Mobility and quality of service across heterogeneous wireless networks

Andrea Calvagna, Aurelio La Corte *, Sabrina Sicari

Department of Computer Science and Telecommunications Engineering, University of Catania, Viale Andrea Doria 6, 95125 Catania, Italy

Abstract

The ability to guarantee Quality of Service (QoS) is one of the key issues in the creation of a telecommunication system. In this paper, we deal with some aspects of providing QoS in wired cum wireless communication environment, focusing on the specific feature of guaranteeing session continuity when the wireless networks, to which a mobile terminal connects to, are also heterogeneous. The terms that define what QoS is should, in this case, be reviewed in order to account also for all aspects related to user mobility inside such an environment. In this paper, such a new framework for the representation of QoS is proposed. Moreover, a wireless mobility test case is also presented. This is based on appropriate middleware we implemented, which allows a mobile terminal to experience real wireless IP mobility while moving on a large spatial scale between different sites, scattered over a metropolitan area.
© 2004 Elsevier B.V. All rights reserved.

Keywords: QoS; Wireless access; Session continuity; Heterogeneous system; WLAN internetworking

1. Introduction

The telecommunications market has shown a positive trend in the last few years. New services, tools, cultural and production opportunities and legislative awareness have all produced growth in the ICT (Information and Communication Technologies) sector, providing new possibilities towards an ever more digital lifestyle.

One of the most attractive features is the convergence of voice and data networks, the multimedia services and the wireless communication systems based on the Internet. However, this latter was originally designed to support best effort applications which only required reliability in communications. To this end, a transport protocol was used for handling occasional loss and corruption of data packets. Other communication

* Corresponding author. Tel.: +39 0957382356; fax: +39 0952507016.
 E-mail addresses: andrea.calvagna@unict.it (A. Calvagna), lacorte@unict.it, aurelio.lacorte@unict.it (A. La Corte), sabrina.sicari@unict.it (S. Sicari).

qualities and guarantees, like timeliness, were considered to be of less or negligible importance, and so a best-effort service model suffices. But now, the advent of multimedia services and the convergence of circuit-switched and packet-switched networks in IP-based infrastructures, has asked for the provision of real-time services on the Internet and for the implementation of network mechanisms supporting QoS objectives on a terminal-to-terminal basis [1].

QoS is related to the performance of the communication environment, which is characterized by quantitative measurement criteria in terms of subjective and objective user satisfaction [2]. There are no dedicated resources provided in traditional Internetworking data transport, which are largely packet switched, that allow for highly flexible and efficient data transport. The degree of resource sharing and multiplexing implies best-effort quality over a timely delivery of service, meaning that, there is no timeliness guarantee. Highly utilized networking and processing resources, which are desirable from a service provider's point of view, may result in QoS deterioration. Generally, two types of countermeasures can be taken: the provision of dedicated resources or adaptation of applications. Beginning with the Intserv and Diffserv models [3–5] and resource reservation protocol (RSVP) [6–9], many QoS control mechanisms and good theories for the maintenance and evaluation of QoS were developed and proposed by the scientific community; just as many books about the architectural, analytical and practical aspects were written [10]. Standardization bodies have faced the problem of defining performance metrics, QoS classes that must be taken into consideration, techniques to measure indicated performance and end-to-end signalling [11].

The result is that today in a wired communication environment based on the Internet Protocol, QoS can be supported by the techniques of resource reservation and those that allow the support of traffic priority [10]. However, the designers of the Internet Protocol implicitly assumed that users were stationary and did not take user mobility into account.

In a mobile communication environment, users wish to access the information they want at any given moment irrespective of their location. Furthermore, mobile computer users wish to access all the services available on the Internet, including multimedia services, with the same quality of access available in a wired environment—as though these services were provided by a desktop computer. Probably, in the near future mobile terminals will be equipped with interfaces that support multiple radio access technologies and that are, therefore, capable of operating in a communication environment that is heterogeneous [12]. So, one of the most exciting challenges will be the provision of services with QoS guarantees to mobile users connected to the Internet anytime, anywhere, with anybody/anything.

In a mobile wireless communication environment the problem of guaranteeing QoS to users or applications is more complex than in a wired communication environment. As is well known, with respect to a wired communication environment, the wireless channel is usually characterized by a lower bandwidth and a greater packet loss rate. In addition, the QoS provided by the wireless network, both in terms of throughput and packet loss rate, may change abruptly over time due to geographic impairments (e.g., physical obstructions) meteorological conditions, and so on. Also, the mobile user may move between cells characterized by different number of served users and, hence, with different available bandwidths. Even if the mobile user temporarily stands motionless the bandwidth available to it may vary as a consequence of the mobility of other users. Moreover, because of user mobility, rerouting of data packets may be necessary, and consequently, a possible variation of resources even in the wired part of the connection may occur. If the user moves between wireless networks with different access protocols, the protocol architecture must be capable of supporting sudden and sometimes abrupt variations in the characteristics of the connection in order to maintain the communication.

The task of providing consistent end-to-end QoS is even more complex when the end-to-end path of an IP session crosses multiple administrative domains. This is the case of third-generation (3G) wireless infrastructures, like UMTS, which

already adopt IP in the core network, and claim to guarantee QoS in both the access and core network [15]. Actually, appropriate mappings of QoS service attributes into policy rules, and a common framework for applying policy-based QoS control, have to be introduced to support end-to-end QoS in these systems [12–15]. In this paper we specifically focus our attention on the goal of seamlessly supporting uninterrupted connectivity to user applications while moving across multiple, possibly overlapping, heterogeneous wireless access systems, which is actually one of the most interesting problems and, at the same time, one of the key issues for QoS metric definition in a global communication environment. Moreover, from the users' point of view, session continuity is one of the fundamental parameters by which QoS is evaluated. From the service providers' point of view, it is one of the basic parameters on which resource reservation, rerouting, and other optimizations of the resource path may be carried out.

