Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2008, Article ID 763264, 9 pages
doi:10.1155/2008/763264

Research Article

Multiradio Resource Management: Parallel Transmission for Higher Throughput?

Alessandro Bazzi, Gianni Pasolini, and Oreste Andrisano

WiLab, IEIIT-BO/CNR, DEIS, University of Bologna, V.le Risorgimento 2, 40136 Bologna, Italy

Correspondence should be addressed to Gianni Pasolini, gianni.pasolini@unibo.it

Received 30 November 2007; Accepted 23 April 2008

Recommended by Moe Win

Mobile communication systems beyond the third generation will see the interconnection of heterogeneous radio access networks (UMTS, WiMax, wireless local area networks, etc.) in order to always provide the best quality of service (QoS) to users with multimode terminals. This scenario poses a number of critical issues, which have to be faced in order to get the best from the integrated access network. In this paper, we will investigate the issue of parallel transmission over multiple radio access technologies (RATs), focusing the attention on the QoS perceived by final users. We will show that the achievement of a real benefit from parallel transmission over multiple RATs is conditioned to the fulfilment of some requirements related to the kind of RATs, the multiradio resource management (MRRM) strategy, and the transport-level protocol behaviour. All these aspects will be carefully considered in our investigation, which will be carried out partly adopting an analytical approach and partly by means of simulations. In this paper, in particular, we will propose a simple but effective MRRM algorithm, whose performance will be investigated in IEEE802.11a-UMTS and IEEE802.11a-IEEE802.16e heterogeneous networks (adopted as case studies).

1. INTRODUCTION

It is a shared opinion among researchers that mobile communication systems beyond the third generation (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, to pursue not only the "always best connected" paradigm, but also to increase 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 make wireless local area networks (WLANs) and 3G cellular networks interact are currently under standardisation (see, e.g., [1, 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 multiaccess 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 multiple-resources management, which in the following will be referred to as multiradio 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 multiradio transmission diversity (MRTD) [3, 4], which consists in the splitting of the

data flow exchanged by two end-to-end entities over more than one RAT.

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

The aim of this paper is, in particular, to investigate the benefits and the critical aspects of the "parallel transmission MRTD without data duplication" in a tight-coupled heterogeneous network in the case of best effort traffic.

An example investigation of "parallel transmission MRTD" is reported, for instance, in [5], 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 [5], 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.

Investigations on MRTD are also carried out in [6, 7], where the emphasis is on the exploitation of the radio channel diversity on a per packet basis, not considering, however, the impact of protocol layers higher than the data link.

Other studies on parallel transmission focus on the physical layer only, for instance [8, 9].

Differently, in this paper we consider the whole protocol stack, from the physical layer to the application one, with particular reference to the Transport layer protocol which, as will be seen in the following, deserves a particular attention in multiple RATs scenarios.

More in general, the achievement of a real benefit from "parallel transmission MRTD" is conditioned to the fulfilment of some requirements related to the kind of RATs, the MRRM strategy, and the transport-level protocol behaviour. All these aspects have been carefully considered in our investigation, which has been carried out partly adopting an analytical approach and partly by means of simulations.

The paper is organized as follows. In Section 2, the scenario considered for our investigations is outlined along with the assumptions and the description of the investigation methodology. In Section 3, the issue of the transport protocol behaviour with multiple RATs is addressed. In Section 4, an analytical investigation on the achievable performance level is carried out. In Section 5, an original MRRM strategy is proposed and its effectiveness is assessed. Finally, in Section 6 the final conclusions are drawn.

2. INVESTIGATION ASSUMPTIONS AND METHODOLOGY

In this paper, the three most relevant actual or upcoming RATs have been considered as case studies: the well-known wideband code division multiple access (WCDMA), UMTS technology for 3G cellular communications [10], the IEEE802.11a technology for WLANs [11], and the IEEE802.16e technology (also known as Mobile-WiMax) for broadband mobile access [12].

The scenario considered in this paper consists of a tight-coupled heterogeneous access network constituted by two RATs, either WLAN-UMTS or WLAN-WiMax.

The assumptions we made with reference to this scenario are summarized hereafter:

Technologies

As far as the three above-mentioned communication technologies are concerned, the following choices and assumptions have been made in the rest of the paper.

(1) *UMTS*. The WCDMA version of UMTS was considered, with a channelisation bandwidth of 5 MHz in the 2 GHz band. The 384 kbps bearer has always been assumed for data transmissions.

(2) *WiMax*. We considered the IEEE802.16e Wireless MAN-OFDMA version operating with 2048 OFDM subcarriers and a channelisation bandwidth of 7 MHz in the 3.5 GHz band; the time division duplexing (TDD) scheme was adopted as well as a frame duration of 10 milliseconds and a 2 : 1 downlink:uplink asymmetry rate of the TDD frame.

