WIRELESS COMMUNICATIONS AND MOBILE COMPUTING
Wirel. Commun. Mob. Comput. 2005; **5**:603–617
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/wcm.331

A cost-based approach to vertical handover policies between WiFi and GPRS

Andrea Calvagna*,† and Giuseppe Di Modica

Dipartimento di Ingegneria Informatica e delle Telecomunicazioni, Università di Catania, Viale A. Doria 6, 95125 Catania, Italy

Summary

To implement seamless mobility inside an integrated, multiple (e.g., GPRS/WiFi) access system, a vertical handover policy has to be devised. This is usually done at the mobile terminal, allowing it to be customized from an end-user's perspective, in order to fit individual needs/preferences. We propose a new approach in taking vertical handover decisions, which are not anymore exclusively based on the knowledge of the available access networks ' characteristics but also on higher level parameters which fall in the transport and application layers. To this extent, in this paper a model has been realized and simulations have been run in order to evaluate the impact of the vertical handover and its frequency on a set of typical user's network applications/services. We also take into account the user preferences in terms of cost and quality of service. We believe this approach reflects the optimal settings from the user's point of view with regard to his running services and applications. Our aim is to understand how to define a metric to be used in order to devise a solution which should try to balance the overall cost of vertical handovers with the actual benefits they bring to actual user's networking needs. This way, each mobile user could autonomously apply the handover decision policy, which is more convenient to his specific needs. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: wireless; mobile; handover; heterogeneous networks

1. Introduction

The Fourth Generation (4G) networks concept supports wide-band data and telecom services for mobile users roaming across multiple, wireless and wired, integrated access networks. Basically, the purpose is that of combining all the existing heterogeneous wireless networks into a single, interoperable system, being IP protocol the 'glue' between the set of underlying radio access and physical layers. Within this overall scenario, it is foreseeable that many access technologies, even with very diverse profiles (in terms of bandwidth, latency, security, etc.) will often be available in overlapping areas. Users will also be equipped with terminals capable of multiple access interfaces, or provided with a dynamically reconfigurable access interface [7], allowing them to seamlessly take advantage of more than just one physical access connection, even at the same time. Particularly, in the context of overlapping, heterogeneous, access networks, specific strategies must be devised to control the triggering of 'vertical' handovers [10],

*Correspondence to: Andrea Calvagna, Dipartimento di Ingegneria Informatica e delle Telecomunicazioni, Università di Catania, Viale A. Doria 6, 95125 Catania, Italy.
†E-mail: Andrea.Calvagna@unict.it

between the available access networks, which affect the overall performance of application sessions running on the mobile user's device. Also, fast handover procedures are needed to meet the application's requirements. In fact, the handover latency may heavily affect the continuity of application sessions (let us think about the strong delay requirements imposed by real-time applications).

In this context, current literature is focused on optimizing vertical handovers from the network-level point of view only [8,4]. In contrast, the specific design of vertical handover policy may deeply impact on the performance of the user applications at the user's terminal. As an example, a triggering policy which simply switches to any 'better' (e.g., lower latency) access network as soon as it is available, could disappoint the user with possibly frequent connection discontinuities and, depending on the running applications (e.g., non real-time), would not necessarily improve the performance. Other non-network-level parameters and variables, such as the available connection's cost, the user's mobility pattern, the kind of the applications running on the mobile terminal, should be accounted for when considering what is 'better' in overall from the end user point of view. In this paper, we modeled a possible 4G-network scenario and performed some simulation work to assess the impact of vertical handovers on the main kinds of IP-based application flows.

The paper is structured as follows. In Section 2 we give an overview of the handover protocols. In Section 3 the concept of 'cost' of vertical handovers, from the point of view of the transport and the application layers, is introduced. Section 4 presents the proposal of a user-centric middleware architecture. In Section 5 we describe a model of interworking between WLAN and GPRS, in which simulations, whose results are presented in Section 6, have been carried out. In Section 7 the implementation of a real wireless mobility framework is described. Finally, Section 8 concludes our work.

2. Handover Management Approaches

It is common agreement in the scientific community that handover process complexity will still increase as we move from 3G to 4G systems, which are even more integrated with data networks than their predecessors. The need for augmented knowledge about context in the MH is considered the key to enable scenarios of seamless connectivity to highly integrated heterogeneous wireless networks. In References [15,16], examples of policy-based solutions involving 2.5G, 3G, and 4G systems integrated with wireless data networks are presented, which build around a policy-based scheme. In accordance with our vision, they show that a flexible MCHO policy-based approach is a light weight solution appropriate for future mobile devices.

A classification of the different approaches to handover management can been made, based on the entity that is in charge of controlling the handover procedures [12]. It is then possible to mention three main types of handover managements,

- Network controlled handover (NCHO)
- Mobile assisted handover (MAHO)
- Mobile controlled (policy-based) handover (MCHO)

The first type is known as the centralized approach, as there is one single logical entity (the network) controlling handovers for all the users. As the handover decision making process moves to a decentralized approach (i.e., moving from NCHO to MCHO), handover latency decreases, but the quantity of context information available to perform a handover decision decreases also.

In an NCHO protocol, the network triggers handovers typically comparing the received signal strength (RSS) from all the MH measured at a number of base stations (BSs). In this scenario, a network entity is in charge of managing every optimal handover decision for all the MHs. These are usually optimized to load-balance the overall network and maximize call-admission probability of each network cell. This network entity itself is a bottleneck, thus, this type of handover is not suitable for a high density of users or a rapidly changing environment due to the associated unpredictable delays. NCHO is used in first-generation analog systems such as total access communications system (TACS), and nordic mobile telephone (NMT).

In MAHO distributed handover decision processes, the MH makes measurements, and the network makes decision. The network gathers information from all of the MHs, and accordingly schedules the handovers. Second generation mobile systems, like GSM, employ this handover approach.

In MCHO, despite the MH is completely in control of the handover process, it still implements a purely network-level-based policy, build to minimize handover latencies for high mobility inside micro-cellular systems. The MH does not have any information about the signal quality of other users. The MH

measures the signal strengths from surrounding BSs and interference levels on all channels, since handover must not cause interference, and simply triggers handover to the stronger BS available. This type of uncontrolled behavior may lead to the well-known ping-pong problem (i.e., MH wasting energy looping back and forth between two BSs) MCHO is the highest degree of handover decentralization.

To summarize, handover decentralization allows for faster handover decisions, and does not burden the network, but handovers policies are very simple and based on a very narrow 'view' about network status. Centralized handover decision processes allow complex resources optimization strategies to be applied on the whole network, but while complexity increases, performance decreases. On the contrary, handover decentralization (i.e., MCHO) has the disadvantage of being based only on small local-context information. This makes harder, if not impossible, for example to implement in the MH smart handover decision strategies that take into account also for the status of our cell (or surrounding cells) neighbors. In order to do that, active support by the network, providing 'extended' status information to the MH, would be required.