The main facets of the QoS problem are briefly represented in a layered general scheme. This scheme was designed taking into account the heterogeneous nature of the networks that a mobile user may cross as he travels. Moreover, a case study is presented which relates to the support of session continuity inside a wireless mobility framework, made up of wireless islands in the different university campus sites in Catania, which are interconnected using a broadband fiber optic network extended over a wide metropolitan area. Such a framework allows a mobile terminal equipped with a PDA device, for example, to experiment with real wireless IP mobility while moving on a large spatial scale, by means of middleware that takes advantage of the three main wireless technologies: Bluetooth, WiFi and GPRS.

The paper is organized as follows. In Section 2 the communication reference scenario is presented. In Section 3 a brief overview of the problem of supporting session continuity, and how to overcome it, is given. In Section 4 the proposed model for representation of QoS is presented. Some details and numerical results about the test performed in the case study are given in Section 5. Finally, in Section 6, we drawn our conclusions.

2. Reference scenario

In this paper we take as reference scenario a communication system made up of a wired broadband network interconnecting two or more types of wireless access networks (GPRS, WiFi, Bluetooth). Each wireless network is subdivided into different administration domains. In this scenario, a mobile user accesses the wired network through a wireless access point. Keeping in mind the need for mobility, mobile users have at their disposal mobile terminals (MTs), equipped with interfaces that support multiple radio access technologies to adequately adapt to a scenario that is heterogeneous. Some hosts (the Base Stations—BSs) act as a radio access interface, thus allowing MTs located in a cell to access the wired network. Cells from any type of wireless system may partially overlap, but we assume that even if a MT can exchange control information with more than one BS from each wireless system, being equipped with devices to support multiple radio access technologies, it only exchanges user information with one BS at a time. We assume that the area in which the MT moves is covered by at least one wireless network, depending on a radio electrical coverage structure organized in micro and macro cells.

Fig. 1 illustrates the example network scenario: three types of wireless networks (in the example given, Bluetooth, WiFi and GPRS) cover some areas
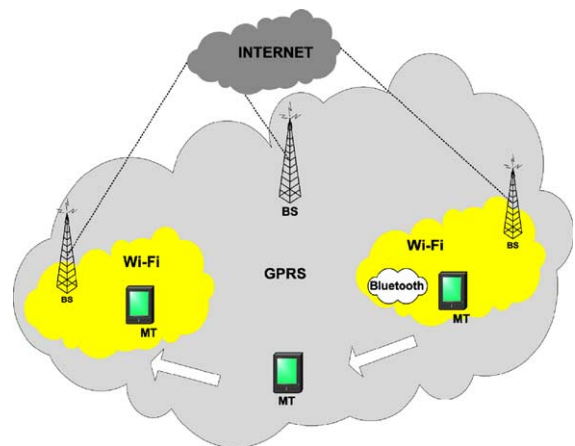

Fig. 1. Reference scenario.

of interest: each of the wireless networks is connected to a wired backbone based on IP, and the BSs of each type of network cover a different area, in such a way that an MT in any position can decide, depending on the appropriate strategy, to which BS do connect.

The aim in the depicted heterogeneous environment is to provide global seamless service coverage to a specific area, allowing access to the service independent of location. In the assumed scenario an MT can carry out different types of handovers and in particular:

- *Intrasystem Handover*. In this case, the MT passes from one wireless system cell to another cell in the same system. The access points are part of the same administration domain.
- *Intersystem Handover*. In this case, the MT passes from one wireless system cell to another cell in the same system. In this case the access points are part of different administration domains.
- *Heterogeneous System Handover*. In this case, the MT passes from a wireless system cell to a cell in a different wireless system.

The problem of session continuity is completely different in each of the three cases. In the first case, the mechanisms that allow us to keep the session open are usually intrinsically present in the link layer. In the case of the intersystem handover, the homogeneous nature of the wireless network allows us to efficiently use network layer solutions such as Mobile IP (MIP) to maintain session continuity.

The most interesting case, with regard to global mobility on a large scale, is that of heterogeneous system handover. To carry out such a handover the following are necessary:

- *Bilateral Protocol Mechanisms* supporting the communication between two equivalent logic entities that are part of the systems involved in the handover. Let's call these entities *Gateways*. In particular, we will define as *Home Gateway*, the gateway of the system of origin, and *Foreign Gateway*, the one of the network to which the user migrates.

- *Support for the transport of flows to/from the two domains*. Traffic flows originated by mobile terminals inside, i.e., the foreign WiFi environment need to be routed through the GPRS environment toward the Home Gateway or the Internet destination host. Therefore, appropriate support for interoperable routing of IP flows across heterogeneous networks must be provided.

Currently, the most viable solution is tunnelling, in that it preserves the IP context, thus keeping the session active. Obviously, the whole handover procedure has to guarantee that certain parameters (delay, throughput, etc.) must be fulfilled, therefore a set of QoS metrics will have to be defined. If MT moves out of coverage and enters into a cell from another wireless system, it would be necessary to use a tunnel between the Home Gateway and the Foreign Gateway. In this way, if the IP session is ongoing, we are able to maintain it without any interruption in the service. Fig. 2 shows where the Gateways are located in the assumed reference scenario.

We also assume that the handover between the different wireless systems can be carried out on the base of the perceived QoS itself. A MT can, i.e., decide to carry out a heterogeneous system handover to maintain the established level of
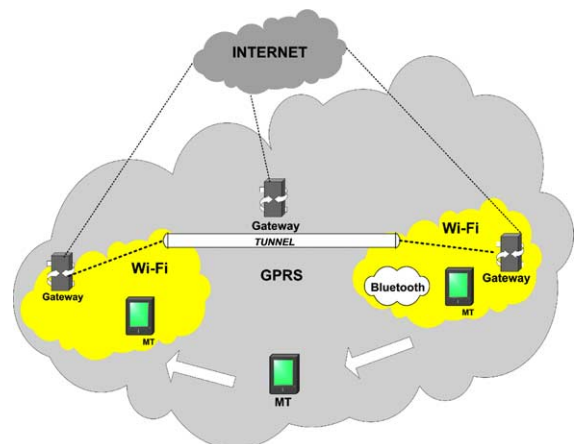

Fig. 2. Reference scenario: gateways and tunnelling between gateways.

