Target access router selection in advanced mobility scenarios

D. Di Sorte, M. Femminella, L. Piacentini, G. Reali

Dipartimento di Ingegneria Elettronica e dell'Informazione (D.I.E.I.), University of Perugia
via G. Duranti 93, 06125 Perugia, Italy
{disorte, femminella, piacentini, reali}@diei.unipg.it

Abstract

In order to provide nomadic users with QoS-enabled services, advanced mobility management will prove to be of fundamental importance in the future Internet. Mobile IP is the reference protocol supporting layer 3 mobility. However, it has been widely recognised that it could perform poorly, especially with QoS demanding applications. In order to improve Mobile IP performance, the discovery and selection of the target access router to hand over to will play a crucial role.

In this paper, we first give an overview of the ongoing work of IETF in this field, and then we present and evaluate an approach for Candidate Access Router Discovery and Target Access Router selection. The process is fully distributed, multicast-based, and allows timely intra and inter-access router handover, which can, in turn, be intra and inter-technology.

***Index terms*:** IP mobility & QoS, Multicast, Candidate Access Router Discovery, Access Router Selection

INTERNAL TECHNICAL REPORT - UNIVERSITY OF PERUGIA

1 Introduction

The widespread use of the Internet has produced the convergence of heterogeneous wired/wireless networks through a unified IP-based architecture. In addition, the diffusion of applications requiring improved IP network support (e.g., real time services) is expected in the near future. Finally, the need for "information anywhere anytime" is the driving force behind the introduction of wireless communication and portable computing devices. Thus, mobile computing is the merging of recent advances in computing and communication technologies with the aim of providing mobile users with a seamless, ubiquitous computing environment.

Within such a framework, handover functions are extremely important and challenging, in particular if handover must be seamless. In [1], the Authors define the *seamless* handover as "*a handover in which there is no change in service capability, security, and quality*".

IP mobility protocols (i.e. Mobile IPv4/v6, [5][6]) enable mobile nodes (MNs) to execute IP layer handover between access routers (ARs). It is well known that basic Mobile IP (MIP) protocols perform poorly, in particular when supporting real time applications. A number of different approaches to improve MIP have been proposed so far:

1. micro-mobility solutions [7], which aim to limit the handover range of MIP procedures, thus reducing handover latency and signalling burden;
2. context transfer solutions [3][4], the goal of which is to quickly re-establish information states (context) associated with the MN in the new AR upon handover, thus avoiding re-initiation of the signalling procedure to set-up the service from scratch;
3. solutions that minimise packet loss and delay, by modifying MIP protocols especially during registration procedures (fast handover and smooth handover) [10][11].

All the proposed enhancements assume that the new AR to hand over to is known. This is indeed the missing step in the overall seamless handover procedure.

In advanced mobility scenarios, it is important to discover the set of potential target ARs and to select the most appropriate one before handing over. A candidate access router discovery (CARD) procedure collects information about the ARs that are candidates (CARs) for the MN's handover [2][8][9]. It is then possible to identify the Target AR (TAR), i.e., the one that best matches the MN's requirements and the CARs' capabilities.

The CARD requires two main functions to be performed:

1. reverse address translation, that is the binding of the layer 2 identifier (L2 ID) of a new

access point (AP) with the IP address of the CAR connected to it. This would allow inferring knowledge about layer 3 coverage from layer 2 information, thus speeding-up MIP handovers. For this purpose, MNs are supposed to be able to listen to the L2 beacons transmitted by the surrounding APs and to learn their L2 IDs from them;

2. discovery of CAR capabilities, since they are the driving force behind the handover decisions.

In this paper, our first goal is to describe the different approaches proposed within the IETF activities [2][8][9], which now tend towards the definition of both a distributed and a server-based solution (both described in [2]) within the framework of the Seamoby WG [12]. It is worth noting that IETF solutions describe mechanisms which collect and communicate information to enable wireless access selection, but which do not define any type of TAR metric/algorithm.

The second goal of this work is to propose and analyse a procedure for CAR discovery and TAR selection.

We start assuming that the network is controlled by a single operator (Fig. 1 shows the reference network scenario) and that the provider of the IP connectivity (Connection Service Provider, CSP) is also the owner of the network infrastructure (Network Service Provider, NSP). The consequence of this assumption is that without any commercial agreement, there are no logical reasons for CSP to induce mobile customer to change access domain. This would clearly mean a loss of traffic and consequently a loss of revenue in favour of competitors. In addition, why should a CSP provide third parties (i.e., other competitor CSPs) with confidential information regarding its own network access (current bandwidth, security policies, etc)? Consequently, our CARD procedure is confined within a single administrative domain. Then, we also indicate an extension to manage the case of different NSPs coordinated by specific agreements.

Our research is built upon three basic concepts: mobility, reconfigurability, and QoS (Quality of Service).

1. Mobility refers to the capabilities of MNs to access network services from different locations while in motion.
2. Reconfigurability means that terminals include specific capabilities to access heterogeneous radio access technologies (RATs). In our view, the MN may be modelled according to different categories. This classification mainly regards the operation mode of the radio section of the mobile terminal: the first class includes single mode terminals, and the second

class includes multi-mode terminals. Consequently, it is important to use an appropriate handover classification. In this respect, we follow the mobility-related terminology in [1]. In particular, we underline the distinction of layer-2 and layer-3 handover. The former does not alter packet routing at the IP layer, whereas the latter does have an impact on it.

3. QoS is a basic requirement of the next generation application services which are expected to run over IP wireless networks.

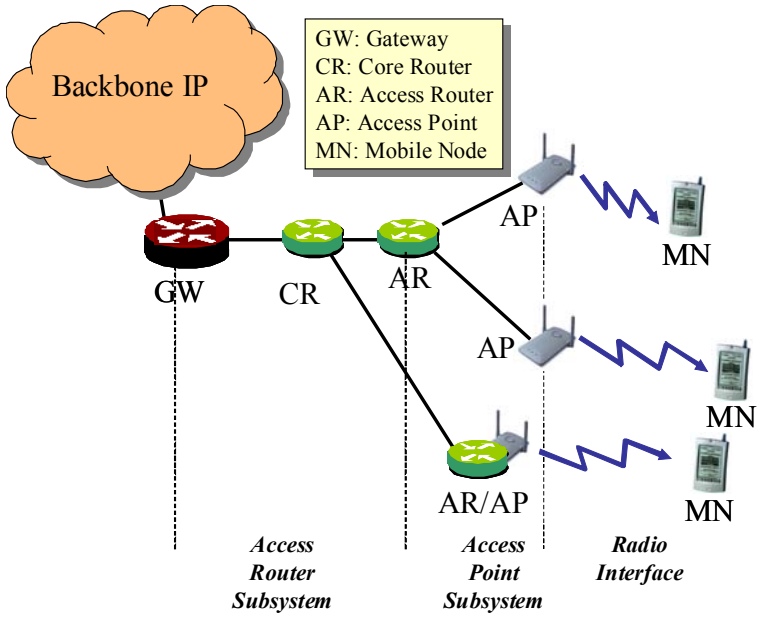


Fig. 1: Reference network scenario.

The basic points and innovations in our proposal are summarised below.

As regards the CARD solution, the proposed approach is fully distributed and enables a local (i.e., relevant to an AR) map of the surrounding wireless coverage to be dynamically self-constructed. The solution is based on the MNs listening to beacons from the surrounding APs (this is a minimum requirement for all L2 wireless technologies). The principle innovation of our CARD approach with respect to the last Seamoby proposals [2] is the use of multicast transmissions for address translation within a global high level multicast group, and a service capability update within local, low-level multicast groups. We use multicast in a push-mode in order to reduce strongly the latency due to explicit queries. Our choice of multicast is also inspired by the intra-domain scope of our CARD/TAR approach, since it is well known that inter-domain multicast transmissions might imply serious problems. In fact, multicast in today's deployment is usually administratively scoped, i.e., multicast transmissions would not leave the

originating domain. To overcome such a drawback, specific agreements among domains involved in the information transfer should be defined. We also present an extension of the CARD procedure in an inter-domain scenario.

As regards the TAR algorithm, which is not explicitly dealt within the Seamoby WG work, we present a novel solution that permits layer 2 and layer 3 handovers. To enter into closer detail, it supports intra and inter-AR handovers, which, in turn, can be intra and inter-technology.

In homogeneous networks, handover is typically driven by metrics strictly related to the received power level [18]. In heterogeneous mobile environments, more complex metrics combining a higher number of parameters have to be defined (e.g., price, bandwidth, priority, power consumption, reliability, etc) [19][20][21].

In this paper, we define a TAR metric which takes into account bandwidth availability and received power level, along with a statistical factor which helps to drive inter-technology handover. As regards the TAR selection procedure, contrary to the IETF proposals, we assume that it is carried out at the current AR. This implies (i) a substantial power saving at MNs, (ii) a reduction in the complexity of terminal equipment, (iii) the management of critical service information only within the fixed network among ARs (security issues), and (iv) a bandwidth saving on wireless links.

We stress that our CARD/TAR solution is suitable for a scenario with single mode and/or multi-mode terminals.

Our goal is to show the effectiveness of the proposed solution in the very challenging situation in which all terminals have a single, active radio interface. This clearly means that multi-mode terminals may have only one interface turned on. The rationale of this choice is that it permits a power saving, but it also implies the remarkable constraint that MNs are not allowed to scan for beacons of different RATs from the one currently being used, thus they cannot obtain information from them. To overcome such a problem, we enhance the solution by means of a statistical approach, the aim of which is to infer about layer 2 coverage.

This paper is structured as follows. In the next section, we summarise the current state of the art about CARD. In section 3, our CARD procedure is described in detail and located within the framework of the IETF activity. In addition, we provide some considerations on security aspects. Finally, we briefly describe how to extend the CARD procedure towards an inter-domain scenario. In section 4, the TAR approach is presented together with the selection metric used. In

section 5, we analyse the performance of the procedure by means of simulations; in particular we show some results about the probability of handover success. Finally, section 6 reports some concluding remarks, together with the description of future work to be carried out.

2 Related Works

As said previously, a CARD solution must cover two different steps: (i) reverse address translation from L2 IDs; (ii) the discovery of service capabilities of related wireless access.

An initial, straightforward solution is to embed the IP address, IP prefix and service capabilities directly in L2 beacons, as suggested in [22]. This type of procedure has two main drawbacks. The former is strictly theoretical: encapsulating layer 3 data within layer 2 control frames explicitly violates protocol layer separation rules. The latter is considered from a practical point of view: if the above strategy were to be used for novel radio technologies, it would at the same time require standard modifications to existing technologies (e.g., IEEE 802.11b).

Consequently, it is necessary to identify a more complex and viable alternative. To overcome the above-mentioned drawbacks, it is necessary to define a network-assisted CARD process. More precisely, a preliminary, straightforward assumption is that each AR should be able to know and keep in a local cache the state of neighbouring wireless accesses (i.e., the pairs AR-AP) that are able to offer an alternative wireless access to an MN currently under its own coverage. On the assumption that the configuration of a wireless access network may be dynamic, then both these steps imply a continuous, dynamic exchange of information among the network entities involved in the CARD process.

Generally speaking, each AR is associated with a number of APs, and, consequently, the coverage area associated with an AR is the one provided by its APs. Two ARs are defined to be neighbours if their wireless coverage areas overlap each other.

The state to be kept by the AR should contain the L2 IDs of the neighbouring APs, the IP address of the relevant AR, and the related service capabilities associated with the pair (AR, AP). To reduce the complexity of table management, the entries of the local cache of an AR are soft states, and unless refreshed within a given amount of time (lifetime), they are deleted.

