# 3GPP TR 23.705 V0.7.0 (2013-08)

*Technical Report*

## 3rd Generation Partnership Project;
## Technical Specification Group Services and System Aspects;
## System Enhancements for User Plane Congestion Management
## (Release 12)

The present document has been developed within the 3rd Generation Partnership Project (3GPP TM) and may be further elaborated for the purposes of 3GPP.
The present document has not been subject to any approval process by the 3GPP Organizational Partners and shall not be implemented.
This Report is provided for future development work within 3GPP only. The Organizational Partners accept no liability for any use of this Specification.
Specifications and Reports for implementation of the 3GPP TM system should be obtained via the 3GPP Organizational Partners' Publications Offices.

Keywords

<keyword[, keyword]>

*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

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.

# 1 Scope

The objective of this Technical Report is to study and define system enhancements for user plane congestion management based on the stage-1 normative requirements defined in 3GPP TS 22.101 [3] for User Plane congestion management.

Based on the technical analysis, any needed enhancements/updates to 3GPP functions and interfaces will be identified.

Normative specifications will be developed based on the conclusions of this Technical Report.

# 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 TR 21.905: "Vocabulary for 3GPP Specifications".

[2] 3GPP TR 41.001: "GSM Release specifications".

[3] 3GPP TS 22.101: "Service principles".

[4] 3GPP TS 23.060: "General Packet Radio Service (GPRS); Service description; Stage 2".

[5] 3GPP TR 23.800: "Study on Application Based Charging; Stage 2".

[6] 3GPP TS 24.312: "Access Network Discovery and Selection Function (ANDSF) Management Object (MO)".

[7] 3GPP TS 29.212: "Policy and Charging Control (PCC); Reference points".

[8] 3GPP TS 23.401: "General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access".

[9] 3GPP TR 22.805: "Feasibility study on user plane congestion management".

# 3 Definitions and abbreviations

## 3.1 Definitions

For the purposes of the present document, the terms and definitions given in TR 21.905 [1] and the following apply. A term defined in the present document takes precedence over the definition of the same term, if any, in TR 21.905 [1].

**RAN user plane congestion:** RAN user plane congestion occurs when the demand for RAN resources exceeds the available RAN capacity to deliver the user data for a period of time. RAN user plane congestion leads, for example, to packet drops or delays, and may or may not result in degraded end-user experience.

NOTE 1: Short-duration traffic bursts is a normal condition at any traffic load level, and is not considered to be RAN user plane congestion. Likewise, a high-level of utilization of RAN resources (based on operator configuration) is considered a normal mode of operation and might not be RAN user plane congestion.

NOTE 2: RAN user plane congestion includes user plane congestion that occurs over the air interface (e.g. LTE-Uu), in the radio node (e.g. eNB) and/or over the backhaul interface between RAN and CN (e.g. S1-u).

**User-impacting congestion:** User-impacting congestion occurs when a service that is delivered to a user over the default bearer or a dedicated bearer does not meet the user's expected service experience due to RAN user plane congestion. The expectation for a service delivery is highly dependent on the particular service or application. The expected service experience may also differ between subscriber groups (e.g. a premium subscriber may have higher expectations than a subscriber with the cheapest subscription). RAN resource shortage where the RAN can still fulfil the user expectations for a service delivery is not considered to be user-impacting congestion; it is rather an indication for full RAN resource utilization, and as such a normal mode of operation.

NOTE 3: It is up to the operator to determine when a service satisfies the user's expected service experience.

**Unattended traffic:** see definition for "Unattended Data Traffic in [3]. See also the discussion in clause 4.9.1 in [9].

**Attended traffic:** see definition for "Attended Data Traffic in [3]. See also the discussion in clause 4.9.1 in [9].

## 3.3 Abbreviations

For the purposes of the present document, the abbreviations given in TR 21.905 [1] and the following apply. An abbreviation defined in the present document takes precedence over the definition of the same abbreviation, if any, in TR 21.905 [1].

# 4 Assumptions and Architectural Requirements

## 4.1 Assumptions

Editor's Note: This clause will define the underlying assumptions of the work.

## 4.2 Architectural Requirements

Editor's Note: This clause will define the architectural requirements based on the normative stage-1 requirements defined in TS 22.101.

# 5 Key Issues

Editor's Note: For each key issue identified, the clause will capture the "General description and assumptions" (sub-clause 1). Different architecture solutions to address the key issues will be documented in Clause 6.

Editor's Note: The key issues defined in this clause are intended to help the architecture solution definition (e.g. by providing some guidelines for the solution descriptions). It is not expected that all the key issues defined here are relevant for all solutions. Solutions defined in Clause 6 shall clearly define which of the key issues they cover and address.

## 5.1 Key Issue #1: RAN User Plane congestion mitigation

### 5.1.1 General description and assumptions

The majority of mobile data traffic (e.g. Internet or over-the-top services traffic) is currently delivered over the default bearers. This key issue addresses aspects how the system can effectively mitigate RAN user plane congestion in order to overcome the negative impact on the perceived service quality for such data traffic.

The congestion mitigation measures include traffic prioritization, traffic reduction and limitation of traffic, and shall be able to manage user plane traffic across a range of variables including the user's subscription, the type of application, and the type of content.

A key challenge for congestion mitigation is to support subscribers with different service requirements (e.g. premium, flat rate or roaming users) and application traffic with different traffic characteristics (e.g. long-lived and short-lived traffic flows) without increasing the system-wide signalling overhead significantly.

The following aspects should be considered by a solution addressing this key issue:

- The type of congestion mitigation measures, i.e. QoS/QoE control/adjustment through traffic prioritization, traffic reduction or traffic limitation based on the congestion status.

- The location of congestion mitigation measures (e.g. in UE, in RAN, in Core, in both, or in connected IP networks such as IMS or Packet-switched Streaming Service).

- The criteria to decide which flows will be subject of traffic mitigation measures (e.g. the user's subscription class, the type of application or the type of content).

- The information that are needed to effectively enforce the mitigation measure (e.g. the RAN congestion status, the impacted users, the type of traffic – e.g. attended vs. unattended) and how this information could be obtained.

NOTE: Depending on the congestion mitigation measure and enforcement point, different information is needed.

- The way operators are able to control congestion mitigation through policies.

## 5.2 Key Issue #2: RAN User Plane congestion awareness

### 5.2.1 General description and assumptions

NOTE 1: This key issue does not exclude any solution proposal; solution proposals that do not require any form of RAN user plane congestion awareness do not need to address this key issue.

NOTE 2: Congestion awareness means awareness of congestion onset, continuance and abatement.

In order to address RAN user plane congestion, the following system capabilities are required according to TS 22.101 [3]:

- allow the network "to adjust the QoS of existing connections/flows and apply relevant policies to new connections/flows depending on the RAN user plane congestion status and the subscriber's profile";

- allow the network "to reduce the user plane traffic load (e.g. by compressing images or by adaptation for streaming applications)" based on RAN congestion status and according to operator policies; and

- allow the network "to limit traffic from operator-controlled and/or third-party services based on RAN user plane congestion status for a UE".

Editor's Note: It is FFS how to derive architecture requirements from this system level requirements.

To support these system capabilities, some network elements outside the RAN may need to become aware of the congestion status.

The following aspects should be considered by solutions that propose some form of RAN congestion awareness:

- Where in the network is awareness of RAN user plane congestion required?

- What information on the congestion (e.g. severity of congestion, etc.) is required to enforce appropriate mitigation measures?

- Which level of granularity for congestion awareness is required?

- In case the congestion status needs to be reported from the RAN towards other system entities:

  - What is congestion and how is it detected?

  - How often and when does the congestion status need to be indicated?

NOTE 3: Short-term congestion should not be indicated.

  - What information needs to be indicated (e.g. severity of congestion or cell information), also taking into account the balance between signalling/processing overhead and benefits (e.g. preciseness)?

  - How is the congestion status be indicated, i.e. in the user plane or in the control plane) and over which interfaces?

# 5.3 Key Issue #3: Differentiated treatment for non-deducible service data flows in case of RAN user plane congestion

## 5.3.1 General description and assumptions

A very common way of dealing with RAN user plane congestion is to throttle certain customers and/or application data flows to preserve higher priority traffic. This requires the ability to enforce per subscriber and/or per application QoS policies.

To some extent the current 3GPP QoS architecture already supports this feature. To that purpose a combination of the following mechanisms can be used:

- Different QCI values, with different Priority levels, can be allocated to the bearers (in particular the default bearer) opened by different classes of subscribers. As an example the operator could use QCI 8 for the default bearer of a "premium" subscriber and QCI 9 for the default bearer of a "basic" subscriber.

- Different applications, or different data flows exchanged by a specific application (e.g. video, audio, file sharing and chat), can be mapped to different bearers. As an example, for a specific class of subscribers, or for any subscriber, the operator could map Internet applications like browsing, ftp and peer-to-peer file sharing to the default bearer, and use dedicated bearers with higher priority for data flows, like for example media streaming, that would benefit of preferential treatment in case of congestion in RAN.

With this approach differentiated treatment for specific applications, or application data flows, in case of RAN user plane congestion can be achieved if such applications, or application data flows, can be mapped to separate bearers; unfortunately this is problematic for applications exchanging data flows for which Service Data Flow (SDF) templates cannot be deduced. Non-deducible SDFs cannot be described by SDF templates or can be described by SDF templates but these SDF templates cannot be applied to unambiguously or efficiently control the application traffic. Applications with non-deducible SDFs are for example those using (potentially many) very short-lived parallel UDP and/or TCP data flows, for which service data flow filters detected via ADC (Application Detection and Control) rules are too short-lived to allow PCC system to control them using SDF templates. Other examples can be found in section 5.1 of 3GPP TR 23.800 [5].

Based on current specifications, for applications with non-deducible SDFs mapping different applications, or application data flows, to different bearers to achieve traffic handling differentiation is possible using predefined PCC rules provisioned into the PCEF, but this approach has the following limitations:

- It only works in the downlink direction.

- It requires application detection to be performed by the PCEF. Deployment scenarios where application detection is performed by a TDF are not supported.

- Roaming scenarios with local-breakout are not supported.

The target of this key issue is to study possible solutions to achieve differentiated treatment in case of congestion in RAN for applications, or application data flows, with non-deducible SDFs. Solutions addressing this key issue should allow for traffic handling differentiation in both uplink and downlink direction and should support scenarios with TDF as well as roaming with local-breakout.

NOTE 1: What is the feasible level of granularity for traffic handling differentiation depends on the application and the transport layer on which the application is layered. For example differentiating the treatment of individual application data flows is not feasible for the applications that multiplex multiple data flows over a single TCP connection, because slowing down or dropping segments for one of the data flows would cause head-of-line blocking for all other data flows sharing the same TCP connection.

NOTE 2: Whether there are use cases of operator's interest requiring support for differentiated treatment of application data flows multiplexed over a single TCP or UDP flow is to be determined.

## 5.4 Key Issue #4: Video delivery control for congestion mitigation

### 5.4.1 General description and assumptions

Mobile network operators identify mobile video as one of the main contributing factors to congestion in mobile networks.

Reducing the rate of video applications during congestion periods is a very effective congestion mitigation measure and can reduce the traffic load in a congested RAN significantly. It should be noted that various approaches exist to reduce video flow rates in the network today, ranging from simple bandwidth limitation or scheduling for adaptive video applications (e.g. DASH) to explicit rate adaptation using CDN, video transcoding or change of manifest file(s) for adaptive streaming protocols. The most appropriate approach depends on the precise video application (e.g. adaptive versus non-adaptive video codecs) and transport protocol (e.g. TCP vs. UDP).

The 3GPP community continues to support the existing end-to-end adaptive bitrate video streaming technologies, specifically 3GP-DASH defined by 3GPP and also adopted by MPEG.

Since the user's service experience depends a lot on the video flow rate (e.g. low rates result typically in a poor service experience), it is important that the operator can control according to the subscription level what delivery rate it provides for a particular user under a certain load situation. For example, during a low congestion period, an operator may still want to offer its gold level subscribers a very good video service experience, whereas a certain reduction of the video quality is acceptable for silver and bronze level subscribers (e.g. the next lower video codec). However, when the congestion becomes more severe, the operator may also want to limit the video flow rate of its gold level subscribers somehow, while still maintaining a better video quality than for its silver and bronze level subscribers.

This key issue is about how the operator can manage (based on RAN, Core Network and/or application layer mechanisms) the delivery of individual video application flows, according to the user's subscription level and current RAN congestion level. Solutions for different video application types (adaptive and non-adaptive) and transport protocols (TCP and UDP) are considered.

NOTE 1: Interaction of potential solutions with existing end-to-end adaptation mechanisms (TCP, DASH etc.) should be documented.

NOTE 2: If different solutions for different video application types are adopted, the network shall be able to identify the type of traffic and the correct mitigation measure.

## 5.5 Key Issue #5: Uplink Traffic Prioritization

### 5.5.1 General description and assumptions

One key aspect of RAN congestion mitigation is the capability for the system to prioritize certain traffic. There are two types of prioritization:

1. Per-flow prioritization:

- It should be possible to identify, differentiate and prioritize uplink traffic from different applications in order to provide these applications with appropriate service quality during RAN user plane congestion.

2. Per-user prioritization:

- It should be possible to prioritize uplink traffic from different users based on subscription type, e.g., differentiate between traffic generated/received by gold users vs. normal users.

There are certain applications that generate much traffic in the uplink direction, like peer-to-peer applications, gaming, video conferencing, etc. Solutions should be considered for both uplink traffic and downlink traffic. If different solutions are used for UL and for DL, coexistence of the solutions should be evaluated. For instance, solutions could allow that a bi-directional data flow receives equal priority (e.g. high/low) in both uplink and downlink, particularly for the case when both directions are congested. Similar applies for per-user prioritization.

