An IMS Architecture and Algorithm Proposal with QoS Parameters for Flexible Convergent Services with Dynamic Requirements

A dissertation presented
by

Miguel Andrés Navarro Patiño

Advisor:

Yezid Enrique Donoso Meisel, PhD

In Partial Fulfillment of the Requirements for the of

Master

In the subject of

Systems and Computing

Systems and Computing Engineering Department
Faculty of Engineering
Universidad de los Andes
January 2010

Acknowledgment

First of all, I would like to thank my advisor Yezid Donoso; this work would not have been possible without his support and guide. I would like thank my parents, my sister, and my brother for their unconditional support during all these years. I would also like to thank my friends at Universidad de los Andes and the Systems and Computing Department directors during the last two years for their support. Finally, I would like to thank professor Harold Castro for letting me be a part of the COMIT research group and for his comments and suggestions about this work.

Contents

List of figures

List of Tables

Abstract

Quality of Service (QoS) provisioning is one of the main requirements in the 3GPP IP Multimedia Subsystem (IMS) and it has been addressed in different works since the beginning of the IMS standardization process. As a result of the fixed and mobile networks evolution, parameters standardized in IMS have changed constantly until the specification of the Policy and Charging Control (PCC) architecture that integrates IMS QoS and Charging functionalities. However, current IMS QoS specifications still have some limitations to handle service flexibility that is required to provide Internet services over IMS. In this work, we propose an enhanced IMS QoS architecture to support efficient QoS providing for flexible services with dynamic requirements. This proposal is compared against different approaches to evaluate their behavior under network saturation conditions. Simulations results show that the architecture we propose achieves efficiency and flexibility, maintaining the number of blocked and active sessions, and increasing the number of high priority sessions activated in a saturated network.

1. Introduction

The concept of convergent networks is the result of an evolution process followed by fixed and mobile networks. These networks have been under different trends of evolution, which led to the introduction of the *IP Multimedia Subsystem (IMS)* as the accepted network architecture to offer convergent services with guaranteed requirements in *Quality of Service (QoS)*, charging, security, and roaming. How to provide QoS on IMS networks is a problem that has been studied since the first IMS standardization given by the 3rd Generation Partnership Project (3GPP) in Release 5 [1]. Since then, many proposals have been presented attempting to solve different problems of providing QoS in IMS networks. However, many of them are limited to concept definitions that still need to be related with functional entities in different layers of the IMS architecture.

In the past twenty years, the evolution process followed by fixed networks has been related to the Internet's development. The Internet has influenced this process by opening a new market of services and opportunities, leading to different technological advances that allow carriers to offer those services. At the beginning of the 90's, communications were voiced-centric and fixed line carriers offering expensive voice services mainly dominated the communications market. As the number of Internet users was increasing, and it became a residential-access service, some substantial changes began to have an effect on the communications market. In the first place, the content transmitted on the web became more relevant among users and applications, such as e-mail and web browsers, constituted most of the Internet activity. At that time, the role of Internet users was understood as content consumers, defining the first web model *Web1.0* [2]. Then, new technology was introduced and Internet carriers could start offering higher transmission rates to residential users. Having services with transmission rates above 1 Mbps, Internet users started accessing to a

more graphic-oriented content. After year 2000, new multimedia applications became available offering services like video streaming, voice over IP (VoIP), video over IP, television over IP (IPTV), and peer-to-peer (P2P) file sharing [3]. With this new set of applications, the role of Internet users was modified from being an exclusive content consumer, to a content creator and consumer. This new model where users share and offer content is known as *Web2.0*. When voice started being transmitted over the Internet, the fixed line communications market was affected and it started to compete with services offered as free, excluding the Internet connection costs. As a result of this competition, service providers started looking for new and more efficient ways to offer services to their customers at lower prices, focusing on the transport networks they were using. Those traditional fixed line networks worked in the *Circuit-Switched (CS)* domain, which was very inefficient compared to the Internet's *Packet-Switched (PS)* domain. Subsequently, the first step considered by many carriers was to transform their CS networks to the PS domain. This domain migration from CS to PS in fixed networks is known as *Next Generation Networks (NGN)* [4].

At the same time that fixed networks were evolving towards NGN, mobile networks were experiencing its own evolution process. At the beginning of the 90s, industrialized countries had deployed analog networks for mobile communications, a period which is generally referred as the first generation (1G). Many technological restrictions, which were transmitted to customers as very expensive services and unaffordable devices, made it difficult to offer a mass service for mobile communications. Just until mobile second generation (2G), when GSM digital networks targeted a mass market for voice call services, mobile networks started positioning in the communications market [2]. Later, mobile carriers offered data services such as mobile Internet access, with the introduction of the General Packet Radio Service (GPRS). Despite data transmission rates were higher, prices and lack of applications made these services unpopular among users. That tendency continued until the third generation (3G) Universal Mobile Telecommunications System (UMTS) technology became available, opening a new range of services for carriers. At this point is where IMS emerges. IMS arose from the vision of integrating the flexibility of

Internet's services and the mobility of cellular networks. However, mobile networks were already offering data services with 3G networks. In consequence, IMS objective went beyond just offering services and it also included new requirements that could not be guaranteed in regular Internet services. Therefore, IMS included the required standards to guarantee security, QoS, charging, and roaming [5].

IMS was first introduced as the subsystem in charge of session control for IP services in 3G networks, and for this reason, its evolution process towards IP in mobile networks could be compared to NGN's evolution process in fixed networks. Although, user and service mobility represented a very high additional value to NGN and since mobility was already considered in IMS session control, NGN would have needed to integrate many of IMS session control features. In the end, IMS was accepted as the unifying standard Core Network (CN) for IP convergent services, increasing its initial scope to include fixed networks as an additional access network [4].

Currently, the specification of the fourth generation (4G) in mobile networks is ongoing. It began when 3GPP defined a program to standardize a next generation wireless network based on PS data transmission, in which the Long Term Evolution (LTE) program covers the design of the radio network and the air interface architecture. Later, the Service Architecture Evolution (SAE) program, also referred as the Evolved Packet Core (EPC), started working on defining the core network. When the two programs where combined, the new program was named the Evolved Packet System (EPS). The objectives followed in this new program are aligned with IMS objectives, since LTE is a new access network that may be integrated in the network architecture, and EPS still has IMS integrated in its architecture following the original vision of providing multimedia services anywhere, anytime and in any device. In this point, networks working in these programs are referred as Next Generation Mobile Networks (NGMN) [2][6].

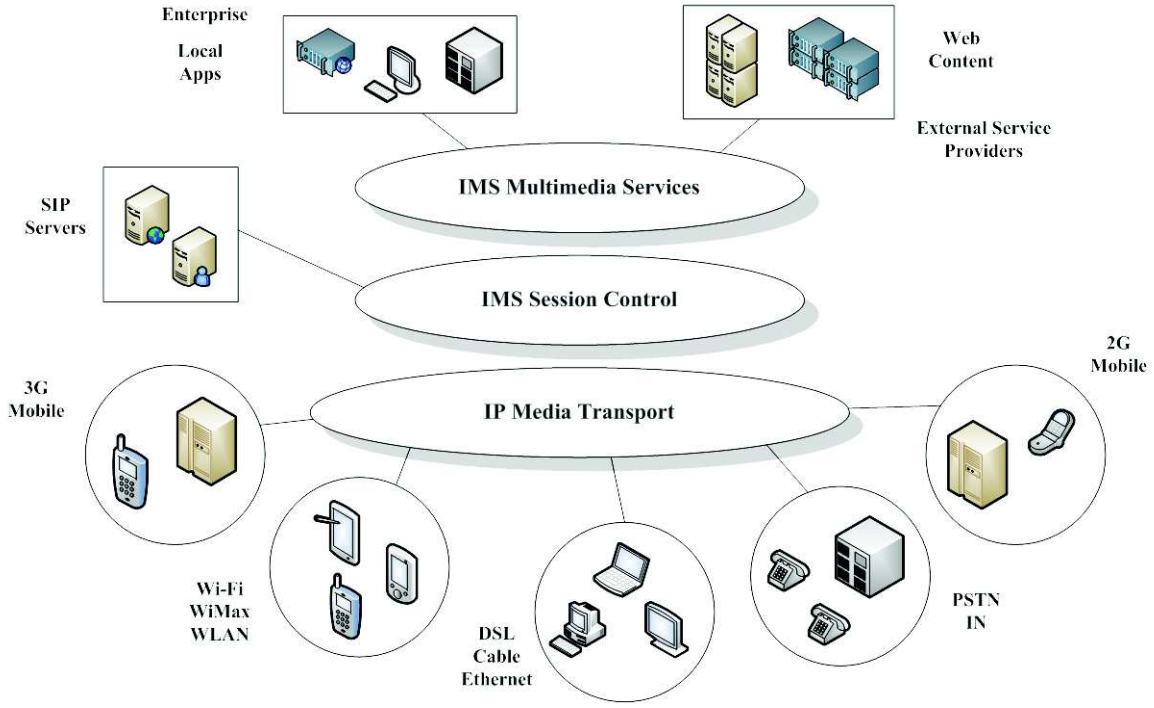


Figure 1. IMS Layers and access networks

IMS has been through a very dynamic standardization process, in which parameters defined for QoS have changed constantly. Following the layered approach specified by 3GPP for IMS, which is depicted in Figure 1, QoS was defined as a transversal requirement to all layers and it does not only concerns the transport layer as it does on the Internet. The problem IMS faces to guarantee QoS requirements becomes more complex as it is considered in a higher layer. IMS should be able to recognize different parameters for session establishment, including the user's profile, the type of media involved, and the access network. The QoS provided should be different according to the parameters defined in session establishment and during the time the session is active. Many scenarios may be presented to illustrate the QoS complexity in IMS; for example, a user may request an IPTV service to watch a football game on his mobile phone on a 3G network and during the time the session is active under those conditions, QoS should be provided accordingly; but if the same user in the same active session enters a Wi-Fi area, QoS parameters provided by the IMS network should be modified according to the parameters given by the Wi-Fi

network. In this case, changes in QoS parameters may not be evident since requirements in 3G and Wi-Fi networks for mobile devices could be similar; however, what would happen if the user does not enters a Wi-Fi area, but instead, he wants to transfer the active IPTV session to his personal computer connected to a fixed broadband access network and he also wants to watch the football game in high definition? In that case, besides transferring the session to a new device, the IMS network should be able to update the QoS parameters in every architectural layer considering the new situation.

The problem of providing QoS at the IP Media Transport layer is the same as it is in the Internet. It has been covered by several authors and the models of *Integrated Services (IntServ)* and *Differentiated Services (DiffServ)* have already been studied under different contexts. Both models apply for IMS networks; nevertheless, DiffServ model's ability to keep minimal information about the network state makes it more scalable compared to IntServ [3]. As a result, 3GPP defined DiffServ as the QoS model for the IP Media Transport layer [1][7].

For the upper layers, 3GPP has also specified the mechanisms for QoS providing. Since IMS Release 7 specification, 3GPP introduced the Policy and Charging Control (PCC) architecture, which continued until Release 9 as the mechanism for determining QoS and charging for convergent services. Although, the PCC architecture specification gives the definition of the entities involved and their basic functions, there is still much work to do to cover different scenarios and to guarantee QoS requirements. In [8], 3GPP standardized the QoS parameters applied in the service level, and also introduced the concepts of service priority and pre-emption capability and vulnerability, in order to support conflict handling between services in a state of network saturation.

The main objective of this work is to define an enhanced IMS QoS architecture, in order to support QoS providing for flexible services with dynamic requirements in an efficient way. Then, we defined an architecture that supports service relocation between different QoS levels, based on information about priority, pre-emption and the service capability to be flexible. To achieve this, we defined a new QoS parameter called the *Service Flexibility Bit*

(SFB) and a new entity named the *QoS Level Relocation Function (QoS-LRF)* in the PCC architecture.

1.1 Related Work

Several works about the IMS PCC architecture have been presented focusing on enhancements for charging and QoS functions. In IMS, QoS may be studied according to the different architectural layers, starting with the session control layer and their effects on the application and service layers. In [9], authors propose an approach to IMS policy control based on session policies. In this work, they present service integration using common functions provided by IMS, and horizontal integration as the methodology applied for multimedia service development. In addition, they show different issues affecting the current approach to policy control, and focus their work in problems affecting horizontal integration.

