

QoS Issues in Mobile IP: Challenges, Requirements and Solutions

Runtong Zhang
Lab of Computer and Network Architectures
Swedish Institute of Computer Science
Box 1263, SE-16429 Kista, Sweden
E-mail: runtong@sics.se

Abstract: Guaranteeing Quality of Service (QoS) in the Internet is a topic of active research. The technologies that have drawn the most attention are several different models - the IntServ, the DiffServ and the MPLS models. IntServ uses the per-flow approach to provide guarantees to individual streams, while DiffServ provides aggregate assurances for a group of applications, while MPLS tries to provide the efficiency and simplicity of IP routing together with the high speed switching of ATM by integrating the label-swapping paradigm with network layer routing. Sometime, ATM itself is also considered as one QoS technology because its powerful ability of handling Internet traffics with different QoS requirements. However, all these models have been designed to work for fixed Internet so far. There exists an urgent and important requirement today to study and identify the possible modifications that need to be made to make them suitable for the coming mobile Internet. In this paper, we aim to develop a thorough understanding of the unique opportunities and challenges, problems, requirements and candidate solutions, that arise in modifying the existing QoS models in order to enable them to efficiently work in mobile Internet. Some existing work is outlined as a survey, while some new ideas and proposals are presented from the research viewpoint.

Index terms: mobile IP, quality of service (QoS), problems, solutions

1. Introduction

Mobile and wireless access to the Internet is becoming more and more popular due to the rapid spread of wireless access technologies with various speeds and ranges, such as IRDA-LANs, IEEE 802.11, Wireless ATM, GPRS, UMTS, etc. Devices used include mobile phones, PDAs, notebook computers and so on. Mobile IP [RFC2002] could become a common platform for mobile access, regardless of the underlying access technology, providing solutions to the many security, routing, and address problems.

An important issue in the Internet, and consequently in every network connected to it, is to support for multimedia applications (video, voice). The applications have specific requirements in terms of delay and bandwidth which challenge the original design goals of IP's best effort service model, and call for alternate service models and traffic management schemes that can offer the required quality of service (QoS, [NX99]). The term "QoS" in the Internet is a topic of active research. In this area, there are some different understandings to this concept because of the different viewpoints of research and application areas. Generally speaking, QoS designates a set of parameters, intended to represent measurable aspects of the subjective "user perceived quality". Criteria taken into account [E800] involve concepts such as service availability, retainability and integrity, transmission characteristics, as well as subjective estimates. There are several approaches proposed to offer QoS to the Internet users, among which Integrate Services (IntServ, [RFC2210]), Differentiated Services (DiffServ, [RFC2474, 2475]) and Multi Protocol Label Switch (MPLS, [RFC3031]) are most known. IntServ uses the per-flow approach to provide guarantees to individual streams, DiffServ provides aggregate assurances for a group of applications, while MPLS tries to provide the efficiency and simplicity of IP routing together with the high speed switching of ATM by integrating the label-swapping paradigm with network layer routing. Sometime, ATM itself is

also considered as one QoS technology because its powerful ability of handling Internet traffics with different QoS requirements.

In fact, people from different areas have different understandings to the QoS concept. For instance, ITU-T, in recommendation E.771 [E.771], established the criterion that the quality of a mobile call for given service should be that of a fixed call plus an additional switching stage. Recommendation E.800 [E.800] defines QoS in a qualitative manner as the collective effect of service performance, which determines the degree of satisfaction of a user of the service. In the meantime, IP (Internet protocol) people sometime just consider the QoS concept as IntServ and DiffServ models. In fact, the communication (ITU-T) definition of QoS generalizes the IP one. In this paper, we may consider the QoS concept in a general sense, while the IntServ and DiffServ might be treated with stresses.

A common feature of the QoS models is the service contracts, either explicit or implicit, between the users and Internet Service Providers (ISPs) should be established. However, most of aforementioned QoS models are designed for fixed network computing environment. Due to the characteristics of mobile computing environment, the contracts based on these models can hardly be implemented and some enhancements to adapt them to be applied in the mobile computing environment are urgently needed.

The QoS solution for Mobile IP should satisfy obvious requirements such as scalability, conservation of wireless bandwidth, low processing overhead on mobile terminals, providing hooks for authorization and accounting, and robustness against failures of any Mobile IP-specific QoS components in the network. While it is not possible to set quantitative targets for these desirable properties, the QoS solutions must be evaluated against these criteria, and even the development of wider band technology and application software, and much more.

Widely speaking, in addition to the aforementioned QoS models, most the research work on the enhancement of the Internet service ability, even at every network layer, falls in the realm of QoS. However, to make this paper really focal and efficient, the "QoS work" in a wider sense is not considered here.

In the literature, considerable attention has been paid on the QoS issues in a mobile computing environment. The references listed in the end of this paper are some examples in this direction, but far from exhaustion. A comprehensive study can be found in the survey works [S96] and [C01].

Most of the work about the QoS issues on the mobile Internet thus far is either case study or general survey. In this paper, we aim to develop a thorough understanding of the opportunities and challenges, problems and candidate solutions, that arise in modifying the existing QoS models in order to enable them to efficiently work in the mobile Internet. Some existing work is outlined as a survey, where there are references excited. When there are no references excited in the paper, the authors' viewpoints on new research trends are proposed. We apply our considerations to Mobile IPv6 [JP01], while mobile IPv4 [RFC2002] may also be applied.

