End-to-end Quality of Service in IP Multimedia Subsystem using DiffServ

HDIP Course Project

By:

Umit Aygun
Yassine Kacemi
Masood Khosroshahy

Supervisor:

Prof. Noëmie Simoni

June 2006

Table of contents

1. Introduction

The communications industry as a whole is undergoing an evolutionary transformation, whereby the line between fixed-mobile broadband service providers are blurring, and where in the past subscribers have historically had multiple service provider relationships, are now able to get most of their communications services provided by a single provider.

While today's owners of multimedia-capable, multi-purpose mobile communication devices are demanding rich-media, interactive services which can take greater advantage of the technical capabilities of their devices, the traditional issues of network infrastructure—connecting the pipes to boxes, is giving way to new issues of service delivery and execution infrastructure, which requires running industry standards based services across multiple platforms, networks, and applications.

In order to meet the demands of this fast changing business climate, Content Service Providers (CSPs) need a horizontal network infrastructure which will allow them to rapidly develop, deploy, and deliver a large number of new services, which in many cases will have been developed by a 3rd party content and service provider[1].

IMS – IP Multimedia Subsystem – is an international, recognized industry standard specification defined by the 3rd Generation Partnership Project (3GPP) in Release 5 & 6, originally for 3G UMTS mobile networks. It specifies interoperability and roaming; provides bearer control, charging and security. The standard supports multiple access types – including GSM, WCDMA, CDMA2000, Wireline broadband access and WLAN. Because of its general applicability outside the wireless access domain, other standards bodies that have subsequently adopted the majority of the 3GPP IMS specifications as the underpinning of their own architectural standards. These forums include the 3rd Generation Partnership Project 2 (3GPP2) under the Multi-Media Domain (MMD) specifications, the Open Mobile Alliance (OMA), and the European Telecommunications Standard Institute (ETSI) [2].

IMS enables services to be delivered in a standardized, well-structured way that truly makes the most of layered architecture. At the same time, it provides a future-proof architecture that simplifies and speeds up the service creation and provisioning process, while enabling legacy interworking. For users, IMS-based services enable person-to-person and person-to-content communications in a variety of modes – including voice, text, pictures and video, or any combination of these – in a highly personalized and controlled way.

What is more, IMS is well integrated with existing voice and data networks, while adopting many of the key benefits of the IT domain. Since the primary concern is IP and application layer issues, non-mobile network operators, such as fixed-line operators and cable operators, are also beginning to adopt IMS as part of their broader move to all-IP networks. On longer term, IMS enables a secure migration path to an all-IP architecture that will meet end-user demands for new enriched services. That makes IMS a key enabler for fixed-mobile convergence and value-based charging. And for those reasons, IMS will become preferred solution for fixed and mobile operators' multimedia business.

IMS is not new in that its underlying technologies and concepts have been discussed by standards and technology groups for some time. But what is new is that the IMS specifications have gone through two 3GPP releases, with increasing adoption by CSPs, as well as vendors and Network Equipment Providers (NEPs) supplying the associated network equipment, applications and devices. IMS delivers a reusable platform for new service experimentation,

deployment, and integration, resulting in the expansion of the types of communications services available to consumer and enterprise end-users [1].

The IMS standard is based upon the widely adopted Internet standard technology called Session Initiation Protocol (SIP). SIP is at the heart of the IMS network architecture, providing the real-time, peer-to-peer, multiparty and multi-media capabilities of IMS. The application services layer of IMS networks support SIP interfaces and IMS SIP application servers, to reduce the complexity of IMS applications as well as to lead to enhanced, feature rich network services.

IMS specifies a core set of network functional entities, which support access to the SIP-based communication services provided by CSPs. Instead of inventing new protocols, IMS builds upon existing Internet protocols, as specified by IETF, such as SIP, Session Description Protocol (SDP), and Diameter, which enables the creation of a complete and robust real-time, peer-to-peer, multimedia network architecture. IMS provides for network elements, such as CSCF, HSS, MRF, and others, whose functionality and external and intra-IMS interfaces have been standardized.

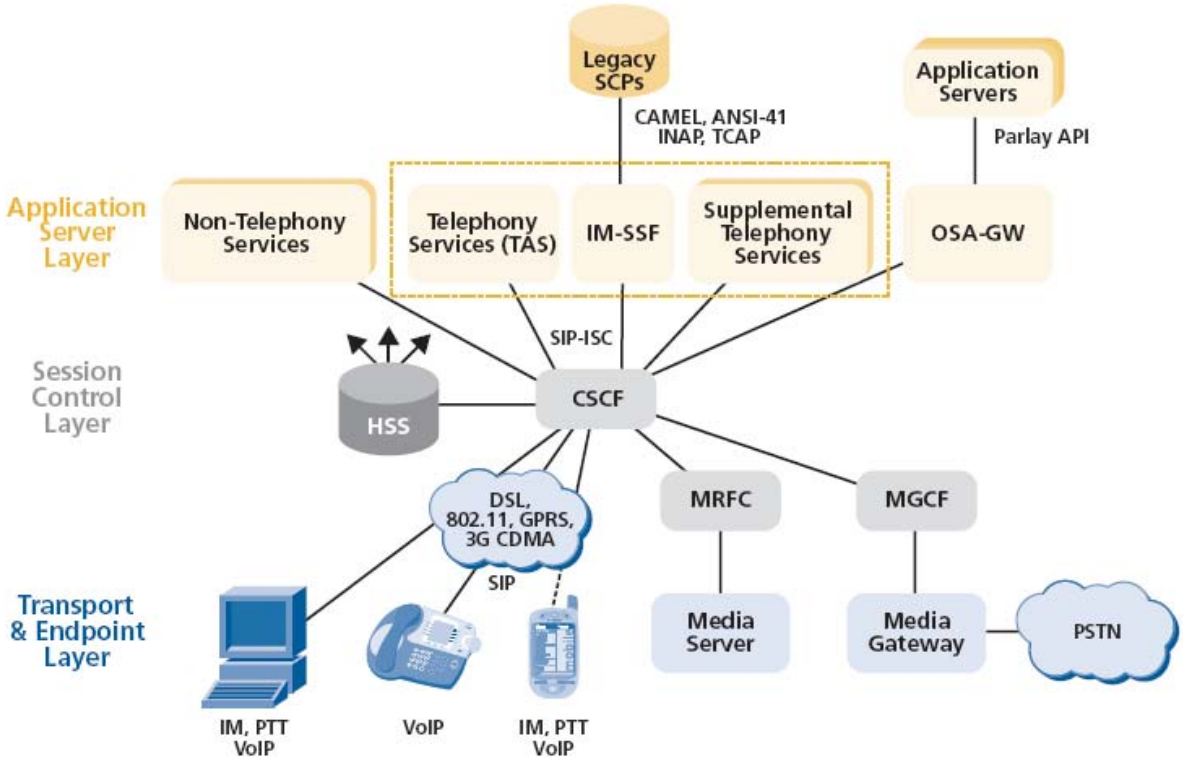


Figure 1 Simplified View of the IP Multimedia Subsystem (IMS) [3]

1.1. Benefits For Service providers

The key benefits derived by service providers from deploying IMS networks is new and increased revenue streams, and reduction in capital and operating expenses. By consolidating application interfaces into the application server, the creation of new multimedia services can be developed and delivered in a very short time-to-market cycle while dramatically reducing the support cost of the applications. IMS enables the creation of new services which were not possible previously, or might have been too costly and complex to implement, such as PoC or video sharing.

Because IMS supports roaming between different networks, new services developed to a single platform can be made available across multiple access networks. This will enable CSPs

to increase customer loyalty, increase ARPU from their installed base, and reduce churn. IMS also enables CSPs to monetize the fast pace of multimedia-enabled mobile device development, and end-user's changing needs. This requires the ability for CSPs to mix-and-match, and integrate different services to come up with new services. IMS enables CSPs to take an existing voice-based application, and integrate with multimedia sharing and video-enabled services. Or Web-based applications can be mobile/real-time/multimedia-enabled[1].

Using IMS, operators can adopt a strategy of first exploring the opportunities of IP multimedia, and then taking appropriate steps to mass-market IP multimedia services, according to market and business motivations. The hard lessons of the Internet bubble have brought us back to sound business logic, based on increased revenues and cost control. The introduction of new services and capabilities must not disturb the current profitable mix of telephony services. They should rather use it as a base for a superior user experience making it even more compelling.

By deploying an IMS network architecture, CSPs can reduce the need to build-out multiple silos of network elements each time they add a new service. By deploying a horizontal, IP-based, converged service delivery architecture based on the IMS standard, CSPs can implement new services on existing network infrastructure, reducing the costs associated with new equipment purchases. And in the longer term, IMS supports CSPs need to reduce the costs and complexities of managing multiple, parallel network elements, reducing their overall operating expenses.

IMS provides sound, business-focused evolution options for delivering attractive, easy-to-use, reliable and profitable multimedia services. It also enables operators achieve fixed–mobile convergence. Strategies are in place for operators to begin rolling out IMS-based services that take advantage of fast, flexible service creation and provisioning capabilities, while also providing for legacy interworking and combinational services that make the most of existing investments. Operators can then build onwards toward the all-IP vision of offering rich, multi-access multimedia services.

1.2. Benefits For End users

With IMS, end-users will be opened up to a new world of communication services which they might have associated mainly with the PC/Internet world, such as instant messaging and presence. In addition, there are new features in IMS services which end-users might never have thought about. For end-users, the benefits include richer, multimedia user experiences, roaming, new IP-based services, simplified identity management, personalization, ease-of-use, security and mobile-fixed-Internet integration. With the proliferation of rich-media capable mobile devices, both consumer and enterprise end-users have become very savvy about the personalized, interactive, and near real-time demands they have about their day-to-day communication services[1].

1.3. Features & Capabilities

The IMS architecture specifies a number of common functions and service enablers which can be reused across multiple access networks to enable multimedia services.

Multimedia session management

Multimedia session control and management in IMS is made possible through the use of SIP as the standard session control protocol. IMS enables the media session between to end-points to consist of any type of media content, and IMS also enables a session to be dynamically modified at run-time. This means media types can be added/dropped dynamically, depending the on the nature of the application[3].

Quality of Service
IMS provides CSPs with a standardized network element, the Policy Decision Function (PDF), which controls and monitors the packet network traffic into an IMS network from a GPRS and UMTS network. Through the PDF, IMS enables CSPs to deliver real-time IP network services at specified QoS levels.

Mobility management
IMS provides the HSS and CSCF elements to enable mobility management. The HSS is the data store for subscriber registration and location information, which is supplied to the CSCFs for session set-up and management, and message forwarding to IMS and non-IMS networks.

Service control
IMS networks address service control through the HSS and CSCF elements. As an end-user registers into the IMS network, the CSCF downloads the Subscriber Service Profile (SSP) from the HSS, which contains each individual's services provisioning information. For each subscriber, the SSP enables CSCFs to know which services need to be executed, in which order, address of the appropriate IMS application server(s), and the order in which the application server needs to execute the specified services. IMS enables CSPs to implement a common service control, execution and interaction platform for all services and subscribers accessing their networks[1].

Access-aware networks
Different services have different requirements. In order for different services to be executed properly, the network has to be aware of the different characteristics of the access methods. Multi-access functionality is inherent in the IMS architecture. This will enable the delivered service to be adapted to the characteristics and capabilities of the currently selected device and its network access method[2].

Standard interfaces
With IMS, 3GPP has delivered a standardized architecture and interfaces for deploying multimedia IP services across multiple access networks. This facilitates the development of new and innovative SIP/IMS services by 3rd party developers and service providers, independent of the IMS network deployments by CSPs, thereby fostering cross-network service integration, interoperability, and roaming.

Safe communication
With IMS, operators can implement end-to-end communications services built around a number of IMS security and network architecture cornerstones. These include the fundamental IMS attribute that *operator-controlled* services are provided to *authenticated* users. The originating operator has end-to-end responsibility in the operator community: no services are delivered to anonymous or untrustworthy end-users, and no service requests are relayed from anonymous and untrusted operators and enterprises. In addition, payload (primarily non-voice and video) is checked for viruses. Access domain security is provided through user authentication and Single Sign On[2].

Simple access to services
Once authenticated through an IMS service, the user is able to access all the other IMS services that he is authorized to use. Authentication is handled by the CSCF as the user signs on. When it receives a service request, the SIP Application Server (AS) can verify that the user has been authenticated.

