Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. VII (2012), No. 1 (March), pp. 123-134

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

M. Navarro, Y. Donoso

Miguel Navarro
Universidad de los Andes
Bogotá, Colombia
E-mail: mignavarro@egresados.uniandes.edu.co

Yezid Donoso
Universidad de los Andes
Bogotá, Colombia
E-mail: ydonoso@uniandes.edu.co

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, the 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.
Keywords: Convergent services, IP Multimedia Subsystem, Quality of Service

1 Introduction

Networks for convergent services are the result of an evolution process followed by fixed and mobile networks. As a result of different trends of evolution, the IP Multimedia Subsystem (IMS) was introduced as the accepted network architecture for convergent services with guaranteed requirements, such as QoS, security, charging, and roaming. QoS provision 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]. 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 the Next Generation Network´s(NGN) evolution process in fixed networks. However, since IMS was already considering session control features, it 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 [2] [3]. Currently, with the specification of the fourth generation (4G) in mobile networks, 3GPP introduced the program named Evolved Packet System (EPS), which combines the Long Term Evolution (LTE) program,

and the Evolved Packet Core (EPC) program. IMS objectives continued having a leading role in EPS, since LTE is considered as a new access network that may be integrated to the network architecture, and the core in EPS integrates IMS architecture. At this point, networks working with these programs are referred as Next Generation Mobile Networks (NGMN) [4] [5].

In IMS, the problem regarding QoS provision at the IP Media Transport layer is the same as it is defined for the Internet. Several authors have already covered this problem and the models of Integrated Services (IntServ) and Differentiated Services (DiffServ) have 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. As a result, 3GPP defined DiffServ as the QoS model for the IP Media Transport layer [1] [6]. For upper layers, 3GPP has also specified the mechanisms for providing QoS. 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 in order to cover all possible scenarios and to guarantee QoS requirements. In [7], 3GPP standardized the QoS parameters applied in the service level, and also introduced the concepts of service priority and pre-emption capability and vulnerability, which support conflict handling between services in a state of network saturation. In spite of this concept and function definitions, their relation to the main functional entities in IMS layered architecture is still an ongoing process.

The main objective in 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.

The remainder of this paper is organized as follows. In Section 2, we describe some related work. In Section 3, we present the PCC architecture. In Section 4, we propose an enhanced QoS architecture and a heuristic algorithm to validate the architecture. In Section 5, we present the architecture and performance evaluation. The discussion of the results is presented in Section 6. Finally, Section 7 contains our conclusions and directions for further study.

2 Related Work

Related work about QoS in IMS has been presented prior the standardization of the PCC architecture in IMS Release 7. The main focus is on the heterogeneity introduced by different access networks and discrepancies between QoS classes in all of them. This problem is analyzed in [8], where authors present the work developed by 3GPP and ETSI TISPAN in QoS provisioning for IMS. With regard to the session control layer from the IMS architecture, authors emphasize on the importance of the Policy Decision Function (PDF) for 3GPP specifications, a function that is later performed by the PCC architecture. The transport layer is also considered, presenting the benefits and weaknesses' in DiffServ core networks. In the end, a practical implementation on a real network is stated and given for further study. After the PCC architecture specifications where given, several works 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. With this methodology, they

are allowed to combine service functions together in order to provide a specific functionality, in contrast to the traditional vertical service integration, which basically provides all the functionality with one service module. There are more studies concerning different problems in QoS on IMS, like [10], [11], and [12]; however, the problem introduced by dynamic QoS requirements, service level relocation, and their effect in the transport network, has not been considered.

3 Quality of Service in IMS

The IMS PCC architecture specified for Release 9 in [7] 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 [13], and the Common Open Policy (COPS) protocol defined in [14]. 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. The functions included in the PCC architecture to control the QoS are the following: resource allocation, event triggering, media flow establishment, and gating control.

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 [7].

