3GPP TR 23.818 V0.10~~9~~.0 (2007-02~~1~~)

Technical Report

3rd Generation Partnership Project;
Technical Specification Group Services and Architecture;
Optimisations and Enhancements for
Realtime IMS communication
(Release 7)

Keywords

3GPP, Architecture, real-time

3GPP

Postal address

3GPP support office address

650 Route des Lucioles - Sophia Antipolis
Valbonne - FRANCE
Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Internet

http://www.3gpp.org

Contents

Foreword ... 9
Introduction ... 9
1 Scope ... 10
2 References ... 10
3 Definitions, symbols and abbreviations ... 10
3.1 Definitions ... 10
3.2 Symbols ... 11
3.3 Abbreviations ... 11
4 Analysis of IMS session establishment procedures ... 11
4.1 Problem Description ... 11
4.1.1 General ... 11
4.1.2 Current IMS session setup procedure when real time media is required ... 11
4.2 Solution analysis ... 15
4.2.1 Session Establishment in GPRS IP-CAN ... 15
4.2.1.1 UE initiated media bearer establishment at SDP Answer ... 15
4.2.1.2 Network requested media bearer establishment at SDP Answer ... 16
4.2.1.3 Network requested media bearer establishment at SIP INVITE request ... 17
4.2.1.4 UE initiated media bearer establishment at SIP INVITE request ... 17
4.2.2 Session Establishment in IMS ... 18
4.2.2.1 Session establishment with resources indicated as available at initial INVITE ... 18
4.3 Conclusion ... 19
5 Analysis of Operator Controlled QoS ... 19
5.1 Problem Description ... 19
5.2 Solution analysis ... 20
5.2.1 Solution option 1 ... 20
5.3 Conclusion ... 20
6 Analysis of impact of non call related IMS signalling ... 20
6.1 Problem Description ... 20
6.1.1 General ... 20
6.1.2 Technical overview of the presence service ... 21
6.2 Solution analysis ... 22
6.2.1 Limiting traffic load ... 22
6.2.2 Reducing message size ... 22
6.2.3 Supporting different prioritisation of the non-call related signalling through the IP-CAN ... 23
6.3 Conclusion ... 23
7 Analysis into mechanisms to inform of loss of signalling bearer transport through the IP-CAN ... 23
7.1 Problem Description ... 24
7.2 Solution analysis ... 24
7.2.1 Behaviour of PCC architecture upon being informed of loss of the ability to communicate with the UE ... 24
7.2.1.1 AF (e.g. P-CSCF) requests establishment of an AF Session for IMS Signalling ... 24
7.2.1.2 AF (e.g. P-CSCF) is notified of IMS signalling bearer events ... 25
7.2.1.3 AF (e.g. P-CSCF) terminates the subscription to IMS signalling bearer events ... 25
7.2.2 Behaviour of P-CSCF upon being informed of loss of the ability to communicate with the UE ... 26
7.3 Conclusion ... 26
7.3.1 AF (e.g. P-CSCF) subscription to IMS signalling bearer events ... 26
7.3.2 Behaviour of P-CSCF upon being informed of loss of the ability to communicate with the UE ... 26
8 Analysis and identification of dynamic allocation of users to application servers ... 26
8.1 Problem Description ... 26
8.2 Solution analysis ... 28
8.2.1 General ... 28

8.3 Conclusion 28
9 Identification of stage 2 impacts for multimedia telephony 28
9.1 Introduction of the Telephony Application Server (TAS) 28
9.1.1 General 28
9.1.2 Standards Impacts 28
9.1.3 Conclusion 29
9.2 Identification of multimedia telephony 29
9.3 Recommended session establishment flows for multimedia telephony 29
10 Analysis of efficient interworking with other VoIP networks 29
10.1 Problem Description 29
10.2 Solution analysis 29
10.3 Conclusion 29
11 Analysis of general domain selection function 30
11.1 General principles 30
11.2 Problem description 30
11.3 Solution analysis 30
11.3.1 SDS Requirements 30
11.3.2 ADS Requirements 31
11.3.3 Relationship between SDS and ADS 31
11.4 Conclusion 32
12 Personal Network Management 33
12.1 General 33
12.3 Procedures on Interfaces 34
12.3.1 HSS – PNM AS/gsmSCF(CAMEL service for PNM) 34
12.3.2 S-CSCF – PNM AS 34
12.3.4 HSS – S-CSCF 35
12.3.5 HSS – GMSC 35
12.4 Interaction 35
12.5 Conclusion 36
13 Continuity of IMS-based Services 37
13.1 General 37
13.2 PS-PS Session Continuity 37
13.2.1 General 37
13.2.2 Potential Solution for PS-PS Session Continuity 38
13.2.2.1 PS-PS Session Continuity signalling flow 39
13.3 Conclusion 40
Annex A: Analysis of operator controlled QoS impact on GPRS 41
A.1 Solution analysis of impacts of mechanisms for operator controlled QoS in a GPRS IP-CAN 41
A.2 Network Requested Secondary PDP Context Activation procedure 41
A.3 Indication of media flows using a TFT 42
A.4 Indication of the support of NRSPCA in UE and the IP-CAN 43
A.5 Conclusion 44
Annex B: Analysis of impact of presence 45
B.1 Estimating presence traffic volumes 45
B.1.1 A presence traffic model 45
B.2 Impact on application layer and UTRAN 47
B.2.1 Impact on the application layer 47
B.2.1.1 Presence interferes the call related signalling 47
B.2.1.2 Presence influence SigComp compression ratios 47
B.2.2 Impact on UTRAN 48

Annex C: Solutions for the dynamic allocation of users to application servers 50
C.1 General 50
C.2 Overview of potential solutions 50
C.3 Flexible application server selection – HSS storage of selected application server 50
C.3.1 Solution Description 50
C.3.1.1 SIP initiated SIP-AS allocation 50
C.3.1.2 Ut interface based SIP-AS allocation 52
C.3.1.3 De-allocation of user from a SIP-AS 54
C.3.2 Solution Analysis 55
C.4 Hierarchical application server – Application server storage of selected application server 56
C.4.1 Solution Description 56
C.4.2 Solution Analysis 57
C.5 Dynamic assignment of application server by S-CSCF caching 57
C.5.1 Solution Description 57
C.5.2 Conclusion 59
Annex D: Network-initiated bearer control and PCC, high level description 60
D.1 Introduction 60
D.2 Solution overview 60
D.2.1 Abbreviations 60
D.2.2 Functional model 60
D.2.3 Additions to 3GPP protocols 62
D.3 Use cases 63
D.3.1 PDP context use cases 63
D.3.1.1 Network-Initiated Bearer setup 63
D.3.1.2 Network-initiated PDP context modification 65
D.3.2 End-to-end use case 66
D.3.2.1 Provisioning phase 66
D.3.2.2 IP-CAN session setup phase 67
D.3.2.3 Service setup phase, IMS call (Rx control) 67
D.3.2.3.1 Successful call setup: Normal case 67
D.3.2.3.2 Successful call setup: B-side SDP modification case 70
D.3.2.3.3 Unsuccessful call setup: No resources on A-side 71
D.3.2.3.4 Unsuccessful call setup: No resources on B-side 71
D.3.2.3.5 Call clearing and bearer termination 72
D.3.2.4 Service setup phase, RTSP streaming (Rx control) 72
Annex E: Change history 73
~~Foreword ..... 6~~
~~Introduction ..... 6~~
~~1 Scope ..... 7~~
~~2 References ..... 7~~
~~3 Definitions, symbols and abbreviations ..... 7~~
~~3.1 Definitions ..... 7~~
~~3.2 Symbols ..... 8~~
~~3.3 Abbreviations ..... 8~~
~~4 Analysis of IMS session establishment procedures ..... 8~~
~~4.1 Problem Description ..... 8~~
~~4.1.1 General ..... 8~~
~~4.1.2 Current IMS session setup procedure when real-time media is required ..... 8~~
~~4.2 Solution analysis ..... 12~~
~~4.2.1 Session Establishment in GPRS IP-CAN ..... 12~~
~~4.2.1.1 UE initiated media bearer establishment at SDP Answer ..... 12~~

Formatted: French (France)
Field Code Changed
Formatted: French (France)
Field Code Changed
Formatted: French (France)

~~4.2.1.2 Network requested media bearer establishment at SDP Answer ..... 13~~
~~4.2.1.3 Network requested media bearer establishment at SIP INVITE request ..... 14~~
~~4.2.1.4 UE initiated media bearer establishment at SIP INVITE request ..... 14~~
~~4.2.2 Session Establishment in IMS ..... 15~~
~~4.2.2.1 Session establishment with resources indicated as available at initial INVITE ..... 15~~
~~4.3 Conclusion ..... 16~~
~~5 Analysis of Operator Controlled QoS ..... 16~~
~~5.1 Problem Description ..... 16~~
~~5.2 Solution analysis ..... 17~~
~~5.2.1 Solution option 1 ..... 17~~
~~5.3 Conclusion ..... 17~~
~~6 Analysis of impact of non call related IMS signalling ..... 17~~
~~6.1 Problem Description ..... 17~~
~~6.1.1 General ..... 17~~
~~6.1.2 Technical overview of the presence service ..... 18~~
~~6.2 Solution analysis ..... 19~~
~~6.2.1 Limiting traffic load ..... 19~~
~~6.2.2 Reducing message size ..... 19~~
~~6.2.3 Supporting different prioritisation of the non-call related signalling through the IP-CAN ..... 20~~
~~6.3 Conclusion ..... 20~~
~~7 Analysis into mechanisms to inform of loss of signalling bearer transport through the IP-CAN ..... 20~~
~~7.1 Problem Description ..... 21~~
~~7.2 Solution analysis ..... 21~~
~~7.2.1 Behaviour of PCC architecture upon being informed of loss of the ability to communicate with the UE ..... 21~~
~~7.2.1.1 AF (e.g. P-CSCF) requests establishment of an AF Session for IMS Signalling ..... 21~~
~~7.2.1.2 AF (e.g. P-CSCF) is notified of IMS signalling bearer events ..... 22~~
~~7.2.1.3 AF (e.g. P-CSCF) terminates the subscription to IMS signalling bearer events ..... 22~~
~~7.2.2 Behaviour of P-CSCF upon being informed of loss of the ability to communicate with the UE ..... 23~~
~~7.3 Conclusion ..... 23~~
~~7.3.1 AF (e.g. P-CSCF) subscription to IMS signalling bearer events ..... 23~~
~~7.3.2 Behaviour of P-CSCF upon being informed of loss of the ability to communicate with the UE ..... 23~~
~~8 Analysis and identification of dynamic allocation of users to application servers ..... 23~~
~~8.1 Problem Description ..... 23~~
~~8.2 Solution analysis ..... 25~~
~~8.2.1 General ..... 25~~
~~8.3 Conclusion ..... 25~~
~~9 Identification of stage 2 impacts for multimedia telephony ..... 25~~
~~9.1 Introduction of the Telephony Application Server (TAS) ..... 25~~
~~9.1.1 General ..... 25~~
~~9.1.2 Standards Impacts ..... 25~~
~~9.1.3 Conclusion ..... 26~~
~~9.2 Identification of multimedia telephony ..... 26~~
~~9.3 Recommended session establishment flows for multimedia telephony ..... 26~~
~~10 Analysis of efficient interworking with other VoIP networks ..... 26~~
~~10.1 Problem Description ..... 26~~
~~10.2 Solution analysis ..... 26~~
~~10.3 Conclusion ..... 26~~
~~11 Analysis of general domain selection function ..... 27~~
~~11.1 General principles ..... 27~~
~~11.2 Problem description ..... 27~~
~~11.3 Solution analysis ..... 27~~
~~11.3.1 SDS Requirements ..... 27~~
~~11.3.2 ADS Requirements ..... 28~~
~~11.3.3 Relationship between SDS and ADS ..... 28~~
~~11.4 Conclusion ..... 29~~

Formatted: English (United States)

Formatted: English (United States)

~~12 Personal Network Management 30~~
~~12.1 General 30~~
~~12.2 Overall architecture 30~~
~~12.3 Procedures on Interfaces 31~~
~~12.3.1 HSS – PNM AS/gsmSCF(CAMEL service for PNM) 31~~
~~12.3.2 S-CSCF – PNM AS 31~~
~~12.3.4 HSS – S-CSCF 32~~
~~12.3.5 HSS – GMSC 32~~
~~12.4 Interaction 32~~

~~Annex A: Analysis of operator controlled QoS impact on GPRS 35~~
~~A.1 Solution analysis of impacts of mechanisms for operator controlled QoS in a GPRS IP-CAN 35~~
~~A.2 Network Requested Secondary PDP Context Activation procedure 35~~
~~A.3 Indication of media flows using a TFT 36~~
~~A.4 Indication of the support of NRSPCA in UE and the IP-CAN 37~~
~~A.5 Conclusion 38~~

~~Annex B: Analysis of impact of presence 39~~
~~B.1 Estimating presence traffic volumes 39~~
~~B.1.1 A presence traffic model 39~~
~~B.2 Impact on application layer and UTRAN 41~~
~~B.2.1 Impact on the application layer 41~~
~~B.2.1.1 Presence interferes the call related signalling 41~~
~~B.2.1.2 Presence influence SigComp compression ratios 41~~
~~B.2.2 Impact on UTRAN 42~~

~~Annex C: Solutions for the dynamic allocation of users to application servers 44~~
~~C.1 General 44~~
~~C.2 Overview of potential solutions 44~~
~~C.3 Flexible application server selection – HSS storage of selected application server 44~~
~~C.3.1 Solution Description 44~~
~~C.3.1.1 SIP-initiated SIP-AS allocation 44~~
~~C.3.1.2 Ut interface based SIP-AS allocation 46~~
~~C.3.1.3 De-allocation of user from a SIP-AS 48~~
~~C.3.2 Solution Analysis 49~~
~~C.4 Hierarchical application server – Application server storage of selected application server 50~~
~~C.4.1 Solution Description 50~~
~~C.4.2 Solution Analysis 51~~
~~C.5 Dynamic assignment of application server by S-CSCF caching 51~~
~~C.5.1 Solution Description 51~~
~~C.5.2 Conclusion 53~~

~~Annex D: Network-initiated bearer control and PCC, high level description 54~~
~~D.1 Introduction 54~~
~~D.2 Solution overview 54~~
~~D.2.1 Abbreviations 54~~
~~D.2.2 Functional model 54~~
~~D.2.3 Additions to 3GPP protocols 56~~
~~D.3 Use cases 57~~
~~D.3.1 PDP context use cases 57~~
~~D.3.1.1 Network Initiated Bearer setup 57~~
~~D.3.1.2 Network initiated PDP context modification 59~~
~~D.3.2 End-to-end use case 60~~

Formatted: English (United States)

Formatted: English (United States)

~~D.3.2.1 Provisioning phase 60~~
~~D.3.2.2 IP-CAN session setup phase 61~~
~~D.3.2.3 Service setup phase, IMS call (Rx control) 61~~
~~D.3.2.3.1 Successful call setup: Normal case 61~~
~~D.3.2.3.2 Successful call setup: B-side SDP modification case 64~~
~~D.3.2.3.3 Unsuccessful call setup: No resources on A-side 65~~
~~D.3.2.3.4 Unsuccessful call setup: No resources on B-side 65~~
~~D.3.2.3.5 Call clearing and bearer termination 66~~
~~D.3.2.4 Service setup phase, RTSP streaming (Rx control) 66~~

~~Annex E: Change history 67~~

Foreword

This Technical Report has been produced by the 3rd Generation Partnership Project (3GPP).

The contents of the present document are subject to continuing work within the TSG and may change following formal TSG approval. Should the TSG modify the contents of the present document, it will be re-released by the TSG with an identifying change of release date and an increase in version number as follows:

Version x.y.z

where:

x the first digit:

1 presented to TSG for information;

2 presented to TSG for approval;

3 or greater indicates TSG approved document under change control.

y the second digit is incremented for all changes of substance, i.e. technical enhancements, corrections, updates, etc.

z the third digit is incremented when editorial only changes have been incorporated in the document.

Introduction

During 3GPP Release 5 and Release 6, significant and considerable advancements have been made towards developing a reusable infrastructure for multimedia communication with the IP Multi-media core network Subsystem (IMS). The success of this has been demonstrated through the adoption of the IMS by other standardisation bodies (e.g. TISPAN, OMA, etc.), some of which have finalised a service definition utilising the IMS.

Study within the RAN groups is progressing the support of the transport of the voice media over HSDPA\EUL. While the efficient transport of voice over the air interface is a major catalyst for the development cellular IP multimedia telephony, some further system aspects (such as definition of supplementary services for multimedia telephony; consideration to call establishment time, and interference of an ongoing telephony call due to other services; handling of the loss of the signalling PDP context) are required in order to provide an efficient and inter-operable service.

This technical report captures the results of a study into potential system optimisations and enhancements required for mass market realtime communication.

When the feasibility of individual items have been assessed and concluded, it is expected that the specification work can proceed, if appropriate, without waiting for all of the objectives to be concluded.

1 Scope

The scope of the technical report is to capture the results of a study into the optimisations and enhancements of the system for mass market real-time communication (IMS multimedia telephony).

The objective is to provide a study into optimisations and enhancements for the support realtime services based on IMS with regards to the following aspects:

- Analysis of IMS session establishment procedures (e.g. signalling flows, bearer establishment) in order to reduce call establishment time for multimedia telephony to obtain the same, or at least similar, characteristics as exists for CS telephony;
- Analysis of impacts of any non call related IMS signalling (e.g. due to Presence) on the efficiency and service aspects of active real time communication sessions and the establishment of such sessions;
- Analysis into mechanisms to inform the IMS of loss of the signalling bearer transport through the IP-CAN;
- Analysis and identification of architecture and information flow impacts due to the dynamic allocation of users to Application Servers, including analysing any potential impacts at initial registration, session establishment and provision of user data in the HSS;
- Identification of any stage 2 impacts in order to support multimedia telephony services;
- Efficient interworking with other VoIP networks, e.g. regarding call establishment time and simplified call flows.

The study is intended to provide conclusions on the above aspects with respect to future normative specification work.

2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non-specific.
- For a specific reference, subsequent revisions do not apply.
- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document *in the same Release as the present document*.

[1] 3GPP TS 22.259: " Service requirements for Personal Network Management (PNM); Stage 1".

3 Definitions, symbols and abbreviations

3.1 Definitions

For the purposes of the present document, the following terms and definitions apply.

Definition format

<defined term>*: <definition>.*

example: text used to clarify abstract rules by applying them literally.

3.2 Symbols

For the purposes of the present document, the following symbols apply:

Symbol format

<symbol> <Explanation>

3.3 Abbreviations

For the purposes of the present document, the following abbreviations apply:

Abbreviation format

<ACRONYM> <Explanation>

4 Analysis of IMS session establishment procedures

Editors Note: This section covers the objective "Analysis of IMS session establishment procedures (e.g. signalling flows, bearer establishment) in order to reduce call establishment time for multimedia telephony to obtain the same, or similar, characteristics as exists for CS telephony".

4.1 Problem Description

4.1.1 General

When IMS is used for similar services as currently are provided via CS/PSTN, users do not want to experience an increased call establishment time. An analysis of the IMS session establishment procedures (e.g. signalling flows, bearer establishment) is required in order to identify possible enhancements to reduce call establishment time for multimedia telephony.

4.1.2 Current IMS session setup procedure when real time media is required

The following flows and principles can be used as a reference for further discussions on how to optimise and enhance the IMS session setup procedure when real time media is used.

The current IMS session setup procedure in TS 23.228 uses the following principles:

1. The resource reservation (if required) of the IP-CAN bearer appropriate for the real time media can be initiated when the UE consider it has enough information, i.e. at the sending of the SIP INVITE request or when the SDP answer is known

NOTE: It is up to the UE when to initiate the resource reservation, but the reservation may fail if done before receiving the SDP answer due to policies applied at the IP-CAN GW (e.g. due to SBLP).

2. The user is alerted, about an incoming multimedia session, when the resources for real time media are available

3. The authorization of QoS resources can be done on SDP offer and/or SDP answer on terminating side, and on SDP answer on originating side.

4. Both pre-conditions attributes and/or SDP direction attribute should be used to indicate when media can be sent. When resource reservation is required and the initial SDP Offer indicates that resources are not met, both pre-conditions and SDP direction attributes shall be set in the SDP.

5. Approval of QoS by the policy network is done when the SDP answer indicate that the media is active.

6. Media may be sent from a UE as soon as other UE has indicated that media can be received.

Editor's Note: This section is planned to contain the current End-to-End flows for an IMS session setup based on the recent Rel-6 session setup optimization activities by CT1.

The flow in figure 4.1 and figure 4.2 below is an example of a normal IMS session setup when real-time media is to be used and the appropriate resources for the real time media has to be reserved.

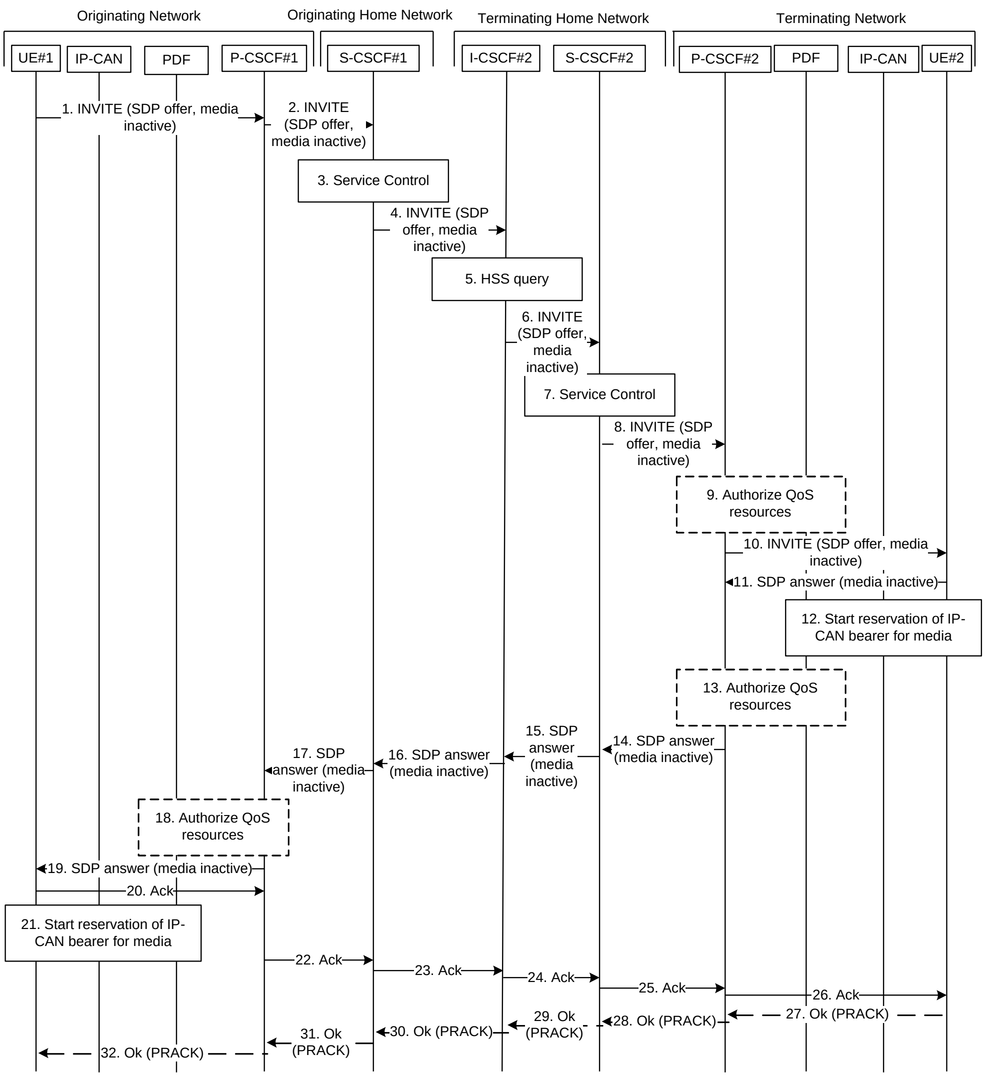

Figure 4.1: IMS session setup baseline

1. UE#1 sends the SIP INVITE request, containing an initial SDP, to the P CSCF#1 determined via the P CSCF discovery mechanism. The initial SDP may represent one or more media for a multi-media session, and one or more media may have a real time characteristic. Both both pre-conditions and SDP direction attributes indicates that the UE#1 cannot receive the real-time media at this point.