(3) *WLAN*. The IEEE802.11a WLAN technology has been considered as foreseen by the specification, that is, with a channelisation bandwidth of 20 MHz in the 5 GHz band and a nominal transmission rate going from 6 Mbps to 54 Mbps.

Since our interest is focused on the access network side, in this paper we assumed that packet losses and delays introduced by the core network are negligible. Packet losses and delays introduced by the access network have been, on the contrary, accurately taken into account.

MRRM

We assumed that, according to the principle of "parallel transmission MRTD," each user can simultaneously operate with both available RATs by means of a multimode user terminal.

Here we considered, in particular, the parallel transmission "without data duplication" modality. This means that the data flow of a single communication is split into two disjoint subflows addressed to the two different RATs.

We made the (realistic) assumption that the entity performing MRRM is periodically informed on the number of IP packets transmitted by each technology as well as on the number of IP packets still waiting (in the data link level transmitting queues) to be transmitted; by the knowledge of these parameters a decision on the traffic distribution over the two RATs is taken, as detailed later on.

Service

In this paper, we did not consider other traffic categories than the best effort one; 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 RAT singularly.

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.

Investigation methodology

Results have been obtained partly analytically and partly through simulations, adopting the simulation platform SHINE that has been developed in the framework of several research projects at WiLab, Bologna, Italy [13]. The aim of SHINE is to reproduce the behaviour of RATs, carefully considering all aspects related to each single level of the protocol stack and all characteristics of a realistic environment. This simulation tool, described in [14], has already been adopted to investigate a UMTS-WLAN heterogeneous network in the case of "dynamic switching MRTD" (see [15], e.g.).

Performance metric

The performance metric we adopted to investigate the above-described multiple RATs scenario is the throughput provided by the integrated network. As we focused our attention, in particular, on best effort traffic, we assumed that the TCP protocol is adopted at the transport layer and we derived, as performance metric, the TCP level throughput perceived by the final user.

Let us observe, now, that a huge number of different TCP versions are available nowadays; as will be shown in the following section, the choice of the particular TCP version adopted in the considered scenario is not irrelevant and must be carefully considered.

3. TRANSPORT LEVEL ISSUES

The most widespread versions of the TCP transport protocol (e.g., New Reno (NR) TCP [16]) 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 and, as a consequence, a reduction of the throughput achievable at the transport level.

This aspect of the TCP behaviour has been deeply investigated in the literature (e.g., [17, 18]) and modern communication systems often include a reordering entity at the data link level of the receiver side (see, e.g., the WiMax standard [12]) to prevent possible performance degradation.

Let us observe, now, that when "parallel transmission MRTD" is adopted, each RAT works autonomously at data link and physical levels, with no knowledge of other active RATs. During the transmission phase, in fact, the packets flow coming from the upper layers is split into subflows that are passed to the different data link level queues of the active RATs and then transmitted independently one of the others.

It follows that the out-of-order delivery of packets and the consequent performance degradation are very likely, owing to possible differences of the queues occupation levels as well as of the medium access strategies and the transmission rates of the active RATs.

The independency of the different RATs makes very difficult, however, to perform a frame reordering at the data link level of the receiver and, at the same time, 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 this problem, is advisable in multiple RATs scenarios.

Here, we considered the adoption of the delayed duplicates New Reno version of TCP (DD-TCP) [18], which simply delays the transmission of TCP acknowledgments when an out-of-order packet is received, hoping that the missing packet is already on the fly. The drawback of this solution is, of course, that the fast recovery-fast retransmit phases are delayed also when they are necessary.

The DD-TCP differs from the NR-TCP only at the receiving side of the transport level peer-to-peer communication; this implies that the NR-TCP can be maintained at the transmitter side. Thus, this solution could be adopted, at least, on multimode user terminals, where the issue of out-of-order packet delivery is more critical owing to the higher traffic load that usually characterises the downlink phase.

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.

As our aim was to highlight only the effect of the transport-level behaviour, the heterogeneous network considered for this specific investigation was somewhat anomalous: the two considered RATs were, in fact, both IEEE802.11a WLANs whose access points (APs) were located in the same place. Since the two simultaneous connections provide the same throughput, the MRRM strategy we adopted in this case randomly distributed TCP/IP packets between the two RATs with equal (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 (i.e., of a single RAT) is also shown

