopted in order to take advantage of the multi-access capability.

From the viewpoint of the heterogeneous network architecture, the simplest solution is the so-called loose-coupling: different networks are connected through gateways, still maintaining their independence. This scenario, that is based on the mobile IP paradigm, is only a little step ahead the current situation of completely independent RATs, and does not allow seamless handovers between two RATs.

A more interesting and promising solution is the so-called tight-coupling: in this case different RATs are connected to the same controller and each of them supports a different access modality to the same “core network”. This solution is significantly more complex but will allow fast handovers and a really effective network resources management, which in the following will be referred as multi radio resource management (MRRM).

As far as MRRM is concerned, it is straightforward to understand that the availability of a heterogeneous access network adopting the tight coupling architecture will make possible to take advantage of the multi-radio transmission diversity (MRTD) [3], which consists in the data flow split between two communication entities over more than one RAT.

MRTD, in particular, can be accomplished in a twofold manner: 1) dynamic switching between the available RATs and 2) parallel transmission [3]; in the former case the entity performing MRRM selects the RAT via which data units are going to be transmitted, while in the latter case there is a concurrent usage of RATs for the same data flow (with or without data duplication for the transmission over the different RATs).

An example investigation of “parallel transmission MRTD” is reported, for instance, in [4], where the provision of video streaming and web browsing services is considered, and the most relevant data (video base-layer and www main-objects, which are only a small fraction of the total but of great importance) are carried by an UMTS RAT, whereas a WLAN, which is faster but less reliable, is used to transmit video enhancement-layers and www inline-objects.

In this paper, differently from [4], we do not assume that the data splitting is performed by the traffic source on the basis of the data importance. Here, on the contrary, we did the more realistic assumption that the traffic source (which could be far from the end user) does not know whether multiple RATs are available at the user side or not.

We assumed, therefore, that the possible data splitting is performed locally at the Network level, by the entity managing the RATs (if more than one) covering the user region. This is even more realistic considering that users could be moving, thus dynamically entering or exiting multiple RATs areas. The aim of this paper is, in particular, to investigate the benefits and the critical aspects of “parallel transmission MRTD without data duplication” in a tight-coupled heterogeneous network in the case of best effort traffic.

The three most relevant actual or upcoming RATs have been considered as case studies: the well known UTRA-FDD/UMTS technology for cellular communications, the
widespread IEEE802.11a technology for WLANs and the IEEE802.16e technology (also known as Mobile-WiMAX) for broadband mobile access.

As will be seen in the following, the achievement of a real benefit from “parallel transmission MRTD” is conditioned to the fulfillment of some requirements related to the kind of RATs, the MRRM strategy and the TCP protocol behavior. All these aspects have been carefully considered in our investigation, which has been carried out by simulation.

Results have been obtained, in particular, by means of the simulation platform SHINE, that was developed in the framework of several research projects at WiLab, Bologna, Italy [5]. Its aim is to reproduce the behavior of RATs, carefully considering all aspects related to every single levels of the protocol stack and all characteristics of a realistic environment. This simulation tool, described in [6], has already been adopted to investigate an UMTS-WLAN heterogeneous network in the case of “dynamic switching MRTD” (e.g., [7]).

The paper is organized as follows. In Section II the scenario considered for our investigations is described; in Section III the issue of Transport protocol behavior with multiple RATs is addressed; in Section IV original MRRM strategies are proposed and, finally, in Section V the final conclusions are drawn.

II. SCENARIO

The scenario here considered consists of a tight-coupled heterogeneous network with two RATs, either UMTS-IEEE802.11a or IEEE802.16e-IEEE802.11a; since our interest is focused on the access network side, we assumed that packet losses and delays introduced by the core network are negligible.

As far as UMTS is concerned, in particular, the 384kbps bearer of UTRA-FDD was considered and a channelization bandwidth of 5MHz in the 2GHz band was assumed.

As for WiMax, a 7MHz channelization with 2048 OFDM subcarriers was assumed in 3.5GHz band; the TDD duplexing scheme was considered as well as a frame duration of 10ms and a downlink:uplink asymmetry rate of 2:1.

Finally, the IEEE802.11a WLAN was considered as foreseen by the specification, that is, with a channelization of 20MHz in the 5GHz band and a transmission rate going from 6Mbps to 54Mbps.

We assumed that, according to the principle of “parallel transmission MRTD”, each user can simultaneously operate with both available RATs by means of multi-mode user terminals. Here we considered, in particular, the parallel transmission “without data duplication”.

In this paper we did not consider other traffic categories than best effort; users were connected to both RATs at the same time, ideally expecting to perceive a total throughput as high as the sum of those possible with each of them singularly.