According to the MCHO strategy, every MH autonomously decides, for its own convenience, when and to which BS to trigger the handover. But of course, the sum of every single MH's convenience will hardly result in an efficient global network resources management. There is not a central entity that helps distribute the traffic across the network when congestion is about to occur. Thus, in general the main disadvantage of MCHO strategies is that the overall resource management is not under the direct control of the network operator and, what is worse, nor by anyone.

For this reason, NCHO strategies are the most convenient, at least from the network operator's perspective. These considerations apply to scenarios in which the MHs roam within the same administrative domain, or between domains owned by the same network operator. The scenarios and the issues addressed in this work are somewhat different, since they contemplate not only different access technologies, but also different network operators owning the domains traversed by the MHs.

Let us think about one single administrative domain, in which the operator offers to their customers two different wireless accesses (e.g., GPRS and WLAN). The considerations regarding the NCHO still applies, even though the resource management task is more challenging due to the presence of two heterogeneous wireless access technologies. The management of the vertical handovers between the two networks can be carried on by the network operator according to resource management criteria. But, in the more general case of two heterogeneous wireless access networks (say Net_1 and Net_2) owned by two different network operators (say Op_1 and Op_2), things get more complicated. If the case that Op_1 and Op_2 have signed roaming agreements, and yet the NCHO is employed, the vertical handover strategy agreed by the two operators might disappoint their customers. In the case that Op_1 and Op_2 are competitors, and therefore do not cooperate for the vertical handover process, from the network point of view a vertical handover is nothing but the termination of a *call* in one of the two networks and the origination of a new *call* in the other network, rather than a transfer of the *call* between the two networks. In both the depicted cases, we argue that either NCHO strategies could be disappointing for the customers.

To conclude, we observe that a policy-based handover strategy is the best fit to maximize user's satisfaction, which in a heterogeneous, multi-access, and multi-vendor mobile network environment should be considered as the first priority. Thus, it should involve the users' preferences and the current application-level networking needs as one force driving the handover decision process. On the other hand, if this would be actively supported by the access networks, by letting the MH easily gain relevant network status data, this cooperation would also allow for applying wise (e.g., load-balancing) overall resources management across the network.

In a mid-way scenario that we assume in the remaining work, access network operators still prefer the NCHO approach inside their controlled domains, but it is possible to leverage from a MCHO strategy across overlapping domains, applying policy-based decision algorithms at least to 'vertical' handovers, that is, handovers between two or more different and overlapping access networks.

3. The *Cost* of Vertical Handovers (Vertical Handover Policies)

The overall, network-level cost of performing a vertical handover between two types of access systems could be expressed in terms of the latency and bandwidth gaps between the two. Also, the location update latency imposed by the installed IP mobility

manager (e.g., MIP [9]) has to be considered as an additional delay. What we now have asked instead is: how can we measure how this vertical handover will influence the behavior of running applications? In other words, what is the application-level cost of the very same *network layer* vertical handover, which may be very diverse depending on how it will impact on the application and the protocol it uses to manage end-to-end data flows (UDP, TCP, RTP, HTTP, FTP, etc.).

Thus, we shift the focus from the network layer to the transport and above layers. Of course, techniques can be devised to make vertical handovers as seamless as possible to this point of view. It is a matter of fact that a vertical handover to/from a different wireless network causes changes in the connection throughput and end-to-end RTT (*round trip time*), and frequent fluctuations of the connection parameters may affect negatively the performances of the application sessions running on the mobile user's device.

As a consequence, it may be argued that the final effectiveness of a vertical handover policy should have to be valued over the whole length of a mobile user's work session, and from a higher-level point of view. Obviously, the ideally optimal schedule of vertical handovers cannot be exactly computed unless the running applications, the access networks topology and the user's movement pattern inside it are all well known beforehand. By the way, it is possible to use abstract modeling to derive a more generally applicable, near optimal, solution.

4. An Architecture to Support the User Preferences

Providing 4G mobile users with the capability to seamlessly and conveniently keep their active work sessions across a network of multiple, possibly overlapping, heterogeneous radio access technologies, requires as a fundamental step to define what 'best connected' means at the user-level, that is, from their personal point of view. We argue that this user-level definition of what is 'best' in such context may not be universally and statically determined. In contrast, it is a dynamic concept strictly related to many factors including:

- the attributes characterizing the current set of applications the mobile user is running in his mobile-terminal (let's call it the 'running applications profile');
- the attributes characterizing the access conditions offered by the set of locally available radio networks (the 'available access profiles');
- the overall network topology;
- the user's movement pattern.

Despite all of the above mentioned (except for the last one) are somewhat network-resource usage related attributes, the user-level 'settings' may reflect also strong implications outside the purely functional domain, that is, the cost. In fact, private companies are willing to profit from their offer of 4G access networks and services, which will thus have significant fees associated to their usage. As a consequence, money should be one of the key parameter driving user-level optimization of commercial networks' resource usage. To summarize, we argue that a 4G mobile user should be able to use his terminal to dynamically evaluate locally offered network services (access/data) and optimize functional (performance of running applications) and nonfunctional (cost of services) requirements according to his specific needs. In this section, we propose a possible framework architecture for the deployment of a ABC system [5,6].

4.1. Overview

In Figure 1 the complete stack of layers of our envisioned architecture is shown. It is intended to run in a mobile host having, at the lower-level, multiple-network physical interfaces. No changes at the IP

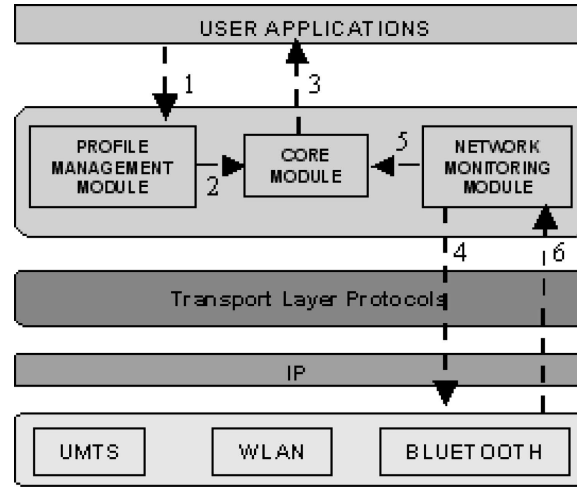


Fig. 1. (1) Interaction between the user and his user's profile; (2) The user preferences as an input to the network selection process; (3) The feedback to the user applications; (4) The selection of the wireless network; (5) The data from the networks as an input to the network selection process; (6) The monitoring of the available networks.

and transport-layer protocols are required, insuring perfect compatibility with current Internet standards. Our architecture just adds a extra software layer over the transport layer interface, in order to implement and provide the user-centric handover policy feature up to the context of user-space network applications. The internals of this software layer are structured in three main submodules whose:

- networking module, in charge of wireless access network detection and monitoring.
- Profile management module, in charge of letting the user get/set the details of his mobile networking preferences on the base of (1) the collected info about cost/benefits of available access services, (2) his current application's networking requirements and (3) his free will.
- Core module, which is in charge of correctly applying the user specified policy about wireless access network selection and consequent application session management.