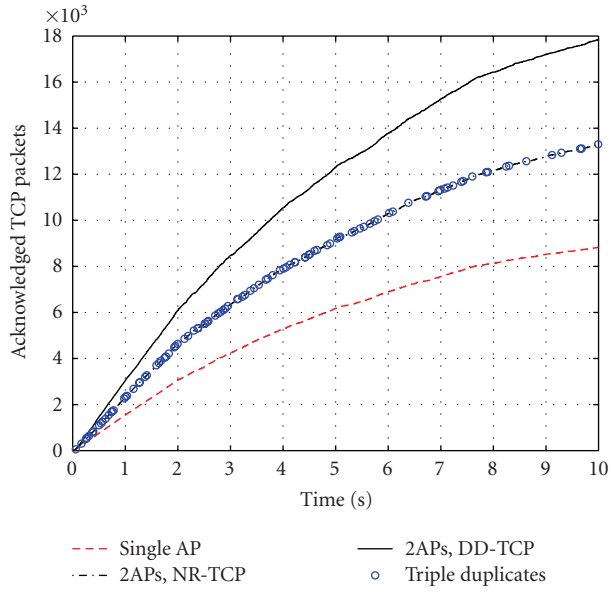

Figure 1: Acknowledged TCP packets of the download performed by one user that moves away from 2 colocated WLAN APs versus time; single RAT connection compared to parallel transmission over two RATs adopting either New Reno or delayed duplicates New Reno as TCP protocols. Triple duplicate events are marked with "o."

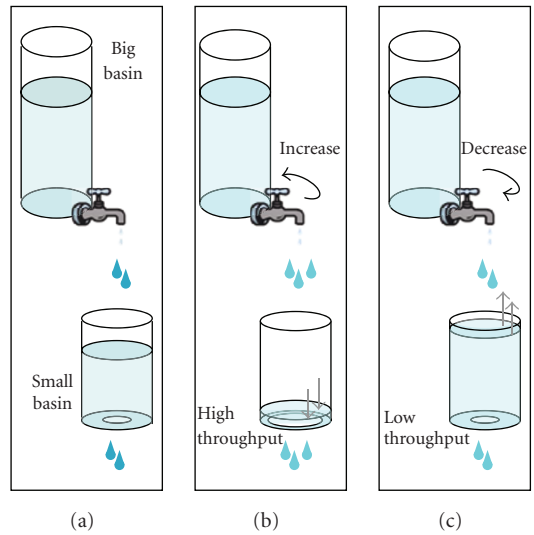

Figure 2: Representation of the TCP mechanism.

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 a user that, starting from the APs position, moves away at a speed of 3 m/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.

Observing Figure 1, it is important to notice that triple duplicates are generated only when NR-TCP over two RATs 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 here observed 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 10 seconds).

As can be observed, triple duplicates heavily 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 amount of TCP packets acknowledged in a given time interval, which is strictly related to the provided throughput, is doubled with respect to the single connection case. Let us underline that this is not a trivial result, since we are splitting a single TCP flow over two independent technologies and reassembling it directly at the TCP level of the receiver.

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.

4. THROUGHPUT ANALYSIS

Let us consider, now, a really heterogeneous network, which is in general constituted by RATs whose characteristics could be very different in terms, for instance, of medium access strategies and 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 clarify this statement, a brief digression on the TCP protocol behaviour is reported hereafter, starting from a simple metaphor.

Let us represent the application-level queue as a big basin (in the following, big basin) filled with water that represents the data to be transmitted (see Figure 2(a)). Another, smaller basin (in the following, small basin) represents, instead, the data path from the source to the receiver: the size of the data link level queue can be represented by the small basin size and the transmission speed by the width of the hole at the small basin bottom.

In this representation the TCP protocol works like a tap controlling the amount of water to be passed to the small basin in order to prevent overflow events (a similar metaphor

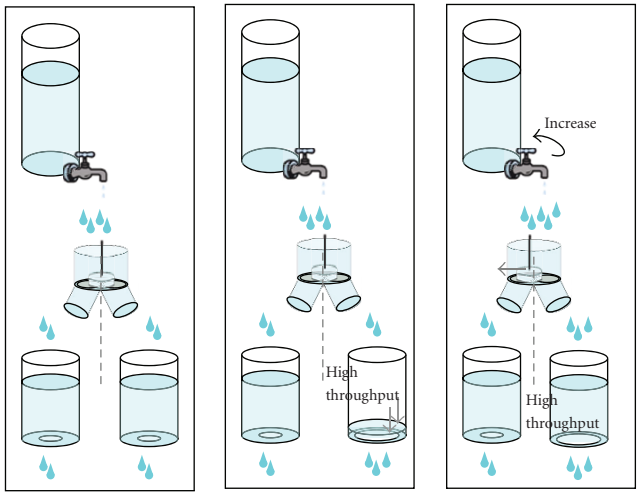

Figure 3: Representation of the TCP mechanism with parallel transmission over two RATs.

is used, e.g., in [19]). It follows that the water flow exiting from the tap represents the TCP level throughput, and the water flow exiting from the small basin represents the data link level throughput.

