LOST?

S. Baudet, C. Besset-Bathias, P. Frêne, N. Giroux

Quality of service management is a key feature to support various applications in UMTS networks.

QoS implementation in UMTS networks

Introduction

The second generation Global System for Mobile Communication (GSM) was designed primarily to provide circuit-switched speech and data services, as well as some packet-oriented teleservices, such as the Short Message Service (SMS). As far as the core network was concerned, Quality of Service (QoS) relied primarily on the quality of the fixed circuit-switched network. The task of the Base Station Subsystem (BSS) was to allocate circuit-switched radio traffic channels and preserve established calls by implementing reliable handover procedures for users moving between cells. Data services remained marginal in terms of the generated traffic, hence QoS was mainly envisaged in terms of the busy hour blocking probability, call drop rate and subjective voice quality. Frequency planning, transport resource and radio dimensioning were key issues in providing the appropriate radio coverage and traffic capacity.

Recent developments, like Adaptive MultiRate (AMR) speech codecs and Tandem Free Operation (TFO), are aimed at improving QoS for circuit-switched speech. Following development of the Internet, the General Packet Radio Service (GPRS) was introduced to offer efficient end-to-end wireless packet data services in GSM networks. Despite the efforts of the standardization bodies to refine QoS handling, taking into account the needs for application and subscriber differentiation, today's GPRS systems are generally considered as "best effort". However, relative QoS levels can be provided under the control of the core network, as demonstrated by Alcatel's implementation of the Serving GPRS Support Node (SGSN). Third generation systems will considerably improve this situation for all types of service, including speech, data and multimedia, since QoS handling is one of the underlying concepts of the system specifications drawn up by the Third Generation Partnership Project (3GPP).

Why is QoS Needed in UMTS?

Second generation mobile systems were developed for voice transport at a time when 75% of traffic in Europe was voice. Consequently, they were optimized to preserve voice isochronism and to offer telephone services, including Intelligent Network (IN) services. The impact of the Internet has led to a dramatic growth in data traffic, making it necessary to deploy a mobile network capable of transporting both Internet and voice traffic. The main reasons for this integration are to minimize the operating and maintenance costs of mobile networks, to optimize bandwidth usage, to support new multimedia applications and to take advantage of improved coding schemes and compression algorithms. Third generation mobile networks have been designed to transmit packet- and circuit-switched applications on the same medium, be it radio or terrestrial.

An important feature of the Universal Mobile Telecommunications System (UMTS) is that information generated by independent sources can be efficiently multiplexed on the same transmission medium. UMTS supports traffic with very different bandwidth and QoS requirements. Traffic generated by data transfer services and Internet access is essentially bursty and unpredictable. Although data transmission between machines is loss sensitive, it is usually not sensitive to end-to-end delay or jitter. On the other hand, speech (and, more generally, real-time applications) requires strict limits on the transmission delay, but can cope with reasonable loss rates. For example, end-to-end delay for voice must be less than 400 ms, even with echo cancellation.

A major challenge for the UMTS infrastructure is to carry various types of application on the same medium, while meeting the QoS objectives. As well as meeting the needs of the user who is only interested in the end-to-end QoS perceived at application level, it is essential that the system uses the transmission resources efficiently. This requirement applies not only to the scarce radio spectrum, but also to terrestrial transmission resources, and especially the access part which must provide a cost-effective transfer service while minimizing investment and operating costs. Thus it is highly desirable to achieve some statistical multiplexing gain. In particular, transmission links and the radio interface must be loaded as heavily as possible while meeting the QoS requirements. Therefore it is important to identify mechanisms that optimize the load.

Tab. 1 UMTS traffic classes

QoS Class	Transfer delay requirement	Transfer delay variation	Low bit error rate	Guaranteed bit rate	Example
Conversational	Stringent	Stringent	No	Yes	VoIP, Video-conferencing, Audio-conferencing
Streaming	Constrained	Constrained	No	Yes	Broadcast services (audio, video), News, Sport
Interactive	Looser	No	Yes	No	Web browsing, Interactive Chat, Games, m-commerce
Background	No	No	Yes	No	E-mail, SMS, database downloads, transfer of measurements