In [10], an enhancement to the NGN architecture is described to support Control Plane (CP)-enabled transport networks. In this proposal, authors concentrate their work on the IMS session control layer in order to integrate NGN to the CP's functionality. Nevertheless, their proposal is based on NGN Release 1 standard given by the International Telecommunication Union – Telecommunication Standardization Sector (ITU-T), and they do not take into account the 3GPP PCC architecture and its integration to NGN.

In [11], authors consider the problem of network usability when different QoS levels are offered and they propose an optimized network model for QoS provisioning and management. Therefore, a scenario in which a single operator deploys different access networks to offer multimedia services was considered with the 3GPP System Architecture Evolution (SAE) and IMS as the service delivery platform. Despite facing a similar problem and introducing a new entity to handle QoS information, this proposal did not covered dynamic requirements and the possibility of relocating a service in a different QoS level.

There are more works studying different problems in QoS on IMS, like [12], [13], and [14]; however, the problem introduced by dynamic QoS requirements, service level relocation, and their effect in the transport network, has not been considered.

2. Quality of Service in IMS

The IMS specification's scope includes requirements to guarantee different aspects in QoS provisioning, such as different types of QoS mechanisms, definitions about the timing in which those mechanisms are applied, and also multiple network domains covered in the IMS specification. To meet these requirements, 3GPP includes specifications for resource management that allow admission control and traffic policing for the different sessions entering the network. It was also necessary to distinguish between different moments for session establishment in which the QoS requested is enforced, since this process may occur before session activation, during session establishment and while the session is already active. In addition, requirements like multi-domain networks and emergency services were also considered in IMS. Many carriers are still in the process of migrating their networks to the PS domain, then it could be possible to find CS and PS domains when considering end-to-end networks; even if there was an end-to-end PS transport network, there would be many challenges to assure QoS in carrier's borders.

2.1 3GPP Policy and Charging Control Architecture

The IMS PCC architecture specified for Release 9 in [8] comprises high-level functions for both Charging and QoS. This architecture associates functions previously carried by the Flow Based Charging (FBC) and the Service-Based Local Policy (SBLP) mechanisms which were separated in previous releases. The evolution process that leads to the PCC

architecture starts in Release 5, with a policy framework specification based on the IETF's Policy Management Architecture standardized in [15], and the Common Open Policy (COPS) protocol defined in [16]. Then, in Release 6, 3GPP specifies the Service-Based Local Policy (SBLP) mechanism to differentiate QoS parameters in the service level. Later, in Release 7, the PCC architecture was first introduced, including charging functions related to the QoS decisions and the allocated resources. Finally, in Release 9 the PCC architecture includes some new specifications. Functions included in the PCC architecture to control the QoS are the following: resource allocation, event triggering, media flow establishment, and gating control. Figure 2 shows the entities and reference points associated with these functions, followed by their description.

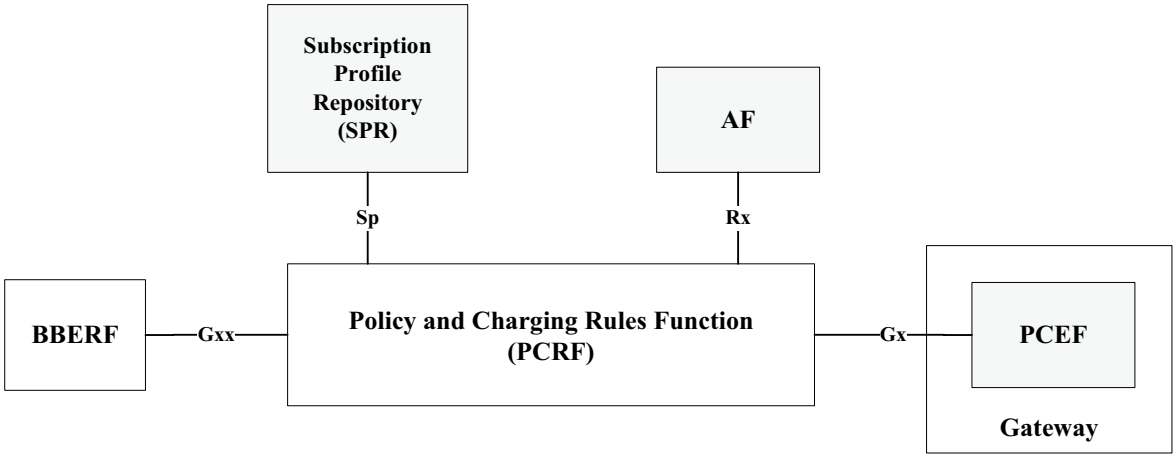


Figure 2. Policy and Charging Control (PCC) Architecture

A. PCC entities and reference points

a) Policy and Charging Rules Function (PCRF)

The PCRF is in charge of making decisions; in particular policy control decisions and flow based charging functionalities. In addition, it decides how a certain service data flow is treated by the PCEF and authorizes QoS resources. Its decisions may be based on

information received from other entities or based on its own pre-defined information. In the end, decisions taken by the PCRF are transmitted to the PCEF.

b) Policy and Charging Enforcement Function (PCEF)

The PCEF is the entity in charge of policy enforcement, QoS handling and service data flow measurement, besides other charging functionalities. It is located at the access network Gateway (e.g. GGSN in GPRS). The PCEF manages service data flows according to the information transmitted in the PCC rules by the PCRF or its pre-defined QoS configuration.

c) Application Function (AF)

In IMS networks, the AF is performed by the P-CSCF. This element offers applications that require dynamic policy and charging control. It also provides some session-related information required by the PCRF to make its decisions.

d) Subscription Profile Repository (SPR)

The SPR contains all subscriber and subscription related information needed by the PCRF. It is not necessarily a centralized entity; it may also be a distributed database across the network serving different PCRF entities. The SPR uses the Sp reference point to communicate with the PCRF, but it is not yet specified in Release 9 [8].

e) Bearer Binding and Event Report Function (BBERF)

It has the main functionalities of bearer binding, event reporting to the PCRF, and information exchange with the PCRF about the access network.

f) Reference Points/Interfaces

In Release 9, new interfaces were introduced to connect entities that evolved from previous releases. Interfaces Gq and Go evolved into Rx and Gx, respectively. One of the main differences with the new entities specification is that COPS is no longer the protocol used

to exchange policy decision information and DIAMETER is now the protocol used to send information from the AF/P-CSCF until it reaches the Gateway, through the Rx and Gx interfaces.

Moreover, the PCC Rule concept was introduced as the main information element of the PCC architecture. It is defined as the information required enabling user session detection, policy control implementation and proper charging [17]. Network operators may configure PCC rules as predefined or dynamic. Predefined rules are directly provisioned into the PCEF, unlike dynamic rules, which are provisioned by the PCRF via the Gx interface. The information related to the QoS and contained in the PCC rules is shown in Table I.

Table I. PCC Rule QoS Information

Information Name	Description
Service Data Flow Detection	PCC rules related to the service data flow detection, define the information required for detecting packets belonging to a service data flow.
Service Data Flow Template	A list of service data flow filters for the detection of the service data flow
Precedence	Determines the order, in which the service data flow templates are applied at service data flow detection
QoS Policy Control	PCC rules associated with QoS policy control define how the PCEF shall apply policy control for service data flow
Gate Status	Indicates if the service data flow detected by the service data flow template may pass (opened Gate) or it shall be discarded (closed Gate) at the PCEF
QoS Class Identifier	Indicates the identifier for the authorized QoS parameters for the service data flow
Uplink/Downlink Maximum bit rate	The UL/DL maximum bit rate authorized for the service data flow, respectively
Uplink/Downlink Guaranteed bit rate	The UL/DL guaranteed bit rate authorized for the service data flow
ARP	Allocation and Retention Priority for the service data flow

The PCC rules in this table define the information required for detecting and controlling service data flows [8].

B. Service-level QoS Parameters

The PCC architecture includes the specification of four service-level QoS parameters: QoS Class Identifier (QCI), Allocation and Retention Priority (ARP), Guaranteed Bit Rate (GBR), and Maximum Bit Rate (MBR). These parameters define QoS features that will be taken into account for further implementations of functions performed by PCC entities [8].

a) QoS Class Identifier (QCI)

The QCI is a scalar number, associated to a network element, used to describe the packet forwarding treatment in terms of performance characteristics. This value needs to be pre-configured by the operator directly into the element. Since there may be many characteristics associated to the QCI values, 3GPP standardized four characteristics: resource type, priority, packet delay budget, and packet error loss rate. Figure 3 shows the scope of the standardized QCI characteristics and Table II shows the relations between the QCI values and the standardized QCI characteristics.

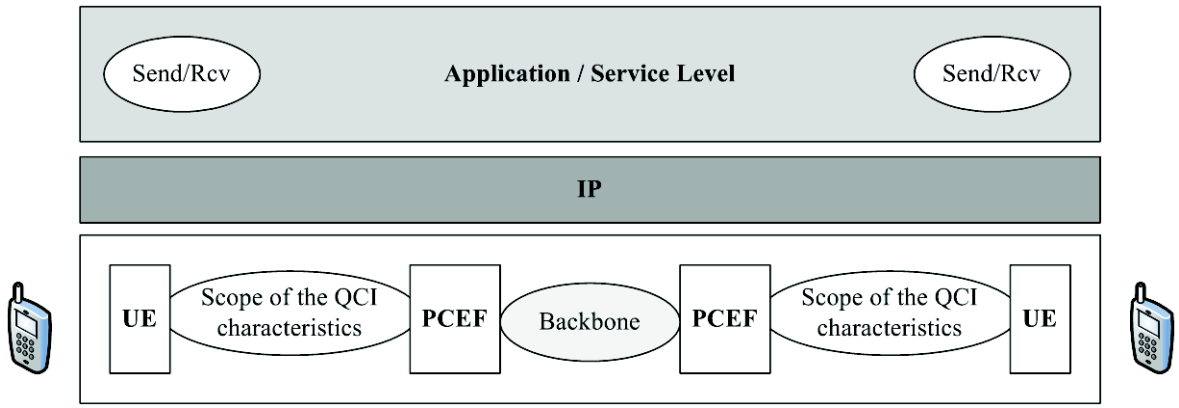


Figure 3. Scope of the standardized QCI characteristics for communication between two users. Modified from [8]

Resource Type: it defines if a session has an associated Guaranteed Bit Rate (GBR) or a Non-Guaranteed Bit rate (non-GBR) parameter to the transport network. GBR and non-GBR sessions may be related to dynamic and static policy control, respectively.

Priority: it is associated to every QCI, taking 1 as the highest priority level. It is used to distinguish between sessions entering the network or sessions that are already activated, from the same or different users.

Packet Delay Budget (PDB): it defines an upper bound for the delay a packet may experience between the UE and the PCEF. This value is intended to support link layer scheduling functions. The PDB may be understood as a maximum delay with a confidence level of 98% and it is regularly expressed in units of milliseconds.

Packet Error Loss Rate (PELR): it defines an upper bound for a rate of non congestion related packet losses. The PELR is also used to support link layer configurations.

Table II. Standardized QCI Characteristics

QCI	Resource Type	Priority	Packet Delay Budget	Packet Error Loss Rate	Example Services
1	GBR	2	100 ms	10^{-2}	Conversational voice
2	GBR	4	150 ms	10^{-3}	Conversational video (live streaming)
3	GBR	3	50 ms	10^{-3}	Real time gaming
4	GBR	5	300 ms	10^{-6}	Non-conversational video (buffered streaming)
5	Non-GBR	1	100 ms	10^{-6}	IMS signaling
6	Non-GBR	6	300 ms	10^{-6}	Video (buffered streaming) TCP-based (www, e-mail, chat, ftp, p2p file sharing, progressive video, etc)
7	Non-GBR	7	100 ms	10^{-3}	Voice Video (live streaming) Interactive gaming
8	Non-GBR	8	300 ms	10^{-6}	Video (buffered streaming) TCP based
9	Non-GBR	9	300 ms	10^{-6}	Video (buffered streaming) TCP based

The objective of the mapping of QCI values and standardized characteristics is to ensure that IMS services will receive the same minimum level of QoS in different kind of networks and in roaming scenarios [8].

b) Allocation and Retention Priority (ARP)

The ARP parameter incorporates information about the priority level, pre-emption capability (PEC) and pre-emption vulnerability (PEV). The priority level has a range of values from 1 to 15, in which 1 is the highest possible value. In the same way, values from 1 to 8 should be assigned to services with priority treatment in the network, and values from 9 to 15 should be used for roaming services. In the case of PEC and PEV, they are defined as the capability of a session to get resources that are already assigned to another session with lower priority lever, and as the vulnerability of a session to allow the loss of resources that are already assigned from another session with higher priority level, respectively. The values of the PEC and PEV parameters are set as "yes" or "no".

c) Guaranteed Bit Rate (GBR)/Non-Guaranteed Bit Rate (non-GBR)

This parameter indicates whether a session has reserved bit rate resources or not. It is associated to the resource type characteristic of the QCI.

d) Maximum Bit Rate (MBR)

The MBR parameter indicates the maximum bit rate authorized for a session.