As long as the small basin is characterised by a wide hole, as depicted in Figure 2(b), the tap can increase the water flow, reflecting the fact that when a high data link level throughput is provided by the communication link, the TCP level throughput can be correspondingly increased.

When, on the contrary, a small hole (→ a low data link level throughput) is detected, the tap (→ the TCP protocol) reduces the water flow (→ the TCP level throughput), as described in Figure 2(c). This way, the congestion control is performed, and the data link level queues saturation is avoided.

Now the question is: what happens when two basins (i.e., two RATs) are available instead of one and the water flow is equally split between them?

Having in mind that the tap has to prevent the overflow of either of the two small basins, it is easy to understand that, in the presence of two small basins with the same hole widths, the tap could simply double the water flow, as depicted in Figure 3(a). Reasoning in terms of throughput and multiple RATs, this is the case investigated in Figure 1, where two equal and equally loaded RATs were considered.

In the presence of a small basin with a hole wider than the other (see Figure 3(b)), on the other hand, the tap behaviour is influenced by the small basin characterised by the lower emptying rate (the leftmost one in Figure 3(b)), which is the most subject to overflow. This means that the availability of a further "wider holed" basin is not fully exploited in terms of water flow increase. Reasoning in terms of TCP protocol, in fact, the congestion window moves following the TCP level acknowledgments related to packets received in the correct order. This means that, as long as a gap is present in the received packet sequence (one or more packets are missing because of a RAT slower than the other), the congestion window does not move at the transmitter side, thus reducing the provided throughput.

Coming back to the water flow metaphor, it is immediate to understand that, in order to fully exploit the availability of the further, "more performing," small basin, the water flow splitting modality must be modified in such a way that the water in the two small basins is kept at almost the same level (see Figure 3(c)). This consideration introduces in our metaphor the concept of resource management, which is represented in Figure 3(c) by the presence of a valve which dynamically changes the subflows discharge.

This concept, translated in the telecommunication-correspondent MRRM concept, will be thoroughly worked out in the remainder of the paper. To do this, however, an analytical formulation of TCP protocol behaviour in the presence of multiple RATs is needed, which is reported in the following subsection.

4.1. Throughput analytical derivation

Starting from the above-reported considerations, we can derive a simple analytical framework to model the average throughput T perceived by the final user in the case of two heterogeneous RATs, denoted in the following as RAT_A and RAT_B, managed by an MRRM entity which splits the packets flow between RAT_A and RAT_B with probabilities P_A and $P_B = 1 - P_A$, respectively.

Focusing the attention on a generic user, let us denote with T_i the maximum data link level throughput supported by RAT_i in the direction of interest (uplink or downlink), given the particular conditions (signal quality, network load due to other users,...) experienced by the user. Dealing with a dual mode user, we will denote with T_A and T_B the above-introduced metric referred to RAT_A and RAT_B, respectively.

Let us assume that a block of N transport-level packets of B bits has to be transmitted and let us denote, furthermore, with O the amount of overhead bits added by protocol layers from transport to data link. After the MRRM operation, the N packets flow is split into two subflows of, in average, $N \cdot P_A$ and $N \cdot P_B$ packets, which are addressed to RAT_A and RAT_B.

It follows that, in average, RAT_A and RAT_B empty their queues in $D_A = (N \cdot (B + O) \cdot P_A)/T_A$ and $D_B = (N \cdot (B + O) \cdot P_B)/T_B$ seconds, respectively.

Thus, the whole N packets block is delivered to the considered user in a time interval that corresponds to the longest between D_A and D_B.

This means that the average TCP level throughput provided by the integrated access network to the final user can be expressed as

$$T = \begin{cases} \dfrac{N \cdot B}{D_A} = \dfrac{T_A}{P_A}\xi, & \text{when } D_A > D_B, \text{ that is when } \dfrac{T_A}{P_A} < \dfrac{T_B}{P_B}, \\ \dfrac{N \cdot B}{D_B} = \dfrac{T_B}{P_B}\xi, & \text{in the opposite case, when } \dfrac{T_A}{P_A} \geq \dfrac{T_B}{P_B}, \end{cases} \tag{1}$$

or in a more compact way as

$$T = \min\left\{\frac{T_A \xi}{P_A}, \frac{T_B \xi}{P_B}\right\}, \tag{2}$$

where the factor $\xi = B/(B + O)$ takes into account the degradation due to the overhead introduced by protocol layers from transport to Data Link.

Let us observe, now, that the term $T_A\xi/P_A$ of (2) is a monotonic increasing function of $P_B = 1 - P_A$, while the term $T_B\xi/P_B$ is monotonically decreasing with P_B.