The rest of this paper is organized in the following order. Section 2 studies the general problems and challenges arisen in adapting the QoS concept for existed wireline networks into mobile/wireless networks. In section 3, we outline some new technology terms in mobile QoS specification. The terms are far from exhaustion, and they are useful in exploring the insights of this area. Known concepts are not included. Sections from 4 to 8 are the main parts of this paper and they are devoted to candidate solutions to variously specific problems. The final section concludes this paper.

2. Problems for Supporting QoS in Mobile Computing Environment

With the appearance of a wide range of powerful and inexpensive mobile devices, such as laptop computers, mobile phones or personal digital assistants (PDAs), mobile users are not satisfied with just obtain traditional voice service or offline computation work. They desire to attach to the Internet at any time when they are roaming and receive service at the same level just as stationary users. However, the applications currently provided in the Internet take no account of mobile hosts. They assume that end users are static hosts (i.e.

hosts with fixed network attachment points) and network topology will not change in the absence of node or link failures. The overall mobile network topology changes dynamically as mobile nodes (MNs) move from one point of attachment to another. In addition, the wireless transmission links and mobile device are generally in a worse position compared with their fixed counterparts in providing QoS to the users. All these factors cause considerable problems, which challenge very much the development of Internet.

QoS parameters for typical applications cover various aspects, such like bounds for bandwidth, packet delay, packet loss rate, jitter and much more. Certain additional parameters that deal with problems unique to wireless and mobile networks are required. Generally, they can be enumerated as below.

1. Mobility problem: Mobile IPv6 enhanced mobility will enable seamless handover between multi-access networks. IP telephony or such services will become seamless even in multi-access environment. However when handovers occur in such environment, some applications such as Web browse and file transfer, which use TCP as transport protocol will become problem or its performance will be degraded.

When a mobile node (MN) using mobile IP undergoes handover from one access router to another, the path traversed by MN's packet stream in network may change. Such a change may be limited to a small segment of the end-to-end path near the extremity, or it could also have an end-to-end impact. Further, the packets belonging to MN's ongoing session may start using the new care of address after handover, and hence, may not be recognized by some forwarding functions along the old path that use IP address as a key. Finally, handover may occur between the subnets that are under different administrative control.

2. High bit error problem: The wireless link is unreliable (optimal packet lengths have to be small) and relatively insecure. Packet loss in a mobile environment is an important issue to be considered because of the limited bandwidth of a wireless network and the possible fading and blackout situations that can occur when a mobile moves from one cell to another.

3. Bandwidth limitation problem: Generally speaking, the bandwidth of the wireless link connecting an MN to the static segment of the network is significantly lower than the one of the wired links between static hosts. This causes a serious degradation to the traffic performance especially to the real time applications.

4. Resource constraint problem: Resource constraints always hold in the cases of memory and storage capacity of the mobile device and in terms of their computational power. Mobile hosts have also tight constraints on power consumption, relative to desktop machines, since they usually operate on stand-alone sources such as battery cells. Mobile systems use limited power batteries and hence have power restrictions that must be taken care of. The network interface card of the computer consumes almost 14% of the total power [SK97]. Transmitting and receiving of packets expend power and will have to be controlled when the mobile battery power is low.

5. IP tunneling problem: The General Packet Radio Service (GPRS) architecture comprises a set of GPRS supporting nodes (GSNs), nodes equipped with GPRS compliant protocol stack. Two types of GSNs are defined: the supporting GSN (SGSN) and gateway GSN (GGSN). The latter acts as a logical interface to external packet data networks (PDNs), and consequently to the global Internet, also maintaining routing information used to tunnel IP datagrams to the correct SGSN, while the former is responsible for servicing the mobile stations currently located within its service area. Communication between the SGSN and GGSN is based on IP tunnels [P96-1]. This means that standard IP packets, as soon as they reach a GSN (SGSN and GGSN), are encapsulated in new IP packets (i.e., a new IP header is added) and routed accordingly. The GPRS Tunnel Protocol (GTP), implemented at both the GGSN and SGSN, is responsible for performing this encapsulation. The first point within the GPRS network where IP packets are interpreted (i.e., header information is accessible) is the GGSN. This restriction implies a major difficulty in applying IP QoS frameworks, such as IntServ/RSVP and DiffServ, within the GPRS network. An RSVP message, say, would be transported transparently within the GPRS network toward the GGSN or vice versa, causing no reservations at intermediate nodes (routers). The situation is quite similar if DiffServ is applied. The IP packet passes through the network, and the DiffServ (DS) fields not available

until the packet reaches the GGSN. Another problem arising from this tunnel-based communication is that different data flows addressed to the same terminal (same IP address) are treated in the same manner. Routing within the GPRS core is based on the IP addresses of GSNs. Discrimination of packets in GSNs is based on tunnel identifiers (TIDs). There is a one-to-one mapping between IP addresses and TIDs. Such implementation is not well aligned with the IntServ framework, where micro-flows are a basic concept.

3. New Terminology in Mobile QoS Specification

During these years, considerable research interests have been devoted in the mobile QoS field. As a result, some definitions concerning the QoS issues in the mobile Internet have appeared in the literature. To better understand the latest achievements, we outline some new but promising terms, most of which will be cited in the remainder of this paper.

