5

End to End Quality of Service in UMTS Systems

Wei Zhuang
China Telecom Co. Ltd. (Shanghai)
P.R.China

1. Introduction

About ten years ago, WCDMA [1] based the third generation mobile systems started to be deployed worldwide. Besides to support basic mobile data services such as file transfer and internet surfer, etc., UMTS has one of the most significant archievements which can support a richer variety of services with QoS guarantee, such as video, VOIP, etc. Quality of Service (QoS) is defined as ”the collective effect of service” performance, which determines the degree of satisfaction of a user of the service in the ITU-T recommendation E.800. At a technical level, QoS can be characterized by service availability, delay, jitter, throughput, packet loss rate (Nortel White Paper, 2002).

3GPP has put many efforts to define and standardize a QoS framework for data services, specially IP-based services. The standardization of a UMTS QoS model started in 1999. the development was based on the following key principles: operation and QoS provisioning needed to be possible in the wireless environment, usage of the Internet QoS mechanisms, applications and interoperability. This chapter is aimed to provide an overview of the UMTS end-to-end QoS architecture, describe how the QoS requirements to be realized from top layer to wireless links.

2. WCDMA QoS architecture

QoS standardization in UMTS PS domain enables UMTS to provide data service with end-to-end QoS guarantees. 3GPP proposed a layered architecture for supporting end-to-end QoS. It includes the following key elements (Sudhir Dixit et al., 2001):

- Mapping of end-to-end services provided by the UE, UTRAN, Core Network (CN), and external IP networks;
- Traffic classes and associated QoS parameters;
- Location of QoS functions;
- QoS negotiation;
- Multiplexing of flows onto network resources;
- An end-to-end data delivery model.

[1] Wideband Code Division Multiple Access W-CDMA - the radio technology of UMTS - is a part of the ITU IMT-2000 family of 3G Standards.

The layered UMTS QoS architecture is shown in Figure 1. The UMTS network can provide end-to-end QoS services from a Terminal Equipment (TE) to another TE. A network bearer service describes how to realize a certain network QoS. It is defined by the control signaling, user traffic transport and QoS management functionality, which enabling the provision of a contracted QoS (Sudhir Dixit et al., 2001; 3GPP23107, 2011). As the end-to-end service is conveyed over several networks, the end-to-end bearer service consists of different network bearer services. The end-to-end bearer service can be decomposed into TE/MT local bearer service, the UMTS bearer service and the external bearer service.

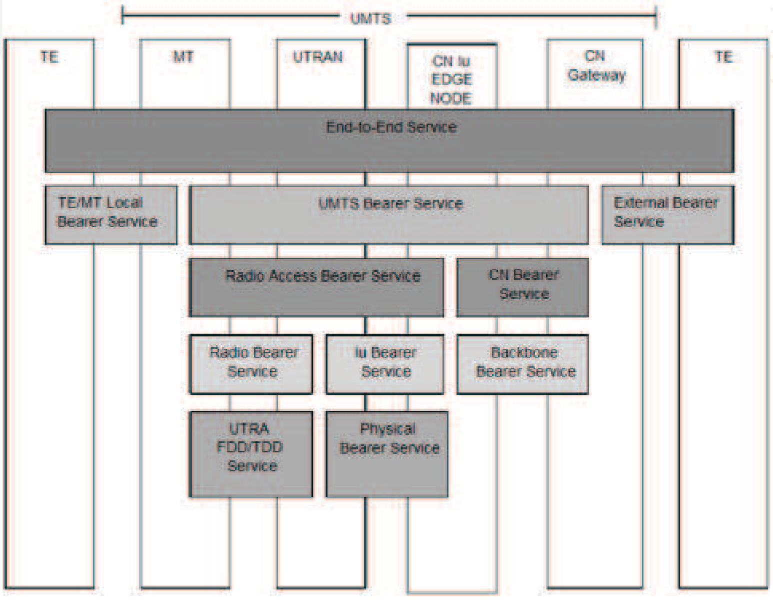


Fig. 1. UMTS QoS architecture

The TE/MT local bearer service provides communication between the TE and MT parts. MT (Mobil Terminal) provides connection to the UTRAN with basic functions, such as radio attachment to 3G network, authenticating the CS/PS domain, mobility management, etc. TE support call control, authenticating the IMS subscription, etc.

The external bearer service deals with the interoperability and interworking aspects with external IP bearer, and provides the appropriate functionality to support it. It is logical located in the GGSN, which is the gateway of UMTS to external network (Sotiris et al., 2002).

UMTS bearer service provides service by using the radio access bearer service (RAB) and the core network bearer service. The detail is given in the following.

2.1 UMTS bear service

The UMTS QoS is provided by the UMTS bearer service. It includes the radio access bearer service and the core network bearer service. They reflect the optimized way to realize the UMTS Bearer Service over the respective cellular network topology taking into account aspects such as mobility and mobile subscriber profiles.

The Radio Access Bearer Service provides confidential transport of signalling and user data between MT and CN Iu Edge Node with the QoS adequate to the negotiated UMTS Bearer Service or with the default QoS for signalling. This service is based on the characteristics of the radio interface and is maintained for a moving MT.

The Core Network Bearer Service of the UMTS core network connects the UMTS CN Iu Edge Node with the CN Gateway to the external network. The role of this service is to efficiently control and utilise the backbone network in order to provide the contracted UMTS bearer service. The UMTS packet core network shall support different backbone bearer services for variety of QoS. And the UMTS bearer service is realized by a GPRS service in the PS domain or a speech/data service in the CS domain.

2.1.1 The radio bearer service and Iu bearer service

The Radio Access Bearer Service is realised by a Radio Bearer Service and an Iu-Bearer Service. The role of the Radio Bearer Service is to cover all the aspects of the radio interface transport. This bearer service uses the UTRA FDD/TDD, which is not elaborated further in this chapter.

The Iu-Bearer Service together with the Physical Bearer Service provides the transport between UTRAN and CN. Iu bearer services for packet traffic shall provide different bearer services for variety of QoS.

2.1.2 The backbone network service

The Core Network Bearer Service uses a generic Backbone Network Service. The Backbone Network Service covers the layer 1/Layer2 functionality and is selected according to operator's choice in order to fulfill the QoS requirements of the Core Network Bearer Service. The Backbone Network Service is not specific to UMTS but may reuse an existing standard.

2.2 QoS requirement

2.2.1 UMTS QoS classes

The layered UMTS QoS architecture requires the definition of QoS attributes for each bearer service. When defining the UMTS QoS classes, the restrictions and limitations of the radio interface have to be taken into account. The QoS mechanism should be simpler than that in wired network due to different error characteristics of the air interface. Table 1 illustrates the QoS classes defined by 3GPP.

The main distinguishing factor between these QoS classes is how delay sensitive the traffic is.

Conversational class

The transfer time of real time conversation scheme shall be low because of the conversational nature of the scheme and at the same time that the time relation (variation) between information entities of the stream shall be preserved in the same way as for real time streams. The maximum transfer delay is given by the human perception of video and audio conversation. Therefore the limit for acceptable transfer delay is very strict, as failure to provide low enough transfer delay will result in unacceptable lack of quality. The transfer delay requirement is therefore both significantly lower and more stringent than the round trip delay of the interactive traffic case. The fundamental characteristic for QoS is to preserve time relation (variation) between information entities of stream and conversational pattern (stringent and low delay) (3GPP23107, 2011).

The most well known use of this scheme is telephony speech (e.g. GSM). But with Internet and multimedia a number of new applications will require this scheme, for example voice over IP and video conferencing tools. Real time conversation is always performed between

Traffic class	Conversational class conversational RT	Streaming class streaming RT	Interactive class Interactive best effort	Background Background best effort
Fundamental characteristics	- Preserve time relation (variation) between information entities of the stream Conversational pattern (stringent and low delay)	- Preserve time relation (variation) between information entities of the stream	- Request response pattern - Preserve payload content	- Destination is not expecting the data within a certain time - Preserve payload content
Example of the application	voice	streaming video	Web browsing	background download of emails

Table 1. UMTS QoS classes

peers (or groups) of live (human) end-users. This is the only scheme where the required characteristics are strictly given by human perception.

Streaming class

Streaming class is characterised by that the time relations (variation) between information entities (i.e. samples, packets) within a flow shall be preserved, although it does not have any requirement on low transfer delay (3GPP23107, 2011). This scheme is one of the newcomers in data communication, raising a number of new requirements in both telecommunication and data communication systems. It is a one-way transport. A user can use this class to watch (listen to) real time video (audio).

The delay variation of the end-to-end flow shall be limited, to preserve the time relation (variation) between information entities of the stream. But as the stream normally is time aligned at the receiving end (in the user equipment), the highest acceptable delay variation over the transmission media is given by the capability of the time alignment function of the application. Acceptable delay variation is thus much greater than the delay variation given by the limits of human perception. The fundamental characteristics for streaming class QoS is to preserve time relation (variation) between information entities of the stream.

Interactive class

Interactive traffic is the other classical data communication scheme that on an overall level is characterized by the request response pattern of the end-user. At the message destination there is an entity expecting the message (response) within a certain time. Round trip delay time is therefore one of the key attributes. Another characteristic is that the content of the packets shall be transparently transferred (with low bit error rate). The fundamental characteristics for interactive class QoS is to request response pattern, preserve payload content.

This class is applied when an end-user (either a machine or a human) is using on line requesting data from remote equipment (e.g. a server). Examples of human interaction with the remote equipment are: web browsing, data base retrieval, server access. Examples of machines interaction with remote equipment are: polling for measurement records and automatic data base enquiries (tele-machines).

Background class

Background traffic is one of the classical data communication schemes that on an overall level is characterised by that the destination is not expecting the data within a certain time. The scheme is thus more or less delivery time insensitive. Another characteristic is that the content of the packets shall be transparently transferred (with low bit error rate). The fundamental characteristics for background class QoS are a) destination is not expecting the data within a certain time; b) preserve payload content.

When the end-user, that typically is a computer, sends and receives data-files in the background, this scheme applies. Examples are background delivery of E-mails, SMS, download of databases and reception of measurement records.

2.2.2 UMTS bearer service attributes

UMTS bearer service attributes describe the service provided by the UMTS network to the user of the UMTS bearer service. A set of QoS attributes (QoS profile) specifies this service. At UMTS bearer service establishment or modification different QoS profiles have to be taken into account.

Traffic class ('conversational', 'streaming', 'interactive', 'background')

Traffic class is a type of application for which the UMTS bearer service is optimized. By including the traffic class itself as an attribute, UMTS can make assumptions about the traffic source and optimise the transport for that traffic type.