In the following, a description of each functional block is given, with particular emphasis given to the profile management module.

4.2. Wireless Access Network Detection and Monitoring

The detection and the monitoring of the wireless access networks visited by the end user while roaming is one of the most challenging problem to be faced. The *network monitoring module* is in charge of interfacing to all the network cards that the user terminal is supplied with, and continuously gathering information about the availability and the reliability of the different wireless access channels. A mechanism is needed to collect the heterogeneous information from the different networks and uniform them (for instance, by means of normalizing procedures). Furthermore, other network relevant data can be collected at this stage. The network operator might want to provide some useful information concerning its brand, the actual bandwidth occupancy, the level of network QoS that can be supplied at that moment to that user (e.g., in terms of bandwidth availability and packet delay), the coverage area granted by that specific wireless access point.

4.3. User Profile Management

One of the main ideas that drove us throughout the design of the middleware is giving the end user as much control as possible on the selection of the wireless network that best meets his preferences. In our opinion, user satisfaction [14] is to be taken in great account when trying to enable the concept of 'always best connectivity'. The perception of 'best' connectivity changes from user-to-user, and strictly depends on the relative value that the user gives to one aspect of the connectivity with respect to another. The *user profile management* module gives the user the chance to statically specify his own preferences, and dynamically modify them whenever he needs. The preferences are locally stored (i.e., in the terminal device's memory, or in the USIM), and its management does not involve any remote interaction. For instance, the user might wish to (1) save on the connection cost by preferring, when available, the networks whose connection costs are lower, according to the fares of the contracts that he has subscribed to; unavoidably, this strategy would reduce the number of accessible networks, and definitely no guarantee is given about the sessions' continuity (application sessions might be abruptly broken down because no connection is available according to the user preferences); (2) guarantee his ongoing application sessions as much as possible, no matter the connection costs; (3) find some compromise between cost savings and sessions' continuity; (4) specify the set of applications that must be preserved in the case that the network performances degrade (e.g., because of an handover); (5) specify whether some applications must be alerted whenever the network conditions change. The *network selection process* is given access to these preferences that, together with the data collected from the network, are used to make the 'best' network choice. It is worth noting that a network selection (i.e., an handover to a different network) can be performed not only because of the sudden unavailability of the currently accessed network, but also because the latter no more meets the actual preferences expressed by the user.

4.4. Wireless Access Network Selection and Application Session Management

This is the core policy module. The network selection process module was in charge of continuously monitoring the networks. To this end, data revealing the actual status-condition of the available networks was periodically retrieved form the network-detection module. The *core module* will combine such data with the ones retrieved form the user profile. Depending on whether the QoS expressed by the user in his

preferences can be sustained by the current network conditions, a change of the current network can be triggered, by selecting the next from the available ones, or some applications might be alerted to let them accordingly adapt to the new network conditions. In the first case, the new network will be selected according to the user's need (as specified in the user profile), and in particular the one that guarantees a reliable connection support will be chosen (according to the data gathered from the each network interface). In the second case, a feed back of the changed network conditions is provided to some applications that the user wishes to preserve. On their turn, the applications will try to adjust their parameters, so that a minimum functionality is granted as specified by the user. Let us focus on the following scenario. Suppose that the user is accessing a network providing him with a good connection in terms of available bandwidth. He decides to run an adaptable video stream session, and sets the parameters of the service accordingly to the high bandwidth that he is given (e.g., high frame and bit rate, full color, etc . . . .). Suddenly the current network conditions vary in such a way that the current session can no more be sustained. According to the user preferences, two different solutions are possible: (1) the client-side video stream application is alerted about the changed network conditions, so that a downgrading in the application parameters is imposed (e.g., the frame and bit rate might be lowered, the audio suppressed, etc . . . .); (2) a new network is looked up, able to sustain the current application session, either in full mode (high frame and bit rate) or in half mode (low frame and bit rate), depending, once again, on the user convenience.

5. Model Description

We have developed a model implementing a subset of the requirements for the interworking between WLAN and GPRS. A MIP-like distributed mobility protocol has been designed and integrated in the model to support the roaming of mobile hosts (MHs) in the WLAN and in the GPRS domain. A cellular IP (CIP) [3] derived protocol is used to take care of micro-movements management. In particular, an hysteresis-based strategy has been conceived for horizontal handovers triggering.

The protocol does not deal with authentication or security issues. Rather, it focuses on the support of the continuity of MHs ongoing transport sessions while it changes wireless access due to its frequent movements. Smart strategies have been devised to handle both horizontal and vertical handovers, aiming at guaranteeing the MHs ongoing connections against sudden disruptions due to the MHs crossing the boundaries of adjacent radio covered regions. No matter whether the MH traverses the boundary of regions served by homogeneous or heterogeneous wireless access points, the protocol is in charge of scheduling the time for triggering the handover, based on the assessment of the MH future movements and on the evaluation of the wireless channels' conditions.

According to Reference [10], we refer to a WiFi-to-GPRS handover as an upward vertical handover; such an handover is in general triggered when the MH moves out of reach from a narrow-coverage (but broadband) network while already inside an overlaying wide-coverage (and narrowband) network. The common upward vertical handover strategy is very straightforward. Beacon packets from every available WiFi access points (APs) are constantly monitored, with particular attention to the ones collected by the AP that is currently serving the MH. Whenever the signal strength level of these beacons falls below a given threshold, meaning that the current radio signal is going to be lost, a new AP is chosen, among the monitored ones, in order to handover to it all MH ongoing sessions (in this case, an horizontal handover would be triggered). If none is available, or their radio signal strength is too weak, an upward vertical handover is triggered to the GPRS network. The MH data connections might be heavily affected by this kind of vertical handover, since (by definition) it is just the consequence of a lack of connectivity (radio 'silence') experienced by the MH, whose handling over to the wide-coverage network is not immediate but requires also another, so called, handover latency. Conversely, we refer to a downward vertical handover when the MH leaves a wide-coverage network to move to a narrower one (GPRS-to-WiFi handover).

Triggering of this kind of vertical handover is usually not a need, as was the previous case, but just an opportunity. As a consequence, it is less disruptive than the previous, given that the MH, during the handover phase, can keep its ongoing connections alive in the overlaying network until the handover procedure has been completed. Let us now assume that the MH is currently accessing the GPRS network, and that the MH is approaching a region covered by a WiFi AP.

In our model, the following threshold-based strategy is adopted to schedule the time for a downward

vertical handover. As soon as the MH senses the first beacon from the WiFi-network card that it is equipped with, a timer is started. The average signal strength of the received beacons (ARSS) is monitored until the timer expires. At expiration time the signal's ARSS is checked against a given value T_h, representing the strength level that the ARSS at least has to equal in order for the signal to be assessed reliable. The greater T_h, the harder for a signal to be evaluated reliable.