Service interoperability

IMS enables the reuse of inter-operator relations. Rather than develop different interconnect relations and agreements for each service, IMS enables a single inter-operator relationship to be established and built upon for each service [2]. Once IMS is in place, access to other users' services is an IMS network issue, common to all IMS personal services, as shown in Figure 2. The requesting user's operator service does not need to be involved in routing the request. The inter-operator network-to-network interface is established in IMS , and the general IMS inter-operator service agreement, routing, service network access point and security are all reused.

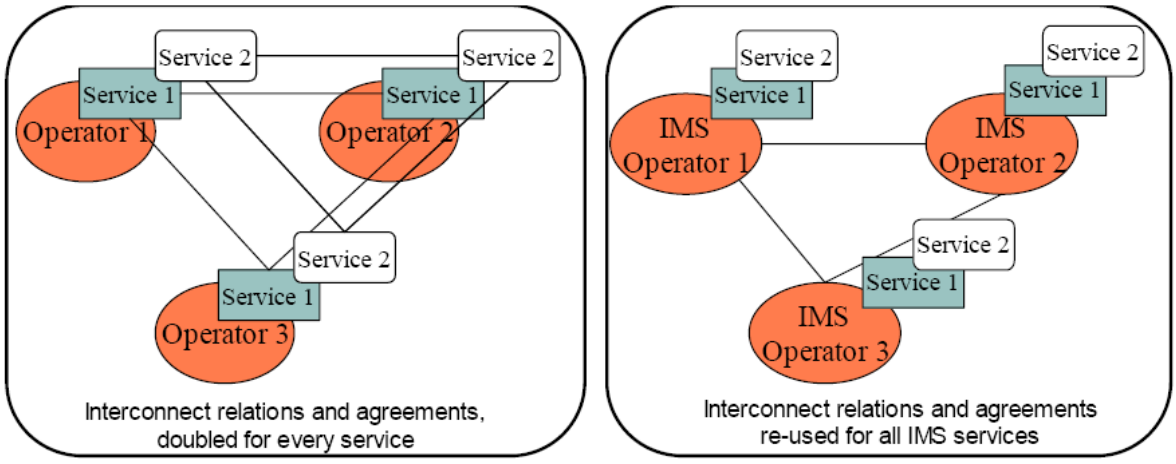


Figure 2. The difference in service interoperability between a pre-IMS network and IMS enabled operators [2]

2. Architecture

The traditional vertical network structure – with its service-unique functionality for charging, presence, group and list management, routing and provisioning – is very costly and complex to build and maintain. Separate implementations of each layer must be built for every service in a pre-IMS network, and the structure is replicated across the network, from the terminal via the core network to the other user's terminal.

IMS provides for a number of common functions that are generic in their structure and implementation, and can be reused by virtually all services in the network. Examples of these common functions are group/list management, presence, provisioning, operation and management, directory, charging and deployment.

IMS offers a network architecture where software infrastructure, through the use of standards, enable network elements to look and feel like general purpose servers. With the 3GPP Release 5 and Release 6 specifications, IMS enables many network functionalities to be reused and shared across multiple access networks, allowing for rapid service creation and delivery. This opens the network for off-the-shelf application servers and IDE tools. The architecture consists of [1]:

1. Service Layer,
2. Control Layer,
3. Connectivity Layer.

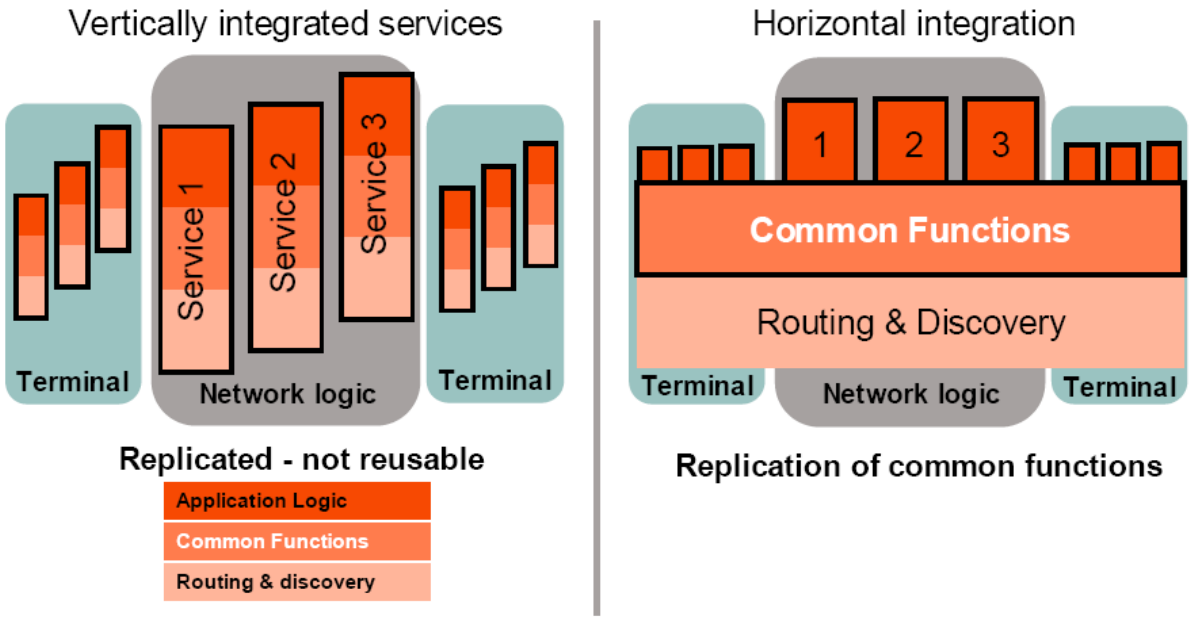


Figure 3 : How IMS enables the move from vertical 'stove-pipe' service implementations to a horizontally layered architecture with common functions [2].

2.1. Service Layer

IMS specifies a SIP-based common interface, IMS Service Control (ISC) by which applications hosted on SIP, Parlay/OSA and CAMEL application servers interact with the IMS core network. The main integration point between IMS application servers and the IMS core network is through the Serving Call Session Control Function (S-CSCF) network element.

SIP Application Servers

Hosts and execute services, as well as influence and impact the SIP session on behalf of the services.

Telephony Application Server

The IMS architecture supports multiple application servers for telephony services. The Telephony Application Server (TAS) is a back-to-back SIP user agent that maintains the call state. The TAS contains the service logic which provides the basic call processing services including digit analysis, routing, call setup, call waiting, call forwarding, conferencing, etc.If the calls are originating or terminating on the PSTN, the TAS provides the SIP signaling to the MGCF to instruct the media gateways to convert the PSTN TDM voice bit stream to an IP RTP stream and to direct it to the IP address of the corresponding IP phone[3].

As part of executing the telephony call model, the TAS provides the Advanced Intelligent Network (AIN) call trigger points. When a call progresses to a trigger point, the TAS suspends call processing and checks the subscriber profile to determine if additional services should be applied to the call at this time. The subscriber profile identifies which application servers should be invoked. The TAS formats a SIP IP Multimedia Service Control (ISC) message and passes call control to the appropriate application server. This mechanism can be used to invoke legacy AIN services or to invoke new SIP based applications servers.

A single IMS can contain multiple TASs that provide specific features to certain types of endpoints. For example, one TAS might provide the IP Centrex business features (i.e., private dialing plans, shared directory numbers, multiple call appearances, Automatic Call Distribution (ACD), attendant services, etc.). Another TAS might support PBXs and provide

advanced Virtual Private Network (VPN) services. The multiple application servers can interwork using SIP-I signaling to complete calls between the different classes of endpoints.

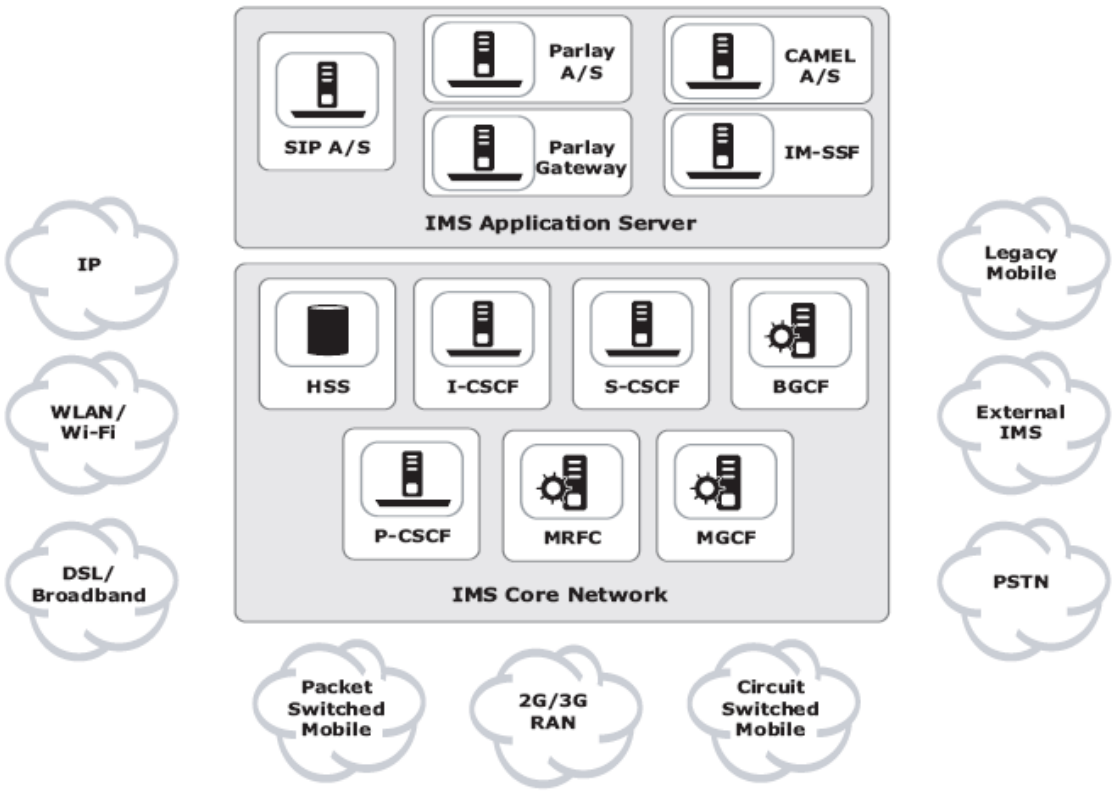


Figure 4 : IMS network architecture [1]

IP Multimedia – Services Switching Function (IM-SSF)

The IP Multimedia – Services Switching Function (IM-SSF) provides the interworking of the SIP message to the corresponding Customized Applications for Mobile Networks Enhanced Logic (CAMEL), ANSI-41, Intelligent Network Application Protocol (INAP) or Transaction Capabilities Application Part (TCAP) messages. This interworking allows the IP Phones supported by IMS to access services such as calling name services, 800 services, Local Number Portability (LNP) services, one number services, and more.

Supplemental Telephony Application Servers

The application server layer can also contain standalone independent servers that provide supplemental telephony services at the beginning of a call, at the end, or in the middle, via triggers. These services include click to dial, click to transfer, click to conference, voice mail services, IVR services, VoIP VPN services, prepaid billing services, and inbound/outbound call blocking services.

Non Telephony Application Servers

The application server layer can also contain SIP based application servers that operate outside of the telephony call model. These application servers can interwork with endpoint clients to provide services such as IM, PTT, or presence-enabled services.

Open Service Access – Gateway (OSA-GW)

The IMS architecture allows service providers the flexibility to add services into their VoIP networks by interacting with legacy applications or by integrating SIP-based application servers that they purchase or develop themselves. In addition, service providers want to allow their customers to develop and implement services that leverage the VoIP network resources.