QoS, even if the user has not left the area of coverage of the original wireless system. Therefore, it has not registered a session interruption, but has experienced a deterioration in service.

3. An overview on "session continuity"

On the Internet, today, there is a certain degree of mobility. A user can move from one site to another and has essentially connectivity and the same set of Web services available everywhere. This is generally referred to as "nomadicity" [16]. Unfortunately with nomadic mobility, users have to shut down all application sessions and restart them when they connect at the new point of attachment to the network.

For many users of the Internet this type of mobility suffices, but wireless data networks bring the potential for an enhanced mobility experience. With the proliferation of IP-based mobile devices, like PDAs, cell phone and laptops, the wireless users' biggest desire, now, is to be able to remain connected to the Internet or to their enterprise network at all times and in all places. There are a large number of IP access networks that support wireless mobility of terminal hosts; for example, wireless Personal Area Networks (PAN) and Local Area Networks (LANs), cellular Metropolitan Area Networks (MANs) and satellite and cellular Wide Area Networks (WANs). Nevertheless, if compared, all these wireless IP access technologies show very diverse technical characteristics (bandwidth, delay, bit error rate, bit energy consumption, coverage/availability, [17]), and require very diverse hardware interfaces and software drivers in order to work properly. In a large-scale network environment, featuring *seamless mobility*, the user should be allowed to roam freely between the different types of wireless access networks, with minimal (possibly zero) user intervention.

Thus, in order to implement such a mobile environment it is necessary to deal with some issues both on the network-side and the terminal-side of the network. The first of which is the integration of heterogeneous access networks. Today, it is not feasible to assume we will ever come to an ideal scenario where only one "perfect", globally spread, standard access network technology will exist. This would be quite an easy scenario for seamless mobility deployment. But, today, we instead see many wireless access techniques, with their evolution or brand new, different solutions appearing and succeeding over time. Thus, having some degree of interoperability between them is the essential requirement that we need to cope with at the network-side of a seamless mobile environment. This involves, as a minimum, agreement on common naming and addressing schemes for the following main entities: users, devices, and services. Then, each access network can privately implement its mapping of those entities into actual physical users and network devices or locations while still preserving the full environment features. Also, a common *inter*-network-level information unit format is needed, and in this case the IP packet format is very most likely to be the one, de-facto standard.

On the terminal-host side, a number of diverse hardware interfaces are required: the more there are, the richer the seamless mobility that may be experienced by the final user. Today, a modern mobile user can leverage from diverse overlapping wireless access technologies over distinct spatial scales [18], in order to always be on-line. Last, but not least, operating system, integrated support for multiple data link level interfaces, is definitely required to really seamlessly implement large-scale mobility of terminal hosts, featuring, i.e.: (1) "always best connected" traditional and vertical inter-system handover management; (2) configuration and user preferences of management tools; and (3) OSI transport-to-network layer enhancements in order to multiplex IP data flows to/from multiple network interfaces that are likely to change over time, dynamically.

To support these changes, there is a need to actively re-direct the IP flows, while preserving their *context*. Moving from the network up to the application layer, the IP flow context includes but is not limited to, security context, protocol information, policy, QoS specification, flow endpoint identifiers, congestion indication/management, header compression and accounting information.

The most common and simple example of context is supplied by the two end points of a

TCP-based application session (i.e., http browsing), that are the IP addresses and the transport-layer port numbers of each end point, which is a tuple that uniquely identifies that connection. Any change in those identifiers tears down the connection and breaks the session continuity. When a mobile terminal moves from one point of attachment (wired or wireless) to another, it is likely to have at least one new IP address assigned. This change in IP address will usually break any ongoing TCP session.

Moreover, the QoS requirements of a given session flow could be very different from application to application (see Table 1), and may even be incompatible with a given access network, or at least require substantial trade-offs in order to adapt to the available resources. What is important to note here, is that we do not want to deal with the IP flow servicing problems that may occur inside any specific network system, nor with the technical details of its underlying access technology and QoS support implementation. We want to take into account only the additional burden that has to be introduced overall to seamlessly support mobility across heterogeneous systems, that is, across systems relying on different access technologies. For true mobility to be achieved it is extremely important that the user sees that the application session persists without timing out because of a heterogeneous-system handover.

While the property known as ''Session Continuity'' [19,20] refers to mechanisms that ensure that active transport or application layer sessions are not broken due to mobility, seamless mobility is achieved when the session continuity is maintained even as the mobile device changes its point of attachment or interface type. So, a mobile terminal could be moving from a fixed Ethernet 802.3 connection to an IEEE 802.11 wireless LAN (WLAN)

interface or into a wide-area cellular interface such as GPRS or UMTS, without session interruption.

The solution to the seamless mobility problem can be classified according to the OSI layer at which it is implemented: at link layer, network layer, application layer, or in all of them, with different impact on the involved networks and terminals.

- *Application layer mobility* essentially confines the burden of managing a mobile session and the underlying changes of the IP context in the application layer, that is, inside the terminal hosts. For example, FTP, which is commonly used for downloading files, music or video, would have to be enhanced to support mobility. As a big advantage, this approach would require minimal intervention on the network side of the systems. On the other hand, all existing applications would have to be rebuilt to support mobility. As a consequence, such an approach is not considered a viable solution. Also, emerging applications (like ROAMIP or MSMessenger) tend to rely on application-layer user identification, instead of terminals. In contrast to, i.e., Mobile IP [21], these assume that mobile users have permanent identifiers that are not IP addresses. The idea is to implement *user* mobility alone, without necessarily overlapping it with *terminal host* mobility. Mobile nodes dynamically obtain IP address in each foreign network they visit, and the binding between a mobile user's permanent identifier and its host actual IP address is stored by a global location service that operates, also, at the application layer. Thus, transparent network-layer mobility support is not needed to locate users. These considerations motivate the application layer approach supporters, in contrast to the network layer approach.