Since $T_A/P_A < T_B/P_B$ when P_B tends to 0 and $T_A/P_A > T_B/P_B$ when P_B tends to 1, it follows that the maximum TCP level throughput $T_{\max}$ is achieved when $T_A/P_A = T_B/P_B$, that is, when

$$P_A = P_A^{(\max)} = \frac{T_A}{T_A + T_B}, \tag{3}$$

and consequently

$$P_B = P_B^{(\max)} = 1 - P_A^{(\max)} = \frac{T_B}{T_A + T_B}, \tag{4}$$

having denoted with $P_A^{(\max)}$ and $P_B^{(\max)}$ the values of P_A and P_B that maximize T.

Recalling (2), the maximum achievable TCP level throughput is immediately derived as

$$T_{\max} = \min\left\{\frac{T_A\xi}{P_A}, \frac{T_B\xi}{P_B}\right\}\Big|_{P_A = P_A^{(\max)}} = (T_A + T_B)\xi, \tag{5}$$

thus showing that a TCP level throughput as high as the sum of the single TCP level throughputs can be achieved.

Equations (3) and (4) show that an optimal choice of P_A and P_B is possible, in principle, on condition that accurate and updated values of the data link level throughputs T_A and T_B are known (or, equivalently, accurate and updated values of the TCP level throughputs $T_A\xi$ and $T_B\xi$).

4.2. Model validation

In order to validate the above-described analytical framework, a simulative investigation has been carried out considering two different scenarios: the first one integrates a WLAN RAT and a WiMax RAT, while the second one integrates a WLAN RAT and an UMTS RAT.

All wireless access points, 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 by the MRRM entity to the WLAN data link/physical levels with probability P_{WLAN} (which corresponds to P_A of the general analytical framework) and to the other technology (i.e., WiMax in the first case or UMTS in the second one) with probability $1 - P_{\text{WLAN}}$ (which corresponds to P_B of the general analytical framework), both in the uplink and in the downlink.

The simulations outcomes are reported in Figure 4, where the average throughput perceived at the TCP level is shown as a function of P_{WLAN}.

In the same figure, we also reported the curves obtained through (2), in which we assumed that $T_A\xi$ is referred (in both scenarios) to the WLAN RAT, and $T_B\xi$ is referred, depending on the scenario, to the WiMax RAT (WLAN-WiMax scenario) or to the UMTS RAT (WLAN-UMTS scenario). The values of $T_A\xi$ and $T_B\xi$, to be feeded to (2), have been obtained by means of simulations for each one of the considered technologies, obtaining $T_{\text{WLAN}} = T_A\xi = 18.53$ Mbps, $T_{\text{WiMax}} = T_B\xi = 12.76$ Mbps (first scenario) and $T_{\text{UMTS}} = T_B\xi = 0.36$ Mbps (second scenario).

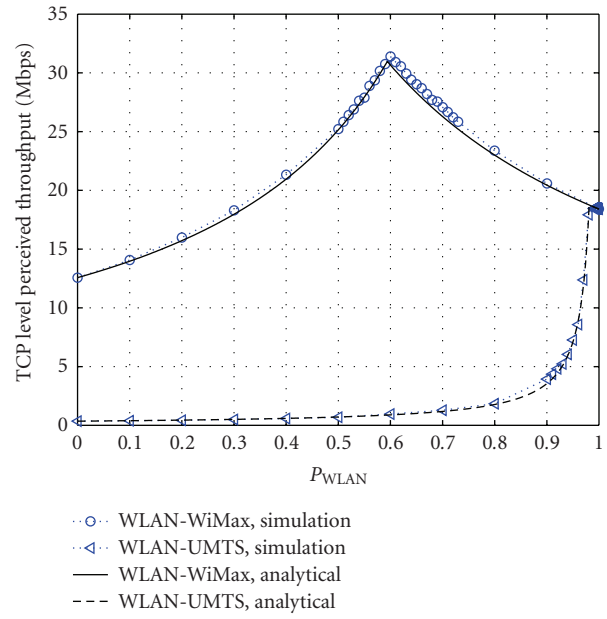

Figure 4: TCP level throughput adopting a WLAN connection and a WiMax or UMTS one, as a function of the probability that the packet is transferred through the WLAN.

With reference to Figure 4, let us observe, first of all, the very good matching between the simulation results and the analytical curves derived from (2), which confirms the accuracy of the whole framework. The accuracy of (3) and (5) can also be easily checked. Focusing the attention, for instance, on the WLAN-WiMax case, it is easy to derive (from (3)) $P_A^{(\max)} = P_{\text{WLAN}} = 0.59$ and (from (5)) $T_{\max} = 31.29$ Mbps, in perfect agreement with the coordinates of the maximum that can be observed in the curve related to the WLAN-WiMax scenario.