For example, an enterprise may want to voice-enable or IM-enable some back office operations to automatically initiate a call or an IM if an order is about to be delivered. This could be triggered by the location information of a wireless PDA carried by the delivery person. However, frequently the enterprise application developers have IT backgrounds and are not familiar with the variety of complex telephony signaling protocols (i.e., SS7, ANSI41, CAMEL, SIP, ISDN, etc.). To provide a simple API for communications services, the Parlay Forum, working closely with the 3GPP and ETSI standards development organizations, have jointly defined a Parlay API for telephony networks. The interworking between SIP and the Parlay API is provided in the Open Services Access – Gateway (OSAGW) that is part of the application server layer of the 3GPP IMS architecture. As described above, other applications servers provide the interworking between SIP and the telephony protocols (ANSI-41, CAMEL, INAP, TCAP, ISUP, etc.). The OSA-GW allows the enterprise-based Parlay applications to access presence and call state information, set up and tear down sessions, and to manipulate legs of a call. The OSA-GW implements the Parlay Framework, which allows the enterprise applications servers to register with the network and manage access to network resources[3].

2.2. Control Layer

Session control is where the network signaling is performed for setting up sessions. The session control layer consists of several core network elements which control and manage session set-up and maintain subscriber user data. Session control also provides interworking between IMS and PSTN/PLNM networks through media servers and gateways**.**

With IMS, users access personal services via a dynamically associated, usercentric, service-independent and standardized access point, the Call Session Control Function (CSCF). The CSCF is dynamically allocated to the user at log-on or when a request addressed to the user is received. Routing to the server is service-independent and standardized. The service architecture is user-centric and is highly scalable.

The CSCF is a SIP proxy and registrar server which manage the registration of IMS User Equipment (UE, also known as terminals, handsets, or SIP phones), and routing of SIP signaling messages to the appropriate IMS application server, or to the appropriate IMS or non-IMS network. Key to optimal SIP message processing by CSCF nodes, and service logic execution by IMS SIP application servers, are high-performance, high-availability features. The CSCF interworks with the transport and endpoint layer to guarantee QoS across all services. IMS specifies 3 different types of CSCFs[3];

Proxy-CSCF (P-CSCF)

Performs SIP proxy server which routes SIP request and response messages of the UE to the Interrogating-CSCF(I-CSCF) determined using the home domain name as provided by the UE. It is the single point of entry for all traffic from the UE into the IMS network. It also sends all subsequent SIP messages received from the UE to the S-CSCF, whose name has been received as part of registration. P-CSCFs are typically located in the user's home network, but are often located in visited networks, and provides the following functionalities:

- Billing information generation
- SIP message compression for latency reduction to reduce the amount of data sent over the radio interface
- IPSec integrity protection for trusted messages
- SIP message verification

Interrogating-CSCF (I-CSCF)
SIP proxy server which queries the Home Subscriber Server (HSS) to obtain the address of the appropriate S-CSCF where the request must be forwarded, if no S-CSCF is currently assigned (e.g., unregistered subscriber), then assigns an S-CSCF to handle the SIP request. It is the first point of contact within the user's home network and routes SIP requests received from another network to the S-CSCF. I-CSCF provides the Topology Hiding Interworking Gateway (THIG) function.

Serving-CSCF (S-CSCF)
It acts like a SIP registrar server which enables the requesting user to access the network services provided by the network operator, and handles all of the SIP signaling between endpoints. The S-CSCF retrieves the subscriber profile from the HSS and interacts with Application Server platforms for the support of services. It also ensures that the media for a session, as indicated by SDP, are within boundaries of subscriber's profile. S-CSCFs are always located in the user's home network, and provides the following functionalities:

- Session control
- Service usage authentication & authorization
- Session context
- SIP message routing

Policy Decision Function (PDF)
PDF is rresponsible for making policy decisions based on session and media-related information obtained from the P-CSCF. It acts as policy decision point for Service-based Local Policy (SBLP) control. Some of policy decision point functionalities:

- To store session and media-related information
- The capability to enable the usage of an authorized bearer (e.g. PDP context)
- To inform P-CSCF when the bearer is lost or modified.
- To pass an IMS-charging identifier to the GGSN and to pass a GPRS-charging identifier to the P-CSCF

Home Subscriber Server
The session control layer includes the (HSS) element provides a central database which stores each subscriber's unique service preferences and information, including current registration information (IP address), roaming information, call forwarding information, etc.. IMS centralizes the subscriber information to enable multiple applications across multiple access networks to share and leverage any given subscriber's status and preference information. HSS also enables operators to better manage and administer subscriber data and provisioning across multiple services across multiple networks. HSS provides IMS service authentication and authorization support, as well as maintain information about the currently assigned S-CSCF for any given user request, and supports interactions with CSCFs and ASs. ***Subscription Locator Function*** **(SLF)** is used as resolution mechanism to find the address of the HSS that holds the subscriber data.

Transport Signalling Gateway Function (T-SGW)
This component serves as the PSTN/PLMN termination point for a defined network. Terminates, e.g. the call control signalling from GSTN mobile networks (typically ISDN) and maps call related signalling from/to PSTN/PLMN on an IP bearer and sends it to/from the MGCF.It also provides PSTN/PLMN IP transport level address mapping.

Roaming Signalling Gateway Function (R-SGW)
The role of the R-SGW concerns only roaming to/from 2G/R99 CS and the GPRS domain to/from the R5-6 UMTS tekeservices domain and the UMTS-GPRS domain and does-not involve the multimedia domain. We can summarize the main functions as [8];

- To ensure proper roaming, the R-SGW performs the signalling conversion at transport level (conversion: Sigtran SCTP/IP vs. SS7 MTP) between the legacy SS7 based transport of signalling and the IP-based transport of signalling. The R-SGW does not interpret the MAP/CAP messages but may have to interpret the underlying SCCP layer to ensure proper routing of the signalling.
- To support 2G/R99 CS terminals; we use R-SGW services to ensure transport interworking between the SS7 and the IP transport of MAP_E and MAP_G signalling interfaces with a 2G/R99 MSC/VLR[8].

Breakout Gateway Control Function (BGCF)
BGCF selects;
- The network in which PSTN breakout is to occur.
- A local MGCF or a peer BGCF.

Media Gateway Control Function (MGCF)
The MGCF serves as the PSTN/PLMN termination point for a defined network. Its defined functionality will satisfy the standard protocols/interfaces to [8]:
- Control parts of the calls state pertain to connection control for media channels in a MGW.
- Communicate with CSCF
- Select the CSCF depending on the routing number for incoming calls from legacy networks
- Perform protocol conversion between the legacy (e.g. ISUP, R1/R2, etc.) and the R00 network call control protocols
- May process out of band information such as DTMF signaling received in MGCF which it may forward to the CSCF or MGW.

Media Gateway Function (MGW)
The MGW serves as the PSTN/PLMN transport termination point for a defined network and UTRAN interfaces with the CN over Iu interface. It may terminate bearer channels from a switched circuit network (i.e. DSOs) and media streams from a packet network (e.g. RTP streams in IP network). Over Iu, the MGW may support media conversion, bearer control and payload processing (e.g. codec, echo canceller conference bridge) for support of different Iu options for CS services, AAL2/ATM based as well as RTP/UDP/IP based. The main functions can be summarized as [8];
- Interaction with MGCF, MSC server and GMSC server for resource control
- Ownership and resources handling, e.g. echo cancellers, etc.
- Ownership of codecs
- May detect events (i.e. bearer loss, DTMF digits, etc.) and notifies the MGCF.
- May perform DiffServ Code Point (DSCP) markings on the IP packets sent towards the UE

Multimedia Resource Function Controller (MRFC)
The MRFC performs;
- Controls the media stream resources in the MRFP.
- Interprets information from an AS via the S-CSCF (using SIP) and controls the MRFP accordingly.
- Communication with the CSCF for service validaation and for multiparty/multimedia sessions
- May be co-located with an AS to provide capabilities such as conference services.

Multimedia Resource Function Processor (MRFP)
Under the control of MRFC the functions of MRFP can be summarized as:

- Mixes (e.g. for multiple parties), sources (for multimedia announcements) and processes (e.g. audio transcoding) media streams.
- Performs bearer control (with GGSN and MGW) in cases of multiparty/multimedia conferencing
- Provide tones and supports DTMF within the bearer path.
- Notifies the MRFC when an event has occurred such as DTMF digit collection.

MSC and Gateway MSC Server
The MSC server includes mainly the call control and mobility control parts of a GSM/UMTS MSC. It has responsibility for the control of MO and MT 04.08CC CS domain calls. It terminates the user-network signalling (04.08 + CC + MM) and translates it into the relevant network-network signalling. The MSC server also contains VLR to hold the mobile subscriber's service data and CAMEL-related data, controls the parts of the call state that pertain to connection control for media channels in an MGW [8].

The GMSC server comprises primarily the call control and mobility control parts of a GSM/UMTS GMCS. An MSC server and an MGW make up the full functionality of an MSC, while the Gateway MSC and a GMSC server and an MGW make up the full functionality of a GMSC.

2.3. Connectivity Layer

The network connectivity layer consists of routers, switches, media servers and media gateways for converting VoIP bearer streams to the PSTN TDM format. This layer provides a common pool of media servers which can be shared across multiple applications and services including conferencing, playing announcements, collecting in-band signaling tones, speech recognition, speech synthesis, etc. This layer also initiates and terminates SIP signaling to set up sessions and provide bearer services such as conversion of voice from analog or digital formats to IP packets using Realtime Transport Protocol (RTP). The IMS network architecture supports connectivity with all types of access networks, whether it's IP, broadband/DSL/cable, Wi-Fi, circuit-switched mobile, packetswitched mobile, legacy mobile, or external IMS networks.

3. Application examples

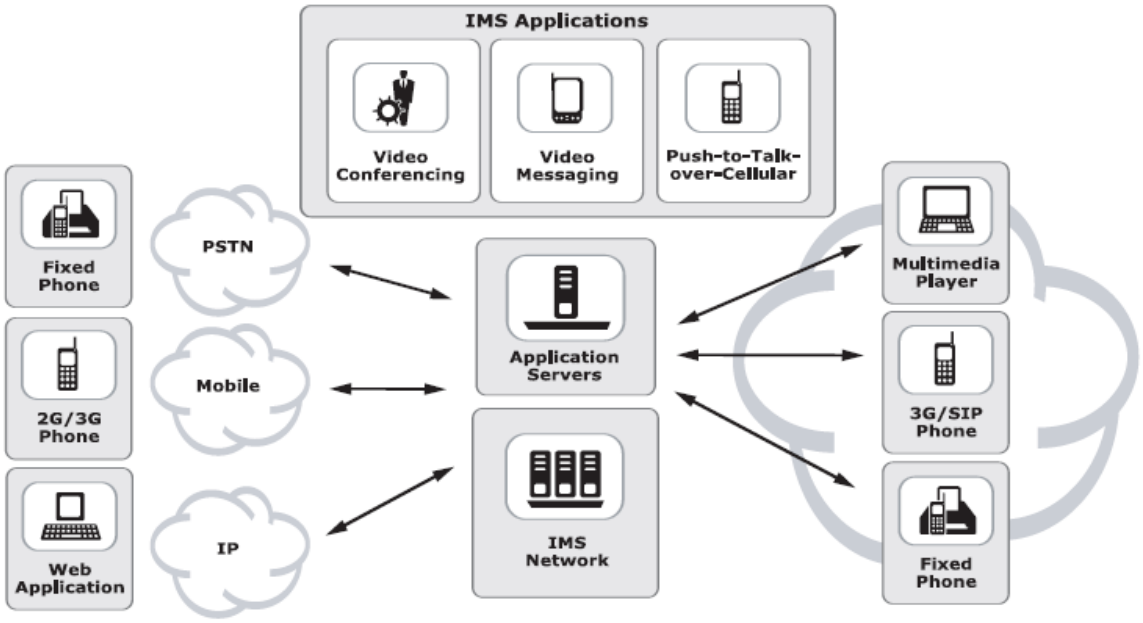


Figure 5 : Application Examples [1]

Push to talk over cellular (POC)
Push to Talk over Cellular (PoC) is one of the first IMS based applications that are available in the wireless network. It operates entirely in the packet-switched domain and is based on IMS service enablers and common functions. IMS enables PoC services through presence, instant messaging, billing, single sign-on, and central OA&M processes [1].

Multimedia conferencing
IMS supports multimedia conferencing services through it's QoS feature, which enables a higher quality user experience, and through multimedia session management features, which enables the session set-up for each individual participant to be managed separately, thereby matching voice and video quality to each participant's device capabilities.

Voice-video messaging
IMS enables voice-video messaging with its standardization on SIP, and via the CSCF and MRF elements.

