QoS INTERPRETATION IN 3^{RD} GENERATION WIRELESS/MOBILE SYSTEMS

P. I. Philippopoulos, C. E. Georgopoulos, E. D. Sykas

Department of Electrical Engineering and Computer Science
National Technical University of Athens (NTUA)
9 Heroon Polytechniou Str., Zographou
157 73 Athens, GREECE
PH: +30-1-772-1480 FAX: +30-1-772-2534
e_mail: cgeorg@telecom.ntua.gr

Abstract - UMTS, viewed as the wireless access part of a backbone ATM-based broadband services fixed network, serves the goal for future integrated communications. One of the major issues for the deployment of UMTS high-bandwidth and real-time multimedia applications, is the specification of acceptable Quality of Service (QoS) requirements for mobile users. In wired ATM networks a traffic contract guarantees the connection the network should provide to a fixed user. Due to the relatively unpredictable nature of user roaming into a non-homogenous system with different wireless cell structures like UMTS, the concept of an initial traffic contract may become meaningless. This paper addresses these issues and proposes a generic framework for the interpretation, and adaptation of a mobile QoS for 3rd generation wireless/mobile systems.

I. INTRODUCTION

The term "Quality of Service" designates a set of parameters, intended to represent measurable aspects of the subjective "user perceived quality". Criteria taken into account ([2]) involve concepts such as service availability, retainability and integrity, transmission characteristics, as well as subjective estimates.

The efficiency of a network in providing services to users includes many factors influencing QoS. In ATM networks, a key characteristic is their ability to provide statistical multiplexing between different users. As the user/data sources exhibit statistically definedbehaviour, there is a certain probability that all active users cannot be serviced, at a given instant. Due to this, ATM services, even in a network with no faults, are not defined in absolute terms, but in terms of *service objectives*. The traffic generated by a user must be defined using a *traffic descriptor*. To ensure proper user data behaviour, the network monitors the received data stream in real time (UPC function) and provides an agreed QoS level for the user-traffic that qualifies. The traffic description and the requested *QoS Class* form part of a *traffic contract*, i.e. a definition of the service obligation that the network has towards a user. To further ensure that the traffic contract is met, the network measures the actual QoS achieved, with *in-service* and *out-of-service* measurements.

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), and changing network charactaristics (e.g. cells providing different maximum QoS levels), are making the concept of an initial traffic contract supported for the lifetime of a call almost meaningless. This paper addresses these issues and proposes a generic framework for interpreting and maintaining a mobile QoS. It should be noted that QoS aspects are still under study even for fixed networks.

II. MOBILE QoS INTERPRETATION

ITU-T, in recommendation E.771 ([1]), established the criterion that the quality of a mobile call for a given service should be that of a fixed call plus an additional switching stage. Recommendation E.800 ([2]) 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.*

Studies on UMTS services carried out in various projects (RACE II MONET, MBS), have focused on collecting QoS objectives mainly defined for fixed networks and circuit switched enviroments. 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

0-7803-5565-2/99/$10.00 © 1999 IEEE

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 are presented in the following, regarding the nature of such service guarantees and taking into account the packet nature of future UMTS services.

The 'Raw QoS' Approach

If *no explicit QoS guarantees* per service are provided, the traffic descriptor could be taken only as an indication of the resources needed. System design and local operating conditions determine 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 connections*, effectively stating 'QoS may vary in times'. An estimate of this variance could also be provided to the user to decide upon accepting or rejecting the call. A possible QoS vs time graph for this case is shown in*figure 1.*

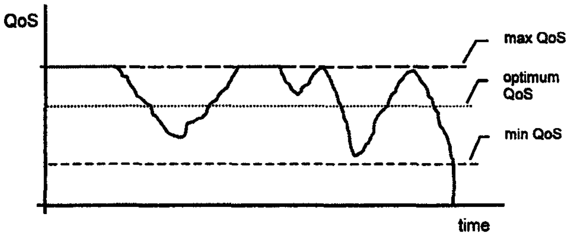

Figure 1: "Raw QoS" Approach

The *max QoS* and *min QoS* lines indicate a fixed operability range, depending on network design, for the specific service. The *optimum QoS* line represents an adequate behaviour of the specific service as perceived by the average user. This case is similar to the way 2d generation mobile systems have handled QoS in their early stages. It could be argued, that an appropriately fast and dimensioned network with an abundance of resources, would always provide a "raw QoS" equal to the optimum one, or even better. However, this is only an evolution far-end. Assuming that the radio parts of the system can adapt to many different operating conditions, and considering multimedia services with stringent time/loss restrictions envisaged for 3d generation systems, some more advanced interpretations would be required in UMTS/MBS.