1. Mobile QoS reference point: In [E800] it is stated that QoS measures are only quantifiable at a service access point. However, real time feedback and network response to QoS fluctuations would require the establishment of a point of reference of redefining and measuring the QoS for a mobile connection in terms of network performance. In fixed ATM networks, this point resides at the entrance to the ATM switch, i.e. at the ATM B-UNI. For a mobile system however, the overall QoS objective should include the air interface for to be meaningful to the end-user. We will assume this point residing at the interface of the fixed radio access system towards the mobile terminal.

2. Mobile QoS components: A connection involving at least one mobile user can be viewed as the concatenation [T97] of fixed and wireless links. A mobile QoS (M-QoS) therefore, comprises of a. a fixed network component (F-QoS), relating to QoS objectives for the wireline (ATM) links, and b. an air-interface component (AIF-QoS), relating to QoS objectives for the wireless (radio) links. Handover (HO) associated parameters, of spatial (e.g., HO rate per cell) or temporal (HO rate per call) significance, will play a dominant role in the performance of mobile networks. It seems reasonable, that HO relates more closely to radio calculations, rather than fixed ones. Nevertheless, certain types of HO operations affect the fixed part of the network (e.g., inter-switch handover). To resolve this, we assume that all quality issues relating to the interaction of a wireless access part of the network, with its fixed counterpart in a static operation mode, can be grouped separately (in AIF-QoS) from those directly associated with HO. Therefore, we introduce an additional logical refining: a handover component (HO-QoS), relating to all quality issues directly influenced by user mobility.

3. Mobile QoS parameters: QoS objectives should include appropriate metrics. A clear distinction is made between network performance parameters that can be objectively measured and subjective QoS parameters depending on user perception. The most indicative QoS metrics are the ones mostly affected by network performance. Considering their scope, metrics could be classified as call level and transport level QoS parameters.

4. Loss profiles: Due to the high loss characteristics of the mobile networks, it will advantageous to applications if they can characterize a way in which packets should be dropped in such cases. The QoS parameter "loss profiles" [S96] gives applications an opportunity to choose between a bursty loss and distributed loss in case of an overloaded situation. An audio application may choose to have a bursty loss because the output is still tangible if a few words are dropped. A distributed loss is better for a video application because it will appear as flicker on the screen. Considerable attention has paid in the distinction of "loss profiles", most of which are known as "TCP optimization".

5. Power Level: Because, as discussed, the base station (BS) needs to be aware of the power situation in the mobile so that it can change the way data is sent to the mobile. The "power level" [SK97] parameter informs the BS about the battery power situation in the mobile and the BS changes the way it schedules packets based on the power profiles. The profile categorizes the way packets must be sent in a low power situation. While some applications would like to reduce the average rate of sending data others may choose to send some important packets as in layered video.

6. In [CK01] a new IPv6 option called "QoS Object" is introduced depending on the context, the QoS Object is included as a Destination Option or a Hop-by-Hop Option in IPv6 packets carrying Binding Update and Binding Acknowledgment messages. When included as a Hop-by-Hop Option, QoS Object triggers certain QoS procedures at the intermediate network domains. This document describes these QoS procedures for the cases of best-effort, MPLS, DiffServ and IntServ domains, which practically cover all the cases of QoS enabled network domains that would be available in near future. QoS Object is included, depending on the context, either as a Destination Option or as a Hop-by-Hop Option along with the packets carrying Binding Update and Binding Acknowledgment options. The basic idea is to include QoS Object as a Hop-by-Hop option along with the binding message that travels in the same direction (HA to MN, CN to MN or MN to CN) as that of MN's QoS-sensitive packet stream. As this packet traverses different network domains in the end-to-end path, the QoS Object is examined at these network domains to program QoS support for the MN's data packets.

7. Probability of seamless communication: Maintaining a reservation when a mobile moves between regions is a challenge because of possible blackout situations during handover. A scheme is required to define how smooth this transition should be since it affects the QoS of an application. The "probability of seamless communication" parameter [S96] defines the nature of breaks that can be allowed in the service. Based on this parameter, advance buffering at the neighboring cells must be made so that the data is available when the mobile moves into that region.

8. Shadow cluster: The "shadow cluster [LAN97]" concept is a concept used to improve resource allocation and call admission in ATM-based wireless networks. This idea can be used to allocate resources that need to be reserved for call handovers, and to determine if a new call should be admitted to a wireless network based on the call's requirements and local traffic conditions. The shadow cluster concept is targeted for ATM-based wireless networks with a micro/nano-cellular architecture, where service will be provided to users with very diverse requirements. In these networks, and as a consequence of the small cell sizes, mobile users will typically experience a high number of cell handovers during their connections' lifetime. With shadow clusters, the QoS of mobile calls can be improved by reducing the number of dropped calls during handovers, and by disallowing the establishment of new calls that are highly likely to later result in a dropped call. The framework of a shadow cluster system is completely distributed, and can be viewed as a message system where a mobile terminal informs the base stations in the neighborhood about its requirements, position, and movement parameters, so that the base stations project future demands, reserve resources accordingly, and admit only those calls that can be supported adequately.

9. Mobiware: "Mobiware" is a novel QoS-aware middleware platform proposed in [CRLS97], which operates between the application and radio ATM link layers. At the heart of the Mobiware platform lies a QoS controlled handover algorithm which exploits the inherent scalability of audio and video flows. Implicit in the term "QoS controlled" is the notion that audio and video flows can be represented as multi-layer scalable flows and adapted during handover to meet fluctuating network conditions based on a user-supplied QoS adaptation policy. Novel aspects of the Mobiware QoS controlled handover algorithm include the use of mobile soft-state and hard-state to represent mobile flows, aggregation techniques for handling and transporting mobile flows to and from mobile devices, and re-routing and QoS re-negotiation anchor points which limit the impact of small-scale mobility on the wireline networks.