2. P CSCF#1 forwards the INVITE request to S CSCF#1 along the path determined upon UE#1's most recent registration procedure.

3. Based on operator policy S CSCF#1 validates the user's service profile and may invoke whatever service control logic is appropriate for this INVITE request. This may include routing the INVITE request to an Application Server, which processes the request further on.

4. S CSCF#1 forwards INVITE request to I CSCF#2.

5. I CSCF#2 performs Location Query procedure with the HSS to acquire the S CSCF address of the destination user (S CSCF#2).

6. I CSCF#2 forwards the INVITE request to S CSCF#2.

7. Based on operator policy S CSCF#2 validates the user's service profile and may invoke whatever service control logic is appropriate for this INVITE request. This may include routing the INVITE request to an Application Server, which processes the request further on.

8. S CSCF#2 forwards the INVITE request to P CSCF#2 along the path determined upon UE#2's most recent registration procedure.

9. Based on operator policy P CSCF#2 may initiate a procedure to authorize the resources necessary for this session.

10. P CSCF#2 forwards the INVITE request to UE#2.

11. UE#2 accepts the session with an SIP response that includes the SDP Answer. The SDP answer is sent to P CSCF#2.

12. UE#2 may reserve a dedicated IP-CAN bearer for media based on the media parameters UE#2 aims to include in the SDP answer. Note that the sequential ordering of 11 and 12. does not indicate that these steps are necessarily performed one after the other. The flow show that UE#2 send the SDP answer before IP-CAN resources for the media is available.

13. Based on operator policy P CSCF#2 may initiate a procedure to authorize the resources necessary for this session.

14. - 19. The SDP answer response traverses back to UE#1.

18. Based on operator policy P CSCF#1 may initiate a procedure to authorize the resources necessary for this session.

20. - 26. UE#1 acknowledges the SDP answer with an Ack (the Ack may be either ACK or PRACK depending on which SIP response the SDP answer was included in), which traverses back to UE#2.

21. UE#1 may reserve a dedicated IP-CAN bearer for media based on the media parameters received in the SDP answer. Note that the sequential ordering of 20 and 21 does not indicate that these steps are necessarily performed one after the other.

27. – 32. If the UE#1 acknowledge the SDP answer with a PRACK then UE#2 would acknowledge the PRACK with a 200 OK.

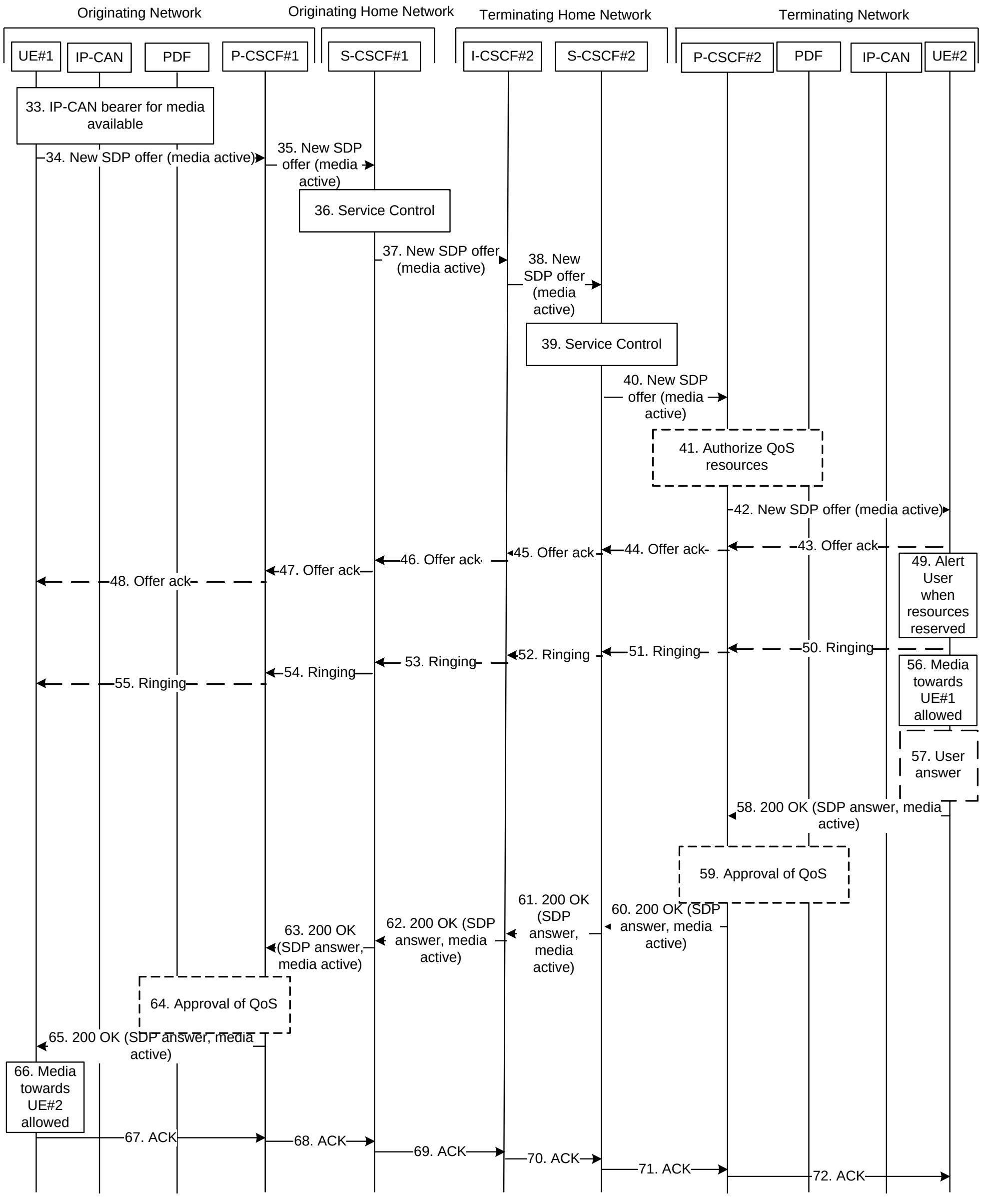

Figure 4.2: IMS session setup baseline

33. The IP-CAN bearer for the media becomes available.

34. – 42. UE#1 sends a new SDP Offer; containing an SDP with the media set to active (i.e. the IP-CAN resources for the media is available). The new SDP Offer traverses the set route to UE#2.

43. – 48. Depending on the SIP request used for the "new SDP Offer" and on the status of the IP-CAN resources the UE#2 acknowledge the "new SDP Offer".

49. The UE#2 may alert the user when the IP-CAN resource becomes available.

50. - 55.. UE#2 may optionally generate a ringing message towards UE#1.

56. UE#2 may at this point send media (e.g. ring tone) towards UE#1

57. The user answers the call

58-65. UE#2 accepts the session with a 200 OK. The 200 OK may (in case Offer ack in step 43 was not sent) include an SDP answer with the media set to active. Note: if the SDP Offer in step 42 was accepted with an SDP answer in step 43 and the resource reservation was not finalized, then the SDP is an SDP Offer in a SIP UPDATE request.

59 and 64. P CSCF#1 and P-CSCF#2 indicates that the resources reserved for this session should now be approved for use

66. UE#1 may at this point send media towards UE#2

67-72. UE#1 acknowledges the SDP answer (200 OK) with an ACK, which traverses back to UE#2

4.2 Solution analysis

4.2.1 Session Establishment in GPRS IP-CAN

4.2.1.1 UE initiated media bearer establishment at SDP Answer

In Figure 4.2a a high-level end-to-end call establishment flow is depicted. The bearers for the media streams are set up by the UEs through the Secondary PDP Context Activation procedure as defined in e.g. 3GPP TS 23.060, at the reception of the first SDP Answer.

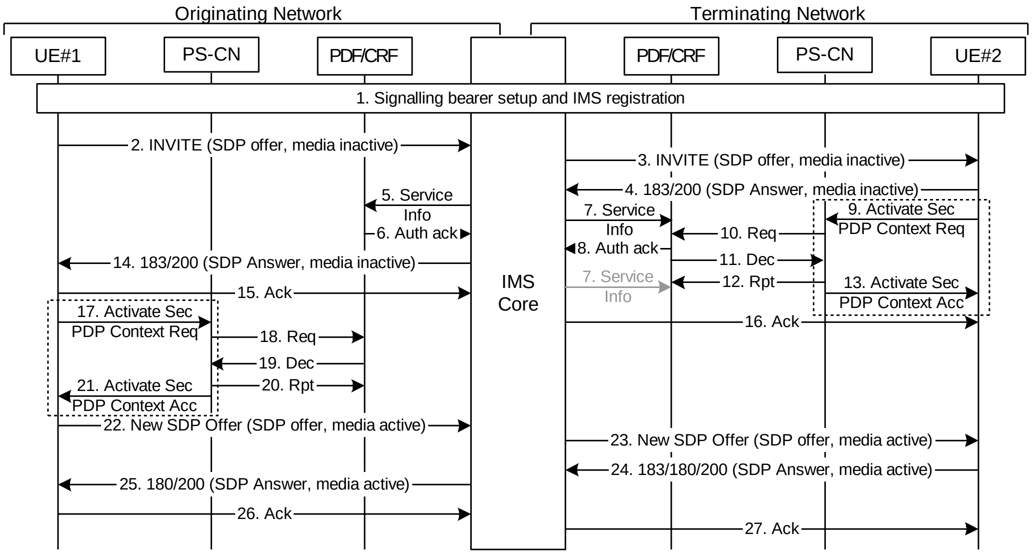

Figure 4.2a: End-to-end call establishment flow in a GPRS IP-CAN using UE initiated media bearer establishment

4.2.1.2 Network requested media bearer establishment at SDP Answer

In Figure 4.3 a high-level end-to-end call establishment flow with network requested media bearer establishment at SDP Answer, is depicted.

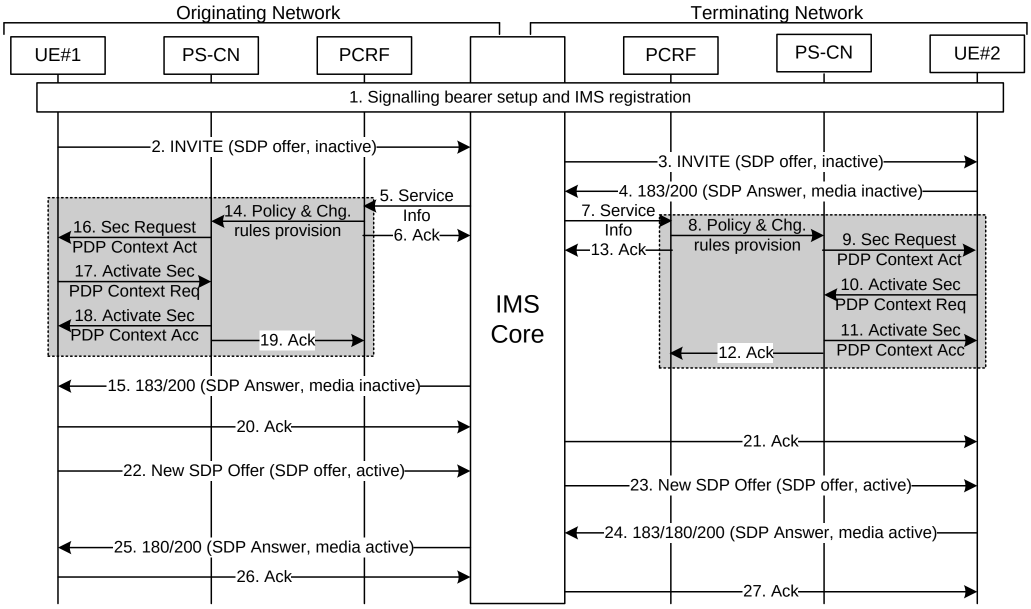

Figure 4.3: End-to-end call establishment flow in a GPRS IP-CAN using network requested media bearer establishment

With network requested bearer establishment, as shown in Figure 4.3 above, the network becomes responsible for triggering the media bearer establishment in the GPRS IP-CAN. On the originating side the media bearer establishment can be initiated from the PCRF as soon as Service Info is received and acknowledged by the PCRF (step 5 and 6). On the terminating side the PCRF can initiate the media bearer establishment as soon as it receives Service Info (step 7).

The network requested media bearer establishment procedure would be a new procedure for the GPRS IP-CAN. In GPRS it could be seen as an extension of the existing Secondary PDP context Activation procedure, i.e. a Network Requested Secondary PDP context Activation (NRSPCA) procedure. To ensure backward compatibility with entities not supporting the procedure, an indication whether the procedure is supported should be passed between the involved entities, i.e. UE, SGSN, GGSN and possibly PCRF. The indication should be added to the PDP Context Activation Procedure in the request and response, i.e. the UE would indicate the support and SGSN and GGSN may modify the indication when passing on the request to indicate whether the entity support the procedure. The indication also needs to be included in the RAU procedure when SGSN is changed.

The PCRF may make use of the information whether the procedure is currently supported, and the PCRF could in the response indicate whether the procedure should not be used by setting the indication to “procedure not supported”.

NOTE: If support is indicated the UE would assume that the network requested media bearer establishment procedure would be used for all IMS services, when applicable.

4.2.1.3 Network requested media bearer establishment at SIP INVITE request

The network requested media bearer may be established at the INVITE request. The flow could then look like in figure 4.4 below.

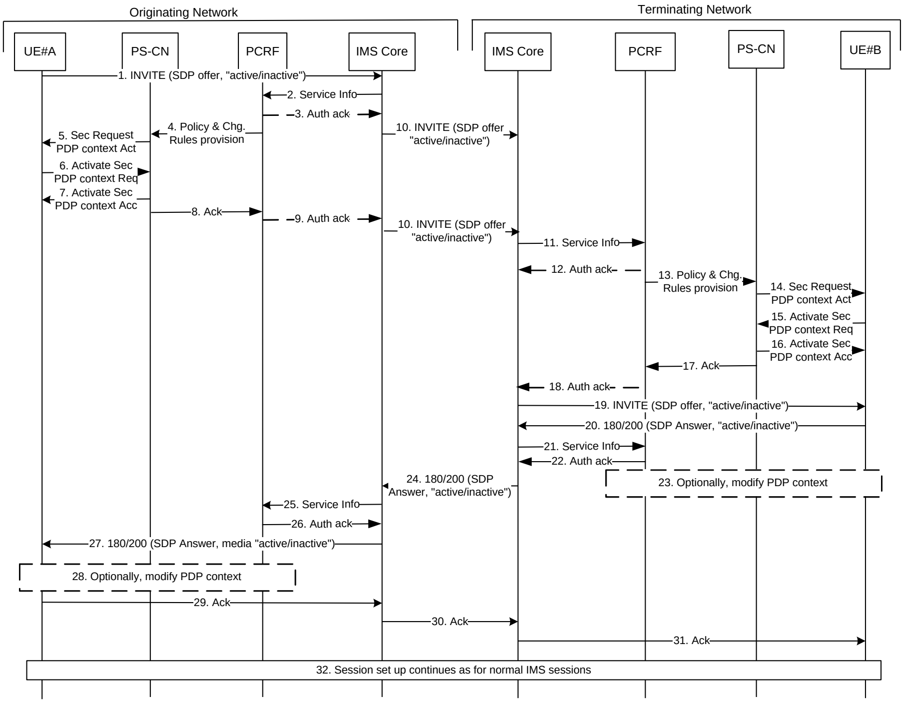

Figure 4.4: End-to-end call establishment flow in a GPRS IP-CAN using network requested media bearer establishment at initial SDP Offer

If the network requested bearer is established at the SIP INVITE Request (initial SDP Offer) instead of at the SDP Answer as shown in subclause 4.2.1.2, then the P-CSCF needs to interact with the PCRF at the reception of the SIP INVITE request. The PCRF could then either respond directly or wait until the appropriate resources are reserved at the IP-CAN. The P-CSCF forwards the SIP INVITE request when receiving the acknowledgement from the PCRF.

4.2.1.4 UE initiated media bearer establishment at SIP INVITE request

In Figure 4.4a a high-level end-to-end call establishment flow is depicted. The bearers for the media streams are set up by the UEs through the Secondary PDP Context Activation procedure as defined in e.g. 3GPP TS 23.060, at the sending or reception of the SIP INVITE request.

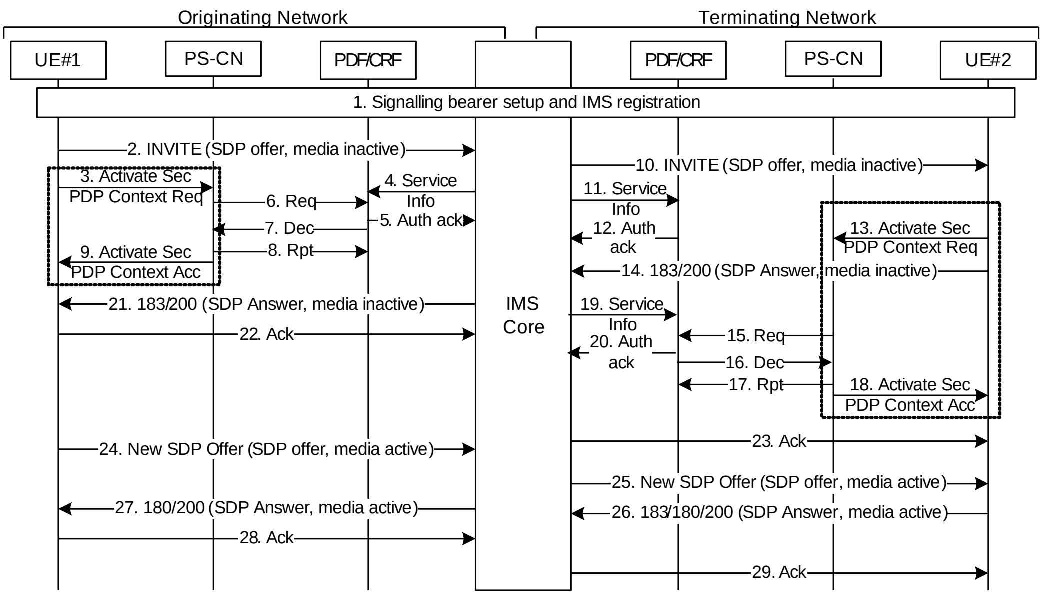

Figure 4.4a: End-to-end call establishment flow in a GPRS IP-CAN using UE initiated media bearer establishment at SIP INVITE

4.2.2 Session Establishment in IMS

4.2.2.1 Session establishment with resources indicated as available at initial INVITE

When the media is e.g. voice or video, a certain QoS from the IP-CAN is required. If the UE indicates that the resources are available from the initial INVITE, as indicated in the figure 4.5 below, the resources needs to be reserved in a way that avoids bad perception of the quality of the media.

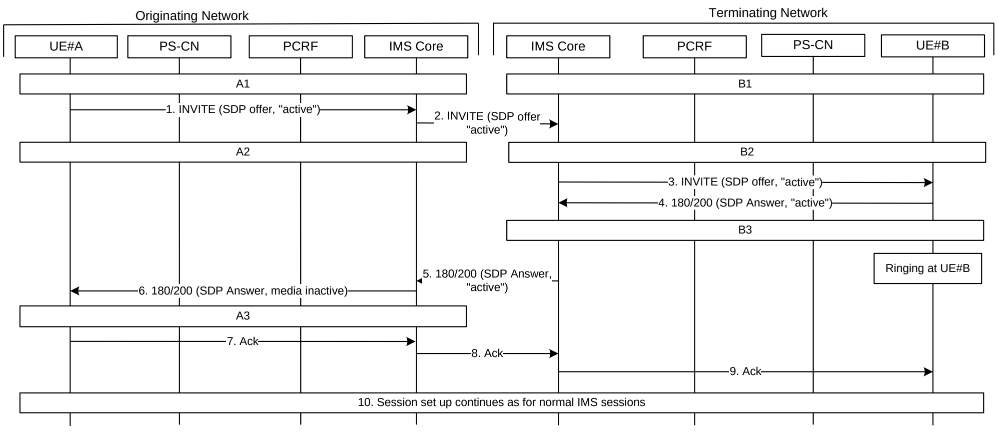

Figure 4.5: IMS Session setup indicating resources available in the initial INVITE/SDP Offer

The resources can be reserved at different times during the IMS session setup, as shown in the figure 4.5 above and described below.

A1 and B1: If the resources are reserved before sending the INVITE, then the actual media and codecs that eventually will be used is not known. The authorization of resources reserved at A1 and B1 would not be able to rely on any SIP/SDP information, i.e. either subscription information would have to be used or only best effort resources would have to be used. Also, there would be an additional delay if the resource reservation and sending of the INVITE is done sequentially. It is proposed to not further analyse this option.