For uplink, techniques for per-user prioritization and per-flow prioritization may be performed in different entities. For instance, the eNB could perform per-user prioritization, since it is in charge of providing UL scheduling grants to each UE, while the UE may be involved in performing per-flow prioritization based on operator/NW instructions.

# 6 Solutions

Editor's Note: This clause is intended to document architecture solutions. Each solution should clearly describe which of the key issues it covers and how.

## 6.1 CN-based Solutions for RAN user plane congestion management

### 6.1.1 General architectural requirements

The following is the list of architectural requirements to address RAN user plane congestion by CN-based solutions:

1. The network shall support RAN user plane congestion information transfer from the RAN to the Core Network.

2. The solutions shall specify the RAN user plane congestion information sent to the Core Network.

3. The Core Network shall be able to use the RAN user plane congestion information in order to select and apply congestion mitigation measures for addressing the RAN user plane congestion.

NOTE: Usage of RAN user plane congestion information will be described as part of the CN-based solution's description, e.g., optimization over all flows/users in a cell.

4. The solutions shall address UE mobility aspects.

5. The solutions shall address roaming UEs.

6. The solutions should avoid additional overload in the network (e.g. signalling overload).

7. The solutions should document interaction aspects between RAN, CN and transport layer/application layer congestion mitigation measures, if applicable. Performance aspects (e.g., measurement averaging time) may be provided.

8. The solutions should document whether the mitigation measures are applicable for uplink and/or downlink traffic.

### 6.1.2 General description, assumptions and principles

This solution addresses key issues #1 and #2 on congestion mitigation and congestion awareness. If not indicated otherwise, the term "congestion" refers to "RAN user plane congestion". The solution is based on the following principles:

Congestion Detection:

P1) The RAN informs relevant CN function(s) about the RAN user plane congestion.

NOTE: The RAN implementation for predicting or detecting RAN user plane congestion is outside the scope of 3GPP.

Editor's Note: The semantics of the congestion notification of RAN user plane congestion is FFS.

Editor's Note: It is FFS how different levels of congestion can be derived.

Editor's Note: It is FFS whether per cell or per bearer granularity is used for congestion feedback.

P2) Congestion is indicated to the CN in order to enable CN function(s) to mitigate congestion (e.g. by enforcing mitigation measures that reduce/limit/block some traffic transmit to/from impacted users).

P3) The CN is made aware of which users are contributing to or are affected by the RAN user plane congestion.

P4) Congestion (abatement) should be indicated in a lightweight but timely way.

Congestion Mitigation:

P5) The user plane congestion management solution supports one or more of the required congestion mitigation schemes (i.e. traffic prioritization, limiting, gating and reduction on application and service-level) to allow flexible operator deployment based on their operational requirements.

P6) Decisions to apply congestion mitigation measures on user traffic may take into account operator policies and subscriber information.

P7) Congestion mitigation measures based on traffic prioritization, limiting and reduction are enforced in the CN. They may also be applied at the service level, based on operator policies. Congestion mitigation based on traffic prioritization may also be applied in the RAN in order to take into account real-time radio channel information. Congestion mitigation should not negatively impact the service experience of users who are not in a congested RAN area.

### 6.1.3 High-level operation and procedures

A high level view of operation and procedures of the proposed solution is shown in Figure 6.1.3-1.

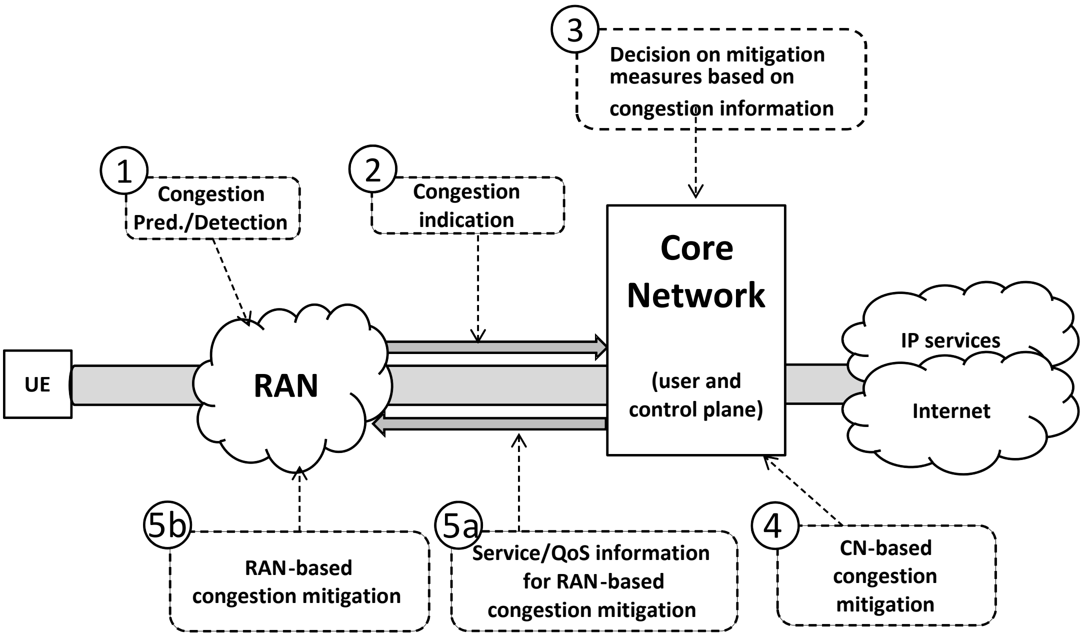
3
Decision on mitigation measures based on congestion information
1
Congestion Pred./Detection
2
Congestion indication
Core Network
(user and control plane)
UE
RAN
IP services
Internet
5b
RAN-based congestion mitigation
5a
Service/QoS information for RAN-based congestion mitigation
4
CN-based congestion mitigation

**Figure 6.1.3-1: User-plane Congestion Management – High-level View**

NOTE 1: The numbers do not necessarily imply a temporal order.

NOTE 2: Step 5a and 5b are optional for solutions that are based on a CN only approach.

1. Congestion prediction/detection based on actual resource shortage or predictive algorithms in the RAN (P1).
2. Congestion indication to the CN (P2, P3, P4).
3. Selection of mitigation measures (e.g. policy rule provisioning) (P5, P6).
4. CN-based congestion mitigation (e.g. traffic limitation, gating, compression) (P5, P7).
5. Measures for RAN-based congestion mitigation (P5, P7).

   a. Optional Service/QoS information to enable traffic differentiation in the RAN based on existing QoS measures.

   Editor`s note: It is FFS how RAN user plane congestion awareness can also be exploited to optimize the performance of potentially agreed RAN-based congestion mitigation solutions. For example, the congestion information could be used to enable packet classification required to mark downlink packets, in order to minimize the performance impacts on the GGSN/PGW or the TDF.

   b. Optional RAN-based congestion mitigation (e.g. traffic prioritization, scheduling).

## 6.1.4 RAN Congestion Detection Solutions

### 6.1.4.1 General description, assumptions, and principles

The following terms are introduced:

- The **congestion level**, which is derived in a RAN node based on RAN measurements.

- RAN user plane Congestion Information (**RCI**), which indicates the congestion level from RAN to the CN.

The congestion information should provide the CN with sufficient information to apply the appropriate congestion mitigation measures.

RAN user plane detection and reporting is based on the following principles:

P1) The complexity in the RAN should be low.

P2) RCIs indicate the level of RAN user plane congestion as a scalar value to the CN.

P3) Operators should be able to flexibly configure the detection parameters of the congestion levels indicated in the RCI.

This is achieved as following:

- The congestion level is detected in the RAN node. Congestion level should provide a meaningful metric for the severity of the congestion. The congestion level is derived based on operator configurations.

- The congestion level is indicated to the CN as a scalar value in the RCI.

The CN performs congestion mitigation by deciding which congestion mitigation measure is taken according to the current RCI (e.g. by activating a policy for congestion mitigation according to the reported RCI).

### 6.1.4.2 High-level operation and procedures

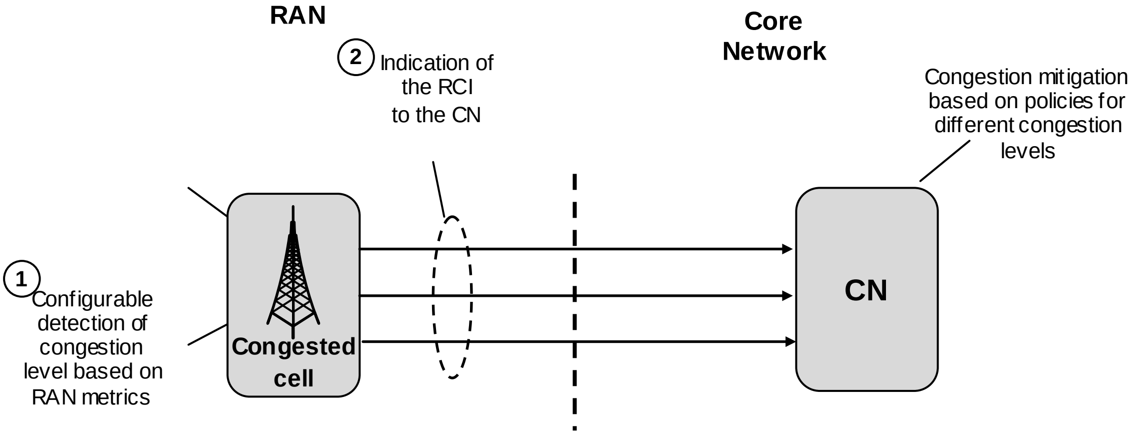
RAN
2 Indication of the RCI to the CN
Core Network
Congestion mitigation based on policies for different congestion levels
1 Configurable detection of congestion level based on RAN metrics
Congested cell
CN

**Figure 6.1.4.2-1: High-level operational principle of RAN congestion detection and reporting**

The high-level operation steps are as following:

1. The RAN detects the congestion level, based on monitoring of RAN resources and related metrics. Averaging over time and/or over bearer/UE-specific metrics should be applied in order to derive a stable expression of congestion. The congestion level is determined based on operator configurations.

2. The RCI is reported to the CN as a scalar value. How this information is sent, and whether RCIs are reported per bearer or cell is not part of this solution.

NOTE: The mobile operator configures the policies for congestion mitigation in the CN in such a way that it reacts appropriately to the RCI, i.e. by activating a policy for congestion mitigation according to the received RCI. In the operator's network, both RAN and CN should have a consistent interpretation of RCI values.

## 6.1.5 RAN Congestion Reporting Solutions

### 6.1.5.1 Solution 1.5.1: RAN User Plane congestion reporting by GTP-U extension

#### 6.1.5.1.1 General description, assumptions, and principles

The RAN nodes include the RAN Congestion Information (RCI) in a GTP-U header extension of the uplink packet to convey the RAN user plane congestion information to the CN GWs such as GGSN/PGW.

At minimum, the RCI comprises of:

- The RAN user plane congestion notification.

- The location of the congested RAN, such as the CELL ID, may also be included in the extension.

Editor's Note: Whether the Cell ID and what additional information is required in RCI is FFS.

The user plane core network nodes such as the GGSN/PGW will inspect the GTP-U header and obtain the congestion information. Therefore, the GGSN/PGW node will know which of the served users/bearers are affected by the congestion.

Editor's Note: How to deliver the RCI within the CN with PMIP-based S5/S8 is FFS.

The congestion is detected based on the monitoring of the RAN network elements. Once the congestion is detected, the RCI is included in all the uplink GTP-U packets.

NOTE: In case where there is no uplink traffic, then the current RCI is indicated to the CN once the next uplink packet is sent.

For the home routed roaming case, it should be possible to configure the VPLMN so that the RCI is not reported from VPLMN to HPLMN.

Editor's Note: Whether in case of home routed roaming it is sufficient to disable reporting of RCI for all HPLMNs or whether it is required to enable/disable RCI reporting for specific HPLMNs and how to achieve this is FFS.

Editor's Note: Whether and how the CN passes RCI to other network elements (e.g. PCRF, OCS, TDF, AF) is FFS.

In RAN sharing scenario, the RAN nodes decide whether CN entities require RCI in GTP-U header or not based on per PLMN configuration. Moreover, the RAN nodes need to generate the congestion information in consideration of RAN sharing configuration.

The CN performs congestion mitigation measures based on received RCI.

Editor's Note: Depending on which other network elements receive RCI (or a subset of RCI), those nodes may perform additional mitigation actions, which are FFS.

#### 6.1.5.1.2 High-level operation and procedures

The solution procedures are the following (see Figure 6.1.1.5.1.2-1):

1) The congestion indicator is reflected in the uplink data traffic packet. The packet header is included with the RCI (RAN Congestion Information) which includes the level of congestion and potentially also the location information (e.g. Cell ID)

2) The GGSN/PGW investigates the GTP-U header and obtains the congestion information.

3) GGSN/PGW may report the congestion to other network nodes.

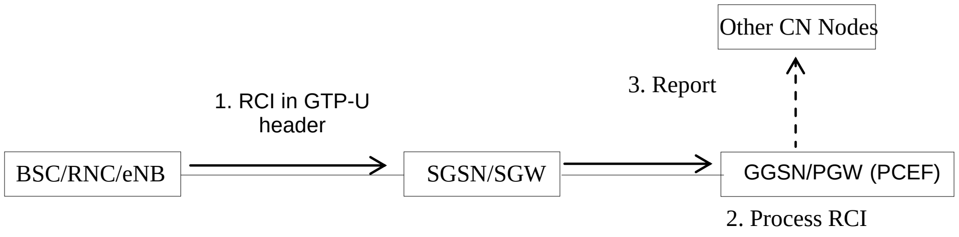

**Figure 6.1.5.1.2-1: User-plane Congestion Management – High-level View**

### 6.1.5.1.3 Event reporting on Gx

In order to enable dynamic policy control for user plane congestion management as described in next subclause, the reporting step 3 is assumed to be done by an extension of the PCC event reporting mechanism on Gx. The following definition is used:

**User plane congestion event report:** A notification provided by PCEF to PCRF indicating the occurrence/change of user plane congestion; it contains at minimum the level of congestion and may contain information about the scope.