We made the (realistic) assumption that the entity performing MRRM is periodically informed on the number of packets transmitted by each technology as well as on the number of packets waiting in each Medium Access Control (MAC) level queue; by the knowledge of these parameters a decision on the traffic distribution over the two RATs is taken, as detailed later on.

In order to make easier the interpretation of numerical results, in the following we considered, without loss of generality, only one active user performing an infinite file download.

III. TRANSPORT LEVEL ISSUES

The most widespread versions of the TCP Transport protocol (e.g. New Reno (NR) TCP) work at best when packets are delivered in order or, at least, with a sporadic disordering. A frequent out-of-order delivery of TCP packets originates, in fact, useless duplicates of Transport level acknowledgments; after three duplicates a packet loss is supposed by the Transport protocol and the Fast Recovery - Fast Retransmit phase is entered at the transmitter side; this causes a significant reduction of the TCP congestion-window size, which is directly related to the throughput achievable at the Transport level.

This aspect of TCP’s behavior has been deeply investigated in the literature (e.g. [8] and [9]) and modern standards often include a re-ordering entity at the MAC level (see e.g. the WiMax standard [10]) to prevent possible performance degradation.

However, when “parallel transmission MRTD” is adopted, each RAT works autonomously at MAC and physical levels, without knowledge of other active RATs, hence no reordering could be possible at MAC level. Nonetheless, it would be preferable to avoid, for the sake of simplicity, the introduction of an entity that collects and reorders TCP level packets coming from different RATs. For this reason, the adoption of particular versions of TCP, especially designed to solve the Triple Duplicate problem, is advisable.

Here we considered the adoption of the Delayed Duplicates New Reno version of TCP (DD-TCP) [9], which simply delays the transmission of the TCP acknowledgment when an out-of-order packet is received, with the hope that the missing packets is already on the fly. Of course, the drawback of this solution is that the Fast Recovery-Fast Retransmit phases are delayed also when they are necessary.

In order to investigate the impact of DD-TCP on the performance achievable with the “parallel transmission MRTD”, here we considered a downlink best effort connection simultaneously exploiting two RATs. Since our aim is to highlight only the effect of the Transport level behavior, the heterogeneous network considered in this section is somewhat anomalous: the two considered RATs are, in fact, both IEEE802.11a WLANs, whose Access Points (APs) are located in the same place. Since the two simultaneous connections provide the same throughput, the MRRM strategy we adopted randomly distributes TCP packets between the two RATs with uniform (i.e., 50%) probability.

The outcome of this investigation is reported in figure 1, where the amount of acknowledged TCP packets is reported as a function of the time for both DD-TCP and NR-TCP. The case of a single AP is also shown for comparison purpose (in this case DD-TCP and NR-TCP provide the same performance); the circles (‘o’) indicate the Triple Duplicates events.

To derive the results reported in figure 1 we considered an user that, starting from the APs position, moves away
at a speed of 3m/s. It follows that increasing time instants correspond to increasing distances from the APs and, as a consequence, to a decreasing slope of the curves, which is induced by the WLAN link adaptation strategy that, as the user moves away from the APs, selects more reliable but slower modulation/coding schemes.

What is more important is that, as can be observed in figure 1, Triple Duplicates are generated only when NR-TCP is adopted and that they occur during the whole simulated time interval, no matter the distance from the APs (i.e., independently on the signal quality); this means that all Triple Duplicates are a consequence of out-of-order packet delivery events. We verified in fact that, thanks to the WLAN Automatic Repeat reQuest (ARQ) mechanism, no Data Link level fragment, and consequently no TCP/IP packet, is lost in the investigated scenario during the whole simulation time, even when the maximum distance is reached (after 10s).

As can be observed, Triple Duplicates greatly affect the achieved performance level; the comparison with the curve related to a single AP shows, in fact, that the number of acknowledged packets is not doubled when considering NR-TCP with two RATs.

When DD-TCP is adopted, on the contrary, no Triple Duplicate event occurs and the throughput of the single connection is doubled; let us stress that this is not a trivial result, since we are splitting a single TCP flow over two independent technologies and resembling them directly at the TCP level at the receiver side.

Please note that the DD-TCP protocol was chosen, among other possibilities, since it is a very simple solution. It is beyond the scope of this paper to investigate the most suitable TCP version to overcome the Triple Duplicate problem in multiple RATs networks.

IV. MRRM STRATEGIES

Let us consider, now, a really heterogeneous network, which is in general constituted by RATs whose characteristics could be very different in terms of transmission rates.