Combined A1/A3 and B1/B3: Best effort resources could be "always" available. The initial media would then be transferred on the best effort bearer and when the SDP answer is received the appropriate resources could be reserved. This approach is used by PoC, see TR 23.979. There may be a risk of starting of with best effort resources for real-time media, as that may give a bad perception (e.g. media clipping) of the quality until appropriate IP-CAN resources have been reserved for the media or "ghost ringing" may occur if the appropriate resources are not eventually available at A3.

A2 and B2: If the resources are reserved when sending the INVITE, the actual media codecs may not be known unless a limited set of codec and codecs modes are used in the initial offer or the resources are reserved according to the most demanding codec property for each media component. It could be seen as a waste of resources if the resources are reserved at the INVITE, but that might not be a problem as the resource will be released if the call is not answered and e.g. if using shared resources that is probably even less of a problem.

Combined A2/A3 and B2/B3: This is the same as "A2 and B2", except that the resources used at the IP-CAN is modified according to the SDP answer, to ensure an efficient resource usage. That might be beneficial if dedicated resources used.

It should be allowed to reserve resources at e.g. A1 and B1, but for optimized resource utilization the following is recommended:

To apply appropriate QoS for potential early media, the resource reservation should be initiated at A2 and B2 (or B3), i.e. the options "A2 and B2" or "Combined A2/A3 and B2/B3" are the preferred options for reserving resources if the initial INVITE indicates that resources are available even though they are not reserved yet.

However, the UE should be aware whether the initial INVITE is allowed to indicate that resources are available even though they are not reserved yet.

Editor's Note: It is FFS which procedure to use in a GPRS IP-CAN, see subclause 4.2.1

4.3 Conclusion

5 Analysis of Operator Controlled QoS

Editors Note: This section covers "Enhancement of IMS real-time communication through Operator Controlled QoS".

5.1 Problem Description

The provisioning of QoS becomes increasingly important with larger volumes of traffic in the 3GPP PS domain, especially with the introduction of 3G High Speed access and the trend towards flat-rate charging for certain services such as Internet access. One drawback with the current 3GPP QoS architecture in that it doesn't put the operator in control of QoS to a degree that is desirable and possible. This may, for example, slow down the process for deployment of time critical IMS based applications due to lack of QoS control in a consistent manner, increase call setup times and in general give a less positive end-user experience of QoS dependent services.

5.2 Solution analysis

5.2.1 Solution option 1

The problems described above are tied to the exclusive lack of flexibility of the QoS negotiation procedure in 3GPP. They can be addressed by enhancing the system with control for the operator over the QoS negotiation procedures.

Basic principles for an enhanced QoS model:

a) QoS level to be used over the 3GPP access is based on what service is requested. The service is defined based on IMS signalling and SDP parameters (e.g. media components, protocol, Service ID (if present), etc)

b) The operator pre-configures the QoS level to use for different services

c) The network provides the UE with information for binding the uplink traffic to the correct bearer with the right QoS level.

5.3 Conclusion

It is proposed to start normative specification work for network initiated QoS for GPRS, i.e. Network-Requested Secondary PDP Context Activation (NRSPCA) and network-controlled uplink packet filters (UL TFT), using draft CR against 23.060(with the intent to cover the solution within one CR, if possible). This TR will document the overall high level description including the PCC aspects, dual mode aspects and application and services.

6 Analysis of impact of non call related IMS signalling

Editors Note: This section covers the objective ·Analysis of impacts of any non call related IMS signalling (e.g. due to Presence) on the efficiency and service aspects of active real time communication sessions and the establishment of such sessions;·

6.1 Problem Description

6.1.1 General

Signalling related to SIP session establishment and tear-down should be prioritized over media for instance, to enable successful connection of emergency sessions in a loaded cell. The means available to ensure that traffic over the signalling bearer is prioritized over other traffic when using 3GPP Release 99 QoS is the use of the signalling indication and the choice of traffic handling priority of the bearer.

Table 6.1.1 shows an example of how data may be prioritized due to the bearer traffic class, traffic handling priority and the use of the signalling indication.

Table 6.1.1: Example of traffic priority

Traffic Priority	Traffic class	Traffic handling priority and signalling indication
1	Interactive	1 and signalling indication set to 'yes'
2	Conversational	N/A
3	Streaming	N/A
4	Interactive	2 and signalling indication set to 'no'
5	Interactive	3 and signalling set to 'no'
6	Background	N/A

SIP is used for more than just multimedia telephony related signalling. In 3GPP, SIP is also used for non-multimedia related signalling such as transfer of presence updates and short text message. From a bearer perspective this type of traffic cannot be distinguished from multimedia telephony related signalling. Therefore applying a traffic priority according to Table 5.1.1 has the result that presence updates and short text messages will have higher priority than delay sensitive conversational media.

Especially the fact that presence is given the highest possible priority is unfortunate. Presence messages are in general generous in size and when the number of users on the buddy list and the number of possible presence states increase the number of presence messages transactions increases.

Assuming the table above and applying IMS multimedia telephony over access networks with limited peak throughput in combination with the presence enabler may then have the consequence that the potentially large presence update messages can cause audible distortions of the conversational voice stream when a UE receives presence updates during a multimedia telephony session.

Presence is also a service that may perform message exchanges in the background without user interaction. In crowded inner-city areas many presence enabled users (maybe many more users than the number of channels the cell can provide) may be located in the same cell. If the presence updates have the highest possible priority and there is many UE performing background presence message exchanges in the same area the result may be that admission control or lack of radio channels cause unnecessary blocking of income bringing sessions.

Intensive non-Multimedia session related signalling interleaved with multimedia session-related signalling that have the same priority may also result in increased SIP session set-up times if the non-multimedia session related signalling interferes with the multimedia session related signalling.

The probability that e.g. presence and short text messaging traffic interferes with the multimedia telephony session establishment signalling should be a function of how often the users are involved in a Multimedia Telephony call and the amount of traffic the presence and short text messaging networks creates. However, presence may be implemented in such a way that every time a user place a call or receive an invitation to a call, the presence client signals that the user is busy. Such implementation should always create traffic that interferes with the multimedia session-related signalling.

6.1.2 Technical overview of the presence service

Due to the possibility of "automatic" message exchange without user interaction, it is very likely that the SIP based presence service will create the major part of the non-call related IMS signalling (at least when the penetration of the presence service has become large). This sub-clause shortly explains how the SIP based presence service works and for further information about the intensity of which presence traffic may be sent is presented in Annex B.

Here follows a short explanation of Figure 6.1.

- A user (here referred to as the watcher) subscribes (sends a SIP Subscribe) to the presence server to subscribe to the presence service.
- The presence server notifies the watcher (gets SIP Notify) about the users on his/her buddy list (here referred to as presentities) published state. The SIP Notify messages may be sent as a response to the subscription to (i.e. the start of) the presence service or when a presentity change its presence state or when the watcher by user interaction updates want to update the buddy list.
- When a presentity change his/her state the presence client updates the state of the presence server by transmitting a SIP Publish message.

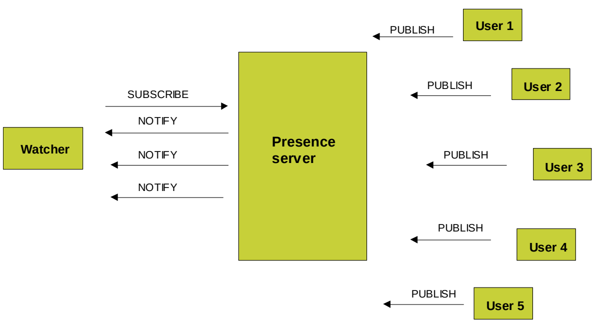

Figure 6.1: High-level view of the SIP based presence service

Presence message exchanges should in general create more downlink SIP Notify transactions then SIP Publish transactions:

- From a system perspective the reason is that one user may be on many buddy lists, thus every published presence state change (transmission of a SIP Publish in uplink) will create many SIP Notifications sent to the different watchers in their downlink direction.
- From a user perspective the reason is that a watcher may have many presentities on his/her buddy list, given that the presentities in average change presence state as often as the watcher many more SIP Notifys are received by then SIP Publish are sent from the users terminal.

The SIP Publish and the SIP Notify messages are all acknowledged by SIP 200 OK. This means that the number of SIP messages sent in the downlink and uplink are equal. But the size of the SIP messages differs. A typical SIP Publish or SIP Notify message size may range from 1500-4000 bytes depending on the content of the XML body (see [OMA-TS-Presence_SIMPLE-V1_0-20060214-C]) while the SIP 200OK may be in the region of 400-800 bytes. Thus, presence message exchanges should in general create more downlink traffic load than uplink traffic load.

6.2 Solution analysis

6.2.1 Limiting traffic load

The traffic load created by presence is a function of the number of presentities on the buddy-list, the number of presence states (see Annex B) and also the presence model used. The following presence models are commonly used:

- Push
  - Updates presence status at the watcher when state change (the presence model used in the traffic model above)
  - Good interactivity but creates potentially lot of traffic
- Throttling
  - The notification messages are grouped together at the server and cumulative notifications are periodically sent to the watchers. (If updates have occurred)
  - Less interactivity than Push, creates less traffic than Push.
- Pull
  - The watchers have to manually pull the server for presence updates
  - Lower interactivity than Push may create less traffic than Push (depends on user behaviour)

It is recommended that the presence client on a 3GPP terminal use the pull or throttling model when the presence client is not active on the watchers screen to limit traffic for users not actively monitoring their buddy lists.

It should be noted that the presence service is defined by the Open Mobile Alliance (OMA). Decisions on including procedures to limiting traffic load needs to be taken by OMA:

6.2.2 Reducing message size

The IETF has developed a SIP dictionary [RFC3485] to increase compression of "ordinary" SIP terms. A similar effort would to develop a presence dictionary for SigComp.

Having a presence dictionary for SigComp can provide means for:

- increase compression ratio - reduced traffic & delay;
- decreases the impact of presence on SigComp with dynamic compression.

If such dictionary is developed by IETF, 3GPP needs to support the presence dictionary by including support for it in [3GPP TS24.229].

6.2.3 Supporting different prioritisation of the non-call related signalling through the IP-CAN

In order to ensure that non-call related signalling (such as presence) does not interfere with call establishment, or the media for ongoing calls, to either the user receiving the non-call related signalling, and other users sharing the same transmission resources (e.g. the same call in a cellular network), a means is required to transport the non-call related signalling through the IP-CAN at a lower priority than the real-time media and call related signalling. This is to allow the IP-CAN to provide a different handling in congestion situations, while still allowing the non-call related signalling to get through when there is sufficient transport capacity to do so.

The different prioritisation could be achieved by placing the non-call related signalling a separate IP port to the rest of the SIP signalling; or through the use of the differentiated service code point (DSCP) IP header. (The DSCP head is used to provide differentiated treatment at the IP transport level).

Placing the non-call related signalling on a separate IP port would imply the following:

- The UE would have to register multiple contacts, one for call related signalling and another for non-call related signalling;
- The multiple contacts would imply multiplying the number of security associations between the UE and the P-CSCF; the current registration procedure would be impacted;
- The multiple contacts might imply multiplying the number of sigcomp flows;
- The multiple contacts would be propagated into the S-CSCF and would be in the scope of the normal forking behaviour; this would require additional means to suppress/change the normal forking logic;
- The multiple contacts related to one UE might require signalling means to associate them to avoid misconception during registration and for subscribers to registration event package;

Transport the non-call related signalling with a separate DSCP value would imply the following:

- The P-CSCF would require logic to differentiate call and non-call related SIP signalling in order to set respective DSCP values;
- The GGSN would map different DSCPs to different QoS PDP contexts;
- The UE would map the uplink non-call related traffic to the different PDP contexts.

The use of the DSCP approach limits the standardisation required, even though it does have some limitations. As the DSCP code points are sometimes mapped at network boundaries, the DSCP approach works best if the GGSN and the P-CSCF are in the same network (as described in standards today) or at least the SLA between the networks takes into account the DSCP values for the call signalling and the non-call related signalling.

6.3 Conclusion

Transport level solutions based upon finding means to prioritise the non-call related signalling differently through the IP-CAN (e.g. over the air interface) are considered to be a possible approach in order to solve this problem. However, SA2 has not been able to complete a workable solution following this approach before completion of Release 7.

It is proposed that the analysis and the definition of a solution for this study item are further deferred to Release 8.

7 Analysis into mechanisms to inform of loss of signalling bearer transport through the IP-CAN.

Editors Note: This section covers the analysis into mechanisms to inform the IMS of loss of the signalling bearer transport through the IP-CAN.

7.1 Problem Description

Knowledge of the "Loss of signalling bearer transport" through the IP-CAN are essential both when the signalling bearer is used to convey signalling for an established session as well as when there is not a session established yet.

If an initial request is sent from the IMS to the terminating user and there is a failure of the bearer that transports the signalling, it takes 64xT1 timer before the IMS stops repeating the request if the calling UE does not clear the session. This will lead to unnecessary tying up resources and with long waiting times for the calling user. For the cases where there are services such as "communication diversion: communication forwarding on mobile subscriber not reachable" operating for the user, then the not reachable timer will place a maximum "waiting time" for calling user. The determination of "not reachable" for such cases will always be based on time values, consuming unnecessary resources while the network keeps repeating the request to the UE.

7.2 Solution analysis

7.2.1 Behaviour of PCC architecture upon being informed of loss of the ability to communicate with the UE

The basic concept of the proposed solution relies on the PCC architecture being defined for Release 7. The PCC infrastructure is able to enforce at the bearer plane (e.g. GGSN) a specific QoS based on service parameters negotiated at the SIP signalling plane (e.g. P-CSCF).

PCC infrastructure is also capable of reporting bearer level events to the application level. The subscription/notification framework being implemented for PCC can make it possible for a P-CSCF to know that a particular media component is not being delivered due to e.g. a loss of the corresponding dedicated PDP context. However, notifications of bearer level events not related to actual media bearers (i.e. IMS (SIP) signaling bearers) are not currently considered.

Therefore, alignment of P-CSCF and PCC procedures would be required in order to be capable of providing notifications of the loss of the ability to transport the IMS signalling. This would require the basic following additions to current P-CSCF and PCC procedures.

- P-CSCF is able to request the establishment of an AF Session specific for IMS signalling, in the absence of session information (e.g. SDP). This would allow the AF to request PCC control procedures (e.g. subscription to notification of bearer level events) for IMS signalling.
- Associated new processing rules at both AF (e.g. P-CSCF) and PCRF to manage this AF Session specific for IMS signalling, including processing rules for the establishment, notification of events and termination.

7.2.1.1 AF (e.g. P-CSCF) requests establishment of an AF Session for IMS Signalling

In order to be able to be notified of the loss of the ability to transport IMS signalling, the AF (e.g. P-CSCF) shall be able to request the initiation of an AF session specific for IMS signalling and in the absence of session information (e.g. at the reception of an initial SIP REGISTER). This would allow the AF (e.g. P-CSCF) to subscribe to bearer level events of the associated IMS signalling.

The establishment process of an AF session specific for IMS signalling should be similar to the establishment of a traditional AF session (i.e. related to media IP flows) but requires specific new processing rules at the different entities involved. Figure 7.2.1 shows the message flow for P-CSCF establishment of an AF Session specific for IMS signalling during an initial SIP REGISTER.

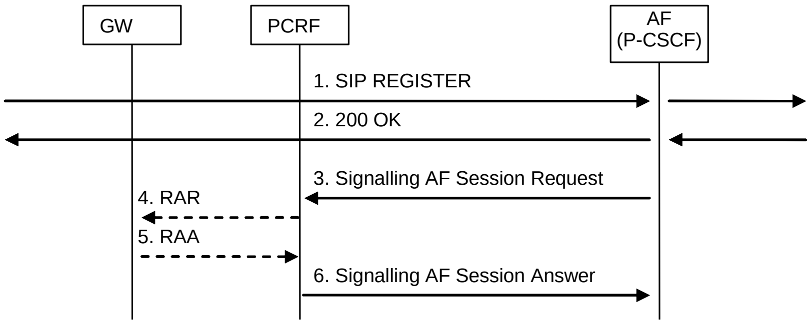

Figure 7.2.1: P-CSCF requests establishment of an AF Session for (SIP)IMS Signalling

1. The user initiates a SIP Registration procedure.
2. The SIP Registration procedure is completed successfully (user has been authenticated and registered within the IMS Core NW).
3. The AF (e.g. P-CSCF in this case) requests the establishment of an AF session related to the IMS signalling.

Editor's note: It is FFS whether dynamic filter, QoS and/or charging characteristics shall be enforced for the IMS signalling. Depending on the above requirements, the request of an AF Session specific for IMS signalling may not have to include any Media-Component-Description AVP, neither the AF-Charging-Identifier AVP.

Apart from the UE IP Address and other basic Diameter AVPs (e.g. Session-Id), the request of an AF Session specific for IMS signalling shall basically include instead a new SIP-Signalling-Indicator AVP and the Specific-Action AVP (requesting subscription to INDICATION_OF_TERMINATION_OF_BEARER). This shall be understood by the PCRF as an indication that the AF (e.g. P-CSCF) is willing to subscribe to IMS signalling bearer events.

4. If the PCRF has not previously subscribed to the required bearer level events from the IP-CAN, then the PCRF shall do so. One possible solution is PCRF activation of a dynamic PCC rule specific for SIP signalling.
5. PCEF (e.g. GGSN) confirms the subscription to bearer level events.
6. PCRF confirms the establishment the AF Session specific for IMS signalling.

Editors Note: It is for further study whether, during this process, the AF and/or PCRF can request the network initiated establishment of IP-CAN resources for the transport of the IMS signalling.

7.2.1.2 AF (e.g. P-CSCF) is notified of IMS signalling bearer events

If by any chance the PDP context utilized for IMS signalling (for GPRS, Bearer-Usage equal to "GENERAL" or "IMS SIGNALLING") is lost and communication for IMS signalling is not possible to the UE, the PCEF will notify this event, impacted PCC rule(s), to the PCRF, which in turn will be able to provide an indication to the AF (e.g. P-CSCF).

7.2.1.3 AF (e.g. P-CSCF) terminates the subscription to IMS signalling bearer events

This section takes care of the procedure to terminate the AF Session specific for IMS signalling, in normal conditions (e.g. the user is de-registered from the IMS Core NW).

Upon the reception of a SIP REGISTER message indicating that the user shall be de-registered from IMS, the P-CSCF shall then initiate the request for termination of the corresponding AF Session specific for IMS signalling by triggering a Session Termination Request Diameter message over Rx interface.

7.2.2 Behaviour of P-CSCF upon being informed of loss of the ability to communicate with the UE

Upon being informed of the loss of the ability to contact the UE for IMS signalling, the P-CSCF could take the following actions:

1. Do nothing
2. If there is an ongoing call, clear the call (which would likely happen anyway as in such a case the medias for the session are likely to have been terminated as well).
3. Clear any ongoing calls and reject any terminating call attempts towards that UE, and continue to do so until the UE re-registers.
4. Request the S-CSCF to de-register the contact information for the terminal that can no longer be contacted.

Consideration for the need of a timer in the P-CSCF is required.

Rejecting the ongoing call (in order to inform the peer user that communication is no longer ongoing) and to immediately reject new attempts to contact the terminal increase the user experience for the calling user as it allows any subsequent behaviour (e.g. supplementary service invocation) to occur immediately. The registration state will be cleaned up after the re-registration period, or when the UE performs a new initiate registration.

Little extra seems is gained by having the P-CSCF request the S-CSCF to de-register that contact.

7.3 Conclusion

7.3.1 AF (e.g. P-CSCF) subscription to IMS signalling bearer events

Within the Release 7 architecture, it shall be possible that the AF (e.g. P-CSCF) is able to receive notifications of events related to the IMS signalling bearer. This requires that the AF (e.g. P-CSCF) is able to request the initiation of an AF session even in the absence of service information (e.g. at the reception of an initial SIP REGISTER) in order to subscribe to such events.

It is proposed that Release 7 PCC specifications define the required functionality taking the solution analysis in this TR as the base for the necessary specification work within the Release 7 PCC work item.

7.3.2 Behaviour of P-CSCF upon being informed of loss of the ability to communicate with the UE

Upon being informed that signalling transport to the UE is no longer possible, the P-CSCF shall clear ongoing calls (if any) to that UE and immediately reject incoming calls to the UE indicating that the user is not reachable.

When the UE detects that has no means to communicate with the IMS network, it should re-establish a new bearer for IMS signalling and perform an IMS registration.

This capability will be included in TS 23.228.

8 Analysis and identification of dynamic allocation of users to application servers

8.1 Problem Description

One aim of the IMS is to be able to reduce the operational cost of a network. The complexity of operating a network increases with the number of supported subscribers, and one contributor will be the management of allocating subscribers to application servers for the same set of services, where there is a requirement for a user to be assigned to an application server longer than the duration of one session. This would occur when there is data which is to be retained together with the processing resources longer than a single session (i.e. sticky data). This will become more complex as both the number of application servers increase for a single IMS communication service (due to the need to support an increasing number of subscribers), as well as handling the application servers required for different IMS communication services; in particular if the application servers come from different vendors, supporting differing characteristics.

To illustrate such complexity; consider a network that contains application servers for the support of PoC and Telephony (i.e. PoC-ASes and TASes). If the network is initially configured such that there is equal number of PoC-ASes and TASes, but later the traffic pattern changes such that more TASs are required, then it will be required to re-allocate the TASs that the subscribers are on, but not the PoC-ASs. The re-allocation of the subscribers amongst the TASs could initially be simply the addition of new TASs to support the new subscribers, however it could also be the situation whereby the traffic model has changed such that the TASs become overloaded, requiring a percentage of the subscribers to be offloaded to other application servers. The traffic model and the characteristics for each service are likely to change independently, and not only depend on the addition of new subscribers.

In order for an S-CSCF to assign or re-assign an appropriate physical service-instance to a user it needs to take into account the Service Availability at an application server. Service Availability consists of the necessary data and logic that allows the S-CSCF to determine the availability for an application server to run the service. For example, the S-CSCF needs to have access to a real-time view of all the current states of all the application servers to identify if the application server has the required service-level capacity to take on an extra user. An application server may not be able to take on another service instance, because it is experiencing service congestion (e.g. required QoS not available to run the service), but the physical server is available (i.e. not experiencing network congestion). Additionally, logic needs to be put in place such that the users are intelligently load balanced among the available pool of Application Servers (for the specific set of services).

The application server name in an iFC (filter criteria) may represent a logical address or physical address. If the application server name represented a physical address, the reallocation of users to application servers will require a "per subscriber modification" - a modification of the iFCs (filter criteria) for all of subscribers with telephony. This effect is even more apparent if the application servers are from different vendors, where vendor specific can be applied amongst the application servers from a single vendors, and the application servers may also have different characteristics (e.g. subscribers/application server) that may make the planning more complicated. It will also be even more complicated when considering more services such that the traffic model and the characteristics for each service may vary independently.