4. Mobility Problem

4.1 Location-Dependency

Mobility creates unique problems in the wireless environment: network system needs to be aware of the whereabouts of the mobile user so that incoming calls can be delivered to the customer seamlessly. In the wireline case, routing decision is based on the network-prefix

portion the IP Destination Address field in the IP header. In the untethered connectivity, the MN may move away from its home link. Therefore it is necessary to facilitate the mobile routing whilst roaming. In addition, during a connection, the MN may change from cell to cell within the same service network due to user mobility in a single session. Mobile IP allows applications to be connected all the time even when the mobile user is roaming [RFC2002].

The solution for the mobility-created information-delivery problem can be resolved by universal adoption of Mobile IP. The mobile IP enabled network can automatically reroute the incoming applications to the targeted recipients when they are detached from their home system and attached to the visited system.

4.2 Consistent QoS in Mobility

Mobile IP provides users with the roaming capabilities outside of their home networks without having to tear down the connection and re-establish a new session. This feature is called macro-mobility as the handover involves with roaming between servicing wide-area wireless data systems. On the other hand, when mobiles move from one base station coverage area to another, Intra-system handovers take place. Mobility management within the same system is called micro-mobility. While Mobile IP handles well in the macro-mobility scenario, the protocol has limitations in the intra-system micro-mobility.

1. Lost packets during handover: In the 3G mobile IP hard handover scenario, when a mobile roams from one cell site to another, there is no mechanism to allow smooth forwarding of packets from Home Agent to the new Foreign Agent. Packets might be lost due to the hard handover switchover.

2. Re-negotiation of QoS requirements in the new cell: Anytime a mobile moves from one cell to another, it triggers a new QoS reservation from Home Agent to Foreign Agent. There is no guarantee of honoring QoS requirements in the new cell due to resource constraint.

3. High signaling overhead: Whenever mobile moves to the new cell, it needs to send notification to Home Agent so that new incoming packets can be delivered to the appropriate subnet. This results in high signaling overhead.

The possible solution to the micro-mobility problems may lie in the following IETF proposals: (1) Mobile IP Route Optimization: Route Optimization attempts to reduce the "Triangle Routing" situation as proposed by [PJ99]. (2) Cellular IP: In Cellular IP, location information is stored in the distributed database. It uses gateway routers with mobile node entries for micro-mobility. The benefit of Cellular IP is that the lost in-flight packets problem can be eliminated [CGKTWY00], and (3) HAWAII: HAWAII is a domain-based approach for intra-system mobility manager [R99]

4.3 Mobile QoS Interpretation

In a wireless/mobile environment, available QoS may vary significantly over time. Widespread mobility (i.e., changes in the geographical density of new calls and handover operations) is making the concept of an initial traffic contact supported for the lifetime of a call almost meaningless.

Studies on UMTS services have focused on collecting QoS objectives mainly defined for fixed networks and circuit switched environments. Others have handled QoS mainly from a call blocking or handover blocking probability viewpoint. These QoS objectives are radically different from those used in fixed ATM, but still contribute to the overall QoS perceived by the user. The scope of an overall QoS service objective, and its implications on the underlying ATM transport is also unclear. Most data services today use upper layer protocols to ensure

error-free transmission. The QoS level required from the underlying transport should only ensure acceptable throughput and low residual error level for the overall service. This is quite different from assigning the overall QoS objective to the transport function alone.

It is evident that a different interpretation of QoS objectives is needed as compared to fixed networks. Three possible alternatives [PGS99] are presented in the following, regarding the nature of such service guarantees and taking into account the packet nature of future UMTS series.

1. The "raw QoS" approach: If no explicit QoS guarantees per service are provided, the traffic descriptor could be taken only an indication of the resources needed. System design and local operating conditions determined the actual QoS attained. This fact should be made known to a user by a network that provides, for instance, a special QoS class for mobile connection, effectively stating QoS may vary in times". This variance could also be provided to the user to decide upon accepting or rejecting the call.

2. The network provides average QoS guarantees per service and attempts to fulfill them in all cases, but with the exception of certain distinct occurrences: e.g. a user handing over to a different cell type or environment. This case could be covered by a special mobile subclass of each QoS class indicating that, due to radio constraints and/or other network conditions, QoS deterioration may be encountered. A certain threshold for accepting the service behavior could be decided by the user at call setup.

3. The "renegotiated QoS level" approach: The network provides flexible but positive guarantees for a given service and renegotiates the QoS requested to match the anticipated performance. A user (or the network itself) can then disconnect the call if the new QoS offered does not meet certain requirements. This case could be covered by a mobile subclass of each QoS class indicating a number of acceptable QoS levels as possible outcome of a QoS renegotiation process. The network initiates the renegotiations mechanism, each time QoS deviates from an initial (optimum) negotiated level and adapts QoS within a range of pre-decided acceptable levels.

4.4 Resource Allocation Policy

In fixed communication networks, resources need to be reserved only along a predetermined connection path. In addition, resource allocation is static for the lifetime of the call, since the QoS cannot be explicitly renegotiated by the user. For a mobile system however, the path of a connection changes dynamically. Resource allocation is probably the most time consuming function during handover. Furthermore, handover is the aspect of call handing that imposes the most processing load in the signaling network and the network switches. Therefore, the scope of a QoS guarantee and the allocation of resources need not be considered together.