The following assumption is taken:

- The PCRF shall be able to subscribe to congestion event reports based on severity levels.

Editor’s note: equivalent functionality for PMIP is FFS.

### 6.1.5.1.4 Policy control of congestion mitigation

The following behaviour is foreseen:

- As long as PCEF has an activated congestion mitigation policy available, it should apply a mitigation measure with matching congestion level on affected traffic
- The enhancement of congestion mitigation handling with pre-provisioned congestion mitigation policies in PCEF can be done as exemplarily shown in figure 6.1.5.1.4-1.

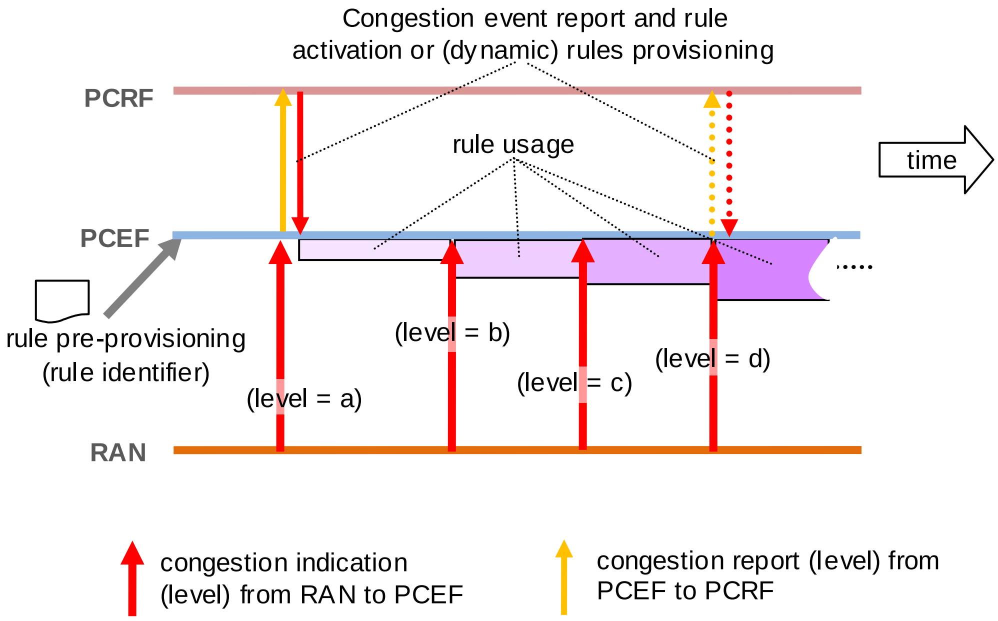
Congestion event report and rule activation or (dynamic) rules provisioning
PCRF
rule usage
time
PCEF
.....
rule pre-provisioning (rule identifier)
(level = a)
(level = b)
(level = c)
(level = d)
RAN
congestion indication (level) from RAN to PCEF
congestion report (level) from PCEF to PCRF

**Figure 6.1.5.1.4-1: possible behaviour of pre-provisioned congestion mitigation policies in PCEF (in combination with dynamic policy handling)**

Editor's note: it is FFS if another behaviour with pre-provisioned user plane congestion mitigation policies is required.

With the behaviour in figure 6.1.5.1.4-1 PCRF will always be in control of which congestion mitigation policies are active in PCEF. Furthermore, PCRF is always able to receive all congestion reports of interest for its policy decisions. In case that PCRF chooses not to subscribe to all congestion reports (for optimisation reasons), it may not always be aware of the currently enforced congestion mitigation policy.

#### 6.1.5.1.5 Impact on existing entities and interfaces

The RAN nodes (BSC/RNC/eNodeB)

- Include RCI defined in this solution in the uplink packet.

The core network user plan elements (GGSN/PGW)

- Recognize the congestion indicator.

#### 6.1.5.1.6 Solution evaluation

### 6.1.5.2 Solution 1.5.2: C-plane Signalling for RAN user plane congestion reporting

#### 6.1.5.2.1 Clarification of terminologies

*RAN user plane Congestion Information (RCI):* This is the information about RAN user plane congestion, e.g., RAN user plane congestion level, RAN user plane congested direction (radio uplink/downlink).

*RCI signalling:* The signalling is used as the means for conveying RCI from RAN to CN. The signalling can be done on a per EPS bearer basis or in an aggregate way as described below.

- *EPS bearer level RCI signalling:* RCI will be conveyed from the RAN to the CN for each EPS bearer. For instance, if RCI is specified on a cell level basis, a signalling message will be sent per EPS bearer even if all messages include the same RCI. The number of signalling messages is equal to the number of EPS bearers that are being served by the same cell.

- *Aggregating RCI signalling:* A single signalling message contains the RCI for multiple EPS bearers belonging to the same UE or even the RCI for EPS bearers of multiple UEs that are served by the same cell.

NOTE 1: The details of "aggregating RCI signalling" are described in clause 6.1.5.2.3.2.

#### 6.1.5.2.2 General description, assumptions, and principles

This solution addresses the key issue on "RAN User Plane congestion awareness".

This solution provides an aggregating RCI signaling mechanism for the RAN to report the RAN user plane congestion information to the CN by using:

- Existing C-plane signalling interfaces: S1-MME, S11, S5/S8, Gx, Rx, and Sd; and

- Existing C-plane signalling protocols: S1-AP, GTP-C and DIAMETER.

When the eNodeB is congested, the eNodeB sends the RAN user plane congestion information to the PCRF via the MME, the SGW and the PGW. The PCRF then decides whether to initiate the IP-CAN Session Modification procedure in order to assist the RAN to mitigate the RAN user plane congestion situation. In addition the PCRF decides whether to forward congestion information to the AF and TDF.

Depending on the operator's congestion mitigation policy, it may not be necessary to have "RCI signalling" for all EPS bearers. An operator shall be able to specify policy for RCI signalling for individual EPS bearers, e.g., activating or deactivating the RCI signalling for the EPS bearer. According to the policy for RCI signalling, the eNodeB sends the RCI to the PCRF only for those EPS bearers that have "RCI signalling" activated.

NOTE 1: Policy for RCI signalling is not used to configure eNodeB to send either an EPS bearer level RCI signalling or an aggregating RCI signalling. Policy for RCI is used to activate or deactivate the RCI signalling for the EPS bearer. Choosing which EPS bearer to be activated for RCI signalling is out of scope of solution, since it is operator and vendor specific.

Policy for "RCI signalling" can be configured either statically or dynamically.

- Static configuration: The policy for "RCI signalling" is pre-defined and stored in advance at the eNodeB and the MME, for example, via the OAM plane or manually configured when deploying the eNodeB and the MME.

- Dynamic configuration: The policy for "RCI signalling" is decided by the PCRF and can be updated dynamically. The policy for "RCI signalling" shall be included in the EPS bearer context information.

In this solution, only the dynamic configuration for EPS bearer level RCI signalling is discussed, since static configuration for EPS bearer level RCI signalling is not necessarily to be standardized.

The signalling for the RAN user plane congestion information from the eNodeB towards the PCRF shall be done on a per EPS bearer basis. For congestion event reporting on Gx and policy control of congestion mitigation considerations in subclauses 6.1.5.1.3 and 6.1.5.1.4 apply.

The policy for "RCI signalling" shall include "Reporting action for RCI signalling (e.g., start, stop)". To further reduce RCI signalling messages and to avoid unnecessary RCI signalling messages that may not lead to any decision at the PCEF/PCRF, the policy for "RCI signalling" may include conditions that trigger the eNodeB to send a RCI signalling message:

- Minimal bit rate of incoming traffic carried over the EPS bearer: Operator may decide not to apply a congestion mitigation measure for the EPS bearer that carries little amount of traffic (e.g., chatting), and thus no RCI signalling for such EPS bearer.

- Minimal congestion level that an operator is interested in for the given EPS bearer.

NOTE 2: In case, there is no policy for RCI signalling available at the eNodeB, it behaves according to operator's configuration.

The RCI should include:

- Congestion level;

- User identity (e.g., eNB UE S1AP ID and MME UE S1AP ID on S1-MME interface, IMSI on S11, S5/S8 and Gx interfaces );

- EPS bearer ID;

- Direction of user plane congested direction (e.g., radio uplink, radio downlink);

- Optionally user location information (e.g., Cell ID);

NOTE 3: How the congestion level is specified is out of scope of the solution description.

### 6.1.5.2.3 High-level operation and procedures

The solution consists of two procedures:

- Procedure for dynamic configuration of policy for RCI signalling from the PCRF to the eNodeB

- Procedure for "RCI signalling" from the eNodeB to the PCRF

#### 6.1.5.2.3.1 Procedure for dynamic configuration of policy for "RCI signalling"

The following procedures specified in TS 23.401 are used to convey the policy for "RCI signalling" from the PCRF to the eNodeB.

- Procedure for E-UTRAN Initial Attach (subclause 5.3.2.1): In step 14 to 17, the policy for "EPS bearer level RCI signalling" is included in the EPS bearer context information.
- Procedure for Dedicated bearer activation (subclause 5.4.1): In step 1 to 4, the policy for "EPS bearer level RCI signalling" is included in the EPS bearer context information.
- Procedure for PDN GW initiated bearer modification with bearer QoS update (subclause 5.4.2.1): In step 1 to 4, the policy for "EPS bearer level RCI signalling" is included in the EPS bearer context information.
- Procedure for UE requested PDN connectivity (subclause 5.10.2): In step 4 to 7, the policy for "EPS bearer level RCI signalling" is included in the EPS bearer context information.

The procedures mentioned above are mainly used for conveying policy for RCI signalling when establishing a new EPS bearer or when modifying QoS parameters of existing EPS bearers.

If an operator decides to only update the policy for "RCI signalling" of existing EPS bearer, e.g., from deactivating to activating the RCI signalling, or vice versa, a new procedure as shown in the Figure 6.1.5.2.3.1-1 is needed.

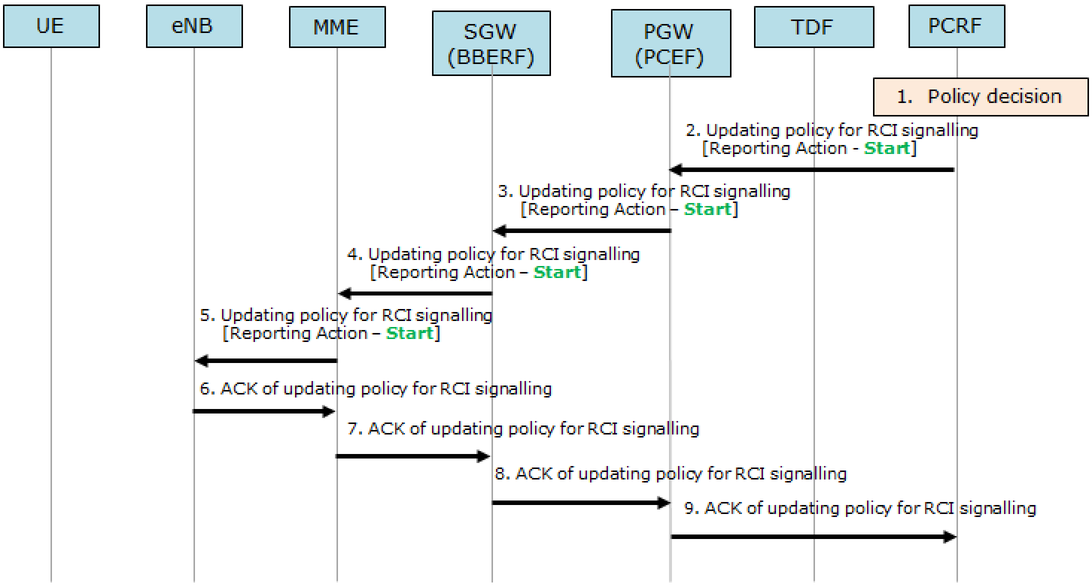
UE
eNB
MME
SGW (BBERF)
PGW (PCEF)
TDF
PCRF
1. Policy decision
2. Updating policy for RCI signalling [Reporting Action - Start]
3. Updating policy for RCI signalling [Reporting Action - Start]
4. Updating policy for RCI signalling [Reporting Action - Start]
5. Updating policy for RCI signalling [Reporting Action - Start]
6. ACK of updating policy for RCI signalling
7. ACK of updating policy for RCI signalling
8. ACK of updating policy for RCI signalling
9. ACK of updating policy for RCI signalling

**Figure 6.1.5.2.3.1-1: Updating policy for "RCI signalling"**

1) Based on the operator's policy, the PCRF decides to activate the "RCI signalling".

2) The PCRF sends PCC rules that apply to the given bearer with the policy for "RCI signalling" to the PGW.

3) The PGW forwards the EPS bearer context information with the policy for "RCI signalling" to the SGW.

4) The SGW forwards the EPS bearer context information with the policy for "RCI signalling" to the MME. The MME stores the policy.

5) The MME forwards the EPS bearer context information with the policy for "RCI signalling" to the eNodeB. The eNodeB stores the policy.

6) The eNodeB acknowledges the policy for RCI signalling to MME.

7) The MME acknowledges the policy for RCI signalling to SGW.

8) The SGW acknowledges the policy for RCI signalling to PGW.

9) The PGW acknowledges the policy for RCI signalling to PCRF.