Therefore, when an MN, under coverage of an AR (current AR), obtains L2 IDs from the beacons of the surrounding APs, it passes them to its current AR. Then, the current AR is in charge of providing the MN with either the IP addresses of CARs and the relevant service capabilities (if the TAR algorithm is run in the MN), or directly the TAR (if the selection

algorithm is carried out in the current AR). This means that, apart from the signalling exchange to resolve L2 addresses between an MN and its current AR, a capability exchange also has to be performed. Consequently, the complete message exchange among the MN, its current AR and a CAR is depicted in Fig. 2. Such an exchange is associated with the phase of the CARD process when the coverage map within the current AR has been completed (*"steady phase"*). The capability exchange between the current AR and CARs (steps 3 and 4 in Fig. 2) may be performed either at an MN request or on expiry of its capability lifetime. CARD ([2]) and dyCARD solutions ([9]) assume that MNs are in charge of triggering/performing the TAR selection.

As regards the secure message exchange, in order to avoid third-party attacks, appropriate IPsec Security Associations (SAs) between an MN and its current AR, and between the current AR and the CAR must be performed [2]. Integrity and authentication for all information from the fixed network (i.e., originating from ARs) must be verified. Confidentiality might also be supported. Furthermore, in order to defeat Denial of Service (DoS) attacks, entities should implement a rate limiting policy concerning the processing of queries.

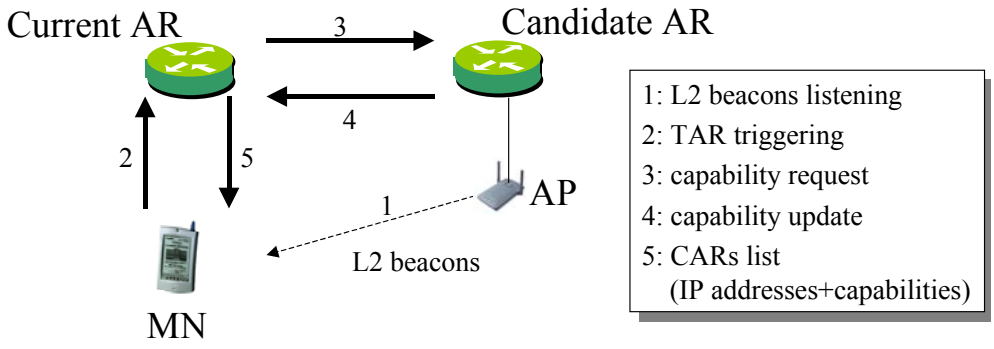


Fig. 2: Steady phase: message exchange for address mapping and capability update.

A very challenging point of a CARD approach is the discovery of the IP addresses of neighbouring ARs ("*discovery phase*"), that is the (L2 ID)→(IP address) mapping when the information is not yet stored in the local cache. In other words, the discovery phase consists of discovering the neighbouring ARs/APs for each AR, so as to be able to build a map of the wireless coverage in the surrounding area at both layer 2 and layer 3. This would also enable a more complete TAR selection process to be performed, which would be able to select not only the new AR, but also the relevant AP.

It is worth noting that a manual configuration of such a mapping within the ARs would not be a good solution for the following reasons:

1. it is not feasible in large wireless networks;
2. it would not permit the exploitation of a variable-topology wireless access network;
3. the coverage area of an AR (i.e., of its APs) could not be easily and precisely defined.

Therefore, a solution for the discovery phase has to be defined with the aim to dynamically build the coverage map at each AR.

In the following, we report the solutions proposed in literature, set within two different general schemes:

1. handover-based solutions, which exploit the first plain MIP handover to infer knowledge about surrounding wireless coverage;
2. solutions which start from the L2 beacon listening of MNs to infer coverage information:
   a. a centralised approach, based on the use of a centralised server;
   b. a distributed approach, based on the use of application level protocol.

2.1 Handover-based CARD approach

The basic idea of a handover-based solution [2][9] is that two ARs are discovered to be neighbouring after the occurrence of a plain handover event between them. More in detail, an AP connected with another AR is discovered after a plain MIP handover which implies a layer 2 handover towards it. After the handover, the MN has the task of sending the new current AR additional signalling information (router identity message) containing the IP address of the old AR, together with the L2 address of the old AP. Thus, the current AR can create a new entry in its cache for the new wireless access. Moreover, a suitable message exchange between the new current AR and the old AR could also make the old AR capable of recognising the current AR as a new neighbour, so as to update its own coverage information as well.

It is worth noting that, according to the MN report, the new current AR can check with the old AR to see whether the MN was indeed attached to it.

A reasonable assumption is that the tables maintained in ARs have to be shared among all MNs, since to maintain a separate table for each MN, though more secure, would imply very high scalability problems, as an MN would be less motivated to provide false information [14].

In the handover-based approach, the first handover between two ARs (bootstrap handover) cannot use CARD, and this may be a weakness for delay-sensitive applications, especially in a

dynamic access network topology.

In addition, the time needed to complete the discovery phase could be highly affected in case of a dense wireless coverage, i.e., when a number of APs cover the same area. In fact, in this case, MNs are driven by the TAR to hand over towards those APs already discovered, and perform a non-driven handover (discovery event) only when there is no alternative.

2.2 Centralised L2 beacon-based approach

This solution was proposed in both [2] and [8]. The basic idea is that ARs must register with a centralised server, indicating its own IP address together with all the L2 IDs of the APs associated with it. This server is a database that is dynamically updated by ARs. Its task is to process queries from ARs to solve L2 IDs and, therefore, to contribute to the building of the wireless coverage map within ARs. L2 IDs are communicated by the MNs to the current AR. Malicious MNs could communicate false information (e.g., L2 IDs of APs which do not have a wireless coverage overlapping with the current AR and which have been listened to in the distant past). This would imply incorrect wireless coverage information within routers, and this might cause inefficiencies in the overall procedure. Therefore, as highlighted in [14], if MNs cannot be trusted, an L2 beacon-based procedure is inefficient from a security point of view.

The server-based solution proposed by the Seamoby WG foresees the extension of the CARD protocol to support an AR-server message exchange, whereas in [8], the Authors exploit the SLP (Service Location Protocol) architecture with the centralised Directory Agent. ARs work as SLP User Agents to send service requests to the Directory Agent.

It is worth remarking that the server clearly represents both a single point of failure and a performance bottleneck for the procedure.

This centralised discovery clearly introduces additional signalling messages between ARs and the server to solve layer 2 addresses. Such a message exchange between ARs and the server must also employ appropriate SAs to ensure integrity, authentication and confidentiality.

2.3 Distributed L2 beacon-based approach

If there is no deployed centralised server, then each AR has to be able to process queries to solve L2 addresses. In [8], the Authors exploit the SLP distributed architecture, where each AR works as both SLP Service Agent and SLP User Agent. Also in this case, integrity, authentication, and confidentiality must be ensured for the signalling exchange among SLP User Agents.

3 The CARD Approach

In this section, we present a proposal for performing the CARD procedure within an administrative domain.

Before entering into the details of the proposed approach, it is necessary to specify the capabilities of the mobile terminal under examination. For the sake of generality, we assume that MNs are either classical single-mode terminal or adaptive terminals. The latter may be either re-configurable terminals based on software defined radio (SDR) capabilities or simply multi-mode terminals. In any case, we assume that they are capable of listening to layer 2 beacons transmitted by surrounding APs. However, in the case of multi-mode terminals, we assume that such terminals do not scan for beacons of different RATs from the one currently used. The rationale of this assumption is that, with a number of network interfaces, scanning for beacons could result in a large drain on power for multi-mode terminals.

The proposed CARD procedure is network-assisted, distributed and based on multicast transmissions.

As regards the discovery phase, we follow a distributed L2_beacons-based approach, thus avoiding the bootstrap handover problem during the discovery phase by means of multicast queries. To this end, the network operator defines a multicast group (MG_{OP}), including all the ARs that currently provide wireless connectivity. These ARs act as multicast hosts, whereas the functions of multicast routing are performed by the routers in the core network. In other words, ARs are the network entities exchanging multicast information. MG_{OP} is used to exchange information about address mapping (IP addres-L2 ID) among ARs.

Then, in order to reduce the time needed to accomplish the TAR process, the procedure provides for continuous updates of the service capabilities, thus avoiding updating them upon handover requests. For this purpose, multicast transmission is used once again. At the lowest hierarchical level, the *ith* access router AR_i builds up another multicast group (MG_i), which includes all ARs with a coverage area overlapping with the coverage area of AR_i. Clearly, we consider the coverage area of each AR as the union of the coverage areas of all APs connected to it. This MG_i is used by the AR_i to efficiently distribute information about the service capabilities of its APs to the geographically adjacent ARs.

3.1 The steady phase

Each AR keeps the information regarding wireless coverage of neighbouring ARs in a local

cache (namely CARD table). We use the following notation: $AP_{i,j}$ denotes the *j*th AP connected to the *i*th AR (AR_i). An example of CARD table within AR_h is reported in Table 1, where:

1. the upper rows of the table specify: (a) the L2 IDs and the RAT identifier of the APs connected to the AR; (b) the relevant service capabilities; (c) a statistic parameter which provides some sort of coverage information at layer 2 (e.g., with reference to Table 1, $Pr_{AP(h,s)\rightarrow AP(h,p)}$ is an estimation of the probability of making a successful layer 2 handover from $AP_{h,s}$ to $AP_{h,p}$). This parameter is clearly used only if the APs refer to different RATs, since in this case it is not possible to rely on beacon listening;
2. the other rows specify: (a) the L2 ID and the RAT identifier of each geographically adjacent AP; (b) the corresponding IP address of the relevant ARs; (c) the associated service capabilities and the parameter reporting the probability of successfully completing an handover. It is worth noting that in this case the probability of success refers to layer 3 handover, since two different ARs are involved. With reference to Table 1, the value of the parameter $Pr_{AP\ (h,s)\rightarrow AP\ (x,k)}$ is the probability of making a successful (layer 3) handover between AR_h and AR_x, due to the MN transfer from $AP_{h,s}$ to $AP_{x,k}$ at layer 2.

IP_{AR}	**$AP_{L2\ ID}$**	**RAT**	**SC**	**Prob $AP_{h,s}$**	**Prob $AP_{h,t}$**	**….**	**Prob $AP_{h,n}$**
AR_h	$AP_{h,p}$	0	$SC_{h,p}$	$Pr_{AP(h,s)\rightarrow AP(h,p)}$	$Pr_{AP\ (h,t)\rightarrow AP\ (h,p)}$	…	$Pr_{AP\ (h,n)\rightarrow AP\ (h,p)}$
	…	…	…	…	…	…	…
AR_x	$AP_{x,k}$	2	$SC_{x,k}$	$Pr_{AP\ (h,s)\rightarrow AP\ (x,k)}$	$Pr_{AP\ (h,t)\rightarrow AP\ (x,k)}$	…	$Pr_{AP\ (h,n)\rightarrow AP\ (x,k)}$
	…	…	…	…	…	…	…
AR_y	$AP_{y,j}$	1	$SC_{y,j}$	$Pr_{AP\ (h,s)\rightarrow AP\ (y,j)}$	$Pr_{AP\ (h,t)\rightarrow AP\ (y,j)}$	…	$Pr_{AP\ (h,n)\rightarrow AP\ (y,j)}$

Table 1: Example of CARD table stored in AR_h.

To reduce the complexity of table management, the entries of the CARD table are soft states which are deleted if they are not refreshed within a given time interval. This is particularly

helpful in the case of a dynamic network access topology.

As regards statistical information, let us consider for instance $Pr_{AP\ (h,s)\rightarrow AP\ (x,l)}$; such a value is simply evaluated as the number of successful handover events between these two APs (and ARs), i.e., from $AP_{h,s}$ to $AP_{x,l}$, divided by the total number of attempts. Both the numerator (successful handovers, $N_{HO,SUCC}$) and the denominator (total number of handover attempts, $N_{HO,TOT}$) are initialised to 1 for each column in the table stored in AR_h, so that the corresponding probability is, in turn, initialised to 1/1=1. However, as previously said, these fields are meaningful only for APs exploiting different RATs.