Table 1
QoS specs for the main classes of Internet applications

Class	Application	Bandwidth (b/s)	Delay bound (ms)	Loss rate
CBR	Voice	32k–2M	30–60	10^{-2}
nrt-VBR	Digital video	1M–10M	Large	10^{-6}
rt-VBR	Video conference	128k–6M	40–90	10^{-3}
UBR	File transfer	1M–10M	Large	10^{-8}
ABR	Web browsing	1M–10M	Large	10^{-8}

• *Network layer mobility* hides the changes in IP address of network attachments from the upper layers, so that applications are essentially unaware of mobility enhancements. This kind of approach also simplifies the problem of providing mobility to all already existing applications. Actually, it requires just one upgrading of the operating system on the terminal host, rather than individually dealing with applications. Nonetheless, it has a more dramatic impact on the network side, where adequate (Home/Foreign Agent) support must be set up in the routers of each featured access LAN. Mobile IP is currently the most acknowledged and deployed IP mobility solution implemented at the network-layer. While it is considered to be a scalable approach in the context of the Internet, it was originally designed to support IP mobility across what once was a wired-only network context.

This obsolescent aspect of its design leads to at least a few big issues with respect to the implementation of the seamless mobility concept. First of all, session continuity is not supported. Actually, Mobile IP uses address translation to ensure that mobile nodes are reachable via permanent IP addresses. Mobile IP, just as the (wired) Internet paradigm itself, assumes that IP addresses are unique identifiers for both the terminal and its actual point of network attachment. As a consequence, changing network location implies also sending explicit terminate notification to all active transport-layer connections. This, in turn, means that even the application sessions owning those transport connections should be stopped and rescheduled in the new host location (and, what is more, the rescheduling is usually left up to the user). This is clearly in contrast with our objective: in the context of a *wired cum wireless* seamless and mobile IP access network, being able to distinguish between hosts and host locations is an essential requirement in order to preserve the IP context of applications while moving from one place to another. The IPv6 protocol natively supports this type of differentiation allowing for doubled address fields (source/destination host and source/destination loca-

tion) directly in the packet header, coupled with adequate route processing algorithms in IPv6 routers. IPv6-based enhancements of the Mobile IP protocol for wired cum wireless networks, at least partly overcoming the above issues, have already been proposed [22,23] but, as of today, the IPv6 Internet backbone is still far from being widely deployed.

In principle, approaching the IP mobility issue from the network layer will always lead to solutions that are easily scalable but, at the same time, not really suited to supporting high mobility rates. Actually, they will clearly tend to show poor performance with respect to more federated/localized mobility management approaches, which better fit the space locality of a rapidly moving terminal host.

• *At link layer mobility* can be supported by means of network interface drivers natively handling terminal mobility. As an example, IEEE 802.11 WLANs provide link-layer mobility: a device moving across an 802.11 access point within the same distribution system is able to maintain its sessions uninterrupted. Nevertheless, the very nature of this approach itself confines its scalability inside one single-access LAN. In addition, link layer mobility solutions for seamless mobility across heterogeneous access media would be extremely complex to realize. Thus, we can conclude that this kind of approach should be conveniently applied only to homogeneous (single-access technology-based) network scenarios.

To summarize this brief comparison, to develop and deploy a network layer based mobility solution may be generally considered the more viable and convenient approach. Nevertheless, existing network mobility protocols, like Mobile IP, have to be enhanced in order to provide for the extra requirements of a seamless, mobile, heterogeneous but integrated, *wired cum wireless* environment. From the final user perspective, these include at least:

• *Application session continuity support in wired cum wireless access contexts*. Some work in this direction has been recently proposed [22,24],

but a lot more remains to be done to guarantee application session persistence while performing fast system handovers between heterogeneous access networks (i.e., WiFi to/from GPRS).

- *Seamless and smart inter-system handover management*. The user should always be able to attach to the most convenient local network without any explicit intervention. The automated procedure should continuously optimize for terminal energy consumption, user movement pattern, network topology (if known), available access networks latencies, bandwidths, medium reliability, QoS support, security policies and servicing tariff plan.
- *The session should be adapted to possibly different network conditions*. Network, transport (and application) layer enhancements to support QoS-requiring flows, possibly moving across multiple, heterogeneous, IP network systems are needed. Agreement on a top-level QoS definition is a fundamental step in order to achieve this.

4. Quality of service

QoS can be defined in many ways and can include various aspects and different sets of service requirements, such as performance, availability, reliability and security. The parameters that describe QoS can be defined in a deterministic or stochastic way or through average values at suitable time intervals [10]. The meaning of QoS and its parameters can take on a meaning significantly different, depending on the point of view from which you wish to evaluate it. Customers, service providers and telecommunication engineers all view QoS in a different way, using performance metrics to evaluate QoS that might be different from each other.

In [25] a general model is presented. In this model the notions of *Intrinsic*, *Perceived* and *Assessed* QoS is illustrated. *Intrinsic* QoS is strictly determined by transport network design and the provision of network access, termination and connection [25,26]. The approach used for defining QoS is similar to that used by IETF in the definition of QoS [7,27,28] and to the approach used by

ITU and ETSI for defining the concept of Network Performance [29–32].

Perceived QoS takes into account how the service is perceived by the customer. It is a subjective type of quality, and therefore can be defined both through technical and objective parameters, linked to the parameters that represent the Intrinsic QoS and through the users' expectations of a specific service. Apart from the technical aspects, linked to the notion of Network Performance, the approach followed by the ITU and ETSI in the definition of QoS is fundamentally based on the perception of the same by the user [29,31]. Completely different, however, is the approach of the IETF that, as already mentioned, treats the problem of QoS as an intrinsic QoS problem and absolutely pays no attention to perceived QoS.

The *Assessed* QoS represents a vision of perceived QoS viewed from a higher level than perceived QoS and depends on, unlike Perceived QoS, various factors such as service price and client satisfaction with customer care [25]. Neither ITU nor ETSI nor IETF deal with assessed QoS. An overview of commonly used terminology related to quality of service in IP networks and a comparison among the approaches used by IETF, ITU and ETSI for defining QoS is shown in [26]. In this paper, starting from [26], we introduce a model that allows us to highlight the aspects of QoS in a heterogeneous and wireless communication environment. Hence, the term QoS is used with many meanings, ranging from the user perception of the service, to a set of connection parameters necessary to achieve particular service quality. Let us consider the layered structure for the representation of QoS shown in Fig. 3.