C. Quality of Service Classifications for transport networks

Up to this point, we have presented the specifications given by 3GPP for QoS provisioning at a service level involving the IMS session control and multimedia services layers. As mentioned earlier, DiffServ is the QoS model defined for the IMS media transport layer, therefore an association is needed between DiffServ's parameters and the service-level QoS parameters discussed in the previous subsection. To define that association, 3GPP includes QoS classes for UMTS networks in the QoS concept and architecture specification given in

[7]. There are four UMTS QoS classes: conversational, streaming, interactive, and background. The principal characteristic that differentiates between these classes is delay sensitivity, going from the most sensitive (conversational class), to the less sensitive (background class). Having a characteristic to differentiate between classes, many services could be classified according to their specific requirements. Table III presents the main characteristics of the four classes and service examples.

Table III. UMTS QoS Classes

Traffic Classes	Characteristics	Services
Conversational Real Time	Preserve time relation (variation) between information entities of the stream (i.e. minimum delay) Conversational pattern (stringent and low delay)	Voice, video telephony, video conferencing
Streaming Real Time	Preserve time relation (variation) between information entities of the stream	Streaming video, IPTv
Interactive Best Effort	Request response pattern Preserve payload content	Web browsing
Background Best Effort	Destination is not expecting the data within a certain time Preserve payload content	Background download of e-mails, file transfers, SMS

Modified from [7] and [4]

The relation between UMTS QoS classes and DiffServ parameters is presented in [4], based on the GSMA specification for the GPRS Roaming eXchange (GRX). Additional distinguishing factors were included besides delay sensitivity, such as jitter, packet loss, and Service Data Unit (SDU) error ratio. Table IV presents the GSMA specification for DiffServ's parameter interpretation into the four QoS classes.

Table IV. GSMA Specifications for QoS Mapping in GRX

3GPP QoS Services		DiffServ	QoS Requirements on GRX			
Traffic Class	Example Services	DiffServ PHB	Max. Delay (ms)	Max. Jitter (ms)	Packet Loss (%)	SDU Error Ratio
Conversational	VoIP, video, conferencing	EF	20	5	0.5	10^{-6}
Streaming	Audio and video streaming	AF4	40	5	0.5	10^{-6}
Interactive ("near real-time")	Transactional Services	AF3	250	N/A	0.1	10^{-8}
Interactive	Web browsing	AF2	300	N/A	0.1	10^{-8}
Interactive	Telnet	AF1	350	N/A	0.1	10^{-8}
Background	E-mail download	BE	400	N/A	0.1	10^{-6}

Modified from [4]

3. Proposed Architecture

In the previous section we presented the QoS specifications in IMS on a service level and how they are associated to DiffServ parameters in the media transport layers. We focus on congested networks that need diverse mechanisms to solve conflicts between the different sessions trying to access the network. In current IMS specifications, these mechanisms are based on information contained in the ARP QoS parameter: priority, PEC, and PEV. Nevertheless, it is not completely specified how these parameters are used to solve conflicts, and because of this, configurations may be applied according to each carrier on its own convenience. The problem when this information is not specified, is that each carrier may apply its own configuration following 3GPP indications about service priority levels, but missing to have congruent configurations will lead to increase the probability of rejecting incoming and active user sessions.

DiffServ assigns a percentage of the network capacity to each Per-Hop Behavior (PHB), based on previous information the carrier knows about their users demands [19]. Despite having accurate information about their users demands, the dynamism introduced by IMS services, makes it very difficult to collect that information for one operator, and when relations between different operators are also introduced, there may be several scenarios in which many sessions will be rejected, adversely affecting the quality of experience (QoE) perceived by the user. At the same time that IMS introduces dynamic services, those services allow some flexibility in their QoS requirements. Flexibility could be used to define mechanisms that not necessarily resolve session conflicts by blocking or canceling sessions when there are not enough resources. We use the concepts blocking and canceling to differentiate the time when a session is rejected from the network; when a new session is trying to enter the network and that request is denied, we name it blocking the session, and

when the session is already activated by the time it is removed from the network, we name it canceling the session.

We define an enhanced IMS QoS architecture that supports flexible services and their relocation in the QoS level assigned at the IP media transport layer. First, relying on the service-level QoS parameters standardized for the PCC architecture [8], we specified a new parameter named the Service Flexibility Bit (SFB) that reflects the service capability of being relocated in a different QoS level. The SFB can be set to "1" or "0", when a session accepts being relocated or not, respectively. The enhanced PCC architecture that we propose is depicted in Figure 4. This architecture introduces a new entity called QoS Level Relocation Function (QoS-LRF), which is in charge of making decisions about session relocation in the QoS levels.

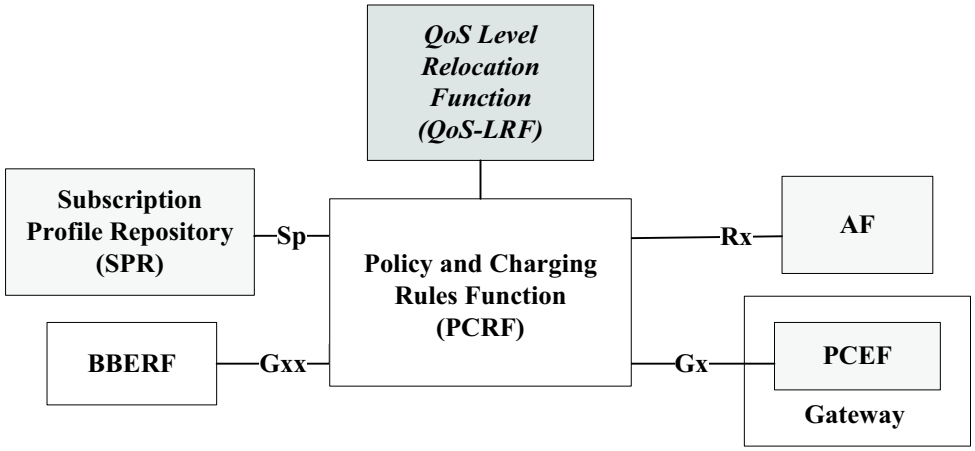


Figure 4. The enhanced PCC Architecture

3.1 QoS Level Relocation Function (QoS-LRF)

The QoS-LRF uses the information given by the SFB, priority level, PEC, and PEV, in addition to parameters about the transport network state, to decide whether a session is going to be relocated and where. In order to define how the QoS-LRF uses the information

to make the decision, we define the mapping of these parameters according to Tables II and IV, and the results are presented in Table V. First, we take the six QoS transport levels defined in Diffserv by each PHB and we assign them a priority level between 1 and 9, which is the specified range of values. To assign these values, we joined the corresponding services between Tables II and IV, starting with IMS signaling that has the highest priority value, and then the different services according to their QoS level. After that, we defined the information required from each service and that is considered by the QoS-LRF to make the relocation decisions. At this point, we divided QoS parameters in two classes: parameters associated to the QoS level and parameters associated to the session. Finally, we reduced the QoS level parameters to the Bandwidth (BW) requirement in order to reduce the problem complexity and to maintain the values specified in Table IV. The session-specific information is shown in Table V.

Table V. QoS-LRF Information

QoS Level	Priority Level	Example Services	Session Information
EF: Conversational	1	IMS Signaling	
	2	VoIP	
	3	Video conferencing	QoS level BW capacity
AF4: Streaming	4	Audio and video streaming	QoS level BW available
AF3: Interactive	5	Transactional services	PEC/V
AF2: Interactive	6	Web browsing	SFB
AF1: Interactive	7	Telnet	Service BW
BE: Background	8	E-mail	
	9	Web browsing	

According to the QoS level classification and the services that would be using each of these levels, we define the session relocation as the possibility of reserving the required network resources on a different QoS level, and transferring the session to a different level in order to provide the service according to the QoS parameters specified for the new level. The main objective of this feature included in the QoS-LRF is to benefit the session with higher priority in each QoS level, and also to optimize network resources offering the possibility to use other QoS level resources,

We use the pre-emption functions specified with the PCC architecture, the PEC and PEV parameters, which give us the possibility to use other session's resources and reserve them for a different session with higher priority level. The introduction of the SFB gives us the possibility of using the pre-emption functions in the other QoS levels before blocking the activation of a new session, or before canceling an active session with lower priority level. The heuristic algorithm used by the QoS-LRF is given in Algorithm 1.

Algorithm 1 New sessions entering the network

```
(1)  A new session enters the network
(2)  if qos-level = EF then
(3)      if availability in EF then
(4)          resources are reserved in EF
(5)          the new EF session is activated in EF
(6)      else if PEC is activated and enough resources from EF users with
(7)      lower priority and PEV activated then
(8)          resources are released from the selected EF users
(9)          EF sessions are relocated in AF (*)
(10)         released resources in EF are reserved for the new EF session
(11)         the new EF session is activated in EF
(12)     else if SFB is activated then
(13)         the PEC, PEV and SFB values from the new EF session are
(14)         saved as historical values
(15)         PEC = 1
(16)         PEV = 0
(17)         SFB = 0
(18)         if availability in AF then
(19)             resources are reserved in AF for the new EF session
(20)             the new EF session is activated in AF
(21)         else if PEC is activated and enough resources from AF
(22)         users with lower priority and PEV activated then
(23)             resources are released from the selected AF users
(24)             AF sessions are relocated in BE (*)
(25)             released resources in AF are reserved for the new EF
(26)             session
(27)             the new EF session is activated in AF
(28)         else
(29)             the new EF session entering the network is rejected
(30)     else
(31)         the new EF session entering the network is rejected

(32) else if qos-level = AF then
(33)     if availability in AF then
(34)         resources are reserved in AF
```

(35) *the new AF session is activated in AF*
(36) **else if** availability in EF **then**
(37) the PEC, PEV and SFB values from the new EF session are
(38) saved as historical values
(39) PEC = 0
(40) PEV = 1
(41) SFB = 1
(42) resources are reserved in EF
(43) *the new AF session is activated in EF*
(44) **else if** PEC is activated **and** enough resources from AF users with
(45) lower priority and PEV activated **then**
(46) resources are released from the selected AF users
(47) *AF sessions are relocated in BE (*)*
(48) released resources in AF are reserved for the new AF session
(49) *the new AF session is activated in AF*
(50) **else if** SFB is activated **then**
(51) the PEC, PEV and SFB values from the new EF session are
(52) saved as historical values
(53) PEC = 1
(54) PEV = 0
(55) SFB = 0
(56) **if** availability in BE **then**
(57) resources are reserved in BE for the new AF session
(58) *the new AF session is activated in BE*
(59) **else if** PEC is activated **and** enough resources from BE
(60) users with lower priority and PEV activated **then**
(61) resources are released from the selected BE users
(62) *the sessions from the selected BE users are rejected*
(63) released resources in BE are reserved for the new AF
(64) session
(65) *the new AF session is activated in BE*
(66) **else**
(67) *the new AF session entering the network is rejected*
(68) **else**
(69) *the new AF session entering the network is rejected*

(70) **else if** qos-level = BE **then**
(71) **if** availability in BE **then**
(72) resources are reserved in BE
(73) *the new BE session is activated in BE*
(74) **else if** availability in AF **then**
(75) the PEC, PEV and SFB values from the new EF session are
(76) saved as historical values
(77) PEC = 0
(78) PEV = 1
(79) SFB = 1
(80) resources are reserved in AF
(81) *the new BE session is activated in AF*

```
(82)            else if the new BE session priority level is 8 (highest priority in the
(83)            BE level) and enough resources from BE users with lower priority
(84)            and PEV activated then
(85)                    resources are released from the selected BE users
(86)                    the sessions from the selected BE users are rejected
(87)                    released resources in BE are reserved for the new BE session
(88)                    the new BE session is activated in BE
(89)            else
(90)                    the new EF session entering the network is rejected
(91)
(92)    end if
```

As seen in Algorithm 1, when a new session is going to be relocated, the PEC, PEV and SFB values are saved as historical values in order to recover them when resources become available at the original QoS level. In addition, when those values are saved, the new values assigned depends on how the relocation is being done; for example, if a new AF session finds enough resources at the EF level, its PEV and SFB parameters are set to "1", so that if a new EF session enters the networks, the AF session could be relocated in a different level or rejected, but just until the EF level resources are required. On the other hand, when EF and AF sessions are relocated, they go to a lower QoS level, then the PEC parameter is set to "1" and the SFB is set to "0", so that the session that is being relocated can use resources from sessions with lower priority and with the PEV parameter activated. In addition, when the session is relocated it cannot be relocated again. This means that a session cannot be transferred two levels below its initial QoS level and we will not find EF sessions in the BE level. The relocation algorithms for EF and AF sessions are given in Algorithms 2 and 3, respectively. Finally, when users leave the network and finish their sessions, if they forced other sessions relocation and those sessions are still active, they can be relocated at their initial QoS level with the historic PEC, PEV and SFB values.