It is straightforward to understand that in this case the random distribution of packets with uniform probability over the different RATs would hardly be the best solution. Indeed, to fully exploit the availability of multiple RATs and get the best from the integrated access network, an efficient MRRM strategy must be designed, able to properly balance the traffic distribution over the different access technologies.

In order to get some insight on the proper traffic balance, a simulative investigation has been carried out considering two different heterogeneous access networks: the first one integrates WLAN and WiMax RATs while the second one integrates WLAN and UMTS RATs.

All wireless access point, that is, the WLAN AP, the UMTS Node B and the WiMax Base Station, are placed in the same position and the single user here considered is located near them (this means high perceived signal to noise ratio). Packets are probabilistically passed to the WLAN MAC/Physical levels with probability $P_{WLAN}$ and to the other technology (i.e., WiMax or UMTS) with probability $1 - P_{WLAN}$ (both in the uplink and in the downlink).

The simulations outcomes are reported in figure 2, where the throughput perceived at the TCP level is shown as a function of $P_{WLAN}$.

With reference to the curve related to the WLAN-WiMax case, it can be noted that a joint use of the two technologies allows to reach (with $P_{WLAN} \approx 0.6$) a throughput that is nearly the sum of the throughputs provided by a single WLAN RAT or a single WiMax RAT ($P_{WLAN} = 1$ and $P_{WLAN} = 0$ respectively). It can also be noted that efficiency falls down quickly if a wrong choice of $P_{WLAN}$ is made.

Observing the curve related to the WLAN-UMTS heterogeneous network, it can be noted that the high difference
of the throughputs provided by the two RATs makes the TCP behavior so inefficient that the adoption of the WLAN technology alone is the best solution; a significant performance degradation can be noted, in fact, if $P_{WLAN}$ is lower than $\sim 0.98$. For this reason, in the rest of the paper we will focus on the WLAN-WiMax heterogeneous network only.

A. Dynamic traffic distribution:

The above discussed results showed that, depending on the characteristics of the considered RATs, there exists an optimal choice for the traffic distribution policy, which depends, in particular, on the TCP level throughput that every single RAT can provide to an user with respect to the TCP throughput provided by the other RATs.

Obviously, the TCP throughput provided to a single user by a given RAT depends on a number of dynamically changing parameters, such as the amount of served user (which affects the MAC level queue occupation), their position (which could affects the Physical level transmission rate through a link adaptation algorithm, if present), etc.

Focusing again the attention on the two heterogeneous networks previously considered, the question is: how can the entity performing MRRM dynamically and automatically select the correct value for $P_{WLAN}$?

In this paper we propose original MRRM strategies, that we called Tx/Qu and Smooth-Tx/Qu, and compare their performance with a benchmark case. More specifically, the following MRRM strategies are considered and compared in the following:

1) *Random*: packets are randomly distributed with equal probability among active connections (please note that a random distribution corresponds to $P_{WLAN} = 0.5$ and observe that in figure 2, referring to the WLAN-UMTS case, this is absolutely a wrong choice). This policy is considered only for comparison purpose;

2) *Transmissions over Pending Packets (Tx/Qu)*: packets are always passed to the technology with the higher value of the ratio between the number of transmitted packets and the number of packets waiting in the MAC queue; thus, system queues are kept filled proportionally to the transmission speed;

3) *Smoothed Transmissions over Pending Packets (Smooth-Tx/Qu)*: it is an evolution of the Tx/Qu strategy. The only difference is that in this case the number of transmitted packets is halved every $T$ seconds (we adopted $T = 0.125s$); periodically halving the amount of transmitted packets allows to reduce the impact of old transmissions, thus improving performance in a scenario where users are moving (and transmission rates could change).

In figure 3 the above detailed MRRM strategies are compared in a scenario consisting of a heterogeneous network with one IEEE802.11a AP and one WiMax Base Station located in the same position. The user is performing an infinite file download and does not change its position; its distance from the co-located AP/Base Station is reported on the $x$-axis, while the average perceived TCP level throughput is reported on the $y$-axis.

Before discussing the results reported in figure 3, a preliminary note on the considered distance range ($0 - 30m$) is needed.

Let us observe, first of all, that WiMax is a long range communications technology, with a coverage range in the order of kilometers. Nonetheless, since our focus is on the heterogeneous WLAN-WiMAX access network, we must consider coverage distances in the order of tenths of meters (i.e., the coverage range of a WLAN), where both RATS are available; that is why the $x$-axis of figure 3 ranges from 0 to 30 meters.