In general, the service capability has to be intended as the set of parameters (available bandwidth, price, security parameters, etc.) that characterises the network service from the MN to the AR port towards the core network. In our approach, we consider as service capabilities the available bandwidth (referred to in the following as SC) of the link from the wireless interface of an AP to the output port of the relevant AR only. The reason for this choice is that, as this network is under the control of a single operator, other factors, such as connection price and security associations, seem to be of less interest than the available bandwidth, which is the basic parameter to perform admission control for QoS-enabled services.

As regards service capabilities update, the use of low level multicast groups aims to manage the geographical proximity of ARs[1]. In particular, the AR (e.g., AR_h) managing the multicast group (MG_h) multicasts update messages of the service capabilities associated with its APs. This operation may take place periodically and upon significant variations of the service capabilities associated with an AR_h-$AP_{h,x}$ pair. Such an update arrives at all the ARs joining MG_h, that is all the ARs which have coverage areas partially overlapped with that of AR_h. This process enables ARs to continually update the service capabilities of their neighbours. In other words, at a TAR request, messages n° 3 and n° 4 in Fig. 2 are unnecessary, since the service capability update process runs in the background and is based on the local exchange of multicast messages. Such a message exchange is reduced in a network scenario that is not highly dynamic.

For what concerns the information about the coverage, it is not provided to ARs in a static way, but it is dynamically learned on the basis of the knowledge of the L2 connectivity from MNs. In the following, we show how a new entry is added to the table.

3.2 The discovery phase

We present a procedure capable of automatically self-constructing the geographical coverage mapping at each AR.

At the highest level, the network operator defines a multicast group (MG_{OP}), including all the ARs that currently provide wireless connectivity and act as multicast hosts. MG_{OP} is used to resolve the IP address of the AR from the L2 ID of one of its APs. For this purpose, when an AR starts offering wireless connectivity through some APs, our proposal is that it has to join the MG_{OP} and to multicast to all other participant ARs its IP address and the L2 IDs of the active APs under its IP scope. In turn, one of the participants (e.g., the last AR that joined the MG_{OP} and is still active) replies in unicast with the address mapping relevant to all participant ARs. Clearly, since the coverage area of the new AR does not typically overlap with coverage of all ARs, only a subset of such an information may be useful to build the CARD table. Nevertheless, it is worth noting that the amount of data to exchange in the network and to manage within ARs is very limited and simple. For instance, in the case of an IPv6 network with 20 ARs, each one with 10 active 802.11 APs, the multicast packets have a payload of 76 bytes, whereas the unicast reply has a maximum payload of 1444 bytes. If the number of ARs increases tenfold, then the maximum size of unicast is equal to about 15 Kbytes, i.e., about a quarter the maximum IP datagram size. This is also the maximum storage requirement within ARs to maintain these data.

Clearly, active ARs are in charge to communicate over MG_{OP} all variations in their radio coverage (e.g., activation/deactivation of APs). We remark that this mechanism allows speeding up the address resolution phase, avoiding the latency associated with the interrogation to a remote database.

Thus, once each active AR has the complete address mapping in memory (address list[2]), the main steps of the discovery phase are the following, as illustrated in Fig. 3:

1. when an MN, located under coverage of the *s*th AP connected to AR_h ($AP_{h,s}$), enters the coverage area of another AP (say, $AP_{z,k}$) connected to AR_z, if the APs involved use the same RAT, the MN listens to the beacons transmitted by this new AP;
2. the MN notifies its current AR (AR_h) of the L2 ID of the new AP, by means of the currently

---

[1] It is worth to note that, if the operator prefers to limit the number of multicast groups to be managed within the domain, then, at each AR, the usage of multicast might be easily replaced by *n*-times unicast, since the IP addresses of adjacent ARs are known. The drawback is the consequent loss of efficiency in terms of network resources.

[2] Please note that this list is not the CARD table (which is more complex and structured), but it separately stored.

serving AP ($AP_{h,s}$);

3. if the detected AP does not appear in any row of its table, AR_h gets the IP address directly from the address list and asks AR_z for service capabilities relevant to $AP_{z,k}$. In addition, it invites AR_z to join its own multicast group (namely, MG_h) and sends the service capabilities associated with $AP_{h,s}$;
4. AR_z sends a unicast reply, containing the service capabilities associated with the AR_z-$AP_{z,k}$ pair and the invitation to AR_h to join MG_z;
5. the process ends when AR_z and AR_h join the multicast groups MG_h and MG_z, respectively.

Clearly, the reciprocal invitations to join the local multicast groups are sent only when two ARs are discovered neighboring, i.e., at the first time they discover to have two APs overlapping.

As regards step n° 2, a question arises: when does the MN communicate to the current AR the list of L2 IDs? This event surely occurs at the time of a TAR event. In this case, the MN communicates the list of the L2 IDs (referred to as the TAR beacon list in the following) listened to a very short time ago (i.e., the last T_{beacon} seconds), because these identities represent candidate wireless access for a possible handover, and must be considered for the TAR selection[3].

As regards step No. 3, if the L2 ID is not found in the address list, this implies two different cases: (i) the AP does not belong to the CSP (i.e., belongs to a competitor); (ii) the procedure, for some unexpected reasons, failed. Then, AR_h is allowed to issue a multicast request over MG_h with the L2 ID of this AP, asking for an answer from the (unknown) AR which manages it, which, in turn, in charge to send the unknown mapping. This choice is reasonable to limit the scope of the signalling towards geographically adjacent ARs. If the MG_h is not created yet, then the query is either forwarded to a AR randomly chosen among those in the address list or multicasted over MG_{OP}. In any case, if the AP belongs to a competitor CSP, then this means that no answer is received, and such an L2 ID is stored in a "black list" in order to avoid future unnecessary queries.

Then, message No. 4 produces the update of the CARD table maintained in AR_h (see Fig. 4).

It is worth noting that the signaling load needed to carry out this procedure could be lowered.

[3] Optionally, the MN could also maintain a list of all L2 IDs received whilst under the current AR handling. This list may be useful in building the wireless coverage map, and has to be deleted at layer 3 handover. The MN is in charge of sending the current AR the list of the L2 IDs not yet previously communicated. This may occur either (i) at a TAR event (together with, and clearly separated by the TAR beacon list) or (ii) periodically, with a time period

Once two ARs have discovered to be neighbors (i.e., the above procedure has been executed for the first time), in principle they do not need to exchange messages for communicating each other the specific APs overlapping. In fact, they already know all L2 IDs associated with each AR from the address list. In addition, they have already reciprocally joined the relevant local multicast groups. This implies that the SC of the newly added AP in the CARD table will be available upon the next update from the peer AR (recall that each AR multicasts the SCs relevant to all its APs). The drawback is that if a TAR selection has to be executed before such a scheduled update, the SC relevant to such an AP misses and thus it has to be retrieved on request. To sum up, the procedure enables the geographical proximity of an AR to be associated with its participation in a given multicast group. The advantage of such a procedure is that it is able to automatically self-construct this geographical mapping and also to react to variations in coverage (e.g., activation/deactivation of APs).

It is worth noting that the procedure enables the coverage map to be extended across all RATs under the assumption that each RAT is configured as active on a subset of the terminals moving in the relevant area. In addition, it is worth to note that an MN can receive beacons (and send them to the current AR) from all RATs during inter-technology handover attempts when scanning different RATs.

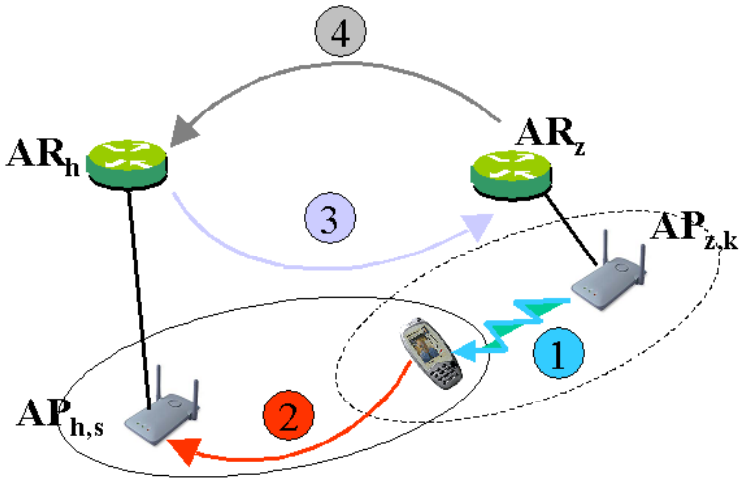
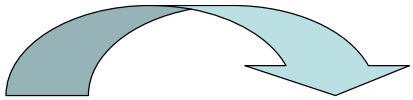


Fig. 3: Message exchanges when a new AP is detected by an AR by means of beacon listening by an MN.

$\widetilde{T}_{beacon}$, only if *new* L2 IDs have been listened to. The latter option would result in a larger use of signalling and would be especially useful to better follow wireless network changes in the transient phase.



IP_{AR}	$AP_{L2\ ID}$	RAT	SC	Prob $AP_{h,s}$	...
AR_h	$AP_{h,p}$	0	$SC_{h,p}$	$Pr_{AP(h,s)\rightarrow AP(h,p)}$	...
	...	...	...	...	...
AR_x	$AP_{x,k}$	1	$SC_{x,k}$	$Pr_{AP(h,s)\rightarrow AP(x,k)}$	...
	...	...	...	...	...

IP_{AR}	$AP_{L2\ ID}$	RAT	SC	Prob $AP_{h,s}$	...
AR_h	$AP_{h,p}$	0	$SC_{h,p}$	$Pr_{AP(h,s)\rightarrow AP(h,p)}$	...
	...	...	...	...	...
AR_x	$AP_{x,k}$	1	$SC_{x,k}$	$Pr_{AP(h,s)\rightarrow AP(x,k)}$	...
	...	...	...	...	...
AR_z	$AP_{z,k}$	2	$SC_{z,k}$	$Pr_{AP(h,s)\rightarrow AP(z,k)}$	...

Fig. 4: CARD table update.

3.3 Security considerations

As underlined in [16][17], in order to extend MIP for advanced network services, some additional mechanisms are needed by the foreign agent not only to verify the identity of the mobile node, but also to authorise connectivity based on local policy or on the ability to pay. RFC 2977 [16] describes an infrastructure which enables AAA servers to authenticate and authorise network access requests from MNs, whereas, in [17], the authors propose extensions to MIP registration messages to create Mobility SA between the MN and its home agent, and/or between the mobile node and a foreign agent. Consequently, since we are proposing a procedure to support advanced network services, we therefore assume that MNs are trustworthy. This means that coverage information provided by them and used by the network to build wireless coverage maps is reliable.

Note that, under this assumption, false or inexact coverage information may be provided from inaccurately configured MNs only. This could imply a mistaken expansion of the CARD tables and additional signalling traffic in the network. Thus, possibly incorrect TAR selections might be performed. These incorrect handover decisions may affect: (i) inaccurately configured MNs only for an intra-technology handover; (ii) other MNs for an inter-technology handover, but only for a

limited number of times, i.e., until the successful handover probability between two APs approaches zero, thus implying a low value of the TAR metric associated with a given wireless access and thus a very low probability to perform the relevant handover attempt. Therefore, the presence of such a statistical parameter also permits ARs to maintain the information stored in their CARD tables coherent with the real coverage overlapping. In addition, as mentioned above, CARD tables entries are soft states. This means that incorrect (and hopefully sporadic) information is not refreshed, and will then be deleted after a given lifetime.