An obvious approach would be to reserve or preallocate resources to a user (at call set-up time). That would only be needed/used after a handover. While this would have a positive impact on the QoS, access to preallocate resources may potentially be denied to other users, thus leading to inefficient use of network resources. The extent of preallocation can vary from the overall path, to just some limited area of the access network (e.g., all neighboring cells). It will still be difficult to meaningfully extend a QoS guarantee beyond the area where resources have been preallocate. One promising solution is to replace static pre-reservation with dynamic resource allocation (DRA). While DRA increases network utilization, the critical issue is the selection of renegotiations strategies (e.g., determining instants to renegotiate resources) and bandwidth estimation or prediction for the future reservation model (e.g., [SG98]).

4.5 Passive reservations

Mobility introduces the need to make reservations in advance between neighboring base stations (BSs) because a mobile may move into one of their regions. Making reservations just

before an MN moves into a cell may be disadvantageous because: (1) resources may not be available, and (2) data can be delayed/lost during the period it takes to setup a new reservation.

Passive reservations [MS98] may be used here to maintain reservations during mobility. Passive reservations are made from the BS of the cell where the MN currently resides to all neighboring BSs. The MN may move into any one of these cells and use the resource. This means the reservations made to some of the neighboring cells may not be used. To eliminate the wastage of resources that can occur, passive reservations can be used by other MN in the cells until the MN in question needs to use the reservation.

4.6 Billing

Billing is based on SLSs [RFC2745], which are either pre-configured or dynamically setup if a SLS negotiation protocol is used. In addition to the (usually) static SLS negotiation between home/foreign links and their ISPs, additional billing and accounting procedures must be provided for the case that an MN visits a foreign network and requests to use the SLSs of the foreign tunnel based virtual private network (VPN) with certain QoS support.

In general, the mobile IP node needs some mechanisms to indicate or signal to some BB in the foreign network that it desires a certain QoS. This could be supported by sophisticated protocols or signaling protocol extensions. The following alternatives could be a basis of such a solution:

1. A new special signaling protocol.
2. Special Mobile IP protocol options to be exchanged between MH and FA/first-hop-router/BB.
3. MNs could request reservations via RSVP. These reservations can be accepted or rejected by some local router dependent on whether the SLA of the foreign networks is sufficient or not. For this scenario, the concepts developed by the RSVP admission and police work group of IETF can be applied. A problem with RSVP might be that RSVP is able to support receiver-driven reservations only. Probably for sender-driven reservations, RSVP needs to be modified slightly.
4. Layer-4-Switching concepts or other signaling protocols: Another approach could be to avoid explicit signaling support for requesting a service and to try to identify flows, which shall get a DiffServ service. Those flows could be H.323 flows, HTTP flows or any other long-lived or high-volume traffic flows. Such flow identification functionality could be installed at a foreign agent or first-hop-router. The task of such a router is identify flows, assign DiffServ Codepoints (DSCPs) and request the establishment /modification of a SLS at the nearest BB or directly at the ISP. This concept is similar to the MF classification concept.

In any case, the foreign network needs some means to get the money back from the mobile node. Access and services could be pre-negotiated and paid in advance or after sending an invoice. An alternative could be that the mobile node carries some electronic cash (e-cash) and pays with this e-cash when requesting the desired service.

4.7 Network Provisioning in Mobile Environments

QoS can only be provided if the backbone networks of the ISPs are well designed and provisioned. However, network provisioning is a relatively complex task. Network provisioning especially becomes very difficult in highly dynamic environments, in particular in networks where the location and the QoS requirements of the end systems may change very quickly such as in mobile environment. Where as a stationary host knows at the beginning of its communication, this is not the case with mobile users. A mobile host/user might start with enough bandwidth and then move to a network that cannot provide enough capacity to fulfill its reservation request. In the case, the service of the mobile hosts or of other users in the visited network have to be degraded. Such a service degradation might make sense in the context of adaptive applications. As result, Premium Service [RFC2745] might be difficult to provide in a mobile environment.

One possible approach is to implement a DiffServ service to the mobile users according to the users' choices of priorities or applications' property, and then charge the users according to feedback preference information [ZM00]. This scheme is promising because it may adaptively provide the mobile users QoS services to the highest possible degree no matter how the dynamics of network provisioning are.

5. High Bit Error Problem

5.1 Unpredictable Bit Error Rate

Due to mobility, wireless applications cannot use a consistent data rate. Fading will affect the bit error rate (BER) of the channel and hence actual data rate. Data rates cannot be evenly distributed across the entire base-station coverage area. This is due to two reasons: (1) the mobile cannot achieve the setpoint bit-energy (E_b) at the target level demanded by base station, and (2) the service provider may opt for allowing the data rate to drop for users in the fringe areas to achieve better capacity. As a result of variable BER (due to fading), a mobile may be granted 144 kbps when it is close to the base station with small shadow fading. But if the user is in the fade zone or fringe of cell, the data rate will drop considerably.

This uneven data rate is a QoS problem: the mobile subscriber will experience inconsistent delays depending on where they are located relative to the base station antenna. To cope with the fading-related uneven data rate problem, some solutions are:

1. Downlink transmit power concentration: to allow maximum possible data rate to a particular user at a given time.
2. Better radio resource management: during the PPP/PCP negotiation phase, the number of QoS classes is conveyed in a particular link. Resources can be better allocated before the actual payload data is sent.