Algorithm 2 EF session relocation

(1) the PEC, PEV and SFB values from the EF session are saved as historical values
(2) PEC = 1
(3) PEV = 0
(4) SFB = 0
(5) **if** availability in AF **then**
(6) resources are reserved in AF for the EF session
(7) *the EF session is activated in AF*
(8) **else if** PEC is activated **and** enough resources from AF users with lower
(9) priority and PEV activated **then**
(10) resources are released from the selected AF users
(11) *AF sessions are relocated in BE (*)*
(12) released resources in AF are reserved for the EF session
(13) *the EF session is activated in AF*
(14) **else**
(15) *the EF session rejected*

Algorithm 3 AF session relocation

(1) the PEC, PEV and SFB values from the EF session are saved as historical
(2) values
(3) PEC = 1
(4) PEV = 0
(5) SFB = 0
(6) **if** availability in BE **then**
(7) resources are reserved in BE for the AF session
(8) *the AF session is activated in BE*
(9) **else if** PEC is activated **and** enough resources from AF users with lower
(10) priority and PEV activated **then**
(11) resources are released from the selected BE users
(12) *the sessions from the selected BE users are rejected*
(13) released resources in BE are reserved for the AF session
(14) *the AF session is activated in BE*
(15) **else**
the AF session is rejected

4. Performance Evaluation

The evaluation presented in this section is based on simulations of architectural models in different scenarios. We define three architectural models in order to have different values to compare results and to have the opportunity to observe improvements given by the session relocation feature and the SFB. Then, the scenarios present different network states, varying times and service requirements for sessions entering the network.

4.1 Architectural models

The first architectural model is the reference point that gives standard values to compare results obtained with models 2 and 3. This model implements neither the session relocation feature, nor the SFB functionality. For this reason, its behavior under congestion conditions is similar to current 3G networks. It looks if there are enough resources and if there are not, the new session is rejected.

The second architectural model does implement the session relocation feature in case there are resources available in a higher level, and it also implements the pre-emption functions for using resources in the same QoS level. The sessions, which resources are released to be used by a higher priority session, are rejected. In this model, a session can only be upgraded to a higher level, so that the QoS provided is not reduced from the original requirements.

The third architectural model comprises all the functionality that we propose for the QoS-LRF. Besides implementing the second model's functionality, it implements the SFB that allows using the pre-emption functions in a lower level before rejecting the session. In this

model before rejecting any session, even sessions with lower priority and the PEV parameter activated, if the SFB is activated there is a possibility to use resources from a lower QoS level.

The relevant information we want to obtain from simulations is the number of rejected and active sessions; with this information we are able to analyze the benefits of the proposed architecture. The first architectural model gives standard values to calculate a percentage error for other models using (1).

$$\delta = \frac{V_{exp} - V_{std}}{V_{std}} * 100\% \quad (1)$$

Table VI. Deterministic Parameters

Variable	Description
N	Number of Monte Carlo simulations
λ [users/time]	Process rate
T	Simulation time
$event\_num = \lambda * T$ *[sessions]*	Number of sessions
Cap_EF	Level EF capacity
Cap_AF	Level AF capacity
Cap_BE	Level BE capacity

Using MATLAB®, we simulate implementations of the three architectural models and the network behavior. Arrival of network users is simulated as a Poisson Stochastic Process with a rate parameter λ, using Monte Carlo simulations [20]. We define the simulation deterministic parameters as shown in Table VI. Table VII shows the bandwidth requirements defined for services in the different priority levels. Afterwards, we specify the random parameters of the simulation, such as arrival time, length of the session, QoS level,

priority levels, session requirements, PEC/PEV, and SFB; Table VIII shows the ranges and distributions for the random parameters.

Table VII. Services Priority Level and Bandwidth Requirements

Priority Level	QoS Level	Service	Bandwidth
2	EF	VoIP	32 Kbps
3	EF	Video conference	1 Mbps
4	AF	Streaming	512 Mbps
5	AF	Transactional services	1 Mbps
6	AF	Web browsing	64 Kbps
7	AF	Telnet	8 Kbps
8	BE	E-mail	1 Mbps
9	BE	Web browsing	1 Mbps

Table VIII. Ranges and Distributions for Random Parameters

Parameter	Distribution	Ranges
Arrival time	Uniform	[1,T]
Session lengh	Normal	$N(\mu, \sigma)$ according to the scenario
QoS level (type of service)	According to the scenario	{EF, AF, BE}
Priority level	Uniform	According to the QoS level
Bandwidth	Uniform	According to the QoS level and the priority level
PEV/PEC	Uniform	{0,1}
SFB	Uniform	{0,1}

4.2 Simulation scenarios

Deterministic simulation parameters were fixed for all scenarios. They were selected to achieve the objective of simulating the architecture in a saturated network and therefore, having the opportunity of studying the model's behavior in that state.

The process rate λ was set to 0.95 simulations per period. Then, the simulation time was set to 2000 sec; it can also be interpreted as any other consistent unit of time. With this values,

for each simulation 1900 users try to access the network with a random service, and a service duration time following a normal distribution with μ=300 sec and σ=200 sec. Afterwards, levels capacities are set to 20 Mbps and according to the bandwidth requirements from Table VII, the state of network saturation may be achieved with at least 60 sessions in a worst case scenario. Finally, type of service is selected as the parameter that changes for each scenario and all the other parameters are defined as random with equal probability for every value in its range.

a) Scenario 1

This is the basic scenario, in which sessions are generated with the same probability for the three QoS levels.

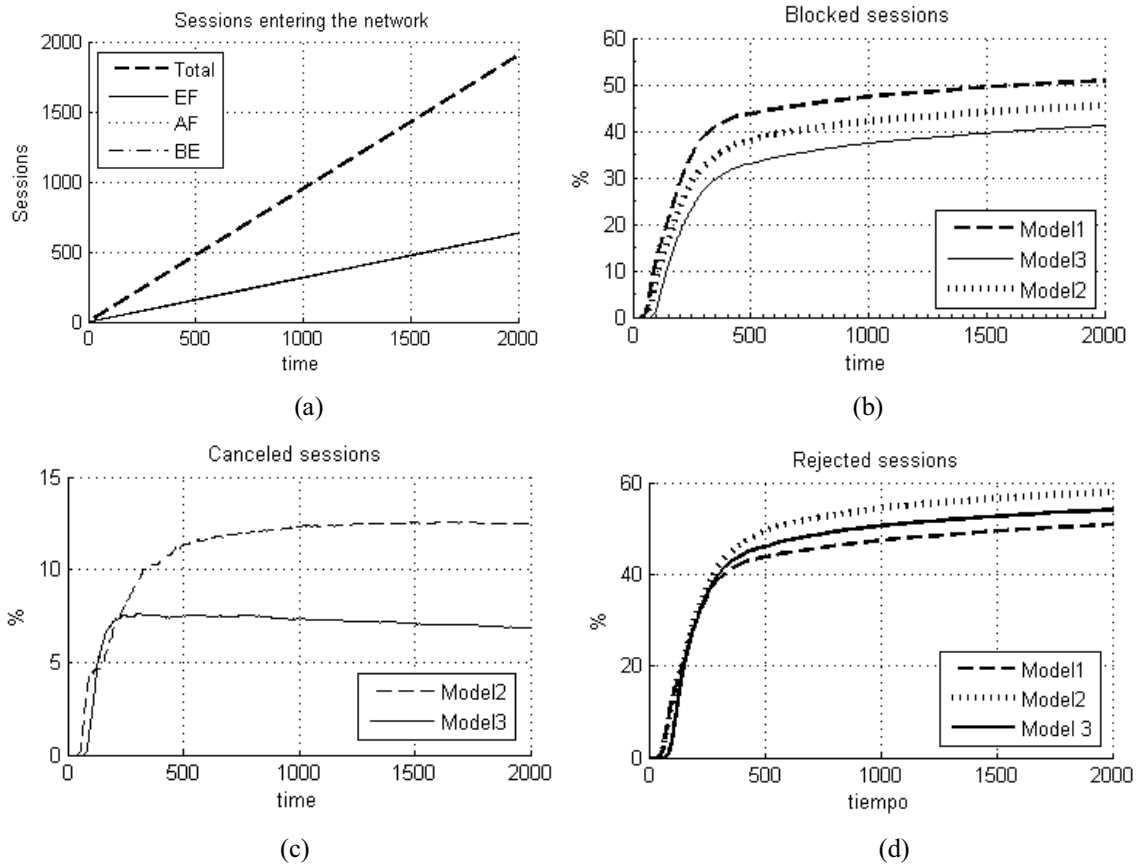
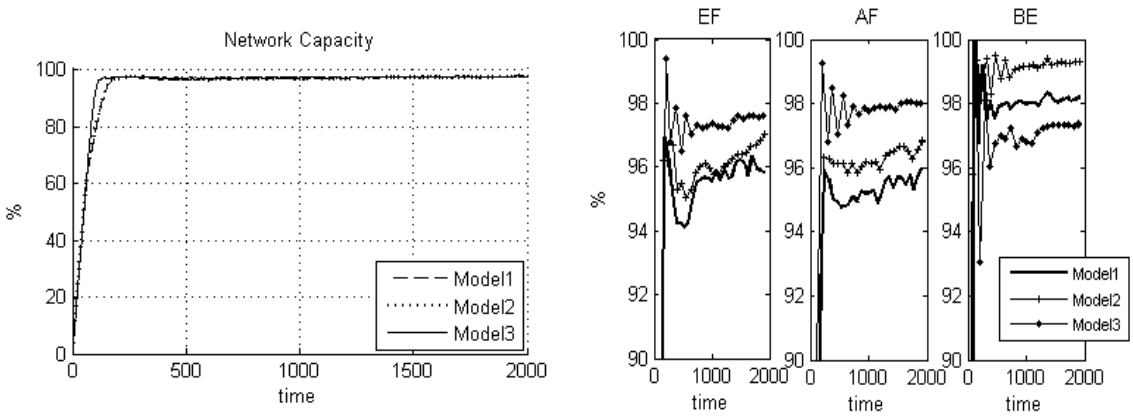


Figure 5. Information about sessions arriving and leaving the network in Scenario 1

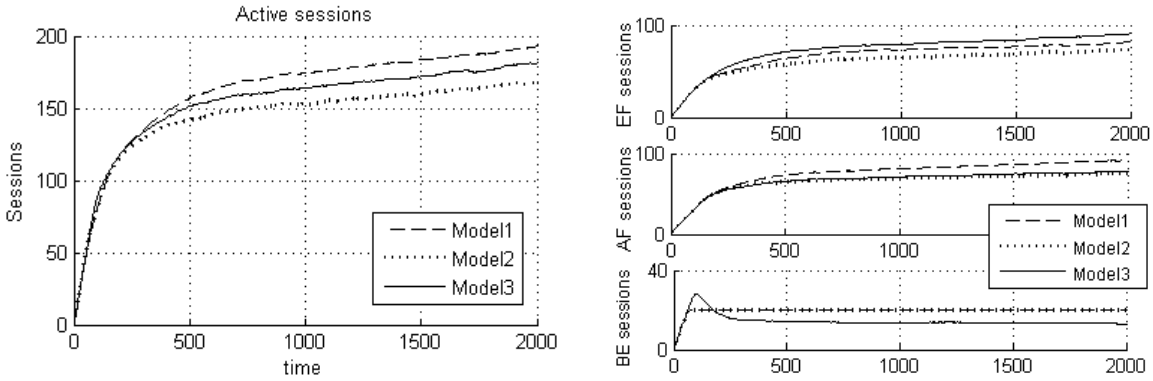
Part (a) shows the number of sessions arriving over time and the type of service requested. Parts (b) and (c) show the accumulated percentage of blocked and canceled sessions, respectively. Part (d) shows the accumulated percentage of rejected sessions.