Figure 6.1.5.2.3.1-1 depicts an example for activating RCI signalling. For deactivating the "RCI signalling" of the EPS bearer, the same procedure as described in Figure 6.1.5.2.3.1-1 is applied. The only difference is that the "Reporting Action" for the policy for "RCI signalling" is to be set to "Stop" instead.

For the case of intra E-UTRAN handover, the same procedures as specified in sub-clause 5.5.1 in TS 23.401 are used to transfer the EPS bearer context information, which includes the policy for "RCI signalling", from the source eNodeB to the target eNodeB.

For the case of home routed roaming, when MME receives the policy for RCI signalling message from the SGW, MME shall figure out whether the UE is served by a PGW in a different PLMN (e.g. looking into the APN information). According to the roaming agreement between VPLMN and HPLMN operators, MME decides whether to further provision the policy to eNodeB. If RCI signalling is not allowed to be shared with the HPLMN operator, MME shall discard the policy for "RCI signalling" received from SGW.

NOTE: In case there is no policy for RCI signalling available at the eNodeB for the roaming UE, it behaves according to operator's configuration.

###### 6.1.5.2.3.2 Procedure for aggregating RCI signalling

Figure 6.1.5.2.3.2-1 illustrates the procedure for conveying RAN user plane congestion information to the CN.

1) The eNodeB monitors its RAN user plane congestion situation and detects whether it is congested or not.

2) Once RAN user plane congestion is detected, the eNodeB sends the RAN user plane congestion information to the MME by a new S1-AP message or via a new information element in an existing S1-AP message. RCI delivered over S1-MME includes:

- Congestion level;
- List of user's identities (eNB UE S1AP ID and MME UE S1AP ID) of UEs that are located in the same cell and have "RCI signalling" activated at least for one EPS bearer;
- EPS bearer ID(s) of UEs that have RCI signalling activated;
- Direction of user plane congestion (e.g., radio uplink, radio downlink).

For each EPS bearer of all UEs in the list sent by the eNodeB, the MME stores the congestion level and direction of user plane congestion.

3) For each UE in the list sent by the eNodeB and for each active PDN connection belonging to the same UE, the MME notifies the SGW about the RAN user plane congestion information by a new GTP-C message or by a new parameter in an existing GTP-C message. RCI delivered over S11 includes congestion level, user identity (IMSI), EPS bearer ID(s), direction of user plane congestion (e.g., radio uplink, radio downlink), optionally user location information (e.g., Cell ID).

4) The SGW notifies the PGW about the RCI by a new GTP-C message or by a new parameter in an existing GTP-C message. RCI delivered over S5/S8 includes congestion level, user identity (IMSI), EPS bearer ID(s), direction of user plane congestion (e.g., radio uplink, radio downlink), optionally user location information (e.g., Cell ID).

NOTE1: In case that the UE is served by multiple PGWs, the number of the RCI signalling message(s) should be equal to the number of PGWs serving this UE via this SGW.

5) The PGW acknowledges the notification of RAN user plane congestion to SGW.

6) The SGW acknowledges the notification of RAN user plane congestion to MME.

7) The PGW notifies the PCRF about the RCI.

8) The PCRF acknowledges the notification of RAN user plane congestion to PGW.

9) The PCRF makes a decision on how to mitigate the RAN user plane congestion and may initiate IP-CAN Session Modification procedure in order to provide mitigation policies to the PCEF/TDF or decides to forward congestion information to the AF and/or TDF.

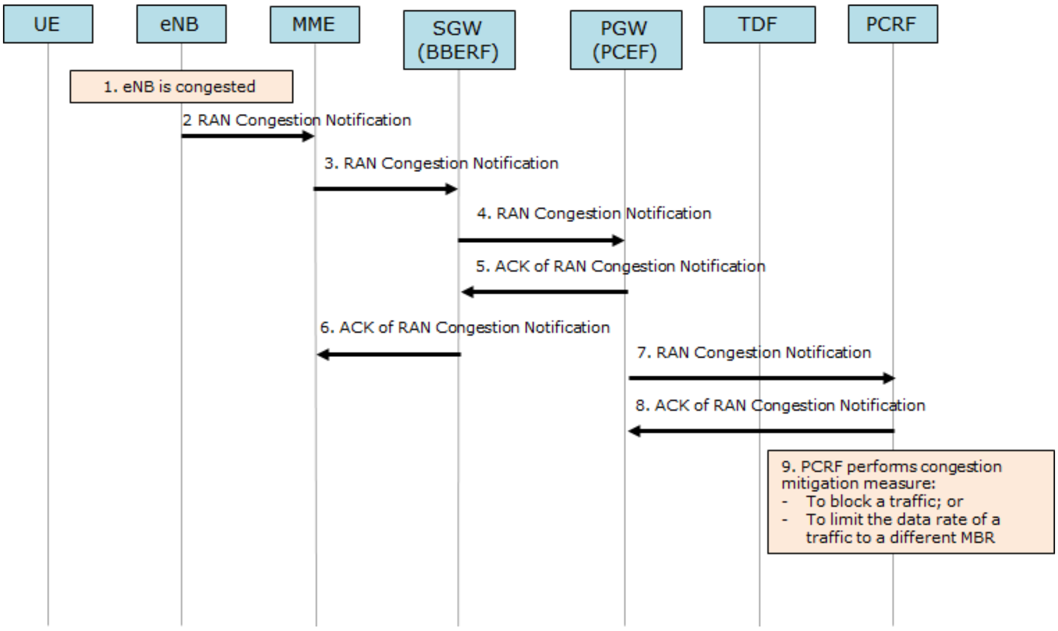
UE
eNB
MME
SGW (BBERF)
PGW (PCEF)
TDF
PCRF
1. eNB is congested
2 RAN Congestion Notification
3. RAN Congestion Notification
4. RAN Congestion Notification
5. ACK of RAN Congestion Notification
6. ACK of RAN Congestion Notification
7. RAN Congestion Notification
8. ACK of RAN Congestion Notification
9. PCRF performs congestion mitigation measure:
- To block a traffic; or
- To limit the data rate of a traffic to a different MBR

**Figure 6.1.5.2.3.2-1: Procedure for RAN user plane congestion notification from the RAN to the CN**

In case that the RAN user plane congestion level is changed or abated, the same procedure as described in Figure 6.1.5.2.3.2-1 is applied. The only difference is that the value for Congestion level parameter is now set with a new value.

For the case of intra E-UTRAN handover, the same procedures as specified in sub-clause 5.5.1 in TS 23.401 [8] are used to transfer the bearer context information from the source eNodeB to the target eNodeB. The bearer context information includes the policy for RCI signalling and the congestion level of the source eNodeB. This enables the target eNodeB to know whether the RAN user plane congestion information shall be reported to CN for the newly handover UE. If the RCI signalling is activated and there is a change in congestion level comparing with the source eNodeB, the target eNodeB sends an aggregating RCI signalling to the MME as described in step 2.

For the case of a UE which performs the Service Request procedure, as specified in sub-clause 5.3.4.1 in TS 23.401 [8], the MME sends the bearer context information to the eNodeB via the existing S1-AP Initial Context Setup Request message. The bearer context information also includes the policy for RCI signalling and the RAN user plane congestion level that are stored at the MME. The eNodeB uses the same procedure as described in Figure 6.1.5.2.3.2-1 to report the RAN user plane congestion notification to the PCRF, if the RCI signalling is activated and congestion level of the current serving eNodeB changes comparing with the previous congestion level received from the MME.

NOTE 2: When to send the aggregating RCI signalling to MME from eNB can be configured by the operator.

Editor's Note: Supporting of PMIP-based S5/S8 is FFS.

#### 6.1.5.2.4 Impact on existing entities and interfaces

#### 6.1.5.2.5 Solution evaluation

### 6.1.5.3 Solution 1.5.3: RPPF based RAN user plane congestion reporting

#### 6.1.5.3.1 General description, assumptions, and principles

This solution addresses the key issue of "RAN user plane congestion awareness".

A new logical function entity, RAN Payload Perceive Function (RPPF), is proposed to collect RAN user plane congestion information and further report to PCRF for the purpose of congestion mitigation.

The PCRF may then report over Rx UE congestion information to applications that have subscribed to this information

A new reference point Np is introduced between RPPF and PCRF to pass on the RAN user plane congestion information to PCRF.

### 6.1.5.3.2 High-level operation and procedures

#### 6.1.5.3.2.1 Architecture

Figure 6.1.5.3.2-1 shows the proposed UPCON architecture.

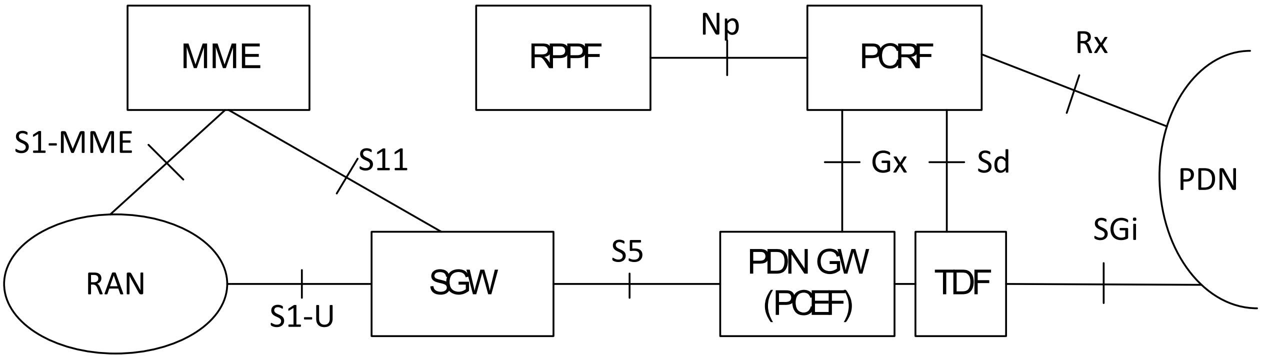

**Figure 6.1.5.3.2-1: UPCON Architecture**

New Reference Point:

Np: The reference point between RPPF and PCRF.

RAN User Plane Congestion Information (RUCI) includes following information elements:

(1) Congestion/Abatement location information (e.g. Cell ID);

(2) Congestion level

Editor's Note: It is FFS to determine if more information is needed (e.g. to optimize the congestion mitigation operation)

The functionality of RPPF:

- Collecting and processing RAN's cell congestion information from OAM;
- Communicating with PCRFs serving the PLMN for RAN user plane congestion information reporting.

Editor's Note: It is FFS whether RPPF can collect information from entities other than OAM.

Editor's Note: It is FFS which mechanism is used to determine the UE location in order to be able to determine the UEs affected by congestion of specific cells.

#### 6.1.5.3.2.2 Procedure to Report RAN User Plane Congestion Information to CN

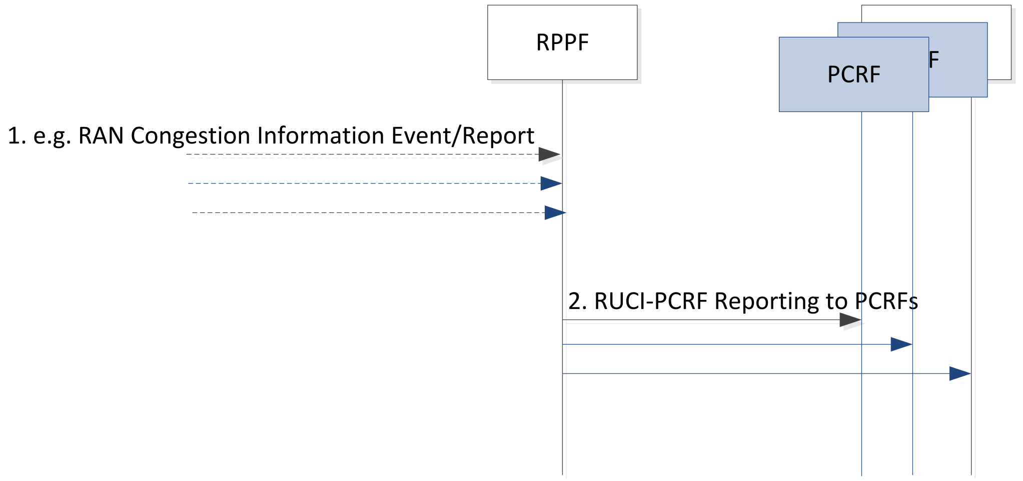
RPPF
PCRF
F
1. e.g. RAN Congestion Information Event/Report
2. RUCI-PCRF Reporting to PCRFs

**Figure 6.1.5.3.2.1-1: Procedure to Report RAN User Plane Congestion Information to CN**

1. Based on network operation policy, for example, an event/report will be sent to RPPF due to radio node/cell user plane congestion or abatement is reached to a pre-configured engineered thresholds with the indication of the affected area (e.g. Cell); another example is that RPPF may solicit the RAN User Plane Congestion Information based on an engineered interval.

2. RPPF reports the RUCI-PCRF congestion status to PCRFs that serve the PLMN.

#### 6.1.5.3.3 Impacts on existing entities and interfaces

The impact on PCRF:

- The PCRF should be enhanced to collect RUCI from RPPF;
- The PCRF should be enhanced to determine congestion policy based on RUCI, subscriber profile, type of application, type of content, etc.

#### 6.1.5.3.4 Solution evaluation

Editor's note: It is FFS.

## 6.1.6 RAN Congestion Mitigation Solutions

### 6.1.6.1 Solution 1.6.1: Policy-based Congestion Mitigation

#### 6.1.6.1.1 General description, assumptions, and principles

This solution addresses key issues #1 ("RAN User Plane congestion mitigation") and #4 ("Video delivery control for congestion mitigation"). It describes a general scheme how PCRF can be involved for congestion mitigation based on policy decisions, with the PCRF providing policies to different network entities performing congestion mitigation, based on congestion awareness.