It is also worth noting that a remarkable implication of our design choices is that confidential network data does not travel over the wireless interface. This implies an improved network protection from the action of possible malicious MNs. Effects of possible DoS attacks can be limited by means of rate limiting policies about query processing at the current ARs.

Clearly, to support unicast and multicast secure message exchange among entities, an IPsec-based solution has to be exploited to ensure authentication, integrity and confidentiality. The extension of the standardised point-to-point IPsec architecture towards multicast transmission is currently being scientifically investigated (e.g., see [15]).

3.4 An extension of the CARD solution towards an inter-domain scenario

So far, the reference scenario of this paper is characterised by a single Network Service Provider (NSP) which is the owner of the network infrastructure. It is also in charge of providing mobile terminals with IP connectivity. In this sense, the NSP also may be considered as a Connection Service Provider (CSP). Consequently, the scope of the CARD procedure in this scenario is limited to the wireless coverage provided by the NSP/CSP, since there are no reasons that could lead the operator to support the mobile customer to change the access domain. In fact, this would clearly mean a loss of traffic and consequently a loss of revenue in favour of competitors.

However, there is another typical network/business model which deserves to be considered. Such a scenario is characterised by a CSP which, under specific contractual agreements with a number of different NSPs, is allowed to access their network infrastructure in order to be able to provide mobile subscribers with wireless IP connection. This means that an MN currently handled by the CSP may access Internet services through ARs belonging to different administrative domains. Consequently, the scope of the CARD process has to be extended to support handovers among those administrative domains which have stipulated the agreement with the considered CSP, and provide overlapping radio coverage. In order to fulfil this additional task, the CARD process

should be modified to overcome drawbacks due to multicast transmission through different administrative domains. A possible solution is briefly described below.

The first step is the introduction of a new network entity within each NSP, named as CARD gateway (CGW), the goal of which is to manage inter-domain CARD signalling exchange. We assume that CGW belongs to the MG_{OP}, so that it has knowledge of address mappings within the NSP.

As regards the discovery phase, CGW is polled by a requesting AR only if it did not succeed in resolving a specific L2 ID[4]. Then, on reception of the query regarding a given L2 ID from the AR, the CGW forwards (unicasts) such a request towards the CGWs of those NSPs that may have a radio coverage overlapped with the one of the current NSP (at this level, a manual configuration of CGWs may be envisaged). One of the polled CGWs finds in its copy of the address list the requested mapping and replies communicating the IP address of the AR relevant to the L2 ID.

The message exchange to support the service capabilities update in the steady phase is also performed through CGWs. More precisely, if ARs of different NSPs (e.g., AR_z of NSP_2 and AR_k of NSP_3) are present in the CARD table of a given AR_h of NSP_1, then the MG_h includes CGW_1, which is in charge of forwarding service capabilities updates towards CGW_2 and CGW_3. In order to avoid wasting network resources, CGW_1 has to maintain mapping between resolved pairs of L2 IDs- IP addresses and the CGWs which provided such information. At the same time, CGW_2 and CGW_3 must configure proper multicast groups within their NSPs in order to deliver a capability update relevant to *external* ARs with the minimum amount of signalling.

To go into more detail, the main steps of the inter-domain CARD procedure are summarised below:

1. when an MN located under the coverage of $AP_{h,s}$ enters the coverage area of $AP_{z,k}$, if the involved APs use the same RAT, the MN listens to the beacons transmitted by $AP_{z,k}$;
2. the MN notifies AR_h of the L2 ID of $AP_{z,k}$;
3. if the detected AP does not appear in any row of its table, AR_h issues an intra-domain multicast request over MG_h with the L2 ID of such an AP;
4. if no answer is received within a time-out, then AR_h sends a unicast request to its CGW (namely, CGW_1);

[4] See also the discussion at the end of sub-section 3.2.

5. CGW_1 forwards (unicasts) the request towards the CGWs of those NSPs that may have a radio coverage overlapped with the one of NSP_1;
6. CGW_2 on behalf of CGW_1 is able to forward the request of AR_h to the right AR (i.e., AR_z);
7. AR_z within NSP_2 sends a unicast reply, containing the notification of its IP address and of the service capabilities associated with the AR_z-$AP_{z,k}$ pair;
8. CGW_2 (i) forwards the reply to CGW_1, and (ii) creates (or updates, if already present) a local multicast group associated with AR_h (that issued the request to CGW_1) and invites AR_z to join this group. Such a group is used for future service capability updates from AR_h within NSP_2;
9. CGW_1, in turn, (i) forwards the reply to AR_h, and (ii) creates an association between AR_z and CGW_2;
10. the answer of AR_h consists of: (i) the notification of its IP address, of the L2 ID of the AP involved in the process ($AP_{s,h}$) and of the service capabilities associated with the AR_h-$AP_{h,s}$ pair; (ii) the invitation to CGW_1 to join its own multicast group MG_h;
11. at this time, steps 8, 9, and 10(ii) are repeated (exchanging index 1 with 2 and vice versa, and h with z and vice versa) to include AR_h within the destinations of capability updates from AR_z.

Clearly, SAs among CGWs must be implemented to ensure authentication, integrity and confidentiality of data exchange.

We underline that the depicted procedure to manage inter-domain handover might also be useful to manage a single, large administrative domain as a number of smaller, coordinated sub-domains.

4 The TAR Approach

As regards the TAR selection procedure, we assume that it is performed at the current AR. This implies (i) substantial power saving at MNs, (ii) avoiding highly complex terminal equipment, (iii) managing critical service information only among ARs (security issues), (iv) avoiding bandwidth waste on wireless links, (v) it is compliant with some specific traffic management policies of the operator.

Each handover (layer 2 and/or layer 3) could be an inter-technology handover. Consequently, it could be that the “best” AR is reachable only through an AP of a different RAT from the one currently used by the MN. In this case, since the MN is unaware of being under coverage of such

an AP, it is impossible to simply rely on beacon listening to make the "best" AR choice. We have solved this problem by means of a stochastic approach that enhances the basic CARD approach based on beacon listening only. In fact, it is suitable only for single mode terminals and multi-mode terminals with all different radio interfaces turned on.

4.1 Handover management

TAR selection may be triggered by three different events:

1. upon explicit request by the customer;
2. periodically, as decided by the current AR, with period T_{TAR}, which can be dynamically adjusted by the AR on the basis of the current load, with the aim of spreading it evenly over the APs. A viable proposal is that each MN is assigned a period equal to

$$T_{TAR} = \max\left\{ T_D \cdot \left(1 - e^{-\delta \frac{SC}{C}} \right),\ T_{TAR,\min} \right\}, \tag{1}$$

   where SC is the service capability, i.e., the amount of bandwidth available in the link from the wireless interface of the AP to the output port of the relevant AR, C is the capacity of the AP, T_D and δ are design parameters. The higher the available network resources, the longer the time period T_{TAR}. The value of T_{TAR} is first determined at connection set-up, then it is updated at each TAR event according to the current load of the AR-AP pair the MN is attached to. When the amount of traffic increases, the frequency of TAR events also increases (up to a maximum fixed value of $1/T_{TAR,\min}$) with the aim of increasing the reactivity of the CARD/TAR procedure to track system changes. Clearly, this implies a more aggressive load balancing action, but at the cost of increased complexity in terms of signalling and computational burden;
3. upon explicit request from the MN, when it detects that the power level of the received signal is rapidly decreasing. In particular, this request is sent when the received power PW is below a threshold PW_{opt}. We consider the MN outside the coverage area when PW falls below the receiver sensitivity, PW_{min}.

Since the handover is managed by the AR, the MN is required to send its handover preferences to the current AR in accordance with the approach proposed within the IETF working group Seamoby [2], whenever one of the events listed above occurs. Preferences for events 1 and 3 listed above are sent together with the handover request; for event 2, the MN is expected to send

them to the AR upon explicit request, in order to permit it to execute the TAR.

In our scenario, three types of information are sent as preferences:

1. which types of RAT the MN is able to hand over to (this information could also be maintained by the AR in the MN context);
2. the set of APs (i.e., their L2 IDs) the beacons of which have been received by the MN during the latest T_{beacon} seconds (i.e., the TAR beacon list); clearly, this information is relevant only to the surrounding APs of the same RAT of the MN;
3. the received power level *PW* of the surrounding APs of the TAR beacon list.

At this point, the AR has the information needed to run the TAR on behalf of the MN (note that it could be necessary to solve some L2 IDs and consequently discover some service capabilities, as depicted in Fig. 3). If different ARs are CARs, a subset of them will have to be selected as target access routers. The proposed TAR algorithm, presented below in section 4.2, aims to select a subset of maximum *N* CARs (*N* is a design parameter), according to a score standing, computed by means of a metric. The AR-AP pair at the top of the list (with the best score) is the first choice. The current AR transfers the MN context [3][4] to all the ARs returned by the TAR. In addition, the current AR notifies the MN of the list and relevant order of ARs (with the relevant APs) to which it should hand over to. At this point, the MN can try the handover. Two types of handover can occur: intra-technology or inter-technology.

As regards the former, the handover involves an AR that manages an AP belonging to the same RAT as the one to which the MN is currently attached. As is explained in more detail in the next section, this procedure might either re-select the current AP (as handing over is not convenient) or an AP included in the list of preferences sent by the MN. Thus, the MN almost certainly[5] maintains radio connectivity, and as it knows the IP address of the destination AR, it may successfully conclude the handover.

It is worth noting that the candidates are not only the APs whose beacons have been communicated by the MN (i.e., those with the same RAT as the one currently active in the MN), but also the APs with other RATs, the coverage areas of which are supposed to overlap with that of the current AP. As mentioned above, this information is inferred from the relevant values of probability of achieving a successful handover stored in the CARD table of the current AR. The

[5] We rely on the fact that the position of the MN is not notably changed in the last T_{beacon} seconds; in addition the TAR selection algorithm takes into account the received power from a given AP.

current AR notifies the MN of the list of ARs (with the relevant APs) to hand over to, together with the relevant order. At this point, the MN can try the handover (intra-technology or inter-technology). If the MN performs a layer 2 inter-technology handover, the old AR (in the case of a concurrent layer 3 handover) or the current AR must be made aware of any successful and failed inter-technology handover. This enables this particular AR to update the value of the handover success probability in its CARD table. In the former case (layer 2 and 3 handover), the new, current AR sends a positive acknowledgement to the old AR whenever a handover of a previously handled MN is successful. In the latter case (intra-AR layer 2 handover), the current AR does not change, and it recognises the layer 2 hand-over between two of its APs. As regards any failed handover attempts, the old AR does not receive any positive acknowledgement, thus it is also able to infer these failures from the TAR list communicated to the MN and from the positive acknowledgement it may have received. Consequently, the AR is always able to update the values of successful handover probability in its CARD table. It is clear that the additional processing and signalling burden is the price to pay in order to make the procedure capable of managing not only intra-technology, but also inter-technology handovers.

Finally, it is important to consider that since the transferred context is short-term (soft states), if the MN does not perform a layer 3 handover by a given amount of time, the context is deleted.

In the following sub-section, we describe the TAR selection algorithm, the aim of which is to provide the N most appropriate CARs for the handover. The selection process is driven by a suitable metric which takes into account the load balancing criterion, the received power level, and the estimated values of successful handover probability between involved APs.