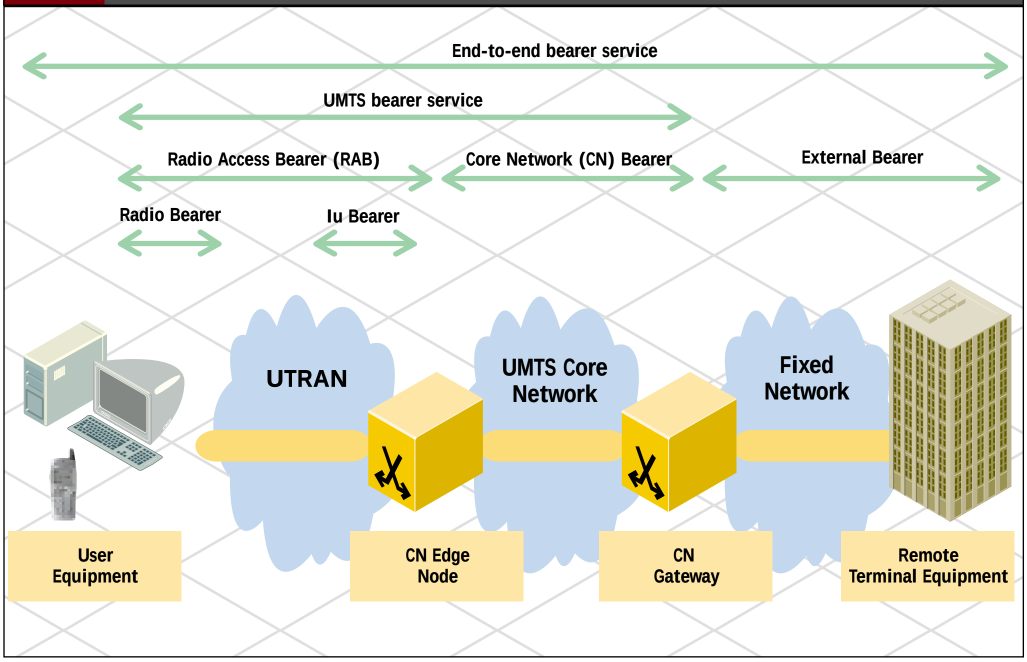
To meet these requirements, 3GPP has defined four QoS classes (TS 23.107): conversational, streaming, interactive and background. Typical applications and a summary of the QoS requirements are given in *Table 1*.
The UMTS QoS architecture defined by 3GPP relies on the bearer services characterized by QoS attributes, as illustrated in *Figure 1*. From the UMTS Terrestrial Radio Access Network (UTRAN) standpoint, this means offering Radio Access Bearer (RAB) services between the user equipment and the core network. A RAB relies on two other types of bearer service, which are provided by the UTRAN on its external interfaces:

- Radio Bearer (RB) services, between the user equipment and UTRAN.
- Iu bearer services on the Iu interfaces between the UTRAN and the core network.

The core network provides a core network bearer service between the UTRAN and external fixed networks, such as the Public Switched Telephone Network (PSTN) or Internet.
Each bearer service is defined by its QoS attributes. For example, the UMTS bearer service attributes are shown in *Table 2*, with their relevance for each traffic class (itself one of the attributes).
UMTS applications depend on the deployment of advanced end-to-end QoS mechanisms:

- On the radio side, techniques such as power control and radio admission control take the QoS profiles (set of UMTS bearer service attributes) into account.
- At network level, efficient integration is performed at all layers, including layer 2 and layer 3, and in every crossed multiplexing point, that is, in every UMTS node and backbone infrastructure. Asynchronous transfer over today's packet networks loses the timing structure of the flow and introduces random delay and delay variations. Large buffers are needed in routers to avoid congestion at burst level. However, the use of large buffers in UMTS nodes can introduce unacceptable delays for real-time applications. This leads to more complex systems than those used by first generation routers with their single First In, First Out (FIFO) queues and blind multiplexing.
- QoS mechanisms, such as bearer control, are also needed at the application level.

Fig. 1 UMTS QoS architecture



QoS Handling in the UMTS Core Network

There are three basic timescales for traffic management in UMTS networks:

- Capacity planning and network dimensioning determine the numbers and configuration of the Mobile Switch-

Tab. 2 UMTS bearer service QoS attributes				
Traffic class	**Conversational**	**Streaming**	**Interactive**	**Background**
Maximum bit rate	X	X	X	X
Delivery order	X	X	X	X
Maximum Service Data Unit (SDU) size	X	X	X	X
SDU format information	X	X		
SDU error ratio	X	X	X	X
Residual bit error ratio	X	X	X	X
Delivery of erroneous SDUs	X	X	X	X
Transfer delay	X	X		
Guaranteed bit rate	X	X		
Traffic handling priority			X	
Allocation/ retention priority	X	X	X	X

ing Centers (MSC), SGSN, Gateway GPRS Support Node (GGSN), etc, as well as the required bandwidth of the UMTS interfaces. This process largely depends on the operator's traffic mix, service level agreements and network topology (number of nodes, number of sites, etc). The impact on and changes to the network infrastructure typically occur on a daily or monthly basis.
- Each time a call is setup, Call Admission Control (CAC) determines whether or not the new call can be accepted while guaranteeing the QoS of established calls. The CAC function computes and allocates an equivalent bandwidth for the duration of the call, which is typically minutes for voice calls or hours for Internet Protocol (IP) sessions.
- Policing, scheduling and congestion mechanisms are performed each time a packet is sent and/or received (typically microseconds). Policing and buffer acceptance mechanisms are algorithms that decide when a packet arrives, whether or not it can be accepted. Scheduling algorithms decide when and which packet to send first.

The QoS features supported by the UMTS core network determine its ability to differentiate between services offered to subscribers. This capability is offered end-to-end to ensure that the necessary resources are allocated to provide an adequate service to a subscriber while preserving fairness to other subscribers and guaranteeing the negotiated quality of service. *Figure 2* sums up the end-to-end Alcatel QoS approach for the core network:

- *At the application layer*, CAC applies in every multiplexing point. Isolation of traffic between UMTS subscribers is offered within the GGSN and SGSN by means of Weighted Fair Queuing (WFQ) as well as enhanced congestion control (Weighted Fair Buffer Allocation). Policing is also performed at the GGSN.
- *At layer 3*, classical IP differentiated services are used.
- *At layer 2*, Multi Protocol Label Switching (MPLS) can be used on the Gn and Gi interfaces, while the Asynchronous Transfer Mode (ATM) is used on the Iu interface.
- QoS provisioning is also performed via the network and service management platforms.