Maximum bit-rate (kbps)

Maximum bit-rate (kbps) is the maximum number of bits delivered to UMTS at a SAP (Service Access Point) within a period of time, divided by the duration of the period. The traffic is conformant with Maximum bit-rate as long as it follows a token bucket algorithm where token rate equals Maximum bit-rate and bucket size equals Maximum SDU size.

The algorithm is well known as "Token Bucket Algorithm" which has been described in IETF. It is a reference algorithm for the conformance definition of bitrate. This may be used for traffic contract between UMTS bearers and external network/user equipment. In the algorithm, "tokens" represents the allowed data volume, for example in byte. "Tokens" are given at a constant "token rate" by a traffic contract, are stored temporarily in a "token bucket", and are consumed by accepting the packet. This algorithm uses the following two parameters (r and b) for the traffic contract and one variable (TBC) for the internal usage.

- r: token rate, (corresponds to the monitored Maximum bitrate/Guaranteed bitrate).
- b: bucket size, (the upper bound of TBC, corresponds to bounded burst size).
- TBC (Token bucket counter): the number of given/remained tokens at any time.

According to a token bucket, conformance can be defined as: "Data is conformant if the amount of data submitted during any arbitrarily chosen time period T does not exceed (b+rT)".

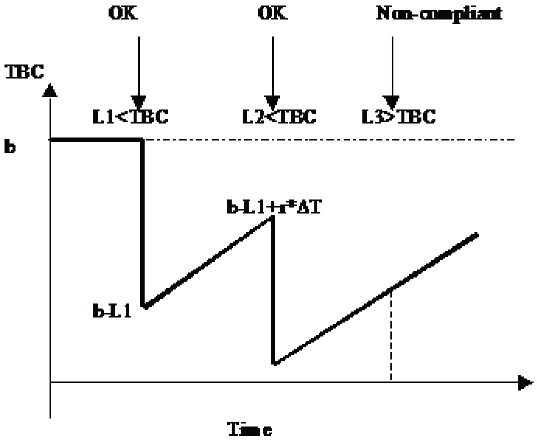


Fig. 2. Operation example of the reference conformance algorithm

The algorithm is described here (Figure 2, (3GPP23107, 2011)). Token bucket counter (TBC) is usually increased by "r" in each small time unit. However, TBC has upper bound "b" and the value of TBC shall never exceed "b". When a packet i with length Li arrives, the receiver checks the current TBC. If the TBC value is equal to or larger than Li, the packet arrival is judged compliant, i.e., the traffic is conformant. At this moment tokens corresponding to the packet length is consumed, and TBC value decreases by Li. When a packet j with length Lj arrives, if TBC is less than Lj, the packet arrival is non-compliant, i.e., the traffic is not conformant. In this case, the value of TBC is not updated.

The Maximum bitrate is the upper limit a user or application can accept or provide. All UMTS bearer service attributes may be fulfilled for traffic up to the Maximum bitrate depending on the network conditions. The downlink of the radio interface can use maximum bitrate to make code reservations. Its purpose is:

1. to limit the delivered bitrate to applications or external networks with such limitations;
2. to allow maximum wanted user bitrate to be defined for applications able to operate with different rates (e.g. applications with adapting codecs).

Guaranteed bitrate (kbps)

Guaranteed bitrate (kbps) is defined as: a guaranteed number of bits delivered by UMTS at a SAP within a period of time (provided that there is data to deliver), divided by the duration of the period. The traffic is conformant with the guaranteed bitrate as long as it follows a token bucket algorithm where token rate equals Guaranteed bitrate and bucket size equals Maximum SDU size.

UMTS bearer service attributes, e.g. delay and reliability attributes, are guaranteed for traffic up to the Guaranteed bitrate. For the traffic exceeding the Guaranteed bitrate the UMTS bearer service attributes are not guaranteed. Guaranteed bitrate may be used to facilitate admission control based on available resources, and for resource allocation within UMTS.

Delivery order (y/n)

Delivery order indicates whether the UMTS bearer shall provide in-sequence SDU delivery or not. This attribute is derived from the user protocol (PDP type) and specifies if out-of-sequence SDUs are acceptable or not. Whether out-of-sequence SDUs are dropped or re-ordered depends on the specified reliability.

Maximum SDU size (octets)
Maximum SDU size (octets) means the maximum allowed SDU size. The maximum SDU size is used for admission control and policing.

SDU format information (bits)
SDU format information (bits) is a list of possible exact sizes of SDUs. UTRAN needs SDU size information to operate in transparent RLC protocol mode, which is beneficial to spectral efficiency and delay when RLC re-transmission is not used. Thus, if the application can specify SDU sizes, the bearer is less expensive.

SDU error ratio
SDU error ratio indicates the fraction of SDUs lost or detected as erroneous. SDU error ratio is defined only for conforming traffic. It is used to configure the protocols, algorithms and error detection schemes, primarily within UTRAN.

Residual bit error ratio Residual bit error ratio indicates the undetected bit error ratio in the delivered SDUs. If no error detection is requested, Residual bit error ratio indicates the bit error ratio in the delivered SDUs. It is used to configure radio interface protocols, algorithms and error detection coding.

Delivery of erroneous SDUs (y/n/-)
Delivery of erroneous SDU (yes/no/-) indicates whether SDUs detected as erroneous shall be delivered or discarded. 'Yes' implies that error detection is employed and that erroneous SDUs are delivered together with an error indication, 'no' implies that error detection is employed and that erroneous SDUs are discarded, and '-' implies that SDUs are delivered without considering error detection. It is used to decide whether error detection is needed and whether frames with detected errors shall be forwarded or not.

Transfer delay (ms)
Transfer delay (ms) indicates maximum delay for 95th percentile of the distribution of delay for all delivered SDUs during the lifetime of a bearer service, where delay for an SDU is defined as the time from a request to transfer an SDU at one SAP to its delivery at the other SAP. Transfer delay can be used to specify the delay tolerated by the application. It allows UTRAN to set transport formats and ARQ parameters.

Traffic handling priority
Traffic handling priority specifies the relative importance for handling of all SDUs belonging to the UMTS bearer compared to the SDUs of other bearers. Within the interactive class, there is a definite need to differentiate between bearer qualities. This is handled by using the traffic handling priority attribute, to allow UMTS to schedule traffic accordingly. By definition, priority is an alternative to absolute guarantees, and thus these two attribute types cannot be used together for a single bearer.

Allocation/Retention priority
Allocation/Retention priority specifies the relative importance compared to other UMTS bearers for allocation and retention of the UMTS bearer. The Allocation/Retention

Priority attribute is a subscription attribute which is not negotiated from the mobile terminal. Priority is used for differentiating between bearers when performing allocation and retention of a bearer. Where there is no enough resource, the relevant network elements can use the Allocation/Retention Priority to prioritize bearers with a high Allocation/Retention Priority over bearers with a low Allocation/Retention Priority when performing admission control.

Source statistics descriptor ('speech'/'unknown')

Source statistics descriptor ('speech'/'unknow') specifies characteristics of the source of submitted SDUs. Conversational speech has a well-known statistical behaviour (or the discontinuous transmission (DTX) factor). By using source statistics descriptor, a network element can know whether the SDUs of a UMTS bearer generated by a speech source or not. UTRAN, the SGSN and the GGSN and also the UE may, based on experience, calculate a statistical multiplex gain for use in admission control on the relevant interfaces.

2.3 QoS management for UMTS bearer service

In the section, an overview of QoS functions is described which is used to establish, modify, and maintain a UMTS bearer service with a specific QoS. The allocation of these functions to the UMTS entities indicates the requirement for specific entity to enforce the QoS commitments negotiated for the UMTS bearer service. UMTS is split into user plane and control plane for easy expanding in the future. So QoS management functions are also split into user plane and control plane. All of the QoS management functions in both planes (control and user plane) will ensure the provision of the negotiated service between the access points of the UMTS bearer service. The end-to-end service is provided by translation/mapping with UMTS external services.

2.3.1 QoS management in control plane

The QoS management functions in control plane are shown in Figure 3. The QoS functions for UMTS bearer service include service manager, translation function, admission/capability control and subscription control in the control plane. These functions are used to establish and modify a UMTS bearer service through signaling/negotiating with UMTS external services, establishing/modifying UMTS internal services.

Subscription control checks the administrative rights when an UMTS bearer service user requires a service with the specified QoS. It is located at CN EDGE.

Admission capability control will maintains available resource information of a network entity, and resource allocated to UTMS bearer service. When receiving an UMTS bearer service request or modification request, admission / capability control function determines whether the required resources can be provided or not. If the network can provide resources, it will reserve these resources. This function also checks the capability of a network entity, i.e. whether the specific service is implemented and not blocked for administrative reasons. It is located at MT, UTRAN, CN EDGE and Gateway.

Translation function is used to convert between internal service primitives and external protocols. It is located at MT and Gateway. At MT and Gateway, translation function converts between UMTS bearer service attributes and external network QoS attributes.

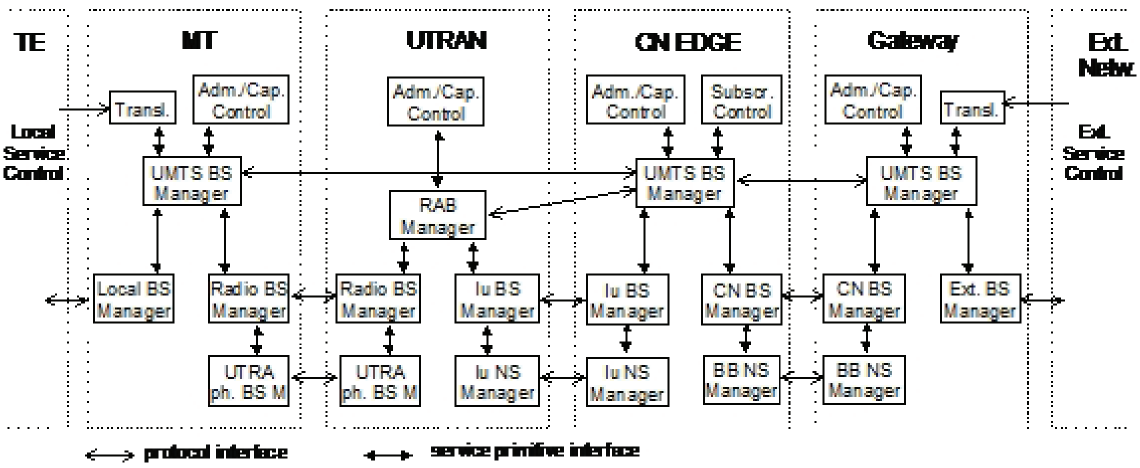


Fig. 3. QoS management function for UMTS bearer service in the control plane

Service manager co-ordinates the related functions in control plane to establish, modify and maintain the service. All user plane QoS management functions are supported by service manager with the relevant attributes. The service manager may perform an attribute translation to request lower layer services. Service manager at UMTS bearer service level is located at MT, CN EDGE and Gateway. The UMTS BS manager can signal among each other and via the translation function with external instances to establish / modify a UMTS bearer service. The UMTS BS manager will interrogate with its associated admission / capability control whether the network entity supports a specific requested service and whether the required resource is available. The UMTS BS manager at CN EDGE also has to verify with the subscription control the administrative rights for using the service. Based on the layered UMTS QoS architecture, UMTS bearer service manager will translate the UMTS bearer service attributes into attributes of the lower layer service manager. For example, the UMTS BS manager of the CN EDGE will translate the UMTS bearer service attributes into RAB service attributes, Iu bearer service attributes, and CN bearer service attributes. Each low layer will provide service to upper layer service manager.