QoS Class Identifier (QCI) The QCI is a scalar number associated to a network element and it is 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, packetdelaybudget, and packeterrorlossrate.

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".

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.

Maximum Bit Rate (MBR) The MBR parameter indicates the maximum bit rate authorized for a session.

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 [6]. 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.The relation between UMTS QoS classes and DiffServ parameters is presented in [2], based on the GSMA specification for the GPRS Roaming eXchange (GRX). This relation includes additional distinguishing factors in addition to delay sensitivity, such as jitter, packet loss, and Service Data Unit (SDU) error ratio.

4 Proposed Architecture

In the previous section we presented the QoS specifications in IMS on a service level and how they are associated toDiffServ 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 [15]. 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. At the same time that IMS introduces dynamic services, those services allow some flexibility in their QoSrequirements. 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 [7], 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 1. 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.

The *QoS Level Relocation Function* (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

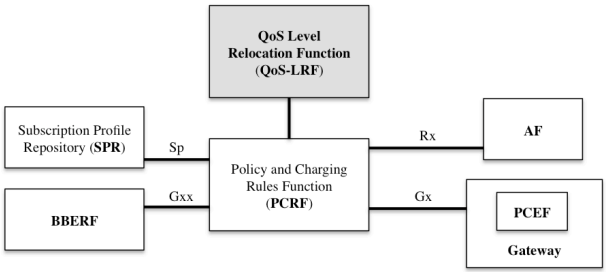


Figure 1: The enhanced PCC architecture

according to the standardized QCI characteristics [7] and UMTS QoS classes and their relation to DiffServ parameters [2]. First, we take the 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, 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 andthat 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, maintaining the relation between UMTS QoS classes and DiffServ parameters, as described in the previous section.

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.

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 Algorithm 4. 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

Algorithm 1 New sessions entering the network

```
A new session enters the network
if qos-level = EF then
  if availability in EF then
    resources are reserved in EF/ the new EF session is activated in EF
  else if PEC is activated and enough resources from EF users with lower priority and PEV activated then
    resources are released from the selected EF users/ EF sessions are relocated in AF (*)/ released resources
    in EF are reserved for the new EF session/ the new EF session is activated in EF
  else if SFB is activated then
    the PEC, PEV and SFB values from the new EF session are saved as historical values/ PEC = 1 /
    PEV = 0/ SFB = 0
    if availability in AF then
      resources are reserved in AF for the new EF session/ the new EF session is activated in AF
    else if PEC is activated and enough resources from AF users with lower priority and PEV activated
    then
      resources are released from the selected AF users/ AF sessions are relocated in BE (*)/ released resources
      in AF are reserved for the new EF session/ the new EF session is activated in AF
    else
      the new EF session entering the network is rejected
    end if
  else
    the new EF session entering the network is rejected
  end if
else if qos-level = AF then
  if availability in AF then
    resources are reserved in AF/ the new AF session is activated in AF
  else if availability in EF then
    the PEC, PEV and SFB values from the new EF session are saved as historical values/ PEC = 0/
    PEV = 1/ SFB = 1/ resources are reserved in EF/ the new AF session is activated in EF
  else if PEC is activated and enough resources from AF users with lower priority and PEV activated then
    resources are released from the selected AF users/ AF sessions are relocated in BE (*)/ released resources
    in AF are reserved for the new AF session/ the new AF session is activated in AF
  else if SFB is activated then
    the PEC, PEV and SFB values from the new EF session are saved as historical values/ PEC = 1/
    PEV = 0/ SFB = 0
    if availability in BE then
      resources are reserved in BE for the new AF session/ the new AF session is activated in BE
    else if PEC is activated and enough resources from BE users with lower priority and PEV activated
    then
      resources are released from the selected BE users/ the sessions from the selected BE users are rejected/
      released resources in BE are reserved for the new AF session/ the new AF session is activated in BE
    else
      the new AF session entering the network is rejected
    end if
  else
    the new AF session entering the network is rejected
  end if
else if qos-level = BE then
  if availability in BE then
    resources are reserved in BE/ the new BE session is activated in BE
  else if availability in AF then
    the PEC, PEV and SFB values from the new EF session are saved as historical values/ PEC = 0/
    PEV = 1/ SFB = 1/ resources are reserved in AF/ the new BE session is activated in AF
  else if the new BE session priority level is 8 (highest priority in the BE level) and enough resources from
  BE users with lower priority and PEV activated then
    resources are released from the selected BE users/ the sessions from the selected BE users are rejected/
    released resources in BE are reserved for the new BE session/ the new BE session is activated in BE
  else
    the new EF session entering the network is rejected
  end if
end if
```