This solution focuses only on policy-based congestion mitigation, and does thus not depend on how congestion awareness is achieved in the PCRF (e.g. if the congestion information is signalled off-path or if they are indicated on-path via the P-GW).

NOTE: The term "congestion information" is used here as a generic term and the detailed information elements are left to the congestion awareness solution.

### 6.1.6.1.2 High-level operation and procedures

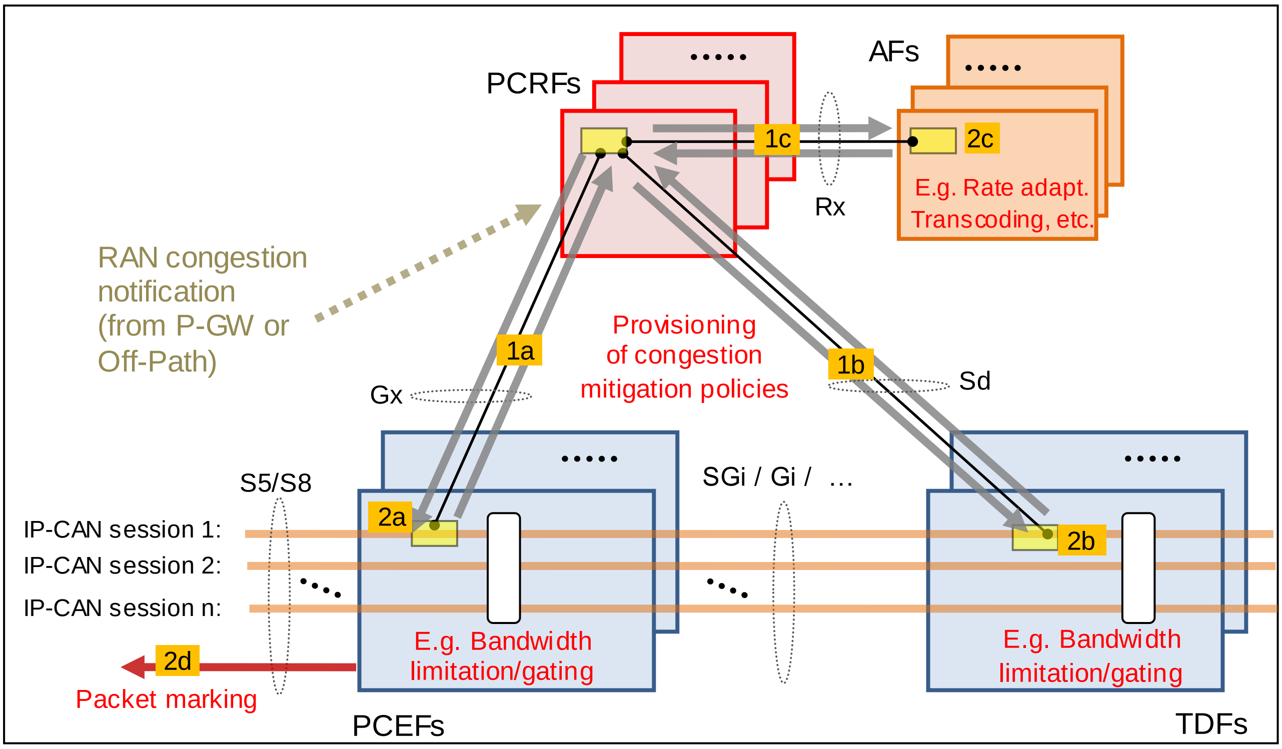

**Figure 6.1.6.1.2-1: Overview of congestion mitigation based on policy decisions.**

NOTE 1: The numbers do not necessarily imply a temporal order.

NOTE 2: If TDF is deployed, congestion mitigation policies may be provisioned to both PCEF and/or TDF.

The procedural steps are:

1. The PCRF provides policies for congestion mitigation to one or more of the following network entities:

   a) to the PCEF (over the Gx interface);

   b) to the TDF (over the Sd interface) ;

   The policies can be provisioned before RAN user plane congestion occurs or after the PCRF becomes aware of the congestion status (e.g. onset, abatement, level of RAN user plane congestion).

NOTE 3: The PCRF may use subscriber information (e.g. from SPR) as input for the policy decisions.

NOTE 4: In case of network configurations without PCRF involvement, the PCEF and/or TDF can enforce static congestion mitigation policies upon receipt of a congestion notification from the RAN. Different policies may be configured for different congestion levels.

Editor's Note: It is FFS if, and if so, how the TDF receives the congestion notification from the RAN for the deployment scenario described by NOTE 4.

2. The PCRF may also provide – subject to agreement with the AF provider – an indication to the AFs (over the Rx interface).

Editor's Note: The details of the indications / information provided to the AF over the Rx are FFS. The indication could for example include the level of service that is supported (e.g. max. bitrate).

3. Congestion mitigation is performed in different network entities according to the policy decision by the PCRF:

   a/b) The PCEF/TDF can perform bandwidth limitation, prioritization and traffic gating according to the provided policies.

c) The AF (e.g. an application server or proxy) can directly or indirectly support the congestion mitigation, e.g. by adapting the sending rate, through media transcoding or compression, or by delaying push services.

d) Based on policies provided by the PCRF the P-GW/TDF may also perform actions to support congestion mitigation measures in the RAN, e.g. the policy can control when packet marking (such as e.g. proposed by Solution 3) should be performed.

e) The PCRF may limit/reject the authorization of new requests for application flows, based on current procedures.

#### 6.1.6.1.3 Assumptions for extensions of policies for congestion mitigation

With this solution, the following definition is used for extension of the policy framework:

**User plane congestion mitigation policy:** A set of information describing actions in the user plane (in the PCEF/TDF) with the target to reduce the (overall or specific) amount of RAN user plane congestion or to minimize service disruption/service degradation experienced by the user, and corresponding conditions under which they shall be performed. Such a policy may be provisioned statically in PCEF, pre-provisioned in PCEF/TDF and de/activated dynamically by PCRF or provisioned dynamically by PCRF to PCEF/TDF. A user plane congestion mitigation policy refers to a level of congestion. A pre-provisioned or dynamically provisioned user plane congestion mitigation policy may contain an event trigger for a subsequent user plane congestion report.

NOTE 1: for static user plane congestion mitigation policies the same restrictions apply as for current static PCC (defined in TS 23.401 subclause 4.7.5 and TS 23.402 subclause 4.10.1).

Editor's note: it is FFS if further restrictions or conditions apply with static user plane congestion mitigation policies, e.g. with respect to admission control.

NOTE 2: possible mitigation measures may be e.g. bandwidth limitation, packet marking etc. Currently available capabilities in PCC/ADC rules are bandwidth limitation, gating, QoS information and redirection.

With this solution, the following assumptions for extension of policies are used:

- All existing variants of policy provisioning are useful to have also for congestion mitigation.
- For user plane congestion mitigation, an enhancement of existing PCC/ADC rules should be defined. They should contain congestion mitigation measures per congestion level (for one or a set of congestion levels).

#### 6.1.6.1.4 Impact on existing entities and interfaces

Details are FFS.

#### 6.1.6.1.5 Solution evaluation

## 6.2 RAN-based Solutions for RAN user plane congestion management

### 6.2.1 Solution 2.1: Flow Priority-based Traffic Differentiation on the same QCI (FPI)

#### 6.2.1.1 General description, assumptions, and principles

This solution addresses the key issue on "RAN User Plane congestion mitigation". The solution also addresses certain aspects of the key issue on "Video delivery control for congestion mitigation" and certain aspects of the key issue on "Differentiated treatment for non-deducible service data flows in case of RAN user plane congestion".

Based on operator's policies and on the information collected after some form of packet inspection (e.g. shallow packet inspection, L7 DPI, heuristic analysis or others) performed either by the GGSN/PGW or by the TDF, the GGSN/PGW marks each user plane data packet delivered in the downlink direction with a Flow Priority Indicator (FPI) identifying the relative priority of the packet compared to other packets mapped to the same QCI.

For GTP-based interfaces the FPI marking is provided in the GTP-U header of downlink user plane packets.

NOTE 1: The FPI could be defined as a new GTP-U extension header, completely independent from the SCI, or as an enhancement of the GTP-U extension header specified in Rel-11 to convey the SCI. The details are up to stage 3.

Editor's note: If and how the approach can be exploited also in the uplink direction is FFS.

Editor's note: How to deliver the FPI to the RAN with PMIP-based S5/S8 is FFS.

The range of valid FPI values shall be standardized.

The usage of the FPI is expected to be useful for Non-GBR QCIs only.

NOTE 2: According to 3GPP TS 23.203, services using a GBR QCI and sending at a rate smaller than or equal to GBR can in general assume that congestion related packet drops will not occur.

The FPI is not intended to replace the QCI, and no conflicts are foreseen between the FPI and the QCI. The FPI complements the QCI as described below:

- Both the FPI marking of each user plane packet and the Priority level associated to a Service Data Flow (SDF) aggregate via its QCI are used to differentiate between IP flows of the same UE, and are also used to differentiate between IP flows of different UEs.

- Via its QCI an SDF aggregate is associated with a Priority level and a Packet Delay Budget (PDB). As defined in section 6.1.7.2 of 3GPP TS 23.203, if the target set by the PDB can no longer be met for one or more SDF aggregate(s) across all UEs that have sufficient radio channel quality then a scheduler shall give precedence to meeting the PDB of SDF aggregates with higher Priority level.

- If the target set by the PDB can no longer be met for one or more packet(s) belonging to SDF aggregate(s) with the same Priority level (across all UEs that have sufficient radio channel quality) then a scheduler should give precedence to meeting the PDB for the packets with higher FPI.

NOTE 3: The details of scheduling are out of scope of 3GPP but implementations are assumed to ensure that starvation of flows with lower FPI is avoided.

If the usage of the FPI is enabled in the RAN, the packets that do not include any FPI marking should be scheduled according to a default FPI pre-configured in the RAN. The default FPI may be configured per PLMN.

NOTE 4: The default FPI pre-configured in the RAN allows support of home routed roaming scenarios where the FPI is used in the VPLMN but not in the HPLMN. The default FPI pre-configured in RAN also enables deployment scenarios where, based on operator's configuration, only downlink user plane packets belonging to specific applications, or application data flows, are marked by the GGSN/PGW with the FPI, while the rest of traffic is not marked. If the usage of the FPI is not enabled in the RAN, the RAN shall ignore the Flow Priority Indicator if received over the S1-U, S12 or other interface, i.e. the RAN shall treat the user plane packet normally.

The usage of the FPI, in conjunction with the QCI, to prioritize user plane data packets has the following characteristics and peculiarities:

- It is applicable to any RAT, i.e. A/Gb mode GERAN, UTRAN and E-UTRAN.

- Delivery of the FPI in downlink user plane data packets should be supported for both GTP-based and PMIP-based S5/S8.

- The FPI should be included in charging records and transferred over online/offline charging interfaces. This is because the FPI can be used for traffic handling differentiation, hence may affect the user experience of the customer and may be used by the operator to create different service profiles.

- It should be possible for the GGSN/PGW to set the FPI based on subscription. Support for PCC control of the feature is therefore necessary.

If both Rel-11 SIRIG (see section 5.3.5.3 of 3GPP TS 23.060 [4]) and the solution described in this section are enabled in an operator's network, considering that the SCI is defined only for A/Gb mode GERAN while the FPI is applicable to any RAT, the following occurs:

- Both the SCI and the FPI are delivered to A/Gb mode GERAN.
- Only the FPI is delivered to UTRAN and E-UTRAN.

The SCI and the FPI provide complementary information to the RAN:

- The SCI indicates the type of application that generated the user plane packet and may be used by A/Gb mode GERAN to optimize resource allocation, e.g. to avoid allocating more time slots than what the application actually needs.
- The FPI indicates the priority of the user plane packet and may be used by A/Gb mode GERAN to decide which traffic flows should be served first in case of congestion.

Editor's note: It is FFS if it would be beneficial for the solution described in this section to extend the applicability of the SCI to all RATs. With the GGSN/PGW delivering both the SCI and the FPI over any RAT, the RAN would become aware of both the priority and the application type associated to each user plane packet. If and how that could be used to allow for more efficient packet scheduling in case of RAN user plane congestion is to be determined.

Editor's note: The interactions between SCI and FPI in case both are delivered to the RAN are FFS.

As discussed for SIRIG during the Rel-11 timeframe, from a deployment perspective it would be beneficial to also support scenarios where the packet classification required to properly set the FPI is performed by a TDF, rather than the GGSN/PGW. To that purpose a mechanism is required to transfer the outcome of the packet classification process from the TDF to the GGSN/PGW, so that the GGSN/PGW can then use that information to mark packets in the downlink direction. Possible tunnelling/marking mechanisms that could be used to solve this issue are described in 3GPP TR 23.800 [5] Annex B.

The following tunnelling/marking solutions are under consideration to be used between the TDF and the GGSN/PGW in order to classify packets detected by the TDF:

- DSCP

NOTE 5: Marking of DSCP bits for this purpose can interfere with appropriate traffic handling in some operator transport networks. The DSCP marking may also get remarked by routing entities within the operator networks.

- Tunnel which carries DSCP marking implemented in the inner IP packet header

In case of Tunnel which carries DSCP marking implemented in the inner IP packet header option, original DSCP markings used in operator's network are used in the outer DSCP field of the tunnel in order to keep the transport network unaffected. The examples of the tunnels which may carry the DSCP marking are: GRE, IP-in-IP tunnel, depending on implementation.

Editor's note: The additional tunnelling options (e.g. GTP-U) are FFS and can be exploited in the future.

Editor's note: It is FFS if and how RAN user plane congestion awareness can be exploited to optimize the solution described in this section. For example an option to be investigated is the possibility to enable the packet classification required to properly set the FPI only in case of RAN user plane congestion, in order to minimize the performance impacts on the GGSN/PGW or the TDF.