4.2 The TAR selection algorithm

The metric used represents the criterion which provides the quality measure $M_{TAR}\left(AP_{h,s}, AP_{z,k}\right)$, of a supposed handover from the current $AP_{h,s}$ to a target $AP_{z,k}$. TAR driven handovers may occur only towards the APs which have an entry in the tables managed by the current AR. Our proposal is that the score associated with a candidate AR_z-$AP_{z,k}$ pair for a handover from the $AP_{h,s}$ is given by

$$M_{TAR}\left(AP_{h,s}, AP_{z,k}\right) = f_1\left(SC_{z,k}\right) \cdot f_2\left(PW_{z,k}\right) \cdot f_3\left(Pr_{AP(h,s)\rightarrow AP(z,k)}\right), \quad (2)$$

where the proposed functions f_i, $i=1,2,3$ are:

$$f_1(SC_{z,k}) = \begin{cases} \min\left\{1, e^{\beta\left(\frac{SC_{z,k}+(-1)^j B}{C}\right)} - 1\right\} & if\ SC_{z,k} \geq (1-(-1)^j)B/2 \\ 0 & otherwise \end{cases}, \quad (3)$$

$$f_2(PW_{z,k}) = \begin{cases} 1 - e^{-\gamma\left(\frac{PW_{z,k}-PW_{\min}}{P_T - PW_{z,k}}\right)} & if\ PW_{z,k} > PW_{\min} \\ 0 & otherwise \end{cases}, \quad (4)$$

$$f_3(Pr_{AP(h,s)\to AP(z,k)}) = Pr_{AP(h,s)\to AP(z,k)} = \frac{N_{HO,SUCC}}{N_{HO,TOT}}. \quad (5)$$

Since all the three proposed functions range between 0 and 1, the final score will also range between 0 and 1.

With reference to equation (3), B represents the MN bandwidth demand to satisfactorily support the current communication session, $SC_{z,k}$ is the service capability of $AP_{z,k}$, C is the capacity of $AP_{z,k}$, and β is a design parameter. The value of j is set to 1 for all the candidates but the current one, for which j=0. The higher the value of β, the higher the score associated with the available bandwidth ($SC+(-1)^jB$) after the hypothetical execution of the handover, normalised by the capacity of the AP taken into account. In addition, if the new network access (i.e., the new AR-AP pair) cannot accommodate such a traffic flow with the necessary bandwidth, its score is zero. This is a sort of admission control function executed only at the AR.

A discussion regarding the choice of the two functions f_2 and f_3 will be further discussed in the next section.

With reference to (4), P_T is the standard value of the transmission power associated with a RAT, $PW_{z,k}$ is the received power from the $AP_{z,k}$, and γ is a design parameter similar to β. Clearly, the higher the value of γ, the higher the sensitivity of the score function to power levels below the optimum value PW_{opt}.

With reference to (5), $Pr_{AP(h,s)\to AP(z,k)}$ is the estimation of the handover success probability from $AP_{h,s}$ to $AP_{z,k}$. This probability value should approximate the percentage of the $AP_{h,s}$ coverage area which overlaps the $AP_{z,k}$ coverage area. This metric gives a low score to the APs, candidates for inter-technology handover, with low overlapping coverage areas with the one of the current AP. It is worth noting that, for APs of the RAT currently used, f_3 is always set to 1 (since the overall process relies on beacon detection), whereas, in the case of different RATs, f_2 is always

set to 1 (since the MN does not estimate any power level).

The best N candidates (AR-AP pairs) are selected. In order to have at least one CAR with radio connectivity of the current RAT, the best AR-AP pair in the TAR beacon list must be included at least at the Nth rank of the handover list. If such a candidate is present at a higher position than N, the Nth rank is freed from this constraint. This policy allows minimising the probability of handover failure after scanning all the N candidates provided by the current AR. In fact, only if a substantial movement of the MN has occurred in the last T_{beacon} seconds, the intra-technology handover could fail. Clearly, only the AR-AP pairs with a score higher than the one of the current wireless network access are considered. Then, the handover list is passed to the MN for handover attempts.

Finally, to determine the failure probability, Pr_F, of the CARD/TAR procedure, we denote Pr_E as the probability that the TAR process returns an empty TAR list, Pr_N as the probability of successfully handing over within the Nth attempt, and Pr_{Rec} as the probability that the MN, from the coverage of the previous AP, finds another AP to hand over to by means of self-reconfiguration, i.e., with a plain MIP handover without any assistance from the network. Then, it is easy to find that

$$Pr_F = (1 - Pr_E) \cdot (1 - Pr_N) + Pr_E \cdot Pr_{Rec}, \quad (6)$$

where $(1 - Pr_E) \cdot (1 - Pr_N)$ is the probability that the handovers towards the TARs fail, and $Pr_E \cdot Pr_{Rec}$ is the probability that a plain MIP handover succeeds despite an empty TAR list.

5 Numerical Results

In this section, we present some of the results obtained from a simulation trial, the aim of which was (i) to test the effectiveness of the proposed CARD/TAR approach in terms of handoff success probability, and (ii) to test the sensitivity of performance to variations of system parameters.

First we illustrate the network topology used in the simulation together with the mobility model adopted, and then we discuss the numerical results obtained by simulation.

5.1 Simulation scenario

The network topology simulated consists of a single administrative domain with a number of APs and ARs. We envisage the presence of three different RATs[6]. The main peculiarities of RATs are reported in Table 2. The coverage areas of APs are assumed to be circular with a maximum radius denoted as $r_{\max}$. The power threshold PW_{opt} corresponds to the one perceived at a distance equal to 75% of the maximum coverage radius and it is calculated by using a quadratic attenuation law. This simple propagation model does not invalidate the generality of the CARD/TAR results. In fact, what is important for analysing the proposed algorithms is to define the boundaries of the coverage areas and the power threshold, regardless the specific propagation laws, whose study is beyond the scopes of this paper. The value of the parameters of the model has been chosen in order to focus on the performance of the proposed approach.

The whole network area is a square with a side measuring 150 meters. We run simulations over three different coverage maps.

In the first coverage configuration (referred to as "sparse" coverage map), all the area is covered by at least one RAT. This assumption implies that only a multi-mode terminal can be moved anywhere within the area without losing the wireless connection. The number of APs is 44 (27 of RAT_1, 11 of RAT_2, 6 of RAT_3) and the number of ARs is 8.

The second coverage (referred to as "medium" coverage) is characterised by a higher amount of wireless access resources than in the previous one. The total number of APs is 50 (30 of RAT_1, 13 of RAT_2, 7 of RAT_3) and the number of ARs is 10.

Finally, the third network configuration (referred to as "dense" coverage) is characterised by a coverage which ensures that the whole area is covered by at least one AP for each RAT; this means that each RAT provides a full coverage of the area. Consequently, a single mode terminal can also be moved anywhere within the area without losing the wireless connection. The number of APs is 72 (42 of RAT_1, 19 of RAT_2, 11 of RAT_3) and the number of ARs is 14.

[6] Here, we assume "ideal" layer 2 technologies, without considering either the details relevant to transmission aspects (e.g., fading and shadowing) or MAC protocol features.

	RAT_1	RAT_2	RAT_3
f (frequency)	5.4 GHz	2.4 GHz	2 GHz
$\mathbf{r_{max}}$ (maximum radius)	20 m	30m	40 m
C (capacity)	10 Mb/s	5 Mb/s	2 Mb/s
$\mathbf{G_T}$ (transmission gain)	1	1	1
$\mathbf{G_R}$ (reception gain)	1	1	1
$\mathbf{P_T}$ (transmitted power)	100 mW	100 mW	100 mW
$\mathbf{PW_{opt}}$ (power threshold)	8.7 nW	19.5 nW	15.8 nW
$\mathbf{PW_{min}}$ (receiver sensitivity)	4.9 nW	11 nW	8.9 nW

Table 2: Characteristic parameters of radio access technologies.

The association AR-APs in the network has been designed according to the following criteria: (i) an AR has to be connected to APs of different RATs (at least 2); (ii) APs connected to the same AR should be very close to each other, possibly with overlapping coverage; (iii) the minimum distance between two APs of the same RAT is equal to the maximum radius reported in Table 2. This condition is used to limit the coverage overlapping of APs with the same technology. In addition, the coverage maps have been designed so that the wireless network resources are uniformly distributed over the area. The output capacity of ARs is set at 45 Mbps.

Simulation results are relevant to homogeneous flows. Each flow is associated with a bandwidth value (B) equal to 128 Kbps. The scenario is characterised by a high density of MNs. The number of MNs is 1800. An MN can be in one of the three following states: (i) disconnected (DISC); (ii) connected (CONN); (iii) in communication (COMM). Calls are generated by a Poisson arrival process with average frequency λ, and the call duration is modelled as an exponentially distributed random variable with mean $1/\mu$. In order to test the procedure and to limit the number of rejected calls and failed handovers due to lack of resources (especially in the middle of the square area where the density of MNs is higher than at the borders), we loaded the network with an average amount of traffic equal to 30% of the total capacity. We set the average call duration equal to 5 min (i.e., $\mu = 1/300$ s^{-1}), thus λ is equal to 2.50 s^{-1}, 2.96 s^{-1}, and 4.2 s^{-1}, corresponding to an average load of 740, 888, and 1258 Erlangs, relevant to the sparse, medium, and dense coverage, respectively.

As regards the CARD/TAR, we ran three simulations with a maximum number of handover attempts towards TARs equal to 2, 3, and 4, respectively.

The TAR selection is triggered by either the MN, when the received power level from the current AP drops below the threshold PW_{opt}, or periodically by the current AR, with period T_{TAR}, computed from (1), with values of T_D equal to 15, 30, and 45 seconds, $T_{TAR,\min} = 0.1 \cdot T_D$ seconds, and $\delta = 2 \cdot \ln 10$ (so that $T_{TAR}(SC/C = 0.5) = 0.9 \cdot T_D$). Moreover, T_{beacon} was set to 0.5 seconds. The power control in MNs is assumed to be executed every second.

In addition, the parameters β and γ in the TAR function were selected according to the following criteria:

- for each RAT_i, i=1,2,3, the value of β is set in such a way that when the amount of the available bandwidth after the potential hand-over (*SC*-*B*, for the new AR, and *SC*+*B*, for the current one) is higher than, or equal to $0.8 \cdot C_i$, C_i being the full capacity of RAT_i, then the weight associated with the service capability is equal to 1. This implies that the value of β is equal to β=0.866 for all RATs. For values of service capability lower than $0.8 \cdot C_i$, such a weight rapidly decreases, and consequently the importance of the load balancing criterion increases. This law was chosen in order to force the handover towards those candidates which were proportionately less loaded;
- for each RAT_i with i=1,2,3, the value γ_i is set so that the function $f_2(PW_{opt,i})$=3/4; in other words, when the power level is equal to the value of $PW_{opt,i}$ which characterises the RAT_i itself, we set the weight associated with the received power level equal to 3/4. For values of received power lower than $PW_{opt,i}$, this weight rapidly decreases. Consequently, the power-based criterion becomes important when the received power level from an AP is lower than PW_{opt}. The final result gave values of γ as $\gamma_1=6.1 \cdot 10^9$, $\gamma_2=2.7 \cdot 10^9$, and $\gamma_3=3.3 \cdot 10^9$, for RAT_1, RAT_2 and RAT_3, respectively. In this way, we take into account different characteristics of RATs, i.e., the signal strength that, as all channel specific factors, cannot be straightforwardly compared.

The curves relevant to f_1 and f_2 are reported in Fig. 5.