initial QoS level with the historic PEC, PEV and SFB values.

Algorithm 2 EF and AF session relocation algorithms

EF session relocation
the PEC, PEV and SFB values from the EF session are saved as historical values
PEC = 1 / PEV = 0 / SFB = 0
if availability in AF **then**
  resources are reserved in AF for the EF session
  the EF session is activated in AF
else if PEC is activated **and** enough resources from AF users with lower priority and PEV activated **then**
  resources are released from the selected AF users
  AF sessions are relocated in BE (*)
  released resources in AF are reserved for the EF session / *the EF session is activated in AF*
else
  the EF session rejected
end if

AF session relocation
the PEC, PEV and SFB values from the EF session are saved as historical values
/ PEC = 1 / PEV = 0 / SFB = 0
if availability in BE **then**
  resources are reserved in BE for the AF session
  the AF session is activated in BE
else if PEC is activated **and** enough resources from AF users with lower priority and PEV activated **then**
  resources are released from the selected BE users
  the sessions from the selected BE users are rejected
  released resources in BE are reserved for the AF session / *the AF session is activated in BE*
else
  the AF session is rejected
end if

5 Architecture and 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.

5.1 Architectural models

The first architectural (M1) model is the reference point that gives standard values to compare results obtained with models 2 (M2) and 3 (M3). This model implements neither the session relocation feature, nor the SFB functionality, and 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. This reference to current networks is based on the analysis presented on [8] regarding DiffServ networks and QoS resource management. The second architectural model, M2, implements 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, M3, 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).