The purpose of this scheme is to filter out all the weak and intermittent signals coming from nearby WiFi APs, thus avoiding unnecessary downward vertical handovers that might result in further heavy sessions' disruptions. In fact, if at a given time a downward vertical handover were performed to a WiFi AP, whose fading radio signal soon revealed too weak to support the MH connection, a new upward vertical handover should immediately be triggered to divert the MH communication sessions back to the GPRS network. In such scenario, given the higher handover latency time needed to complete an upward vertical handover, the well know ping-pong effect is even more disruptive than in scenarios where only horizontal handovers occur.

Simulations have been carried out in a NS2 [13] model that we have developed. The MH moves within a large area according to the well known mobility scheme named 'random walking model'; the pattern of the MH movements have been generated being the parameters max-speed set to 20 m/s and the pause time set to 0 s. Several 802.11 APs have been placed in such a way to ensure islands covered by radio signal to the roaming MH. The transmission power of the APs has been set up to a value that guarantees overlapping areas of radio signals. Furthermore, regions not covered by any radio signal ('holes') have been intentionally left in between radio-covered islands. In such holes the MH will take profit of the overlaying GPRS network.

6. Simulation Results

6.1. Used Traffic Models

In Reference [11] a classification of the traffic running over the Internet is given. Elastic traffic includes TCB-based applications like Telnet, FTP, P2P file sharing, e-mail, and Web browsing. Even though reliability is a crucial QoS parameter for these application, throughput may be considered as performance metric as well. Inelastic traffic is generated by real-time services (voice and video) and, in general, by all the data services to which both timing and throughput are relevant parameters in order to meet the QoS requirements.

As far as the elastic traffic is concerned, a further classification can be introduced, based upon the level of user interactivity imposed by the semantics of each application. The transfer of a long-sized file, as well as the downloading of an e-mail with a big attachment, generates long-lived TCP transmissions with no interaction from the user. Conversely, applications like Web browsing and Telnet envisage a tight user interaction but generate a lot of short-lived TCP transmissions.

The inelastic traffic category enumerates applications like VoIP, MPEG, and H.263 video sources. Given the strict timing requirements, they reside on top of the UDP transport layer and usually generate transmissions at constant or variable bit rate. In the proposed analysis we make an effort to cover all the categories of Internet traffic, by employing several traffic source models.

6.2. Performance and Metrics

Let S_m be the minimum signal strength level in order for a packet to be correctly sensed by the MH when it roams in a WiFi domain. For each category of simulation, several simulations (each one lasting 900 s) will be run, respectively setting T_h to different values. In particular, one simulation will be run by setting the threshold to such a value that the MH will never abandon the GPRS connection (let us call this conservative setting 'connection-safe'). Another simulation is run by instructing the handover algorithm to search for just WiFi access points, even if that means that the MH will experience 'black-outs' in its connections (let us call this settings 'bandwidth greedy'). Other simulations are run by setting the T_h to values $1.1 \times S_m$, $1.4 \times S_m$, and $10 \times S_m$, respectively. A complete set of simulations is thus run, ranging from the most connection-conservative configuration to the most bandwidth-greedy. Results will show how varying the T_h affects in different ways the performance of the running applications.

We can monitor the performance of the TCP-based applications by observing how fast the number sequencing of the TCP segments increases and/or how big is the amount of transferred packets. At the end of a simulation, the greater the sequence number (or, the bigger the amount of downloaded bytes), the better the performance of the TCP. As far as the UDP-based

applications are concerned, the relevant metrics that we will monitor are the total number of lost packets and the packet delay as a function of the time. In order for UPD to show good performance, both the packet loss and the fluctuation's rate of the end-to-end packet delay are to be kept as low as possible.

6.3. TCP-Based Simulations

6.3.1. FTP session

The first group of simulations is focused on the study of the dynamics of the TCP protocol, during vertical handovers, when a file transfer is going on both upwards (from WiFi to GPRS) and downwards (GPRS to WiFi) between the MH and a corresponding host (CH) in the Internet. The FTP session is set up to start at time $t = 1.0$ s, and lasts until the end of the simulation: it is a typical, non-interactive, long-lived TCP transmission. In Figure 2, the sequence number's progression of the TCP segments received by the CH is plotted for different configurations of handover threshold. In the figure, each curve has a variable slope in time, depending on the current accessed network. In particular, when the MH is accessing the GPRS network, the curve's slope keeps low: actually in this phase the high end-to-end delay imposed by the channel limits the growth of the TCP number sequencing (i.e., the TCP sender's congestion window imposing the flow control increases very slowly). Conversely, when the MH switches to the WiFi network, the TCP re-estimates the end-to-end link delay and the transmission rate increases. Figure 2 for almost each simulation clearly shows a long permanence of the MH within the GPRS network from time $t = 300$ s to $t = 600$ s.

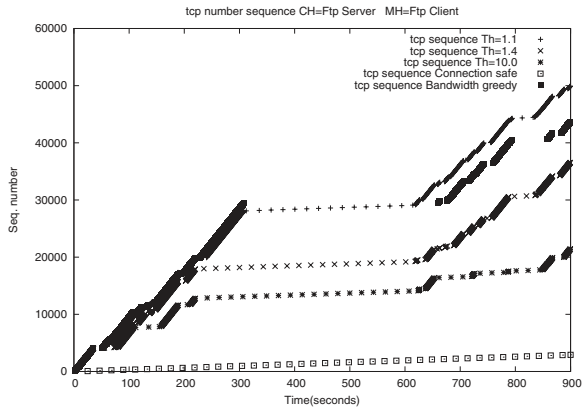


Fig. 2. The TCP number sequencing.

The simulation with the best performance is the one with $T_h = 1.1 \times S_m$. We notice that its performance does not differ much from that with 'bandwidth-greedy' (BG) settings. Even if the latter experiences long 'black-out' periods, the TCP protocol is, however, able to greatly evolve whenever the WiFi access is available (thanks to the higher bandwidth and the lower end-to-end delay). Conversely, the simulation with 'connection-safe' (CS) settings does not experience connection blackouts, but can only benefit from a connection with limited bandwidth and a high end-to-end delay (the GPRS one).

We can conclude that for long-lived TCP transmissions (like, for instance, a file transfer), 'bandwidth-greedy' handover strategies seems to give better performance. For this kind of transmissions it is more advisable to profit of a connection with high bandwidth and low end-to-end delay, even if intermittent, rather than a permanent connection but 'poor' in terms of bandwidth and end-to-end delay.

6.3.2. Http session

In this set of simulations an http session is started between the MH (the http client) and the CH (the http server). The client's page request generation process follows the Pareto distribution model. This model has been used in order to simulate the typical behavior of the user that browses the Web: a new page request is issued after the relevant information contained in the previously downloaded page has been read by the user. The average time spent by the user to read a page (i.e., the average *Toff* in the Pareto distribution) is set to 15 s, while the average web page size is set to 8 Kb. Obviously, the resulting *Ton* is variable, since it depends on the size of the page (whose average is fixed to 8 kb) and on the transfer rate of the page itself (which of course strongly depends on the network currently accessed by the user).

The traffic model that is being simulated is of *short-lived* type, and a *loose-user interaction* is observed. With respect to the FTP simulation, instead of a single long TCP session, several short TCP sessions will be started. In Figure 3, the total amount of the downloaded traffic for each simulation is shown.