(a) (b)

Figure 6. Network Capacity Scenario 1

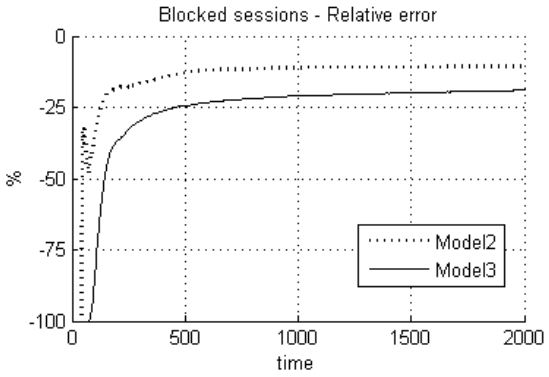
Part (a) shows the percentage of usage considering the full network capacity. Part (b) shows the percentage of usage for each QoS level in the network.



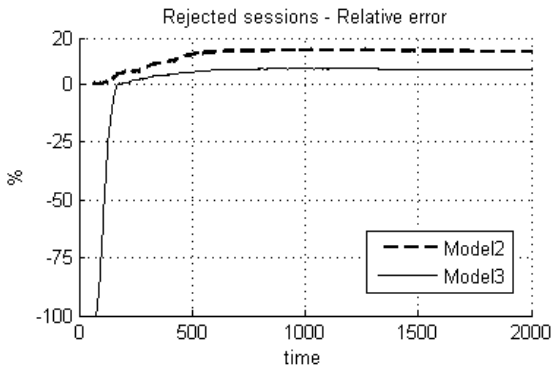
(a) (b)

Figure 7. Instantaneous number of active sessions in the network - Scenario 1

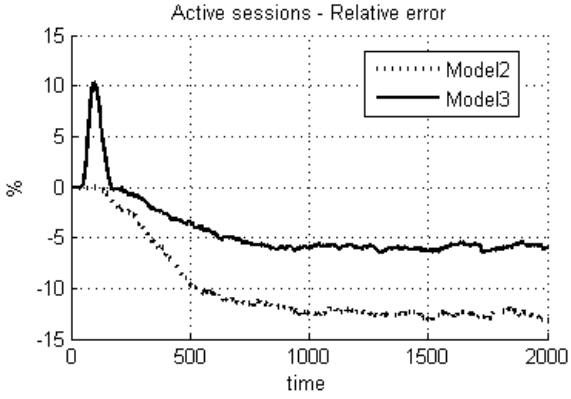
Part (a) present the total number of active sessions in the network, and part (c) shows this value for each QoS level.



(a)



(b)



(c)

Figure 8. Relative error having M1 as the reference model to compare M2 and M3 - Scenario 1

Par (a), (b), and (c) show the relative error for blocked, rejected, and active sessions, respectively.

b) Scenario 2

This scenario maintains the same values from scenario 1, except for the probability distribution used in the type of service generation. In this case, there is a 60% probability of generating an EF session, and a 20% probability for generating an AF or a BE session.

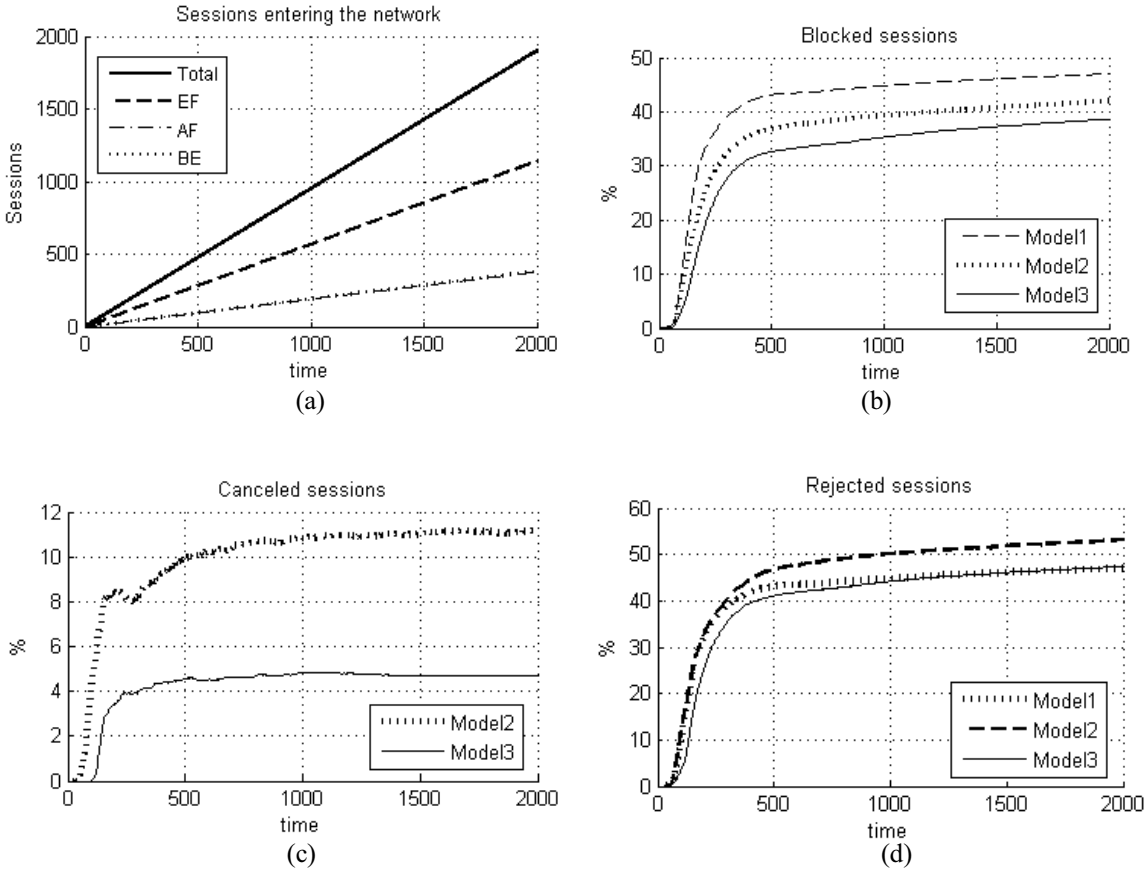


Figure 9. Information about sessions arriving and leaving the network in Scenario 2

Part (a) shows the number of sessions arriving over time and the type of service requested. Parts (b) and (c) show the accumulated percentage of blocked and canceled sessions, respectively. Part (d) shows the accumulated percentage of rejected sessions

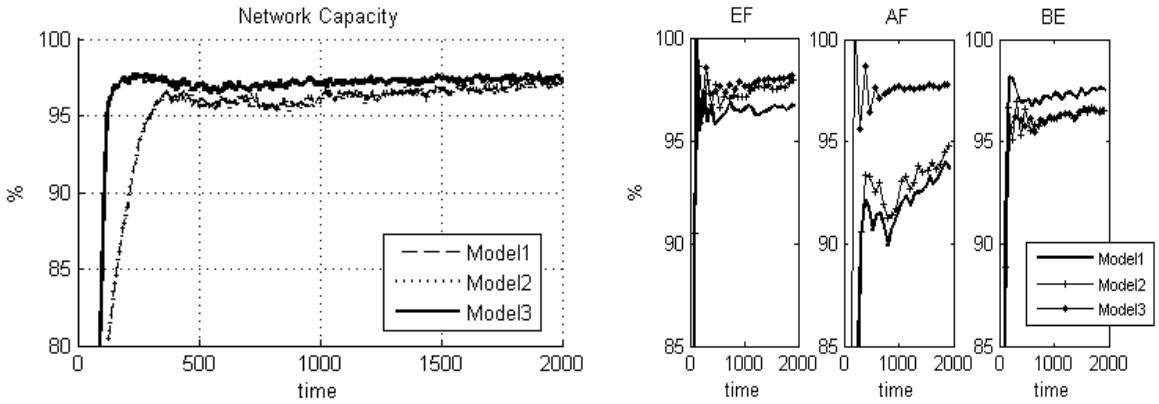


Figure 10. Network capacity - Scenario 2

Part (a) shows the percentage of usage considering the full network capacity. Part (b) shows the percentage of usage for each QoS level in the network.

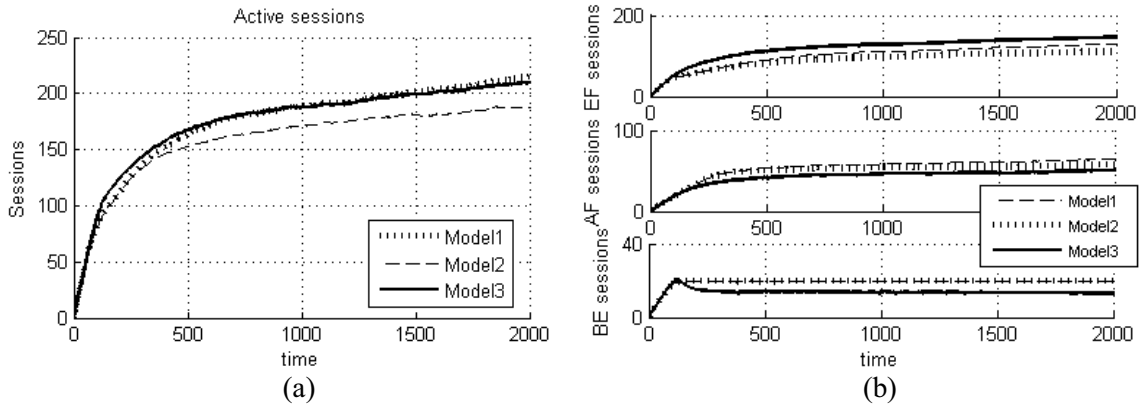
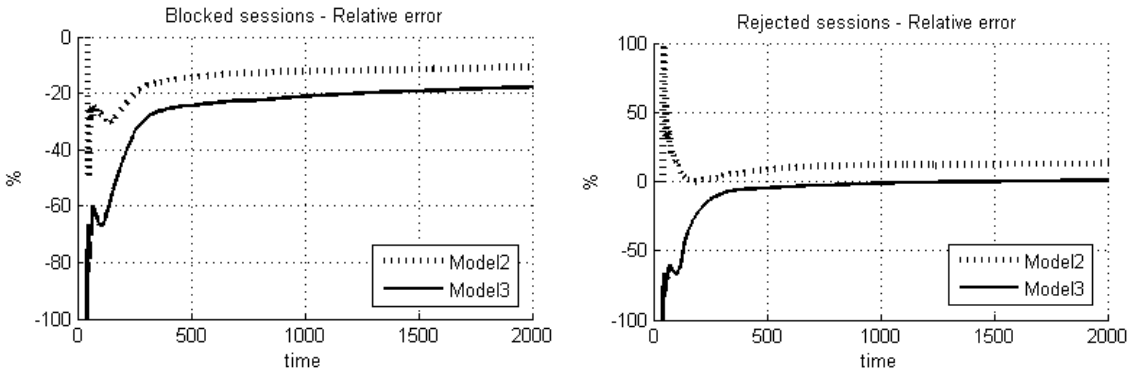


Figure 11. Instantaneous number of active sessions in the network - Scenario 2

Part (a) present the total number of active sessions in the network, and part (c) shows this value for each QoS level.