### 6.2.1.2 High-level operation and procedures

Overall the solution would work as described below (see Figures 6.2.1.2-1 and 6.2.1.2-2):

- In case the packet classification is performed by the GGSN/PGW, upon packet classification the GGSN/PGW derives the FPI to be provided in downlink user plane data packets based on configuration or based on the FPI parameters received from the PCRF within the corresponding PCC Rule. In case the packet classification is performed by the TDF based on configuration or based on ADC rules received from the PCRF, the TDF marks the packet according to the result of the packet classification. Then, GGSN/PGW performs FPI marking based on PCC rules which take into account the result of packet inspection received from the TDF and then provides the FPI marking in the downlink user plane data packets.

- When receiving the FPI in a user plane packet, the SGSN, or the Serving Gateway (SGW), copies it, without modifying its value, into a correspondent information element over Gb, Iu or S1. In order to support roaming scenarios, the FPI should be forwarded over Gb, Iu or S1 together with the HPLMN ID and additional information, added by the SGSN or SGW, which indicates whether the FPI is assigned by a GGSN/PGW in the Home PLMN, by a GGSN/PGW in the Visited PLMN or by a GGSN/PGW for which the FPIs are coordinated across the different operator group PLMNs and the serving PLMN of the SGSN or SGW (Operator Group GGSN). Absence of additional information is an indication of a VPLMN provided FPI.

NOTE: The SGSN or SGW determines and indicates “Operator Group GGSN” based on local configuration.

- For roaming subscribers, based on local configuration, and taking into account the HPLMN ID and the GGSN/PGW location information provided by the SGSN or SGW , the RAN may remap the FPI received in the downlink user plane packet to a value locally configured in the RAN. The RAN uses the FPI associated to each downstream user plane packet and the QoS parameters associated to the bearer, such as the QCI, to prioritize the packets delivered over the air interface.

Editor’s note: The current description of the usage of the FPI in roaming scenarios is aligned with what was defined in Rel-11 for SIRIG, where remapping of the SCI values in downlink user plane packets is performed by the GERAN access in VPLMN. Considering that the FPI, differently from Rel-11 SCI, is applicable to all RATs, it is FFS whether other solutions should be considered (e.g. remapping of the FPI at the SGW, usage of GTP firewalls, or others).

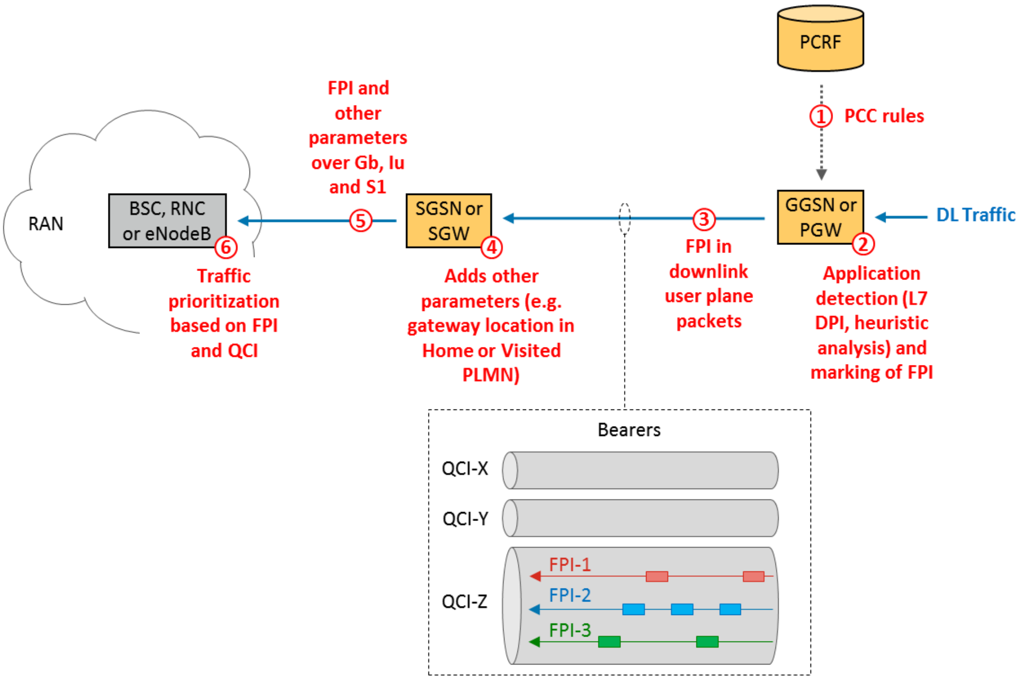
PCRF
1 PCC rules
GGSN or PGW
DL Traffic
2
Application detection (L7 DPI, heuristic analysis) and marking of FPI
3
FPI in downlink user plane packets
SGSN or SGW
4
Adds other parameters (e.g. gateway location in Home or Visited PLMN)
FPI and other parameters over Gb, Iu and S1
5
RAN
BSC, RNC or eNodeB
6
Traffic prioritization based on FPI and QCI
Bearers
QCI-X
QCI-Y
QCI-Z
FPI-1
FPI-2
FPI-3

**Figure 6.2.1.2-1: RAN congestion mitigation based on the FPI with packet classification performed by the GGSN/PGW**

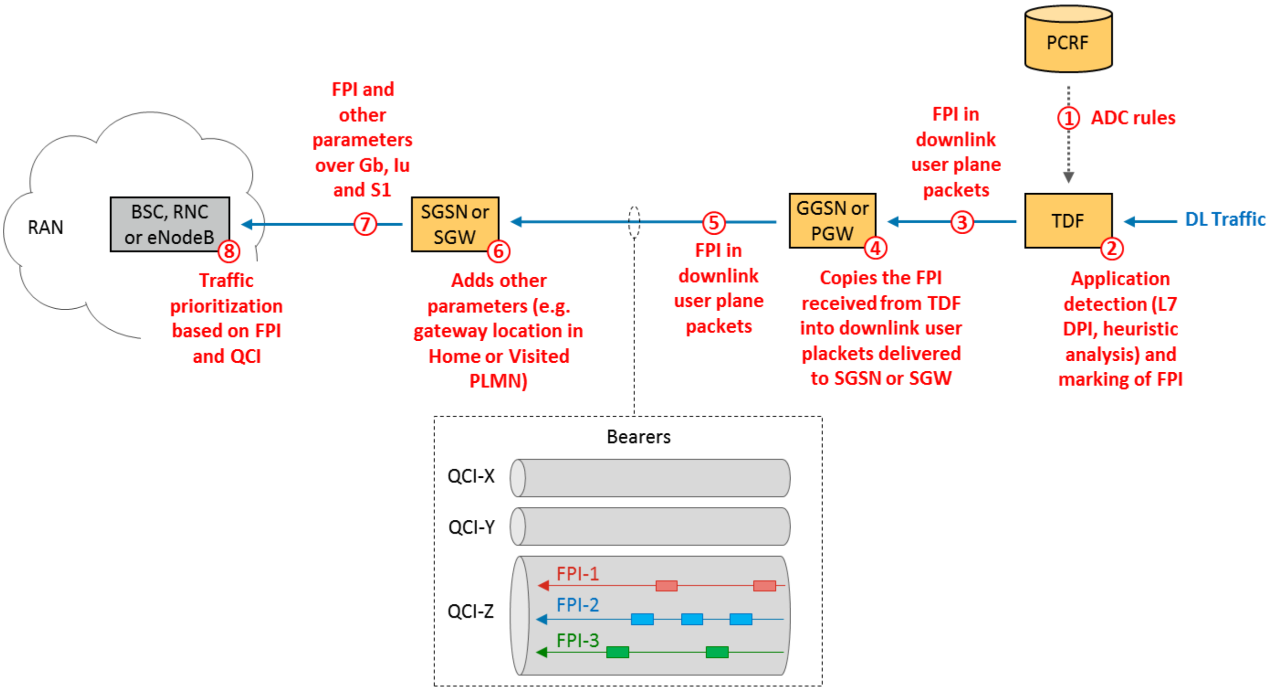
PCRF
1 ADC rules
DL Traffic
TDF
2
Application detection (L7 DPI, heuristic analysis) and marking of FPI
FPI in downlink user plane packets
3
GGSN or PGW
4
Copies the FPI received from TDF into downlink user plackets delivered to SGSN or SGW
5
FPI in downlink user plane packets
SGSN or SGW
6
Adds other parameters (e.g. gateway location in Home or Visited PLMN)
7
FPI and other parameters over Gb, Iu and S1
RAN
BSC, RNC or eNodeB
8
Traffic prioritization based on FPI and QCI
Bearers
QCI-X
QCI-Y
QCI-Z
FPI-1
FPI-2
FPI-3

**Figure 6.2.1.2-2: RAN congestion mitigation based on the FPI with packet classification performed by the TDF**

### 6.2.1.3 Impact on existing entities and interfaces

GGSN and PGW:

- Marking of the Flow Priority Indicator (FPI) in downlink user plane data packets based on the configuration or the policies received from the PCRF and the information collected after some form of packet inspection.

- Inclusion of the FPI in CDRs and transfer the FPI over online/offline charging interfaces.

- In case the TDF is deployed for packet classification, taking into account the received packet classification for determining the FPI value which is then provided in the downlink user plane data packets.

TDF:

- Marking of the downlink user plane data packets based on the configuration or the policies received from the PCRF and the information collected after some form of packet inspection.

- Inclusion of the FPI in CDRs and transfer the FPI over online/offline charging interfaces.

NOTE: This can be done if TDF marks the classified packets in the same way as PCEF will mark FPI in the downlink packets. This can be achieved by having appropriate configuration at the TDF or appropriate ADC Rule setting by the PCRF.

SGSN and SGW:

- When receiving the FPI in a packet, the SGSN, or SGW, copies it, without modifying its value, into a correspondent information element over Gb, Iu or S1.

- Together with the FPI, the SGSN, or SGW, provides to the RAN the HPLMN ID and additional information, which indicates whether the FPI is assigned by a GGSN/PGW in the Home PLMN, by a GGSN/PGW in the Visited PLMN or by a GGSN/PGW for which the FPIs are coordinated across the different operator group PLMNs and the serving PLMN of the SGSN or SGW (Operator Group GGSN).

PCRF:

- Provision of PCC/ADC Rules to control FPI marking on per subscriber and/or per application basis.

OCS and OFCS:

- Support for charging differentiation based on the FPI.

BSC, RNC and eNodeB:

- Usage of the FPI, in conjunction with the QCI, to prioritize the packets delivered over the air interface.

Editor's note: The impacts on existing entities and interfaces with PMIP-based S5/S8 are FFS.

#### 6.2.1.4 Solution evaluation

Editor's note: The solution evaluation is FFS.

### 6.2.2 Solution 2.2: Flow and bearer QoS differentiation by RAN congestion handling description (FQI)

#### 6.2.2.1 General description, assumptions, and principles

Editor's Note: This sub-clause should identify the key issues address by this solution.

This solution addresses key issues #1, #2 and certain aspects of key issues #3, #4 and #5. The solution applies to non-GBR bearers.

The PGW/GGSN may mark downlink data packets with FQI – Flow QoS Index, identifying a specific RAN treatment that these packets should receive. The marking is done based on operator's policies and on the information collected after some form of packet inspection (e.g. shallow packet inspection, L7 DPI, heuristic analysis or others) performed either by the GGSN/PGW itself or by the TDF. There is full flexibility in how the traffic flows are mapped to FQI markings in the core network. A number of criteria can be used such as:

- Service category (such as web, file download, video, etc.)
- Application (such as YouTube, Skype, etc.)
- Subscription (such as Gold, Silver, Bronze)
- Header fields (such as a range of IP addresses or port numbers)
- Usage policies (such as heavy user, light user)
- Any combination of the above.

For GTP-based interfaces the FQI marking is provided by the GGSN/PGW in the GTP-U header of downlink user plane packets.

In case the TDF performs packet inspection, the GGSN/PGW performs FQI marking based on PCC rules which take into account the result of packet inspection received from the TDF and then provide the FQI in the downlink user plane data packets within the GTP-U header.

Editor's Note: How to deliver the FQI to the RAN with PMIP-based S5/S8 is FFS.

The RAN handling of a given traffic class at a certain congestion level is described by the RAN Congestion Handling Descriptor (RCHD) as will be described below. The traffic class of a flow belonging to a specific user is determined by the combination of QCI corresponding to the radio bearer and the FQI packet marking of the traffic flow. For each QCI, a traffic class is also defined by the QCI in combination with no FQI packet marking. For each traffic class, separate RCHDs are provided for the set of congestion levels {low, high}. Hence, the RCHD describes the RAN handling per QCI, per FQI, per congestion level.

Editor's Note: The number of congestion levels to be defined is FFS.

NOTE 1: One example for defining downlink traffic classes is that traffic flows with QCI=9 are differentiated by different FQI values. Another example for defining both downlink and uplink traffic classes is that traffic flows are differentiated into bearers with non-standardized QCI values, and no FQI marking is used. Other examples for defining traffic classes using a combination of FQI and QCI values (both standardized and non-standardized) are also possible.

NOTE 2: Certain QCIs may be excluded from the RCHD based description. In that case, QoS differentiation is based on the QCI only.

In case of congestion, i.e., when the resource demand of traffic flows exceeds the available capacity, the RAN performs allocation of resources as described by the QCI characteristics and the RCHDs of the flows. The QCI based differentiation is applied first. The RAN then tries to allocate resources as described by the RCHDs of the flows corresponding to the lowest congestion level, within the bounds of the QCI characteristics; if that is not feasible it tries to apply the RCHDs at a higher congestion level. The RAN applies the lowest congestion level to the set of traffic flows that is feasible within the bounds of the QCI characteristics. Hence the QCI characteristics of traffic flows always take precedence over the RCHDs of the traffic flows in determining the resource sharing.