Once again, the best performance is obtained by setting $T_h = 1.1 \times S_m$. This time, the simulation with BG settings shows the worst performance, while the one with CS settings is not greatly penalized as it was in the FTP simulation. The http-traffic model, in fact, generates short-lived TCP connections. The application does not greatly benefit from the higher

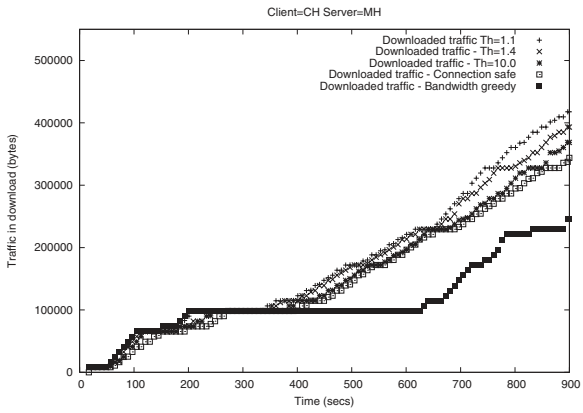


Fig. 3. Downloaded traffic during the http session.

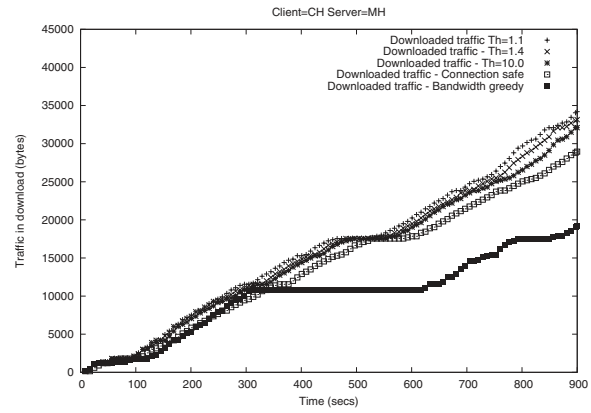


Fig. 4. Downloaded traffic during the telnet session.

bandwidth available in the WiFi domain, given that the most of time is spent by the user to read the downloaded page, whilst the connection is exploited for a very little fraction of time. That is, why the simulation in which the MH has a permanent connection to the GPRS network performs better than the simulation in which the MH connects only to the WiFi network. In fact, in the latter, from time $t = 270$ s to time $t = 620$ s the MH connection gets stuck because of the WiFi black-outs (i.e., the MH is out of the range of any WiFi access point), while the former can benefit from the GPRS connection.

The conclusion that we draw is that the connection parameters (the bandwidth and the end-to-end delay) have a lower influence on the overall performance when short-lived TCP transmissions are considered.

6.3.3. Telnet session

We describe the results that we obtained by simulating a telnet session. The considered traffic model resembles the one of http sessions. The user, in fact, interacts with his terminal, but this kind of interaction is tighter than the one observed for the Web-browsing session. Furthermore, the size of the packets exchanged between the telnet client program and the server one are much smaller than those exchanged in http sessions. We refer to this kind of traffic as a *very short lived* one with *tight-user interaction.* In Figure 4 the measured performance for each simulation is reported. All the simulations, except the one with BG settings, seems to show almost equal performance. Once again, the strategy of searching at any cost for the WiFi access does not pay.

6.3.4. UDP-based simulations

Simulations have been run to evaluate the impact of vertical handovers on the performances of applications built on top of the UDP protocol. In particular, RTP sessions have been set up between the MH and the CH. A CBR application running on the CH sends packet at a rate sustainable by the capacity of the link that has the narrower bandwidth (i.e., the GPRS link). This is to make sure that no packet gets lost due to congestion in any queue of the traversed domains' routers. In Figure 5, the total number of lost packets for each simulation are shown. The graph refers to the number of packets that have been lost during handovers, both horizontal and vertical. The more the threshold increases, the less packets are lost. Of course, the simulation with CS settings gives the best result (no packet lost), while the one with BG settings (not shown in the graph) experiences a great packet loss (more than 8000 packets get lost).

In Figure 7(a,b,c) the measured end-to-end packet delay is plotted for the simulations with $T_h = 1.1 \times S_m$, $T_h = 10.0 \times S_m$ and the one with bandwidth-greedy settings have been reported. During the 900 s of

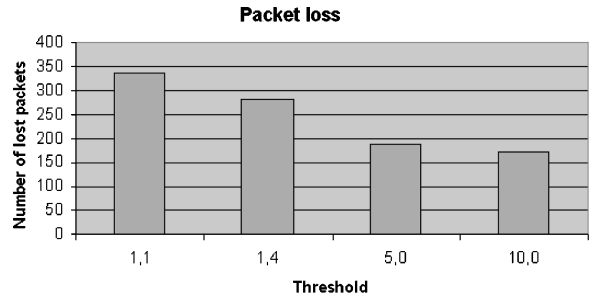


Fig. 5. Total number of lost packets.

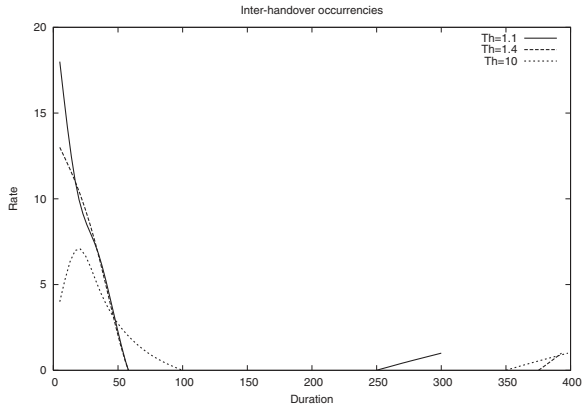


Fig. 6. Occurrences of inter-handover times.

simulations, the MH switches several times from the WiFi network to the GPRS. The ongoing CBR session undergoes frequent packet delay fluctuations (from 0.018 s in the WiFi link to 0.55 in the GPRS link) as far as the experienced throughput and the RTT are concerned. In particular, for this specific simulation, only the end-to-end RTT is affected by the frequent vertical handovers, given that the transmission rate of the CBR session does not exceed the GPRS link capacity. The graphs show that the rate of the packet delay fluctuations lowers as soon as the threshold increases. As far as the BG simulation is concerned, almost no fluctuation is observed, since the packet delay is almost constantly set to 0.018 s; but of course a lot of packets are lost during the black-out periods. The results of the simulation with CS settings (whose graph is not reported) showed a constant packet delay almost equal to half the RTT of the GPRS link.