5.2 Distinction of the Causes of Packet Losses

TCP is one of the most important transmission protocols in the Internet and it works under an implicit assumption that all packet losses are due to congestion. Accordingly, TCP reduces its congestion window before retransmitting packets, and backing off its retransmission timer. However, this assumption is not accurate when a TCP connection traverses a wireless link that a significant fraction of packet losses may be due to transmission errors, handover and etc. This will unnecessarily result in severe throughput degradation and very high interactive delays when packets are lost for reasons other than congestion. TCP enhancements can be considered as a special but important aspect of Internet QoS. Because of the importance of congestion control in the wireless Internet, various attempts have been made to enhance TCP functions. Recently, along with the rapid development of satellite or mobile technologies, improving TCP's performance on wireless links has been becoming a challenge and an active research topic.

Among all the amendments to TCP, slow start, congestion avoidance, fast retransmission and fast recovery have been proven very efficient and mature, and well applied. Another bright example is the Explicit Congestion Notification. Fast-TCP is recently proposed, and its control philosophy is different from all the existed methods. However, some problems still exist even in these enhanced mechanisms. For example, most of them need some delay time (one to three RTT) to react the congestion and need support from TCP terminals. Regarding the TCP enhancements on the wireless platform, considerable efforts are devoted to determining the cause of packet drops and accordingly taking control actions. Generally, forward error correction (FEC), automatic repeat request (ARQ), indirect TCP (I-TCP), explicit loss notification (ELN), selective ACK (SACK), forward ACK (FACK), new ECN, and FR+ are some possible solutions in this category. Recognition accuracy, computation burden and modification to the TCP terminals are general problems in these approaches. For a comprehensive overview, please refer to [MDMMAV00].

6. Low Bandwidth Problem

6.1 Data Burst Control

Many IP-based applications exhibit traffic burstiness: application may lay dormant for a while and then send a burst of data as in the case of email and web browsing. Traditional circuit-switched approach allocates more bandwidth to users than would otherwise statistically multiplex system. Static allocation of bandwidth to data users does not use the precious air resource efficiently. Wireless packet data designers are motivated to produce high data rates for packet data applications.

In the newer wireless standards (3G and some improved 2G systems), a packet data burst mode is introduced to allow better interference management and capacity utilization [E97]. In the burst mode, the number of code channels that may be used by a mobile on the forward or reverse link for the duration of a burst is controlled by the infrastructure. Dynamic infrastructure-controlled burst allocation makes it possible to share the bandwidth efficiently among several high-speed packet data mobiles.

In the IS-95B packet mod, data rate can be adapted based on application needs and Carrier-to-Interference ratio. Each data session is assigned a fundamental channel and can be optionally assign additional 7 supplemental channels for the same application. Each channel operates at 9.6 kbps or 14.4 kbps. The data rate at IS-95B application can be ranging from 9.6 kbps (Rate Set I, one channel only) to 115.2 kbps (Rate Set II, all 8 channels). With this finer quantization of data rates, a better use of channel resources is achieved [NBK00].

6.2 RSVP Scalability

One of the major concerns with soft-state protocols like RSVP is scalability. Many RSVP flows across the wireless link can cause a lot of control information to flow across the wireless link. With wireless networks currently being low bandwidth systems, this is a point of concern.

RSVP refresh messages are periodically sent to detect any underlying network changes like route changes. Since here is only one wireless hop between a mobile and a BS, refresh messages need not be sent as often on the wireless link. This will help reduce the amount of control information on the link.

7. Computation Power Constraint Problem

7.1 Always-Connected

The wireline users are accustomed to power up their devices and initiate connection with networks. An example is the email application where users login only once after they come into the office. In addition, they are used to have NetMeeting where voice and data applications (such as file transfer or whiteboard) are in one shared medium.

Landline application habits will be extended to the wireless world where users do not want to constantly login to check their messages. On the other hand, to be always connected to the network means a separate PPP session needs to be dedicated to this application which is circuit – mode in nature and hence costly. An example of the solution is in IS-95B where mobiles are connected to the network all the time but there are two states: Active and Dormant states [TE1998]. In the Active State, a traffic channel is allocated to the user. Power control between base station and mobile is in place. In addition, one or multiple PPP session(s) are established between IWF (Inter Working Function) and end user. In the dormant state, PPP is in the open state between IWF and mobile. The traffic channel is not allocated to the mobile but IWF needs to have certain knowledge of the dormant mobile.

In the CDMA2000 (3G), two states are added to solve the problems of latency and high signaling overhead: Control hold state and suspended state are added to further accommodate granule signal overheads [K98].

7.2 Computation Migration

The challenges such as low bandwidth, low processing power and low memory capacity of the mobile devices will always exist, compared with the fixed computing devices, no matter how large they are and will be. It seems that no fundamentally efficient solutions so far existed for these kinds of challenges.

Besides the above challenges existing in mobile computing environment, there are some general limitations for any kinds of computer networks which are even not be aware by the human users. For instance, to gain access to the computational and informative world, we normally have to type on the keyboards or click with mice and learn artificial names for the people and resources we wish to access. These kinds of things are very trivial but troublesome. More terribly, the resources and services we are accessing may not be the most suitable ones for us.

Over the last few years, mobile agents have emerged as a powerful paradigm to deal with the situations. A mobile agent is a running program that autonomously decides to change location in order to continue its execution in an environment with better resources. The mobile agent technique will play a key role in the future mobile Internet infrastructure, by take the advantage of computation migration [ZZKM01] in order to save computation power of the mobile equipment.