Let us observe, moreover, the rapid throughput degradation following an uncorrect choice of P_{WLAN}. This means that the correct assessment of P_{WLAN} heavily impacts the system performance.

Focusing the attention on the curve related to the WLAN-UMTS heterogeneous network, we can argue that in the investigated conditions the high difference of the data link throughputs provided by the two RATs makes the TCP behaviour so inefficient that the adoption of the WLAN technology alone is almost the best solution; a significant performance degradation can be noted, in fact, when P_{WLAN} is lower than $\sim$0.98. For this reason, in the next session we will focus on the WLAN-WiMax heterogeneous network only.

Please note that, although we limited our investigation to the case of two active technologies, all conclusions can also be generalised for a greater number of RATs.

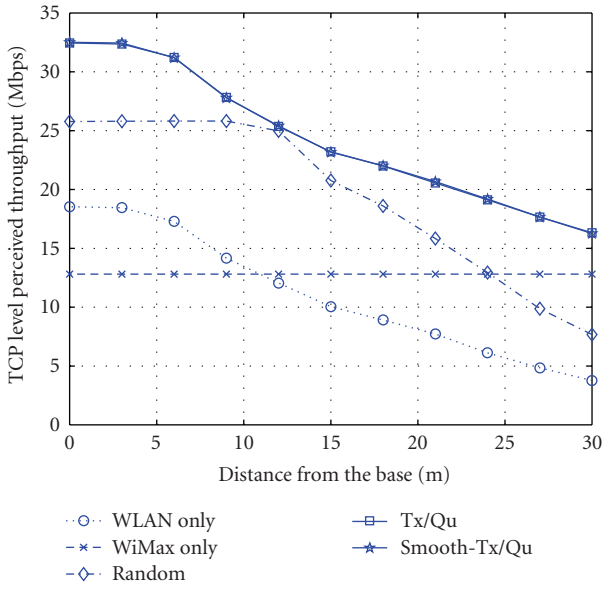

Figure 5: WLAN-WiMax heterogeneous networks. TCP level throughput varying the distance of the user from the AP/base station, for different MRRM schemes. No mobility.

5. MRRM STRATEGIES

In Section 4, we showed that, depending on the characteristics of the considered RATs, there exists an optimal traffic distribution policy for each (dual mode) user, which depends, in particular, on the average throughput that every single RAT can provide to it.

In principle, starting from knowledge of the maximum data link or TCP throughput that can be provided to the user by each RAT, the MRRM entity could perform the optimal traffic balance on the basis of (3) and (4).

Let us observe, however, that the maximum (data link or TCP) throughput that can be provided to a single user by a given RAT is time variable, since it depends on a number of dynamically changing parameters, such as the amount of served users (which affects the data link level queue occupation), its position (which could affect the physical level transmission rate if a link adaptation algorithm is adopted), the presence of interference, and so forth. It follows that its assessment could be difficult and scarcely accurate.

When a new connection is established, in fact, no knowledge of the throughput that the incoming user will perceive is available, hence no optimal traffic balance could be performed at the connection activation. When the communication is ongoing, on the other hand, the not optimal traffic balance performed at the connection setup could bring to an under-utilisation of one (or more) RAT, thus making the related throughput measurement not consistent with the throughput potentially available and consequently preventing the correct choice of the splitting probabilities.

Focusing again the attention to the case study of the two heterogeneous networks previously considered, the question is therefore how to dynamically and automatically select the correct value for P_{WLAN}.

In this paper, we propose an original MRRM strategy, that we called *Smooth-Tx/Qu*, and we compare its performance with those of benchmark cases. More specifically, the following MRRM strategies are considered and compared in the following.

(i) *Random*: packets are randomly distributed with equal probability among active connections (please note that a random distribution corresponds to $P_{\text{WLAN}} = 0.5$ and observing the curve of Figure 4 related to the WLAN-UMTS case, we can argue that in some cases this is absolutely a wrong choice). This policy is considered only for comparison purpose.

(ii) *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 data link queue; thus, system queues are kept filled proportionally to the transmission speed;

(iii) *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_{half} seconds (in our simulations we adopted $T_{\text{half}} = 0.125$ s); periodically halving the amount of transmitted packets allows to reduce the impact of old transmissions, thus improving the achieved performance in a scenario where transmission rates could change (due to users mobility, e.g.).

In Figure 5, 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 colocated 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 5, a preliminary note on the considered distance range (0–30 m) 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 few dozens of meters (i.e., the coverage range of a WLAN), where both RATS are available; for this reason the x-axis of Figure 5 ranges from 0 to 30 meters.

The different curves of Figure 5 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 a user located in the region of interest is always at the maximum achievable level, as shown by the flat curve in Figure 5. As expected, on the contrary, the throughput provided by the WLAN in the same range of distances rapidly decreases for increasing distances.