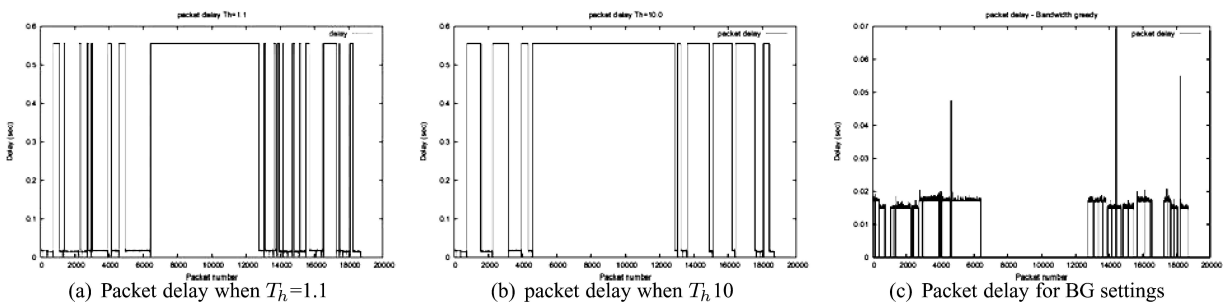
Let us define *inter-handover duration*, or *interval*, as the interval spent by the user within a wireless access network before the next vertical handover occurs. In Figure 6 by varying the duration of inter-handover times we report, respectively the number of occurrences of such intervals with respect to their amplitude in seconds. It can be noticed that by comparing the simulations, the number of times that the *inter-handover interval* is about 10 s, respectively decreases from 18 for $T_h = 1.1$, to 13 for $T_h = 1.4$, to 4 for $T_h = 10$. Generally, when varying the T_h from 1.1 to 1.4, and 10, the occurrences of short *inter-handover intervals* tend to diminish while those of longer ones increase. In the graph, the lower-right points showing durations of, respectively 300, 393, and 400 s represent the long, continuous, permanence of the MH in the GPRS network respectively for each of the three simulations (see also Figure 7(a) and 7(b)).

6.4. Overall Connection *Cost*

The previous simulation have been carried out with the aim of evaluating the impact of vertical handovers just on the performance of the applications running on the user terminal. The results have shown that by varying the handover decision policy the performance of some applications improves while that of others might worsen. According to the application actually running on the user terminal, one could try to find the optimum policy (i.e., the optimum value for the threshold T_h) to maximize its performance. Obviously, this strategy would perform the scheduling of the wireless network, disregarding nonfunctional parameters like, for instance, the connection cost charged by the network operator to the user.

While roaming, the user (or customer) is offered several wireless accesses by different network operators. The fee that he will have to pay depends on the type of contract that he has subscribed (and indirectly, on the roaming agreements between the providers whose networks he is going to access) and of course on the actual usage of the offered connections.

Typically, three different charge models can be applied to the customers:



(a) Packet delay when T_h=1.1 (b) packet delay when T_h10 (c) Packet delay for BG settings

Fig. 7. Packet delay.

Table I. Charge models for wireless networks.

Wireless networks	Charge models
Cellular modems	Time based
High speed circuit switched data	Time and volume based
GPRS, WCDMA	Volume based
Wireless LAN (IR, WiFi)	Free of charge

- Flat monthly rate
- Volume based (function of data transmitted)
- Time based (function of the connection time)

In Table I the mostly applied charge models for some wireless networks are shown.

In order to make some evaluations about the cost of handovers from the purely monetary point of view, we pictured different scenarios.

In the first scenario the customer is charged based on the time spent in the GPRS network, whilst he is not charged any fee for the use of the WiFi network. In the second scenario, the customer is charged on a volume base when accessing the GPRS network, while the use of the WiFi network is free of charge. In the third scenario both the GPRS operator and the WiFi one charge the customer for accessing their network, but in this case the GPRS' charge model is volume based and the WiFi is time based. In the following, we devise three different formulas for the computation of the fee to be charged to the customer in each of the three scenarios and analyze the suitability of the handover policies as related to the described charge models. In Figure 8 we report the time spent by the user in the WiFi and in the GPRS networks, respectively for each of the policies that controls the handovers. Let us consider the first charge scenario. In this specific case the customer is not charged for using the WiFi connection, and therefore

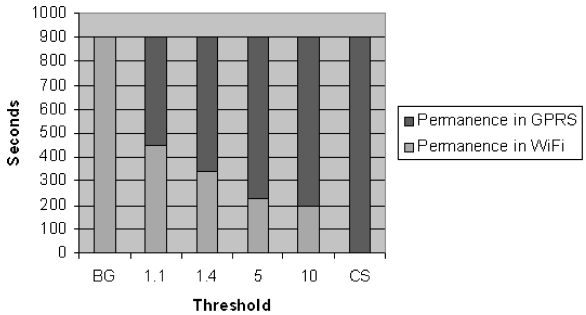


Fig. 8. Time spent by the MH in the GPRS and in the WiFi network.

the overall due fee is proportional to the amount of time that he spends in the GPRS network.

In particular, for a given communication session, let us define the following *cost* function:

$$C = T_{\mathrm{GPRS}} \times c_{\mathrm{GPRS}}(h) \tag{1}$$

where T_{GPRS} is the sum of the permanence intervals spent by the customer in the GPRS access network and c_{GPRS}(h) is the fee per unit of time (second) that the GPRS operator charges to the customer. Equation (1) represents the monetary cost faced by the customer for a given communication session. It is worth noticing that c_{GPRS}(h) may vary, depending on the actual time at which the customer is accessing the GPRS network (think about some charge models, according to which the operator charges higher fees to the customer if he access the network at specific hours of the day). The term T_{GPRS} strictly depends on both the user's movement pattern and the adopted handover decision's policy. Given the time that the customer is accessing the networks (h) and the costs associated to the usage of each single network (c_{GPRS}(h)), by adopting a suitable handover decision's policy (in this case, by tuning the threshold T_{h}) the willingness to pay expressed by the customer can be reasonably satisfied.

However, the adoption of a policy that allows the customer to save money does not give guarantees that the performance of the running application sessions will be preserved. For instance, an handover strategy with BG settings on one hand might satisfy the customer from the connection cost point of view, but will disappoint his expectation of QoS, depending on the application sessions currently going on: this is the case of an http session or a Telnet session, whose performance does not benefit from such a policy (see Figures 3 and 4). Conversely, a conservative policy like the CS one will satisfy that customer who is willing to pay for having its connections as granted as possible, but might reveal ineffective if, for example, a long file transfer is currently going on (see Figure 2).

In the second and in the third *charge* scenarios, the traffic variable comes into play. Obviously, the amount of traffic that is being carried, respectively by the GPRS network and the WiFi network depends on the handover policy that is applied. In Figures 9 and 10 we report the volume of traffic generated by two of the applications under investigation of this work (FTP and Telnet), for three different handover policies (respectively, the policy with $T_{\mathrm{h}} = 1.1$, the CS policy and the BG policy).

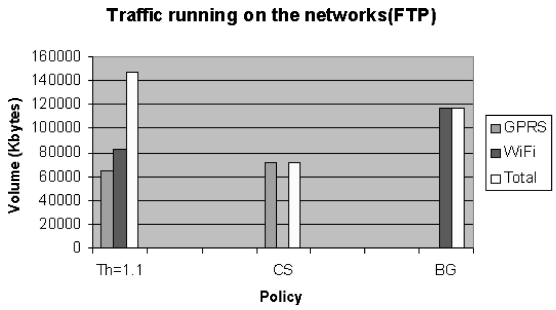


Fig. 9. FTP traffic respectively carried by the GPRS and the WiFi networks.