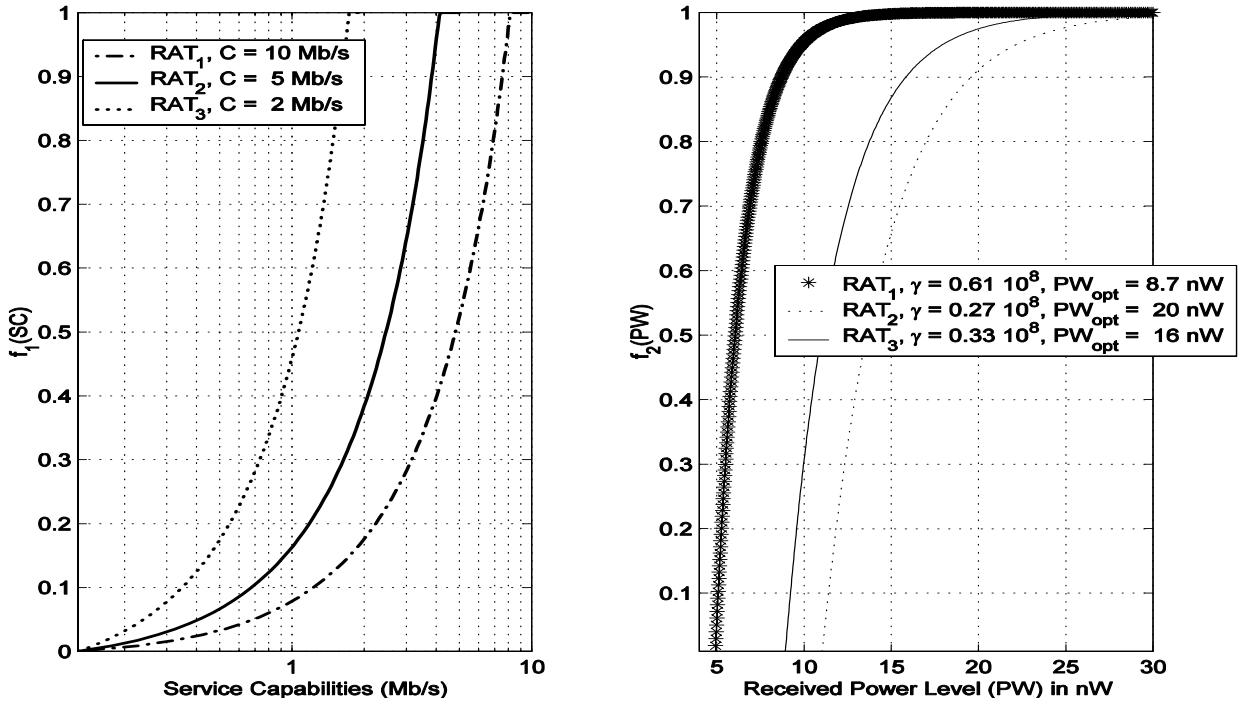


Fig. 5: Metric functions f_1 and f_2.

As regards the update of the service capabilities, ARs multicast the update of the *SC* related to an AP when its current value of available bandwidth differs by at least 5% from the one previously communicated.

The mobility model used for the simulations is the Gauss-Markov model [13], with directional parameters $\alpha = 0.5$, and average speed $s_{ave} = 1.5$ m/s, with the step fixed at 1 s (i.e., the MN position is updated every one and a half meters on average). We adopt this model since it avoids sharp direction changes, by allowing previous speed and direction to influence future mobility.

5.2 Simulation results

Our goal is to verify the effectiveness of the CARD/TAR procedure, in particular as regards the capability of not only self-constructing the coverage map of the whole area, but also of driving both inter and intra-technology handovers.

Initially, the CARD table stored in a generic AR contains only the information relevant to the APs connected to the AR itself, then this table self-constructs as time goes by according to the information provided by the MNs. As regards the data collected to generate statistics, we consider only those relevant to the MNs in the COMM state. For all the figures, we evaluated the relevant 95% confidence intervals; however, as they are very small, they are not shown in the figures to improve their clearness.

Let us first analyse the simulations in the case of dense coverage, with T_D=30 s, and N=3. Fig. 6 shows the cumulative probability of successfully handing over within first, second and third attempt suggested by the current AR after the TAR selection. As expected, we can see that, after a brief transient period needed for self-constructing the CARD tables, the probability of successfully handing over towards the best-scored CARs rapidly converges to values close to 1. Clearly, the successful handover probability increases with the number of attempts.

As regards the efficiency of the procedure to drive handovers, the failure probability of the CARD/TAR procedure versus simulation time is depicted in Fig. 7. Out of a total number of 164182 handovers (i.e., a handover event every 23 seconds per call), 128574 of them (78.31%) are driven by the procedure. The failure probability in the steady state is about 5%. It is worth noting that this quantity is relevant to all TAR events, and not only to those TARs implying a subsequent handover (i.e., the MN remains attached to the current wireless access). In addition, a high number of plain MIP handovers occur during the initial transient phase. For these two reasons, the value of failure probability of the overall CARD/TAR procedure remains satisfying. In fact, a non-zero value of the failure probability is expected due to the intrinsic characteristic of the procedure, which, for inter-technology handovers, is blind to the movement direction of MNs which are driven in handing over. In this regard, Fig. 7 also shows that the probability of failing an inter-technology handover is approximately 26%.

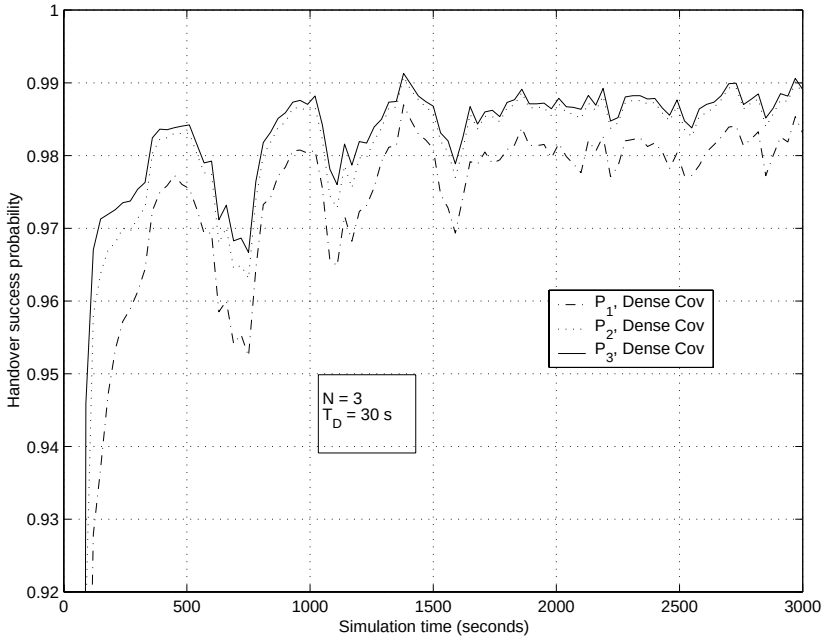


Fig. 6: Probabilities of successful handover: dense coverage case.

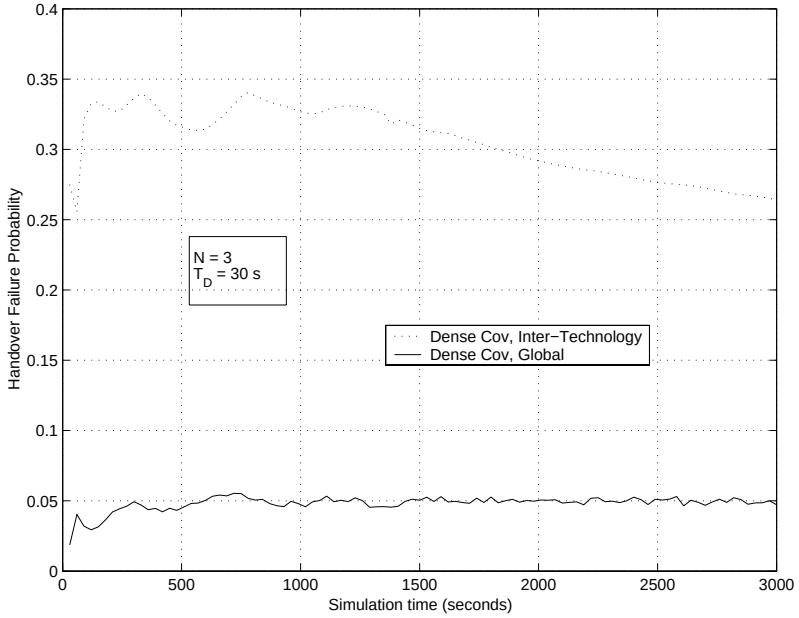


Fig. 7: Probability of failure of the CARD/TAR procedure and of inter-technology handover attempts: dense coverage case.

We noted that the performance of the CARD/TAR procedure is very satisfying for dense wireless coverage. This result is to be expected for two main reasons:

1. inter-technology handovers (the most critical ones) are not so frequent since
    a. each RAT covers the whole area;
    b. the offered traffic load is not heavy. This also means that inter-technology handovers are not so often solicited by a load balancing action.

    The distribution of successfully driven handovers, classified as intra and inter-technology handovers, and their distribution over the three different RATs is summarised in Table 3 for the specific dense coverage taken into consideration. The driven intra-technology handovers are equal to 93.24% of the total driven handovers and equal to 73.02% of the total accomplished handovers. As expected, most handovers occur within RAT_1, which has the largest capacity. On the other hand, driven inter-technology handovers are equal to 6.76% of the total driven handovers and equal to 5.29% of the total accomplished handovers;
2. the overlapping between the coverage areas of different RATs is high enough to keep the failure probability of inter-technology handover attempts sufficiently low as not to degrade the performance of the procedure.

Driven intra-technology handovers	$RAT_1 \rightarrow RAT_1$	65.14%	93.24%
	$RAT_2 \rightarrow RAT_2$	22.27%	
	$RAT_3 \rightarrow RAT_3$	5.83%	
Driven inter-technology handovers	$RAT_1 \rightarrow RAT_2$	1.52%	6.76%
	$RAT_1 \rightarrow RAT_3$	0.37%	
	$RAT_2 \rightarrow RAT_1$	2.27%	
	$RAT_2 \rightarrow RAT_3$	1.09%	
	$RAT_3 \rightarrow RAT_1$	0.44%	
	$RAT_3 \rightarrow RAT_2$	1.07%	

Table 3: Handover distribution: dense coverage case.

Consequently, if we consider a coverage with an amount of resources lower than those in the dense coverage, a deterioration of the performance of the procedure is to be expected. Let us analyse what happens in the case of medium and sparse coverage. Fig. 8, Fig. 9, and Fig. 10 show the cumulative probability to successfully hand over to within the first, second and third attempt, respectively in all the coverage cases. Again T_D=30 s, and N=3. It is clear that performance improves from sparse to medium coverage and from medium to dense coverage. This trend is highlighted in Fig. 11, where the probability of failure increases when compared with the dense coverage case. This is due to a lower amount of wireless network resources (i.e., a lower number of APs and therefore smaller radio coverage), hence to a reduced overlapping of coverage areas of different RATs. As a result, the percentage of driven handovers remains nearly constant for medium coverage (87300 out of a total of 110427, i.e., 79.05%) and rapidly decreases for sparse coverage (67995 out of a total of 98651, i.e., 68.92%). The number of handovers for MNs in the COMM state is higher in the case of dense coverage since the amount of offered traffic is higher than in the other two cases, as mentioned above. As regards the percentage of inter-technology driven handovers, it generally decreases with the coverage area provided by the single RAT in the area (it is equal to 13.37% in the case of sparse coverage). In addition, the probability of failing the CARD/TAR process increases from the 5% for dense coverage up to approximately 5.5% for medium and 8% for sparse coverage.

As regards a rough estimation of the initial transient, the curve analysis indicates values of approximately 10 minutes for all coverage cases.