Click to dial
IMS networks enable click-to-dial services by leveraging the SIP protocol and the 3PCC (3rd Party Call Control) B2BUA (Back-to-Back User Agent) network element, which can establish, manage, and terminate communication sessions on-behalf of two or more SIP user agents [1].

4. QoS in IMS [9]

With interaction between the user plane and the control plane, operators are able to control quality of service, among others. A mechanism to authorize and control the usage of the bearer traffic intended for the IMS media traffic was created; it was based on the SDP parameters negotiated at the IMS session. This overall interaction between the GPRS and the IMS is called a Service-Based Local Policy (SBLP) control. The following figure shows the functional entities involved in the SBLP. PDF and P-CSCF are co-located as was the case in Release 5 of the standard; this is changed in Release 6 and the reference point Gq is standardized between the two elements.

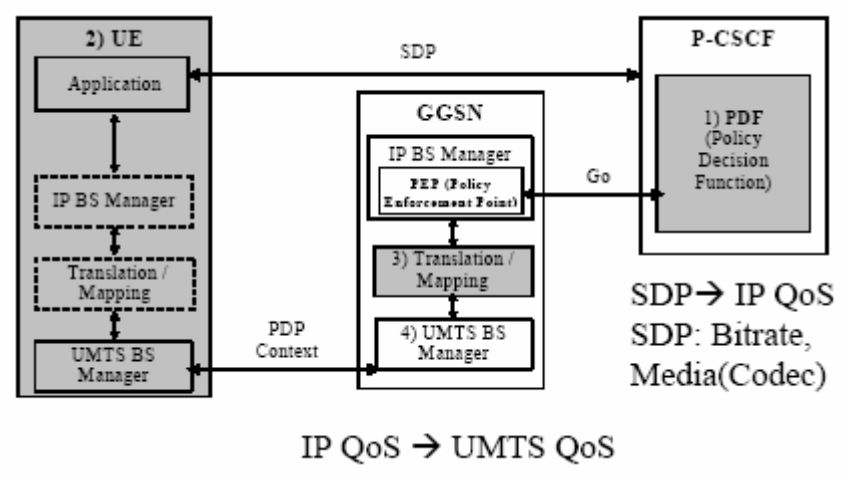


Figure 6 : SBLP entities [5]

Hereafter, we give a brief overview of the functionalities of each entity:

- IP Bearer Service (BS) manager:
  Manages the IP BS using a standard IP mechanism. It resides in the GGSN and optionally in the UE.
- Translation/Mapping function:
  Provides the inter-working between the mechanism and parameters used within the UMTS BS and those used within the IP BS. It resides in the GGSN and optionally in the UE.
- UMTS BS manager:
  Handles resource reservation requests from the UE. It resides in the GGSN and in the UE.
- Policy Enforcement Point:
  Is a logical entity that enforces policy decisions made by the PDF. It resides in the IP BS manager of the GGSN.
- Policy decision function:
  Is a logical policy decision element that uses standard IP mechanisms to implement SBLP in the IP media layer. In Release 5, it resides in the P-CSCF. In Release 6, it is stand-alone entity. The PDF is effectively a policy decision point that defines a framework for policy-based admission control.

4.1. Bearer Authorization

Session establishment and modification in the IMS involves an end-to-end message exchange using SIP and SDP. During the message exchange, UEs negotiate a set of media characteristics. If an operator applies the SBLP, then the P-CSCF will forward the relevant SDP information to the PDF together with an indication of the originator. The PDF notes and authorizes the IP flows of the chosen media components by mapping from SDP parameters to authorized IP QoS parameters for transfer to the GGSN via the Go interface.

When the UE is activating or modifying a PDP context for media, it has to perform its own mapping from SDP parameters to some UMTS QoS parameters. PDP context activation or

modification will also contain the received authorization token and flow identifiers as the binding information.

On receiving the PDP context activation or modification, the GGSN asks for authorization information from the PDF. The PDF compares the received binding information with the stored authorization information and returns an authorization decision. If the binding information is validated as correct, then the PDF communicates the media authorization details in the decision to the GGSN. The media authorization details contain IP QoS parameters and packet classifiers related to the PDP context.

The GGSN maps the authorized IP QoS parameters to authorized UMTS QoS parameters and finally GGSN compares the UMTS QoS parameters against the authorized UMTS QoS parameters of the PDP context. If the UMTS QoS parameters from the PDP context request lie within the limits authorized by the PDF, then PDP context activation or modification will be accepted. The following diagram shows the explained functionality:

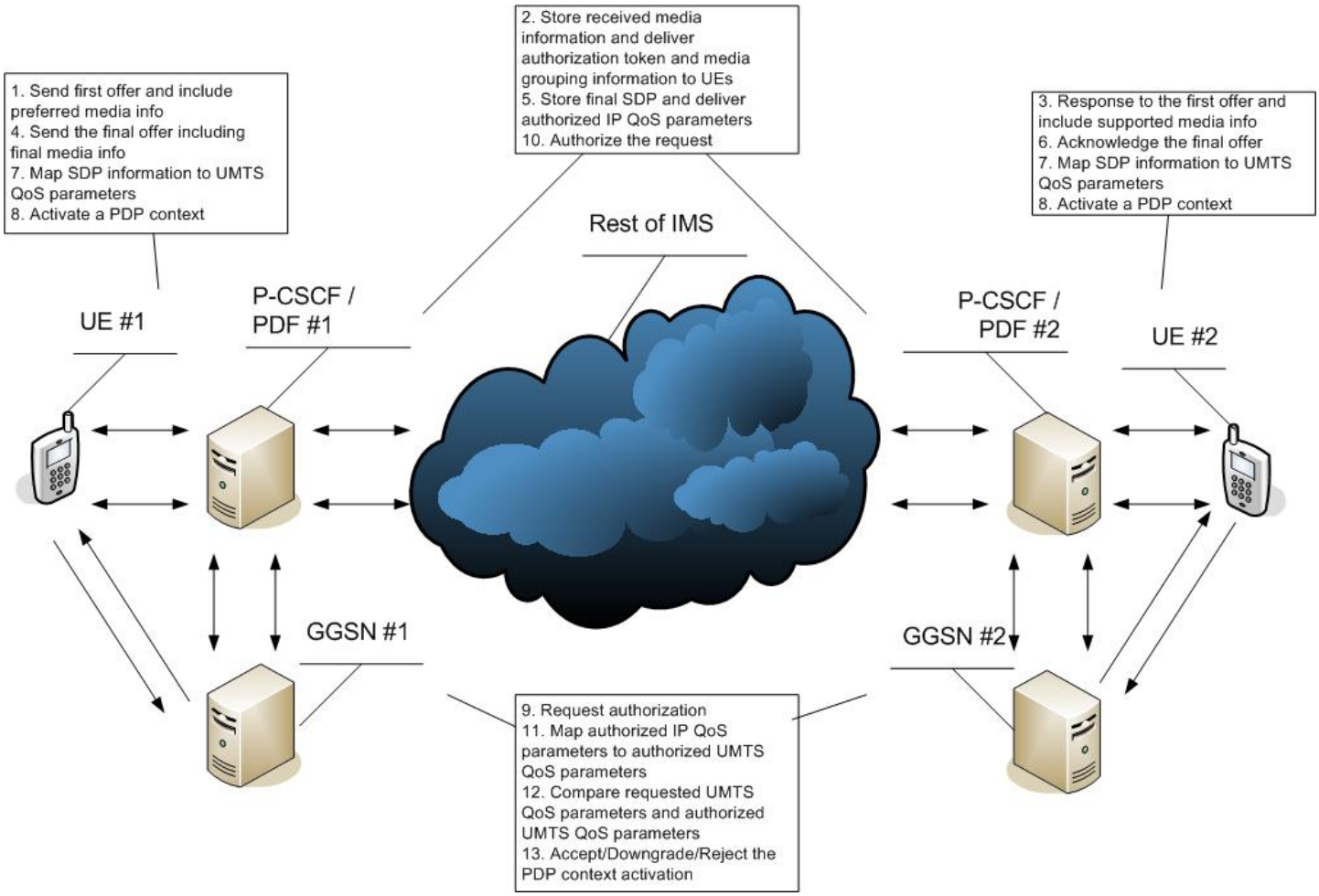


Figure 7: Bearer Authorization using SBLP

4.2. Authorize QoS Resources

During the session setup, the PDF collects IP QoS authorization data which are comprised of:

- Flow Identifier:
  Used to identify the IP flows that are described within a media component associated with an SIP session. A flow identifier consists of the ordinal number of the position of the “m= ” lines in the SDP session description and the ordinal number of the IP flow within the “m= ” line assigned.
- Data rate:
  This information is derived from SDP bandwidth parameters.

- QoS class:
  The QoS class information represents the highest class that can be used for the media component. It is derived from the SDP media description.

PDF derives the data rate value for the media IP flow(s) from the SDP parameters. The PDF maps media-type information into the highest QoS class that can be used for the media. The PDF will use an equal QoS class for both the uplink and the downlink directions when both directions are used. The authorized IP QoS comprises the QoS class and data rate. The PDFs also create the flow identifiers that will be used to create packet classifiers in the GGSNs.

Authorization token:

- It is a unique identifier across all PDP contexts associated with an access point name.
- It is created in the PDF when the authorization data are created.
- It consists of the IMS session identifier and the PDF identifier.
- The UE includes it in a PDP context activation/modification request.
- GGSN uses a PDF identifier within the authorization token to find the PDF that holds the authorized IP QoS information.
- The PDF uses the authorization token to find the right authorized data when receiving requests from the GGSN.

Media grouping

SIP and the IMS allow multimedia sessions to be setup which may comprise a number of different components, such as audio and video. Any particular party may add or drop a media component from an ongoing session. As defined in the standard, all components should be individually identifiable for charging purposes, and it must be possible to charge for each of these components separately in a session.

In Release 5, GGSN is able to produce only one GGSN call detail record (CDR) for a PDP context. Therefore, it is impossible to separate traffic for each media component within the same PDP context. As the current model for charging data generation and correlation does not allow multiplexing media flows in the same secondary PDP context, there must be a mechanism on the IMS level to force the UE to open separate PDP contexts for each media component. For this purpose, a keep-it-separate indication was defined. There is ongoing work in Release 6 to introduce a capability to charge on an IP flow basis. This would allow more freedom to transport media components in the same PDP context.

4.3. Resource Reservation

UE functions –When the UE receives an authorization token within the end-to-end message exchange, it knows that SBLP is applied in the network. Therefore, it has to generate the requested QoS parameters and flow identifiers for a PDP context activation/modification request. The requested QoS parameters include Traffic Class, Guaranteed Bit Rate and Maximum Bit Rate among others:

- Traffic Class –the four different classes defined for UMTS are conversation, streaming, interactive and background. By including the traffic class, UMTS can make assumptions about the traffic source and optimize the transport for that traffic type.
- Guaranteed Bit Rate (GBR) –Describes the bit rate the UMTS bearer service will guarantee to the user or application.
- Maximum Bit Rate (MBR) –Describes the upper limit a user or application can accept or provide. This allows different rates to be used for operation (e.g. between GBR and MBR).

The traffic class values, GBR and MBR for downlink/uplink should not exceed the derived values of maximum authorized bandwidth and maximum authorized traffic class per flow identifier. The maximum authorized bandwidth in the UE is derived from SDP in the same way as was done in the PDF. Also, flow identifiers are derived in the UE in the same manner as in the PDF.

Next, the UE needs to decide how many PDP contexts are needed. The key factors are the nature of media streams (i.e., required traffic class) and the received grouping indication from the P-CSCF. After deriving and choosing the suitable, requested QoS parameters, the UE activates the necessary PDP contexts. The authorization token and flow identifiers are inserted within the traffic flow template information element and the requested QoS parameters are inserted within the QoS information element.