In the second *charge* scenario the customer pays according to the volume of the data generated by the accessed applications and transported by the GPRS network. In this scenario the following equation can be used to represent the cost for a given communication session:

$$C = V_{\text{GPRS}} \times c_{\text{GPRS}}(Kb) \qquad (2)$$

where V_{GPRS} is the global data traffic that has been carried by the GPRS network (expressed in Kb) and c_{GPRS} is the cost associated to the transport of each Kb of traffic. This time, the cost is proportional not just to the time spent in one network or in another, but to the amount of traffic that has been generated along the communication session. According to the customer's willingness to pay and to the pattern traffic that the applications may generate, a suitable handover policy can be chosen to satisfy the customer. For instance, when starting a Telnet session, one could decide to benefit of the connection's preservation provided by the CS policy, which allows the customer to save money (a small amount of data packet is generated) and shows a good performance though (see also Figure 4). Conversely, a BG policy is to be preferred

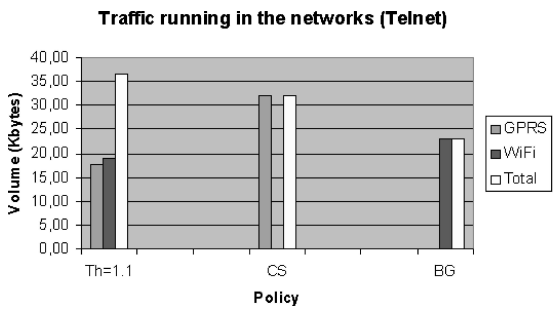


Fig. 10. Telnet traffic respectively carried by the GPRS and the WiFi networks.

if a great amount of data is to be transported (like in the FTP case).

In the presented scenarios, the handover policies that the customer more likely will prefer, according to his willingness to pay, are similar. For instance, in either of the scenario, for FTP session a BG policy is to be preferred. But for sure this strategy has to be revised in the case that congestion may occur in the WiFi network. In fact, in such case, in both the scenarios the performance of the application would undergo a downgrade, but while in the second scenario there will be no impact on the customer's fee, in the first scenario the FTP session will take a longer time to complete, and therefore the customer will cope with a higher connection cost.

In the *third* scenario the customer is charged in either of the networks, but in different ways. For a given communication session, the connection cost is given by the following:

$$C = V_{\text{GPRS}} \times c_{\text{GPRS}}(Kb) + T_{\text{WiFi}} \times c_{\text{WiFi}}(s) \qquad (3)$$

In this scenario there are several variables that come into play. Besides the ones introduced in the previous scenarios, the connection cost for accessing the WiFi network is to be considered. In particular, the comparison between the cost per unit of time for the usage of the WiFi network and the cost per Kb of data transported in the GPRS network is of particular relevance. Depending on this comparison, on the customer's willingness to pay, and on the traffic pattern of the applications, once again handover policies can accordingly be chosen. If for instance the cost for transporting a *Kb* of data in the GPRS network is much lower than that for using the WiFi network for a unit of time (e.g., a second), then bandwidth intensive applications (like the FTP) might be diverted preferably to the GPRS network (CS policies), even though their performance would be downgraded. Conversely, if the cost of using the WiFi network is much lower than that of using the GPRS, loose interactive applications (like the http) might be supported by intermittent WiFi connections (BG policy), which do not guarantee a great performance (see Figure 3) but make the customer save money though.

7. Real System Implementation

In this section we describe an actual wireless mobility framework [2] we designed and implemented in our

University campus, which is spread over a wide metropolitan area. This is the reference framework of the future implementation of the user-centric middleware presented in Section 3.

The aim of the framework in its current implementation is to let a mobile host, that is a user equipped with a PDA device, experience real wireless IP mobility while moving on a large spatial scale, by means of a middleware that leverages from three main wireless access technologies: Bluetooth, WiFi, and GPRS. In our framework, these are managed as a hierarchy of spatially overlapping access domains. While the user is on the move, the client-side of the middleware running on our mobile PDA device triggers smart switching between the currently best available and more appropriate wireless access path, based on locally collected information.

On the other hand, the network-side part of the middleware manages mobile hosts location paging and routing-update functions inside the whole campus area with an approach which extends from the cellular IP micromobility protocol. We call this approach 'WiFi bridge' since it also adds mobile hosts with the capability to seamlessly travel between disjunct WiFi access domains inside the campus by means of a temporary relay onto the GPRS access domain.

We considered as mobile host a PDA device (IPaq 3870) running Linux. Thanks to a WiFi card installed on it, the device is free to move around into a domain offering 802.11b connectivity and mobility support through our mobility middleware, which extends from the cellular IP (CIP) micromobility protocol, thus experiencing service continuity inside or between spatially contiguous wireless LAN contexts. The PDA we used is capable also of integrated Bluetooth [1] connectivity: this will be a key feature in enabling our mobility framework. In fact, the used PDA device does not expose any integrated GPRS access interface, and PDA expansion units featuring both GSM/GPRS and WiFi radio access interfaces were not yet commercially available at the time of our experimentation. Thus, we used the PDA integrated Bluetooth interface to create a personal area network (PPP/BNEP) link to a separate Bluetooth-enabled mobile phone,which in turn offered also GSM/GPRS connectivity.

To summarize, when the user brings its PDA outside the radio boundaries of a WLAN enabled domain, the device middleware will automatically detect the loss of WiFi connectivity and, as a consequence it will divert all IP connections to the GPRS access network. This will be reached by means of the Bluetooth PAN specifically set-up between the PDA and the GPRS mobile phone. As a result, a user carrying both a PDA with some running Internet applications and a Bluetooth-GPRS enabled cellular phone in his pocket can seamlessly experience wireless IP mobility inside our metropolitan campus area. The logical scenario which abstracts the physical context is depicted in Figure 11.

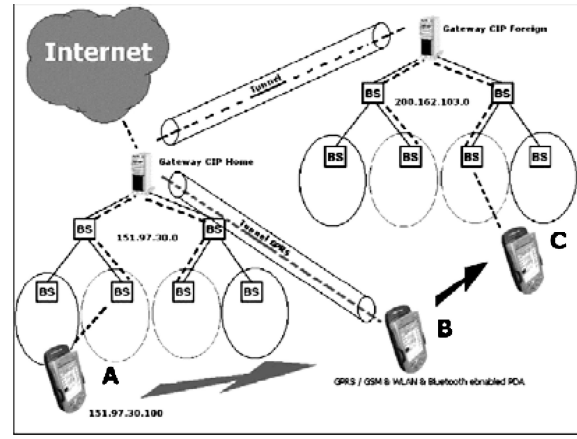


Fig. 11. Considered system scenario: MH moves out from its home domain (A) toward a distant foreign domain (C) preserving connections to the Internet through GPRS mean-while (B).