2.3.2 QoS management in user plane

The Figure 4 shows the QoS management functions of UMTS bearer service in the user plane. They are mapping function, classification function, resource manager and traffic conditioner. They are used to maintain the data transfer characteristics according to the commitments established by the UMTS BS control functions.

Mapping function provides each data with the specific marking for receiving the requested QoS at the transfer. It is located at UTRAN, Gateway.

Classification function (Class.) in the MT and Gateway assigns user data units received from the external bearer service or the local bearer service to the appropriate UMTS bearer service according to the QoS requirement of each user data unit.

Traffic conditioner provides conformance between the negotiated QoS for a service nad the data unit traffic. Policing or traffic shaping is used for traffic conditioning. The policing function compares the data unit traffic with the related QoS attributes. Data units not matching the relevant attributes will be dropped or marked as not matching, for preferential

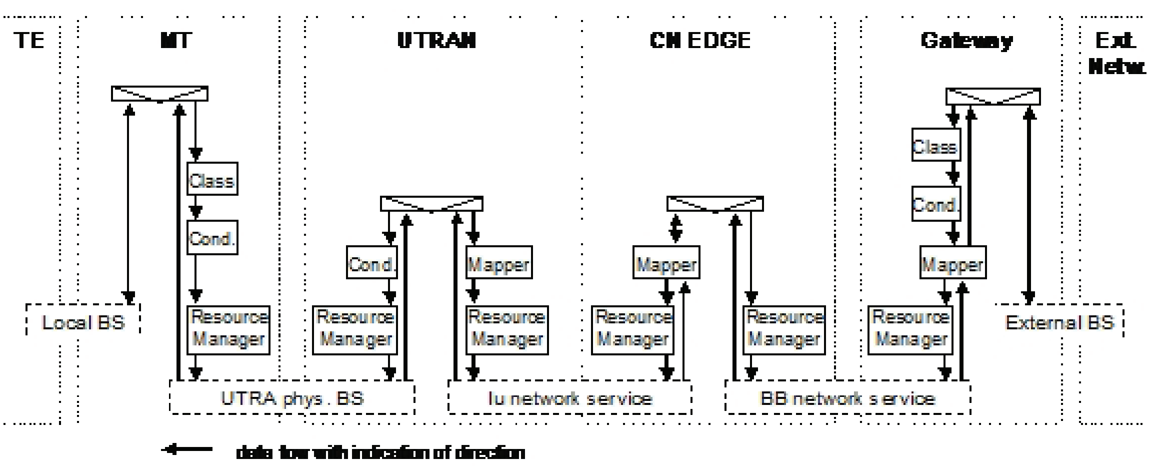


Fig. 4. QoS management function for UMTS bearer service in the user plane

dropping in case of congestion. The traffic shaper forms the data unit traffic according to the QoS of the service. The shaper algorithm is "Token Bucket Algorithm". At MT side, the traffic conditioner (Cond.) provides conformance of the uplink user data traffic with the QoS attributes of the relevant UMTS bearer service. In the Gateway a traffic conditioner may provide conformance of the downlink user data traffic with the QoS attributes of the relevant UMTS bearer service; i.e., on a per PDP context basis. A traffic conditioner in the UTRAN forms this downlink data unit traffic according to the relevant QoS attributes.

Resource Manager distributes the available resources between all services sharing the same resource. The resource manager distributes the resources according to the required QoS. Example means for resource management are scheduling, bandwidth management and power control for the radio bearer. It is located at MT, UTRAN, CN EDGE and Gateway.

2.4 QoS management in UMTS

2.4.1 QoS in CS domain

CS call control will control the QoS in the CS domain. MSC server and CS-MGW will provide QoS related functions. For UMTS release '99 CS-CC, the QoS related bearer definitions of GSM (as defined in bearer capability information element, octet 6 and its extensions) are sufficient.

In the CS domain the UE can only request a certain service with a well defined set of QoS parameters. CS domain uses traditional "circuit switching" technology, i.e. a constant set of resources exclusively dedicated to a connection. All CS domain services will require real-time bearers but differ in bandwidth and delay requirements. Based on the Bearer Capability information element the following services can be identified:

1. speech: from the Information Transfer Capability (ITC) parameter;
2. data, non-transparent: from the ITC and Connection element (CE) parameters;
3. data, transparent: from the ITC and CE parameters.

According to the standard, speech as well as the transparent data service is mapped to the conversational class while the non-transparent data service is mapped to the streaming class.

The MSC-Server is responsible for the service negotiation which includes subscription check and admission control. Furthermore, the QoS parameters corresponding to the service have to be mapped specifically for the interfaces to the UTRAN, GMSC-Server and CS-MGW. To provide QoS the CS-MGW has to perform admission control for the bearer resource which is therefore a part of the call admission control. Additionally, the CS-MGW is responsible for the QoS mappings to the Iu-, CN- and external bearer services.

With the separation of transport and control in the CS domain the resource allocation becomes more flexible. The new transport techniques ATM and IP (which are available for the CS bearer independent domain) allow a more efficient network usage from a parallel transmission of voice and data possibly leading to the consolidation of the whole PLMN (including the PS domain and parts of RAN) on one transport network. The QoS issue in the CS domain with IP or ATM based transport is to guarantee the same QoS as a TDM based PLMN with increased bandwidth efficiency.

2.4.2 QoS in PS domain

Since the PS domain provides packet data services, which are characterized by individual transmission of packets. QoS of different packet service is defined by a set of explicitly defined QoS parameters. So some effort is necessary to assure that packets of one flow are transmitted with guaranteed QoS.

The 3GPP specificationscitep (3GPP23107, 2011) define the QoS management functions in the UMTS bearer service for both control plane and user plane. Establishment of QoS within a UMTS network is achieved through the Packet Data Protocol (PDP) context activation procedure. The user equipment (UE) sends an Active PDP Context Request message to the SGSN, which contains the desired QoS profile, among other parameters. With these QoS attributes the treatment of the packets is sufficiently defined and all packets (or flows) belonging to the same PDP context are handled in the same way by the GPRS bearer service. After the UE sends a PDP context request with explicitly defined QoS parameters, the SGSN will negotiate the QoS parameters which includes subscription check and admission control (capability and resource check). Then the SGSN interacts with the UTRAN and the GGSN to establish the PDP context. The GGSN also performs admission control, i.e. the resource check for the GPRS as well as for the external bearer service. Additionally, the GGSN has to map QoS parameters from the GPRS to the external bearer service.

2.4.3 QoS in IP multimedia subsystem

IP multimedia subsystem (IMS) is introduced in 3GPP Release 5. It is an IP based system overlay on the PS domain. It support Session Initiation Protocol (SIP) based multimedia service. IMS can support end-to-end IP QoS service by using IP based bearer service. The IP based bearer service is supported by MS local bearer service, UMTS bearer service and external service.

Since 3GPP Release 5, the UMTS will support QoS in the IP layer between UE and multimedia application server/UE. The UE and GGSN have important roles in the IP layer QoS framework, they map QoS parameters between IP layer bearer service and UMTS bearer service. The detail will be discussed in the section 3.

For supporting IP layer QoS, 3GPP introduces the policy based QoS management in the IMS. The policy framework is recommended for policy management in IETF. The detail discusses is given in the section 4.

3. End-to-end IP QoS over UMTS

With the evolution of the 3GPP standards, operators want to provide end-to-end QoS enabled services in UMTS. The end-to-end behavior provided by a series of network elements is an assured level of bandwidth that produces a delay-bounded service with no queueing loss for all conforming packet data (RFC2212, 1997). Assuming the network is functioning correctly, these applications may assume that (**?**):

- A very high percentage of transmitted packets will be successfully delivered by the network to the receiving end-nodes. (The percentage of packets not successfully delivered must closely approximate the basic packet error rate of the transmission medium).
- The transit delay experienced by a very high percentage of the delivered packets will not greatly exceed the minimum transmit delay experienced by any successfully delivered packet. (This minimum transit delay includes speed-of-light delay plus the fixed processing time in routers and other communications devices along the path.)

The end-to-end QoS architecture is provided in Figure 1 in section 2. IP level mechanisms are necessary in providing end-to-end QoS services by interacting TE/MT local bearer service, GPRS bearer service and external bearer service. In this section, how to implement end-to-end IP QoS is described.

3.1 QoS mechanisms in IP

Quality of service refers to the nature of the packet delivery service provided, as described by parameters such as achieved bandwidth, packet delay, and packet loss rates (RFC2216, 1999). The Internet, as originally conceived, offers only a very simple quality of service (QoS), point-to-point best-effort data delivery. It means the network just offered available bandwidth and delay characteristics dependent on instantaneous network load. Before real-time applications such as remote video, multimedia conferencing, visualization, and virtual reality can be broadly used, the Internet infrastructure must be modified to support real-time QoS, which provides some control over end-to-end packet delays. From the view of applications, QoS is realized by adequate provisioning of the network infrastructure. In contrast, a network with dynamically controllable quality of service allows individual application sessions to request network packet delivery characteristics according to their perceived needs, and may provide different qualities of service to different applications. There are two basic types of QoS available (qodwhitepaper, 1999):

- Resource reservation (integrated services): network resources are apportioned according to an application's QoS requirement, subject to bandwidth management policy.
- Prioritization (differentiated services): network traffic is classified and apportioned network resources according to bandwidth management policy.