(a) Deterministic Parameters

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

(b) 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

(c) 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	$0, 1$

Table 1: Simulation Parameters

$$\delta = \frac{100\,(V_{exp} - V_{std})}{V_{std}} \tag{1}$$

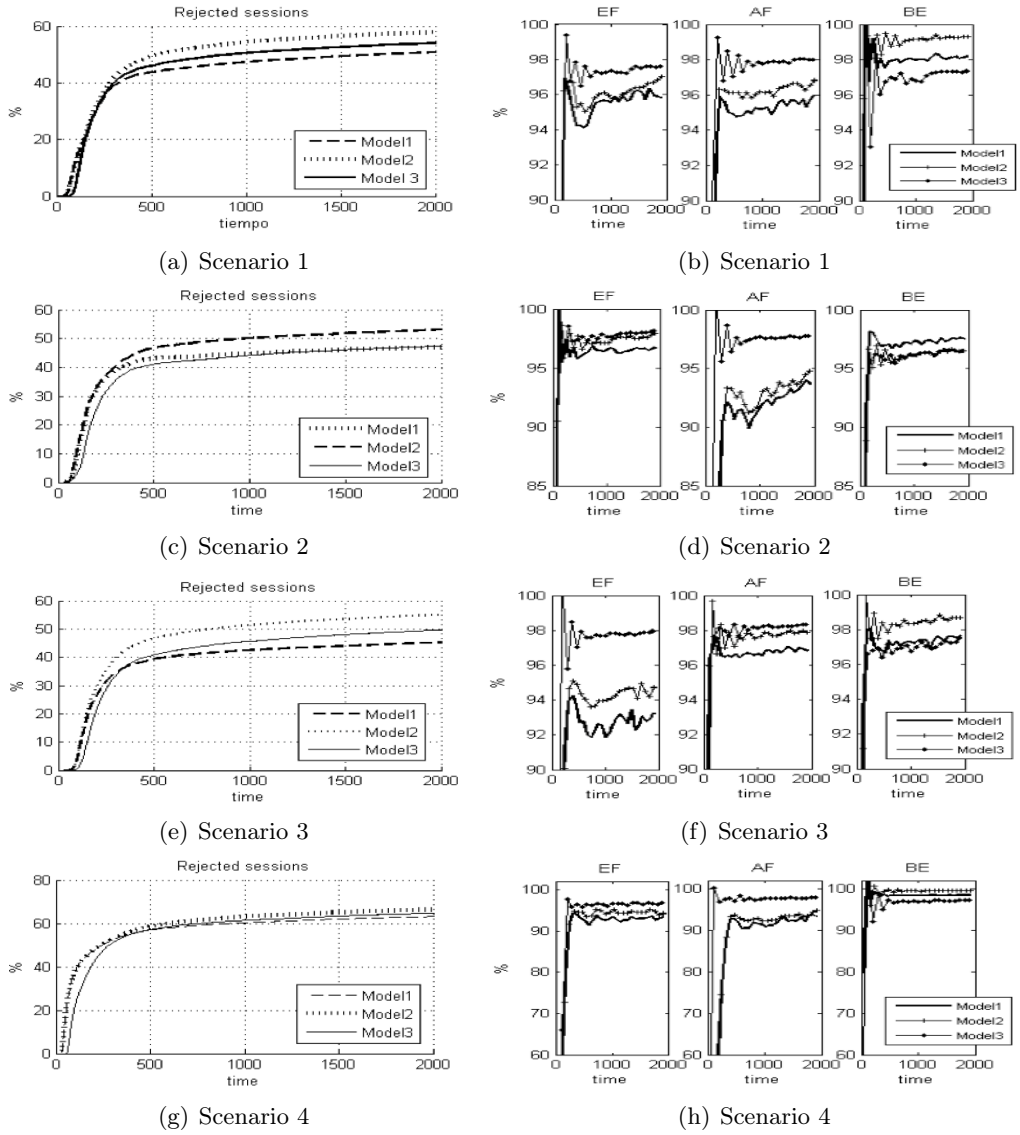
We simulate implementations of the three architectural models and the network behavior. Arrival of network users is simulated as a Poisson Stochastic Process with a rate parameter λ, using Monte Carlo simulations, and we define the simulation deterministic parameters as shown in Table 1(a). Table 1(b) 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 1(c) shows the ranges and distributions for the random parameters.

5.2 Simulation scenarios

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 0(b), 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. Figure 2 presents a basic description of each simulation scenario, and the results obtained for the number of rejected users and the percentage of usage for each QoS level.

6 Discussion

Previous simulations give information about the number of rejected and active sessions 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 inthe 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.