The method for directing the SIP traffic to a specific application server, for a specific user, is based upon the initial filter criteria (iFC). Take, for example, a network with 3 Telephony application servers (TAS), with logical names TAS1.operator.com; TAS2.operator.com and TAS3.operator.com. For such a network, subscribers would be allocated to the different TASes, requiring different iFCs for the different subscribers as the application server name is part of the iFC. These would have to be managed and updated as either the traffic characteristics changes or the characteristics of the application servers change to e.g. support more users per application server. This results in a higher than required OPEX.

In addition to the operational costs, using the iFCs to allocate the subscribers to the application servers has an impact on the network availability. To illustrate this, consider the above example: If TAS1.operator.com has an outage, then all of the subscribers with TAS1.operator.com in the iFC (which is this example is 1/3 of the subscribers) would not receive the telephony service. This results in a lower service performance than required.

It would be desirable to avoid requiring a per subscriber modification in the network when managing the changing characteristics of a network. In order to achieve this, <u>the iFCs</u> for all subscribers <u>with the same service set</u> should remain the same (the service set is realised with IFCs that point to the application servers providing the service in the service set), irrespective of the network characteristics. Such an approach would lead to a reduction in the operational costs, as well as improved in service performance.

The goal is to prevent the need for any changes of data in an S-CSCF (and HSS) when a new application server is introduced into the network. An S-CSCF needs to be informed of the specific capabilities of an Application Server (e.g. what services it provides, how many instances it can run, what services require a permanent assignment of a user to an application server, etc) so that it can be added to the "pool" of available Application Servers that can provide a particular set of services for the subscriber.

In Summary, the key problems are:

1. How is a server selected to support a service for a new subscriber taking into account data and intelligence to allow for load balancing and performance of the network when a pool of application servers are available for assignment.

2. How is a user dedicated to that application server for ongoing communication?
3. How is a user re-assigned to an application server, what is the data that determines that a reassignment is required and where does this data come from?
4. How is the S-CSCF informed of the capabilities of a new application server that allows it to make decisions on re-assignment?

8.2 Solution analysis

8.2.1 General

This section describes procedures for the support of keeping the same IFC for the users with the same services irrespective of the network configuration and is based upon the following principles:

- A user could be served on a number of SIP-AS.
- When a user is not allocated to a SIP-AS, none of the SIP-ASes stored the data for the user (for that service).
- The solution allows that SIP-AS maybe allocated to the user when the network receives the first application facing triggerable event for that user. Such a request could be a SIP registration; a SIP terminating call; an operation over the Ut interface or an operation over other interfaces, the first originating INVITE for the user, etc.
- A SIP-AS can decide when to de-allocate the user from the SIP-AS. This is expected to be at, or sometime after, e.g. de-registering from the network.

A number of proposed solutions are captured in Annex C.

8.3 Conclusion

For the timeframe of a release 7 network, the hierarchical application server approach as described in clause C.4 can be used, as it does not require impacts to the stage 3 specification. The S-CSCF caching does not require stage 3 work either, but is limited to the ISC. Further re-evaluation of this is open in the timeframe of a subsequent 3GPP release.

9 Identification of stage 2 impacts for multimedia telephony

9.1 Introduction of the Telephony Application Server (TAS)

9.1.1 General

A number of the procedures involved in the provision of the multimedia telephony service require the support of a SIP-AS for telephony. Such a SIP-AS is referred to as a Telephony Application Server (TAS). For some services, the telephony application server interacts with a MRFC for e.g. the sending of tones or announcements in different media formats. Following the principles established with messaging and conferencing, the functional split between the MRFC and the TAS is out of scope of the present document. Procedures for the MRFC are described together with those for multimedia telephony.

9.1.2 Standards Impacts

The Telephony Application Server is a SIP-AS providing the network support for the multimedia telephony service. The functional split between the TAS and the MRFC is out of scope of this specification. Any procedures and flows requiring media interaction will show the TAS and the MRFC described together.

9.1.3 Conclusion

The identified impacts in section 9.1.2 are included in TS 23.228.

9.2 Identification of multimedia telephony

Multimedia telephony is an IMS communication service. In principle, there are two approaches that could be taken for the identification of multimedia telephony. One approach is to explicitly identify the SIP requests associated with multimedia telephony though the use of an IMS communication service identifier. A second approach would be to assume that the absence of any IMS communication service identifier is an indication of multimedia telephony.

Explicitly identifying multimedia telephony with an IMS communication service identifier is the recommended approach.

9.3 Recommended session establishment flows for multimedia telephony.

In order to support the mass market deployment of multimedia telephony, it is recommended that there is one recommended flow for each of the following scenarios:

- When the UE and the network support NRSPCA, and a bearer for the media is not established.
- When the UE and the network support NRSPCA, and a bearer for the media is established.
- When the UE and/or the network do not support NRSPCA, and a bearer for the media not is established.
- When the UE and/or the network do not support NRSPCA, and a bearer for the media is established.

When establishment of UE initiated IP-CAN bearer(s) for the media is required and the UE has been made aware of the operator MTSI policies with regards to allowed media for the subscriber, then the principle to reserve IP-CAN bearer(s) at the sending of the SIP INVITE request (see flow in subclause 4.2.1.4 for an example) is the recommended principle. If the policies are not made aware to the UE, then the principle to reserve IP-CAN bearer(s) at the reception of the SDP answer (see flow in subclause 4.1.2 for an example) is the recommended principle.

TS 23.228 should be updated with the principles above.

10 Analysis of efficient interworking with other VoIP networks

Editors Note: This section covers "Efficient interworking with other standardised SIP based VoIP networks, e.g regarding call establishment time and simplified call flows."

10.1 Problem Description

10.2 Solution analysis

10.3 Conclusion

11 Analysis of general domain selection function

11.1 General principles

The need for a general domain selection function has been identified. The goals of this study are the following:

- identifying mechanisms by which the HPLMN and VPLMN operator and the user can determine and influence whether the CS domain or the IMS is preferred for MO and MT voice calls;
- analysing how domain selection mechanisms being developed in the VCC Technical Specification (3GPP TS 23.206) and the CSI Interworking Technical Report (3GPP TR 23.819) can be harmonised into a general network domain selection mechanism for terminating calls – the study is not restricted however to the support of VCC-capable or CSI-capable terminals;
- facilitating the long-term migration of realtime services to IMS by the development of a general mechanism for selecting the domain for originating and terminating voice calls, needed during the transition period.

11.2 Problem description

Domain selection is a functionality required for subscribers that simultaneously and non-simultaneously attach/register in CS domain and IMS with the same MSISDN, or related public user identifiers. Domain Selection functionality is not required for pure CS-only and MMTel-only subscribers.

Domain selection is a functionality to determine which domain (CS or IMS) shall be used to establish a call/session, thereby it changes the normal call routing behaviours in both CS domain and IMS.

11.3 Solution analysis

Domain Selection can be divided into two types of functionality: Service Domain Selection (SDS) and Access Domain Selection (ADS).

Editor's Note: The aim of this separation is to allow easy description of the existing functionality required from the domain selection mechanisms defined in VCC and CSI-Interworking. Whether the resulting general mechanism for domain selection requires separated or distinct SDS and ADS is FFS.

11.3.1 SDS Requirements

SDS selects the service engine that shall be applied for a call. The SDS has a role for originating service domain selection and for terminating service domain selection.

Generic requirements:

- SDS shall be able to take the location information, e.g. user is in roaming or not roaming, state of the UE in CS domain and IMS, user preferences, service subscription and operator policy into account for both originating and terminating calls.

For terminating calls to a user

- SDS selects terminating services be provided either in the CS or the IMS service domain.

For originating calls from a user:

- SDS selects the service domain for originating calls from a user, i.e. it chooses whether originating services shall be provided by the CS or IMS service engine.

11.3.2 ADS Requirements

Access Domain Selection selects either CS access or packet access that is to be used to deliver a call between the UE and the network, as applicable for certain subscriber types e.g. a VCC subscriber which is able to initial or receive the call in CS domain or IMS.

Generic requirements:

- ADS is always the functionality nearest to the user

For terminating sessions to a user

- ADS is always performed after the execution of SDS and the terminating services.
- ADS may be a functionality located in CS domain, and/or in the IMS as well.

The ADS shall be able to take following factors into account for domain selection decision:

- The state of the UE in the circuit switched domain. This state information shall be included: Detached, Attached.
- The state of the UE in the IMS. The state information shall include: Registered, Deregistered.
- The domain used by an existing session.
- The media components included in the incoming IMS multimedia telephony.
- User preferences and operator policy

For origination sessions from a user:

- ADS is a functionality of the UE to choose the either CS, PS or IP-CAN access network to originate the session.

11.3.3 Relationship between SDS and ADS

The relationship between SDS and ADS in the originating side can be shown as below:

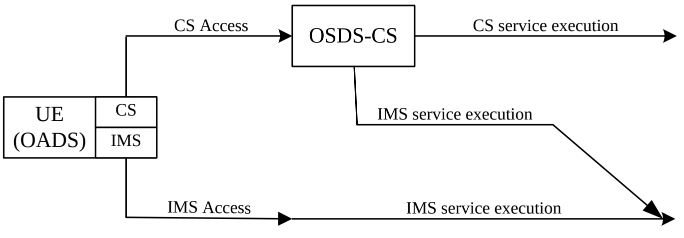

Figure 11.1 Relationship between SDS and ADS in the CS domain and IMS originating side

Note: Till now, no use case is identified to support the necessity for routing originating IMS session to CS domain to execute service logic.

The SDS and the ADS functionality for terminating calls for IMS multimedia telephony in IMS domain have a few things in common. These include:

- Both the ADS and SDS require interfaces to the CS domain for moving a CS terminating call to the IMS (using CAP and potentially MAP)
- Both the ADS and SDS shall be able to query the CS attach status (via the Sh interface or the MAP interface).
- Both are impacted by the user/operator preferences.

The relationship between SDS and ADS in the termination side can be shown as below in figure 11.2. Call termination in CS and IMS core network is shown separately.

In the terminating CS call case, if SDS selects the CS domain for service execution, no ADS is invoked and all calls are delivered by CS access to the UE. In this case there is no flexibility to choose the access domain.

Note: Separating TSDS-CS and CS termination services in the figure does not imply any restriction which technique is used to route calls to IMS.

Formatted: NO

For terminating IMS session IMS services are always selected but the access may be chosen as either CS or IMS.

Note: It is possible to route CS terminating call to IMS for services and still select the CS access.

Formatted: NO

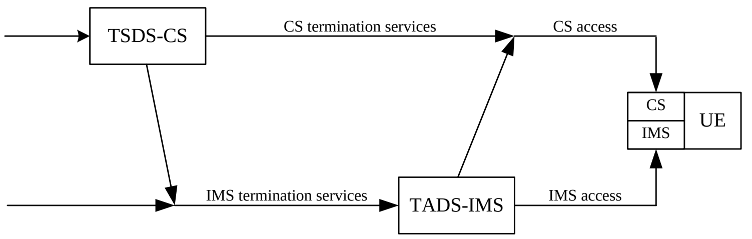

Figure 11.2 Relationship between SDS and ADS in the CS domain and IMS termination side

Formatted: TF

Note: Till now, no use case is identified to support the necessity for routing incoming IMS session to CS domain to execute service logic.

11.4 Conclusion

The service domain selection and access domain selection for different services may have functionality in common, but due to differences they are to be specified for each service.

The service domain selection and access domain selection should be included in the relevant specifications.

Formatted: Normal

12 Personal Network Management

12.1 General

PNM provides a service to manage the Personal Network (PN) of a user comprising a number of registered UEs and PANs. The user is provided with means to re-direct incoming services to UEs and PANs as configured by the user, i.e. is a terminating service. The management functions cover setup, configuration, and operation of Personal Networks. The management and invocation of PNM services is performed via a PNM network entity. This clause describes the architecture for PNM UE redirecting service as described in TS 22.259 [1].

PNM functionality is being provided via an AS in the IMS and a CAMEL service in the CS domain using interfaces available in Rel-7 or earlier Releases. Procedures for re-direction of services and set-up and configuration are identified.12.2 Overall architecture

Formatted: Normal

Working assumptions:

- PNM is realized as AS in the IMS and a CAMEL service in the CS domain
- In the IMS the PNM AS acts as a Routeing B2BUA as defined in 3GPP TS 23.218.
- PNM utilizes the Sh reference points in the IMS and MAP interface in the CS domain for the inquiry of HSS to perform validity check of PNM Registration/De-registration and subscription for terminating a specific service.
- Management procedures between UE and PNM AS are realized via the Ut reference point
- For support of legacy devices the UE shall be able to perform the basic management via USSD according to TS 22.090
- PNM redirection uses existing 3GPP mechanisms

The following figures illustrate the working assumption for the integration of the PNM UE Redirecting Service into the 3GPP system architecture. Figures 12.2.1 and 12.2.2 respectively show the architectures in the IMS and in the CS domain.

Note that the figures are for illustration of the PNM procedures; hence it only shows the components relevant to the PNM procedures.

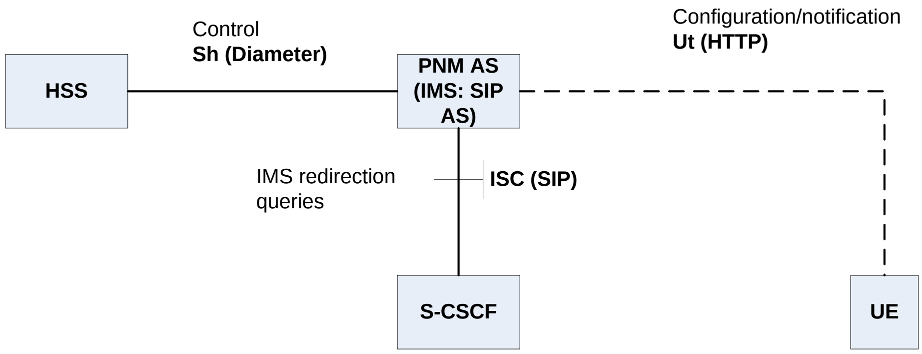

Figure 12.2.1: IMS architecture of the PNM UE Redirecting Service

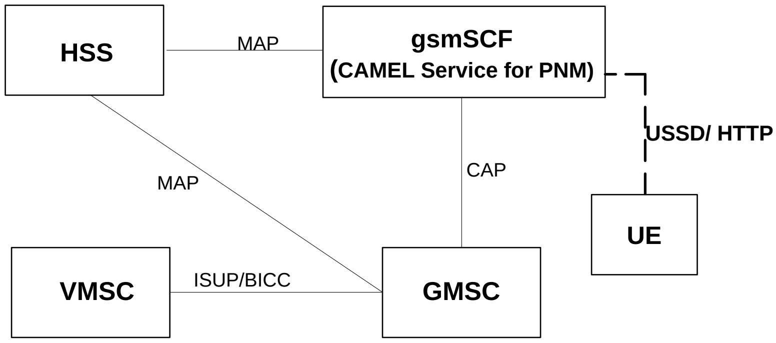

Figure 12.2.2:CS domain architecture of the PNM UE Redirecting Service

12.3 Procedures on Interfaces

In the following basic procedures are mapped to the interfaces of the architecture.

12.3.1 HSS – PNM AS/gsmSCF(CAMEL service for PNM)

Procedures involving the interface HSS – PNM AS/gsmSCF(CAMEL service for PNM):

- Enable the inquiry of HSS for validity check of PNM Registration/de-registration of UE (e.g. identify UEs belonging to a PN, prevent UE registration to more than one PN)
- Enable the inquiry of HSS for validity check of subscription for terminating a specific service

12.3.2 S-CSCF – PNM AS

Procedures involving the interface S-CSCF – PNM AS:

- Enable redirection according to PNM UE Redirecting Service setting
- Re-attempt the delivery of an incoming session towards activated UEs in a decreasing order of priority when the UEs with higher priority fail to establish the session.

12.3.3 UE – PNM AS/gsmSCF(CAMEL service for PNM)

NOTE: The UE should send USSD request to the gsmSCF(CAMEL service for PNM) via the VMSC, VLR and HSS in accordance with TS 23.090.

Procedures involving the interface UE – PNM AS/gsmSCF(CAMEL service for PNM):

- Registration/ de-registration of a UE (either self-registration of the UE or requesting registration of another UE)
- Activation (/de-activation) of a UE for all or for selected terminating services (either self-activation of the UE or requesting activation of another UE)
- Configure priorities for terminating a specific service
- Interrogation of settings in the PNM AS
- Switch to temporary activation settings/ remove+deactivate temporary settings
- Notifications from the PNM AS to the UE (e.g. last active UE for a service is deactivated)
- Update of UE related identities and capabilities (e.g. MSISDNs, URIs, GRUU)

- Invitation to UE for registration or activations
- Exclusion of public identities from redirecting
- Configure a UE of the PN as private UE (private UE of the PN means accessible from members of the PN and from the access list, further details see TS22.259)
- Configuration of the access list for private UEs of the PN

12.3.4 HSS – S-CSCF

Procedures involving the interface HSS – S-CSCF:

- Provision of the PNM AS initial filter criteria in the S-CSCF

12.3.5 HSS – GMSC

Procedures involving the interface HSS – GMSC:

- Provision the T-CSI for triggering of CAMEL service for PNM

12.3.6 GMSC –gsmSCF(CAMEL service for PNM)

Procedures involving the interface GMSC –gsmSCF(CAMEL service for PNM):

- Enable redirection according to PNM UE Redirecting Service setting
- Re-attempt the delivery of an incoming session towards activated UEs in a decreasing order of priority when the UEs with higher priority fail to establish the session.

12.4 Interaction

The Application Servers of PNM, VCC, and CSI may coexist in the same network. In this case all ASs are connected to the S-CSCF but the PNM AS is triggered first. The scenario is similar to the coexistence of VCC and CSI with an IMS Redirecting AS as specified in TS 23.218 and this specification.

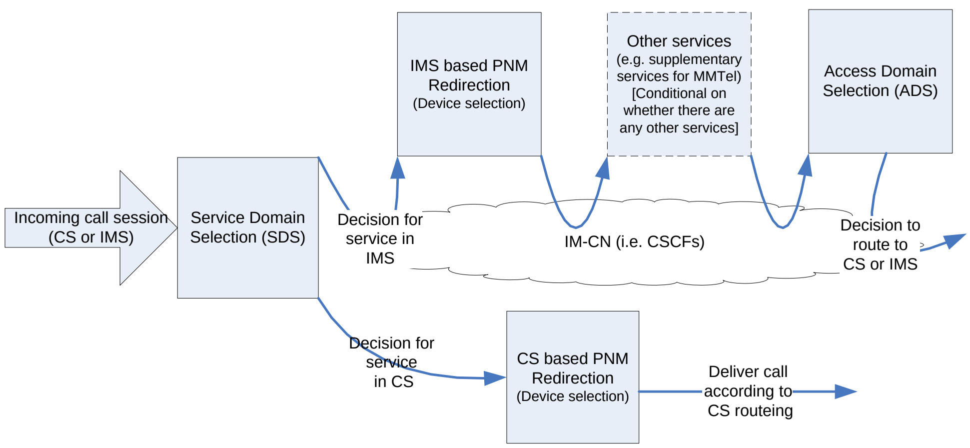

Figure 12.4.1: Example of coexistence of PNM with domain selection and other IMS based services

NOTE: If PNM Redirection changes the target MSISDN or Public User Identity of a call, the terminating handling shall be invoked again for the new public identity, which may include another SDS decision.

In this way the PNM AS selects the UE to which the incoming call will be terminated as configured by the user. Configuration data stored in the PNM AS are a result of administration by the user or service provider.

Access domain selection, implemented as per investigations captured in clause 11, is being invoked after the PNM AS has identified the terminating UE. There is no interaction between PNM and VCC.

The invocation of service logic after UE redirection needs careful implementation to ensure that circular routeing does not occur. A change to the Request-URI may re-invoke the full iFC processing and thereby the PNM AS is reinvoked.

12.5 Conclusion

The IMS-based PNM UE redirecting service has been shown to be possible to implement without any normative changes to existing stage 2 specifications within the control of SA2. The CT level stage 2 specification should take into account mechanisms to inhibit re-invocation of already applied service logic (iFC) when forwarding like services are performed.

NOTE: The synchronisation of the PNM service data and possible interactions between a CS-domain and an IMS-domain redirection service for a single Personal Network were out of scope of the stage 2 study.

Formatted: Heading1

13 Continuity of IMS-based Services

13.1 General

In general, the continuity of IMS-based services refers to the capability of continuing IMS-based communications (including circuit calls served in IMS), as we move from one access network to another access network. The main need for such continuity arises from the fact that mobile terminals roam within a multiplicity of access networks and consequently they occasionally need to change their access technology in order to satisfy several conditions, e.g. meet quality of service conditions, fulfil user preferences and/or operator policies, etc.

Figure 13.1 shows a typical example of a session continuity scenario in which UE-a transfers an IMS multimedia session established over an I-WLAN to a UTRAN/GERAN radio network. From a user point of view, the session needs to be transferred as seamlessly as possible. This creates several challenges however because due to the diverse characteristics of the various access networks it is not always easy to provide consistent service quality across a number of dissimilar access networks. For example, when a multimedia session including a voice component is transferred to an UTRAN/GERAN environment, the voice component might need to be transferred on the CS domain in order to maintain the necessary QoS and resource efficiency levels.

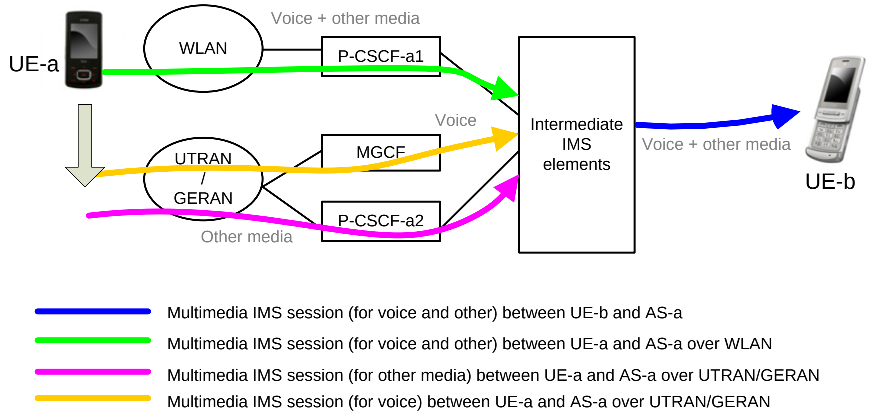

Figure 13.1: An example of a session continuity scenario.

This section discusses some of the issues related to the continuity of IMS-based services and in particular the specific case of PS-PS session continuity.