The both types of QoS can be applied to individual application 'flow' or to flow aggregates, so there are two other methods to characterize types of QoS:

- Per flow: A 'flow' is defined as an individual, uni-directional data stream between two clients (caller and callee), uniquely identified by a 5-tuple (transport protocol, source address, source port number, destination address, and destination port number).
- Per aggregate: An aggregate is simply two or more flows. Usually the flows have something in common (e.g. any one or more of 5-tuple parameters, a label or a priority number, or perhaps some authentication information).

Generally, we can see that there are two methods to support QoS in IP network. One is IntServ (Integrated Service), the other is DiffServ (Differentiated Service). IneServ is Per Flow based QoS control mechanism. Diffserv is Per Aggregate based QoS control mechanism. To accommodate the need for these two types of QoS, there are following QoS protocols and algorithms:

- ReSerVation Protocol (RSVP):
- Differentiated Service (DiffServ)
- Multi Protocol Labeling Switching (MPLS)

3.1.1 IntServ

The Internet integrated services (IntServ) framework provides the ability for applications to choose among multiple, controlled levels of delivery service for their data packets. It can provide hard QoS guarantee to individual traffic flows. To support this capability, two things are required (**?**):

- Individual network elements (subnets and IP routers) along the path followed by an application's data packets must support mechanisms to control the quality of service delivered to those packets.
- A way to communicate the application's requirements to network elements along the path and to convey QoS management information between network elements and the application must be provided.

In the integrated services framework the first function is provided by QoS control services such as Controlled-Load (RFC2211, 1997) and Guaranteed (RFC2212, 1997). The second function may be provided in a number of ways, but is frequently implemented by a resource reservation setup protocol such as RSVP (RFC2205, 1997).

The controlled load service is intended to support a broad class of applications which have been developed for use in today's Internet, but are highly sensitive to overloaded conditions. Important members of this class are the ”adaptive real-time applications” currently offered by a number of vendors and researchers. These applications have been shown to work well on unloaded nets, but to degrade quickly under overloaded conditions. It is equivalent to ”best effort service under unloaded conditions”. The controlled-load service is intentionally minimal, in that there are no optional functions or capabilities in the specification. The service offers only a single function. It is better than best effort, but cannot provide strictly bounded service as guaranteed service.

The controlled-load service can be implemented by using evolving scheduling and admission control algorithms. The implementations are highly efficient in the use of network resources.

Guaranteed service guarantees that datagrams will arrive within the guaranteed delivery time and will not be discarded due to queue overflows, provided the flow's traffic stays within its specified traffic parameters. It is similar to emulate a dedicated virtual circuit. This service is intended for applications which need a firm guarantee that a datagram will arrive no later than a certain time after it was transmitted by its source. For example, some audio and video "play-back" applications are intolerant of any datagram arriving after their play-back time. Applications that have hard real-time requirements will also require guaranteed service.

Guaranteed service does not attempt to minimize the jitter (the difference between the minimal and maximal datagram delays); it merely controls the maximal queueing delay. Because the guaranteed delay bound is a firm one, the delay has to be set large enough to cover extremely rare cases of long queueing delays. Several studies have shown that the actual delay for the vast majority of datagrams can be far lower than the guaranteed delay. Therefore, authors of playback applications should note that datagrams will often arrive far earlier than the delivery deadline and will have to be buffered at the receiving system until it is time for the application to process them.

Guaraneteed service represents one extreme end of delay control for networks. Most other services providing delay control provide much weaker assurances about the resulting delays. In order to provide this high level of assurance, guaranteed service is typically only useful if provided by every network element along the path (i.e. by both routers and the links that interconnect the routers). Moreover, as described in the Exported Information section, effective provision and use of the service requires that the set-up protocol or other mechanism used to request service provides service characterizations to intermediate routers and to the endpoints.

Integrated Services routers uses admission control and resource allocation method to offer QoS guarantee. A token-bucket model is used to characterize the input/output queueing algorithm. It can smooth the flow of outgoing traffic. The IntServ parameters include (qodwhitepaper, 1999):

Token rate (r): The continually sustainable bandwidth (bytes/second) requirement for a flow. It represents the average data rate into the bucket, and the target shaped data rate out of the bucket.

Token-bucket rate (b) : the extent to which the data rate can exceed the sustainable average for short periods of time, or the amount of data sent cannot exceed rT+b (where T is any time period).

Peak rate (p): It is the maximum send rate (bytes/second) if known and controlled. At any time period (T), the amount sent data cannot exceed M+pT.

Minimum policed size (m): The size (byte) of the smallest packet (data payload only) can be generated by the sending application. The size m is not an absolute number. If the percentage of small packets is small, the number m should increased to reduce the overhead estimate. All packets smaller than m are treated as size m.

Maximum packet size (M): The biggest size of a packet (bytes). The M is absolute number. Any packets (size > M) are considered out of spec and may not receive QoS controlled service.

3.1.2 RSVP

For offering IntServ, a way to communicate the application's requirements to network elements along the path and to convey QoS management information between network elements and the application must be provided. A resource reservation setup protocol called RSVP (rfc2205, 1997) is implemented for this purpose. It is a signaling protocol that can provide reservation setup and control to enable the integrated services by using a variety of QoS control, a variety of setup mechanisms.

A simplified RSVP working flow is shown in Figure 5

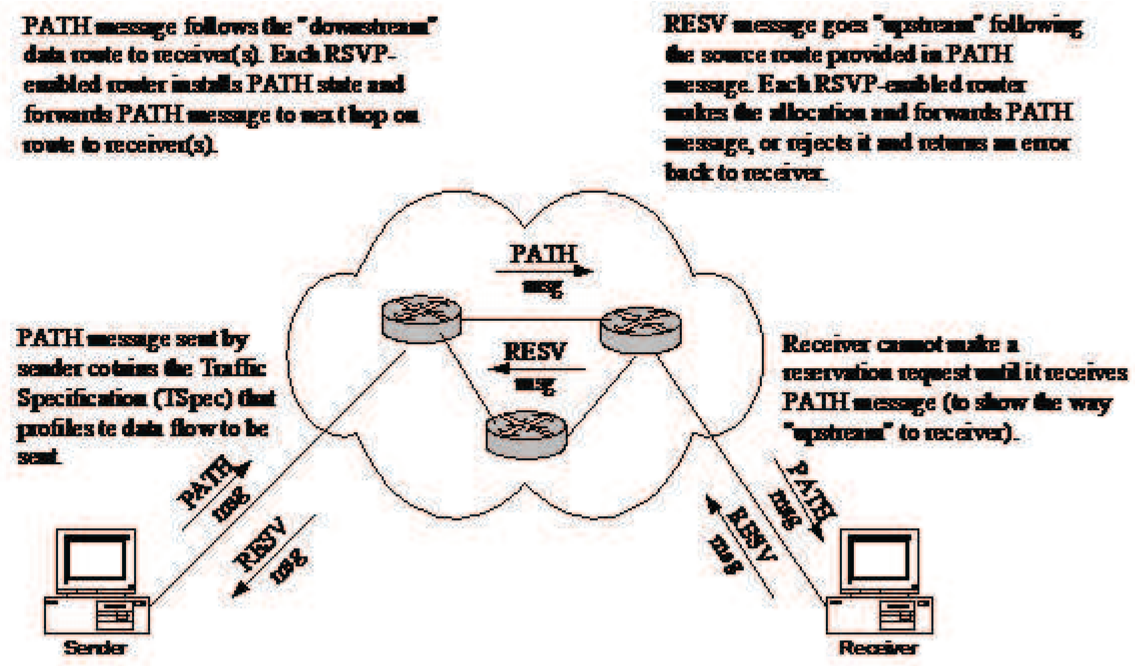

PATH message follows the "downstream" data route to receiver(s). Each RSVP-enabled router installs PATH state and forwards PATH message to next hop on route to receiver(s).
RESV message goes "upstream" following the source route provided in PATH message. Each RSVP-enabled router makes the allocation and forwards PATH message, or rejects it and returns an error back to receiver.
PATH msg
RESV msg
PATH message sent by sender cotains the Traffic Specification (TSpec) that profiles te data flow to be sent.
Receiver cannot make a reservation request until it receives PATH message (to show the way "upstream" to receiver).
PATH msg
RESV msg
PATH msg
RESV msg
Sender
Receiver

Fig. 5. RSVP setup flow

1. When a sender wants to set up a traffic link, it will generate the traffic specification (TSpec), which describes data traffic, such as upper/lower bounds of bandwidth, delay, and jitter. Then RSVP sends out a PATH message containing TSpec to the receiver(s) (unicast or multicast). Along the route, each RSVP-enabled router trigger a "path-state" that includes the previous source address of the PATH message.
2. After receiver receives the PATH message, the receiver sends a RESV message "upstream" to make a resource reservation. The RESV message includes a request specification (RSpec) which indicates what type of IntServ required âĂŞ either Controlled Load or Guaranteed, a filter specification (filter spec) (indicating e.g. the transport protocol and port number). The RSpec and filter spec represent a flow-descriptor that RSVP routers use to identify each reservation.
3. Along the RSVP upstream, RSVP routers use the admission control to authenticate the resource reservation request and allocate the necessary resources when the routers receive the RESV message. If the request cannot be met (due to no enough bandwidth or authorization failure), the RSVP router returns an error back to the receiver. If the request is accepted, the router forwards the RSVP message to the next router.
4. When the last router receives the RESV message and accepts the request, it sends out a confirmation message back to the receiver.
5. There is an explicit tear-down process for a reservation when sender or receiver terminate a RSVP session.

3.1.3 DiffServ

The Integrated Services/RSVP model relies upon traditional datagram forwarding in the default case, but allows sources and receivers to exchange signaling messages which establish additional packet classification and forwarding state on each node along the path between them (rfc1633, 1994). In the absence of state aggregation, the amount of state on each node scales in proportion to the number of concurrent reservations, which can be potentially large on high-speed links. This model also requires application support for the RSVP signaling protocol. Differentiated service is a simple method by classifying services of different applications (rfc2475, 1998). Currently there are two standard per hop behaviour (PHBs) define two traffic classes:

- Expedited Forwarding (EF): Has a single codepoint (DiffServ value). Ef minimize delay and jitter and provides the highest level of aggregate quality of service. Any traffic that exceeds the traffic profile is discarded (?). EF class offers a low jitter, low delay service. UserâĂŹs traffic cannot exceed the agreed peak rate. Otherwise, the packets will be discarded.
- Assured Forwarding (AF): Has four classes and three drop-precedence within each class (a total of twelve codepoints). Excess AF traffic is not delivered with as high probability as the traffic "within profile", which means it may be demoted but not necessarily dropped (?). The AF class is suitable for delay-tolerant applications. The guarantee just implies that the better QoS class will give a better performance than the low-level QoS class. Network operator can define their own per-hop behavior.