QoS mechanisms specific to UMTS are mainly provided by the core network bearer service. Backbone bearer services rely as much as possible on the existing QoS mechanisms offered by the IP Differentiated Services (DiffServ) and MPLS/ATM architectures.

UMTS Call Admission Control

Call admission control is performed in every multiplexing point involved (MSC, SGSN, GGSN, border gateway, media gateway, etc). CAC answers two questions. Can the node accept this new call? Will the node be able to meet the QoS requirements of the new call and of the established calls?

UMTS CAC is implemented using the simple and flexible concept of equivalent bandwidth. The philosophy is to estimate the network resources required to provide the requested QoS and to determine whether these resources are available. If they are available, the necessary resources are reserved. However, if they are not available, a mechanism is initiated to downgrade the QoS. For example, the user could be offered a smaller guaranteed bit rate or lower traffic handling priority.

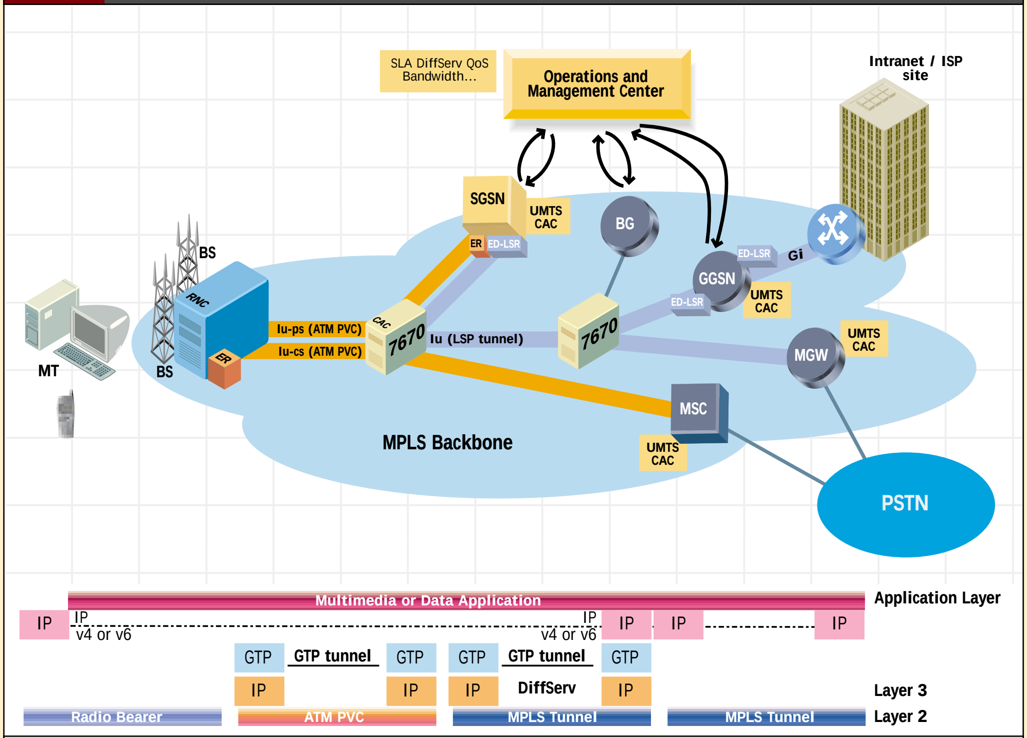
CAC is performed at every Packet Data Protocol (PDP) context activation or circuit call setup, every time the QoS is modified, and in the event of handover or Serving Radio Network Subsystem (SRNS) relocation.

Moreover, to guarantee the QoS at all levels – not only at the application level – CAC takes into account the resources available at the IP transport layer (for instance, the provisioned bandwidth of every DiffServ class). CAC is also performed at layer 2 (ATM and/or MPLS).

QoS Differentiation between User Equipment

Several mechanisms are available to isolate traffic coming from every UMTS user equipment: queuing differentiation and weighted fair buffer allocation within the SGSN and GGSN; policing and traffic flow templates within the GGSN.

Fig. 2 Alcatel QoS approach



BG: Border Gateway
ER: Edge router
LSP: Label Switched Path
MT: Mobile Terminal
BS: Base Station
GTP: GPRS Tunneling Protocol
LSR: Label Switched Route
ED: Edge
ISP: Internet Service Provider
MGW: Media Gateway

GGSN policing
The GGSN is the edge node in the UMTS network, so it receives the incoming traffic and checks that the downlink user data traffic conforms with the QoS attributes of the corresponding UMTS bearer service. The GGSN includes a traffic conditioner, which guarantees traffic conformance. Another role of the GGSN is to filter the incoming traffic according to the Traffic Flow Templates (TFT). These are used to distinguish between different user packets going to the same PDP address which have different QoS requirements identified by different PDP contexts.

Queuing differentiation
Scheduling techniques, such as WFQ or Weighted Round Robin (WRR), are used to resolve any contentious access to a resource. These techniques guarantee a minimum allocated bandwidth for every PDP context. Fair queuing mechanisms also provide an “implicit policing” function since bandwidth is allocated in proportion to the weights. It is known as implicit policing as it only takes any action when a network element is congested, always dividing the bandwidth up in a fair manner. Implicit policing is mandatory to support a QoS commitment: user behavior has to be constrained to that specified in the contract if resources are short so that a misbehaving or malicious user does not adversely affect the QoS delivered to other users.