GGSN functions –When a GGSN receives a secondary PDP context activation request to an access point name for which the Go interface is enabled, GGSN will:

- Identifies the correct PDF by extracting the PDF identity from the provided authorization token. If an authorization token is missing, then the GGSN may either reject the request or accept it within the limit imposed by a locally stored QoS policy.
- Requests authorization information from the PDF for the IP flows carried by a PDP context. This request is a Common Open Policy Service (COPS) request and contains the provided authorization token and the provided flow identifiers.
- Enforces the decision after receiving an authorization decision. The authorization decision is given as a COPS authorization_decision message. The main components of the decision are:
  - Direction indication: Uplink, downlink
  - Authorized IP QoS: Data rate, Maximum authorized QoS class
  - Packet classifiers (also called a gate description): Source IP address and port number, Destination IP address and port number, Protocol ID
- Maps the authorized IP QoS to the authorized UMTS QoS
- Compares the requested QoS parameters with the authorized UMTS QoS. If all the requested parameters lie within the limits, then the PDP context activation will be accepted. If the requested QoS exceeds the authorized UMTS QoS, then the requested UMTS QoS information in downgraded to the authorized UMTS QoS information.
- Constructs a gate description based on the received packet classifier. The gate description allows a gate function to be performed. The gate function enables or disables the forwarding of IP packets. If the gate is closed, then all packets of the related IP flows are dropped. If the gate is open, then the packets of the related IP flows are allowed to be forwarded. The opening of the gate may be part of the authorization decision event or may be a stand-alone decision. The closing of the gate may be part of the revoke authorization decision.
- Stores the binding information.
- May cache the policy decision data of the PDF decisions.

During the secondary PDP context modification, the GGSN may use previously-cached information for a local policy decision in case the modification request does not exceed the previously-authorized QoS. If the GGSN does not have cached information, then it performs aforementioned procedure.

PDF functions –When a PDF receives a COPS request, the PDF validates that:

- The authorization token is valid.
- The corresponding SIP session exists.

- The binding information contains valid flow identifiers.
- The authorization token has not changed in an authorization modification request.
- The UE follows the grouping indication.
- If validation is successful, then the PDF determine and communicate the authorized IP QoS, packet classifiers and the gate status to be applied to the GGSN.

4.4. Other Issues

Approval of the QoS commit function –During the resource reservation procedure, a PDF sends packet classifiers to a GGSN. Based on the packet classifiers, the GGSN formulates a gate to policy-control incoming and outgoing traffic. It is the PDF's decision when to open the gate. When the gate is open, the GGSN allows traffic to pass through the GGSN. Opening the gate could be sent a response to an initial authorization request from the GGSN or the decision can be sent as a stand-alone decision. With a stand-alone decision, an operator can ensure that user-plane resources are not used before the IMS session is finally accepted.

Removal of the QoS commit function –The function closes a gate in the GGSN, when a PDF does not allow traffic to traverse through the GGSN. This function is used, for example, when a media component is put on hold due to media re-negotiation.

Indication of bearer release function –When the GGSN receives a delete PDP context request and the PDP context has been previously authorized via the Go reference point, the GGSN informs the PDF of the bearer release related to the SIP session by sending a COPS delete request-state message. The PDF removes the authorization for the corresponding media component. When the PDF receives a report that a bearer has been released, it could request the P-CSCF to release the session and revoke all the related media authorization.

Indication of bearer loss/recovery –When the MBR value equals 0 kbit/s in an update PDP context request, the GGSN needs to send a COPS report message to the PDF. Similarly, when the MBR is modified from 0 kbit/s, the GGSN send a COPS report message to the PDF after receiving an update from the SGSN. Using this mechanism, the IMS is able to learn that the UE has lost/recovered its radio bearer when a streaming or conversational traffic class is in use in the GPRS system. When the PDF receives a report that the MBR equals 0 kbit/s, it could request the P-CSCF to release the session and revoke all the related media authorization.

Revoke function –This function is used to force the release of previously authorized bearer resources in a GPRS network. With this mechanism the PDF is able, for example, to ensure that the UE releases a PDP context when an SIP session is ended or that the UE modifies the PDP context when a media component bound to a PDP context is removed from the session. If the UE fails to do so within a predefined time set by an operator, then PDF revokes the resources.

Charging identifiers exchange function –The Go reference point is the link between the IMS and the GPRS networks. For charging correlation to be carried out, the IMS layer needs to know the corresponding GPRS layer charging identifier and vice versa. These charging identifiers are exchanged during the bearer authorization phase. An IMS charging identifier is delivered to the GGSN within the authorization decision message, while a GPRS charging identifier is delivered to the PDF a part of the authorization report.

5. Differentiated Services [11]

IETF proposed a framework, called Diffserv, that could support a scalable form of QoS and could provide a variety of end-to-end services across multiple, separately administered domains. Trying to maintain per-flow QoS becomes a monumental task for large networks. DiffServ works at class level, where a class is an aggregate of many such flows. For example, packets coming from a set of source addresses may fall into one class.

5.1. DiffServ Architecture

The RFCs 2474 and 2475 define the fundamental framework of the Diffserv architecture. The scaling properties of the Diffserv architectural framework are achieved by marking each packet's header with one of the standardized codepoints. Each packet containing same codepoint receives identical forwarding treatment by routers and switches in the path. This obviates the need of state or complex forwarding decisions in core routers based on per flow, as is the case with Intserv.

The following figure shows a Diffserv domain with a set of interior (core) routers and boundary (edge) routers. The ingress boundary router is normally required to classify traffic into microflow based on TCP/IP header fields. Diffserv microflows are subjected to policing and marking at the ingress boundary router according to a contracted service level specification (SLS). Depending on the particular Diffserv model, out-of-profile packets are either dropped at the boundary or marked with a different priority level, such as best-effort. These functions are termed as traffic conditioning in Diffserv language. A traffic conditioner is governed by rules that are defined in the traffic conditioning agreement. TCA typically includes traffic characteristics (token bucket parameters may be used for this) and performance metrics (delay, throughput, etc.) as actions required for dropping nonconformant packets.

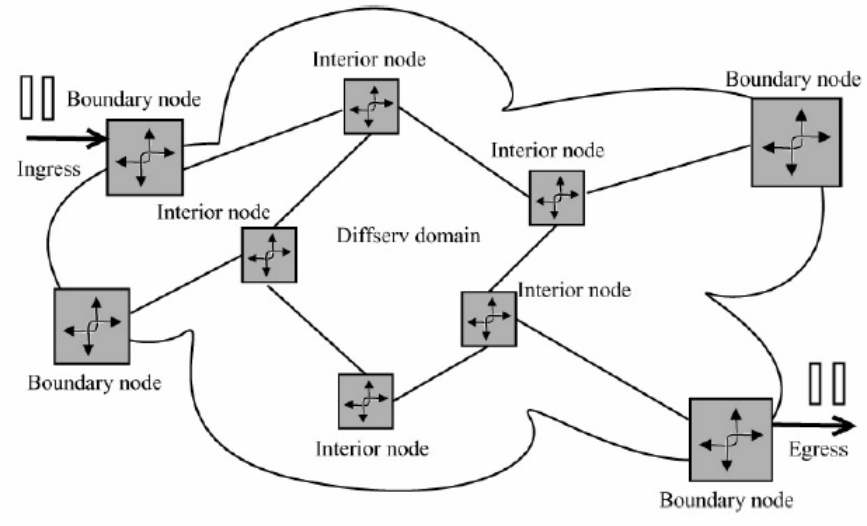


Figure 8. DiffServ Domain

A DiffServ flow along with similar Diffserv traffic forms an aggregate. All subsequent forwarding and policing are performed on aggregates by Diffserv interior nodes. As the interior nodes are not expected to perform an expensive classification function, their ability to process packets at high speeds becomes viable. Enforcement of the aggregate traffic contracts between Diffserv domains is key to providing QoS. The admission control modules must ensure that new reservations do not exceed the aggregate traffic capacity. These features make it possible to provide end-to-end services using Diffserv architecture.

5.2. Per-Hop Behavior

In contrast to Intserv, the Diffserv model does not define any service; it defines certain behaviors a packet may receive at each hop. This is called per-hop behavior (PHB). PHBs are combined with a much larger number of policing policies at the edge routers, to provide a range of services. Many different PHBs can be defined. Note that Diffserv does not standardize any particular queuing discipline. The vendors may use priority queuing, WFQ, or anything they like, as long as the observable behavior meets the PHB specification. In the Diffserv model, several traffic flows are aggregated to one of a small number of behavior aggregates (BAs). Each BA gets treated using the same PHB. Flows identified by the same Diffserv Code Point (DSCP) belong to a BA. A PHB group is a set of PHBs that share a common constraint. Within a group, resources can be allocated relative to each other. Also, the drop precedence of packets may be defined within a group.

RFC2598 has standardized a PHB called expedited forwarding (EF). Using the EF PHB, carriers can develop a service that provides a low loss, low latency, low jitter, and bandwidth assurance through its DS domain. Such a service is also known as premium service. Premium service is intended for traffic that requires a virtual leased line. The virtual leased line is similar to constant bit rate (CBR) traffic. It provides a simple abstraction of a link with minimum guaranteed bandwidth. The EF PHB is defined as a forwarding treatment for a particular Diffserv aggregate where the departure rate of the aggregate's packets from any Diffserv node must equal or exceed a configurable rate. The EF traffic receives this rate independent of the intensity of any other traffic attempting to transit the node. It averages at least the configured rate when measured over any time interval equal to or longer than the time it takes to send an output link MTU-sized packet at the configured rate. The configured minimum rate is settable by a network administrator. If the EF PHB is implemented by a mechanism that allows unlimited pre-emption of other traffic (e.g., a priority queue), the implementation has to include some means to limit the damage EF traffic could inflict on other traffic (e.g., a token bucket rate limiter). Traffic that exceeds this limit is discarded. This maximum EF rate, and burst size if appropriate, is settable by a network administrator. Code point 101110 is used for the EF PHB.

The assured forwarding (AF) PHB group as defined in RFC2597 is the means for a provider to offer different levels of forwarding assurances for IP packets received from a customer. The customer or the DS domain provider separates traffic into one or more of these AF classes according to the services that the customer has subscribed to. Packets within each class are further divided into drop precedence level. A typical example used to describe AF PHB could be to provide different service types such as gold, silver, and bronze. Service providers in this case could guarantee that gold service gets lower delay and loss than other services. This requires allocation of resources such as buffer and bandwidth at routers and switches. Service providers need to perform admission control to ensure that they don't over-commit the provisioned capacity for each service. However, if the level of traffic generated by customers using gold service is very large (i.e., no admission control is performed), then it is likely that the silver customers may experience better service. Non-conformant packets are marked so that if insufficient resources are available, these packets will be dropped.

Four AF classes are defined; where each AF class in each DS node gets allocated a certain amount of forwarding resources (buffer space and bandwidth). Packets are assigned to a queue based on the service class. A scheduler can be configured to assign bandwidth for some queue. Within each AF class, IP packets are marked with one of three possible drop precedence values. In case of congestion, the drop precedence of a packet determines the relative importance of the packet within the AF class. A congested DS node tries to protect packets with a lower drop precedence value from being lost by preferably discarding packets with a higher drop precedence value. Congestion avoidance techniques such as random early

detection may be used for packet dropping from each queue to keep the long-term congestion low while absorbing the short-term burstiness.

There is also the Best Effort (BE) PHB group which has the lowest priority compared to other PHB groups.

5.3. DiffServ Router

Diffserv router needs a series of components such as classifier, meter, marker, shaper, and dropper commonly known as traffic conditioner. Functions of these components are provided in the following:

- *Classifier:* The packet received by the Diffserv router is first classified by a classifier module. The classifier selects packets based on the values of one or more packet header fields. Following are the two types of classification supported by Diffserv:
  - *Multifield (MF) classification:* Supports classification based on multiple fields. It may be similar to the Intserv classification whereby the 5-tuple (source and destination address, source and destination port, and protocol identification) is used to classify packets. This type of classification is required at any Intserv capable router at the edge of a network connecting to a Diffserv domain. The MF classified flows need to be marked by appropriate DSCP either by the egress router of the Intserv domain or by the ingress router of the Diffserv domain. In the latter case, the Diffserv ingress router needs to perform MF classification.
  - *Behavior aggregate (BA) classification:* Sorts packets based on the ToS field that contains the DSCP. This classification is performed in the DSCP core routers and results in faster classification.
- *Marker:* Once the MF classification process is complete, the packet is handed over to the marker. The job of the marker is to insert the appropriate DSCP value in the DS byte so that the packet receives appropriate service (PHB) in subsequent routers. Once the packet has been marked, all downstream routers need to perform only BA classification.
- *Meter:* A meter is used to compare the incoming flow with the negotiated traffic profile and pass the violating packets to the shaper and dropper or remark the packet with lower grade service using a different DSCP. The meter can be used for accounting management of the network.
- *Shaper:* A packet may be sent to the shaper module. This module may introduce some delay in order to bring the flow into compliance with its profile. The shapers usually have limited buffer, and packets that don't fit into the buffer may be discarded. The shaper buffers may accept a burst of traffic and then send it at an acceptable rate to the next hop.
- *Dropper:* A dropper performs a policing function by simply dropping the packets that are out of profile. It is a special instance of a packet shaper with no buffer.