7.3 Signaling Information

DiffServ architecture does not require end-to-end signaling and follows an implicit admission control mechanism. In wireless networks, a simple signaling scheme would be required and advantageous because: (1) static provisioning is not enough because user mobility necessitates dynamic allocation of resources, (2) the sender must know the limitations of the wireless link or better performance, and (3) information on local conditions like power status of the MN etc. need to be sent occasionally between the BS and the MN. A signaling protocol that can be used is a modify ICMP (Internet Control Message Protocol). The modified ICMP protocol is scalable and generates reduced control traffic when compared to RSVP.

Minimal use of the transceiver to send/receive packets is required to conserve power. All data packets need to be sent whereas we can control the amount of control information that is sent on the wireless link. For example all RSVP refresh messages can be sent only in one direction – from the mobile to the BS. This prevents the MN from the receiving unnecessary control messages. Also information like battery power level can be sent as part of these refresh messages.

7.4 Minimize the Interruption in QoS at the Time of Handover

At the time of handover, interruption in QoS would occur if the packets sent by or destined to the MN arrive at the intermediate node in the new end-to-end packet path without that node having information about their QoS forwarding requirement. Then, those packets will receive default forwarding treatment. Such QoS interruption must be minimized. A good metric [C01] for this performance is the number of packets that get served with "default" QoS at the time of handover. The number of such packets must be minimized.

When the care-of address changes upon handover, it may be required to perform some signaling even over the unchanged part of the end-to-end path if the path contains any QoS mechanisms that use IP address as a key to forwarding functions. Examples are filter specs in

the IntServ nodes or packet classifiers at the edges of DiffServ networks. However, double provisioning of resources over the unchanged part of the packet path MUST be avoided.

The QoS mechanism must provide some means (explicit or timer-based) to release any QoS state along the old packet path that is not required after handover. It is desirable that the unwarranted QoS states, if any, along the old path are released as quickly as possible at the time of handover. Note that, during handover, the MN may not always get a chance to send explicit tear down message along the old path because of the loss of link layer connectivity with the old access router.

7.5 Lightweight TCP/IP

Some small devices are often required to be physically small and inexpensive, and an implementation of the Internet protocols will have to deal with having limited computing resources and memory. Implementing a minimal TCP/IP stack is a very smart and efficient idea to solve this problem.

lwIP [D01] is a small implementation of the TCP/IP protocol stack suitable for systems with limited memory and CPU power. The focus of the lwIP TCP/IP implementation is to reduce the RAM usage while still having a full scale TCP. This makes lwIP suitable for use in embedded systems with tenths of kilobytes of free RAM and room for around 40 kilobytes of code ROM.

8. Tunneling Problems

As stated in section 2, in 3G networks, the GPRS core network uses IP tunnels, which makes the applicability of IP QoS schemes troublesome. Some attempts have been devoted to solve this problem. Normally, any effort to enhance the present GPRS QoS mechanism should focus on the GGSN, since this is the first node within the network where IP packets are interpreted.

In [P99], Puuskari proposed to maintain IP and GPRS QoS-related information within the GGSN. Interworking between GPRS and IP QoS schemes takes place in both the MS and GGSN. It is proposed that GPRS implement a QoS scheme general enough to cooperate with both IP QoS schemes, i.e., IntServ and DiffServ. The basic idea is to associate multiple QoS profiles to one PDP context. RSVP flows can be aggregated onto a few profiles, and each DiffServ traffic class can be associated with some particular QoS profile. A slightly different approach, proposed in the same article, is to allow several PDP contexts use the same IP address and to distinguish these contexts from each other based on information other than PDP addresses.

In [MT98], J. Mikkonen and M. Turunen identify the deficiency of IP multimedia (QoS) traffic handling in the GSM architecture. The authors assume the adoption of the IntServ framework for dealing with applications with specific QoS requirements. A dual protocol stack is proposed for the mobile terminal. The use of the GPRS infrastructure is suggested for the handling of controlled load and best effort. Guaranteed load traffic is dispatched by means of the high-speed circuit switched data infrastructure. In the same article, a mapping between GPRS radio link control (RLC) classes and IP QoS classes is suggested. The authors propose changes in the GPRS specification in order to expedite the handling of IP multimedia (QoS) traffic.

9. Conclusion

A lot of research has been done toward finding solutions for the landline QoS. As wireless technology matures and wider bandwidth spectrum is allocated to mobile users, wireless data customers will demand landline-like types of data services. This paper identifies major problems, challenges and requirements in providing QoS-enabled mobile applications and their corresponding candidate solutions. Some existing work is outlined as a survey, while some new ideas and proposals are presented from the research viewpoint.

The mobile IP promises a wide variety of applications to a wide spectrum of potential subscribers. On the other hand, there are many competing service providers offering newer and better services. Mobile IP enable QoS will be one of the ultimate criteria for successful services.