13.2 PS-PS Session Continuity

13.2.1 General

PS-PS session continuity is a special case of session continuity in which continuity of IMS session(s) is required after performing a PS-to-PS handover (for example, a handover from I-WLAN to 3G PS domain or to another IP-CAN, as illustrated in figure 13.2 for UE-a).

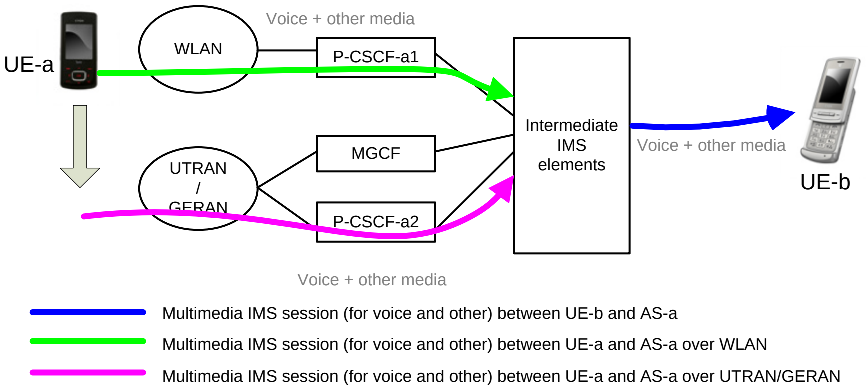

Figure 13.2

When the UE performs a PS-to-PS handover it typically changes its address at the network layer and possibly the outbound proxy (i.e., P-CSCF) that it is connected to. Consequently, it then needs to update its registration binding with the new contact address and also to transfer its ongoing IMS session(s) to the new contact address and possibly change their signalling paths (i.e., dialogs) to use the new P-CSCF.

There might be cases when the UE performs a PS-to-PS handover but it does not change its address at the network layer, for example when conducting a handover from 2G PS domain to 3G PS domain in the same PLMN, or when network mobility mechanisms are used (such as mobile IP). In such cases, communication continuity is achieved by means of lower-layer mechanisms (e.g. Mobile IP) and therefore there is no need to activate mobility mechanisms at the session (SIP) layer since the handover is transparent to this layer.

13.2.2 Potential Solution for PS-PS Session Continuity

Formatted: Heading3

PS-PS session continuity can be enabled by available SIP mobility and routing mechanisms (e.g. GRUU, INVITE with Replaces header, etc) should the session participants support these mechanisms. However, it cannot always be assured that all such SIP mechanisms will be supported by the session participants. This creates the need for providing a network function to support the PS-PS session continuity and effectively handle the relevant interworking aspects. This network function is termed as IMS Session Mobility Function (SMF) and is further described in this section.

The IMS Session Mobility Function (SMF) may be deployed to perform the following functions:

- Acts as a B2BUA anchoring IMS multimedia sessions originated by UE over an IP-CAN.

Formatted: B1

- Acts as a B2BUA anchoring incoming IMS multimedia sessions terminated at UE over an IP-CAN.

NOTE: Incoming and outgoing INVITE requests are routed to the IMS SMF using the iFC mechanism.

- Splits an IMS session into two separate legs, an access leg between the originating UE and the IMS SMF and a remote leg between the IMS SMF and the remote party.
- Hides and/or translates the SIP mechanisms used by the UE to implement session continuity (e.g., GRUUs, INVITE with Replaces) from the remote terminal which might not support those mechanisms (e.g., when the remote terminal does not support GRUUs or Replaces header).
- Terminates session update requests (e.g., UPDATE or re-INVITE requests to add/remove media streams, change or reconfigure codecs, etc.) received from either leg and interworks it with the other leg.
- Terminates session mobility request (e.g., re-INVITE or INVITE w/Replaces to change IP address and possibly change outbound proxy in signaling path) received from either leg and interworks it with the other leg.

Outgoing INVITE w/Replaces requests can either be addressed to the remote user of the original session in which case they are routed to the IMS SMF using iFC mechanism, or they can be addressed to a PSI representing the IMS session mobility service in which case they are routed directly to the IMS SMF.

13.2.2.1 PS-PS Session Continuity signalling flow

Formatted: Heading3

The following simplified example flow describes a possible use case of PS-PS session continuity. UE#1registers over one IP-CAN and establishes a video sharing session with UE#2. Then UE#1 moves and discovers and attaches to a new IP-CAN, registers over that and transfers the ongoing video sharing session to this new IP-CAN.

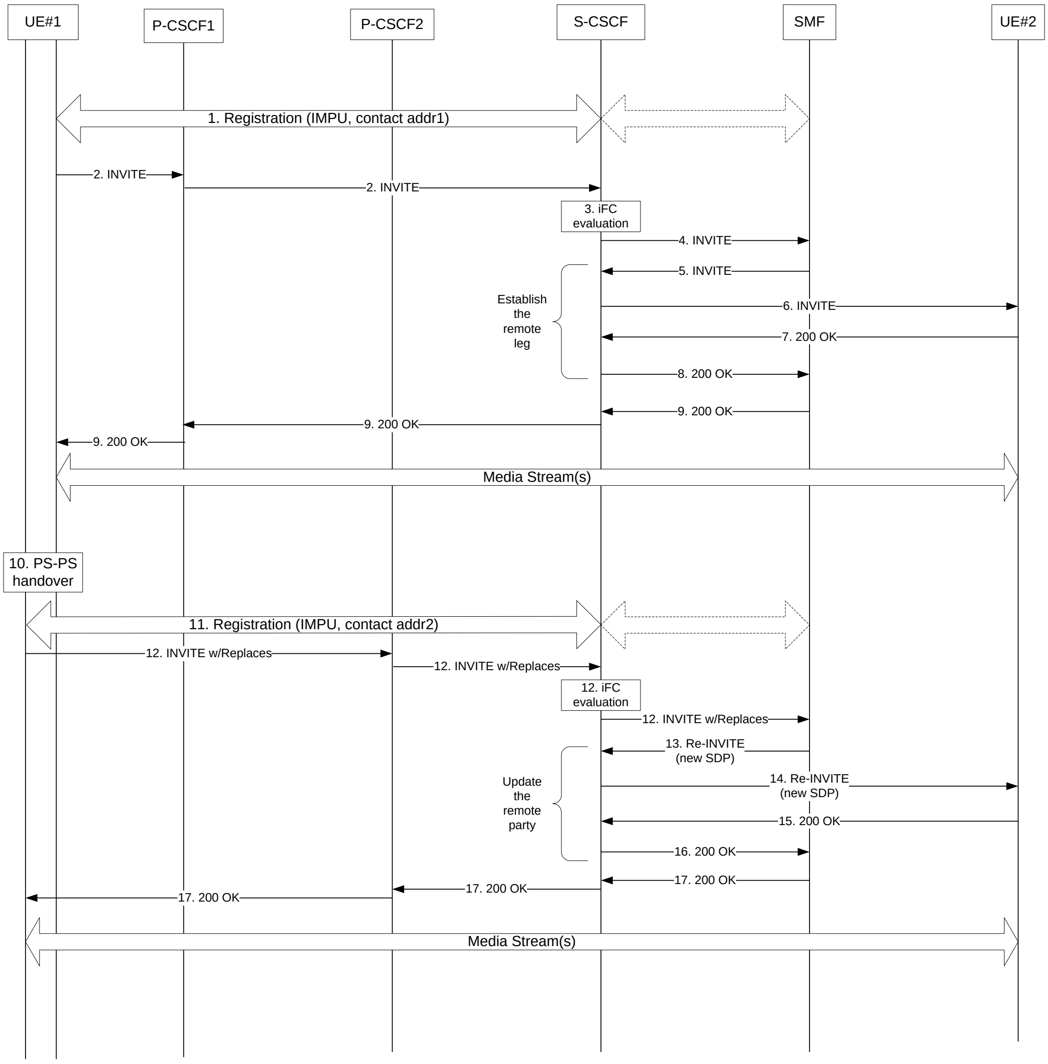

Figure 13.3: PS-PS Session Continuity signalling flow

A step-by-step description of the signalling flow is shown below:

1. UE#1 registers an IMPU with contact addr1 over one IP-CAN (e.g. 3G PS domain). 3rd-party registration may optionally be used.

Formatted: B1

2. UE#1 sends an INVITE request for initiating a video sharing session with UE#2.

3. The S-CSCF evaluates iFC to decide if the INVITE should be forwarded to SMF. Appropriate initial filter criteria could be used in order to forward sessions initiated from a PS domain or IP-CAN to SMF.

4. In this example flow, the S-CSCF forwards the INVITE to SMF.

5-6. Based on operator policy and/or other conditions the SMF decides to anchor this session and to establish the remote leg with UE#2. It therefore acts as a B2BUA and creates another dialog by sending a new INVITE request to UE#2.

7-8. UE#2 accepts the INVITE from SMF and responds with a 200 OK.

9. SMF responds to the INVITE sent by UE#1 in step 4 with a 200 OK. The video sharing session can then be established between UE#2 and UE#1.

10. Later, UE#1 discovers and attaches to a new IP-CAN (e.g. an I-WLAN). In the context of this attachment UE#1 obtains configuration information for the new IP-CAN including a new IP address (addr2). In the example shown in Figure 5.42a UE#1 also discovers a new P-CSCF (P-CSCF#2) that is applicable in the new IP-CAN.

NOTE: The discovery and attachment to the new IP-CAN might be triggered by poor QoS in the old IP-CAN (e.g. as a result of signal deterioration) or by applicable user preferences and/or network policy.

11. UE#1 registers again its IMPU with contact addr2. Again, 3rd-party registration may optionally be used.

Formatted: B1

12. To transfer its ongoing video sharing session from the old IP-CAN (addr1) to the new IP-CAN (addr2), UE#1 sends an INVITE with the Replaces header (this initiates a new dialog), effectively requesting from the remote party to replace the previous dialog settings (including the address(es) for media transport) with the settings in the new INVITE with Replaces header. The INVITE with Replaces header sent by UE#1 includes in SDP a new address (addr2) for the video sharing media. This INVITE creates a new dialog between UE#1 and SMF.

13-14. SMF interprets the content of Replaces header and identifies if there is matching ongoing SIP session that the SMF has previously anchored. In this case the SMF updates this existing dialog with UE#2 by sending a re-INVITE (or UPDATE) message which contains the new SDP payload transmitted by UE#1 in step 12.

15-16. UE#2 accepts the dialog update and the new SDP by sending a 200 OK response to SMF.

17. SMF responds to the INVITE sent by UE#1 in step 12 with a 200 OK effectively accepting the request to transfer the video sharing session from the old contact (addr1) to the new one (addr2). The video sharing session between UE#1 and UE#2 is then continued by used the new contact address of UE#1 (addr2).

13.3 Conclusion

For enabling continuity of IMS-based services several issues have been investigated and discussed. Also, one specific solution has been proposed for the special case of PS-PS session continuity in Rel-7 timeframe. However, it was felt more appropriate to study the general problem of IMS session continuity in altogether as opposed to first introducing an interim and partial solution for the special case of PS-PS session continuity in Rel-7 and then study the rest of the IMS continuity issues in a subsequent release.

Based on the above discussion, it was agreed that the general problem of continuity of IMS-based services should be further studied in a subsequent release.

Formatted: Normal

Annex A:
Analysis of operator controlled QoS impact on GPRS

A.1 Solution analysis of impacts of mechanisms for operator controlled QoS in a GPRS IP-CAN

In order to overcome the limitations identified when using UE initiated media bearer establishment, means should be introduced to allow the network to, based on e.g. the 'SDP Offer information, provide the UE with the appropriate bearer QoS to request for the service. This would provide for an alternative solution, where the PCRF defines an appropriate QoS for a service according to operator policy, and provides this information to the GPRS network, which then provides it to the UE during the network requested media bearer establishment. Since the network is the originator of the request for media bearer establishment it will be able to apply a consistent error handling in case of e.g. failure in establishing the media bearer.

Three main components have been identified as the necessary means to realize operator controlled QoS:

a) Network Requested Secondary PDP Context Activation procedure (NRSPCA)

b) Indication of which media flows shall use the PDP context using a TFT (including indication for any uplink media flows with uplink filter information)

c) Indication of the support of NRSPCA in UE and network

The components and their respective impact on GPRS are further discussed in the following sub-sections.

A.2 Network Requested Secondary PDP Context Activation procedure

The Network Requested Secondary PDP Context Activation procedure (NRSPCA) is a new Session Management procedure for the GPRS IP-CAN. It should preferably be based on the existing Secondary PDP Context Activation procedure [23.060 section 9.2.2.1.1], i.e. a Network Requested Secondary PDP Context Activation procedure.

A proposed stage-2 signaling flow is depicted below:

The Network Requested Secondary PDP Context Activation Procedure allows the GGSN to initiate the Secondary PDP Context Activation Procedure (23.060, 9.2.2.1.1). The Network Requested Secondary PDP Context Activation Procedure is illustrated in Figure A.1.

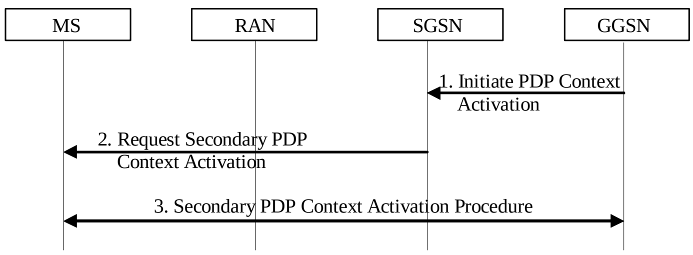

Figure A.1: Network Requested Secondary PDP Context Activation Procedure

1) The GGSN sends an Initiate PDP Context Activation (Linked NSAPI, QoS Requested, TFT, Protocol Configuration Options) message to the SGSN. The QoS Requested, TFT, and Protocol Configuration Options are sent transparently through the SGSN.

2) The SGSN sends a Request Secondary PDP Context Activation (Linked TI, TI, QoS Requested, TFT, Protocol Configuration Options) message to the MS. The Linked TI indicates the TI value assigned to the Activated PDP Context corresponding to the Linked NSAPI previously received as described in step 1 above.

3) The MS initiates the Secondary PDP Context activation procedure as described in 23.060, 9.2.2.1.1. The Linked TI, TI, QoS Requested, TFT, and Protocol Configuration Options sent in the Activate secondary PDP Context Request shall be the same as previously received in step 2 above.

A.3 Indication of media flows using a TFT

When the UE initiates a media bearer establishment in a GPRS IP-CAN through the Secondary PDP Context Activation Procedure (23.060, 9.2.2.1.1), it is able to establish a relation (binding) of the PDP context to a media flow. In the case of UE initiated bearer establishment the binding will most likely be static, i.e. not changed during the life time of the PDP context, and be used for routing of uplink traffic. The routing of downlink traffic can be achieved through the down-link packet filters in the Traffic Flow Template (TFT), situated in the GGSN. There is currently no standardized way to change the uplink traffic binding to PDP Contexts from the network. The PCC architecture uses the TFT only for binding media flows to PDP contexts. For the routing of downlink traffic PCC rules are used instead of the TFT.

When using the Network Requested Secondary PDP Context Activation Procedure there is a need to establish a unique relation between the activated PDP Context and the media flow(s) for which it is activated by the network. There is no obvious and reliable way for the UE to establish such binding without further information from the network. There is furthermore currently no standardized means for the network to ensure that the correct media flow(s) is/are put on a certain PDP Context when using Network Requested Secondary PDP Context Activation.

It is proposed to use the TFT to indicate to the UE which media flows shall use this PDP context. In addition to the downlink packet filters, the TFT shall also contain any uplink filtering information. The uplink filtering information should preferably consist of the same Packet filter attributes as the current downlink packet filters. The TFT can be sent to the UE during the Network Requested Secondary PDP Context Activation procedure. This is depicted in figureºA.2 below.

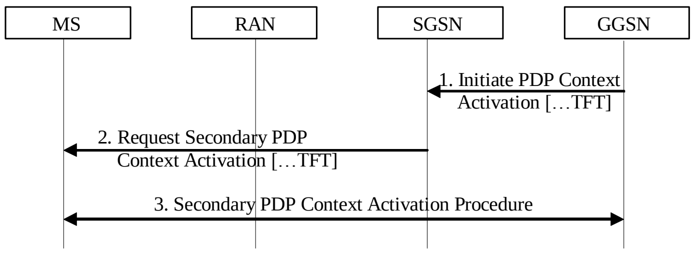

Figure A.2: TFT with packet filtering information sent to the UE in the Network Requested Secondary PDP Context Activation Procedure

In order to enable further operator control it is proposed to add a possibility to update the TFT in the UE from the GGSN by a small modification to the GGSN-Initiated PDP Context Modification Procedure (23.060, 9.2.3.2). This is depicted in figureºA.3 below.

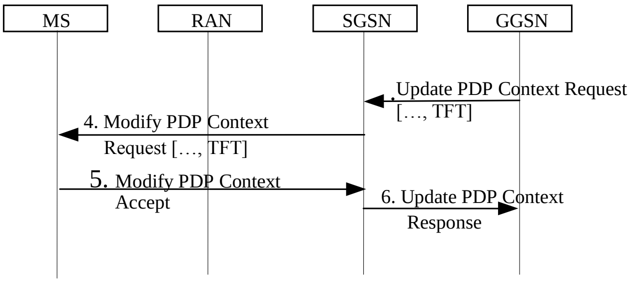

Figure A.3: TFT with packet filtering information sent to the UE in the GGSN-Initiated PDP Context Modification Procedure

A.4 Indication of the support of NRSPCA in UE and the IP-CAN

When the option of Network Requested Secondary PDP Context Activation is added to the GPRS IP-CAN it becomes necessary to introduce means for the UE as well as the different network nodes (SGSN, GGSN and possibly PCRF) to indicate possible support of the Network Requested Secondary PDP Context Activation Procedure.

Without such an indication it could become unclear for e.g. the UE whether or not to expect the network to request setup of the media bearers. This in turn could lead to 'dead-lock' situations with UE and Network both waiting for the other party to start.

The indication should preferably be added to the PDP Context Activation Procedure so that the UE could indicate the support to SGSN. SGSN will add its capability to the Create PDP Context request, and the GGSN could, based on its capability, choose to modify the indication when passing on the request. This is depicted in figure°A.4 below.

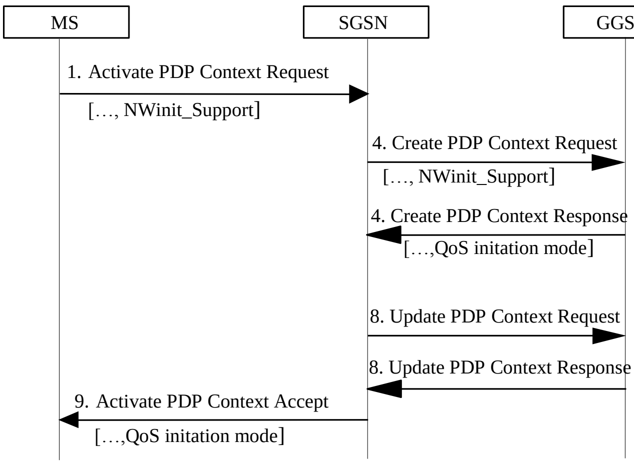

Figure A.4: NRSPCA indication sent in the PDP Context Activation Procedure

The PCRF may make use of the indication to decide the 'QoS initiation mode' to use for the PDP address / APN pair (IP-CAN session). The PCRF may not need to receive or use the indication, e.g. if the PCRF always pushes rules to the GGSN.

If there is a need for explicit support for the NRSPCA procedure at the SGSN, the indication should be included in the Routing Area Update procedure in order to update the indication whenever the SGSN is changed to allow the GGSN and UE to determine support at the current serving SGSN for the NRSPCA procedures.

The indications should serve two purposes: 1) Allow the UE and network to exchange capability information on the support of NRSPCA. 2) Allow the network to indicate the 'QoS initiation mode' for context activations within a PDP Address/APN pair (IP-CAN Session).

It should be an operator choice/option whether to allow either of network- or UE requested establishment only, or to allow a mix thereof. To cater for the different cases it is proposed to define that the indicated 'QoS initiation mode' is one of 'NW only', 'UE only', or 'NW/UE':

- When 'NW only' is indicated as the 'QoS initiation mode' the UE should assume that the network takes care of bearer QoS establishment and modification(s), and that any UE-initiated establishment request may be rejected. This implies that an IMS client in the UE should trigger SDP Offer/Answer indicating that the UE is able to receive the offered media.
- When 'UE only' is indicated as the 'QoS initiation mode' the UE should assume that any bearer establishment or modification must be triggered from the UE. The network may still reject or modify bearers according to operator policies.
- When 'NW/UE' is indicated as the 'QoS initiation mode', meaning that co-existence is allowed, there is a need to establish a clear rule or set of rules to make it possible for both UE and network to know which method to use.

Editor's note: the details of the mechanism for handling the different modes of QoS initiation will need to be revisited before being accepted in the TS

A.5 Conclusion

In this Section it has been shown that operator controlled QoS handling could be realized by introducing mainly three components, all with relatively small impact on the current specifications.

Annex B:
Analysis of impact of presence

B.1 Estimating presence traffic volumes

As previously mentioned, besides call-related signalling, SIP is also used to carry media when the protocol is utilized for the presence and short message services. Due to the possibility of "automatic" message exchange without user interaction, it is very likely that the SIP based presence service will create the major part of the non-call related IMS signalling (at least when the penetration of the service has become large). Thus to understand the intensity of which non-call related IMS signalling is transmitted, it is useful to derive a traffic model for SIP based presence. This sub-clause derives such traffic model for the presence service.

B.1.1 A presence traffic model

Here is a list of traffic model parameters and definitions used:

- Change of state is a presence event
    - Presence events occurs when the presentities change state and thus creates traffic on the interface between the presentities and the presence server.
    - Presence event creates notifications on the interface between the watcher and the presence server
- Frequency of Presence Event**:** f_{PE}
    - Average time until an Presence Event occurs $t_{PE}= 1/ f_{PE}$
- Frequency of notifications**:** f_{Not}
    - Average time until a notification is sent: $t_{Not}= 1/ f_{Not}$
- Frequency of traffic for the presence service per watcher
    - $f_{tot} = f_{Not} + f_{PE}$; where $f_{Not} = f_{PE} *N$; where N is the number of presentities on the watchers buddy list; → $f_{tot} = f_{PE} *N + f_{PE}$

For the purpose of showing the impact of the number of presence states used, in the traffic volume calculations two presence service settings are used;

- one that use two states: registered and un-registered, and
- one that use three states: registered, un-registered and busy in a phone call (Figure B.1 depicts the three state presence model).