Weighted fair buffer allocation
Weighted Fair Buffer Allocation (WFBA), also known as Weighted Random Early Discard (WRED), is used to discard packets in an intelligent and proactive way. It is particularly useful in allowing the Transmission Control Protocol (TCP) to react quickly and efficiently to congestion. The basic principles are as follows. The buffer is divided fairly among the active user equipment. The fraction of the buffer allocated to a user equipment depends on its UMTS bearer service QoS attributes, as defined in *Table 2*. Buffer overbooking per PDP context is allowed, depending on the buffer occupancy. A new packet is accepted only if the user equipment is not already using too much space in the buffer.

Use of IP Differentiated Services at Layer 3

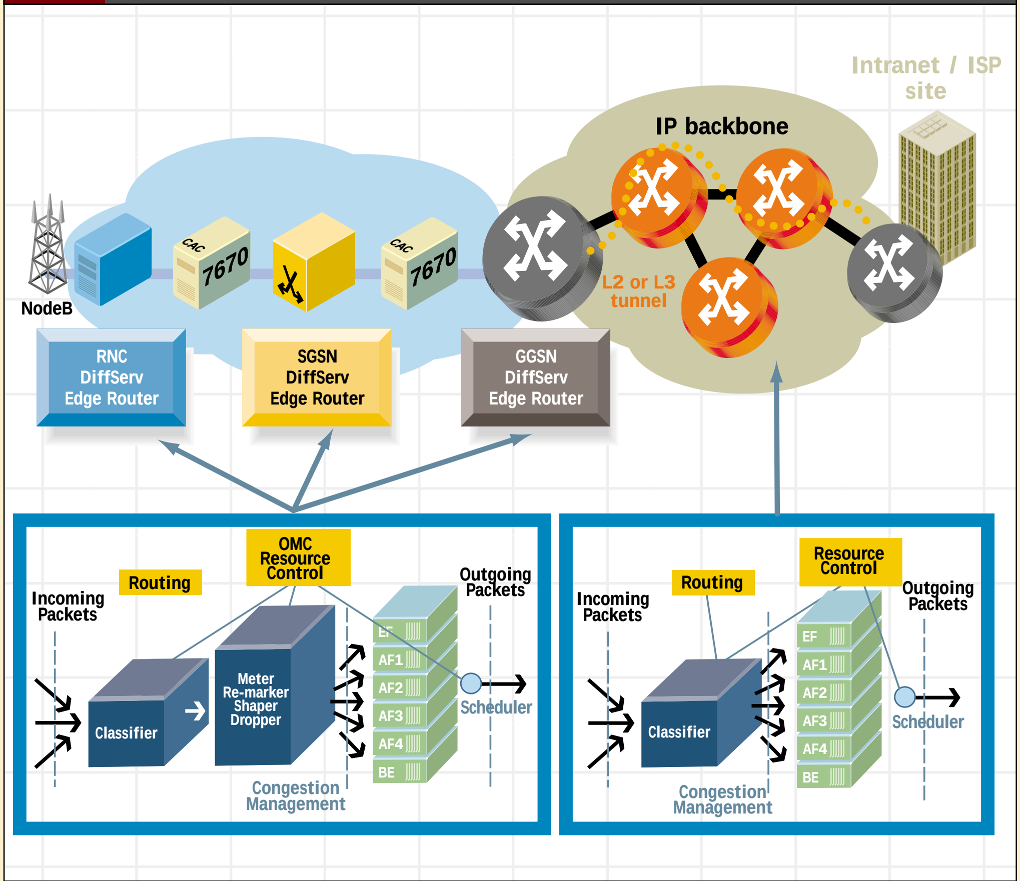
QoS differentiation is necessary to ensure that UMTS traffic is suitably handled within the network. On the UMTS backbone, Internet Engineering Task Force (IETF) Differentiated Services (DiffServ or DS) are used at layer 3, the IP transport layer.
Basically, the IP DiffServ octet contains a Differentiated Services Code Point (DSCP) to identify and select the particular Per-Hop Behavior (PHB) that an IP datagram will receive at a given network node. A DSCP is set in the IP header of each packet.
The main advantages of DiffServ are as follows:

- DiffServ does not need extra signaling at the IP level. All necessary QoS information is already contained in UMTS specific signaling messages (Radio Access Network Application Part, GPRS Tunneling Protocol Control Plane, etc) exchanged within the Public Land Mobile Network (PLMN).
- Network resource allocation is provisioned per aggregated flow (DiffServ class). All the complexity is located at the edge routers.
- DiffServ is largely deployed in IP core networks. It facilitates interoperability of the UMTS PLMN with other networks (e.g. inter-PLMN backbone, Internet service provider, intranet/extranet).

Differentiated services architecture
The Radio Network Controller (RNC), SGSN, GGSN and border gateways act as DiffServ edge routers. As such, they provide classification, aggregation, traffic conditioning, weighted scheduling and priority dropping. DSCP field (re-)marking

is based on data configured via management ("Provider marking") or according to a traffic management policy ("Policer marking"). The meter enforces the flow traffic contract. The shaper delays packets in such a way that they comply with the traffic contract, while the dropper discards non-compliant packets. All other routers in the network offer the QoS mechanisms of typical DiffServ core routers, based on the DSCP field in the IP header. These mechanisms are used to trigger congestion control and per-class queuing.
This DiffServ architecture is shown in *Figure 3*.