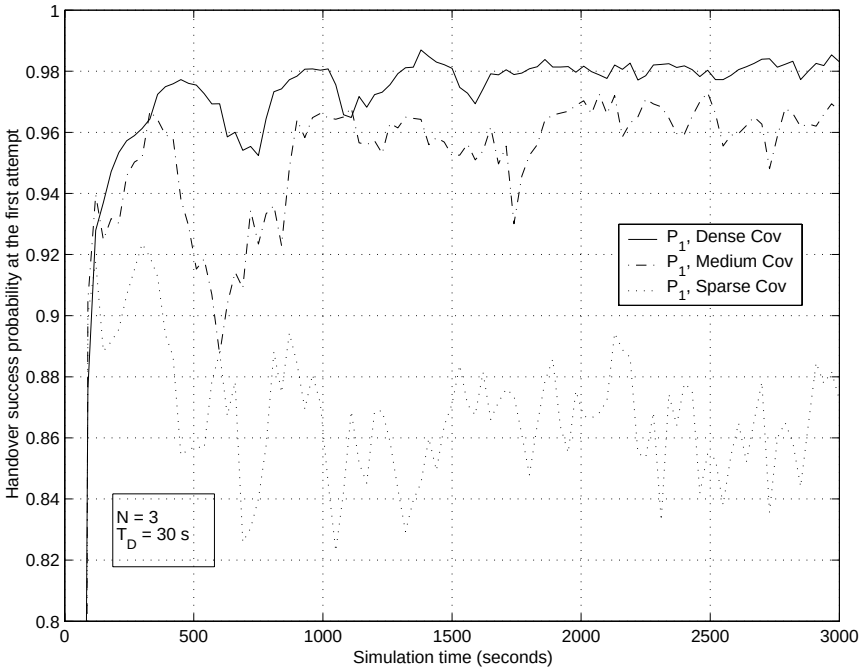


Fig. 8: Probability of successful handover at the first attempt: dense, medium, and sparse coverage.

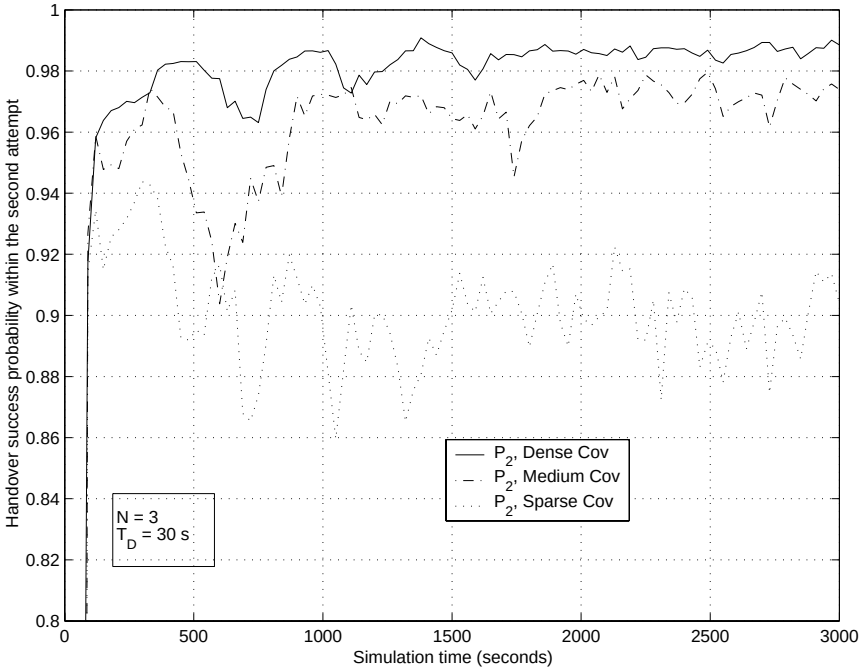


Fig. 9: Probability of successful handover within the second attempt: dense, medium, and sparse coverage.

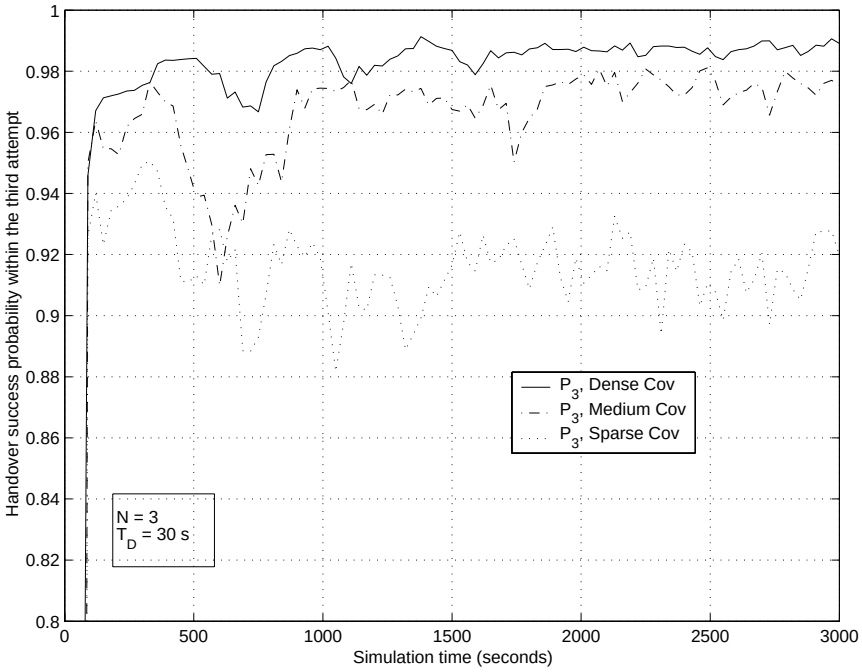


Fig. 10: Probability of successful handover within the third attempt: dense, medium, and sparse coverage.

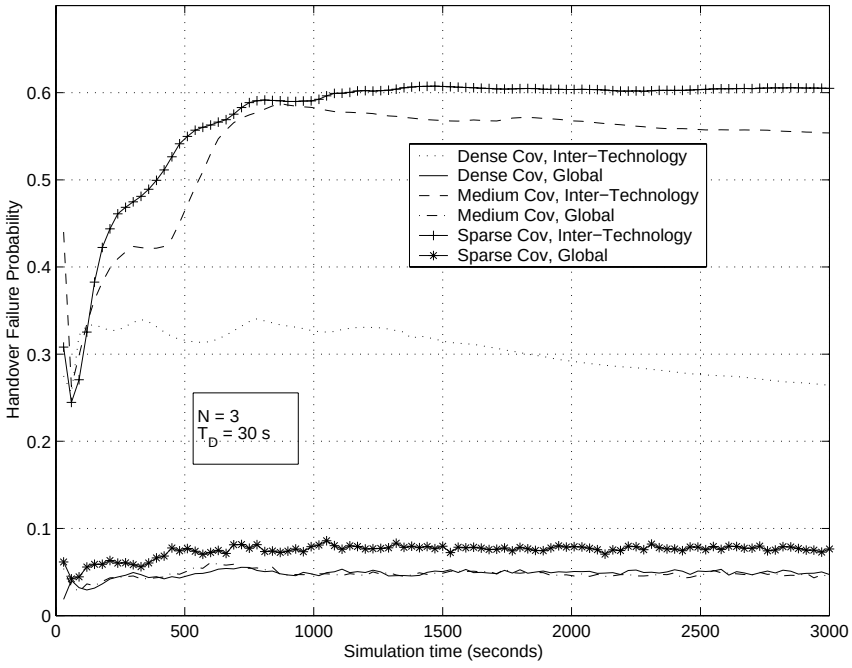


Fig. 11: Probability of failure of the CARD/TAR procedure and probability of failure of inter-technology handovers: sparse, medium, and dense coverage.

Let us now analyse the performance of the procedure in the case of sparse coverage when N ranges in set {2,3,4}, and T_D in set {15,30,45 seconds}. We expect a performance improvement when N increases and T_D decreases. In fact, lower values of T_D imply more frequent TAR actions, thus a better load balancing and better performance of the CARD/TAR procedure, which is able to quickly track the network changes. Moreover, by increasing the value of N means that the TAR list, provided to the MNs by the current AR, includes a higher number of CARs to try to hand over to, and this implies an improvement in the performance of the procedure. Fig. 12 shows the cumulative probability of a successful handover within the first N handover attempts for cases N=2, N=3, and N=4, with T_D=30 s. The improvement is clear for each handover attempt. We noted in particular the improvement of the successful handover probability at the last attempt. This value varies from 0.899 for N=2, to 0.915 for N=3, and, finally, to 0.92 for N=4. This improvement induces only a negligible decrease in the failure probability of the procedure and in the probability of failure of inter-technology handovers.

In addition, as expected from the above considerations, the cumulative probability of successfully handing over within the first, second and third attempt (N is fixed to 3) improves when T_D decreases, as shown in Fig. 13. Moreover, the behaviour of failure probability of the whole procedure and of the inter-technology handover attempts is shown in Fig. 14 (a) and Fig. 14 (b), respectively. For both performance parameters, a slight improvement occurs when T_D decreases.

5.3 Additional considerations

As regards a comparison of performance of our CARD approach with the solutions proposed within the IETF community, it is necessary to distinguish between the "discovery phase" and the "steady phase". We can make the following qualitative considerations:

1. as regards the discovery phase, L2 beacon-based solutions outperform handover-based solutions in terms of duration of the discovery phase itself. This is due to the higher number of events that trigger the discovery of new wireless resources. It is also worth noting that an L2 beacon-based solution does not present the problem of bootstrap handover, which might be critical for both delay-sensitive applications and variable-topology access networks. The price to pay with respect to handover-based approaches is a higher amount of signalling. However, it is worth noting that the discovery phase is generally limited in time, since it is the initial transient in the acquisition of information concerning the surrounding wireless

coverage at ARs. In addition, our specific solution outperforms other L2 beacons based solutions (both server-based solution and distributed solution based on SLP) since it avoids the latency due to explicit queries to a remote entity;

2. as regards the steady phase, in our approach the capabilities update is push-mode performed (i.e., unsolicited) by the relevant AR under noticeable variations (e.g., its available bandwidth is changed by a predefined step value), whereas in the other solutions the update mechanism is pull-mode based (i.e., on demand). Our selected mechanism enables the TAR process to be speeded up, since all the information needed to perform the TAR algorithm is present and updated in the current AR. Moreover, the capability update process is not a merely trivial information flooding through the overall network domain, but it is performed by the localised exchange of multicast messages among participants to low level multicast groups. Finally, it is also worth noting that to carry out the TAR selection in the current AR enables not only that the transmission of confidential network data over the air interface can be avoided, but also wireless bandwidth can be saved.

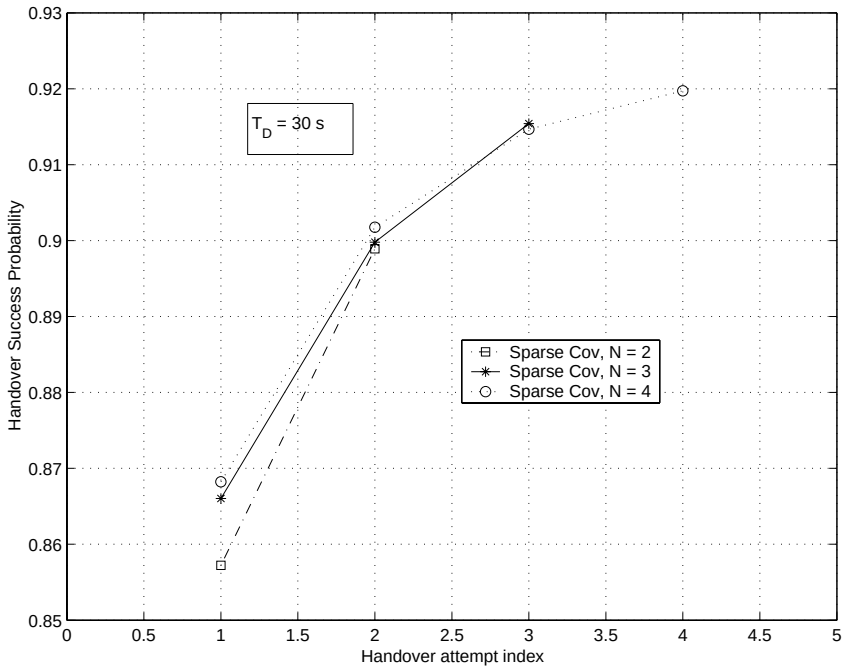


Fig. 12: Probability of successfully handing over versus handover attempts for *N*=2,3,4: sparse coverage.