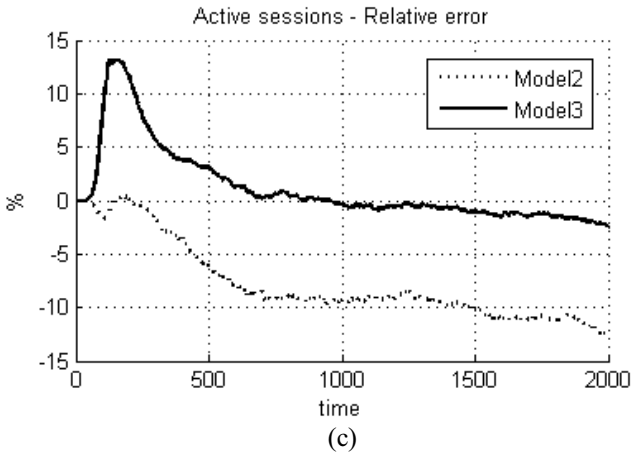


Figure 12. Relative error having M1 as the reference model to compare M2 and M3 - Scenario 2

Par (a), (b), and (c) show the relative error for blocked, rejected, and active sessions, respectively.

c) Scenario 3

In scenario 3, we change the probability, so that AF sessions are generated 60% of the time. EF and BE sessions are generated with 20% probability, and the remaining parameters continue with the same value as in previous scenarios.

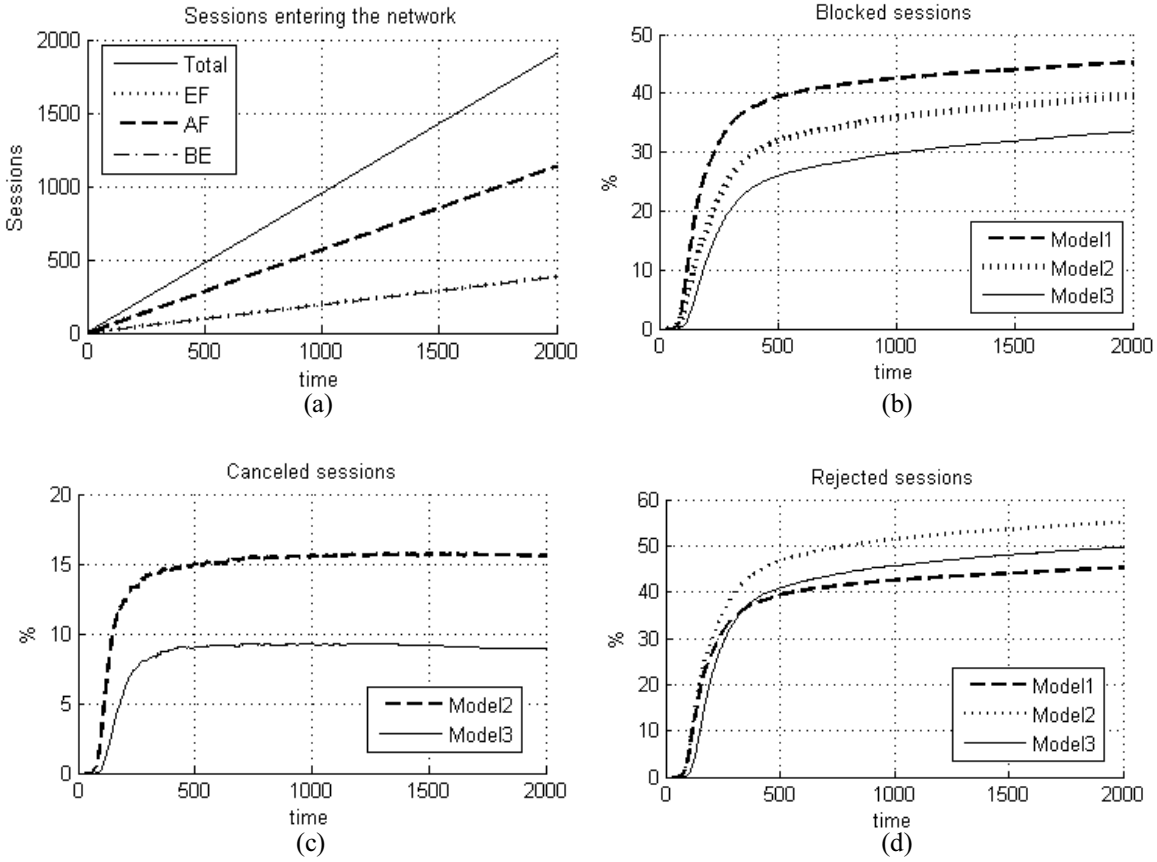


Figure 13. Information about sessions arriving and leaving the network in Scenario 3

Part (a) shows the number of sessions arriving over time and the type of service requested. Parts (b) and (c) show the accumulated percentage of blocked and canceled sessions, respectively. Part (d) shows the accumulated percentage of rejected sessions

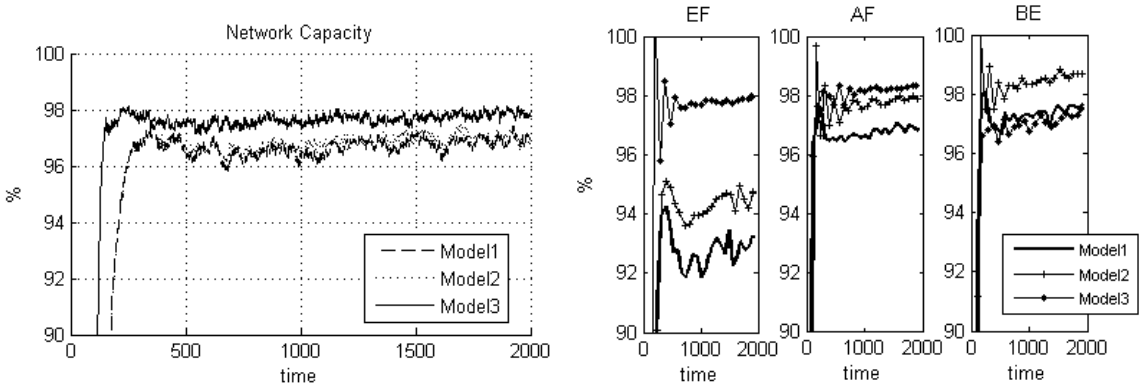


Figure 14. Network capacity - Scenario 3

Part (a) shows the percentage of usage considering the full network capacity. Part (b) shows the percentage of usage for each QoS level in the network.

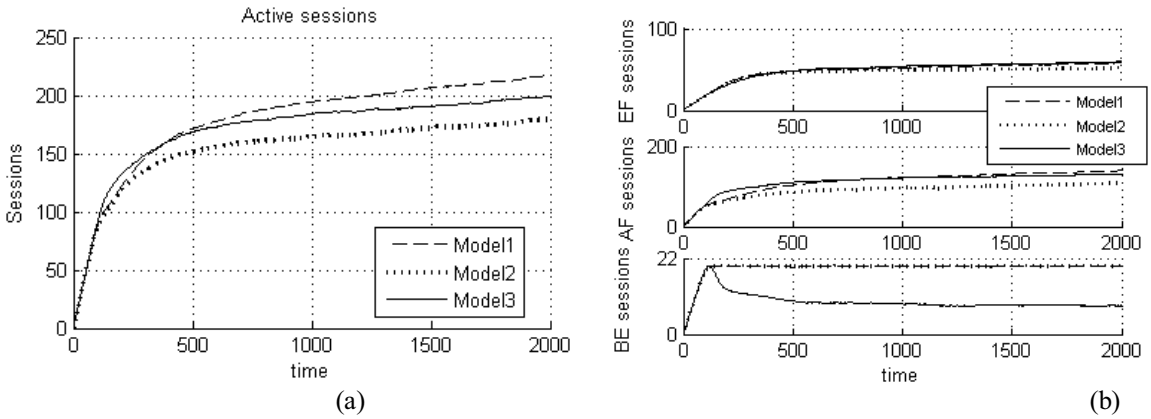


Figure 15. Instantaneous number of active sessions in the network - Scenario 3

Part (a) present the total number of active sessions in the network, and part (c) shows this value for each QoS level.

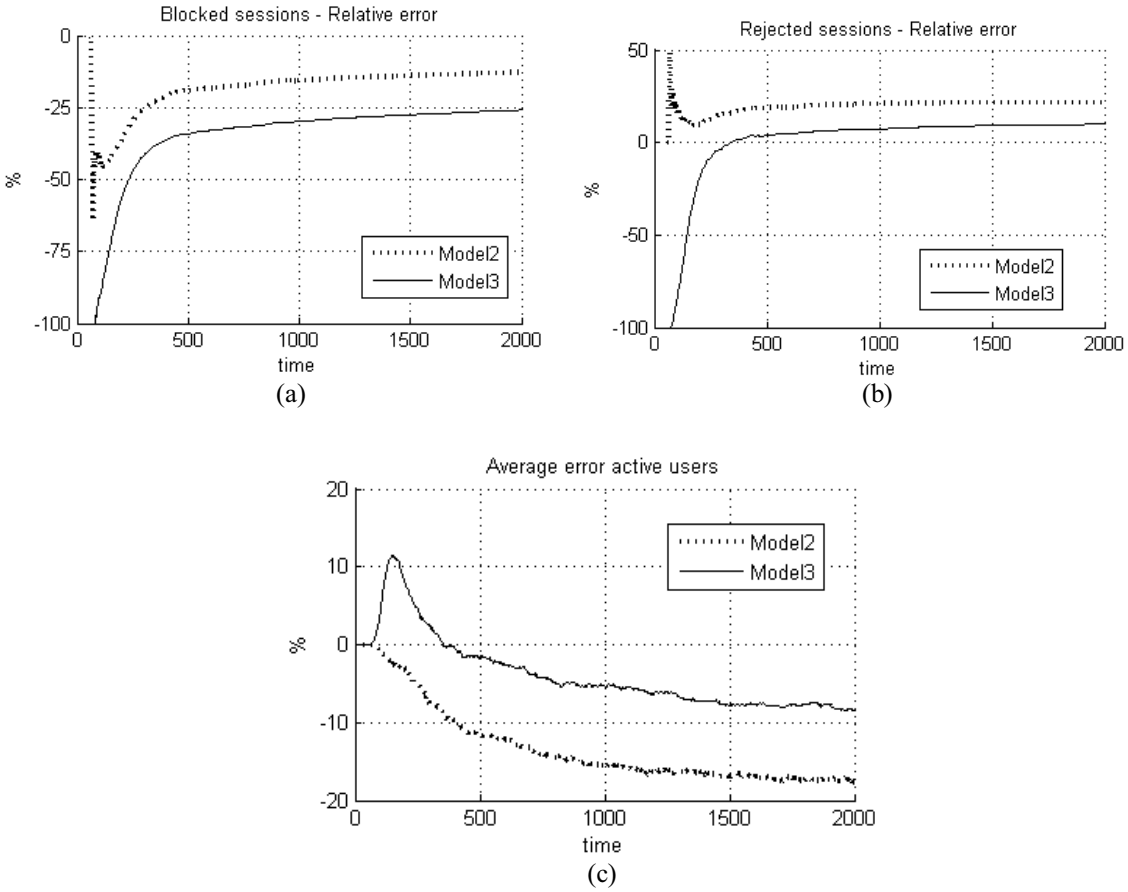


Figure 16. Relative error having M1 as the reference model to compare M2 and M3 - Scenario 3

Par (a), (b), and (c) show the relative error for blocked, rejected, and active sessions, respectively.

d) Scenario 4

Finally, in scenario 4 the probability of generating a BE session is set to 60%, while EF and AF sessions are generated with a 20% probability.