The 'Limited Impairments QoS' Approach

The network provides *average QoS guarantees* per service and attemps 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 behaviour could be decided by the user at call setup. *Figure 2* illustrates a possible QoS vs time graph.

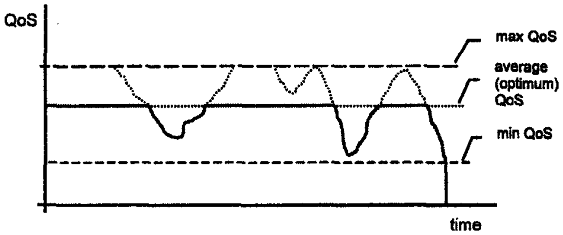

Figure 2: "Limited Impairments QoS" Approach

Constant (or nearly constant) QoS is maintained most of the time, except when physical conditions or increased operation complexity make it impossible. The user can then sign up for a QoS lower than the maximum possible, but a more realistic one and still close to the optimum. By lowering the max. available QoS to an average level (decided by statistics and current operating conditions) the network has a gain in resources, which can be utilized by other users, thus enhancing overall QoS. Resource management should assess that in real time. The key element in this approach is the minimisation of abnormal situations. The capability of the network to treat them as 'isolated' events depends on system planning and the efficiency of the resource management scheme applied.

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 renegotiation mechanism, each time QoS deviates from an initial (optimum) negotiated level and adapts QoS within a range of pre-decided acceptable levels.*Figure 3* illustrates a possible QoS vs time graph for this approach.

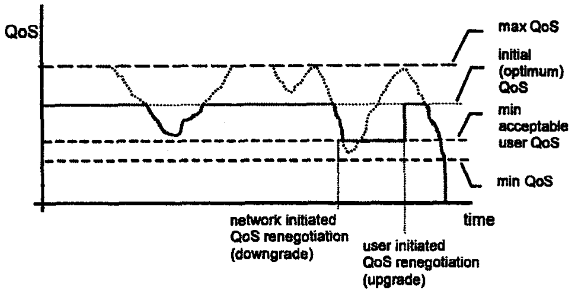

Figure 3: "Renegotiated QoS Level" Approach

When a pre-decided lower threshold (*min user QoS*) is crossed, resource management should be able to stabilise QoS at this level, or else notify the user. In addition, the user could authorize the network to upgrade QoS automatically whenever conditions improve, or decide this him/herself upon network notification. The current definition of ATM QoS classes ([3], [4]) does not provide for renegotiation of the QoS class, but this is quite likely to happen in future versions. The efficiency of the renegotiation process and the limits set to QoS adaptation flexibility, are the decisive features in this approach.

III. MOBILE QoS SPECIFICATION

Assuming a certain interpretation of the QoS objectives, a method must be established for specifying QoS.In the following a definition of a mobile QoS is outlined as an example to highlight and refine on all relative issues.

Mobile QoS Reference Point

In [2], 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* for defining 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.

Mobile QoS Components

A connection involving at least one mobileuser, can be viewed as the concatenation ([5]) of fixed and wireless links. A mobile QoS (M-QoS) therefore, comprises of:

- ♦ a fixed network component (F-QoS), relating to QoS objectives for the wireline (ATM) links.
- • 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 (e.g. 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.

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.

Regarding HO-QoS, primary metrics would be the *blocking probability due to HO* and the *new call blocking probability*. A metric that can be interpreted in terms of *HO request priority* would also be useful. The node level at which HO mechanisms are applied, the controlling scheme employed (e.g. network controlled HO) and the use of special resources (e.g. for 'soft' HOs), may also be included in this objective. Measures of the overall *delay for HO accomplishment* and *time between HOs* are also indicative of the actual impact of HO on applications. Traffic disruption incurred by HO could be translated into *Cell Loss/Cell Sequence Integrity* and *Delay/Delay jitter due to handover*. Priorities concerning the degradation of specific connections' QoS within a call can also be introduced.

Regarding AIF-QOS, the way QoS fluctuation occurs and the metrics appropriate to measure it, depend largely on the specific air interface employed. Parameters that may apply to a generic 3rd generation air interface include: BER, Signal Strength, Frame Error Rate, Packet Delay & Delay jitter, Mean Access Delay, Quality Change probability, Link Loss probability. Besides transmission characteristics, *cell profile* QoS objectives incorporating OAM-status related information may be input to AIF-QoS.

Resource Allocation Policy

In fixed ATM 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 explicitlyrenegotiated by the user. For a mobile system however, the path of a connection changes dynamicaly. Resource allocation is probably the most time consuming function during HO. Furthermore, HO is the aspect of call handling that imposes the most processing load in the signalling network and the network switches. Therefore, the scope of a *QoS guarantee* and the *allocation of resources* need to 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 HO. While this would have a positive impact on the QoS, access to preallocated 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 call path, to just some limited area of the access network (e.g. all neighbouring cells). It will still be difficult to meaningfully extend a QoS guarantee beyond the area where resources have been preallocated. 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 ofrenegotiation strategies (e.g. determining instants to renegotiate resources) and bandwidth estimation or prediction for the future reservasion model (e.g. [6]).

IV. NEGOTIATION & MAINTENANCE

A mobile reference configuration ([7]) assuming an ATM capable end-user device connected to a mobile termination is shown in *figure 4*.

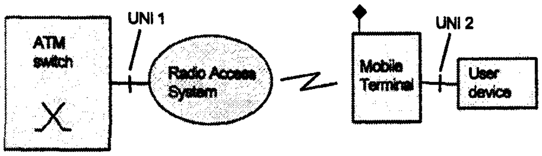

Figure 4: Mobile reference configuration

UNI 1 is the ATM UNI as seen by the ATM switch, while UNI 2 is the user-device viewpoint for ATM. To provide meaningful QoS quarantees, the differences in QoS between the two UNIs must be accounted for. For mobile-to-mobile calls, the effect of both radio parts should be considered. As QoS information should not be processed separately at each end of the connection, a generic *negotiation mechanism*, applicable also to fixed ATM nodes is required. The QoS that can be achieved for a mobile-to-fixed call, could be substantially different from a mobile-to-mobile call. User devices cannot assume that a QoS requested, is automatically provided if the call is accepted - a call could be accepted with significantly reduced QoS objectives. Modifications to the mobile end of the call may apply to the fixed party as well (for mobile-to-fixed calls). Furthermore, to achieve integration with a fixed ATM-based backbone infrastructure, the QoS mechanisms, both for negotiation (using signalling)and maintenance (using transport mechanisms) must be *transparent* to the fixed user.

Renegotiation of QoS would add significantamounts of signalling over the radio interface and the fixed network. A certain degree of relaxation could be achieved by *limiting the renegotiation range* to a set of predefined degrading/upgrading policies, or *restricting the renegotiation process* (which is normally end-to-end) between the mobile terminal and some appropriate point in the access network (e.g. the base station).

To maintain full control of the radio aspects, it is probably preferable to allow only the network to make adjustments. Furthermore, the QoS provided by the radio path, must be (reasonably) stable and predictable. Therefore, an *adaptive QoS-capable radio interface* is required. To protect the invaluable radio resources, a functionality similar to the UPC function should be implemented before the air interface. This function would reside in the UMTS terminal, provided that the terminal can be trusted.

Special considerations also apply for *monitoring QoS*. Since the performance of the radio part is constantly measured anyway, QoS related information for the radio part should be available. On the other hand, the methods typically used to measure ATM performance using a separate connection and sending OAM cells in the user data stream, are not well suited for a heavily loaded wireless network, due to the extra bandwidth required. In lightly loaded networks, the capacity problem is not intense, but the information obtained is not very interesting. Therefore, measuring and estimating QoS during operations could preferably be done separately for the air interface and the fixed network. This is well reflected in the refinement of M-QoS into separate fixed and wireless components

The activity of the actual network QoS control function, can be categorised as follows :

- ❑ **Sustain M-QoS:** If the new status of the radio link and the fixed network path conform to the M-QoS requirements, operations continue with the current parameters.
- ❑ **Degrade M-QoS:** If the new path (fixed and wireless part) indicated e.g. during HO, supports a lower QoS, or if new operating conditions make it

impossible to sustain the required M-QoS, the network performs an appropriate degradation scheme (e.g. local or overall M-QoS refresh), according to the extent and the significance of the cause that effected the degradation. Upon completion, all participating parties are informed about the new M-QoS, and start their own adaptation, if authorised. It would be worth to be able to renegotiate the QoS parameters of the respective existing VCs in the core network, but even if this is infeasible, then a re-establishment of certain parts of the path might be necessary. It is not always clear whether re-opening VCs would be wiser than keeping the initial ones with the original parameters, even after a QoS degradation at the access points.

□ **Upgrade M-QoS:** A possible availability of abundant resources in the new path, or improved operating conditions, could be exploited by upgrading the current M-QoS in a pre-decided manner. However, if the potential upgrade is insignificant, or if the subscriber is constantly moving and just temporarily transiting through a relatively «free» cell, then upgrading would not be justified, in terms of signalling overhead versus expected gain in quality

After a number of adaptations, M-QoS parameters may need to be *refreshed*, to enable an application to re-request its initial QoS guarantees. Otherwise, the QoS the network applies after each adaptation may only be degrading -especially if the network decides that an upgrade would not be worthy or long term, due to e.g. frequent HOs. The above renegotiation mechanism is applicable in the "renegotiated QoS level" approach.

Applications Aspects

Most existing network architectures are geared towards *hiding QoS fluctuations* from applications. A number of techniques (e.g. partitioning, compression, caching) are also employed for *reducing application demands*. An alternative approach, would be one where applications *monitor the network status and adapt* to fluctuating conditions ([8]). In the context of the adaptation approach, an application would need to specify a range of operability via e.g. a flow control mechanism. The application should also indicate a policy for scaling its demands up or down. An efficient service policy should distinguish between real-time and non-real-time oriented services. Considering their special characteristics, static priorities can be introduced between service classes (e.g. CBR>VBR>ABR>UBR). As UMTS design/implementation will not be geared towards supporting a specific set of services, the introduction of a Service Adaptation Layer ([9]) may offer the required transport capabilities to support the whole range of 3^{rd} generation mobile services. It is essential that mobility aspects and functions, such as handover, be hidden from the SAL.

V. CONCLUSIONS

The primary and most important measure of service quality should be customer satisfaction. It is quite reasonable to assume, that the requirements on the QoS for a certain service provided by a network, determine what the end-user may expect in statistical terms. Factors that influence user acceptance of the behavior of a service (e.g. users tend to accept isolated events more easily than repeated ones) should be carefully examined against complex design strategies.

Ínterpreting QoS is not only important from a technical viewpoint, it will also have significant impact on charging practices. If meaningful QoS parameters are used, the user should be charged at the full rate only when the requested QoS behaviour can be maintained. QoS renegotiation would certainly involve adaptive billing policies based on content and not only on time.

In tommorow's mass market of telecommunications, 3^{rd} generation systems services that cannot be well justified in terms of efficiency versus complexity and relevant cost, stand little chances for surviving competition.

REFERENCES

[1] ITU-T Recommendation E.771, "Telephone Network & ISDN QoS, network management and traffic engineering", October 1992.

[2] ITU-T Recommendation E.800, "Terms and Definitions related to Quality of Service and Network Performance including Dependability", August 1994.

[3] The ATM Forum Technical Committee, "ATM UNI Signalling Specification V3.1", September 1994.

[4] The ATM Forum Technical Committee, "ATM UNI Signalling Specification V4.0", July 1996.

[5] C-K Toh, "Mobile QoS for Wireless ATM Networks: An Adaptive Approach", *Proceedings of the ACTS Mobile Summit,* Aalborg, Denmark, October 1997.

[6] William Su, Mario Gerla, "Bandwidth Allocation Strategies for Wireless ATM Networks using Predictive Reservation", *Proceedings of the IEEE Globecom Conference*, Sydney, November 1998.

[7] RACE MoNet report, "UMTS System Structure Document", *CEC deliverable number R2066/BT/PM2/DS/P/113/b1,* 31/12/1995.

[8] Campbell A.T., Coulson G., Hutchison D., "A Multimedia Enhanced Transport System (METS): a Contribution to Discussion on ECFF", *ISO/IEC JTC1/SC6/WG4 N832, International Standards Organisation, UK,* December 1993.

[9] Saidi A. et al., "RAINBOW demonstrator transport chain", *Proceedings of the ACTS Mobile Summit,* Aalborg, Denmark, October 1997.