Fig. 3 IETF DiffServ architecture



Mapping UMTS QoS classes to IP DiffServ classes

A mapping is defined between the UMTS QoS classes and IP differentiated service classes: Assured Forwarding (AF), Expedited Forwarding (EF), and Best Effort Forwarding (BEF). EF, which is characterized by low delay, low jitter and low packet loss, emulates a "leased line". The AF service offers different levels of forwarding assurance. Four independent AF classes are defined; three levels of drop precedence can be used within each AF. The level of forwarding assurance depends primarily on the buffer space and bandwidth allocated to an AF class, the current load of the AF class, the level of the drop precedence class and the level of forwarding assurance defined by the operator.
The mapping between UMTS QoS classes and IP Differentiated Services can be configured by the operator, although a default mapping is provided.

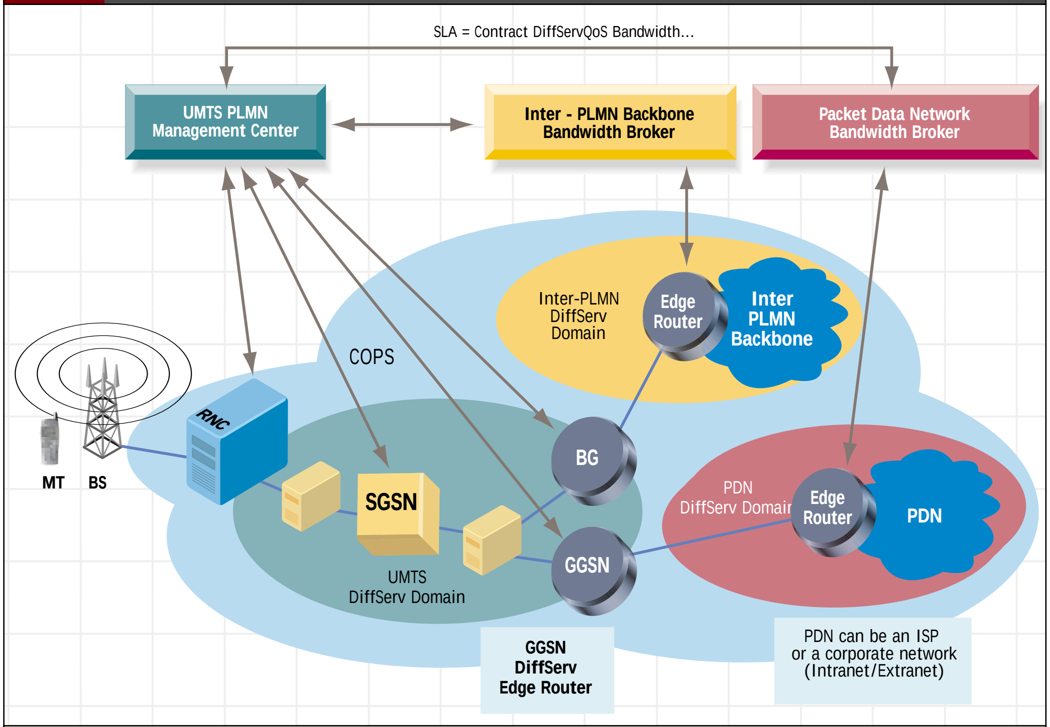
Use of MPLS at Layer 2/3

Alcatel's strategy is to support QoS mechanisms at IP level (layer 3) and above. However, QoS must be provided in an end-to-end way, making it necessary to implement QoS mechanisms at a lower level (layer 2).
One solution is to overdimension the network, as is often proposed for IP / Point-to-Point Protocol / Synchronous Digital Hierarchy networks. However, this approach is not cost-effective and cannot fully protect against congestion. Thus Alcatel is proposing an MPLS architecture for the UMTS core network at layer 2/3.
The main advantage of this architecture is the loose coupling between the transmission technology and the multiple QoS statistical multiplexing system provided by the upper layers. MPLS makes it possible to use any kind of transport network, be it a Wide Area Network (WAN) like ATM or a Local Area Network (LAN) like Gigabit Ethernet. Thus MPLS provides a continuity of QoS policy in the case of a mixed LAN and WAN IP backbone architecture.
In Alcatel's proposal, the SGSN, GGSN and border gateway are edge label switched routers. The UMTS backbone is provided by the Alcatel 7670 core label switched routers. Label Switched Path (LSP) tunnels are established dynamically based on IP routing tables; their bandwidth can be renegotiated according to the user's needs. Network resources are optimized.
When ATM is used under MPLS, the latter inherits all the QoS handling capabilities of ATM. Indeed, ATM fulfills the requirements of the UMTS core network in terms of:

- jitter and delay control;
- bit rate granularity;
- multiplexing flexibility;
- bandwidth efficiency, as a result of the asynchronous nature of the access and the use of fixed length cells.

The operator can benefit from ATM QoS features at the transport layer. ATM CAC reserves some bandwidth for the ATM Virtual Path (VP) or Virtual Circuit (VC) in every node of the UMTS backbone crossed. An ingress policing function checks the incoming traffic rate against the ATM traffic contract. A spacer delays ATM cells when they arrive too early. Packet discard mechanisms intelligently discard ATM Adaptation Layer 5 (AAL5) IP frames when congestion occurs on the ATM backbone. Cell scheduling controls time priority.

Fig. 4 Management and configuration of QoS provisioning



COPS: Common Open Policy Service **PDN:** Packet Data Network

Service Level Agreements

The UMTS PLMN network is a DiffServ domain. Each packet data network can also be a DiffServ domain, as can the inter-PLMN backbone. The PHB policy is uniform within a DiffServ domain. Thus there is a single PHB policy within a UMTS operator's PLMN, and possibly different PHB policies across different domains. The PHB policy can vary across the inter-PLMN backbone and packet data networks.

To guarantee the end-to-end QoS across external networks, Service Level Agreements (SLA) are required between the UMTS network and each external network. SLAs can be signed between the UMTS and every Packet Data Network (PDN) operator, such as a corporate intranet/extranet or Internet Service Provider (ISP), and between UMTS operators and inter-PLMN backbone providers or other PLMN UMTS/GPRS operators.

This architecture is shown in *Figure 4*.

To fulfil the SLAs, management tools are provided for dimensioning the number, type and bandwidth of node interfaces. Observation tools control the use of the network resources within each of the nodes and for each of the interfaces.

QoS Handling in the UMTS Terrestrial Radio Access Network

UTRAN QoS: General Aspects

Radio access bearer services (just called "RAB" in the rest of this article) are established dynamically to support one or several applications for a given mobile user (e.g. speech and Web browsing).

UTRAN is connected to both the circuit- and packet-switched core network domains through the Iu-cs (circuit-switched) and Iu-ps (packet-switched) interfaces, respectively. It can handle one or several RABs per domain and per user at the same time, with each RAB having its own QoS requirements.

To enable the UTRAN to contribute to the provision of end-to-end QoS for the users, each RAB is characterized by QoS attributes which are derived from the characteristics of the application. These characteristics are translated to RAB QoS attributes by the core network. The UTRAN simply obtains the RAB QoS attributes from the core network when the RAB is established.

The role of the UTRAN is then to establish and maintain the RAB with the required QoS levels. The RAB is always established at the request of the core network, which retains ownership of the RAB throughout its life. Once the UTRAN has committed to a given QoS level, this should not be downgraded by the UTRAN without a prior modification request from the core network. In particular, this requirement applies to mobile terminals, which frequently experience varying radio conditions as they move from cell to cell.

Given the basic principle of Wideband Code Division Multiple Access (WCDMA), in which all users share the same resources both in the frequency and time domains, it is essential to maintain a low level of radio interference throughout the system. To this end, power control constrains the transmission power to/from each user between appropriate limits: not too high, to preserve the UMTS cell traffic capacity, and not too low, to avoid excessive transmission errors that would lower the overall bit rate.

In addition, QoS monitoring is essential to ensure that the QoS requirements are met but not exceeded. Smoothing of the traffic flow and congestion avoidance are essential. For example, low priority Non-Real-Time (NRT) traffic may be delayed in favor of high priority Real-Time (RT) services. Radio Admission Control (RAC), radio resource allocation and management, and radio load control functions are key features for handling QoS within UTRAN.

At the lower layers, various transport network layer technologies are foreseen within UTRAN. In the first release of the UMTS standards, ATM is the only technology allowed by the 3GPP standards, while IP is seen as a future option. QoS has also to be considered at the transport network layer.

Radio Admission Control

The RAC function (see *Figure 5*) grants or denies access to radio resources. It is invoked each time the radio resource in a cell has to be shared by one more user or one more RAB, or when a user moves from one cell to another (handover).

When a RAB is established, the RNC maps the RAB QoS attributes onto the radio bearer characteristics and then invokes the RAC function. The characteristics of the required radio bearer are determined on the basis of the traffic class, bit rate requirements, delay requirements, traffic priority parameters, etc. For example, the allocation/retention priority parameter allows low priority calls to be preempted in order to serve higher priority requests.

In the interference-limited UMTS system, one of the roles of RAC is to check whether or not accepting a new RAB will increase the interference beyond an acceptable level, making it impossible to maintain an acceptable radio quality for all established calls. The traffic load also has to be considered: accepting a new RAB should not reduce the throughput and the QoS offered to other users below acceptable limits.

Requests that cannot be served according to these criteria are rejected by UTRAN, or queued if this capability is available.

From the user's viewpoint, it might be more acceptable to have a call rejected at the setup phase, rather than dropped before its normal completion because handover was not possible. Thus, some capacity should be reserved for handover; handover and call establishment can be handled with different priorities.

Radio Resource Allocation and Management

Channelization and scrambling operations are required at the radio physical layer in order to separate the different traffic sources that are sharing the same frequencies. This is outside the scope of the article. Nevertheless, code allocation is linked to QoS to some extent.