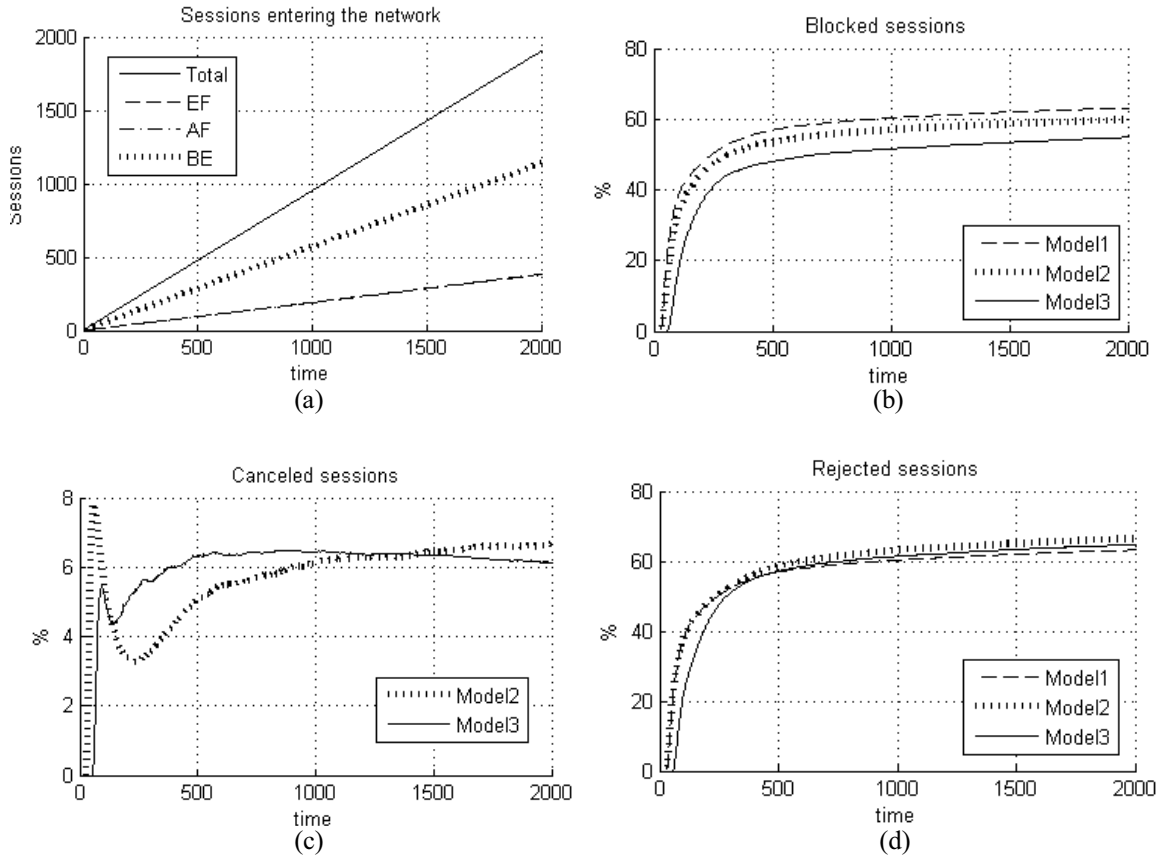


Figure 17. Information about sessions arriving and leaving the network in Scenario 4

Part (a) shows the number of sessions arriving over time and the type of service requested. Parts (b) and (c) show the accumulated percentage of blocked and canceled sessions, respectively. Part (d) shows the accumulated percentage of rejected sessions

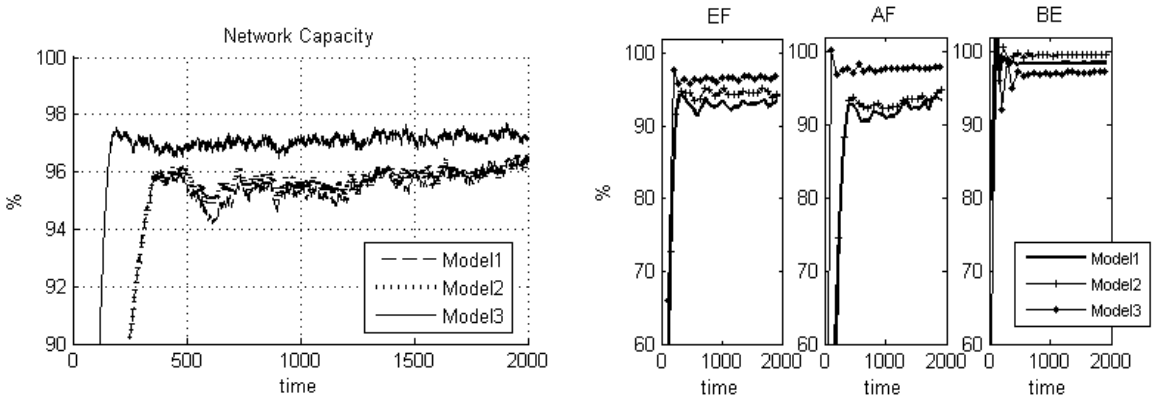


Figure 18. Network capacity - Scenario 4

Part (a) shows the percentage of usage considering the full network capacity. Part (b) shows the percentage of usage for each QoS level in the network.

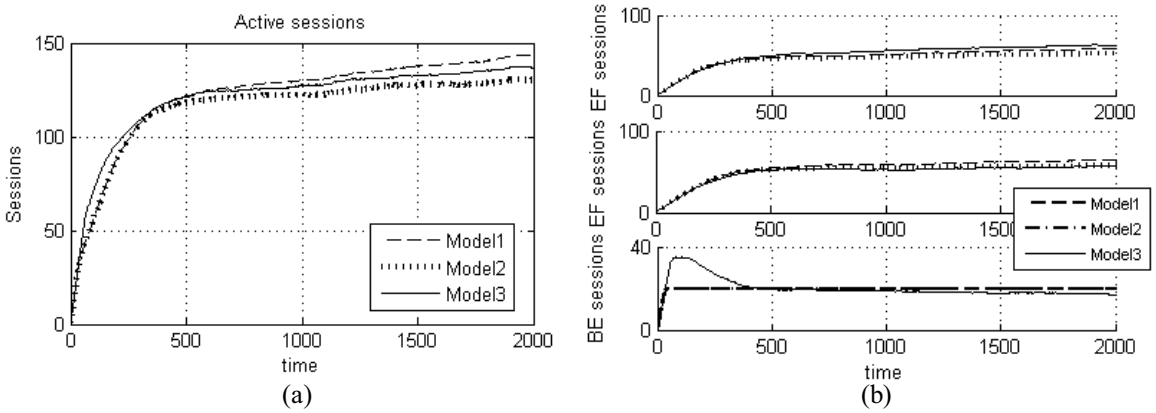
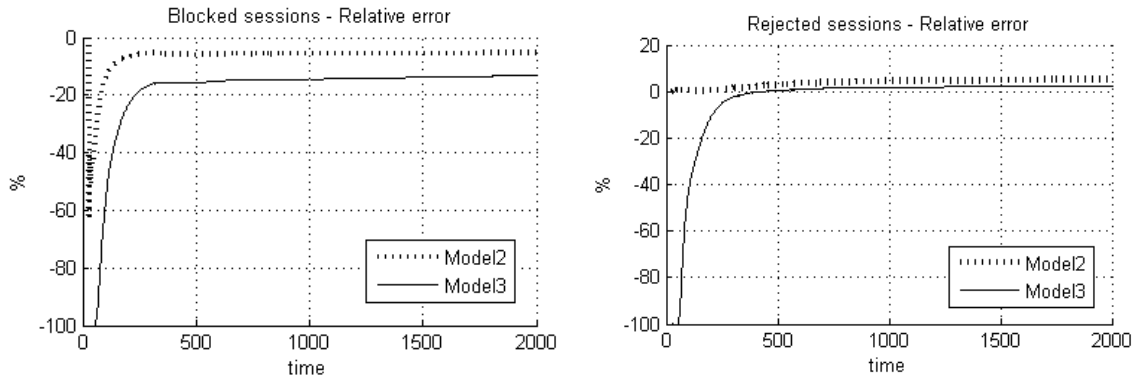
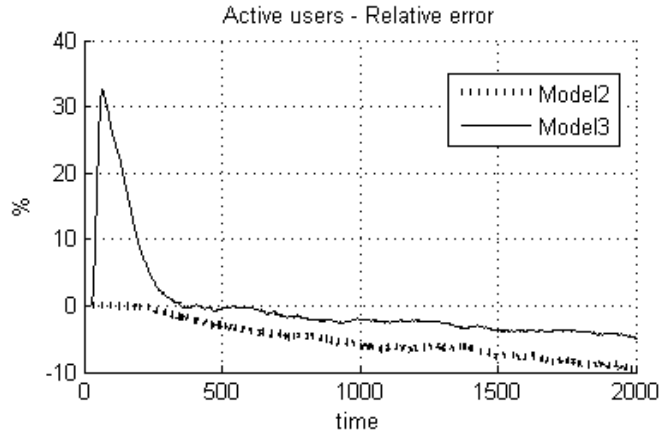


Figure 19. Instantaneous number of active sessions in the network - Scenario 4

Part (a) present the total number of active sessions in the network, and part (c) shows this value for each QoS level.



(a) (b)



(c)

Figure 20. Relative error having M1 as the reference model to compare M2 and M3 - Scenario 4

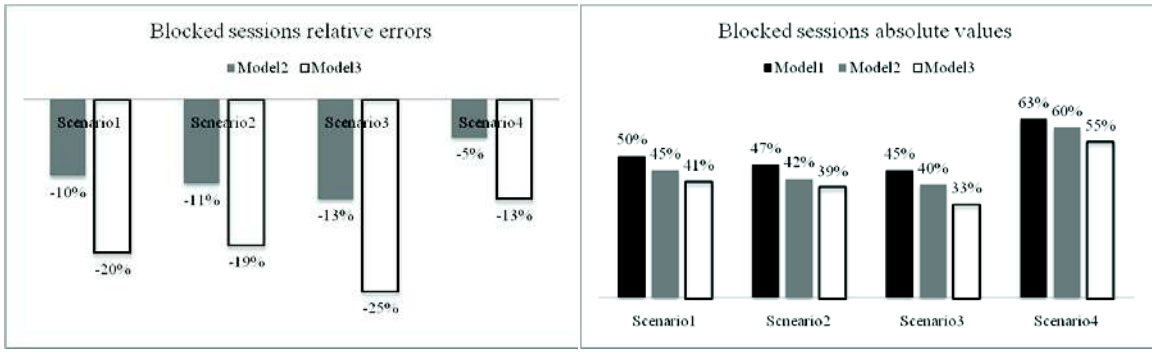
Par (a), (b), and (c) show the relative error for blocked, rejected, and active sessions, respectively.

5. Discussion

Previous simulations give information about the number of rejected and active sessions, network state, and the average error for three models in the four scenarios we selected. The four scenarios were chosen to test the models varying the type of sessions entering the network, a parameter that we selected as the most sensitive to the algorithm proposal because in the simulation it determines how the QoS level resources are used, and it also gives information about the type of carrier. That allows us to analyze the results according to the behavior a carrier would expect.

The first graphics presented, for each scenario, depict the behavior of blocked, canceled, and rejected sessions. These graphics are shown separated because they differentiate the time when an action is performed in the session. As we defined it previously, a blocked session refers to a session that could not be activated in the network due to the lack of resources. A canceled session is counted when a session that was already active in the network is removed from it at because a new session, with the PEC parameter activated and with a higher priority, is going to use the resources from the canceled session. Then, rejected sessions refer to the total number of sessions that leave the network, adding the blocked and canceled values. Comparing results from the different scenarios in the previous section, we observe that Model 3 (M3) reduces the number of blocked sessions in all scenarios compared to Model 2 (M2) and Model 1 (M1). Figure 21(a) presents the final relative error in the four scenarios, having a -25% as the best improvement for the third scenario (generating 60% percent of AF sessions). In scenarios 1 and 2, M3 has approximately the same error of -20%, and finally, in the fourth scenario M3 improves a 13% compared to M1. In addition, if we compare absolute values from the average percentage of blocked sessions, M3 maintains almost 40% blocked sessions in scenarios 1

and 2, then it is reduced to a 33% in scenario 3, and finally in scenario 4, it increments up to 55%. These results are shown in Figure 21(b).