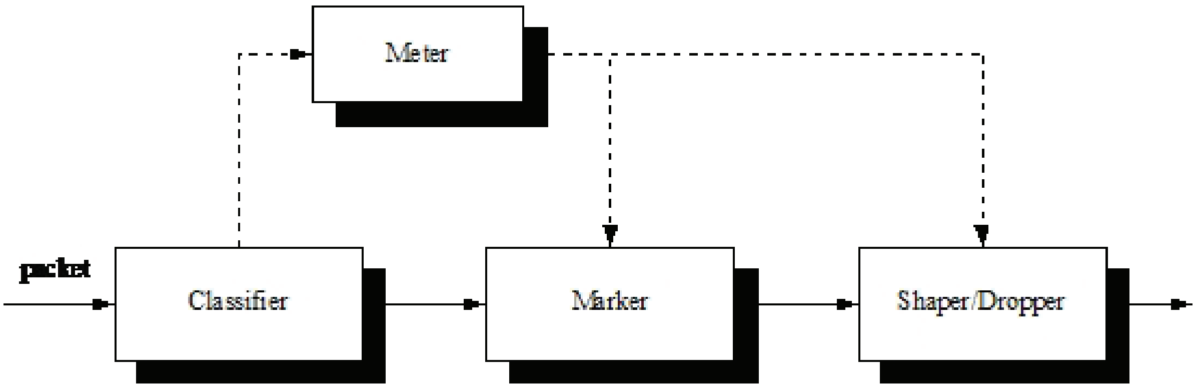


Fig. 6. DffServ architecture

DiffServ offers a simple QoS management method without signaling mechanism. The DiffServ architecture is shown in Figure 6. It includes classifier and traffic conditioner. A traffic conditioner contains the following elements: meter, maker, shaper/dropper. The differentiated services architecture is based on a simple model where traffic entering a network is classified and possibly conditioned at the boundaries of the network, and assigned to different behavior aggregates. Each behavior aggregate is identified by a single DiffServ codepoint (DSCP). Within the core of the network, packets are forwarded according to the per-hop behavior associated with the DiffServ codepoint.

When a traffic flow enters a DiffServ network, the flow is selected by a classifier, which steers the packets to a logical instance of a traffic conditioner. A meter is used to measure the traffic flow agains a traffic profile. A meter is used (where appropriate) to measure the traffic stream against a traffic profile. The state of the meter with respect to a particular packet (e.g., whether it is in- or out-of-profile) may be used to affect a marking, dropping, or shaping action. When packets exit the traffic conditioner of a DS boundary node the DiffServ codepoint of each packet must be set to an appropriate value.

3.1.4 MPLS

MPLS is a key development in IETF that will add a number of essential capabilities to today's best effort IP networks, including

- Traffic Engineer, enhancing overall network utilization by creating a uniform or differentiated distribution of traffic throughout the network.
- Providing traffic with different Classes of Service (CoS)
- Providing traffic with different Quality of Service (QoS)
- Supporting network scalability, providing IP based Virtual Private Networks (VPN)

MPLS borrows the idea from ATM switching. It remains independent of the Layer-2 and Layer-3 protocols. Besides IP, other network protocols (such as IPX, ATM, PPP or Frame-Relay) also can work with MPLS. MPLS resides on routers. When a packet flow enters a edge router of the MPLS domain, all packets are marked to clarify priority with a fixed-length label (20 bits label). The label identifies the packets routing information in this MPLS network, also define the quality of service for the packets.

A MPLS domain includes label edge routers (LERs) and label switching routers (LSRs). The route taken by an MPLS-labeled packet is called the label switched path (LSP). LST is a high-speed router in the core of a MPLS network, which participates in the establishment of LSPs. LER is a router that operates at the edge of a MPLS network. It is used to assign and remove labels when packets enter or exit the MPLS network.

MPSL is similar to DiffServ because it also marks traffic at ingress of a MPLS network, and un-marks at egress gate. However, MPLS marking is used to decide the next hop router while DiffServ marking is used to determine priority in route itself.

3.2 QoS management functions for end-to-end IP QoS

This section describes how to provide Quality of Service in UMTS for the end-to-end services through the TE/MT local bear service, GPRS bearer service and external bearer service shown in the Fig. 1. To provide end-to-end IP QoS, it is necessary to manage the QoS within each domian. An IP BS Manager is used to control the external IP bearer service. Due to the different techiniques used within the IP network,this communicates to the UMTS BS manager through the Translation function.

At PDP context setup the user shall have access to one of the following alternatives, basic GPRS IP connectivity service or enhanced GPRS based services. To enable coordination between events in the application layer and resource management in the IP bearer layer, a logical element, the Policy and Charging Rules Function (PCRF), is used as a logical policy decision element which will be detailed in section 4. It is also possible to implement a policy decision element internal to the IP BS Manager in the GGSN. While interworking with the external network, the RSVP, DiffServ, MPLS will be used.

QoS management functions is shown in Fig. 7 which describes how to control the external IP bearer services and how they relate to the UMTS bearer service QoS management entity.

IP BS Manager uses standard IP mechanisms to manage the IP bearer services. These mechanisms may be different from mechanisms used within the UMTS, and may have different parameters controlling the service. When implemented, the IP BS Manager

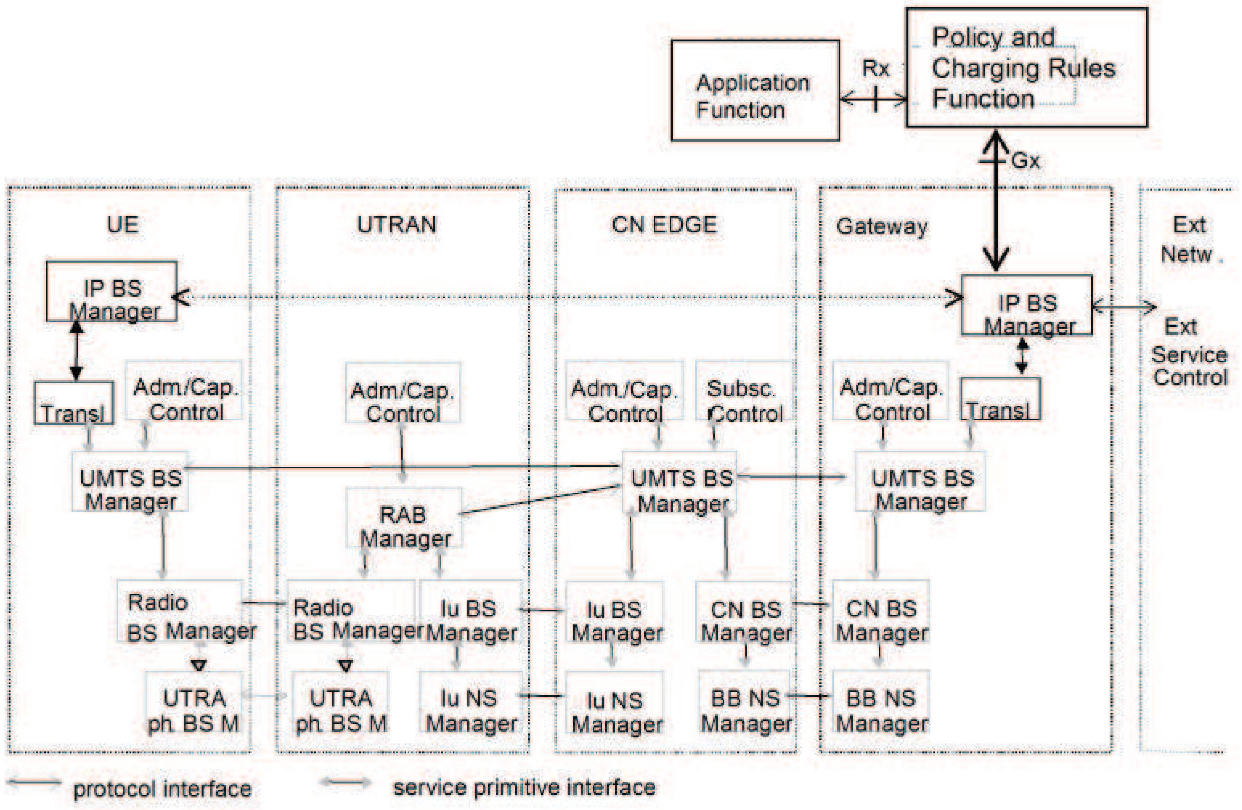


Fig. 7. QoS management functions for end-to-end QoS in UMTS

may include the support of DiffServ Edge Function and the RSVP function. The Translation/mapping function provides the inter-working between the mechanisms and parameters used within the UMTS bearer service and those used within the IP bearer service, and interacts with the IP BS Manager. In the GGSN, the IP QoS parameters are mapped into UMTS QoS parameters, where needed. In the UE, the QoS requirements determined from the application layer (e.g., SDP) are mapped to either the PDP context parameters or IP layer parameters (e.g., RSVP). If an IP BS Manager exists both in the UE and the Gateway node, it is possible that these IP BS Managers communicate directly with each other by using relevant signalling protocols. The required options in the table define the minimum functionality that shall be supported by the equipment in order to allow multiple network operators to provide interworking between their networks for end-to-end QoS. Use of the optional functions listed below, other mechanisms which are not listed (e.g. over-provisioning), or combinations of these mechanisms are not precluded from use between operators. The IP BS Managers in the UE and GGSN provide the set of capabilities for the IP bearer level as shown in table 2. Provision of the IP BS Manager is optional in the UE, and required in the GGSN.

Capability	UE	GGSN
DiffServ Edge Function	Optional	Required
RSVP/IntServ	Optional	Optional
IP Policy Enforcement Point	Optional	Required

Table 2. IP BS Manager capability in the UE and GGSN

4. Policy based QoS management - IP QoS for IMS

This section will provide an overview of policy based QoS management in UMTS IMS. Although the UMTS packet switched (PS) domain can support IP QoS enabled multimedia applications, there are many ways of establishing QoS guaranteed IP multimedia session through a signaling protocol before it can map and reserve the equivalent amount of QoS resources along the data path in the PS domain. In order to support interoperation among UMTS network providers, the IP multimedia Subsystem (IMS) is standardized by the 3GPP to serve Session Initiation Protocol (SIP) signaled IP multimedia services over the UMTS PS domain.