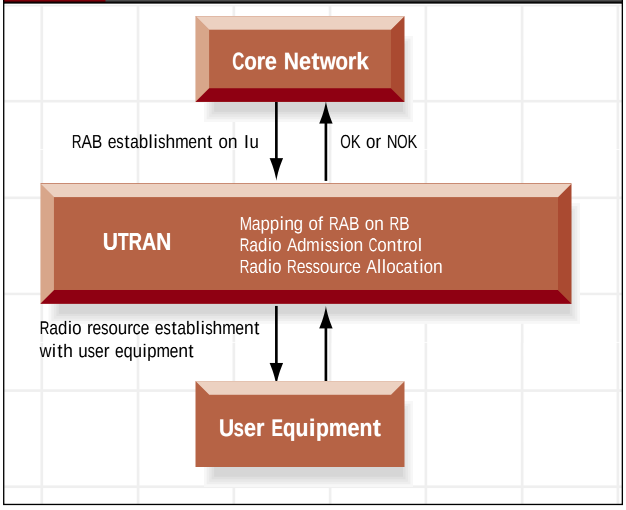
The radio resource allocation function allocates the channelization codes and scrambling codes when requested by the RAC function.

Channelization codes achieve spectrum spreading and are allocated according to the bit rate. The higher the bit rate, the lower the Spreading Factor (SF). Codes can be arranged as a code tree, as illustrated in *Figure 6*. An important point is that when a given code is allocated to a user, neither the codes of higher SFs nor those of lower SFs on the same branch can be allocated to other users, otherwise the receiving side could not properly decode the corresponding signals.

Thus code tree management is of a particular importance, especially for high bite rate services (low SF), to minimize the risk of code shortages. For example, code tree reshuffling is performed to change the code of established calls in order to maximize the traffic capacity of the UMTS cell in terms of the bit rate and number of users.

In addition, as the standards define several types of transport channel, the radio resource allocation and management function selects appropriate types of transport channels, depending on the required QoS attributes.

Fig. 5 RAB service establishment principles



Radio Load Control

This function maintains stability in the radio network by monitoring the load, and detecting and handling overload situations (i.e. when the network is running out of resources). In particular, radio load control guarantees that some algorithms, like power control, remain stable in the event of a temporary overload. Two main effects can occur in such situations. On the uplink, mobiles might increase their transmit power to overcome the increased interference, thus making the situation worse. This is often referred to as the "party effect". A similar effect can occur on the downlink.

Radio overload situations are detected by monitoring the interference level and traffic volume. Congestion avoidance can be achieved using preventive algorithms in coordination with RAC.

Synchronization and Scheduling

Synchronization mechanisms are defined within UTRAN to meet the time requirements for each type of RAB.

Fig. 6 Channel code tree principles

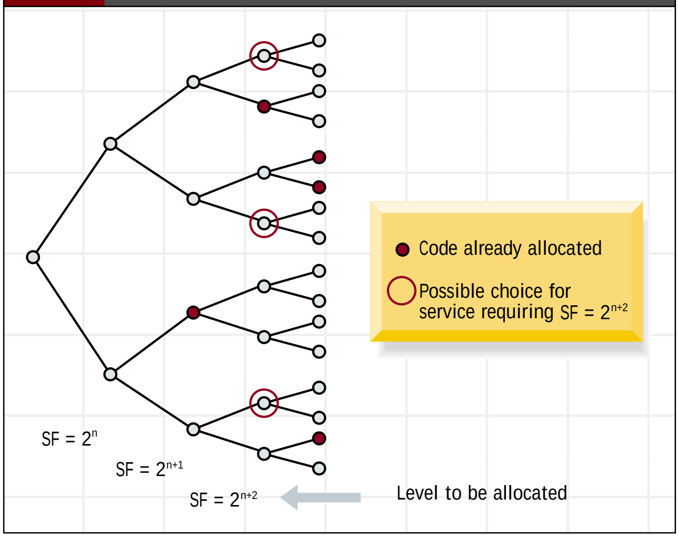
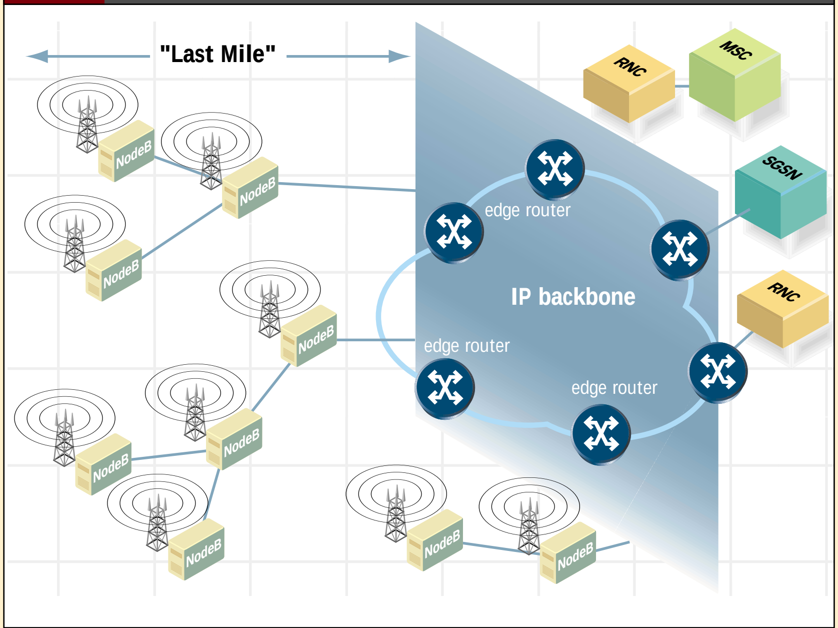


Fig. 7 IP as a transport option in UTRAN



Time adjustment takes place between the Node B and the RNCs to maintain the overall transfer delay across the UTRAN below given limits, which depend on the type of transport channel used within the UTRAN, which itself depends on the RAB QoS attributes. For example, speech services require a short transfer delay (e.g. less than 5 ms within UTRAN), whereas NRT data services can tolerate larger but constrained delays (e.g. 50 to 100 ms within UTRAN). In the case of NRT traffic, flow control may take place between UTRAN nodes.
In the case of downlink traffic, the scheduling of data units over the radio interface is controlled by the RNC, which defines the actual time of transmission of each block.