These components (meter, marker, shaper, and dropper) are also known as traffic conditioners in the Diffserv world. Combination of these components facilitates building a scalable Diffserv network. MF classification combined with metering at the edge is scalable, as the traffic volume is not very high (in comparison to the core). The core network doesn't need to maintain per-flow state, as the classification is performed based on BA. QoS guarantees can be achieved by separating flows using different DSCP and by shaping and policing traffic.

6. QoS support in IMS using DiffServ

The proposed usage of DiffServ QoS method in the context of “End-to-end IMS QoS” is schematically shown in the following diagram. We also draw the attention of the reader to the fact that DiffServ domain could be potentially between any other two elements of the network as well, but the whole concept of utilizing DiffServ will not change in this latter scenario. So, for simplicity, we just position the DiffServ domain in one place, between the GGSN and the IMS network elements, the most probable place of deployment.

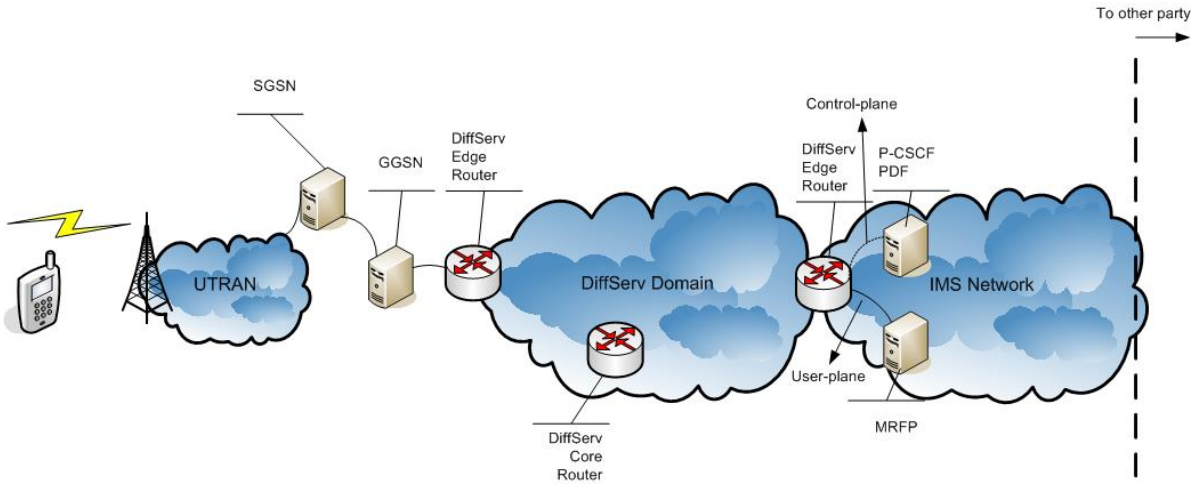


Figure 9. One possible scenario of the place of the DiffServ domain

As mentioned in the previous sections of this study, the primary PDP context is used for IMS signaling and secondary PDP context(s) are used for transmission of media. The mapping between UMTS Traffic Classes and DiffServ Code Points are done according to the following table:

DiffServ DSCP	UMTS Traffic Class	Traffic Handling Priority
EF	Conversational	N/A
AF4	Streaming	N/A
AF3	Interactive	1
AF2		2
AF1		3
BE	Background	N/A

Figure 10. Mapping Rules [10]

According to this mapping rule, primary PDP context will get Interactive UMTS traffic class and will be mapped to AF31 DSCP (AF31 Code Point = 011010) in the entrance into the DiffServ domain. Secondary PDP context(s), if carrying a real-time service, will get Conversational UMTS Traffic Class and will then be mapped to EF DSCP (EF code point = 101110).

In what follows, we will put all the information provided up to here into action, by examining the presented concepts in an actual scenario.

7. End-to-end quality of service scenario in IMS

To make clear QoS management in a IP Multimedia Subsystem, we illustrate a visio (audio and video) session between two end users. First we assume that we have two users UE (1) and UE (2) who initiate this session. UE (1) is a UMTS client connected through a UTRAN and UE (2) is a WLAN client registered to an IEEE 802.11 access point. P-CSCF (1) and P-CSCF (2) are respectively provisioning UE (1) and UE (2) and served the same S-CSCF.

We have focused on our scenario to demonstrate the integration of a Diffserv domain in the IMS transport layer. Hence, the Policy Enforcement Points are the GGSN for the UMTS domain and the IEEE 802.11AP in the other end.

First, the UE (1) performs a GPRS attach to have access to the PS-domain in the UMTS core network. Both SGSN and GGSN are involved in this process. Then UE (1) initiates a primary PDP context activation in order to transport control plan signalling using SIP. The creation of the PDP context along with the corresponding UMTS and Diffserv QoS is done by the GGSN. The primary PDP context is used primarily for signalling, so the QoS Class assigned to it in the GGSN (PEP) is Background and Best Effort in the Diffserv domain

It is important to note, that the two users must register within the IMS before starting a media session. The registration process was not detailed in our example since it is out of scope.

The initiator UE (1) sends an INVITE request to UE (2) to launch both a video and audio session. UE (1)'s request encapsulates the QoS offer in a SDP (1) (Session Description Protocol) including the media wanted, bit rate and the corresponding codecs capabilities.

SDP (1), as shown in figure 11, describes the type of the session to initiate. In our case, UE (1) specifies a video session with a bit rate of 75 Kb/s, MPEG-4 and H263 as video codecs. For the audio session, it requires a 25 Kb/s bit rate, G726 and AMR as audio codecs.

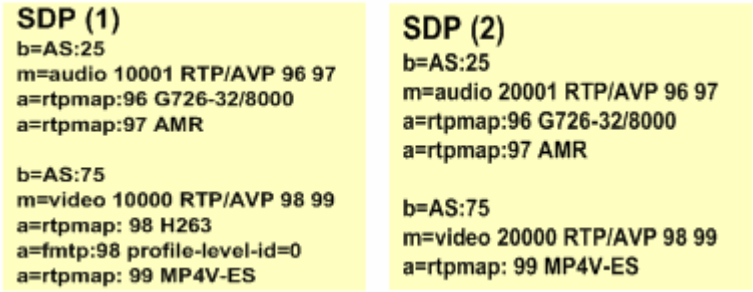


Figure 11: SDP (1) and SDP (2) parameters

The INVITE request is forwarded through P-CSCF (1), S-CSCF and P-CSCF (2) to reach finally UE (2). P-CSCF (1) and (2) checks SDP (1) against their respective local policies for conformance. S-CSCF checks both the user's profiles and local network policy for the QoS parameters (bit rate and codec) requested. In order to successfully initiate the session, SDP (1) must conform to the local policies and its user profile depending on the subscription.

In another scenario, S-CSCF could reject SDP (1) QoS parameters if they are not conforming to the user's profile. In this case, S-CSCF replies with a message "488 Not Acceptable Here" and containing an SDP for the allowed QoS parameters according to the profile and local policy. A P-CSCF behaves in the same way but it checks QoS parameters just against local policy. In our case all the checking are successful.

Upon SDP (1) reception by UE (2), it constructs its QoS parameters answer according to the request. UE (2) answers its SDP (2) within a 183 Session in Progress message. SDP (2) can be either identical to SDP (1) or a subset which means just a part of SDP (1) but conforming to QoS parameters requested by UE (1). As illustrated in the figure, UE (2) chooses only the MPEG-4 as a video codec and discards H263 but remains applicable to the QoS requested.

The SDP (2) message is not checked against local policy and user profile by the P-CSCF and S-CSCF since SDP (1) has been successfully allowed before.

Upon receiving 183 Session in Progress message with SDP (2), P-CSCF (2) makes use of SDP (2)'s parameters to generate the QoS resource authorization. Then, the Policy Definition Function (PDF) uses this authorization to allow any media defined for the session in SDP (2).

Hence, PDF pushes then these authorization parameters into the PEP associated with UE (2). In consequence, all the media meeting of SDP (2)'s requirements will be assured to have the QoS negotiated before.

In order to enforce QoS policies, PDF authorization encloses the following information [3GPP 23.207]:

- A flow filter: Identifying each flow by an IP flow 5-tuples. It includes the destination IP /Port, source IP/Port and the protocol number (UDP in general).
- A data rate and the QoS class mapped to a flow filter.

In consequence, the UE (2)'s PEP (AP) receives SDP (2) QoS parameters through the COPS protocol and add then to its policy and reserves resources. The 5-tuplues flow specification, QoS class and the bit rate will be mapped to a there parallel DSCP to ensure the same QoS.

For example, in the uplink audio flow, the QoS class is conversational and the maximum bit rate of 25 kb/s along with the IP flow 5-tuples. The QoS conversational class is mapped to the Expedited Forwarding (EF) equivalent class in the Diffsev domain. The mapping is performed according to the table shown in figure 12, proposed by 3gpp.

Src Addr	Src Port	Dst Addr	Dst Port	Proto ID	DSCP	QoS Class	Bit Rate
UE (1)	-	UE(2)	20001	17	EF	Conver	25
UE (2)	-	UE (1)	10001	17	EF	Conver	25
UE(1)	-	UE(2)	20000	17	AF4	Stream	75
UE (2)	-	UE (1)	10000	17	AF4	Stream	75

Figure 12: Policy enforcement table parameters

Diffserv DSCP	**QoS Class**
EF	Conversational
AF4	Streaming
AF3	Interactive
AF2	
AF1	
BE	Background

Figure 13: QoS class to Diffserv QoS parameters mapping

In the same way, upon reception of the 183 Session in Progress with SDP (2), P-CSCF (1) defines the QoS parameters authorization that it will be used by the PDF. The PDF will allow all the traffic that will conform to SDP (2) specifications. Hence, both the uplink and down

links are now authorized by the two PDFs conforming to the QoS negotiated. In order that the UE (1) traffic flow be authorized, PDF includes a token authorization to the forwarded 183 Session in Progress. Later, the token will be used by the GGSN (PEP) to reserve resources.

Upon reception of the 183 Session in Progress message, UE (1) maps SDP (2) specifications to their equivalent UMTS QoS parameters. Then, UE (1) activates a secondary PDP context toward the SGSN with the UMTS QoS parameters along with the authorization token given by the PDF. The activation procedure is relayed with the creation of the PDP context between the GGSN and the SGSN.

The GGSN starts to reserve resources by asking the PDF for the allowed parameters to SDP (2) using the authorization token. The exchange between the PDF and GGSN, which acts as a PEP, is done over the COPS protocol. First GGSN send a request message (COPS-REC) including the token, the PDF replies (COPS-DEC) with the QoS parameters (5-tuples, bit rate, QoS class, DSCP). Lastly to complete the 3-way handshake, GGSN confirms with a message (COPS-RPT) the compliance and allocation of the resource for the session. As a result, the secondary PDP context is created and activated for the UE (1) and corresponding to the UMTS QoS negotiated.
Now, all the resources required for the session in the Diffserv domain are allocated in the two PEPs. Hence, the data flows that conform to the policy requirement will be tagged with the corresponding DSCP and assuring the negotiated Diffserv QoS in a per-hop based forwarding.

UE (1) acknowledges the conformance with SDP (2) using a PRACK message (including SDP (2)). PRACK message indicates that session resource requirements are reserved and the session is ready to start. Once the message arrives at UE (2), it replies with “200 OK” for PRACK acknowledging the reception of PRACK from UE (1).

At this point, the two users agree on the SDP negotiated with the corresponding QoS to be used for the visio session. Then, the normal session establishment is performed in the control plane. UE (2) sends a “180 ringing” message to UE (1), followed by a “200 OK” message to accept the INVITE request sent at the beginning of the session. Lastly, concludes the session establishment by replying with an ACK message. The session starts at this level.
In the data plan, an end-to-end quality of service is now assured. In our example, the two access networks define their QoS and map them to the Diffserv domain using correspondence between them. So the flows passing through the GGSN, will be tagged according to the filter and the mappings assigned by the PDF. For example, a RTP audio flow sent from UE (1) will be filtered by the 5-tuples and tagged with the EF Diffserv Code Point. The GGSN checks also the conformance of the bit rate, if the maximum bit rate is exceeded, the packets are dropped.

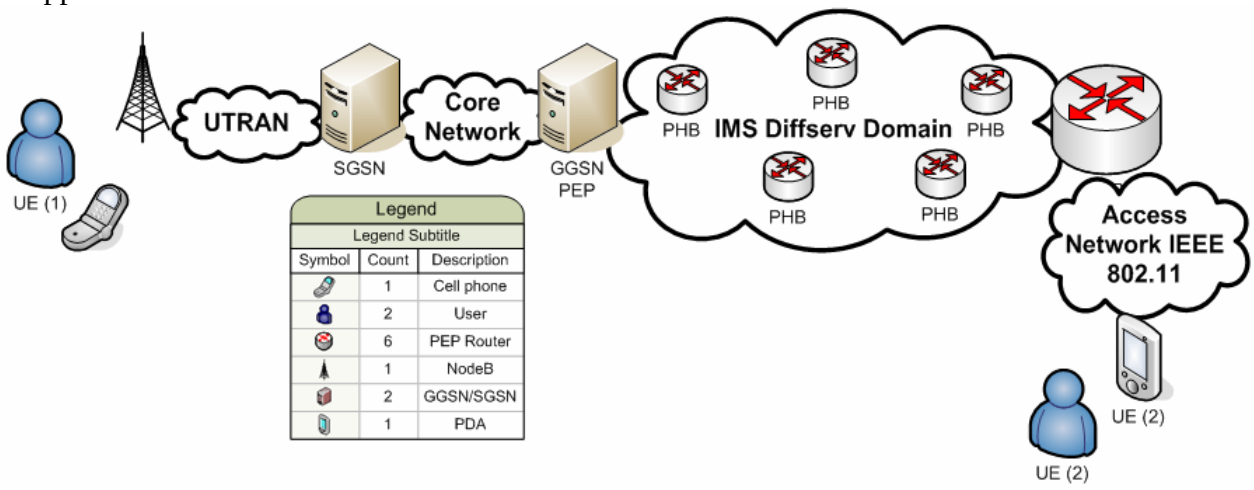
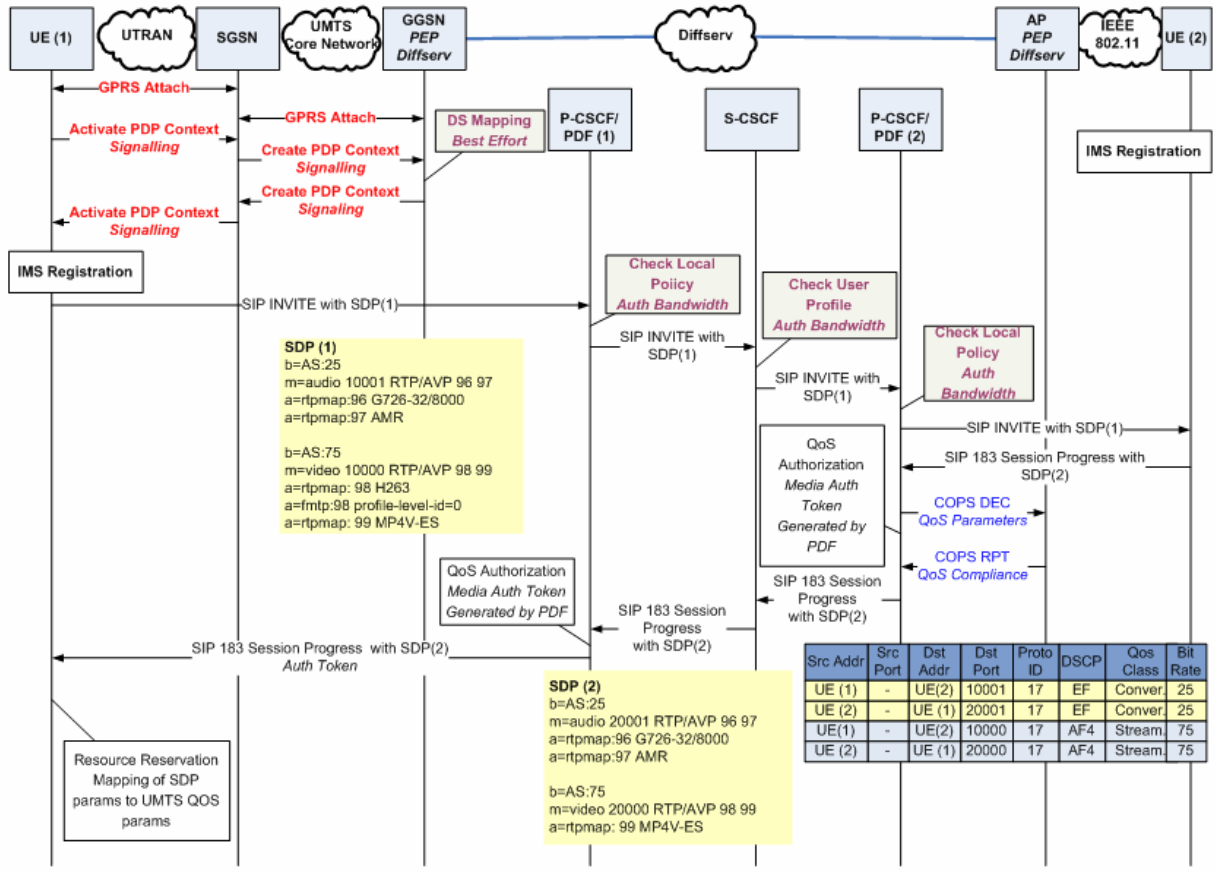
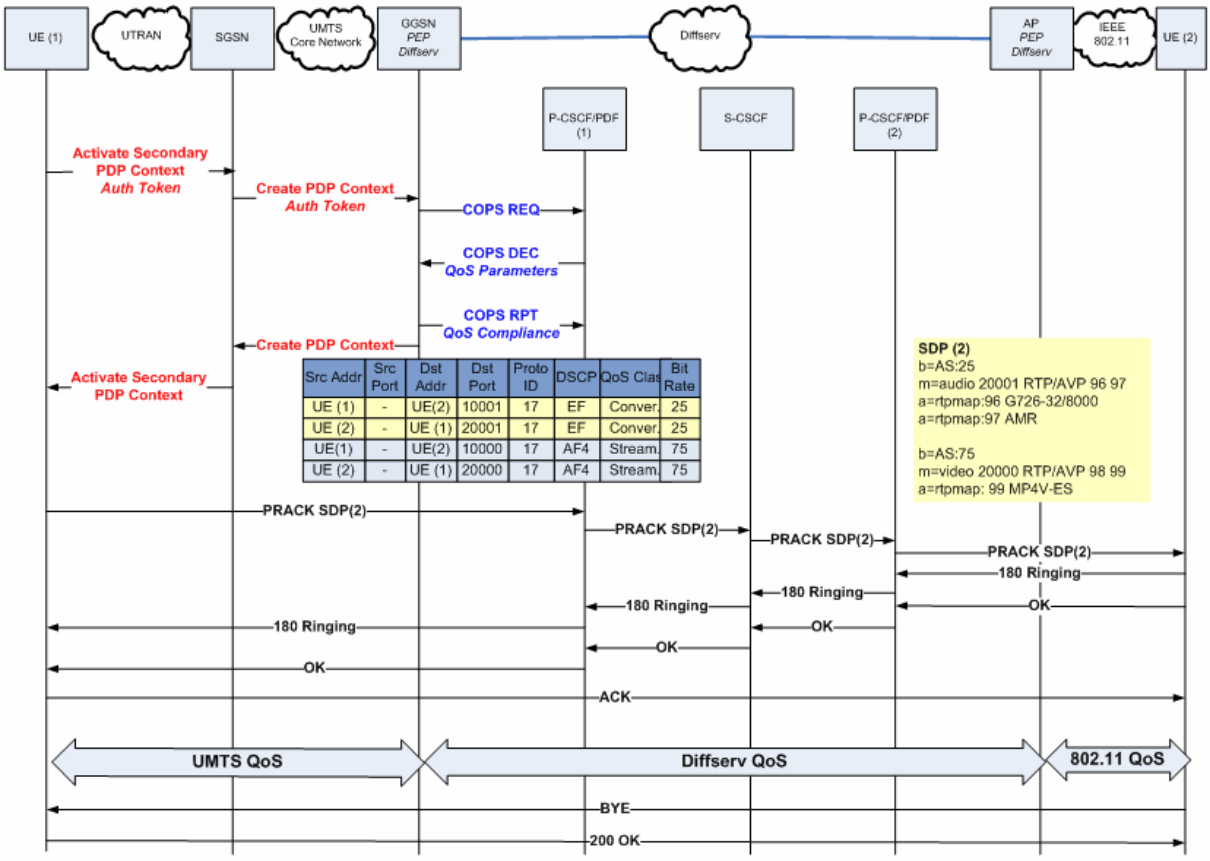


Figure 14: End-to-end QoS negotiation in IMS architecture



Src Addr	Src Port	Dst Addr	Dst Port	Proto ID	DSCP	Qos Class	Bit Rate
UE (1)	-	UE(2)	10001	17	EF	Conver.	25
UE (2)	-	UE (1)	20001	17	EF	Conver.	25
UE(1)	-	UE(2)	10000	17	AF4	Stream.	75
UE (2)	-	UE (1)	20000	17	AF4	Stream.	75

Figure 15: End-to-end QoS negotiation in IMS – Part 1



Src Addr	Src Port	Dst Addr	Dst Port	Proto ID	DSCP	QoS Clas	Bit Rate
UE (1)	-	UE(2)	10001	17	EF	Conver.	25
UE (2)	-	UE (1)	20001	17	EF	Conver.	25
UE(1)	-	UE(2)	10000	17	AF4	Stream.	75
UE (2)	-	UE (1)	20000	17	AF4	Stream.	75

Figure 16: End-to-end QoS negotiation in IMS – Part 2

Annex

New QoS Control Mechanism for Access to UMTS Core Network over Hybrid Access Networks

Creating end-to-end QoS in heterogeneous wireless/wired networks beyond 3G networks in which the access to the IP core network can be accomplished via different kinds of access networks with different technologies is essential for supporting real time application. Today there are new researchs on IMS for new functionalities and interfaces like extensions to the existing SIP signalling to resolve some of the existing problems existing in UMTS that don't let end-to-end QoS control between different technologies and domains.

The central problem in providing consistent end-to-end IP QoS services is the difficulty in configuring network devices like routers and switches to handle packet flows in a manner that satisfies their 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 operators may agree on the QoS requirements of a particular set of IP services, they may not configure thcir network devices in the same way to implement the services due to differences in their network topologies, QoS mechanisms available in the network devices, and other non-technical management requirements.

From the architecture point of view, there is no a way between the access and core network or even between different domain edge proxies to exchange the policies and limitation of their network dynamically and efficiently. On the other hand, from the signalling point of view, in the current session signalling, in the SDP inside of the SIP messages the only QoS parameters that can be indicated by the user are codec and bit-rate and the user can not express exactly his expectation about the QoS level of the required multimedia service; although it doesn't mean that the user receives a bad QoS but the user may wishes to have the choice in selecting the level of QoS for the same service because of the cost or end-device capabilities.

For example, with the current QoS parameters in SDP, "video call" will be exactly mapped to a certain QoS class beyond of user choice but for a long international video call, the caller may desires an acceptable QoS but not a high quality to reduce his costs. One possible solution to that problem is some extensions to SIP to exchange some additional QoS level information to satisfy the user QoS expectation for the requested multimedia service and help different administrative domains (or even different network technologies) negotiate SLA dynamically.

Thus the end-to-end QoS control mechanism defined by UMTS is limited to a single domain and doesn't work well for multidomain data path or inter-technology inter-operation. We can summarize the limitations of the existing system [5]:

[1] There is no E2E resource based admission control: The PCF will authorize all resource reservation if a session at application level can be established. The GGSN can perform local resource based admission control and won't care of service network or external network.

[2] PCF is limited to SIP signaled services: PCF is supposed to be in P-CSCF which is a SIP Proxy. So it can support only SIP based multimedia application. Although in release six this problem will be resolved by defining the proper interface between application servers and PCF.