Our mobility middleware extended the original CIP middleware to allow the MH leverage from the availability of not just one but two radio access interfaces, 802.11b and GPRS via a Bluetooth link, switching between them when appropriate in order to always stay connected. Specifically, the PDA network connections are carried through the GPRS access domain when WiFi access is not available, but are carried back to the WiFi access domain as soon as an available WiFi access point is detected by the PDA middleware.

The whole software architecture in the MH is depicted in Figure 12, showing in dark color the parts that were written from scratch and in light gray those other that only needed to be adapted/modified for the embedded version of the Linux OS. As can be seen, there is a strict correlation between the CIP and the Switcher module, which practically extends its basic functionality (WLAN, intra-domain handover) to allow for a new type of handover between heterogeneous access networks (inter-domain).

Three types of wireless access are considered:

- IEEE 802.11: This module is the wireless IP interface and drive used to establish a data-link layer connection between MH and current BS. It is the only interface actually used by the MH when inside

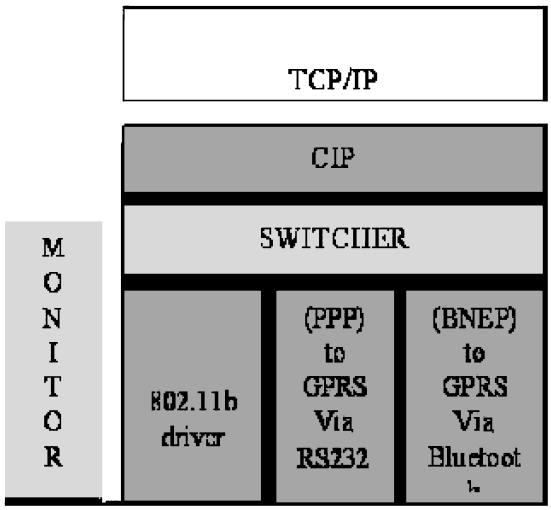


Fig. 12. Software architecture in the MH.

a CIP domain. No substantial modification has been introduced in this module.

- GPRS: The MH uses GPRS to access the network when it is outside a WLAN. The IP packets are encapsulated into point-to-point data-link frames. The other PPP endpoint is the GPRS network interface offered by a mobile phone. Likely, the MH and GPRS mobile phone are physically connected through a serial RS232 cable.
- Bluetooth: Most of the commercial mobile computing terminals do not have GPRS interface. On the contrary, they are expected to offer a Bluetooth (BT) interface. This can be utilized to connect the terminal to a GPRS phone with BT card. The basic network encapsulation protocol (BNEP) functionality of the two devices' BT protocol stack is used to seamlessly transport IP packets into BT data-link frames.

8. Conclusion

Several wireless access systems are today available in most big cities, which are frequently physically overlapping, yet belonging to separate administrative domains. Seamless mobility inside these areas is already possible but requires vertical handover capabilities at the user terminal.

We proposed a new approach in designing vertical handover algorithms, which is not aimed at optimizing resource usage of the two integrated access network from their administrators' point of view, but should try to balance, from the end-user point of view, the overall *cost* of vertical handovers with the actual benefits they bring to his actual networking needs. This approach does not lead to a single optimal handover decision function. Instead, it poses the problem to assess the impact of vertical handovers' scheduling times, as well as their frequency, with respect to the transport level protocols used by network applications.

In this early work, we set-up an example scenario involving two of the most common radio access systems (GPRS and WiFi) and simulated network traffic of the main application types, being so able to contrast the handover strategy with its overall impact on the user's work session. We currently aim at extending this set of initial tests with more varied and even uncommon network scenarios and/or user mobility patterns, in order to derive more generally valid data characterizing this correlation. Specific optimal vertical handover strategies are misleading, if optimized from a single, network-level, point of view. It is of primary importance to model a realistic vertical handover-strategy versus user-satisfaction function, to be used by each mobile user to autonomously evaluate and apply, from time-to-time, the handover decision which is more convenient to his current network needs and actual mobility model.

References

1. http://www.bluetooth.com.
2. Calvagna A, Morabito G, La Corte A. WiFi bridge: Wireless mobility framework supporting session continuity. In *PERCOM*. IEEE, 2003.
3. Campbell AT, Valko AG, Gomez J. Cellular IP. Internet Draft, 1998. draft-valko-cellularip-00.txt.
4. Du F, Ni LM, Esfahanian A-H. Hopover: a new handoff protocol for overlay networks. In *ICC2002*, pages 3234–3239, New York, New York, USA, May 2002. IEEE.
5. Fodor G, Eriksson A, Tuoriniemi A. Providing quality of service in always best connected networks. *IEEE Communications Magazine* 2003; **41**(7): 154–163.
6. Gustaffson E, Jonsson A. Aways best connected. *IEEE Wireless Communications*, 2003.
7. Mitola J. The software radio architecture. *IEEE Communications Magazines* 1995; **33**(5): 26–38.
8. Ylianttila M, Pande M, Makela J, Mahonen P. Optimization scheme for mobile users performing vertical handoffs between IEEE 802.11 and GPRS/EDGE networks. In *Global Telecommunications Conference*, volume 6, pages 3439–3443, San Antonio, Texas, USA, 2001. IEEE.
9. Perkins C. Ip mobility support for ipv4. IETF RFC 3220, Jan 2002.
10. Stemm M, Katz RH. Vertical handoffs in wireless overlay networks. *ACM Mobile Networking and Applications (MONET)* 1998; **3**(4): 335–350.
11. Sun Z, Liang L, Koong C, Cruickshank H, Sanchez A, Miguel C. Internet QoS measurement and traffic modelling. In *2nd International Conference on Conformance Testing and Interoperability (ATS-CONF 2003)*, January 2003.

12. Tripathi ND, Reed JH, VanLandinghman HF. Handoff in cellular systems. *IEEE Personal Communications*, December 1998.
13. UCB/LBNL/VINT. Network simulator—ns (version 2). software tool. www.isi.edu/nsnam/ns/.
14. Vanem E, Svaet S, Paint F. Determining user satisfaction in a scenario with heterogeneous overlaywireless networks. In *Proceedings of IASTED International Conference on Communications and Computer Networks (CCN 2002)*, Cambridge, USA, November 2002.
15. Vidales P, Chakravorty R, Policroniades C. Proton: a policy-based solution for future 4G devices. In *5th IEEE International Workshop on Policies for Distributed Systems and Networks*. IEEE Computer Society: Yorktown Heights, NY, USA, June 2004.
16. Wang HJ. Policy-enabled handoffs across heterogeneous wireless networks. Technical Report CSD-98-1027, 23, 1998.

Authors' Biographies

Andrea Calvagna received his cum laude degree in Computer Engineering from the University of Catania in 1998, and his Ph.D. in Electronic, Computer, and Telecommunication Engineering from the University of Palermo, Italy, in 2001. Since 2001, he is a contract researcher at the University of Catania, where he also serves as a teacher. His current research interests include integration of heterogeneous systems, wireless IP mobility, distributed computing, and P2P networks.