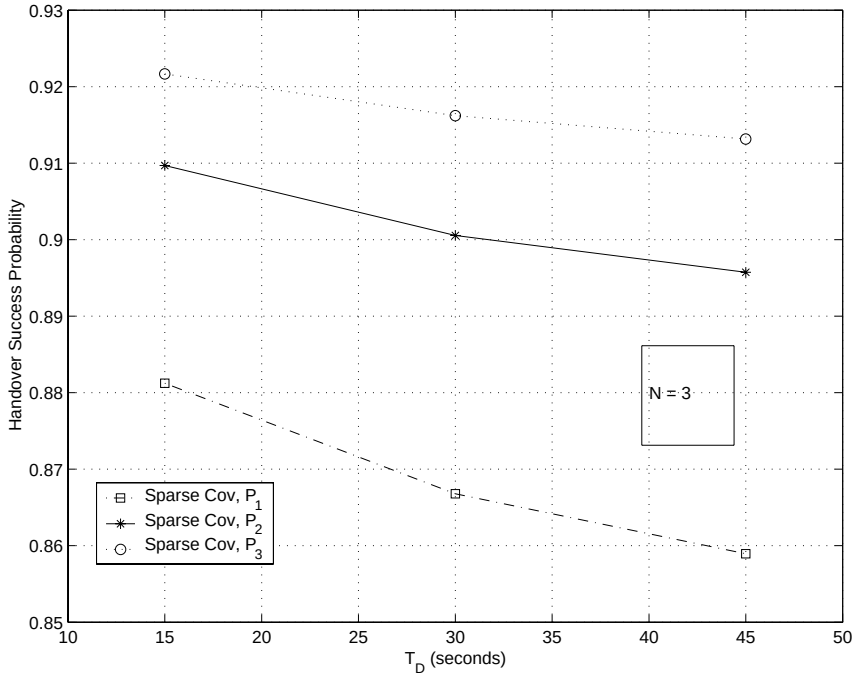


Fig. 13: Probability of successfully handing over within the first, second, and third attempt versus T_D: sparse coverage.

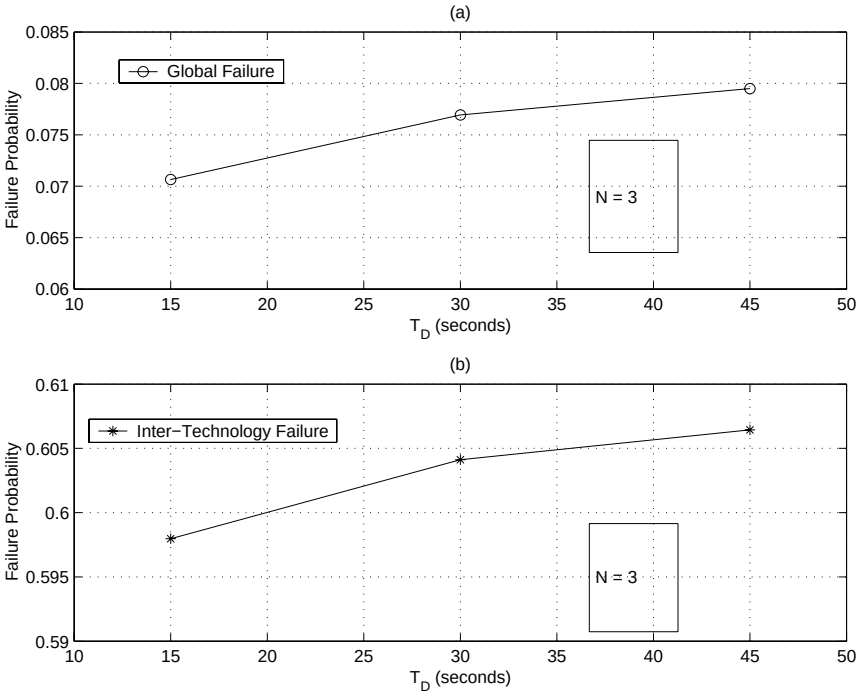


Fig. 14: Sparse coverage: (a) Probability of failure of the CARD/TAR procedure versus T_D; (b) probability of failure of the inter-technology handover attempts versus T_D.

6 Conclusions and Future Work

We have considered a framework in which an administrative domain provides wireless access by means of heterogeneous radio technologies. In this scenario, we have proposed an intra-domain distributed CARD/TAR procedure based on L2 beacon listening, to (i) dynamically self-construct a local map at each AR of the surrounding wireless coverage, and (ii) select the most suitable access to the network.

The main novelties of the paper are: (i) the CARD process is speeded up by the use of push-mode multicast, contrary to pull-mode unicast of IETF solutions; (ii) the CARD procedure is able to provide the TAR algorithm with the necessary information to enable it to drive network-assisted handovers at layers 2 and 3, both intra and inter-technology. This algorithm, used to select the best AR-AP for handing over to, is based on a set of criteria, including load balancing, received power level, and QoS level (namely the bandwidth). The procedure is valid for both single-mode and multi-mode terminals.

In particular, we have considered the very challenging situation of multi-mode terminals with a single network interface turned on; this constraint derives form the need to save power in small, portable devices.

We have analysed the performance of the procedure by simulations. Performance is expressed in terms of handover success probability. We have also evaluated the sensitivity of the procedure to variations of the model parameters. Simulation results show the effectiveness of the procedure, especially when the size of the overlapping AP coverage areas, relevant to different RATs, is considerable. In this case, the percentage of successfully-driven handovers by our CARD/TAR procedure is approximately 78%. Inter-technology handovers, which are clearly the critical ones, are not that frequent (approximately 7% of the total driven handovers). The probability of failure of inter-technology handover attempts is sufficiently low (around 25%) and it does not significantly affect the overall failure probability, which is about 5%.

It is worth stressing that our work aims to provide all the necessary, updated information concerning heterogeneous radio coverage. Our TAR metric pushes inter-technology handovers only to balance the traffic load among ARs. However, the introduction of other TAR factors, such as users' preferences (e.g., willingness to pay), operator strategies (e.g., tariff and specific proprietary management traffic policies, such as different priorities associated with different access segments), could drive TAR decisions more frequently towards inter-technology

handovers. This may occur even though they are generally highly power consuming and potentially critical when real-time services are supported. We stress that this criticality is mainly due to the fact that the selection procedure is blind, but we have achieved the noteworthy result that it is not driven by beacon listening. The performance of inter-technology could highly improve if the physical position of MNs is available as an additional input to the TAR selection process. Furthermore, it is clear that, if multi-mode terminals are allowed to simultaneously scan for beacons on all network interfaces, the procedure is no longer "blind", and failures are avoided. In fact, handover candidates would be selected only within the set of APs communicated to the current AR by MNs in the TAR beacon list. Consequently, the benefits due to its intrinsic characteristics (self-construction of coverage mappings, timely update of service capabilities, load balancing capabilities, power control) would remain unaltered. The price to pay is increased complexity and power consumption in terminals.

Finally, it is also worth noting that the proposed CARD/TAR procedure is able to simultaneously manage both single-mode and multi-mode terminals.

Future research activity will focus on a comparison with IETF CARD solutions concerning: (i) the signalling burden associated with message exchange in the wired/wireless network; (ii) the time that ARs take to estimate the coverage map; (iii) the estimate of handover delay and packet losses.

References

[1] J. Manner, M. Kojo, "Mobility related terminology", *Internet draft*, draft-ietf-seamoby-mobility-terminology-06.txt, work in progress, February 2004.

[2] M. Liebsch, A. Singh, H. Chaskar, D. Funato, E. Shim, "Candidate access router discovery", *Internet draft*, draft-ietf-seamoby-card-protocol-06.txt, work in progress, December 2003.

[3] J. Loughney , M. Nakhjiri, C. Perkins, R. Koodli, "Context transfer protocol", *Internet draft*, draft-ietf-seamoby-ctp-08.txt, work in progress, January 2004.

[4] J. Kempf: "Problem description: reasons for performing context transfers between nodes in an IP access network ", *IETF RFC 3374*, September 2002.

[5] C. Perkins, "IP mobility support for IPv4", *IETF RFC 3344*, August 2002.

[6] D. B. Johnson, C. E. Perkins, "Mobility support in IPv6", in *Proceedings of MobiCom'96*, Rye, NY, USA, November 1996.

[7] A.T. Campbell, J. Gomez, K. Sanghyo, C. Wan, Z. R. Turanyi, A. G. Valko, "Comparison of IP micromobility protocols", *IEEE Wireless Communications*, February 2002.

[8] D. Funato, X. He, C. Williams, A. Takeshita, M.D. Ali, J. Nakfour, "Geographically adjacent access router discovery protocol", *Internet draft*, draft-funato-seamoby-gaard-01.txt, June 2002

[9] D. Trossen, G. Krishnamurthi, H. Chaskar, R.C. Chalmers, E. Shim, "A dynamic protocol for candidate access-router discovery", *Internet draft*, draft-trossen-seamoby-dycard-01.txt, March 2003.

[10] R. Koodli, "Fast hand-overs for mobile IPv6", *Internet draft*, work in progress, draft-ietf-mipshop-fast-mipv6-01.txt, March 2003.

[11] K. El Malki et al., "Low latency handoffs in mobile IPv4", *Internet draft*, draft-ietf-mobileip-lowlatency-handoffs-v4-08.txt, work in progress, January 2004.

[12] IETF Seamoby WG Homepage, http://www.ietf.org/html.charters/seamoby-charter.html.

[13] T. Camp, J. Boleng, V. Davies, "A survey of mobility models for ad hoc network research", *Wireless Communication & Mobile Computing (WCMC)*: Special issue on Mobile Ad Hoc Networking: Research, Trends and Applications, vol. 2, N. 5, pp. 483-502, 2002.

[14] E.Shim, J.P. Redlich, R.D. Gitlin, "Secure candidate access router discovery," *Proceedings of IEEE Wireless Communications and Networking Conference (IEEE WCNC)*, New Orleans, USA, March 2003.

[15] IETF MSEC WG Homepage, http://www.ietf.org/html.charters/msec-charter.html.

[16] S. Glass, T. Hiller, S. Jacobs, C. Perkins, "Mobile IP authentication, authorization, and accounting requirements," *IETF RFC 2977*, October 2000.

[17] C. E. Perkins, P. R. Calhoun, "AAA registration keys for mobile IPv4," *Internet draft*, draft-ietf-mip4-aaa-key-01.txt, work in progress, October 2003.

[18] G.P. Pollini, "Trends in hand-over design," *IEEE Communications Magazine*, March 1996.

[19] P.M.L. Chan, R.E. Sheriff, Y.F. Hu, P. Conforto, C. Tocci, "Mobility management incorporating fuzzy logic for a heterogeneous IP environment," *IEEE Communications Magazine*, December 2001.

[20] K. Pahlavan, et al., "Handoff in hybrid mobile data networks", *IEEE Personal Communications*, 7(2), April 2000.

[21] N.D. Tripathi, "Generic Adaptive Handoff Algorithms Using Fuzzy Logic and Neural Networks," Ph.D. Thesis, Virginia Polytechnic Institute and State University, 1997.

[22] W. Zhang, J. Jähnert, K. Dolzer, "Design and evaluation of a hand-over decision strategy for 4th generation mobile networks," *Proceedings of IEEE Vehicular Technology Conference* (VTC 2003 Spring), Jeju, Korea, 2003.