The central problem of providing consistent end-to-end IP QoS services is the difficulty of configuring the network devices like routers and switches to handle packet flows in a manner that satisfies the requested QoS requirements. Policy-based QoS management is used to control QoS resources in the UMTS IMS.

4.1 Introduction to policy-based QoS network

4.1.1 The need for policy based network

There is a consistent effort to implement new IP multimedia services in UMTS. While the IP based network is well suited for packet data transfer, providing consistent end-to-end IP QoS services is the difficulty of configuring the network devices like routers and switches to handle packet flows in a manner that satisfies the requested QoS requirements. This problem is especially acute when the end-to-end data path of an IP QoS session crosses multiple administrative domains managed by different operators. Although the operators agree on the QoS requirements of a particular set of IP services, they may not configure their network devices in the same way to implement the services due to differences in the network topologies, QoS mechanisms available in the network devices and non-technical management requirements. Thus, there is a need to create a solution that permits network operators, including UMTS network operators, to easily configure their networks to implement consistent IP QoS services without dealing with the complexity of their networks.

Policy-based Networking (PBN) is a novel approach to configure myriad network devices in an administrative domain to implement a set of IP QoS services. Policy-based network will allow the network operator to define, in a succinct and organized fashion, operator policies that automatically effect change on specific equipment in the network environment. The end result is that the end-to-end network performance will meet the general expectations of UMTS service provider environment.

4.1.2 What is policy?

A policy is a set of business rules that guide and determine how to manage network resources. The basic concept is that policy rule(s) describe how network to act when specific condition(s) happen. "Policy" can be defined from two perspectives: (POLICYTERM, 2001). - A definite goal, course or method of action to guide and determine present and future decisions. "Policies" are implemented or executed within a particular context (such as policies defined within a business unit). - Policies as a set of rules to administer, manage, and control access to network resources. [RFC3060] Note that these two views are not contradictory since individual rules may be defined in support of business goals.

Policy can be represented at different levels, ranging from business goals to device-specific configuration parameters. Enforcement of policy ensures that business rules are always followed. Policy rule is a basic building block of a policy-based system. It is the binding of a set of actions to a set of conditions - where the conditions are evaluated to determine whether the actions are performed. [RFC3060] A condition is a set of expressions or objects used to determine whether a given policy rule's action should be performed. A condition answers the question, "when and where do we enforce a policy?" An action defines what to be done to enforce a policy rule, when the conditions of the rule are met. Policy actions may result in the execution of one or more operations to affect and/or configure network traffic and network resources. An action answers the question, "what must be done to enforce a policy?"

A policy also defines how the network's resources are to be allocated among its clients. Clients can be individual users, departments, host computers, or applications. Resources can be allocated based on time of day, client authorization priorities, availability of resources, and other factors. How resources are allocated can be static or dynamic (based on variations in traffic). Policies are created by network managers and stored in a repository. During network operation, the policies are retrieved and used by network management software to make decisions.

4.1.3 Policy framework & architecture

The network operators negotiate Service Level Agreements (SLAs) that describe the sets of IP QoS services that they have mutually contracted to provide. Individual operators will then transform the QoS requirements specified in the SLAs into sets of policy rules that will be applied to their network domains to implement the contracted IP QoS services. The IETF has defined a policy framework (RFC2753, 2001) as shown in Figure 8 to transform the sets of policy rules to network device configurations in an administrative domain. The sets of policy rules are stored in the Policy Repository through the Policy Management Tool. The Policy Decision Point (PDP) retrieves the appropriate policy rules from the Policy Repository in response to policy events that are triggered by the contracted IP QoS services, e.g., the reception of an RSVP message by the Policy Enforcement Point (PEP). It translates the acquired policy rules into a set of QoS mechanism configuration actions that is communicated to the PEP as policy decisions. The PEP then executes the actions spelt out in the supplied decisions to handle the triggering policy events in accordance with the requested IP QoS services. Alternatively, the retrieved policy rules may be returned to the PEP, which is capable of translating them into configuration actions. These policy rules can be cached in the PEP so that similar future triggering policy events can be serviced locally without further interactions with the PDP.

Outsourcing and Provision Model in PBN

> There are two main models for policy management: outsourcing and provisioning. The outsourcing model assumes there is a signaled event in the Policy Enforcement Point (PEP) that must be resolved based on policy criteria. The PEP outsources the decision-making to an external policy decision point (PDP). This outsourcing model is sometimes referred to as "Pull" mode, or "reactive" mode, since the PEP pulls policy decisions from the PDP, while the PDP responds according to the PEP events.
>
> The provisioning model is almost the mirror image of the outsourcing model. In this system, the PDP predicts future configuration needs, and proactively provisions resources accordingly. In other words, rather than responding to PEP events, the PDP prepares

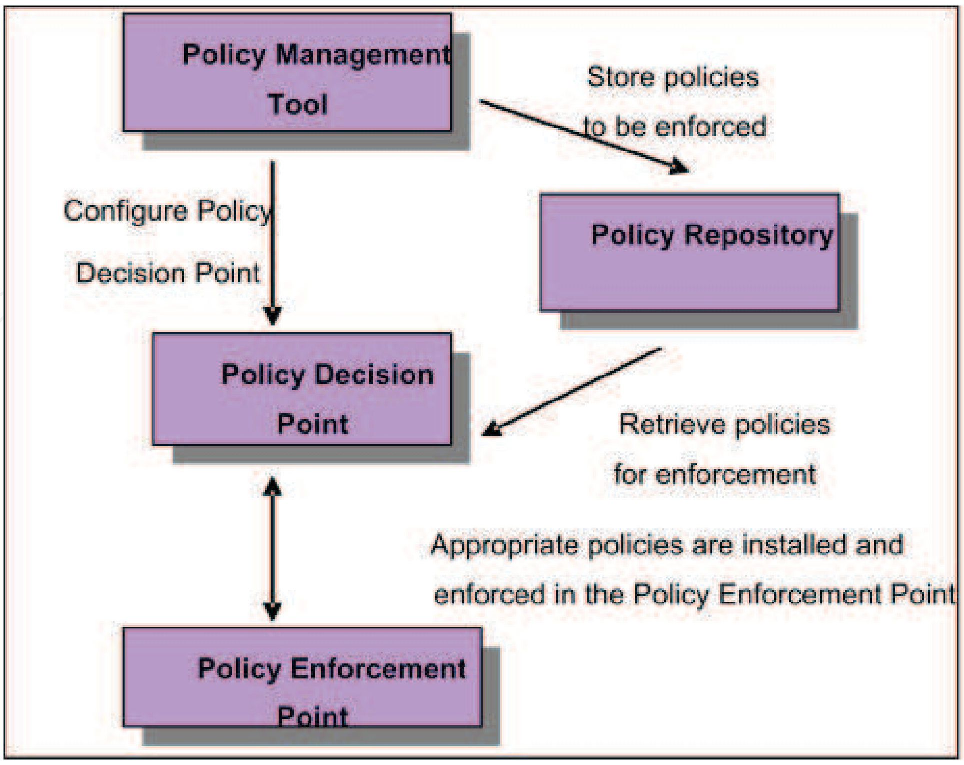

Policy Management Tool
Store policies to be enforced
Configure Policy Decision Point
Policy Repository
Policy Decision Point
Retrieve policies for enforcement
Appropriate policies are installed and enforced in the Policy Enforcement Point
Policy Enforcement Point

Fig. 8. A PBN architecture that is derived from the policy framework specified by the IETF

and "pushes" configuration information to the PEP. This takes place as a result of external events (unrelated to the PEP) such as change of applicable policy, time of day, expiration of account quota, or information from third party (non-PEP) signaling.

Both models employ policy servers as the PDP to control the network devices that enforce the policy (i.e. PEPs). PBN also offers a policy repository for storing policy information accessed by the PDPs in the system. To communicate policy information between PDPs and PEPs, the COPS policy protocol is engaged. Additionally, the LDAP protocol functions to access the policy repository.

Policy Decision Point (PDP)

The PDP is the PBN component that directly controls the network devices or policy enforcement points (see next section). Functionally, the PDP handles policy information that has been entered into the PBN management system. The policy data used by the PDP can either be obtained in real-time upon entry into the management console, or from the policy repository on an as-needed basis. The function of the PDP involves retrieving policy, interpreting policy, detect policy conflicts, receiving policy decision requests from PEPs, determining which policy is relevant, applying the policy and returning the results. It also sends policy elements the PEP.

Policy Enforcement Point (PEP)

Network devices that receive and enforce the decisions from the PDP are referred to as PEPs. In both outsourcing and provisioning policy management models, PEPs receive policy decisions and enforce them at the packet level as data passes through the devices.

4.2 Policy framework in UMTS IMS

To support IP based multimedia services, the IP Multimedia Subsystem (IMS) is introduced in the 3GPP Release 5 specifications. It provisions IP based multimedia services as an extension of the UMTS PS domain (Figure 9). The added IMS functionalities are control functionalities; the user data traffic is still carried by the PS domain. The main advantage of the IMS is that it offers operators a scalable service platform on which new services can be developed rapidly in a flexible way, without requiring any change to the PS domain.

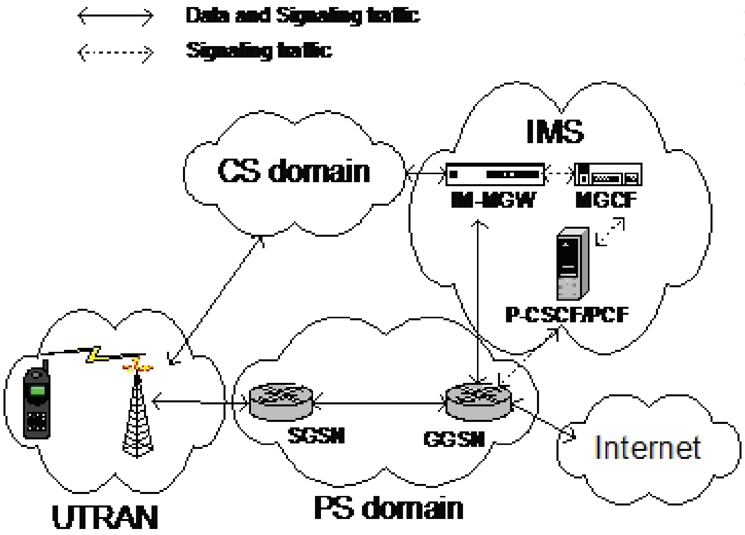


Fig. 9. A simple UMTS network with IMS

Having put in place the functionalities to handle IP multimedia calls, the next big challenge is to ensure that sufficient QoS resources are provided to authorized users in the UMTS network. A policy-based QoS solution is adopted by the 3GPP for this purpose.