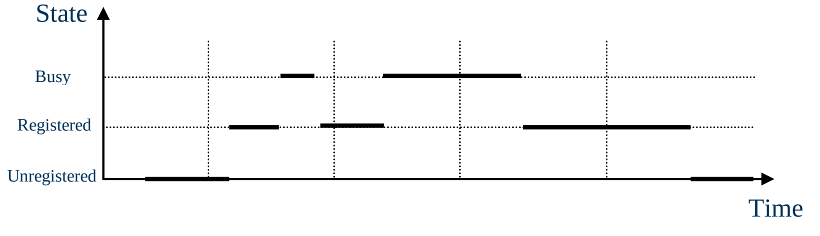

Figure B.1: The three state presence model

Frequencies for the presence state changes may be evaluated by detailed user behaviour studies and scenarios. However, it should be noted that in this paper a set of frequencies are assumed with no detailed analysis behind them. The assumed frequencies for presence state changes are;

- frequency of registering when unregistered (f_{Reg}): 2.27 (i.e. registers after 26 minutes),
- frequency of de-register ($f_{De\text{-}reg}$): 0.21 (i.e. stays registered for 4h 45 minutes)
- frequency of going busy when registered (f_{Busy}): 2 (i.e. assume that the presentity receives two calls per hour)
- frequency of being only registered after being busy ($f_{Busy\text{->}Reg}$): 15 (i.e. assume that the duration of the call is 4 minutes)

When having assumptions on the frequencies for presence state changes, the probability that a users is in a certain state can be calculated by using limiting probability.

For the two state presence service setting the equations for this limiting probability calculation are:

- $P_1 + P_2 = 1$; (where P_2 is the probability that the user is registered and P_1 is the probability that the user is unregistered)
- $P_1 * f_{Reg} = P_2 * f_{De\text{-}reg}$
- $\rightarrow P_1 = 0.08;\ \ P_2 = 0.92$

For the three state presence service setting the equations for this limiting probability calculation are:

- $P_1 + P_2 + P_3 = 1$ (where P_3 is the probability that the user is busy in a call)
- $P_1 * f_{Reg} = P_2 * f_{De\text{-}reg}$
- $P_2 * f_{De\text{-}reg} + P_2 * f_{Busy} = P_1 * f_{Reg} + P_3 * f_{Busy\text{->}Reg}$
- $\rightarrow P_1 = 0.07;\ \ P_2 = 0.82;\ \ P_3 = 0.11;$

The frequency of presence events per watcher can now be calculated by the following equations;

- Two states: $f_{tot} = (N + 1) * (P_2 * f_{De\text{-}reg} + P_1 * f_{Reg})$
- Three states: $f_{tot} = (N + 1) * (P_2 * (f_{De\text{-}reg} + f_{Busy}) + P_1 * f_{Reg} + P_3 * f_{Busy\text{->}Reg})$

Figure B.2 shows the number of presence events per watcher as a function of the number of presentities on the watchers buddy list.

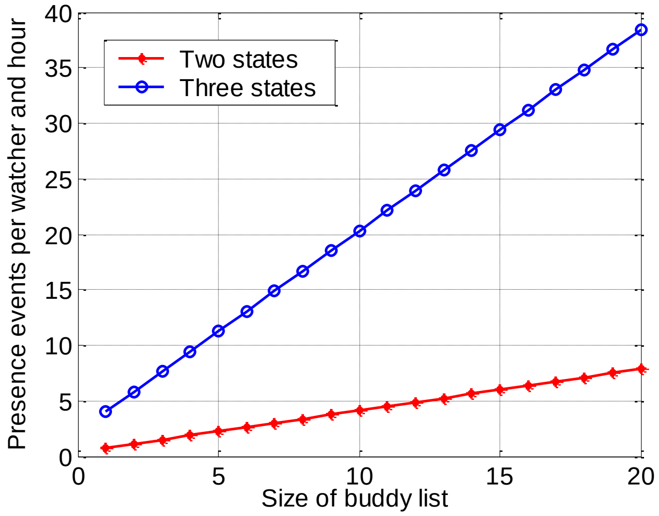

Figure B.2: The number of presence events per hour for one watcher as a function of the size of the buddy list

From the analysis above it can be concluded that the traffic created by the presence service is highly dependent on the size of the buddy lists and the number of possible presence states. Having three states and the frequencies for state changes assumed earlier, and a buddy-list of 20 users you will have 38 presence events per hour meaning that 76 SIP messages will be sent/received by this particular watchers terminal during the hour.

B.2 Impact on application layer and UTRAN

This sub-clause discusses the impacts of non-call related signaling on the application layer and UTRAN.

B.2.1 Impact on the application layer

B.2.1.1 Presence interferes the call related signalling

The presence related signaling share the same bearer as the call-related signaling. Therefore, presence updates may interfere with delay sensitive call-related signaling causing additional delays in the multimedia telephony call set-up. The probability for this to happen is however quite small. From the analysis in previous section we can assume that presence updates will happen with in the interval of one per second minute (this number can easily be reduced, see Sub-clause 6.2.2), while a multimedia telephony call set-up should be concluded in 4-8 seconds. A simple estimation of the risk of having presence interfering with the multimedia telephony call set-up would be 4-8 seconds/120 seconds (~average interval between presence updates). Thus the risk is in the region of 1 per 15-30 call set-ups will be interfered.

Unless the presence client always updates the presence state when the user either places a call or receives an invitation to a call. In that case the presence related signalling will always interfere with the multimedia telephony call related signalling and/or media transfer.

B.2.1.2 Presence influence SigComp compression ratios

As an option SigComp , see [RFC3320] and [RFC3321], may use dynamic compression (see [RFC3321]) meaning that it uses information in previously sent/received SIP messages as states to further increase compression efficiency. When using the presence service, there is thus a risk that presence messages will remove call related information in the stored states thus reducing compression efficiency for the delay sensitive multimedia telephony call set-up messages.

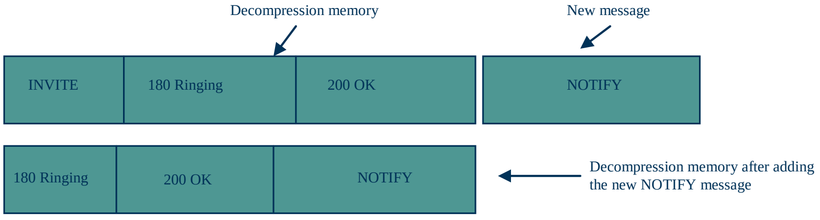

Figure B.3: SigComp decompression memory influenced by presence

B.2.2 Impact on UTRAN

The presence related signaling share the same bearer as the call-related signaling. The bearer used for signalling related to SIP session establishment and tear-down may have a higher allocation/retention priority compared to media bearer, to enable successful connection of emergency sessions in a loaded cell. If a higher allocation/retention priority is used for the signalling bearer, the "automatic" presence message exchange without user interaction threatens to reduce multimedia telephony capacity.

The reason for this is that every time a presence event occurs, a transport channel is needed and thus the UE needs to consume some resources in the network. When using WCDMA and a high allocation/retention priority every presence event will force the UE to go to the radio resource management state Cell_DCH. In Cell_DCH the UE consumes resources like codes and channel elements. Therefore in a loaded cell, the transition of the UE to Cell_DCH due to a presence update may trigger the termination (blocking) of an already connected media bearer that has lower allocation/retention priority.

Every time a presence update triggers the use of a radio channel, the channel will be established, kept and terminated during a certain amount of time. For bursty services with limited amount of data sent/received at every message exchange like presence, the time of establishing, releasing and keeping the channel during the expiry of an inactivity timer is far longer than the actual transmission time of the presence message. This is shown in Figure B.4.

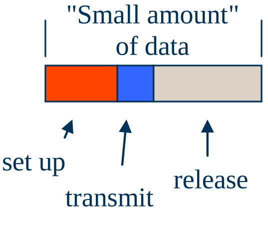

Figure B.4: The overhead created by establishing and releasing the channel is large for presence

The maximum number of presence users, filling the cell completely with presence traffic, can be calculated for WCDMA HSPA. Assumptions:

- Configurable down switch timer from Cell_DCH to Cell_FACH or URA_PCH;
  - WCDMA HSPA: 10 s.
- Channel held ~700 ms during set up assuming URA_PCH (the channel up-switch and cell update procedures).
- Channel held ~500 ms during release (the channel down-switch procedure).
- Transmission time 32 ms for 4 Kbyte message (WCDMA HSPA with 1Mbps).
- UE is in Cell_DCH thus consuming code resources for 11.23 seconds s for a presence message.

For WCDMA HSPA we can assume that 10/16 of the code tree are used for the data channel while 5/16 of the code tree are assigned to the associated DPCH (A-DPCH) or the fractional DPCH (F-DPCH). The A-DPCH consumes 1/256 of the code tree per user giving a maximum of (5/16) / (1/256) = 80 simultaneous users of the A-DPCH . Thus the maximum number of presence users the cell can handle is: 80 (number of simultaneous users) * 3600 (seconds per hour) / (38 (number of presence events) * 11.23 (channel allocation time per presence event)) = 674 presence users for WCDMA HSPA. When using the F-DPCH that is recommended for multimedia telephony the code limitation limit can be multiplied with thus giving 6740 presence users. Note these numbers assume 100% code utilization and no code consumption due to soft handover, so in reality the maximum number of presence users should be less than 674 and 6740.

But for large presence messages as assumed here, the WCDMA HSPA system is more likely power or TTI limited. The power limitation is best found by system simulations and such results are not presented here. But here follows a simple investigation of the TTI limitation. The presence update message was 4000 bytes large and the WCDMA HSPA network provided a throughput of 1 Mbps in average. Thus, the amount of time needed to transmit the presence update message is 32 ms or 16 TTIs. In reality there will be a certain amount of retransmissions, lets assume ~25%. This means that the amount of TTIs used per presence event is 16*1.25 = 20. The amount of TTI per hour are; 500*3600 = 1800000. The amount of TTIs consumed by a user having 38 presence evets per hour are; 38*20 = 760. Thus the maximum number of presence users are; 1800000/760 that equals approximately 2400 users.

It can be noted that for R99 DCHs the system is code limited. Assuming the use of 64 kbps DCHs when an interactive RAB is established; for the R99 DCH case every presence event consumes 1/16 of the cell resources for 2.7 seconds (for R99 DCHs the downswitch timer must be short, here 1 second is assumed) this means that if we have 38 presence events per hour the maximum number of presence users the cell can handle is: 16 (number of DCHs) * 3600 (seconds per hour) / (38 (number of presence events) * 2.7 (channel allocation time per presence event)) = 560 presence users for WCDMA R99 DCH. Note these numbers assume 100% code utilization and no code consumption due to soft handover, so in reality the maximum number of presence users should be less than 560.

Annex C:
Solutions for the dynamic allocation of users to application servers

C.1 General

Annex C captures a number of potential solutions describing the dynamic allocation of users to application servers.

The problem description is described in clause 8.

C.2 Overview of potential solutions

In the clauses below, the following potential solutions are detailed:

- Flexible application server selection – HSS storage of selected application server.
  With this approach, the HSS is the location of where the selected application server is stored. The S-CSCF or a front-end for other interfaces, will perform a selection of the application server if an application server is not selected.
  The details are documented in C.3 below.

- Hierarchical application server – Application server storage of selected application server
  With this approach, the application server is the location of where the selected application server is stored. All initial requests are routed through a "representative AS", which is not included in the links for subsequent SIP messages.
  The details are documented in C.4 below.

C.3 Flexible application server selection – HSS storage of selected application server

C.3.1 Solution Description

C.3.1.1 SIP initiated SIP-AS allocation

In this section, the term "Specific SIP-AS name" is used to represent the FQDN that would uniquely resolve to an IP address of the physical SIP-AS serving the user.

The procedures for allocating a user to a SIP-AS based upon the reception of SIP signalling is shown below in figure C.3-1.

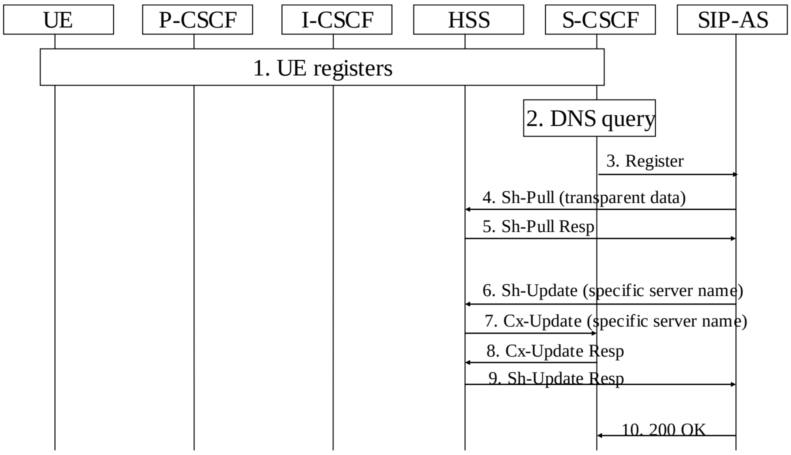

Figure C.3-1: SIP-AS allocation due to SIP registration

NOTE 1: While steps 4-5 and 6-9 are shown as separate information flows over the Sh interface, these could be combined for reasons of efficiency.

1. The UE registers with the network. The service profile is downloaded from the HSS to the S-CSCF. The service profile for the selected service contains a "server name" that could correspond to a number of SIP-ASs, and does not contain a "specific SIP-AS name" representing an allocated SIP-AS.
2. The S-CSCF performs the DNS query on the "server name" and resolves this to one of the IP address which represents one of the SIP-ASs.
3. The S-CSCF sends the 3rd Party register to the SIP-AS over the ISC
4. The SIP-AS requests the subscriber data contained in the transparent data over the Sh
5. The HSS returns the transparent data to the SIP-AS
6. The SIP-AS writes the specific name of the SIP-AS to the HSS
7. The HSS informs the S-CSCF of the specific SIP-AS name.
8. The S-CSCF acknowledges the update
9. The HSS acknowledges the read of the data to the HSS.
10. The 200 OK is returned to the S-CSCF.

NOTE 2: While the above flow is for a SIP registration, the same principle can be applied to any SIP signalling

It can be seen in the flow contained in Figure C.3-1 that :

- The SIP-AS retrieves the subscriber data over the Sh-interface from the HSS. The subscriber data is stored in the transparent data. (steps 4-5).
- The SIP-AS writes the specific name of the selected SIP-AS into the HSS, and the HSS informs the S-CSCF of the specific name of the allocated SIP-AS

This allows the S-CSCF to forward any further relieved flows to the allocated SIP-AS.

If there was a SIP-AS already allocated to the user, then upon registration the S-CSCF would be provided with the name of the specific SIP-AS instead. This applies to IMPUs of the subscriber and to any application server for the subscriber with the same general name for the application server.

C.3.1.2 Ut interface based SIP-AS allocation

The procedures for allocating a user to a SIP-AS based upon the reception of signalling over the Ut interface when no S-CSCF is allocated for a user is shown below in figure C.3-2.

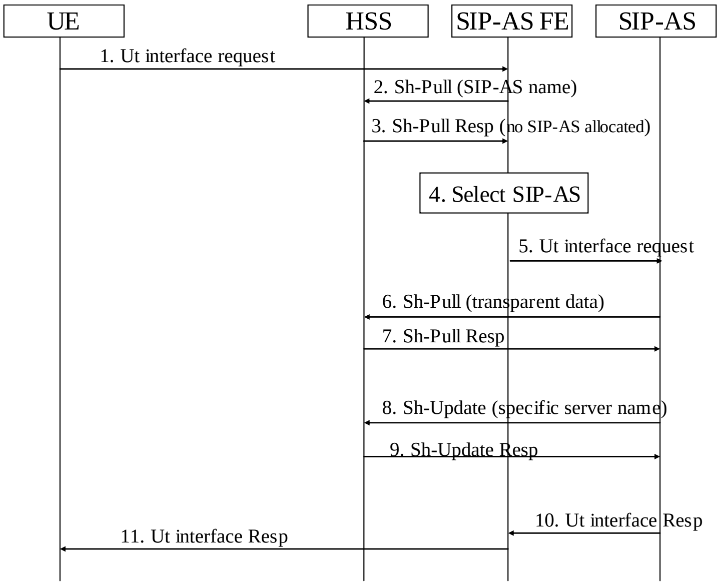

Figure C.3-2: SIP-AS allocation due to upon Ut interface signalling when no S-CSCF is allocated for a user.

NOTE: While steps 6-7 and 8-11 are shown as separate information flows over the Sh interface, these could be combined for reasons of efficiency.

1. The Ut request is sent to the configured address in the terminal – which reaches a SIP-AS front end.
2. The SIP-AS FE queries the HSS for the allocated SIP-AS
3. In this case, as there is not a SIP-AS already allocated, the HSS returns an indication that not SIP-AS has been allocated
4. The SIP-AS front end selects the SIP-AS.
5. The Ut request is sent to the selected SIP-AS
6. The SIP-AS request the subscriber data contained in the transparent data over the Sh interface
7. The HSS returns the transparent data to the SIP-AS
8. The SIP-AS writes the specific name of the SIP-AS to the HSS
9. The HSS acknowledges the read of the data to the SIP-AS.
10. The Ut interface response is returned to the SIP-AS front end.
11. The Ut interface response is returned to the UE.

It can be seen in the flow contained in Figure C.3-2 that :

- Upon the reception of a Ut interface request, the SIP-AS front end contacts the HSS to see if a SIP-AS has already been allocated.

- The SIP-AS retrieves the subscriber data over the Sh-interface from the HSS. The subscriber data is stored in the transparent data. (steps 6-7).
- The SIP-AS writes the specific name of the selected SIP-AS into the HSS

If there was a SIP-AS already allocated to the user, then specific SIP-AS name would be returned to the SIP-AS FE. The SIP-AS FE would return the Ut interface request to the specific SIP-AS.

The procedures for allocating a user to a SIP-AS based upon the reception of signalling over the Ut interface when a S-CSCF is allocated for a user is shown below in figure C.3-3.

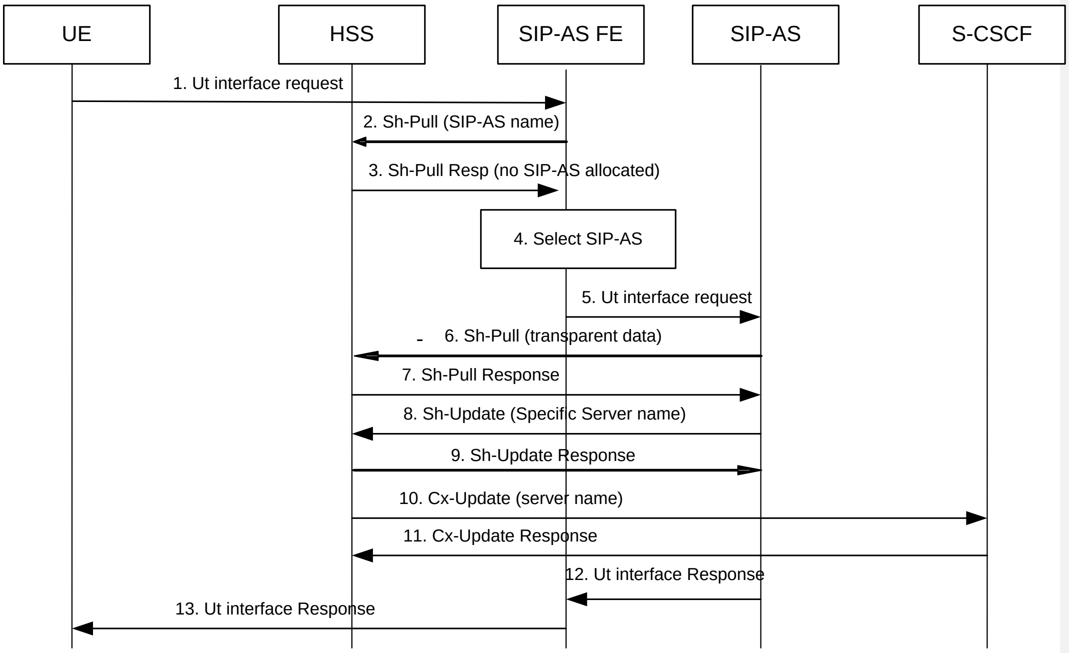

Figure C.3-3: SIP-AS allocation due to upon Ut interface signalling when a S-CSCF is allocated for a user.

Note: While steps 6-7 and 8-11 are shown as separate information flows over the Sh interface, these could be combined for reasons of efficiency.

1. The Ut request is sent to the configured address in the terminal – which reaches a SIP-AS front end.
2. The SIP-AS FE queries the HSS for the allocated SIP-AS
3. In this case, as there is not a SIP-AS already allocated, the HSS returns an indication that no SIP-AS has been allocated
4. The SIP-AS front end selects the SIP-AS.
5. The Ut request is sent to the selected SIP-AS
6. The SIP-AS request the subscriber data contained in the transparent data over the Sh
7. The HSS returns the transparent data to the SIP-AS
8. The SIP-AS writes the specific name of the SIP-AS to the HSS
9. The HSS acknowledges the read of the data to the SIP-AS.
10. The HSS sends a Cx-Update message to the S-CSCF
11. The S-CSCF sends a Cx-Update Response message back to the HSS

Note: Steps 10 & 11 may be performed before step 9 to ensure that S-CSCF is updated before acknowledgement is sent to AS.

12. The Ut interface response is returned to the SIP-AS front end.

13. The Ut interface response is returned to the UE.

C.3.1.3 De-allocation of user from a SIP-AS

The procedure for a SIP-AS to de-allocate a user when no S-CSCF is allocated for user is shown in Figure C.3-4.

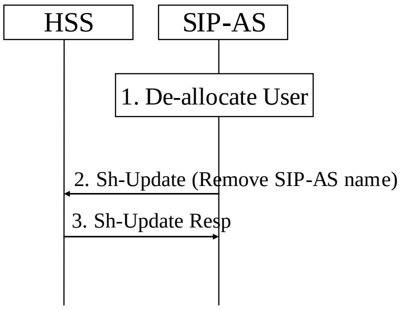

Figure C.3-4: SIP-AS de-allocating a user when no S-CSCF is allocated to the user

1. The SIP-AS decides to de-allocate a user from the SIP-AS.
2. The SIP-AS sends a Sh-Update to the HSS to remove the specific SIP-AS name for the user.
3. The Sh-Update Response is returned to the SIP-AS.

NOTE: The de-allocation of an application server may occur at when a user is considered to be de-registered from the network, though the de-allocation is not restricted to this case and may occur for other reasons.

The procedure for a SIP-AS to de-allocate a user when the user is still registered with the network, i.e. an S-CSCF is allocated for the user is shown below. Such a procedure may be carried out when either the current AS is overloaded due to other users or when the current AS is going down for maintenance.