The two lowest layers are the *Basic Level* and the *Multimedia Level*. The *Basic Level* corresponds with the layer of Intrinsic QoS in the general model proposed in [26], and identifies the QoS parameters that must be considered in any type of connection. These depend on technical aspects and are determined by the type of transport network projected, from connection to termination. QoS is expressed in terms of the bit rate of transferring user data available for the service or target throughput that may be achieved, delay and delay variation (jitter) experienced by user information
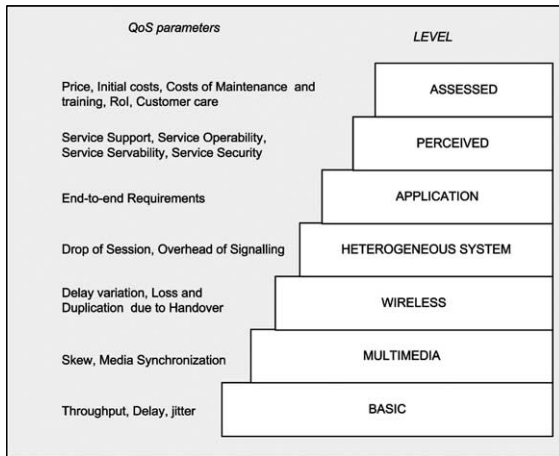
Fig. 3. QoS model.

units while passing through the network and loss rate of user information unit [26].

In wired networks it is possible to assume stationary users and low error rates, which are not valid assumptions in the wireless/mobile environments. Thus, it is not feasible to apply the same solutions in order to provide QoS guarantees, since new obstacles have to be overcome. Specifically:

Mobility: wireless users tend to move around, triggering frequent handovers between adjacent wireless cells. Optimizing the handover procedure in order to provide seamless communication is not an easy task, since resource availability may vary from cell to cell, and even local QoS policies may vary from one administrative domain to the other.

Scarcity of resources: wireless links have much lower bandwidth than wired links, due to physical limitation of the wireless media, even though this disparity is expected to be partly filled in the future.

Unreliability: noise, multipath fading, shadowing, and interference render Wireless channels much more unreliable than wired links. They are also inherently prone to location-dependent, time-varying, and bursty errors.

The *Multimedia Level* is the second layer in the proposed model. This layer was introduced taking into account the requirements of multimedia serv-

ices and basically refers to media synchronization. A multimedia stream is, actually, characterized by multiple monomedia streams related to each other by means of time, spatial or logical relationships that can be altered when the information unit crosses a network and that must be preserved [33]. At this level, the QoS parameters are basically linked to the skew occurring in the multimedia stream, that is the difference between the instantaneous delays of information units belonging to two different monomedia streams that compose the multimedia stream. As measurements of human perception of the above parameters have shown monomedia streams may appear to be "in synch" if jitter and skew are limited to appropriate values [2], the QoS parameters can be expressed as restrictions on the statistic values assumed by skew.

The third level proposed is the *Wireless Level*. This level was introduced to measure user mobility, or more specifically, to take into account the fact that when a mobile terminal passes from one cell to another the rerouting of the information units may induce a sudden variation in the end-to-end delays, and depending on the technique used for the handover, a loss or duplication of the information units that make up each monomedia stream. At this level, we assume that the MT carries out the handover through the homogeneous systems and so, does not carry out a network protocol change during the passage from one cell to another. Therefore, the connection's nominal rate does not change. The QoS parameters can be expressed in terms of the variation of average delay during the passage from one cell to another and in terms of the number of information units that are lost or duplicated during the handover. Also in this case, the QoS parameters can be expressed as restrictions on the statistic values [33].

The fourth level of the proposed architecture is the *Heterogeneous System Level*. This takes into account the fact that the handover can be carried out by heterogeneous systems, in which the mobile terminal has to carry out a change in network access protocol and a mechanism that guarantees session continuity for it to be a seamless handover. QoS, at this level, can be represented by two parameters: the probability that the session will not be maintained, and overhead in transmission

due to the necessary signal to maintain the session. The probability of a drop in the session depends on the type of network access protocol and other factors such as the speed of the MT, nominal rate variation, and so on.

At the next level of the QoS model architecture we find the *Application Level*. This level takes into account the specific characteristics of the service. The parameters of QoS are linked to the application and are expressed in terms of the type of end-to-end requirements. For example, in the case of Web browsing, the QoS parameter is the response time, while in the case of VoIP services, it is the end-to-end delay.

At the next level of the model shown in Fig. 3 we find the *Perceived QoS Level*. This accounts for how the user perceives the QoS in his experience using the service and is influenced by the customer's expectations in contrast to observed service performance [26]. These expectations reflect how the service is supported, and performance in terms of operability, serviceability and security of the service. It is comparable with its namesake in the ITU and ETSI models.

At the highest level of the model is the *Assessed QoS Level*. The Assessed QoS "starts to be seen when the customer decides whether to continue using the service or not" [25]. This depends on the quality perceived by the user, pricing, initial costs, return of investments, responses of customer care, benefits which depend on the service use in economic terms as well as productive ones.

In the following section we will focus on a possible *heterogeneous system level* implementation of the session continuity QoS feature in an experimental testbed. All low-level technical details directly related to implementing QoS to a specific type of media access protocol are outside the scope of this paper.

5. Test framework for mobile QoS

As already stated in Section 3, in order to provide session continuity features in a mixed-technologies wireless access environment, proper entities and protocol enhancements must be introduced, both in the network and the terminal. Moving to-

Fig. 4. Exemplary seamless mobility framework including heterogeneous access systems.

wards this direction, we have implemented an exemplary seamless mobility framework as shown in Fig. 4.

The wireless mobility environment is composed of WiFi wireless access domains spread inside our University Campus network, Fig. 5, whose coverage ranges are much shorter than their separation distances. Thus, WiFi access gaps could be experienced while moving inside the campus. The campus sites are interconnected by means of a broadband fiber optic network backbone, deployed over a wide metropolitan area.

The implemented framework shows three heterogeneous types of wired/wireless access systems:

Fig. 5. Catania Universitary Campus Network.

the WiFi access WLAN, the local GPRS access system and the wired campus backbone. The whole framework integrates the first two access systems by means of *Gateways* (GW), installed in the WiFi network, and connecting the WLAN to the wired Internet. Note that, appropriate IP gateways to/from the Internet are an essential part of any GPRS network infrastructure. Also, note that in order to allow for session continuity whether the terminal on the Internet or inside the GPRS or WiFi domain, proper mechanisms to preserve the IP context have been developed and installed in the mobile terminals, in the WiFi Gateways and in the involved IP routers of our framework. In particular, inter-system handover management features have to be added to the user's terminal system software and to the introduced WiFi Gateways. Both the Internet backbone and the GPRS system implemented network level macro-mobility with Mobile IP, while in the WiFi domain Cellular-IP-based micro-mobility has been enabled. As a result, in the context of this framework a user is allowed to experience mobility and session continuity to an extent directly proportional to how many of the three required access interface types are installed on his terminal. In our experience, this framework has proved to be a flexible and easy way to enable terminal host mobility using common technologies; thus it may be considered as a reference case study to evaluate any perceived QoS metric proposal across heterogeneous access systems.

Every featured WLAN has a Gateway implementing the following basic functions:

1. authentication of terminal hosts; (i.e., based on network interface MAC address) and initial location registration;
2. locations and routes updating, i.e., by means of paging mechanisms;
3. IP packets forwarding;

and also, these *newly introduced* features:

4. tunnelling toward terminal hosts located inside the GPRS system;
5. signalling protocol to manage heterogeneous system handover (also in the terminal host).

The considered access technologies have very diverse properties in terms of bandwidth, transmission media reliability, physical availability, cost, and so on. A terminal, left outside the WLAN coverage range, would normally lose all its ongoing IP connections, whereas in our framework this drawback is avoided. The Mobile Terminal (MT) system software is able to detect the loss of connection and, in the present case study, automatically diverts all connections to the always-on GPRS wireless access network. The Mobile Terminal is a Compaq iPAQ PDA with Linux 2.4.17. The MT is provided with both an IEEE 802.11b PCMCIA WiFi NIC card, and a Bluetooth (BT) interface. The BT connects the PDA to a GPRS mobile phone where, in turn, a properly configured PPP connection tunnels through the GPRS access network. Even though inter-system handovers are performed in the shortest possible time, some problems may still occur, i.e., in terms of loss or duplication of packets, and temporary introduction of extra delays, as shown by our test results. Tests on roaming functionality between WLANs (intra-system) have not been carried out because, in that case, our framework performs a Cellular IP handover procedure, which has already been evaluated in [34,35].

We investigated the impact of handovers on constant bit rate UDP traffic flows, generated by a source host in a fixed subnet and sent to the MT. The packet rate was set to 25 pps while the packet size was set to 100 bytes. Packet losses and packet delays were measured during handovers from WiFi to GPRS, and back. More specifically, during our experiments the MT cyclically performs the following sequence of actions:

1. The MT moves from WiFi to GPRS.
2. The MT remains in the GPRS domain for 7 seconds.
3. The MT moves from GPRS to WiFi.
4. The MT remains in the GPRS domain for 7 seconds.

We let the MT remain within a certain domain for 7 seconds in order to reach the steady state before beginning a new handover procedure.

5.1. Delay measurements

In Fig. 6 we show the delivery delay of UDP packets, during handoffs from WiFi to GPRS and vice versa. In this figure we highlighted the four different stages that the MT experiences.

Initially, the MT is located inside a WiFi domain, experiencing packet delays in the order of 2 ms. When the MT goes outside of WiFi radio range, it performs a handover to the GPRS access network. Packet losses at the MT are experienced in this stage in the order of tenths of packets. They are due to the time gap required to complete the necessary routing update operations in the home network GW. During the subsequent GPRS access stage, the packet delay along this route increases to the high delay values typical of GPRS channels.

Subsequently, we have the GPRS to WiFi handover stage. The behavior of this stage is rather complex since we can distinguish three different time intervals, as can be seen in the Fig. 6. The peak in the network end-to-end delay is due to a sudden consumption of computing resources and should not be taken into account. We observe that the MT still receives UDP packets from the GPRS tunnel for a time lapse, marking the actual end of the handover stage. Only after that does the packet delay fall back to the lower WiFi values. The goal
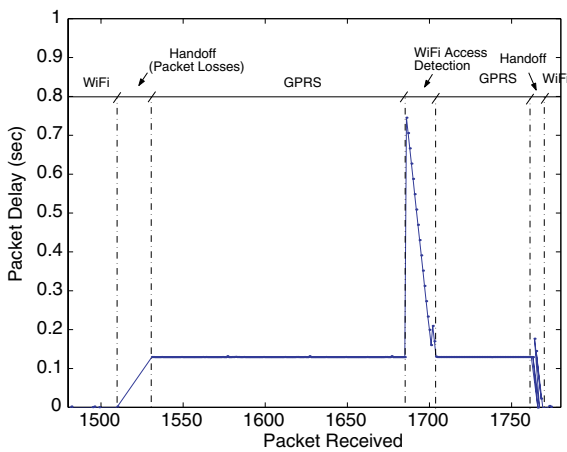
of optimized handover procedures should be to shorten the overall duration of the GPRS to WiFi handover, while avoiding unstable (ping-pong) behavior. It is otherwise possible that when the MT comes to a domain boundary it would indefinitely perform quick handovers back and forth between the two access networks.

The last transient concluding the GPRS to WiFi handover procedure is characterized by oscillations in the packets delays, as shown in Fig. 7. In fact, a *soft* heterogeneous system handover is performed, i.e., for a brief period the MT receives IP packets from both WiFi and GPRS access interfaces. Therefore, as they experience different RTT (round trip times) delays, packets are not received in sequential order. In particular, late packets are those coming from the GPRS tunnel that have been forwarded by the GW on the GPRS tunnel. Note that as a *soft* handover has been performed no packet losses will be experienced by the MT in this stage.

5.2. Loss measurements

Test results, shown in Fig. 8, are taken at different values of the RTT between the MT and its WiFi domain's GW, as well as varying the beacon interval duration, that is the time between two consecutive WiFi *beacon* packets. If compared to a standard WiFi handover, handover towards GPRS has a longer duration because of the time

Fig. 6. Packet delay experienced in heterogeneous (WiFi-GPRS) handover test. Four stages are highlighted: *WiFi access*, *WiFi to GPRS handover*, *GPRS access*, *GPRS to WiFi handover*.
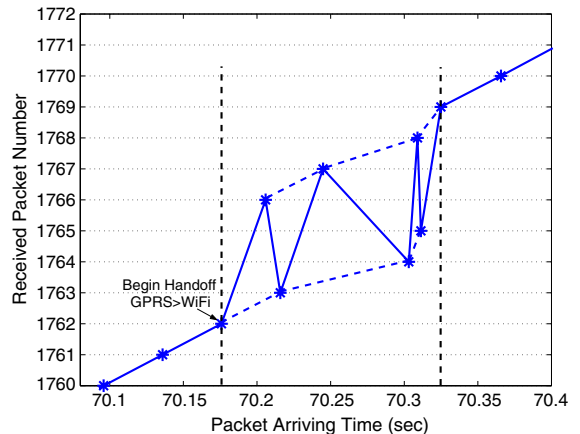
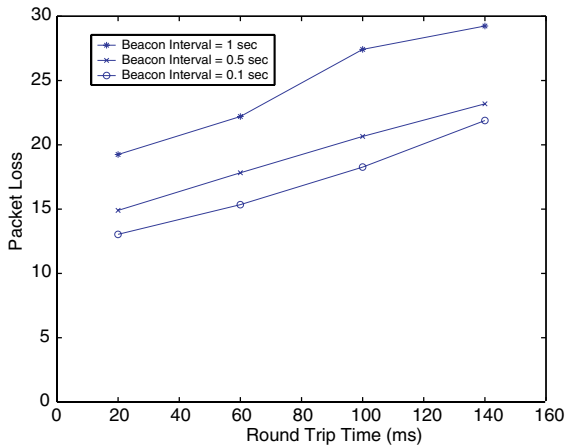Fig. 7. Packet arrival times during GPRS to WiFi soft handover showing non sequential packet arrivals.

Fig. 8. Packet loss measurements result for the WiFi to GPRS handover procedure, with respect to varying *Bi* and *RTT* values.

necessary for the MT to detect that the current position is outside the radio boundaries of its home WiFi domain. We call this delay: *alert delay, (Ad)*. As shown in Fig. 8, packet losses are directly proportional to the values of RTT along the GPRS path and increase as the *Bi* duration increases. We have to stress that, in comparison to the ordinary WiFi to WiFi handover (see [35]), WiFi to GPRS *system* handover causes a larger number of packet losses. This is mainly due to the fact that the *alert delay* must be greater than the beacon delay, to keep the MT from misjudging the ordinary delay of WiFi intra-system handovers as a lack of connectivity due to exit from the WiFi domain boundaries.

6. Conclusions

In this paper we proposed a QoS framework which extends existing QoS standards from IETF, ETSI, I-TUT, to focalize on the new scenario of an integrated heterogeneous system. In this context a comprehensive QoS framework should be specifically designed in order to address seamless mobility/session continuity features of a given environment. We started our analysis from an actual reference scenario and as a result proposed a layered model for mobile quality of service. This

model proposes abstraction of the many underlaying access technologies, whose technical QoS implementation details are hidden in the lower layers. The proposed framework easily allows metering and comparison the level of support for seamless mobility that a wireless IP access environment, possibly heterogeneous, is actually providing. We believe that, in order to achieve wide-area wireless seamless mobility, it is mandatory to first cope with definition of interoperable mobile-QoS support features. These include mobile routing and paging services but, also, a smart and convenient way to seamlessly switch, from time to time, to the most appropriate wireless access path, and related access interface, based on the current application context and user-specified policies.

A case study has been discussed, consisting of an example framework we designed and implemented to enable wireless/mobile IP user experience service continuity while moving in a heterogeneous wireless access environments. The access systems, consisting of 802.11b disjoint areas inside a GPRS network, have very different characteristics in terms of bandwidth and, what is more, present only a transport facility to the upper layers. Performance measures show that a service degradation occurs in terms of packet losses and delay when the MT moves from the WiFi access domain to the GPRS access domain. This is basically due to the bandwidth mismatch between the two environments and could be improved by appropriate setting of certain timers. Despite that, access continuity is preserved. As a natural extension of our work we are currently running tests to study the scalability of the proposed system. We also plan to approach the problem of making applications aware of the currently used network interface in order to adapt, if possible. We believe this is an important requirement, in order to let all these IP access technologies act as a unique infrastructure.

References

[1] N. Seitz, ITU-T QoS standards for IP-based networks, IEEE Communications Magazine 41 (6) (2003) 82–89.
[2] R. Steinmetz, K. Nahrstedt, Multimedia—Computing, Communications and Applications, Prentice Hall International, New York, 1995.

[3] R. Braden, D. Clark, S. Shenker, Integrated Service in the Internet Architecture: an Overview, RFC 1633, 1994.

[4] J. Liebeherr, N. Christin, A QoS architecture for quantitative service differentiation, IEEE Communications Magazine 41 (6) (2003) 38–45.

[5] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, An Architecture for Differentiated Services, RFC 2475, 1998.

[6] D. Estrin, S. Shenker, D. Zappala, L. Zhang, S. Deering, RSVP: a new resource reservation protocol, IEEE Network 7 (5) (1993) 8–18.