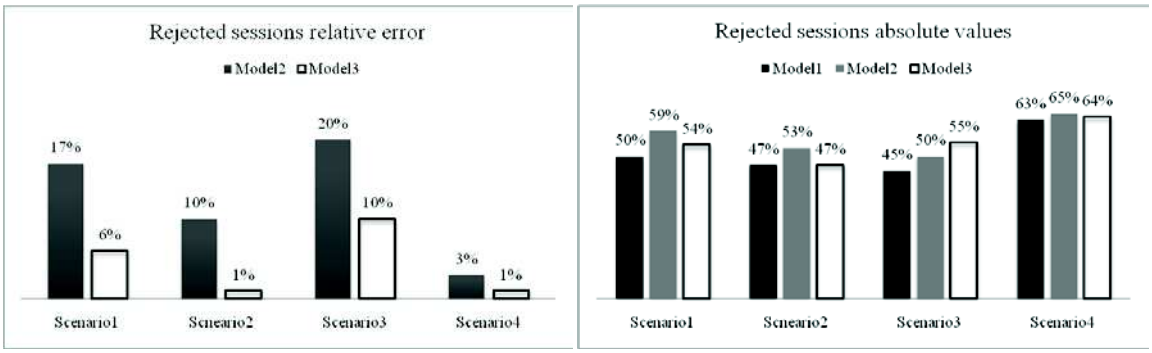


(a) (b)

Figure 21. Simulation results for blocked sessions

Part (a) shows the blocked sessions relative error comparison and part (b) shows the blocked sessions absolute values comparison at t=2000

It is very important for the algorithm not to increment abruptly the canceled session percentage because M1 does not cancel any session, according to its definition once a session reserves resources, they cannot be released until the session is finished by the user. Looking at the results in all scenarios, the number of canceled sessions in M3 may be considered to have a small effect, since the biggest value obtained is approximately 9% of canceled sessions in scenario 3. Besides, M2 always ends having a mayor percentage of canceled sessions than M3, except in scenario 4 where M3 is approximately 2% above M2 at the beginning of the simulation. The small effect of canceled sessions may be confirmed with the percentage of rejected sessions. In scenarios 2 and 4 there is no significant difference for the percentage of rejected sessions between M1 and M3, and in scenarios 1 and 3 there are relative errors of 6% and 10%, respectively. As expected, M2 increases the total percentage of rejected sessions compared to M1 and M3, and it is also very important to remark that the behavior of M3 may maintain the percentage of rejected sessions obtained in M1, the reference model. Figure 22 presents the final results for rejected sessions in relative errors and absolute values.



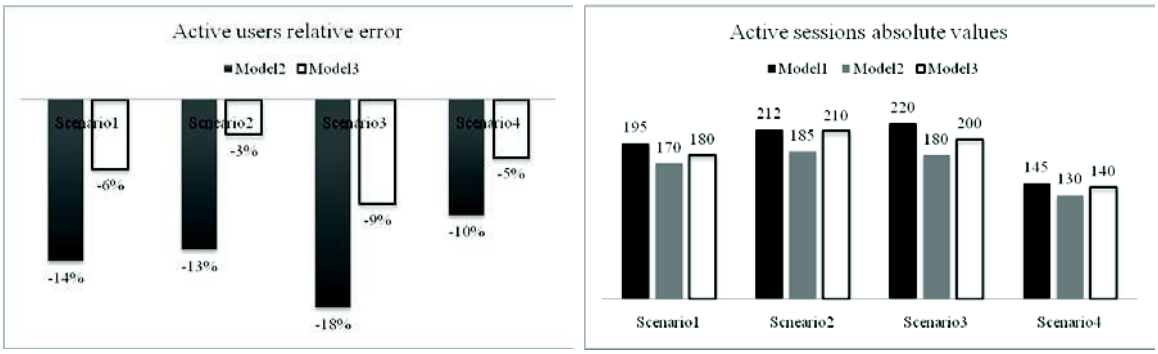
(a) (b)

Figure 22. Simulation results for rejected sessions

Part (a) shows the rejected sessions relative error comparison and part (b) shows the rejected sessions absolute values comparison at t=2000

It is not enough to demonstrate that our proposal maintains the percentage of rejected sessions to validate it, because there would be no reason to select M3 over M1. The value added given by the proposal comes from the number of active sessions and how they are distributed among the QoS levels. There are some differences between the number of active sessions in M1, M2 and M3 along the four scenarios, as shown in Figure 23. Beginning with scenario 1 and 3, they have the biggest differences, having and -6% and -9% relative errors for M3. Then, for scenarios 2 and 4 the relative error for M3 is between -3% and -5%. In this point, M3 algorithm's behavior is better for scenarios 2 and 4 than for scenarios 1 and 3, but it still has a negative relative error, which means that M3 may reduce the number of active sessions. Unlike blocked session data, the information about active sessions is considered instantaneous and in absolute values, not accumulated percentages as before. Then, taking into account the final values is not enough. If we consider the results for the number of active sessions in each QoS level presented in the previous section, we can observe there is a consistent behavior for M3 increasing the number of EF active sessions. Therefore, results obtained with M3 are according to the objectives, increasing the number of active sessions that have the highest priority level, and M2 does not increase the number of active EF session compared to M1 in any scenario. Under saturation conditions,

the number of EF active sessions in M1 is higher than in M2, validating the importance of the SFB implementation in M3.



(a) (b)

Figure 23. Simulation results for active sessions at the end of the simulation

Part (a) shows the active sessions relative errors to Model 1 at t=2000 sec. Part (b) sows the active sessions absolute values at t=2000 sec.

Bringing previous observations together, the simulations results show that our proposal, implemented in M3, may be a feasible implementation for the four considered scenarios, although it has a better behavior in scenarios 2 and 4. In scenario 2, it is very important to see that M3 maintains the same values as M1 for both rejected and active sessions. There is significant reduction of the number of active AF users, but since EF sessions are arriving with three times AF sessions probability, it may be considered as an accepted tradeoff. Then, Scenario 4 gives important results because M3 also maintains very small differences with M1 in rejected and active sessions. In this scenario, the number of BE sessions entering the network is higher compared to the other QoS level sessions. Finally, scenarios 1 and 3 present higher differences between M1 and M3; nevertheless, we consider them feasible scenarios because they maintain the model's objective and increase the number of high priority sessions with a higher tradeoff in the number of sessions rejected from the network.

Finally, analyzing the complexity of the algorithms, we find that in a regular behavior it would have to relocate more sessions for each session entering the network, with

complexity *O(n)*. In a worst case scenario, every session would have to relocate every session at the same QoS level, then the complexity would be $O(n^2)$

Conclusions

In this work we present an efficient and enhanced IMS QoS architecture to support QoS providing for flexible services with dynamic requirements. Our approach follows the 3GPP QoS specifications and is based on the PCC architecture. We propose an architecture including new features in the PCRF entity given by the concept of session relocation and the introduction of the QoS-LRF and SFB. The proposed heuristic algorithms for the QoS-LRF use information already available at the PCRF according to the PCC architecture specifications. These algorithms are evaluated with Monte Carlo simulations that include three models of implementation and four scenarios. The models represent different states of implementation, defining the first model as the reference; the second model is defined as an intermediate implementation, and the third model implements all the proposed elements. Scenarios were defined to study the behavior of the algorithm in networks requesting different types of services associated to a QoS level. The first scenario defined a network in which service sessions were generated with the same probability for the different QoS level; the three remaining scenarios incremented the probability of generating a session for one of the QoS level.

According to the three model simulations, our proposal overcomes the first two models, which offer a valid implementation of current 3GPP PCC architecture specifications. The results obtained for the number of rejected and active sessions validate it, and for this reason, our proposal would have a good performance for carriers with customers requesting more EF and BE services. Furthermore, for carriers with customers requesting all types or services at the same rate, or requesting more AF services, the algorithm achieve the objectives but with some tradeoffs for its implementation that would need to be evaluated.

The architecture proposal achieves the objectives of efficiency and flexibility. Efficiency may be analyzed according to how network resources are used. The objective validation is

given by the simulations showing that implementing our proposal, the number of rejected and active sessions is maintained and at the same time, the number of high priority sessions is increased, then network resources are properly assigned according to the priority level. Flexibility is achieved with the definition of the SFB and the algorithms implementations. They offer the possibility of relocating a session in a lower QoS level, before it is rejected from the network. Other important contribution of this work is that carriers would have the possibility of assigning different priorities to the same service within the same QoS level, and offer the service at different rates controlled by the PCC architecture charging mechanisms. Finally, the complexity of the algorithms, $O(n)$ in regular behavior and $O(n^2)$ in worst case scenario, allows us to conclude that our proposal is scalable in the number of sessions entering the network.

For further study, we will continue with the message flow analysis required to implement our proposal and a prototype implementation. We will also study scenarios involving different carriers and roaming services, which could be implemented in the prototype.

References

[1] 3GPP, "IP Multimedia Subsystem (IMS); Stage 2", Release 5, TS 23.228 V5.15.0, June 2006. http://www.3gpp.org/ftp/Specs/html-info/23228.htm

[2] M. Sauter, *Beyond 3G – Bringing Networks, Terminals and the Web Together*. John Wiley & Sons Ltd., United Kingdom, 2009.

[3] J. F. Kurose and K. W. Ross, *Computer networking: a top-down approach*, Pearson Education Inc, USA, 2008.

[4] R. Copeland, *Converging NGN wire line and Mobile 3G Networks*, CRC Press, USA, 2009.

[5] G. Camarillo and M. A. García-Martín, *The 3G IP Multimedia Subsystem (IMS): Merging the Internet and the Cellular Worlds*, 2nd Edition, John Wiley & Sons Ltd., England, 2006.

[6] T. Magedanz, A. Diez, M. Corici, and D. Vingarzan, "Understanding NGMN and Related Technologies – LTE, EPC and IMS", IMS Workshop 2009, Tutorial 4, Fraunhofer FOKUS. Berlin, Germany, November 2009.

[7] 3GPP, "*Quality of Service (QoS) concept and architecture*", Release 9, TS 23.107 V9.0.0, December 2009. http://www.3gpp.org/ftp/Specs/html-info/23107.htm

[8] 3GPP, "*Policy and charging control architecture*", Release 9, TS 23.203 V9.0.0, March 2009. http://www.3gpp.org/ftp/Specs/html-info/23203.htm

[9] G. Camarillo, T. Kauppinen, M. Kuparinen, and I. M. Ivars, "Towards an innovation oriented IP multimedia subsystem", Communications Magazine, IEEE, volume 45, issue 3, pp. 130-136, March 2007.

[10] F. Baroncelli, B. Martini, V. Martini, and P. Castoldi, "Supporting control plane-enabled transport networks within ITU-T Next Generation Network (NGN) architecture", IEEE Network Operations and Management Symposium - NOMS 2008, pp. 271-278. April 2008.

[11] M. I. Corici, F. C. de Gouveia, and T. Magedanz, "A Network Controlled QoS Model over the 3GPP System Architecture Evolution", The 2nd International Conference on Wireless Broadband and Ultra Wideband Communications - AusWireless 2007, pp. 39. August 2007.

[12] R. Good and N. Ventura, “End to end session based bearer control for IP multimedia subsystems”, IFIP/IEEE International Symposium on Integrated Network Management IM '09, pp. 497-504. June 2009.

[13] M. Ageal, R. Good, A. Elmangosh, M. Ashibani, N. Ventura, and F. Ben-Shatwan, “Centralized policy provisioning for inter-domain IMS QOS”, EUROCON ’09, pp. 1793-1797. May 2009.

[14] S. Tompros, C. Kavadias, D. Vergados, and N. Mouratidis, “A Strategy for Harmonised QoS Manipulation in Heterogeneous IMS Networks”, Wireless Personal Communications, vololume 49, number 2, pp. 197-212. Springer Netherlands, August 2008.

[15] R. Yavatkar, D. Pendarakis, and R. Guerin, “RFC2753 - A Framework for Policy-based Admission Control”, Network Working Group, January 2000. http://www.rfc-editor.org/rfc/rfc2753.txt

[16] D. Durham, J. Boyle, R. Cohen, S. Herzog, R. Rajan, and A. Sastry, “RFC 2748 - The COPS (Common Open Policy Service) Protocol”, Network Working Group, January 2000. http://www.rfc-editor.org/rfc/rfc2748.txt

[17] A. D. Albaladejo, F. C. de Gouveia, M. I. Corici, and T. Magedanz, “The PCC Rule in the 3GPP IMS Policy and Charging Control Architecture”, Global Telecommunication Conference, IEEE GLOBECOM, November 2008

[18] 3GPP, “Quality of Service (QoS) concept and architecture”, Release 9, TS 23.107 V9.0.0, December 2009. http://www.3gpp.org/ftp/Specs/html-info/23107.htm

[19] C. Filsfils and J. Evans, Deploying Diffserv in Backbone Networks for Tight SLA Control, IEEE Internet Computing, volume 9, issue 1, pp. 66-74. January 2005

[20] A. Torres, “Probabilidad, variables aleatorias, confiabilidad y procesos estocásticos en Ingeniería Eléctrica”, Universidad de los Andes, Facultad de Ingeniería, Departamento de Ingeniería Eléctrica y Electrónica. Bogotá 1993