[3] PCF scope limited to GGSN: PCF only serves GGSN as the policy control function and doesn't control other network elements such as inter-domain edge routers. 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.

Architecture of Heteregeneous IP Mobile Networks for E2E QoS

As discussed in previous section, the defined end-to-end QoS architecture by 3GPP has some limitations that can't support E2E QoS for multi-domain data path and in addition, the existing architecture is not flexible enough to support access of different networks with different technologies to the core network. The existing limitations can be divided in two categories:

1) Architectural problems.
2) Weakness of signalling protocols.

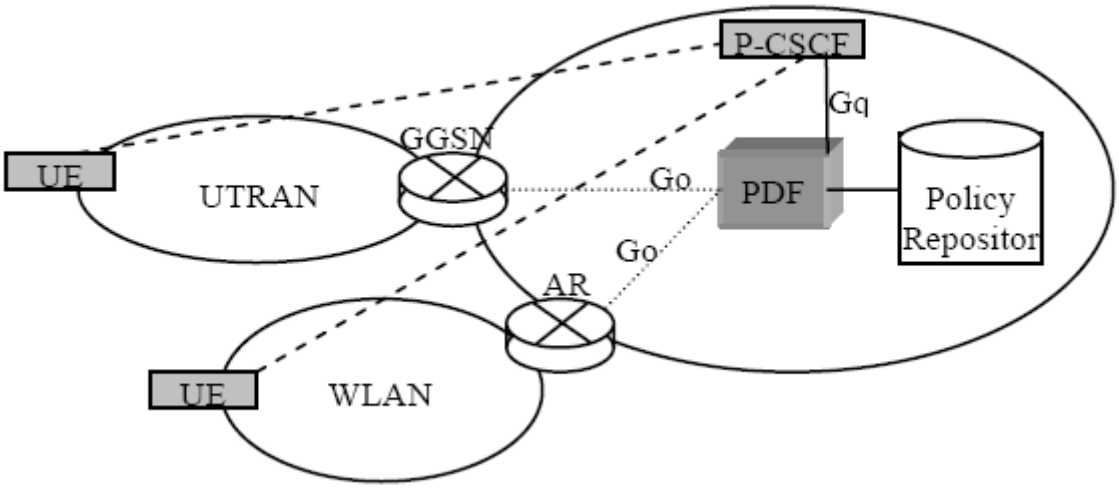


Fig 1A: Modified Architecture for Multi Domain E2E QoS :

All the Edge/Access routers are controlled by the policies defined in UMTS core network [5]

When we are accessing to IMS via another access network we need some more co-ordination between session and bearer layers; because, the QoS signaling and protocol, in addition to availability of resources in access and UMTS-CN can be completely different. For example, the IP QoS protocol in access network can be Intserv and in UMTS-CN can be Diffserv. In addition it is very likely that the four QoS classes defined in the UMTS don't have exact equivalents in other access networks. An architecture where PCF can control the edge router of other access networks is a good solution for the cases that: a) the operators of all access networks are the same or b) there is a big trust between two operators and the access network operator has agreed that the polices be pushed by the core network operator.

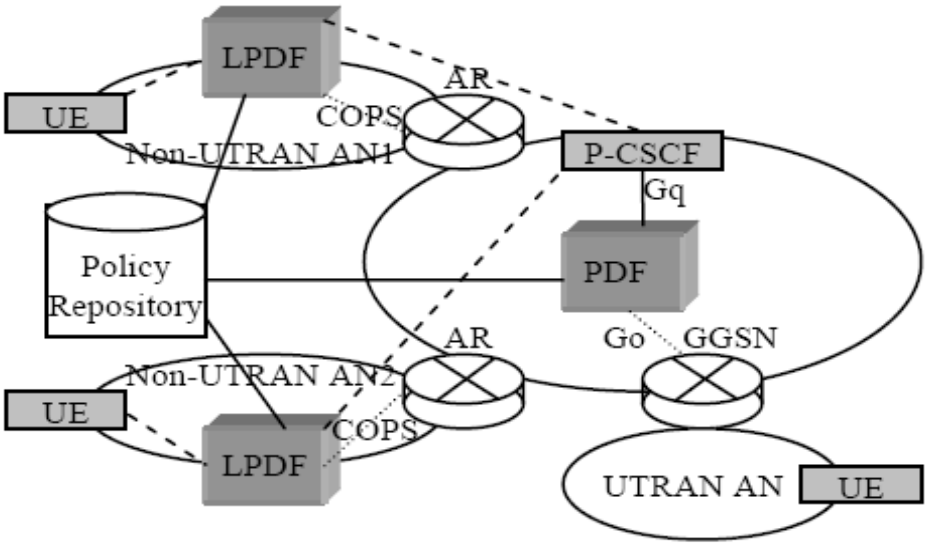


Figure 2A: Modified Architecture for Multi Domain E2E QoS :
The access networks own their Local PDF to control the AR [5]

To cope with this problem two other architectures are proposed: the Local PDF (LPDF) will exchange the policies with the PDF in the IMS (PCF) and control the edge router of the access network. Local Policy Repositories of each accesses network will exchanges their policies with a shared S-PDF and the S-PDF will control the edge routers of all access networks. Each architecture has its benefits and drawbacks and the use of them depend on the policies and capabilities of the access network operators. In the first architecture, for example for the SIP based applications, the L-PDF should support SIP and acts as a SIP proxy and this push more cost but is more dynamic for policy enforcement according to the local policies.

This method is more suitable for the access networks which had had this kind of proxy for their local services regardless of their connection to core network of UMTS; then by upgrading the existing proxies, a flexible and dynamic policy control for end-to-end QoS control will be possible. On the other side, in the architecture, there is no need for supporting the session signaling in the access networks and then the cost will be decreased. But first, the policy exchange can't be as dynamic as the previous architecture and second, the S-PDF may be the bottle-neck of the system.

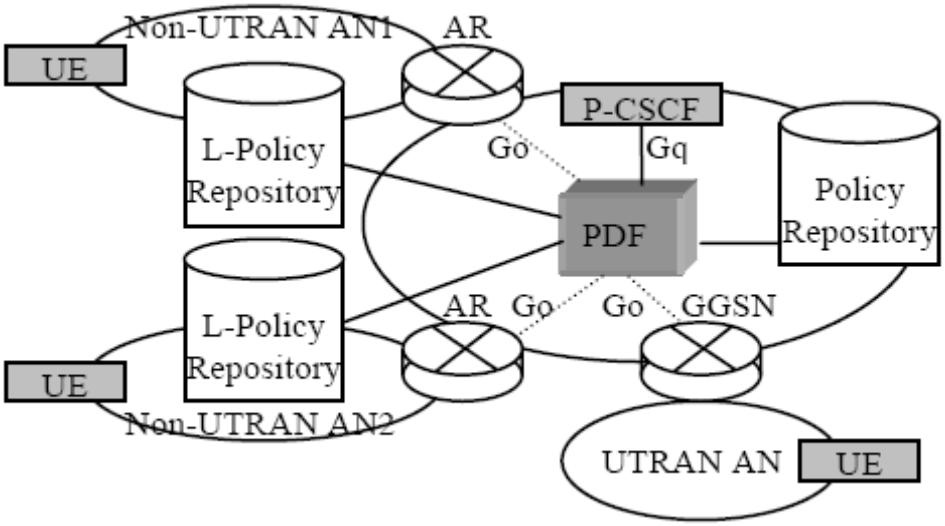


Figure 3A: Modified Architecture for Multi Domain E2E QoS :
The access networks don't have their Local PDF to control the AR.
But defines their policies themselves [5]

New Expansions on Architecture for Enriching Signaling Flow

To reach a proper E2E QoS control over the heterogeneous networks for the multimedia application, a tight co-ordination between bearer and signaling level is necessary. The policy based architecture proposed by 3GPP has some limitation from the view point of signaling and policy rules for this coordination. From the side of policy rules, it should be considered that the PCF as specified in IMS does not evaluate the policy rules and only authorize the services and negotiate the resources locally in application layer.

Table 1: Mapping of SDP media to UMTS QoS Classes defined by 3GPP

Media inside of SDP	**UMTS QoS Class**
Audio	Conversational or Streaming
Video	Conversational or Streaming
Application	Conversational
Control	Interactive Priority 1
Data	Interactive Priority 3
Others	Background

From the side of signaling, we can note that QoS parameters that can be extracted from current SDP in the body of SIP messages are too poor to allow the user to express its

expectation about the QoS level he wishes to receive for each media component. The only QoS parameters that can be extracted from the SDP are codec and bit-rate. So with this level of information there is no way except a one to one mapping between SDP QoS parameter and UMTS QoS classes. Hence, it is impossible for user to have different level of QoS for a certain media. (e.g. Video with low quality). New extensions to SIP can be defined to solve some of the existing problems and facilitate the coordination between bearer and application level for resource reservation and allow the UE to express exactly its required QoS level.

Those extensions proposed by [5] to SDP can be defined in two categories; The Traffic Information (TI) and the Sensitivity Information (SI) are added to the information of an SDP message. TI characterizes the traffic type of the bearer associated with codec (bandwidth, packet size). But SI defines the parameters like end-to-end delay, delay jitter and maximum packet loss that defines the level of quality that a user wish to have.

In [7] an extension to SIP named Q-SIP is introduced where QoS information will be carried by a SIP INVITE message in a manner that keep backward compatibility to the standard SIP elements. The proposed Q-SIP proxies detect these QoS messages and use them for resource reservation. The architecture defined in [7] makes some possibilities to exchange dynamic SLA between end-user and service network but not for inter-domain and inter-technology architecture.

By adding these informations to the SDP(183) message, the PCF in caller domain can be informed about the resource constraints in the destination domain. When this SDP reaches PCF inside of the P-CSCF in caller network, now according to the updated information about local and called domain (service) resource limitations and other negotiated policies in PCF the authorization token will be issued for the user. The authorization token won’t be issued only based on resource negotiation in application layer anymore and the resource availability in local and service domain will be considered too. This is very helpful, because when the resource reservation begins the probability of success resource reservation increase and therefore the signaling load will be decreased. Because in this strategy, if the resources are not available in network elements, it will be detected in signaling stage and authorization token won’t be issued anymore. Second, this method let the user to express its exact expectation about the QoS level.

References

[1] “BEA WebLogic Communications Platform and IP Multimedia Subsystem (IMS)”, BEA White Paper, 2006.

[2] “IMS – IP Multimedia Subsystem”, Ericsson White Paper, 2004.

[3] “IMS Service Architecture, White Paper”, Lucent Tehcnologies, 2005.

[4] “IP Multimedia Subsystem”, Motorola White Paper, 2005.

[5] Mani M. and Crespi N., "New QoS Control Mechanism Based on Extension to SIP for Access to UMTS Core Network over Hybrid Access Networks", IEEE Wireless and Mobile Computing, Networking and Communications, WiMob 2005, Montreal Canada, 22-24 August 2005.

[6] Wei Zhuang, Yung Sze Gan, Kok Jeng Loh, Kee Chaing Chua, “Policy-Based QoS Architecture in the IP Multimedia Subsystem of UMTS”, 2003, IEEE Network.

[7] Stefano Salsano, Luca Veltri, “QoS Control by Means of COPS to Support SIP-Based Applications”, March 2002, IEEE Network.

[8] Jonathan P. Castro, “All IP in 3G CDMA Networks”, 2004, Wiley.

[9] M Poikselka et al., “The IMS”, 2004, John Wiley, ISBN 0-470-87113-X

[10] T. Borosa, B. Marsic, S. Pocuca, **“**QoS support in IP multimedia subsystem using DiffServ”, ConTEL 2003. Proceedings of the 7th International Conference on Telecommunications, 11-13 June 2003 Page(s):669 - 672 vol.2

[11] Sanjay Jha and Mahbub Hassan, “Engineering Internet QoS”, 2002, ARTECH HOUSE, ISBN 1-58053-341-8

[12] M. Ali Siddiqui, Katherine Guo, Sampath Rangranjan and Sanjoy Paul, “End-to-End QoS Support for SIP Sessions in CDMA2000 Networks”, Bell Labs

[13] 3rd Genereation Partnership Project, “End-to-end Quality of Service (QoS) concept and architecture(Release 6)”, 3GPP TS 23.207-v6.6.0