[7] R. Braden, D. Clark, S. Shenker, Integrated Services in the Internet Architecture: An Overview, IETF RFC 1633, July 1993.

[8] Y. Bernet, R. Yavatkar, P. Ford, F. Baker, L. Shang, K. Nichols, M. Speer, A Framework for Use of RSVP with Diff-serv Networks, Internet Draft, draft-ietf-diffserv-rsvp-01.txt, 1998.

[9] Y. Bernet, R. Yavatkar, P. Ford, F. Baker, L. Zhang, A Framework for End-to-End QoS Combining RSVP/Intserv and Differentiated Services, Internet Draft, draft-bernet-intdiff-00.txt, 1998.

[10] V. Firoiu, J. Le Boudec, D. Towsley, Z.L. Zhang, Theories and models for Internet quality of service, Proceedings of the IEEE 90 (9) (2002) 1565–1591.

[11] Y. Maeda, QoS standards for IP-based networks, IEEE Communications Magazine 41 (6) (2003) 80.

[12] G. Fodor, A. Eriksson, A. Tuoriniemi, Providing quality of service in always best connected networks, IEEE Communications Magazine 41 (7) (2003) 154–163.

[13] W. Zhuang, Y.S. Gan, K.J. Loh, K.C. Chua, Policy-based QoS architecture in the IP multimedia subsystem of UMTS, IEEE Network 17 (3) (2003) 51–57.

[14] R. Yavatkar, D. Pendarakis, R. Guerin, A Framework for Policy-Based Admission Control, IETF RFC 2753, January 2000.

[15] S.I. Maniatis, E.G. Nikolouzou, I.S. Venieris, QoS issues in the converged 3G wireless and wired networks, IEEE Communications Magazine 40 (7) (2002) 44–53.

[16] Y. Li, V.C.M. Leung, Supporting personal mobility for nomadic computing over the Internet, ACM Mobile Computing and Communications Review 1 (1) (1997) 22–31.

[17] Wireless World Research Forum, The Book of Visions 2001, WWRF, December 2001. Available from: <http://www.wireless-world-research.org/general-info/BoV2001-final.pdf>.

[18] M. Hannikainen, T.D. Hamalainen, M. Niemi, J. Saarinena, Trends in personal wireless data communications, IEEE Computer Communications 25 (1) (2002) 84–99.

[19] A. Mohammad, Seamless mobility requirements and mobility architectures, in: IEEE Global Telecommunications Conference (GLOBECOM '01), vol. 3, 25–29 November 2001, pp. 1950–1956.

[20] Z.R. Turanyi, Cs. Szabo, E. Kail, A.G. Valko, Global Internet roaming with ROAMIP, ACM SIGMOBILE Computing and Communications Review 4 (3) (2000) 58–68.

[21] C.E. Perkins, Mobile IP: Design Principles and Practices, Addison Wesley, 1998, ISBN 0-201-63469-4.

[22] C. Castelluccia, A hierarchical mobility management scheme for IPv6, Third Symposium on Computers and Communications (ISCC'98), Athens, Greece, 30 June–2 July 1998.

[23] D.B. Johnson, C. Perkins, Mobility support in IPv6, June 1996, draft-ietf-mobileip-ipv6-01.txt.

[24] A. Pappalardo, A. Calvagna, G. Morabito, WiFi mobility framework supporting GPRS roaming: design and implementation, in: IEEE International Conference on Communications, 2003, ICC 11–15 May 2003, vol. 1, pp. 116–120.

[25] W.C. Hardy, QoS Measurement and Evaluation of Telecommunications Quality of Service, Wiley, New York, 2001.

[26] J. Gozdecki, A. Jajszczyk, R. Stankiewicz, Quality of service terminology in IP networks, IEEE Communications Magazine 41 (3) (2003) 153–159.

[27] E. Crawley et al., A framework for QoS-based routing in the internet, IETF RFC 2386, August 1998.

[28] S. Blake et al., An architecture for differentiated services, IETF RFC 2475, December 1998.

[29] ETSI. Network Aspects (NA): General Aspects of Quality of Service (QoS) and Network Performance (NP), Technical Report ETR003, 2nd ed., October 1994.

[30] ITU-T, Terms and definitions related to Quality of Service and Network Performance including Dependability, ITU-T Rec. E.800, August 1993.

[31] ITU-T, General Aspects of Quality of Service and Network Performance in Digital Networks, including ISDN, ITU-T Rec. E.350, March 1993.

[32] ITU-T, Internet Protocol Data Communication Service—IP Packet Transfer and Availability Performance Parameters, ITU-T Rec. Y.1541, February 1999.

[33] A. La Corte, A. Lombardo, S. Palazzo, G. Schembra, Control of perceived quality of service in multimedia retrieval services: prediction-based mechanism vs. compensation buffers, Multimedia Systems 6 (2) (1998) 102–112.

[34] A.G. Valko, A.T. Campbell, J. Gomez, Cellular IP, Internet Draft: draft-valko-cellularip-00.txt, November 1998.

[35] A.G. Valko, A.T. Campbell, J. Gomez, On the analysis of cellular IP Access Networks, in: Proc. 6th IFIP International Workshop on Protocols for High Speed Networks, Salem, 25–27 August 1999.

Andrea Calvagna received his degree in Computer engineering from the University of Catania in 1998, and the 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 teaching assistant. His current research interests include IP mobility, integration of heterogeneous systems, wireless IP communications, distributed computing and P2P networks.

Aurelio La Corte received a degree in electrical engineering from University of Catania and the Ph.D. degree in electronic and computer science in 1988 and 1992, respectively. Since 1994 he has been at the University of Catania, where he is now an Associate Professor of Telecommunications Networks. His current research interests include mobile systems, quality of service management, multimedia traffic modelling and digital signal processing.

Sabrina Sicari received her degree in Electronic Engineering from University of Catania, Italy in 2002. Since 2003 she has been at the University of Catania, where she is currently a Ph.D. student with the Department of Computer Science and Telecommunication Engineering, Her current research interest include mobile communication and network security.