Table 1: TCP level average throughput in Mbps for various distribution schemes in different conditions. Single user, 1 WLAN AP and 1 WiMax Base Station (BS) colocated. 10 seconds simulated.

User position	WLAN only	WiMax only	Random	Tx/Qu	Smooth-Tx/Qu
(1) Still, near the AP/BS	18.53	12.76	25.23	32.28	32.37
(2) Still, 30 m far from the AP/BS	3.81	12.76	7.95	16.35	16.40
(3) Moving away at 1 m/s, starting from the AP/BS	11.83	12.76	20.99	24.94	25.01
(4) Near the AP/BS for half simulation, then 30 m far (instantaneously)	10.04	12.61	14.44	18.43	21.03

The most important results reported in Figure 5, 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. This marginal aspect, which is strictly related to the particular medium access control strategy adopted by the WLAN, is neglected in the analytical framework developed in Section 4.

As a final consideration on Figure 5, we can observe that the *Smooth-Tx/Qu* and the *Tx/Qu* strategies are almost equivalent in this case; this is due to the fact that the related curves have been obtained considering a still user.

The impact of user mobility is immediately evident considering the results reported in Table 1, which are related to the same scenario (single user and a WLAN-WiMax heterogeneous network with colocated WLAN AP and WiMax base station) in different conditions. Four scenarios are, in particular, considered:

(1) the user stands still near the AP/base station (optimal signal reception);

(2) the user stands still at 30 m 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);

(4) the user stands still near the AP/base station for half the simulation time, then it moves instantaneously 30 m far away (reproducing the effect of a high-speed mobility).

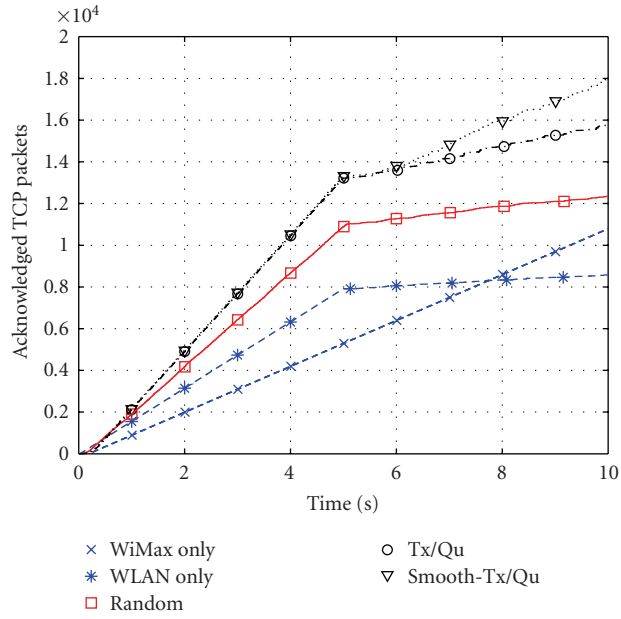

Figure 6: Acknowledged TCP packets of the download performed by one user that instantaneously (after 5 seconds) moves 30 m away from colocated WLAN AP and WiMax BS versus time; single WLAN RAT connection, single WiMax RAT connection, and different distribution strategies are compared.

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 10 seconds 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 30 m, 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 *Smooth-Tx/Qu* method is clearly evident, although the *Tx/Qu* method may be sufficient in most cases.

To get a more accurate picture of the system behaviour in a high mobility scenario, in Figure 6 the amount of acknowledged TCP level packets is shown as a function of the time, in the above-described scenario 4. Please note that the throughput values shown in the fourth row of Table 1 can be obtained from Figure 6 through the following equation: $T = (N_{\text{acked}} \cdot N_{\text{bit}})/D$, where T is the average throughput in

bits per second, N_{acked} is the total number of acknowledged packets, N_{bit} is the number of payload bits per TCP packet (i.e., 1460*8), and D is the total duration of the simulation (i.e., 10 seconds).

Observing the curves related to the *Tx/Qu* and the *Smooth-Tx/Qu* strategies, the effectiveness of the latter approach appears, once more, clearly evident. In the former case, in fact, the splitting probabilities update takes place very slowly in time, thus reducing the total achievable throughput.

6. CONCLUSIONS

In this paper, we faced the issue of RATs integration in tight-coupled heterogeneous networks. The "parallel transmission multiradio 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 access technologies without the need of any modifications to communication protocols.

Here, we proposed original multiradio resource management strategies and derived their performance in extremely relevant scenarios, such as those constituted by WLAN-UMTS and WLAN-WiMax heterogeneous networks.