The different curves of figure 3 refer, in particular, to the three MRRM strategies above described and, for comparison, to the cases of a single WLAN RAT and of a single WiMax RAT.

Obviously, when considering the case of a single WiMAX RAT, the throughput perceived by an user located in the region of interest is always at the maximum achievable level, as shown by the flat curve in figure 3. As expected, on the contrary, the throughput provided by the WLAN in the same range of distances rapidly decreases for increasing distances.

The most important results reported in figure 3, however, are related to the three upper curves (two of them are superimposed), which refer to the previously described MRRM strategies when applied in the considered heterogeneous WLAN-WiMAX access network.

As can be immediately observed, the two dynamic strategies proved to be really effective, greatly outperforming the Random distribution strategy. Please observe that the achievable throughput in these cases is even slightly higher than the sum of the throughputs provided by each technology alone.

At a first glance it could seem strange that a throughput (slightly) higher than the sum of the two throughputs provided in the single RAT cases can be achieved; however, this phenomenon can be easily explained considering the fact that the adopted DD-TCP solution (slightly) reduces the number of
TCP level acknowledgments transmitted in the uplink phase (in average a higher number of packets are acknowledged by a single DD-TCP acknowledgment with respect to NR-TCP). Since in a WLAN the uplink and downlink phases contend for the wireless medium, a reduction of the uplink traffic turns into a downlink throughput increase.

As a final consideration on figure 3, we can observe that in this case the Smooth-Tx/Qu and the Tx/Qu strategies are almost equivalent; this is due to the fact that there is no user mobility.

In table I some results are given which are related to the same scenario (single user and a WLAN-WiMax heterogeneous network with co-located WLAN AP and WiMAX Base Station) in different conditions, including also mobility.

In particular, four situations are considered: (1) the user stands still near the AP/Base Station (optimal signal reception), (2) the user stands still at 30m from the AP/Base Station (optimal WiMax signal, but medium quality WLAN signal), (3) the user moves from the AP/Base Station far away at a speed of 1 m/s (low mobility) and (4) the user stands still near the AP/Base Station for half the simulation time, then it moves instantaneously 30m far away (reproducing the effect of a high speed mobility).

Results are shown for all the above described MRRM strategies as well as for the benchmark scenarios with a single WLAN RAT and a single WiMax RAT and refer to the average (over the 10s simulated time interval) throughput perceived in each considered case.

As can be observed, while the random distribution confirms its poor performance (please note that when it is adopted with the user standing still at a distance of 30m, the perceived throughput is lower than that obtained using WiMax only), the proposed dynamic MRRM methods provide satisfying performance. Focusing the attention on the last case (correspondent to high mobility), the gain achieved with the Smoothed method is clearly evident, although Transmissions over Pending Packets may be sufficient in most cases.

V. CONCLUSIONS

In this paper we faced the issue of Radio Access Technologies (RATs) integration in tight-coupled heterogeneous networks. The "parallel transmission multi-radio diversity" has been particularly investigated with the aim to highlight benefits and critical aspects. Results, obtained through simulations, refer to a TCP session whose traffic is split over different wireless accesses without the need of any modifications to communication protocols.

Here we proposed original multi radio resource management strategies and derived their performance in extremely relevant scenarios, such as those constituted by UMTS-IEEE802.11a and IEEE802.16e-IEEE802.11a heterogeneous networks.

The outcomes of our investigations showed the effectiveness of the proposed strategies, that, in spite of their simplicity, fully exploit the pool of resources provided by the integrated heterogeneous network.

TABLE I

<table>
<thead>
<tr>
<th>User position</th>
<th>WLAN only</th>
<th>WiMax only</th>
<th>Random</th>
<th>Tx/Qu</th>
<th>Smooth-Tx/Qu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stands still, near the AP/Base Station</td>
<td>18.53</td>
<td>12.76</td>
<td>25.23</td>
<td>32.28</td>
<td>32.37</td>
</tr>
<tr>
<td>Stands still, 30m far from the AP/Base Station</td>
<td>3.81</td>
<td>12.76</td>
<td>7.95</td>
<td>16.35</td>
<td>16.40</td>
</tr>
<tr>
<td>Moving away at 1m/s, starting from the AP/Base Station</td>
<td>11.83</td>
<td>12.76</td>
<td>20.99</td>
<td>24.94</td>
<td>25.01</td>
</tr>
<tr>
<td>Near the AP/Base Station for half sim., then 30m far (instantaneously)</td>
<td>10.04</td>
<td>12.76</td>
<td>14.12</td>
<td>19.02</td>
<td>22.30</td>
</tr>
</tbody>
</table>

REFERENCES