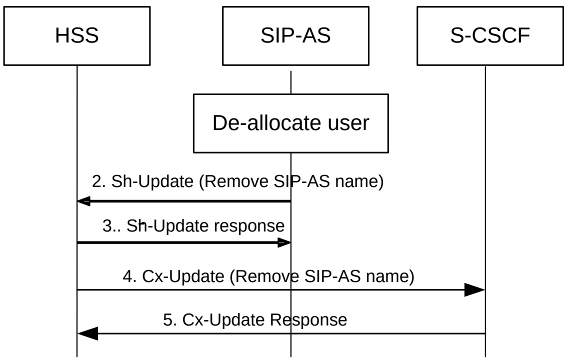

Figure C.3-5: SIP-AS de-allocating a user when an S-CSCF is allocated to the user.

1. The SIP-AS decides to de-allocate a user from the SIP-AS
2. The SIP-AS sends a Sh-Update to the HSS to remove the specific SIP-AS name for the user
3. The Sh-Update Response is returned to the SIP-AS

4. The HSS sends a Cx-Update message to the S-CSCF indicating the removal of the specific SIP-AS
5. The S-CSCF sends a Cx-Update Response to the HSS.

Note: Steps 4 & 5 may be performed before step 3 to ensure that S-CSCF is updated before acknowledgement is sent to AS.

Note: The S-CSCF can do a DNS query as shown in Figure C.3-1 to register the user with an alternate SIP-AS.

IMPACTS TO IMS ENTITIES:

- HSS:
  - The HSS needs to remember the AS that has been allocated to the user for every service
  - The HSS and Sh interface needs to be capable of handling large scale de-allocation of subscribers from a particular AS if it were to go into overload/failover. Otherwise system level sanity may be lost.
  - The Sh interface is required for ASs complying to this approach
- S-CSCF:
  - New signalling needs to defined over the Cx interface to notify S-CSCF of the AS selection and de-allocation.
  - FFS - S-CSCF needs to contain the data and logic that allows it to make decisions on server assignment and re-assignment. For example, if the AS sheds users, that AS should not factor into allocation procedures until that AS is ready to accept further messages.
  - FFS - The S-CSCF needs to have access to the data related to the health of each AS, so it can make decisions on assignment and re-assignment (in the case of unable to contact an AS) of users to application servers.

C.3.2 Solution Analysis

The Flexible application server approach is a scalable, efficient solution that allows multi-vendor deployment without the need of new interfaces.

The solution is efficient in that for traffic (i.e. call establishment) in that it minimises the need of an intermediate nodes between the S-CSCF and the AS performing the logic. This reduces the deployment cost as well as the latency on the ISC interface.

The solution is scalable in terms of re-using the basic scalability provided by the IMS (HSS, CSCFs etc), and the solution is valid independent of the number of instances of the application servers deployed. Means to favour application servers in the same site as the S-CSCF can be applied with efficient usage of DNS (other means may be possible as well).

The solution allows for inter-vendor deployment without requiring new interfaces in the IMS architecture. The IMS This is valid even irrespectively of whether the application server was selected due to SIP related activities (registration, terminating call); or whether it was due to other interfaces such as Ut, or other interfaces that may connect to an application server.

The noted disadvantage is that it requires an application server to employ the Sh interface in order to take advantage of this solution. The weight of this disadvantage is questionable though as solutions arise inside 3GPP that rely on the use of the Sh interfaces (e.g. VCC, ..).

The use of this solution for load distribution requires further study.

C.4 Hierarchical application server – Application server storage of selected application server

C.4.1 Solution Description

One of SIP application servers acts as a load balancer (or a distributor) and other application servers behind it provide the desired service to a user. In this section it is called as the Hierarchical application server architecture. Hereafter the application server acting as a distributor is a representative AS and an ASes at the back of the representative AS are back-end ASes. It is the name of a representative AS that is registered in the iFC. The S-CSCF routes the received request message from the UE to a representative AS according to iFC and a representative AS selects one of back-end ASes and route the request to it.

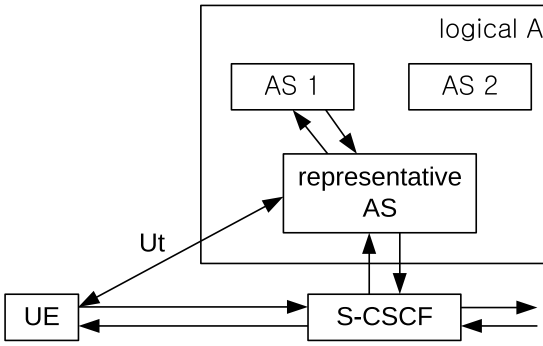

Figure C.4-1: processing initial/standalone request

In figure C.4-1, the S-CSCF routes the initial request from the UE to the representative AS as usual IMS service procedure. The representative AS selects one of back-end ASes and reroutes the request message received from the S-CSCF. Then the selected back-end AS invokes the service logic and returns the message back to the representative AS or the S-CSCF to proceed.

However, a SIP message is usually large and only single additional hop could result in additional routeing delay. SIP dialog consists of the initial request, the subsequent request and the corresponding responses. On receiving the initial request, the application server decides to remain or not in the subsequent requests using the Record-Route header. Therefore, routeing path can be optimized when a representative AS decides not to remain on the path and the forwarding delay will not happen. In some service scenarios, a representative AS doesn't even need to keep the dynamic allocation information because it is already embedded in the Record-Route header included in the initial response. In this way a representative AS can be a state-less SIP proxy server.

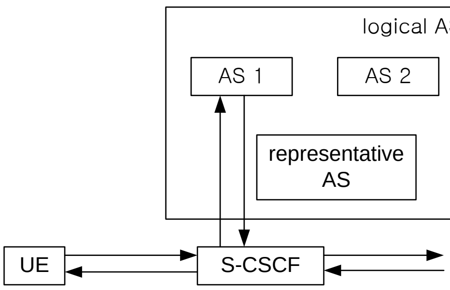

Figure C.4-2: optimized routeing path in subsequent request

In figure C.4-2, the UE directly sends subsequent request to the allocated back-end AS (AS 1) by incorporating explicitly the Route header built from the routeing information received in the response.

IMPACTS TO IMS ENTITIES:

- HSS:
    - None.
- S-CSCF:
    - FFS – depending on scaling solution.
    - FFS - server de-allocation needs to be communicated to the Representative AS.
    - FFS – The Representative-AS needs to contain the data and logic that allows it to make decisions on server assignment and re-assignment. For example, if the AS sheds users, that AS should not factor into allocation procedures until that AS is ready to accept further messages
    - FFS - The Representative-AS needs to have access to the data related to the health of each AS, so it can make decisions on assignment and re-assignment (in the case of unable to contact an AS) of users to application servers.

A new element, the Representative AS, needs to be added to the network. The representative AS would maintain the states of users and their allocated AS. All new sessions are initially routed though the Representative AS.

C.4.2 Solution Analysis

This solution requires an additional traversal of an additional functional entity for the initial signalling and requires the specification of a new interface in order to produce a scalable architecture that supports inter-vendor deployment.

The need of a new interface in order to ensure scalable solution that can be deployed in a mult-vendor scenario arises from the understanding that it is likely that a "single representative AS" is not likely to be sufficient to support all of the signalling for a logical AS in the network. Furthermore, it would be inefficient to have a single point in the network performing the AS selection. A large network deployment would have the application servers distributed across sites, and it would be inefficient to take all traffic via a single site.

Given that it does not seem reasonable to have a single central "representative AS", the alternative is to have multiple instances of a representative AS. Given that an AS may be selected for e.g. Registration, terminating call to unregistered users, or for a SIP-AS, then solution needs to work when the request for a user arrives at any Representative AS. As such, when receiving a new request for a user, that will result in a the allocation of a new AS, then the representative AS has to both distribute this decision to the other representative ASs as well as checking to see if another representative AS is not allocating a user at that particular time. This updating is likely to be quite a load and scalability of such an architecture represents certain challenges. This protocol would require standardisation if multi-vendor deployment is to be supported.

C.5 Dynamic assignment of application server by S-CSCF caching.

C.5.1 Solution Description

The solution "**Dynamic assignment of application server by S-CSCF caching**" is based on a new S-CSCF caching functionality.

The following figure C.5-1 shows the procedures for allocating an AS to a user, using existing DNS (RFC 1035) functionality and using a new S-CSCF caching functionality.

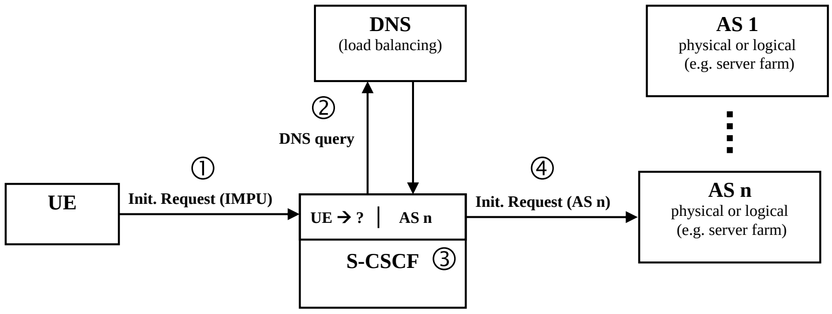

Figure C.5-1: Assignment of AS via DNS query during first Initial service request

First request of a service after IMS registration:

(1) After IMS registration a user sends an initial request to the S-CSCF for requesting a service (served by an AS).

(2) The S-CSCF performs the DNS query on the server name and resolves one (or a prioritised list) of the IP address(es), which represents a physical or logical AS.

(3) The S-CSCF caches the IP address of the assigned AS and stores it during the IMS registration period of the user.

(4) The S-CSCF routes the request to the assigned AS. (Depending on the service the AS could read/write/store user data (e.g. using Sh interface).

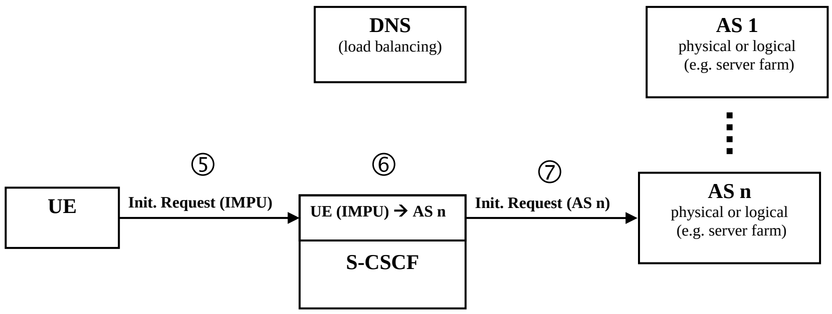

Figure C.5-2: S-CSCF has stored assigned AS for all following service requests

During the registration period of the IMS user, subsequent service requests to the S-CSCF can be routed directly to the assigned AS:

(5) The IMS user requests the service again and sends an initial request to the S-CSCF.

(6) The S-CSCF has stored the IP Address (or a prioritised list) of the assigned AS. There is no longer need to perform a time-consuming DNS query.

(7) The S-CSCF routes the request to the assigned AS. (Depending on the service the AS can reuse prior stored user data).

The AS pre-assignment and storage could be also done after downloading the service profile during the user registration procedure.

C.5.2 Conclusion

While this Solution requires only a small optional modification of the S-CSCF functionality and covers efficiently: the dynamic usage of multi-vendor environment, scalability, load balancing, redundancy issues and optimises call setup times, this solution has not yet address the scenarios when an application server is contactable via interfaces other than the ISC.

Annex D:
Network-initiated bearer control and PCC, high level description

D.1 Introduction

This annex considers services provided by the operator to the subscriber, subject to PCC control, and in particular IMS-based services. In these cases the subscriber is paying for a service experience rather than a L2 bitpipe. It is the responsibility of the operator to deliver the service to the subscriber with sufficient quality, and therefore to determine the adequate L2 bearer QoS mechanisms.

This section describes the additions to Rel7 nodes and protocols in order to provide network-initiated bearer control, and gives an overview of how these mechanisms could be used in an end-to-end IMS use case. Note that it is an operator decision whether to use network-initiated bearer control or not.

NOTE 1: This annex describes the case when the network controls the QoS, i.e. the NW-init mode and for those cases when the network controls the QoS in the mixed-mode.

NOTE 2: The use cases describe the UTRAN case (RAB operations). The solution applies similarly to GERAN.

D.2 Solution overview

D.2.1 Abbreviations

The following abbreviations are introduced in the description of the solution:

NRSPCA Network-Requested Secondary PDP Context Activation – a new SM procedure
ULTFT UpLink Traffic Flow Template – packet filters for a PDP context in the UE, for mapping uplink packets to the PDP context

D.2.2 Functional model

Figure D.2-1 shows the overall model for QoS policy control. The main QoS-related functions of the nodes involved are described below.

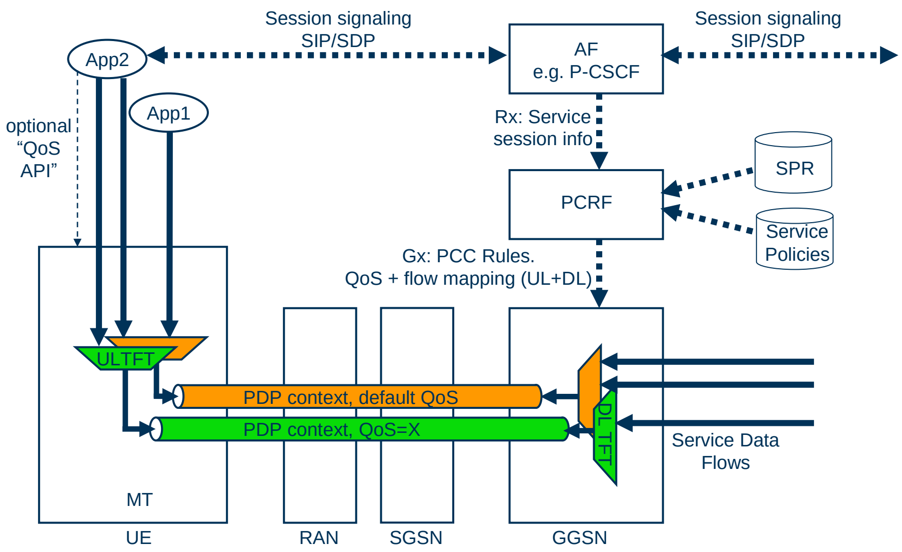

Figure D.2-1: Model for QoS control

Application in UE:

- Initiates/terminates end-to-end session signaling. (In this context, the IMS client would be part of the "application".)
- Acts as if resources are available, when the network has instructed the bearer layer ("MT") to use the NW-init mode
- For the IMS case: initiates sessions with "preconditions=met" (i.e. current qos equals desired qos), and indicates preconditions "not required". (Using "preconditions=not met" or "inactive" would require that the UE can determine when resources later are available. But with NW-init mode, the UE should not need such QoS policy data to determine this condition.)
- Is agnostic to the use of NW-init vs UE-init mode, and may therefore request QoS status from the bearer layer over a QoS API also for NW-init mode. In case of NW-init mode, the bearer layer (MT) then responds that resources are available.

AF:

- Provides service-related information extracted from session signaling to the PCRF, for potential use in deciding bearer QoS for the service data flows

PCRF:

- Mapping operator-provided service to bearer QoS class, and indicate this to GGSN in PCC Rules. The mapping is based on a combination of service session information received on Rx, policies provisioned by the operator and optionally subscriber information. Service session information include information that the AF can derive from the session level signaling, that may be used (according to operator policies) for deriving a suitable QoS class. This may include: media flow descriptions, IMS communication service identifier, destination address for network-provided services (e.g. URL), etc.

- Provides policy for QoS initiation mode per IP-CAN session (PDP address / APN pair for a UE), i.e. "UE-initiated bearer control" or "Network-initiated bearer control" or "Mixed mode".
- Policy-based decision on when a QoS Class (PDP context) should be established (e.g. early or late in an IMS sequence)
- Informs the P-CSCF about the outcome of the resources reservation for the IMS session.

GGSN:

- Initiates NW-init PDP context activation or modifications for a QoS class, based on PCC rules.
- Gate enforcement per service data flow for UL and DL. In particular, contains a DL TFT controlling mapping of DL flows onto PDP context, and an UL TFT, to police the correct mapping of uplink packets.
- Sends ULTFT to be installed in UE for mapping UL flows to correct PDP context.
- Indicates capability of NRSPCA to the PCRF

SGSN:

- Indicates capability of NRSPCA to the GGSN.
- Relays PDP context control signals from GGSN towards MT.

RAN:

- Sets up RABs and RB on demand (as today). Performs admission control based on GBR.

MT:

- Indicates capability of NRSCPA to the network.
- If "NW-init QoS" indicated from network: the UE should not initiate bearer control signaling itself, but should assume that resources will be provided by the network, including initiation of a new PDP context (if needed), modification of QoS of an existing PDP context (if needed), modification of the ULTFT of an existing PDP context (if needed) or release of a PDP context. Applications are informed that resources are available, if querying over a QoS API.

NOTE: The MT is still allowed to perform the release of PDP context procedure if required by events that are not related to a specific service (e.g. detach procedure).

- Stores an ULTFT per PDP context, that maps UL flows to that PDP context.
- If "NW-init QoS" indicated from network: the UE does not need to contain policies for determining bearer QoS based on service / session information. Such policies need only be provisioned to the PCRF in the network. (Any policies provisioned in the MT for UE-init mode, are not used in NW-init mode.)

P-CSCF:

- Continues with the IMS session establishment depending on the outcome of the resource reservation.

Editor's note: The different possibilities are FFS as well as whether and how the P-CSCF knows the requirements of an IMS session.

D.2.3 Additions to 3GPP protocols

The additional features needed for a complete solution for network-initiated bearer control are:

- Network-initiated bearer setup procedure
- Network-controlled uplink filters (ULTFT) in the UE (based on IP address, port numbers)

Editor's note: It is FFS, whether the DLTFT needs to be provided to the UE to allow for a correlation between services and bearers in the UE.

- Capability negotiation for support of NRSCPA

These are proposed to be supported by the following limited additions to 3GPP protocols:

- NRSPCA
  - A new procedure to let GGSN request a new PDP context from the SGSN. Parameters include a QoS profile and the associated ULTFT to be installed in the UE.
  - A new procedure to let the SGSN request a new PDP context from the UE. Parameters include a QoS profile and the associated ULTFT to be installed in the UE.
- Network-controlled modification of ULTFT
  - New information in the network-initiated PDP modification procedure (from GGSN to SGSN, and from SGSN to UE), including the associated ULTFT to be installed in the UE.
- Capability negotiation in IP-CAN session establishment for support of NRSCPA
  - A new parameter in the PDP context activation request from the UE to the SGSN, and from SGSN to GGSN, indicating the capability of the UE to handle NRSPCA and network-controlled ULTFT.
  - A new parameter in the PDP context activation response from the GGSN to the SGSN to the UE, indicating the QoS initiation mode (UE-initiation or NW-initiation or mixed mode)

NOTE: The working assumption is that changing the QoS initiation mode during an IP-CAN session should be avoided. In a realistic deployment, the operator has upgraded all SGSNs before starting to use NW-initiated mode.

Editor's note: It is FFS, whether a simple but brutal mechanism should be standardised, that allows the network to indicate a changed QoS initiation mode with a GGSN-initiated PDP context request, after the completion of an inter-SGSN Routeing Area Update, if the new SGSN has another capability. In some cases, loss of PDP contexts may occur, but since this is a rare event, this is seen acceptable.

- Support on Gx for the above, in particular:
  - Capability indication to PCRF, and bearer control initiation mode indication from PCRF
  - Push of PCC rules for a not yet existing bearer

Below is shown how these components can be used in a PCC framework to enable a complete and unambiguous control according to operator policies, in an IMS use case.

D.3 Use cases

D.3.1 PDP context use cases

The following use cases are used as components in the end-to-end use cases, but are here described in detail, including interaction between SGSN and GGSN.

D.3.1.1 Network-Initiated Bearer setup

Figure D.3-1 shows the signaling sequence for a network-initiated bearer setup.

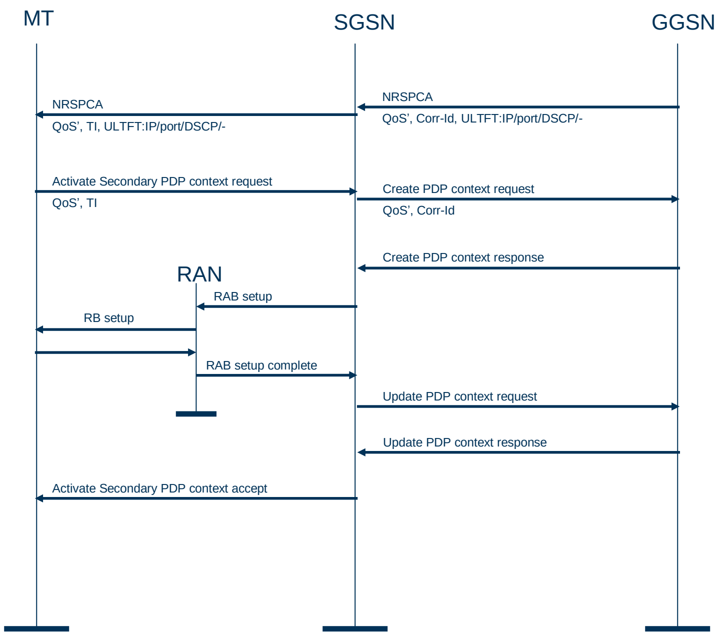

Figure D.3-1: Network-Initiated Bearer setup

The following steps are performed:

- the GGSN initiates a NRSPCA procedure towards the SGSN, including the QoS profile and a correlation identifier (Corr-Id). The ULTFT to be installed in the UE for this PDP context is also included. The ULTFT may be empty, or may be based on IP adresses, port numbers, protocol id , or a combination. The Corr-Id is later used by the GGSN to associate the new PDP context with this NRSPCA.
- the SGSN initiates the NRSPCA procedure towards the UE, including the same information as received from the GGSN, but with a TI (Transaction Identifier) instead of the Corr-Id. The SGSN stores the association of the TI with the Corr-Id.
- the UE stores the ULTFT associated with the PDP context being setup, and then triggers an Activate secondary PDP context request according to Rel6 specs, including the QoS profile and TI received in the NRSPCA.
- the SGSN triggers a Create PDP context request to the GGSN, and includes the Corr-Id associated with the TI received from the UE.
- the GGSN associates the Create PDP context request with the earlier NRSPCA by using the Corr-Id, and sends a Create PDP context response, including the QoS profile to be setup for this PDP context. (Note that this QoS profile may be different from the one received from the UE, e.g. due to resource or capability limitations of the UE.)
- the SGSN requests a RAB from RAN using the QoS profile received from GGSN in the Create PDP context response.
- the RAN sets up the RAB, including necessary Radio Bearer signaling, and returns a completion message.
- the SGSN informs the GGSN about the successful setup with the Update PDP context request procedure, and finalises the PDP context activation with a response to the UE.

D.3.1.2 Network-initiated PDP context modification

Figure D.3-2 shows two use cases for network-initiated PDP context modification: modification of the ULTFT (above) and modification of the GBR, i.e. admission control (below). These are described separately, however, a combined procedure may of course also be triggered.

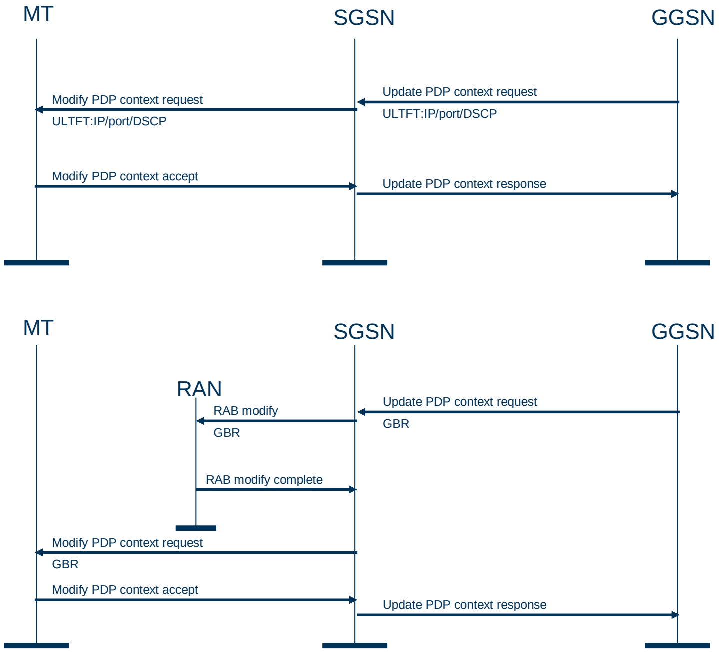

Figure D.3-2: Network-initiated PDP modification, of ULTFT (above) and GBR (below)

The PDP context modification of the ULTFT involves the steps:

- the GGSN triggers a Network-Initiated PDP context modification to the SGSN, including the new ULTFT for the PDP context. The ULTFT may be based on IP addresses, port numbers, protocol id, or a combination.
- the SGSN triggers a PDP context modification with the new ULTFT to the UE
- the UE installs the new ULTFT for this PDP context and responds to the SGSN
- the SGSN responds to the GGSN.

The PDP context modification of the GBR involves the steps:

- the GGSN triggers a Network-Initiated PDP context modification to the SGSN, including the new GBR for the PDP context
- the SGSN triggers a RAB modification with the new GBR to the RAN
- the RAN performs an admission control with the new GBR, and responds to the SGSN.
- the SGSN signals the new GBR to the UE, and waits for the response
- the SGSN responds to the GGSN.

It is assumed that, if the requested GBR could not be admitted, the RAN rejects the request.

D.3.2 End-to-end use case

Notes on the use cases:

- in the figures, the SGSN and GGSN are collapsed into one node. The interaction between them is described in the text.
- On Gx, QoS is described with the QoS class identifier, MBR and optionally GBR.
- Dotted arrows indicate messages that are not always present.
- The use cases should be seen as an example of a possible signaling sequence – not mandating this particular sequence in all implementations. In particular, the shown use cases apply for a particular policy configuration, when Network-initiated bearer control is supported by all nodes, and used by operator policy for all services.

Editor's note: The other use cases, e.g. when the PCRF decides to apply the mixed mode are FFS.

Editor's note: The use cases require also updates for the P-CSCF. The PCRF needs to have the knowledge about the P-CSCF version to be able to decide whether network-initiated bearer control is possible or not.

D.3.2.1 Provisioning phase

- The operator provisions the mapping rules for mapping of services (subject to PCC control) to bearer QoS class into the PCRF.
- These mapping rules may use service data (from Rx), subscription data (provisioned) or a combination as an input to the mapping decision.
- In these use cases, the operator provisions the policy to use network-initiated bearer control whenever possible (UE and SGSN capable).

NOTE: There is no need for the operator to provision terminals with rules how specific applications shall be handled.

D.3.2.2 IP-CAN session setup phase

Figure D.3-3 shows a use case for two UEs connecting to the IP CAN network.

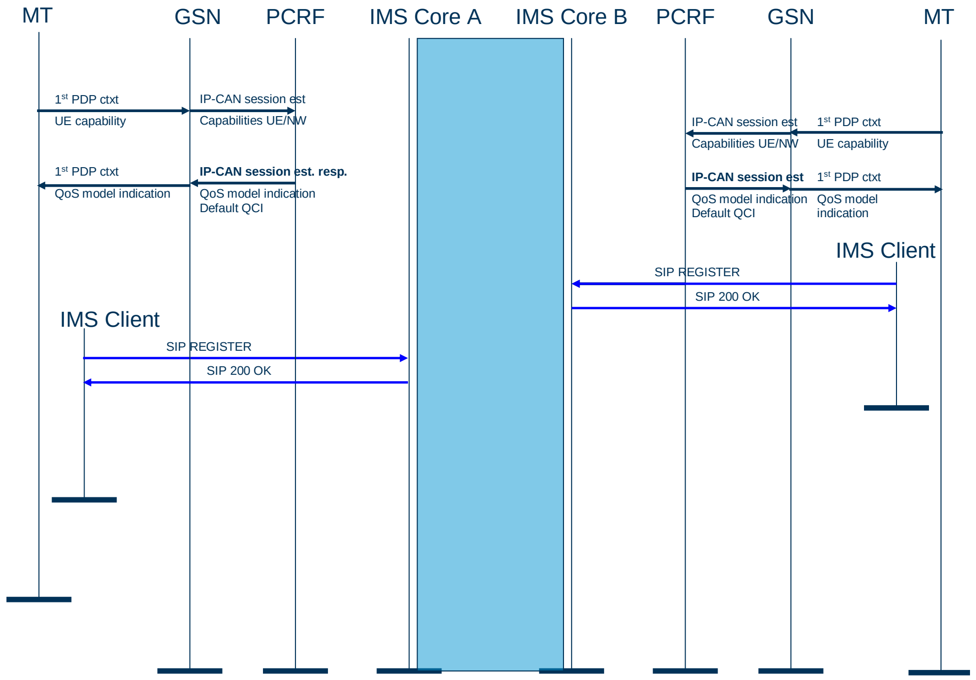

Figure D.3-3: IP-CAN session setup

- At IP-CAN session setup (PDP context activation to a PDP address / APN pair), the UE includes a capability indication indicating it is capable of network-initiated bearer control. In this use case, the PDP context established is a general-purpose PDP context (i.e. the UE does not include the PCO flag indicating that this PDP context would be used only for IMS signaling).
- The SGSN appends its capability indication, and the GGSN forwards UE + network capability indications to the PCRF.
- The PCRF takes the decision to use network-initiated bearer control for this IP-CAN session, and includes such an indication to the GGSN, for forwarding to the UE. Also, a default QoS level is enforced for this general-purpose PDP context.

D.3.2.3 Service setup phase, IMS call (Rx control)

First a thorough description of a normal call setup use case is described. Then some briefer notes are given on other cases for call setup, as well as for call clearing.

D.3.2.3.1 Successful call setup: Normal case

Figure D.3-4 and Figure D.3-5 below shows two use cases for an IMS call setup between two IMS clients, for the basic Network-initiated bearer control solution. Figure D.3-4 shows the case where the bearer (PDP context) for the QoS class needs to be setup. Figure D.3-5 shows the case where the bearer (PDP context) for a given QoS class is already established, triggered by another service data flow using that QoS class.

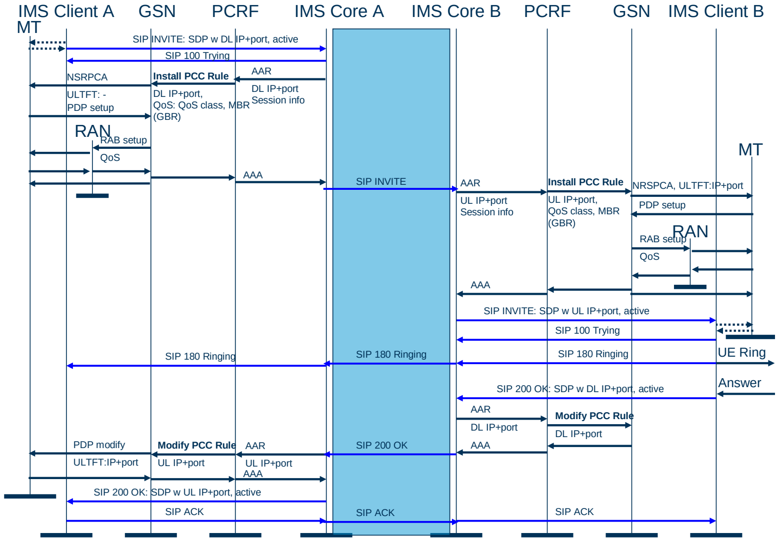

Figure D.3-4: IMS call setup– bearer needs to be setup on demand

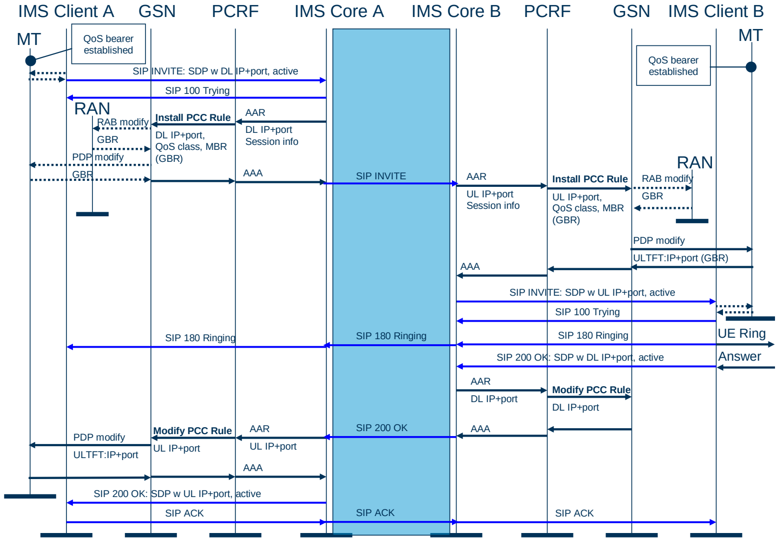

Figure D.3-5: IMS call setup - bearer for QoS class is already setup

- Optionally, the IMS client will check with the terminal before session setup, to check the resource status. (This would be the case of an IMS Client integrated in the UE, developed to work also for the UE-initiated mode.)
- Since "NW-init bearer control mode" is used, the application will be informed that resources are available. The application sets up a session, in the IMS-case with "preconditions=met" (and no "inactive" indication).
- The P-CSCF requests a policy check over Rx, forwarding session service information that is potentially useful for the PCRF for identifying QoS needs for this service:

  - This information may include information derived from the SDP about media bandwidth and type, as well as IMS communication service identifier, optionally an Application Reference and potentially other information relevant to identify the service QoS needs. Also the flow status is included for gating control.

  - The resource reservation already at the SDP Offer is a change compared to normal use cases for Rel6. It is needed to avoid ghost ringing (and depending on when SDP answer arrives, voice clipping), when this short SIP sequence (preconditions="met", indicated with current qos=desired qos) is used. Since the P-CSCF should be unaware of the bearer control initiation mode, it needs to trigger this check for any SDP Offer, where preconditions="met" is indicated.

- The PCRF maps the flows described on Rx, using the session service information from Rx as input to the PCC rules provisioned to it. Optionally, this is combined with subscription data from the SPR. The output is the bearer QoS needed, described by: QoS class identifier, UL+DL MBR and optionally UL+DL GBR.
(MBR is derived from the maximum rates of the SDP. GBR, if used, is derived according to operator policy – it could be set to MBR, or it could be lower.)

- The PCRF creates a PCC Rule including the bearer QoS for each service data flow. The PCC Rule also includes the DL filter to be installed in the GGSN. The PCC Rule is pushed to the GGSN.

- Depending on operator policy for the requested IMS service/media type, the PCRF may respond to the P-CSCF directly ("early Rx response"), to let the Session setup proceed in parallel with the PCC Rule installation, or withhold the response to the P-CSCF until a response from the GGSN is received ("late Rx response"), as shown here:

  - This is a trade-off between risk for ghost ringing (dependent also on network dimensioning) and setup delays.

- The GGSN analyses whether there is a PDP context with the given QoS class to this UE already:

  - If the requested QoS class is not established (case shown in Figure D.3-4):

    - the GGSN initiates a Network-initiated Bearer setup procedure according to Figure D.3-1 above. Since there is no SDP Answer yet with dynamic IP+port to be used for the uplink, the ULTFT is empty.

    - the GGSN installs the PCC Rule with normal PCC procedures (filters, charging key, gate status) and responds to the PCRF.

  - If the requested QoS class is already established (case in Figure D.3-5):

    - if a GBR value is included in the PCC Rule, the GGSN should trigger admission control on this QoS class. The GGSN adds the GBR of the requested PCC Rule to the aggregate GBR already admitted on this QoS class (PDP context), and triggers a GGSN-initiated PDP context modification of the GBR according to Figure D.3-2 above.

    - the GGSN installs the PCC Rule with normal PCC procedures (filters, charging key, gate status) and responds to the PCRF.

- The PCRF responds successfully to the P-CSCF.

- The SIP INVITE is propagated via the IMS network to the terminating P-CSCF

- The terminating P-CSCF triggers a policy check to the PCRF, including service session information, and flow status for gating control.

- As on the originating side, the operator can control the PCRF to respond directly to the P-CSCF (to let the session setup proceed), or to wait for the successful PCC Rule installation (as shown here).

- The PCRF maps the service data flows to bearer QoS according to the terminating operators policies.

- The PCRF creates a dynamic PCC Rule and requests the GGSN to install it. This includes the QoS class identifier, MBR, optionally the GBR and the UL filter.

- As on the originating side, the GGSN checks whether a PDP context for this QoS class already is setup:

  - If the requested QoS class is not established (case shown in Figure D.3-4):

    - the GGSN initiates a Network-initiated Bearer setup procedure according to Figure D.3-1 above. The ULTFT to be installed in the UE for this PDP context is also included, and contains the UL IP+port as described in the SDP offer.

    - the GGSN installs the PCC Rule with normal PCC procedures (filters, charging key, gate status) and responds to the PCRF.

  - If the requested QoS class is already established (case in Figure D.3-5):

    - if a GBR value is included in the PCC Rule, the GGSN should trigger admission control on this QoS class. The GGSN adds the GBR of the requested PCC Rule to the aggregate GBR already admitted on this QoS class (PDP context).

    - the GGSN triggers a Network-Initiated PDP context modification, according to Figure D.3-2 above, including the ULTFT and optionally the GBR.

    - the GGSN installs the PCC Rule with normal PCC procedures (filters, charging key, gate status) and responds to the PCRF.

- The PCRF responds successfully to the P-CSCF.

- The SIP INVITE is forwarded to the terminating IMS client.

- The terminating IMS client may optionally check resource status from the terminal, and is informed that resources are ensured (since NW-init QoS mode is used).

- The terminating IMS client can start ringing, and sends a SIP 180 Ringing message back to the A-client.

- The answer of the user triggers a SIP 200 OK from the terminating client to the originating side. This includes the SDP answer, including the IP+port to be used in the direction towards the terminating client.
  (Note: Here, the SDP answer is shown to be included in the SIP 200 OK. Another possibility is that SDP answer is sent already in SIP 180 Ringing, in which case the related Rx/Gx interactions are triggered at that message. There may however still be Rx/Gx interactions for gating control at the SIP 200 OK.)

- The P-CSCF triggers a policy check with the updated service session information, and including flow status for gating control.

- The PCRF updates the PCC Rule with the DL filter information in the GGSN, and responds to the P-CSCF.

- The SIP 200 OK is forwarded to the originating P-CSCF, which performs a policy check, and includes the SDP answer, and flow status for gating control.

- The PCRF updates the PCC Rule with the UL filter for this flow. The PCRF modifies the PCC Rule over Gx.

- The GGSN updates the PCC rule with the given UL filter. The GGSN triggers a Network-Initiated PDP context modification according to Figure D.3-2 above (the ULTFT modify case). Then the GGSN responds to the PCRF.

- The PCRF responds to the P-CSCF.

- The P-CSCF forwards the SIP 200 OK to the originating IMS client.

- The media can now flow in both directions. The ULTFT in each terminal maps the flow to the PDP context decided by the policy of the operator of that user.

D.3.2.3.2 Successful call setup: B-side SDP modification case

If the terminating client down-negotiates the SDP, then resources reserved with GBR admission control earlier may need to be released. This is done within the policy checking procedures anyway triggered by the SDP Answer on both the terminating and originating side:

- The PCRF determines a lower GBR based on updated session information, includes this in the PCC Rule modification to the GGSN.

- On the terminating side, the GGSN triggers a PDP context modification with the lower GBR to the RAN, via the SGSN.

- On the originating side, the GGSN includes the modified GBR in the PDP context modification that installs the updated ULTFT in the UE.

The principle followed is thus one roundtrip of SIP/SDP signaling, with resource reservation in each access on SDP offer, and potentially with changes on SDP answer. This is in contrast to the principle when using preconditions="not met" or "media=inactive" in the SIP/SDP signaling, when there are two roundtrips of SIP/SDP signaling needed, before and after the resource reservation in the access, to complete the session setup.

D.3.2.3.3 Unsuccessful call setup: No resources on A-side

If the originating RAN rejects the requested GBR for one or more media flows, this reject is propagated to the PCRF. There are two main cases how this then is handled.

Late Rx response case:

- The PCRF has not yet responded to the P-CSCF, and will therefore indicate rejection of the one or more media flows.
- The P-CSCF can have different policies for handling the situation:
  - A basic case would be to reject the SIP INVITE/SDP Offer to the IMS Client, and indicate a temporary failure due to resource reasons, to allow a retry. The SIP client would then retry, potentially with lower bitrates and/or omitting some media types.
  - More advanced cases are FFS, e.g. the P-CSCF disables the rejected media flows by setting the port numbers to zero.

Early Rx response case:

- The PCRF has already responded to the P-CSCF, and the SIP session setup has progressed towards the terminating side. (This case would occur if the PCRF knows that the PCC Rule installation will take long time).
- The P-CSCF will reject the SDP offer according to the above case. In addition, it needs to signal cancellation of the session towards the terminating network.
  - More advanced cases are FFS, e.g. the P-CSCF sends an updated SDP offer disabling the rejected media flows by setting the port numbers to zero.
- In this case, there is a risk for ghost ringing (if ringing starts before cancellation reaches B-client).

D.3.2.3.4 Unsuccessful call setup: No resources on B-side

If the terminating RAN reject the requested GBR for one or more media flows, this reject is propagated to the PCRF. As above, there are two cases for this.

Late Rx response case:

- The PCRF responds with a reject to the P-CSCF.
- The terminating P-CSCF can act as on the originating side, i.e:
  - Basic case: reject SIP / SDP Offer towards originating network, with temporary failure indicating resource limitation. This rejection is propagated to the originating client, which may retry with lower resource requirements.
  - Advanced cases (FFS), e.g. setting port number = 0 on rejected media flows.

Early Rx response case:

- The P-CSCF needs in addition to cancel the ongoing session towards the B-client, with the risk of ghost ringing.
  - More advanced cases are FFS, e.g. sending an updated SDP offer and setting port number = 0 on rejected media flows.

D.3.2.3.5 Call clearing and bearer termination

Since the purpose with Network-initiated bearer control is to provide complete control on bearer QoS from the network side, also QoS-related release of PDP contexts and resources should be initiated from the network. This requires no additional procedures, but the specifications should state that the UE, in case of NW-init mode, should not itself release PDP contexts at e.g. the end of the SIP session. (Note: for other reasons, e.g. at detach, the UE should of course release PDP context(s)):

- The clients send SIP BYE messages, triggering release of the Rx session to the PCRF.

- If the policy is to release unused PDP contexts, the PCRF will signal removal of the PCC Rules to the GGSN. The GGSN will check if no other PCC Rule is using that QoS class, and if not, the GGSN will initiate release of that PDP context.

D.3.2.4 Service setup phase, RTSP streaming (Rx control)

A use case for RTSP-based streaming can be outlined, using a similar solution as above. In this case, the streaming server or streaming proxy will act as the AF towards the PCRF.

Annex E: Change history

Change history							
Date	TSG #	TSG Doc.	CR	Rev	Subject/Comment	Old	New
2005-10					Skeleton	-	v.0.0.0
2005-11	SA2#49				Incorporates agreements from S2-052987; S2-052564; S2-052990; S2-052991	v.0.0.0	v.0.1.0
2005-12	SA2#49				Editorial corrections and inclusion of TS number	v.0.1.0	v.0.1.1
2006-01	SA2#50				Incorporates agreements from S2-060111; S2-060471; S2-060472; S2-060508; S2-060509; S2-060510	v.0.1.1	v.0.2.0
2006-03	SA2#51				Incorporates agreements from S2-061133, S2-061131, S2-061189	v.0.2.0	v.0.3.0
2006-06	SA2#52				Incorporates agreements from S2-061409, S2-061970, S2-061771, S2-061772, S2-061774, S2-061896, S2-061966	v.0.3.0	v.0.4.0
2006-07	SA2#53				Incorporates agreements from S2-062023, S2-062438, S2-062575, S2-062437, S2-062571 and S2-062509	v.0.4.0	v.0.5.0
2006-09	SA2#54				Incorporates agreements from S2-063352, S2-063354, S2-063355	v.0.5.0	v.0.6.0
2006-09	SA2#54				Incorporates agreements from S2-063458	v.0.6.0	v.0.7.0
2006-10	SA2#54				Correct heading numbers in section 12	v.0.7.0	v.0.7.1
2006-11	SA2#55				Incorporates S2-063963; S2-063964; S2-063965; S2-063967; S2-063968	v.0.7.1	v.0.8.0
2007-01	SA2#56				Incorporates S2-070405; S2-070417; S2-07422; S2-070423; S2-070425; S2-070619; S2-070627; S2-070628	v.0.8.0	v.0.9.0
2007-02	SA2#56b				Incorporates S2-070870; S2-070871; S2-070873; S2-070992	v.0.9.0	v.0.a.0