References

- [1] "Telephone network & ISDN QoS, network management and traffic engineering", *ITU-T Recommendation E.771*, Oct. 1992.
- [2] "Terms and definitions related to Quality of Service and network performance including dependability", *ITU-T Recommendation E.800*, Aug. 1994.
- [3] "TIA/EIA-95B, Mobile station – base station compatibility standard for dual-mode spread spectrum systems", October 31, 1998.
- [4] Blake, D. Black, D., Carlson, M., Davies, E., Wang, Z. and Weiss, W., " An architecture for differentiated services", *RFC2475*, December 1998.
- [5] Braden, R., Clark, D., and Shenker, S. (ed.), "Integrated services in the Internet architecture: and overview", *RFC 1633*, 1994.
- [6] Campbell A. T., Raymond, R., Liao, F. and Shobatake, Y., "Supporting QoS Controlled Handoff in Mobeware", *Proc. Winlab 3rd Gener. Wireleas Infor. Net.*, New Brunswick, USA, 1997.
- [7] Campbell, A. T., Gomez, J., Kim, S., Turanyi, Z., Wan, C-Y. and Valko, A., "Design, Implementation and Evaluation of Cellular IP", *IEEE Personal Communications*, June/July 2000.
- [8] Chaskar, H. and Koodli, "A framework for QoS support in Mobile IPv6", *Draft-chaskar-mobileip-qos-01.txt*, work in progress, March 2001.
- [9] Chaskar, H. (Editor), "Requirements of a QoS solution for Mobile IP", *Draft-ietf-mobileip-qos-requirements-00.txt*, work in progress, June 2001.
- [10] Dunkels A., "Minimal TCP/IP Implementation with Proxy Support". MSc Thesis. Technical report T2001:20. SICS, February 2001.
- [11] Ejzak, R., et. al., "Bali: A solution for high speed CDMA data", *Bell Lab Technical Journal*, Summer 1997.
- [12] Johnson, D. and Perkins, C., "Mobility support in IPv6", *draft-ietf-mobileip-ipv6-13.txt*, work in progress, March 2001.
- [13] Knisely, D., et. al., "Evolution of wireless data services: IS-95 to CDMA2000", *IEEE Communications Magazine*, October 1998.
- [14] Levine, D. A., Akyildiz, I. F. and Nagshineh, M., "A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept", *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 1-12, 1997.
- [15] Montenegro, G., Dawkins, S., Kojo, M., Magret, V., and Vaidya, N., "Long thin networks", *RFC2757*, January 2000.
- [16] Mahadevan, I. And Sivalingam, K. M., "An architecture for QoS guarantees and routing wireless/mobile networks", *Proc. ACM Intl. Workshop on Wireless Mobile Multimedia (WOWMoM)*, pp. 11-20, Dallas TX, Oct. 1998.
- [17] Mikkonen, J. and Turunen, M., "An integrated QoS architecture for GSM networks", *Proc. ICUPC'98*, Florence, Italy, 1998.
- [18] Nandaet, S., Balachandran, K. and Kuma, S., "Adaptation techniques in wireless packet data services", *IEEE Communications Magazine*, January 2000.
- [19] Ni, M. and Xiao, X., "Internet QoS: A Big Picture," *IEEE Network*, Vol. 13, No. 2, pp. 8-18, Mar.-Apr. 1999.
- [20] Nichols. K., Blake, S. Baker, F and Black, D., "Definition of the differentiated services field (DS field) in the IPv4 and IPv6 headers", *RFC2474*, December1998.
- [21] Perkins, C. (Editor), "IP mobility support", *RFC 2002*, October 1996.
- [22] Perkins, C., "IP encapsulation with IP", *RFC 2003*, October 1996.
- [23] Perkins, C. and Jonhson, D. D., "Route optimization in mobile IP", *draft-ietf-mobileip-optim-10.txt*, Work in Progress, November 2000.

- [24] Philippopoulos, P.I., Georgopoulos, C.E. and Sykas, E.D., "QoS interpretation in 3rd generation wireless/mobile systems", *Proc.1999 IEEE 49th Vehicular Technology Conference*, Vol. 3, pp. 2059 –2063, 1999.
- [25] Puuskari, M., "Quality of service framework in GPRS and evolution towards UMTS", *Proc. 3rd Euro Mobile Commun. Conf.*, Paris, France, 1999
- [26] Ramjee, R., et. al., "IP micro-mobility support through HAWAII", *Internet Draft*, work in progress, October 1999.
- [27] Rosen, E., Viswanathan, A. and Callon, R., "Multiprotocol Label Switching Architecture" *RFC 3031*, January 2001
- [28] Singh, S., "Quality of Service guarantees in mobile computing", *Computer Communications*, vol. 19, pp. 359-371, April 1996.
- [29] Stemm, M. and Katz, R. H., "Measuring and reducing energy consumption of network interfaces in hand-held devices" *IEICE Trans. on Fundamentals of Electronics, Communications, and Computer Science*, Aug. 1997.
- [30] Su, W. and Geria, M., "Bandwidth allocation strategies for wireless ATM networks using predicative reservation", *Proc. IEEE Globecom*, Sydney, November 1998.
- [31] Toh, C., "Mobile QoS for wireless ATM networks: An adaptive approach", *Proc. ACTS Mobile Summit*, Aalborg, Denmark, Oct. 1997.
- [32] Wroclawski, J., "The use of RSVP with IETF integrated services", *RFC2210*, Sept. 1997.
- [33] Zhang, R. and Ma, J., "On the enhancement to a differentiated services scheme", *Proc. IEEE/IFIP NOMS'2000*, 2000.
- [34] Zhang, R., Zhang, D., Kan, Z. and Ma, J., "Computation Migration based on Mobile IP and Intelligent Agent Techniques", *Proc. ICII'2001*, vol. C, pp. 251-255, November 2001, Beijing, China