The main outcomes of our investigations can be summarised as follows:

(i) the parallel transmission allows a total throughput as high as the sum of throughput of the single RATs;

(ii) the parallel transmission generates a disordering of upper layers packets at the receiver side; this is an issue to be carefully considered when the parallel transmission refers to a TCP connection;

(iii) the performance of parallel transmission is very sensitive to the algorithm adopted to split upper layers packet over the considered RATs;

(iv) when different RATs with remarkable difference in achievable throughput are considered, the adoption of parallel transmission as defined in this paper should be preferably avoided;

(v) the proposed dynamic MRRM strategy, in spite of its simplicity, proved to be really effective, fully exploiting the pool of resources provided by the integrated heterogeneous network.

REFERENCES

[1] 3GPP TS 22.234 v8.1.0, "Requirements on 3GPP system to Wireless Local Area Network (WLAN) interworking," June 2007.

[2] Unlicensed Mobile Access specifications, http://www.umatoday.com/.

[3] K. Dimou, R. Agero, M. Bortnik, et al., "Generic link layer: a solution for multiradio transmission diversity in communication networks beyond 3G," in *Proceedings of the 62nd IEEE Vehicular Technology Conference (VTC '05)*, vol. 3, pp. 1672–1676, Dallas, Tex, USA, September 2005.

[4] J. Sachs, H. Wiemann, J. Lundsjö, and P. Magnusson, "Integration of multiradio access in a beyond 3G network," in *Proceedings of the 15th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '04)*, vol. 2, pp. 757–762, Barcelona, Spain, September 2004.

[5] J. Luo, R. Mukerjee, M. Dillinger, E. Mohyeldin, and E. Schulz, "Investigation of radio resource scheduling in WLANs coupled with 3G cellular network," *IEEE Communications Magazine*, vol. 41, no. 6, pp. 108–115, 2003.

[6] L. Badia, C. Taddia, G. Mazzini, and M. Zorzi, "Multiradio resource allocation strategies for heterogeneous wireless networks," in *Proceedings of the Wireless Personal Multimedia Communications Conference (WPMC '05)*, Aalborg, Denmark, September 2005.

[7] R. Veronesi, "Multiuser scheduling with multiradio access selection," in *Proceedingsof the 2nd International Symposium on Wireless Communications Systems (ISWCS '05)*, pp. 455–459, Siena, Italy, September 2005.

[8] M. Saquib, S. Das, G. Mandyam, and M. Z. Win, "Fade resistant transmission over time-varying wireless channels using parallel sequences," in *Proceedings of the IEEE International Conference on Communications (ICC '04)*, vol. 1, pp. 385–389, Paris, France, June 2004.

[9] M. Saquib and M. Z. Win, "Fade-resistant transmission over time-varying wireless channels," *IEEE Signal Processing Letters*, vol. 11, no. 6, pp. 561–564, 2004.

[10] http://www.3gpp.org/.

[11] IEEE Std 802.11a 1999, "Information technology—telecommunications and information exchange between systems—local and metropolitan area networks specific requirements part 11: wireless lan medium access control (MAC) and physical layer (PHY) specifications amendment 1: high-speed physical layer in the 5 GHz band".

[12] IEEE Std 802.16e-2005 and IEEE Std 802.16-2004/Cor1-2005, "IEEE Standard for Local and metropolitan area networks Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands and Corrigendum 1," February 2006.

[13] Wireless Communications Laboratories, Bologna, Italy, http://www.wilab.org/.

[14] A. Bazzi, G. Pasolini, and C. Gambetti, "SHINE: simulation platform for heterogeneous interworking networks," in *Proceedings of the IEEE International Conference on Communications (ICC '06)*, vol. 12, pp. 5534–5539, Istanbul, Turkey, June 2006.

[15] O. Andrisano, A. Bazzi, M. Diolaiti, C. Gambetti, and G. Pasolini, "UMTS and WLAN integration: architectural solution and performance," in *Proceedings of the 16th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '05)*, vol. 3, pp. 1769–1775, Berlin, Germany, September 2005.

[16] S. Floyd and T. Henderson, "The NewReno Modification to TCP's Fast Recovery Algorithm," RFC 2582, April 1999.

[17] J. C. R. Bennett, C. Partridge, and N. Shectman, "Packet reordering is not pathological network behavior," *IEEE/ACM Transactions on Networking*, vol. 7, no. 6, pp. 789–798, 1999.

[18] M. N. Mehta and N. H. Vaidya, "Delayed duplicate-acknowledgments: a proposal to improve performance of TCP on wireless links," Tech. Rep., Department of Computer Science, Texas A&M University, College Station, Tex, USA, December 1997.

[19] A. S. Tanenbaum, *Computer Networks*, Prentice Hall, Upper Saddle River, NJ, USA, 2002.