Transport Network (Iu, Iu-b, Iu-r)

In Release 3 of the UTRAN standards, the transport network is based on ATM. ATM Permanent Virtual Circuits (PVC) are set up between UTRAN nodes by management platforms. The control planes and the user plane protocols are carried over different ATM VC connections. AAL5 is used for the control planes on the Iu, Iu-b and Iu-r interfaces, as well as for the Iu-ps user plane, whereas AAL2 is used for the user plane on the Iu-b, Iu-r and Iu-cs interfaces.
Except on Iu-ps, the QoS requirements for each service are fulfilled at ATM level using the ATM Transfer Capabilities (ATC), and at AAL2 level by intelligent scheduling and priority handling. In the user plane, AAL2/ATM traffic contracts on the Iu-cs, Iu-r and Iu-b interfaces are deduced from the RAB QoS attributes. On Iu-ps, the QoS requirements for each service are met at the IP layer using DiffServ, and not at the ATM layer.
When implemented, QoS differentiation can realize significant bandwidth savings on the terrestrial interfaces, because real-time tolerant data flows can be transmitted with longer delays than real-time stringent data flows. QoS differentiation can be achieved at the ATM and/or AAL2 levels.

QoS Negotiation between UTRAN and Core Network

3GPP is envisaging that the RNC should be able to negotiate some of the QoS attributes requested by the core network. For example, the RNC could inform the core network that the RAB establishment request could be accepted provided that the bit rate is reduced to some extent. QoS renegotiation might also take place for an established RAB. In any case, it will be up to the core network to decide what is or is not acceptable to the user.

IP Transport Network

In Release 4 of the standards, 3GPP is specifying the use of IP as an alternative to ATM (see *Figure 7*). However, the introduction of IP as a transport technology raises the question as how to efficiently use the small bandwidth of the slow “last mile” links between the IP transport network and the Node Bs while meeting the transfer delay constraints.
To this end, two features are under discussion in 3GPP:

- Segmentation of large frames and interleaving of segments to prevent, for example, small real-time packets from being delayed by large non-real-time packets.
- Multiplexing of small packets within IP frames to reduce User Data Protocol / Internet Protocol overhead.

In the IP backbone, a DiffServ QoS scheme can be envisaged, whereby voice and packet data are treated with stringent delay requirements. Using this scheme, QoS differentiation between the different flows of the RNC/Node B traffic is achieved by implementing priority-based queuing in the nodes, allocating different priorities to the different traffic classes.
Alternatively, several service classes can be used in the IP network (if this capability is available), for example, other DiffServ classes, MPLS traffic engineered paths, etc.

Multimedia Support

Release 5 of the 3GPP standards introduces IP multimedia services with Session Initiation Protocol (SIP) capable mobile. Apart from the multimedia call server and media gateway, this affects QoS handling in the core network in three main ways:

- Resource reservation within the packet-switched domain will be synchronized with the MultiMedia Call Control (MMCS).
- GGSN might have to terminate Resource Reservation Protocol (RSVP) signaling. For example, a fixed SIP phone called by or calling mobile SIP user equipment sends RSVP signaling messages.

• Service domain nodes (home subscriber server, application server) influence QoS provisioning in the call state control function handled by the MMCS through the policy enforcement point and policy decision point.

Conclusion

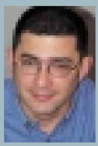
Considering the huge increase in mobile wireless communications, both for speech and data services, providing end-to-end quality of service to support a variety of applications has become of key importance for both operators and users. Since the perceived QoS has a major impact on subscriber satisfaction, the ability to provide an appropriate QoS will be a differentiating factor between UMTS operators.
As an end-to-end UMTS solution provider, Alcatel has developed a set of mechanisms for both the core network and the radio access network. On the core network side, Alcatel implements DiffServ mechanisms at IP level, MPLS functions at layer 2, UMTS call admission control (SGSN, GGSN, MSC, RNC, etc), and QoS network management and provisioning. In the UTRAN, Alcatel has developed radio features that contribute to end-to-end QoS while optimizing the use of both radio and terrestrial transmission resources.
Of course, the Alcatel implementation complies fully with all the relevant standards, supporting all the QoS classes defined in 3GPP TS 23.107 [1]. ■

Reference

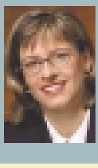
1. "QoS Concept and Architecture", *3GPP Recommendation TS 23.107.*

Serge Baudet is Product Manager within the Evolium joint venture in Vélizy, France.

Claire Besset-Bathias is GPRS Core Network Product Manager at the Alcatel Switching and Routing Division in Vélizy, France.



Patrick Frêne is Director of the Mobile Internet Product Line within the Alcatel Switching and Routing Division in Vélizy, France.



Natalie Giroux is head of the Architecture and Technology Group in the Alcatel Carrier Internetworking Division CTO office in Kanata, Canada.