The RCHD shall be capable of expressing a bitrate which corresponds to the minimal amount of resources allocated to the given traffic flow at a given congestion level. The bitrates corresponding to the lowest congestion level that is feasible in the current resource situation are applied observing the QCI based constraints of the bearers. Once the RAN determines that the bitrate target cannot be achieved on a given congestion level, it tries to apply the bitrates for the next higher congestion level. The RCHD may express the RAN handling by other parameters as well, instead of or in addition to the bitrate.

Editor's Note: It is FFS how the operator can control the allocation of remaining resources. Possible options:

- A sufficiently high number of congestion levels can be defined so that the resource allocation can be made sufficiently accurate for the operator. In this case, the allocation of remaining resources, or the resource allocation if the highest congestion level is not feasible, can be undefined and left to the implementation.

- Some form of priority scheme can be defined.

The RCHD may also describe how the radio channel quality is taken into account in the resource allocation under congestion. A user with a worse channel quality may experience a different performance at a given congestion level compared to a user with a better channel quality. By taking the channel quality into account, it may be possible to control whether a user with worse channel quality is being compensated by additional radio resources and to what extent such a compensation is done. Hence, RCHD parameters such as for example the bitrate may be combined with the consideration of the radio channel quality to determine the actual resource sharing.

The parameters applied for the selected RCHD are considered over an averaging period. The details of how the averaging is performed are implementation specific. The averaging may e.g., take into account how the packet arrivals are distributed over time.

In addition to enabling differentiated handling in congestion scenarios the RCHD may also be used to express an optimized handling of a certain traffic class to the RAN. Besides the RAN handling for general best effort traffic, the use of different RCHDs can for example make it possible to express an optimized handling for a certain types of application/service classes in order to further improve the radio resource utilization and/or user experience.

The RCHD is realized by one or more vendor defined parameters that are configurable via O&M. The RAN is required to enable the configuration of the RCHD on a per QCI, per FQI, per congestion level granularity. The standardization of the FQI values themselves are not necessary. Consistency of the RAN handling in a multivendor environment is ensured by the requirement for the same granularity of RCHD configuration, by the requirement that RCHD is capable of expressing a bitrate which corresponds to the minimal amount of resources allocated to the given traffic flow at a given congestion level, and by the requirement that the RAN applies the lowest congestion level's RCHD that is feasible.

Regarding the relationship of FQI and rel-11 SCI, FQI is backwards compatible to SCI for GERAN and can be regarded as an evolution of SCI. The SCI is typically associated with service category or application based classification, whereas the FQI is meant to allow any type of classification. FQI allows operators to explicitly and quantitatively set the RAN handling at different levels of congestion, which is not supported by SCI. SCI is intended for application specific RAN optimizations, which is possible, although not required by the FQI approach.

It is suggested that the rel-11 SCI mechanism for GERAN is evolved to the rel-12 FQI concept. The rel-11 GERAN SCI based treatment may need to be evolved to implement the RCHD based handling as described above. This evolution is useful in order to harmonize the packet marking treatment for all 3GPP RATs according to the UPCON approach. This evolution is backwards compatible: as long as the packet marking formatting is backwards compatible on stage 3 level, rel-11 SCI implementations and rel-12 FQI implementations can co-exist in the same network, no matter whether some RAN nodes or some CN nodes are of a different release. This means that if there are existing GERAN realizations of SCI which can improve the radio resource efficiency, they can continue to be used in the context of the FQI approach.

The following tunnelling/marking solutions are under consideration to be used between the TDF and the GGSN/PGW in order to classify packets detected by the TDF:

- DSCP

NOTE 3: Marking of DSCP bits for this purpose can interfere with appropriate traffic handling in some operator transport networks. The DSCP marking may also get remarked by routing entities within the operator networks.

- Tunnel which carries DSCP marking implemented in the inner IP packet header

In case of Tunnel which carries DSCP marking implemented in the inner IP packet header option, original DSCP markings used in operator's network are used in the outer DSCP field of the tunnel in order to keep the transport network unaffected. The examples of the tunnels which may carry the DSCP marking are: GRE, IP-in-IP tunnel, depending on implementation.

Editor's note: The additional tunnelling options (e.g. GTP-U) are FFS and can be exploited in the future.

### 6.2.2.2 High-level operation and procedures

Overall the solution would work as described below:

- In case the packet classification is performed by the GGSN/PGW, upon packet classification the GGSN/PGW derives the FQI to be provided in downlink user plane data packets based on configuration or based on the FQI parameter received from the PCRF within the corresponding PCC Rule.

- In case the packet classification is performed by the TDF, upon packet classification, the TDF marks the downlink packets according to the result of the packet classification based on configuration or based on the ADC rule received from the PCRF. Then, GGSN/PGW performs FQI marking based on PCC rules which take into account the result of packet inspection received from the TDF.

- When receiving the FQI in user plane packet, the SGSN, or the Serving Gateway (SGW), copies it, without modifying its value, into a correspondent information element over Gb, Iu or S1.

- In the roaming case, the SGSN or the SGW may remap the FQI to a value used in the VPLMN based on a roaming agreement, or in the absence of a roaming agreement to a value that may be based on the HPLMN. Alternatively, the GGSN/PGW in the HPLMN may also set the FQI based on the VPLMN.

Editor's Note: Which alternative is applied in the roaming case is FFS.

- The RAN handling is determined by the QCI and the RCHD for the given combination of QCI and FQI of the traffic flow for the given congestion level, as described above.

### 6.2.2.3 Impact on existing entities and interfaces

GGSN and PGW:

- Marking of the Flow QoS Index (FQI) in downlink user plane data packets based on the configuration or the policies received from the PCRF and the information collected after some form of packet inspection.

- Inclusion of the information needed to enable charging based on FQI when reporting over online/offline charging interfaces and when performing credit control over online charging interfaces.

- In case the TDF is deployed for packet classification, taking into account the received packet classification for determining the FQI value which is then provided in the downlink user plane data packets.

TDF:

- Marking of the downlink user plane data packets based on the configuration or the policies received from the PCRF and the information collected after some form of packet inspection.

- Inclusion of the information needed to enable charging based on FQI when reporting over online/offline charging interfaces and when performing credit control over online charging interfaces.

NOTE: This can be done if TDF marks the classified packets in the same way as PCEF will mark FQI in the downlink packets. This can be achieved by having appropriate configuration at the TDF or appropriate ADC Rule setting by the PCRF.

SGSN and SGW:

- When receiving the FQI in a packet, the SGSN, or SGW, copies it, without modifying its value, into a correspondent information element over Gb, Iu or S1.

PCRF:

- Provision of PCC/ADC Rules to control FQI marking.

OCS and OFCS:

- Support for charging differentiation on the applied FQI based on the principles for PCC flow based charging.

BSC, RNC and eNodeB:

- Realize packet treatment taking into account the RCHD for the different congestion levels which can be set via vendor specific QoS parameters for a combination of QCI and FQI.

Editor's Note: The impacts on existing entities and interfaces with PMIP-based S5/S8 are FFS.

#### 6.2.2.4 Solution evaluation

Editor's Note: The solution evaluation is FFS.

### 6.2.3 Solution 2.3: Enhancing Existing Bearer Concepts

#### 6.2.3.1 Solution principles

This solution is targeting to solve RAN user plane congestion mitigation by re-using and enhancing the existing bearer concept to cope with RAN overload situations. This solution is based on the following principles and pre-requisites:

- The Core Network is in charge of subscriber and service management (policy control) and is not required to be aware of RAN resources or cell load situation.

- The RAN takes care of congestion handling, resource management (RRM) and performs resource allocations (policy enforcement).

- The QoS and priority on a per subscriber or service level (= policy) is delivered from the Core Network to the RAN via bearer specific signalling.

- The UE supports multiple dedicated bearers, which can be pre-established, e.g. established at time of attachment to the network. Dedicated bearers are used on a per need basis and it is up to the operator how many are pre-established. At least one dedicated bearer is required for moving traffic from the default bearer.

- Deep Packet Inspection functionality in the network (via PCEF enhanced with ADC or TDF) is used to identify application traffic and classify/mark data packets. On a per need basis and at any time this functionality could also be used for radio bearer reconfiguration, e.g. addition of a new dedicated bearer fitting to the detected application class.

- The PCEF performs the bearer binding based on the configured PCC rules and packet classification, i.e. traffic flows are allocated to certain (pre-established) dedicated bearers in downlink direction based on SDF rules and the actual packet marking. These dedicated bearers are adapted to carry certain types of applications e.g. by using pre-defined QCI and ARP values.

NOTE: If the Deep Packet Inspection functionality is integrated in the PCEF, the PCEF can use it for evaluating the bearer binding for SDFs detected via pre-defined PCC rules.

- In case the Deep Packet Inspection is performed by the TDF, the TDF classifies the packets and applies corresponding markings. Then the PCEF, upon receiving those marked packets, performs the bearer binding based on the configured PCC rules and packet classification, i.e. traffic flows are allocated to certain (pre-established) dedicated bearers in downlink direction based on SDF rules and the actual packet marking. These dedicated bearers are adapted to carry certain types of applications e.g. by using pre-defined QCI and ARP values.

- In uplink direction the UE can, without the need for an update of installed TFTs, use the same bearer as the network used in downlink direction for a certain flow. This is applicable in case of DSCP based marking performed by the PCEF. In such a case it is also under consideration that TDF, in order to apply marking of packets sent to the PCEF, uses either

  - DSCP

NOTE: Marking of DSCP bits for this purpose can interfere with appropriate traffic handling in some operator transport networks. The DSCP marking may also get remarked by routing entities within the operator networks.

  - Tunnel which carries DSCP marking implemented in the inner IP packet header

In case of Tunnel which carries DSCP marking implemented in the inner IP packet header option, original DSCP markings used in operator's network are used in the outer DSCP field of the tunnel in order to keep the transport network unaffected. The examples of the tunnels which may carry the DSCP marking are: GRE, IP-in-IP tunnel, depending on implementation.

The solution also addresses the following limitation with the current EPS bearer concept:

- An Application uses (potentially many) very short-lived parallel UDP and/or TCP data flows, for which service data flow filters detected via ADC/PCC rules are too short-lived to allow PCC system to control them using SDF templates (aka application with non-deducible service data flows).

### 6.2.3.2 High-level operation and procedures

The EPS bearer concept allows establishing dedicated bearers in addition to the default bearer. Different QoS parameters (QCI and ARP) can be assigned to each dedicated bearer. This guides the radio scheduler to assign resources to each bearer according to the bearer's priority and the actual cell load, thus is able to reduce the throughput of low-priority traffic in case of congestion.

The radio scheduler is able to differentiate any multi-rate traffic mix, it estimates the resources required for GBR bearers and shares the remaining resources between non-GBR bearers according to traffic priority.

A dedicated non-GBR bearer may carry several applications requiring similar QoS treatment in CN and RAN. The core network can be aware of applications and their QoS requirements by using DPI functionality and assigns applications with similar QoS and priority requirements to one dedicated bearer. This allows the RAN to reduce the throughput of low-priority applications (carried in appropriate dedicated bearers) once congestion occurs without explicit notification and assistance of the core network.

The number of established dedicated bearers per UE, e.g. based on subscriber priority (bronze, silver, gold), is determined by operator policies. Operator can also determine whether and which of the dedicated bearers are pre-established, e.g. at time of attachment to the network.

The basic concept of this solution as shown in the following figure is to combine the load-aware functionality in the RAN (eNB/NodeB) with the application and policy awareness of the core, which is enhanced by DPI functionality to detect certain applications. Two configurations are possible, PCEF enhanced with ADC and TDF:

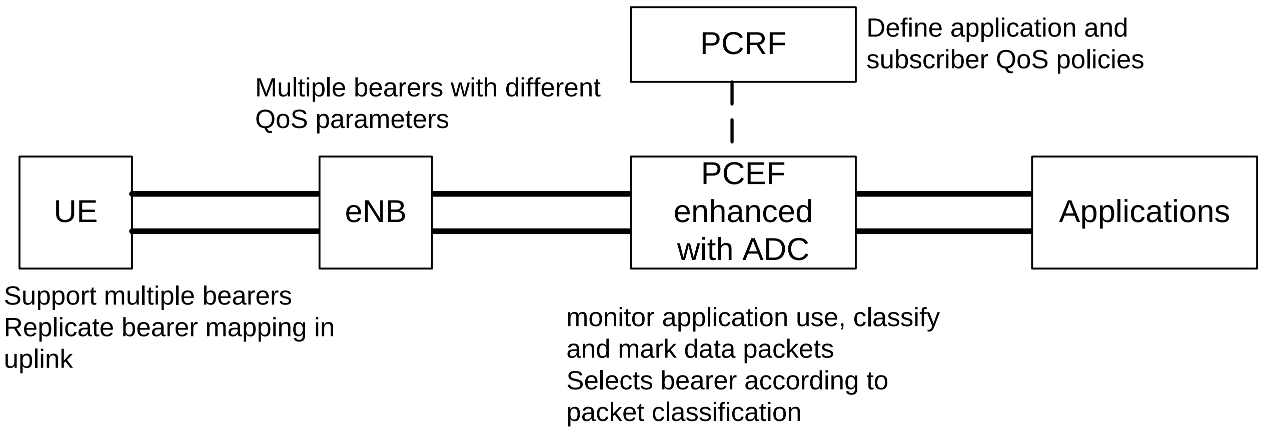
PCRF
Define application and subscriber QoS policies
Multiple bearers with different QoS parameters
UE
eNB
PCEF enhanced with ADC
Applications
Support multiple bearers
Replicate bearer mapping in uplink
monitor application use, classify and mark data packets
Selects bearer according to packet classification