(a) Scenario 1 (b) Scenario 1

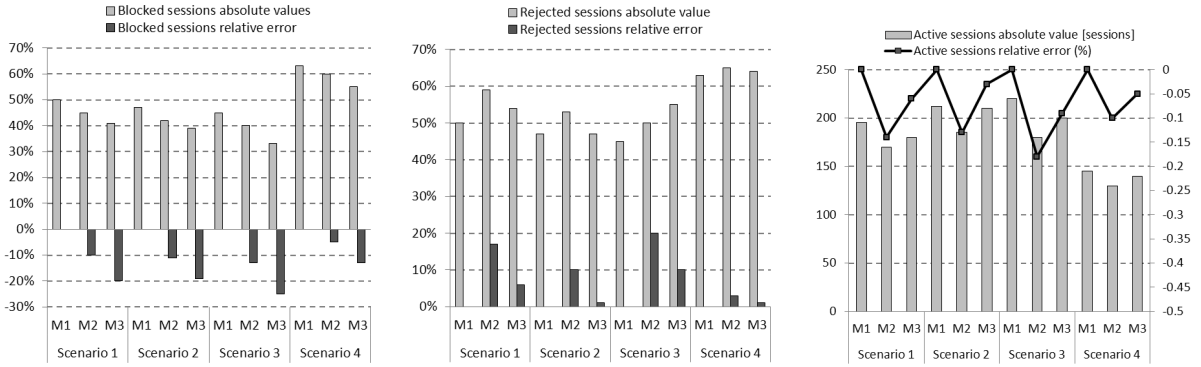
(c) Scenario 2 (d) Scenario 2

(e) Scenario 3 (f) Scenario 3

(g) Scenario 4 (h) Scenario 4

Figure 2: Simulation scenarios. Scenario 1: Basic scenario, sessions generated with the same probability. Scenario 2: Sessions generated with 60% probability for EF, 20% for AF, 20% for BE. Scenario 3: Sessions generated with 60% probability for AF, 20% for EF, 20% for BE. Scenario 4: Sessions generated with 60% probability for BE, 20% for EF, 20% for AF. (a), (c) (e) and (g) show the accumulated percentage of rejected sessions, and (b),(d), (f) and (h) show the Percentage of usage of each QoS-level in the network.

The first graphics presented, for each scenario, depict the behavior of rejected sessions. 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 can see that M3 reduces the number of blocked sessions in all scenarios compared to M2 and M1. This results presented in Figure 3(a), indicates a benefit from implementing M3 over



(a) (b) (c)

Figure 3: Simulation results. Part (a) shows the blocked sessions relative error comparison and, it shows the blocked sessions absolute values comparison at *t=2000sec*. Part (b) shows the rejected sessions relative error comparison, and the rejected sessions absolute values comparison at *t=2000sec*. Part (c) shows the active sessions relative errors to Model 1 at *t=2000sec*, and the number of active sessions at *t=2000sec*.

M1 and M2. Despite having this result for the number of blocked sessions, it is very important for the algorithm not to increment abruptly the canceled session percentage because M1 does not cancel any session. According the definition of M1, once a session reserves resources they cannot be released until the user finishes the session. The small effect of canceled sessions may be confirmed with the percentage of rejected sessions, as depicted in Figure 3(b). Looking at the results in all scenarios, the number of canceled sessions in M3 may be considered to have an effect that we may consider as small if we take into account that M2 always ends having a mayor percentage of canceled sessions than M3. The small effect of canceled sessions may be confirmed with the percentage of rejected sessions form Figure 2. In scenarios 2 and 4 there is no significant difference for the percentage of rejected sessions between M1 and M3; however, 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.

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 this work comes from the number of active sessions and how they are distributed among the QoS levels. 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 in Figure 3(c) 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 M2, validating the importance of the SFB implementation in M3.

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 session's 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. Analyzing the complexity of the algorithms in a worst case basis, it is evident that for every incoming session, the running time will be $O(n)$, where n is the number of active sessions in the network. Considering concurrent sessions entering the network, the running time will be $O(mn)$, where m is the number of sessions entering the network.

7 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. 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 validates 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 contributions 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 worst case running time of the algorithms is $O(n)$, where n is the number of active sessions in the network, and if the possibility of having m concurrent sessions entering the network is considered, the worst case running time is $O(mn)$. 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.

Bibliography

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

[3] 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.

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

[5] 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.

[6] 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

[7] 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

[8] S. Tompros, and S. Denazis. "Interworking of heterogeneous access networks and QoS provisioning via IP multimedia core networks", Computer Networks, Volume 52, Issue 1, pp. 215-227, January 2008.

[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] 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.

[11] 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.

[12] 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.

[13] 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

[14] 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

[15] 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