As mentioned in section 4.1.3, the reference model of a policy-based network consists of two main elements, the PDP and the PEP (RFC2753, 2001). PEPs often reside in policy aware network nodes that carry out actions stipulated by policy rules. The actions taken are based on the decisions of a PDP, which retrieves the policy rules from a repository. The PDP is the final authority, which the PEP needs to refer to for actions to be taken.

In the IMS, the Policy and Charging Rules Function (PCRF) (3G23203, 2008) plays the role of the PDP and online charging and offline charging functions, the Policy and Charging Enforcement Functions plays the role of the PEP. Policy charging and rules function (PCRF) is the node designated in real-time to determine policy rules in a multimedia network. As a policy tool, the PCRF plays a central role in WCDMA networks. Unlike earlier policy engines that were added on to an existing network to enforce policy, the PCRF is a software component that operates at the network core and efficiently accesses subscriber databases and other specialized functions, such as a charging systems, in a scalable, reliable, and centralized manner. The PCRF as the part of the network architecture that aggregates information to and from the network, operational support systems, and other sources (such as portals) in real time, supporting the creation of rules and then automatically making intelligent policy decisions for each subscriber active on the network. Such a network might offer multiple services, quality of service (QoS) levels, and charging rules. In this chapter, we will focus on policy based management functions.

The PCRF communicates with the PCEF via the Gx interface (3G29212, 2008). It allows two modes of operation. In the "push" mode, the PCRF initiates communication with the PCEF and sends the PCEF its decision. In the "pull" mode, the PCEF initiates communication with the PCRF to request a decision for a particular IP flow. The Gx interface and the protocol used for communication on the interface are described in the following.

Figure 10 depicts the relationship between these entities. In the following subsections, each of these network elements will be described.

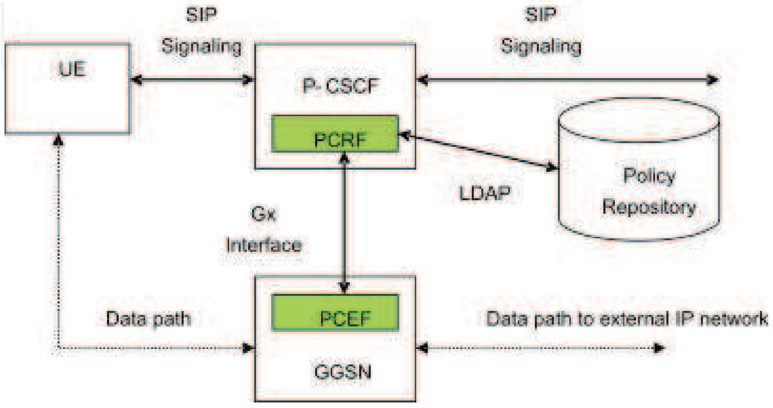


Fig. 10. Policy architecture in UMTS IMS

PCRF in Proxy-CSCF

During the establishment of a SIP session, a P-CSCF is the first contact point in the IMS domain for a UE [3G23228]. Hence it is the natural place to authorize the usage of network resources such as the bandwidth requested by the UE. The QoS requirements of the UE are carried in the Session Description Protocol (SDP) description within a Session Initiation Protocol (SIP) message. Besides the QoS requirements in the SDP description, the PCRF also examines the source and destination IP addresses and port numbers in its decision-making. The PCRF refers to the policy rules, which are generally stored in a policy repository, governing the local domain. It then generates an authorization token that uniquely identifies the SIP session across multiple Packet Data Protocol (PDP) contexts terminated by a GGSN. This token is sent to the UE via SIP messages so that the UE can use it to identify the associated session flows to the PCEF in the GGSN in subsequent transmission of IP packets. This mechanism is consistent with the IETF specification on supporting media authorization in the SIP protocol [RFC3313]. The flow of events in session set-up is described in Section 4.6.

PCEF in Gateway GPRS Support Node (GGSN)

In the PS domain, a GGSN maintains connectivity to other packet switched external networks such as the Internet. From the service point of view, the GGSN controls which IP flows are permitted into the external IP network by policing the IP packets based on their source and destination IP addresses and port numbers [3G23228]. As such, it is logical to embed the PCEF in the GGSN. The role of the PCEF is to ensure that only authorized IP flows are allowed to use network resources that have been reserved and allocated to them.

The policy enforcement function in the GGSN is called a "gate". A gate comprises a packet classifier, a traffic meter, and the relevant packet handling mechanisms for packets that have been matched by the packet classifier. When an IP flow is authorized by the PCRF to use the specified network resources, the PCEF opens the "gate" for the flow and effectively commits the network resources to the flow by allowing it to pass through the packet handling mechanisms (i.e., policing or marking). On the other hand, if an IP flow is not permitted by the PCRF to use the requested resources, the PCEF closes the âĂIJgateâĂİ and drops the IP packets of the flow. This process is called policy-based admission control. It ensures that an IP flow is only allowed to use resources that have been approved by the policy rules. The above process takes place at the IP bearer service (BS) level. The translation/mapping function within the GGSN will map this resource information into the format used by the admission control function at the UMTS BS level.

The PCEF may store decisions in a local policy decision point, thus allowing the GGSN to make the admission control decisions without additional interactions with the PCRF. This will reduce the traffic over the Gx interface and lessen the processing load on the PCRF.

4.3 Policy-based QoS delivery: an example of policy based call control

There are several reasons why a policy-based QoS framework is adopted for the UMTS. Policy-based QoS control allows network operators to configure their network devices easily. It provides a high level view of the network devices and allows the automated translation of business level policies to suitable information for configuring network devices.

UMTS requires a strict authorization of users so that the network resources are not abused. Once authorized and approved, the UMTS must guarantee that the resources are made available to the legitimate users. If these requirements are not met, these users may be denied the use of the resources, leading to dissatisfaction with the quality of service provided. To ensure that this is not the case, all IP multimedia calls must go through the following steps:

1. Authorization of resources;
2. Reservation of resources. This is to make sure that the resources are available when the "phone" rings;
3. Once the called party picks up the "phone", the network resources reserved previously are committed. The charging process is then triggered.

In all these steps, policy rules are used in approving the requests, and the PCRF is the sole approving authority. By changing the policy rules in the PCRF, a network operator can alter the IP multimedia services it offers to its subscribers without having to know the details of its network configuration and the types and mechanisms of the network devices.

To meet the above requirements, two procedures are needed for the establishment of an IP multimedia session in addition to the normal GPRS bearer establishment procedures. These procedures are Authorize QoS Resources and Approval of QoS Commit (3G29212, 2008). Similarly, the procedures, Removal of QoS Commit and Revoke Authorization of QoS Resources, are carried out to reverse the authorization and commitment of QoS resources when an IP multimedia session is terminated. The following provides an overview of the session set-up procedures, in particular, the emphasis is on the additional procedures introduced by the service-based local policy.

4.4 Session establishment procedures

The establishment of an IP multimedia service session with policy control differs from that without policy control in that additional steps are taken to check the policy rules for a decision on whether to grant or deny the required network resources to the session. As the signaling messages used to set up the session take a different path from that used for the data flow, an authorization token and a flow identifier are used to associate the session with its IP data flow (UMTSGO01, 2001). The GGSN, which is located on the data path, relies on this binding information to enforce the policy rules on the IP data flows.

Figure 11 depicts the sequence of events that take place during the establishment of an IP multimedia service session. Note that a number of signaling messages have been omitted for clarity. The events are described in the following paragraphs:

Steps 1-5: The UE sends a session set-up request (i.e., SIP INVITE) to the P-CSCF indicating, among other things, the media streams to be used in the session. This message is routed to the called party via a number of other CSCFs (viz., the caller and callee S-CSCFs) along the signaling path. The S-CSCFs perform the appropriate session control services for the UEs. In particular, they maintain a session state that is needed by the network operator to support the requested service.

Steps 6-14: The called party responds with a provisional SIP 183 response message. This message is routed to the calling party via the same CSCFs along the (reversed) signaling path. When the callee P-CSCF receives this message, it examines the SDP description within the message to determine the QoS parameters requested for the session. The P-CSCF sends the necessary information in this SIP message (e.g., the bandwidth, IP addresses and ports, etc.) to the PCRF for authorization of the session request. If the policy permits, the PCRF responds with an authorization token that can be used to identify the authorized session and resources. The P-CSCF includes the token in the response (SIP 183 message) and forwards it to the callerâĂŹs UE. A similar process is carried out at the caller P-CSCF when it receives the SIP 183 message. This process of authorization by the PCRF and the generation of a token is called "Authorize QoS Request".

Steps 15-22: In between steps 14 and 15, other message exchanges take place between the caller and the callee. However, these are not important in this particular example and are omitted for clarity. The caller's UE starts the resource reservation by sending a PDP Context Activation Request to the GGSN. The authorization token and the flow identifier(s) from the PCRF are included to identify the IP data flow(s) of the session. When the GGSN receives the PDP Context Activation Request, it sends a policy decision request to the PCRF to determine whether the resource reservation request should be accepted. The PCRF uses the token in the message to correlate the request for resources with the media authorization previously granted to the session. The PCRF then sends a decision to the GGSN. If the PCRF approves the resource reservation, the GGSN sends a PDP Context Activation Response to the UE indicating that the resource reservation has been completed. A similar process takes place at the callee's end.

Steps 23-31: In between steps 22 and 23, there are other events, e.g., 180 Ringing, that take place. These events are omitted to prevent cluttering Figure 3-22. When the callee answers the call, a SIP 200 OK message is sent towards the caller. When the SIP 200 OK reaches the P-CSCF, it will approve the QoS commitment by sending a decision to the GGSN. Upon receiving this message, the GGSN opens the gate, thereby effectively permitting the IP data flow to use the resources reserved previously. Once this is done, the GGSN responds to the PCRF with a report on the status of the session. A similar process takes place at the caller's end. When this entire process is completed, the proper resources on the data path have been reserved and committed to the session.

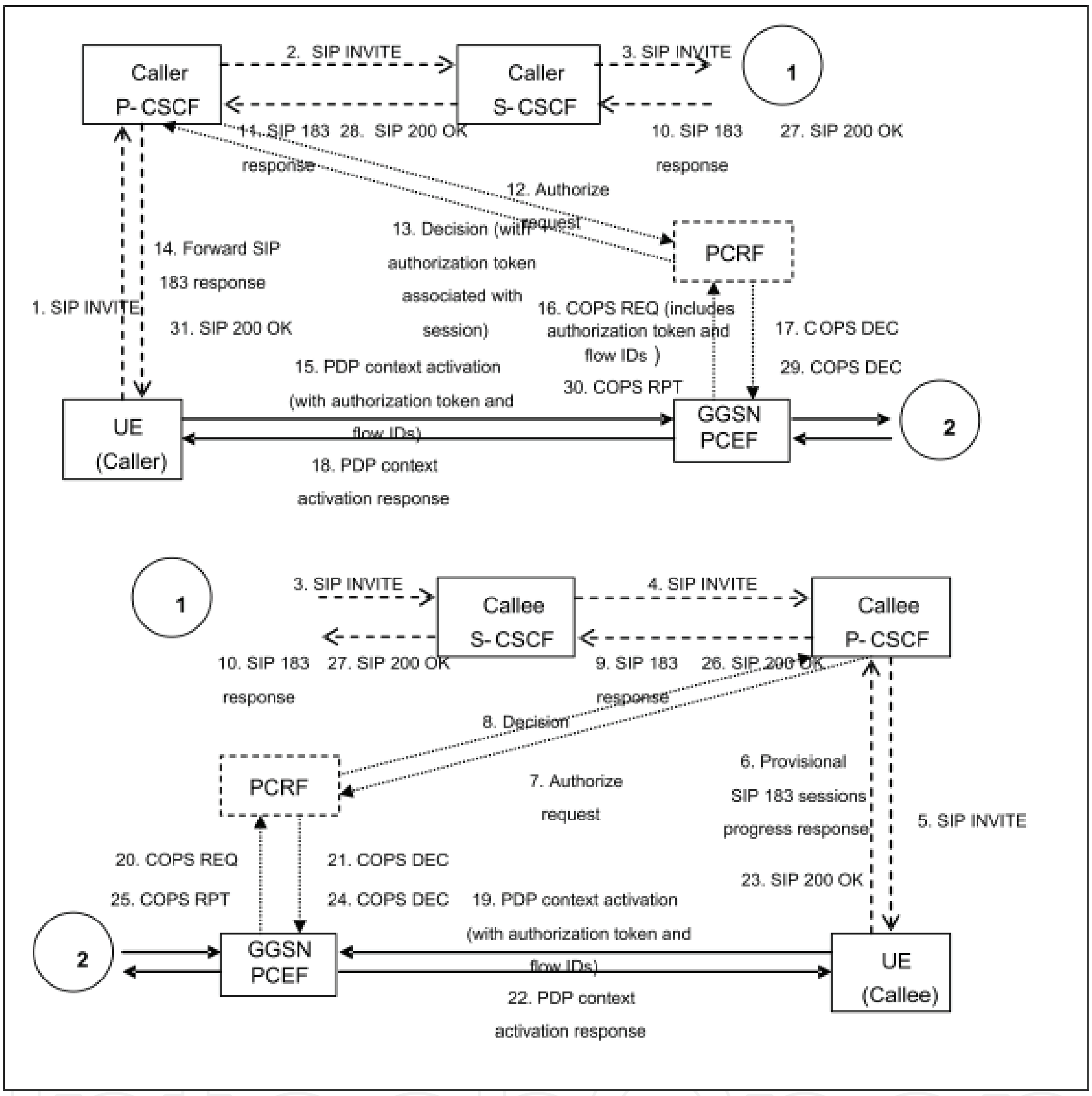


Fig. 11. Session authorization mechanism in a UE-to-UE session establishment process

5. References

Nortel White Paper (2002). Benefits of Quality of Service (QoS) in 3G Wireless Internet, Nortel Networks.

Sudhir Dixit, Yile Guo, Zoe Antoniou (2001). Resource management and quality of service in third-generation wireless network, *IEEE Communication Magazine*, Feb. 2001, pp.125-133.

Sotiris I. Maniatis, Eugenia G. Nikolouzou, & Iakovos S. Venieris (2002). QoS issues in the converged 3G wireless and wired networks, *IEEE Communications Magazine*, Aug. 2002, pp.44-53.

3GPP TS 23.107 (V9.2.0) (2011). Quality of Service(QoS) Concept and architecture (Release 9).

S.Shenker, C.Partridge, R.Guerin (1997), Specification of Guaranteed Quality of Service, *RFC2212*, Sept.1997.

J. Wroclawski (1997), Specification of the Controlled-Load Network Element Service, *RFC2211*, Sept. 1997.

S.Shenker, J.Wroclawski (1999), Network Element Service Specification Template, *RFC2216*, Sept. 1999.

White paper (1999), QoS Protocol & Architecture, *www.qosforum.com*, July, 1999.

R. Braden, D. Clark, S. Shenker, Integrated Services in the Internet Architecture: an Overview, *RFC 1633*, June 1994.

Braden, B., Ed., et. al., Resource Reservation Protocol (RSVP) - Version 1 Functional Specification, *RFC 2205*, September 1997.

S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, An architecture for differentiated services, *RFC 2475*, Dec. 1998.

V. Jacobson, K. Nichols, K. Poduri, An expedited forwarding PHB, *RFC 2598*, June, 1999.

J. Heinanen, F. Baker, Weiss, J. Wroclawski, Assured forwarding PHB group, *RFC 2597*, June 1999.

White paper, Introduction to QoS Policies, *www.qosforum.com*, 1999

Survey on Policy-based networking, *INTAP*.

A. Westerinen, etc. Terminology for Policy-Based Management, <*draft-ietf-policy-terminology-04.txt*>, July, 2001.

B.Moore, E. Ellesson, J. Strassner and A. Westerinen, Policy Core Information Model – Version 1 Specification, *RFC 3060*, IETF, Feb. 2001.

M. Handley, et al., SIP: Session Initiation Protocol, *Internet draft (work in progress)*, <*draft-ietf-sip-rfc2543bis-09.txt*>, Feb. 2002

3GPP TS 29.212 (version 8.3.0), Policy and Charging Control Over Gx Reference Point (Rel 8), Dec. 2008.

3GPP TS 23.228 (version 8.3.0), IP Multimedia Subsystem- Stage 2 (Rel 8), June 2008.

3GPP TS 23.203 (version8.3.0),Policy and Charging Control Architecture (Rel. 8), 2008.

W. Marshall, et al., Private SIP Extensions for Media Authorization, *RFC 3313*, Nov. 2001

D. Durham, J. Boyle, R. Cohen, S. Herzog, R. Rajan and A. Sastry, The COPS (Common Open Policy Service) Protocol, *RFC 2748*, IETF, Jan. 2000.

R.Yavatkar, D.Pendarakis, R.Guerin, A Framework for Policy-based Admission Control, *RFC 2753*, IETF, Jan. 2000.

R. Atkinson, Security Architecture for the Internet Protocol, *RFC 2401*, IETF, Aug. 1995

T. Dierks and C. Allen, The TLS Protocol Version 1.0, *RFC 2246*, IETF, Jan. 1999

K. Chan, D. Durham, S. Gai, S. Herzog, K. McCloghrie, F. Reichmeyer, J. Seligson, A. Smith and R. Yavatkar, COPS Usage for Policy Provisioning, *RFC 3084*, Mar. 2001

L-N. Hamer, K. Chan, H. Syed, H. Shieh and R. Zwart, COPS-PR for outsourcing in UMTS: UMTS Go PIB, *draft-hamer-rap-cops-umts-go-00, IETF*, Nov. 2001

B. Moore, L. Rafalow, Y. Ramberg, Y. Snir, J. Strassner, A. Westerinen, R. Chadha, M. Brunner and R. Cohen, Policy Core Information Model Extensions, *draft-ietf-policy-pcim-ext-06*, Nov. 2001

J. Jason, L. Rafalow and E. Vyncke, IPsec Configuration Policy Model, *draft-ietf-ipsp-config-policy-model-03, IETF*, July 2001

Y. Snir, Y. Ramberg, J. Strassner, R. Cohen and B. Moore, Policy QoS Information Model, *draft-ietf-policy-qos-info-model-04, IETF*, Nov. 2001

S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang and W. Weiss, An Architecture for Differentiated Service, *RFC 2475, IETF*, Dec. 1998

R. Braden, D. Clark and S. Shenker, Integrated Services in the Internet Architecture: an Overview, *RFC 1633, IETF*, June 1994

R. Braden, Ed., L. Zhang, S. Berson, S. Herzog and S. Jamin, Resource ReSerVation Protocol (RSVP) – Version 1 Functional Specification, *RFC 2205, IETF*, Sept. 1997

B Moore, D. Durham, J. Strassner, A. Westerinen, W. Weiss and J. Halpern, Information Model for Describing Network Device QoS Datapath Mechanisms, *draft-ietf-policy-qos-device-info-model-06, IETF*, Nov. 2001

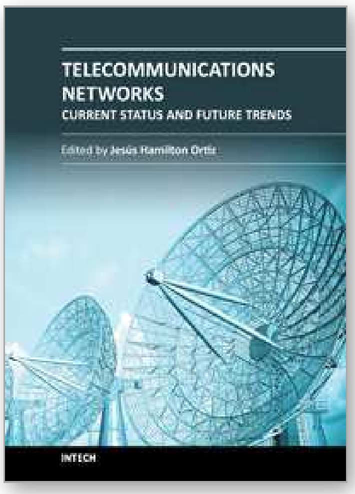
M. Wahl, T. Howes and S. Kille, Lightweight Directory Access Protocol (v3), *RFC 2251, IETF*, Dec. 1997

G. Good, The LDAP Data Interchange Format (LDIF) - Technical Specification, *RFC 2849, IETF*, June 2000

Ebata, M. Takihiro, S. Miyake, et al., Interdomain QoS Provisioning and Accounting, *INET 2000*, Yokohama, Japan, July 2000

K. Nichols, V. Jacobson, L. Zhang, A Two-bit Differentiated Services Architecture for the Internet, *RFC 2638*, July 1999.

B. Teitelbaum, P. Chimento, “QBone Bandwidth Broker Architecture”, work in progress, http://qbone.internet2.edu/bb/bboutline2.html.


TELECOMMUNICATIONS NETWORKS
CURRENT STATUS AND FUTURE TRENDS
Edited by Jesús Hamilton Ortiz
INTECH

Telecommunications Networks - Current Status and Future Trends
Edited by Dr. Jesús Ortiz

ISBN 978-953-51-0341-7
Hard cover, 446 pages
Publisher InTech
Published online 30, March, 2012
Published in print edition March, 2012

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821