**Figure 6.2.3.2-1: Reference Architecture with PCEF enhanced with ADC**

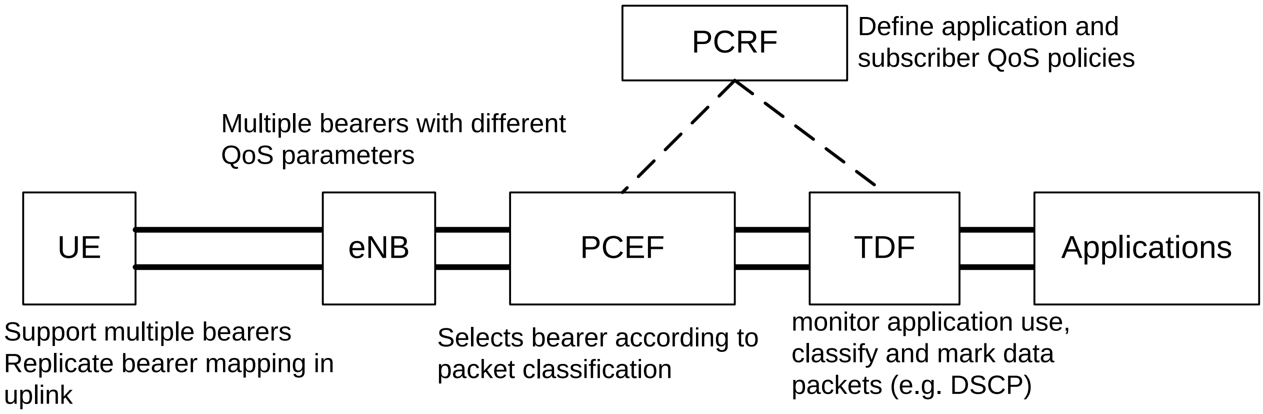
PCRF
Define application and subscriber QoS policies
Multiple bearers with different QoS parameters
UE
eNB
PCEF
TDF
Applications
Support multiple bearers
Replicate bearer mapping in uplink
Selects bearer according to packet classification
monitor application use, classify and mark data packets (e.g. DSCP)

**Figure 6.2.3.2-2: Reference Architecture with TDF**

In order to limit the need for frequent bearer modifications each UE may have a small number of pre-allocated dedicated bearers (at a minimum, one pre-allocated dedicated non-GBR bearer would be needed for selected UEs). In case of PCEF enhanced with ADC, the application detection is done as part of the SDF filter evaluation, which may implicitly entail usage of DPI functionality. In case of TDF, the application detection is provided by the TDF which classifies the packets and applies corresponding marks. The PCEF has SDF filters configured using those marks and the SDF filter evaluation leads to appropriately assigning the marked packets to the pre-established bearers. This can be achieved by using filter rules including ToS classification according to TS 29.212 [7] and marking the packets with DSCPs accordingly.

The allocation/modification of bearers can be further optimized when triggered by subscriber policy which reflects service subscription information; either controlled by the PCRF or pre-defined via local policies in the PCEF. Inactivity timers can be used to remove idle bearers. Dedicated bearers may consume network resources; however with intelligent management the total number of active dedicated bearers can be controlled.

In addition, if the UE performs automatic flow mapping to bearers in uplink direction (which is a new functionality in the UE) allows for reusing the downlink QoS bearer optimization also for uplink congestion mitigation.

| Remote IP | local port | remote port | protocol | DSCP | dedicated bearer | life time state | state |
|---|---|---|---|---|---|---|---|
| 199.239.136.200 | 51452 | 80 | TCP | 12 | 1 | 60s | active |
| 85.183.195.96 | 51455 | 80 | TCP | 12 | 1 | 70s | active |
| 74.125.43.149 | 51459 | 80 | UDP | 12 | 1 | 70s | active |
| 2.18.175.139 | 51470 | 80 | TCP | 14 | 2 | 30s | active |

**Figure 6.2.3.2-3: Example of Flow tracking for automated bearer mapping**

The UE can learn how flows must be mapped to dedicated bearers by simply tracking the flows in downlink direction and assign corresponding packets in uplink direction to the same bearers. The flow table (see example in the figure above) contains all flows detected in dedicated bearers (downlink direction, i.e. mapped by packet core). In uplink packets are mapped according to flow table entries stored in the UE. In that sense each entry emulates an uplink filter, which is not created by signaling, i.e. flow table entries take precedence over TFT filters in the UE. DSCP value is reflected into uplink packets to comply with TFT verification rules in the core network. Flow entries which are aged out can be actively removed by the UE (e.g. TCP FIN packet can trigger flow removal).

Optionally, if the core receives RAN congestion information in band or out band signaling, the information can also be used adjusting the bearer configurations dynamically and at any time, e.g. establishing a new dedicated bearer for certain application traffic.

### 6.2.3.3 Impact on existing entities and interfaces

For subscriber differentiation based on subscription data, the solution doesn't require any standardisation effort in case of DSCP marking usage.

TDF/PCEF:

- For application differentiation, the DPI functionality is required in the network. The DPI functionality can be part of a TDF or a PCEF enhanced by ADC.

- In case of TDF, the derived marking is based on configuration or based on the new parameter received from the PCRF within the corresponding ADC Rule.

UE:

- Needs to support multiple dedicated bearers.

- For uplink congestion mitigation the UE needs to automatically assign packets from certain flows to the corresponding bearers in uplink direction.

Editor's Note: It is for further study, what are the standardization impacts on the UE.

### 6.2.3.4 Solution evaluation

- This solution offers an alternative way to solve key issue #1, i.e. RAN user plane congestion mitigation by re-using and enhancing (e.g. using DPI functionality in the network or improve uplink bearer usage) the existing bearer concept, i.e. no or only minor standardisation effort is required.

- It fully supports congestion handling on subscriber- and application-level.

- Standardized interfaces and procedures for multi-vendor support are re-used. No new interfaces or protocols are required.

- No impacts on RAN foreseen as the existing bearer based QoS control concept are re-used.

- It does not rely on any form of RAN congestion awareness in the core, i.e. no feedback loop is needed and there is no issue with signalling load towards and in the core network. If RAN congestion information is indicated to the CN, bearer usage can be adapted and optimized.

- It works also for fast changing load and congestion situations in RAN. It is much more responsive to congestion and scalable than any feedback-based solution.

- It allows the radio scheduler a full visibility about the traffic demand, so RAN can work in full buffer model and can allocate traffic to available resources according the current radio conditions. It allows the RAN to react on congestion situations without assistance from CN.

- It does not support content-level optimization or adaptation mechanisms, as these are typically building on core network functions. Application-level adjustments would require congestion feedback towards the core network.

- It requires the capability of the UE to support multiple dedicated bearers which is guaranteed within EPS. The number of different prioritisation levels is limited to the UEs capability to support several established dedicated bearers. Furthermore, it depends on operator's bearer configuration policies, e.g. the VPLMN operator might have different bearer policies than the HPLMN operator.

- In order to replicate the optimised downlink QoS control in uplink, the UE is required to perform automatic flow mapping in uplink direction. This requires that the traffic aggregate can be unambiguously indentified by the IP-5-tuple.

- In respect to application detection, this solution has the same implications (i.e. DPI processing load or issues with non-deducible service data flows) as in-bearer marking solutions (e.g. SCI or FPI).

- The proposed multiple dedicated bearer solution allows for re-use of the bearer based QoS mechanism in RAN and CN, thus going beyond pure in-bearer packet prioritisation.

## 6.3 UE-based Solutions for RAN user plane congestion management

### 6.3.1 Solutions for Uplink Congestion Management

### 6.3.2 Solutions for Handling of Unattended Traffic

#### 6.3.2.1 Solution 3.2.1: Unattended traffic limitation in the UE in case of RAN congestion

##### 6.3.2.1.1 General description, assumptions, and principles

This solution addresses part of Key issue #1, in particular, limiting unattended traffic in case of RAN congestion.

Whether an application is running in the foreground or in the background of a device, and therefore whether the traffic the application generates is attended or unattended, is currently only known at the UE. How the UE can detect such traffic is implementation dependent, but techniques may include detecting the UE is not used by the user, e.g. the phone is in a pocket or left on a desk, or detecting applications that are running on the background, e.g., not being displayed to the user.

If the UEs were required to provide to the RAN or CN, for each flow, whether the traffic flow is attended or unattended, this is very likely to produce undesired overhead**.** One possibility is that the UE indicates whether the UE itself is attended or unattended, where all flow are considered attended or unattended respectively, but that would be a very coarse indication and possibly not very useful.

On the other hand, the UE can have the capability of knowing the UE situation (user present/ not present), which application is requesting a connection, and whether the application is running on the background or in foreground (e.g. being displayed to the user).

This document proposes a solution where the UE is responsible for blocking unattended traffic when the network requests it and based on configuration.

##### 6.3.2.1.2 High-level operation and procedures

The solution works on two levels:

- Dynamic indication to UE based on RAN congestion:
  - Network indication to block transmission of certain unattended traffic
  - This indication is dynamic.

Editor's Note: It is TBD whether this indication is an indication of congestion or an explicit indication to not initiate unattended traffic. It is also TBD whether the indication is provided by the RAN or CN. Given ongoing work in RAN2 for 3GPP-WLAN radio interworking, it is FFS whether parts of the design adopted in RAN2 can be reused for this solution.

Editor's Note: Security aspects and network operational impacts of providing such indication need further evaluation.

  - There may also be a time indication of how long to block unattended traffic.

- Configuration in UE:

  - The UE is configured with which applications are subject to being blocked when the NW sends indication above, which application are exempt and optionally default actions for application not explicitly identified.

    - Operators may configure the device, e.g. via OMA DM, using application ID similar as defined for DIDA in TS 24.312 [6] subclause 5.7.

Editor's Note: The details of how the UE is configured are FFS.

The UE behaves as follows. When the UE receives an indication to block unattended traffic, for each application, it checks the configuration for the particular application ID and:

- If the application is subject to being blocked and is identified as unattended, the UE internally blocks uplink traffic generated by the application.

- If the application is exempt from being blocked or is identified as attended, the UE does not block uplink traffic generated by the application.

There is no application impact in this solution.

Editor's Note: It is FFS what the implications are to the applications if the keepalive messages are being blocked when the application is unattended.

Although this solution has direct impact on uplink transmission reduction, it can also reduce traffic load in the downlink. For instance, there are many applications that pull data from the network periodically without user interaction (e.g. e-mail, Facebook, etc.). In that case, the uplink traffic of the request is not large, but potentially the downlink traffic caused by the update may be substantial.

#### 6.3.2.1.3 Impact on existing entities and interfaces

UE:

- Support indication from network.

- Identify and block traffic based on network indication and configuration in the UE.

- Support of new OMA DM configuration.

eNB:

- May have impact depending on how indication is provided to the UE.

MME:

- May have impact depending on how indication is provided to the UE.

S-GW:

- May have impact depending on how indication is provided to the UE.

P-GW:

- May have impact depending on how indication is provided to the UE.

###### 6.3.2.1.4 Solution evaluation

## 6.X Solution X: <Title of Solution>

### 6.X.1 General description, assumptions, and principles

Editor’s Note: This sub-clause should identify the key issues address by this solution.

### 6.X.2 High-level operation and procedures

### 6.X.3 Impact on existing entities and interfaces

### 6.X.4 Solution evaluation

# 7 Evaluation

Editor's note: this clause contains the evaluation of various solutions.

# 8 Conclusions

Editor’s Note: The clause will capture agreed conclusions from the Key Issues and Architecture Solutions clauses.

# Annex A:
Change history

| Change history | | | | | | | |
|---|---|---|---|---|---|---|---|
| Date | TSG # | TSG Doc. | CR | Rev | Subject/Comment | Old | New |
| 2012-11 | SA2#94 | - | - | - | TR skeleton generated for submission at SA2#94 (Approved in S2-124762) | - | 0.0.0 |
| 2012-11 | SA2#94 | - | - | - | Inclusion of documents agreed at SA2#94: S2-124717 and S2-124858.<br>Formatting and Editorial corrections by aligning the clause numbers as per approved TR skeleton in S2-124762.<br>Formatting and Editorial corrections of references. | 0.0.0 | 0.1.0 |
| 2013-02 | SA2#95 | - | - | - | Inclusion of document agreed at SA2#95: S2-130681.<br>Adoption of assigned TR number: 23.705. | 0.1.0 | 0.2.0 |
| 2013-04 | SA2#96 | - | - | - | Inclusion of documents agreed at SA2#96: S2-131539, S2-131491, S2-131358, S2-131399, S2-131400, S2-131492, S2-131493, and S2-131494.<br>Editorial integration of S2-131539 into the structure given by S2-131493.<br>Formatting and editorial corrections. | 0.2.0 | 0.3.0 |
| 2013-06 | SA2#97 | - | - | - | Inclusion of documents agreed at SA2#97: S2-132224, S2-132305, S2-132170, S2-132172, S2-132309, S2-132311, S2-132331, S2-132332, and S2-132333.<br>Formatting and editorial corrections. | 0.3.0 | 0.4.0 |
| 2013-06 | SA2#97 | - | - | - | Update of Version 0.4.0 due to the inclusion of the agreed document S2-132306 that was missed in the update from version 0.3.0 to 0.4.0.<br>Formatting and editorial corrections. | 0.4.0 | 0.5.0 |
| 2013-07 | SA2#98 | - | - | - | Inclusion of documents agreed at SA2#98: S2-132821 (TR restructuring), S2-132822, S2-132962, S2-132973, S2-132976, S2-132977, S2-132978.<br>Formatting and editorial corrections. | 0.5.0 | 0.6.0 |
| 2013-08 | SA2#98 | - | - | - | Correct inclusion of Figure 6.1.5.2.3.2-1. | 0.6.0 | 0.7.0 |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |