3GPP TR 22.805 V2.0.0 (2012-08)

Technical Report

3rd Generation Partnership Project;
Technical Specification Group SA;
Feasibility Study on User Plane Congestion Management
(Release 12)

Keywords
User Plane, Congestion

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.

Introduction

Mobile operators are seeing significant increases in user data traffic. For some operators, user data traffic has more than doubled annually for several years. Although the data capacity of networks has increased significantly, the observed increase in user traffic continues to outpace the growth in capacity. This is resulting in increased network congestion and in degraded user service experience. Reasons for this growth in traffic include the rapidly increasing use of "smart phones" and the proliferation of data applications that they support, and the use of USB modem dongles for laptops to provide mobile or nomadic Internet access using 3GPP networks. As the penetration of these terminals increases worldwide, this trend of rapidly increasing data traffic is expected to continue and possibly accelerate.

Network operators continue to invest in additional network capacity (network entities and connectivity resources) attempting to cope with user data traffic increases that cause user plane congestion. This additional investment is becoming increasingly costly due to the rapid and continuing increases in user data traffic. From a CAPEX or OPEX perspective, this approach is not sufficient. In addition, existing QoS and PCC mechanisms are being deployed but the full effect is still to be seen. It is therefore necessary to study approaches and mechanisms to manage user plane congestion.

1 Scope

This TR considers scenarios and use cases where high usage levels lead to user plane traffic congestion in the RAN, and proposes requirements for handling user plane traffic when RAN congestion occurs. The aim is to make efficient use of available resources to increase the potential number of active users while maintaining the user experience.

Scenarios that will be considered include handling of user plane traffic when RAN congestion occurs based on:

- the subscription of the user;
- the type of application;
- the type of content.

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 23.860: "Enabling coder selection and rate adaptation for UTRAN and E-UTRAN for load adaptive applications; Stage 2"

3 Definitions, symbols 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].

Network: remaining part of the "3GPP system" after excluding the UE.

RAN user plane congestion: the situation where the demand for RAN resources to transfer user data exceeds their capacity to deliver the user data with the expected QoS.

System: "3GPP system" is defined in TR 21.905 [1].

User plane traffic: user data to be transferred between entities connected to the 3GPP network but not used by the 3GPP network for purposes such as set-up and release of data sessions.

3.2 Symbols

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

3.3 Abbreviations

For the purposes of the present document, the abbreviations given in TR 21.905 [x] 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 Scenarios and Use Cases

4.1 Congestion Scenarios

4.1.1 General

What is "user plane" traffic and what is "control plane" traffic?

Examples of "user plane" traffic are keep-alive messages for smart phone applications, TCP synchronization messages, streaming data and HTTP data. "Control plane" traffic includes LTE/EPC-related signalling such as RRC, NAS messages for set-up and release of data sessions, etc.

There is not a 1:1 relationship between the amount of control information and the amount of user data transferred. What is envisaged is that the volume of user plane data may exceed the capacity of the radio technology to transfer user data while the relevant control channels remain uncongested. It is also possible that the volume of control information may exceed the capacity of the control channels while the volume of user data is less than the capacity of the user data channels, e.g., smart phone application keep-alive functions which do frequent set-up and release of data sessions.

In this TR, the focus is on user plane congestion. It is presumed that the control plane is not in congestion. This means that there is available control plane capacity to allow action to be taken to manage user plane congestion.

4.1.2 Identifying user plane traffic using its attributes

This section considers a number of general aspects to provide context for the use cases in subsequent sections.

The aim is to manage user plane traffic when RAN congestion occurs. Therefore, the problem is to select the appropriate user plane traffic flows to be subjected to congestion management. This selection process needs to be based on attributes of the traffic flows.

Approaches to selecting user plane traffic flows to be subjected to congestion management may affect one or more subscribers, one or more applications, or one or more types of traffic.

One approach to managing user plane congestion is to control all the traffic for a given subscriber without further considering the nature of that subscriber’s traffic flows. Operators may choose to offer subscriptions with various levels or tiers of service with differing performance levels when congestion occurs, and use the subscription level to manage user plane congestion by giving higher precedence to users with a higher subscription level.

Therefore, a candidate attribute in identifying the traffic to be managed is the identity of the subscriber sending or receiving it.

Another approach to managing user plane congestion is to control the volume of traffic to and from web sites supporting certain types of applications. Some applications may involve much more data transfer than others. Some applications may have unique traffic profiles peculiar to their functions. For example, social networking web sites may have traffic profiles with higher and lower traffic levels at times of the day different from those for business email or web browsing. Throttling different applications when user plane congestion occurs may be an effective approach for mitigating user plane congestion. Some types of application require near real time handling of traffic (e.g., point of sale terminals) while others may be relatively less time sensitive (e.g., calendar and contact list updates, vending machine reports, email push, etc.) Under user plane congestion conditions, the less time sensitive application traffic should be controlled before the more time sensitive.

Therefore, a second candidate attribute in identifying the traffic to be managed is the type of application.

Another approach to managing congestion is to control certain types of traffic. Some types of traffic inherently demand more network resources and therefore are more impacting on user plane congestion. Web browsing and email may entail short periods of heavy demand but are then likely to be followed by relatively long periods of low or no activity from a user plane traffic point of view. Streaming applications (audio, video, multimedia) are likely to place high demands on the user plane and for longer durations than other types of traffic. An application may involve multiple types of traffic (e.g. a social networking application may involve user browsing among friends' postings, then streaming a video posted by a friend.) Hence this approach may affect some types of traffic for a given application but not others. It may therefore be seen as finer-grained than user plane traffic management controls on an application basis. Controlling this type of traffic, especially when multiple subscribers using this type of user plane traffic are in close proximity to each other, may be an effective approach.

Therefore, a third key candidate attribute in identifying the traffic to be managed is its type.

Each of the three categories of approaches identified, can be applied individually, or combined with one another in a given traffic congestion management situation. For example, a high tier user may be running an application of lesser time sensitivity.

4.1.3 User plane congestion due to full use of cell capacity

When a number of UEs have user plane traffic totalling the cell capacity, and then an additional UE attempts to generate user plane traffic, congestion occurs. This is because the traffic volume exceeds the capacity of the cell. This is illustrated in Figure 1 for UE 3, using example capacities.

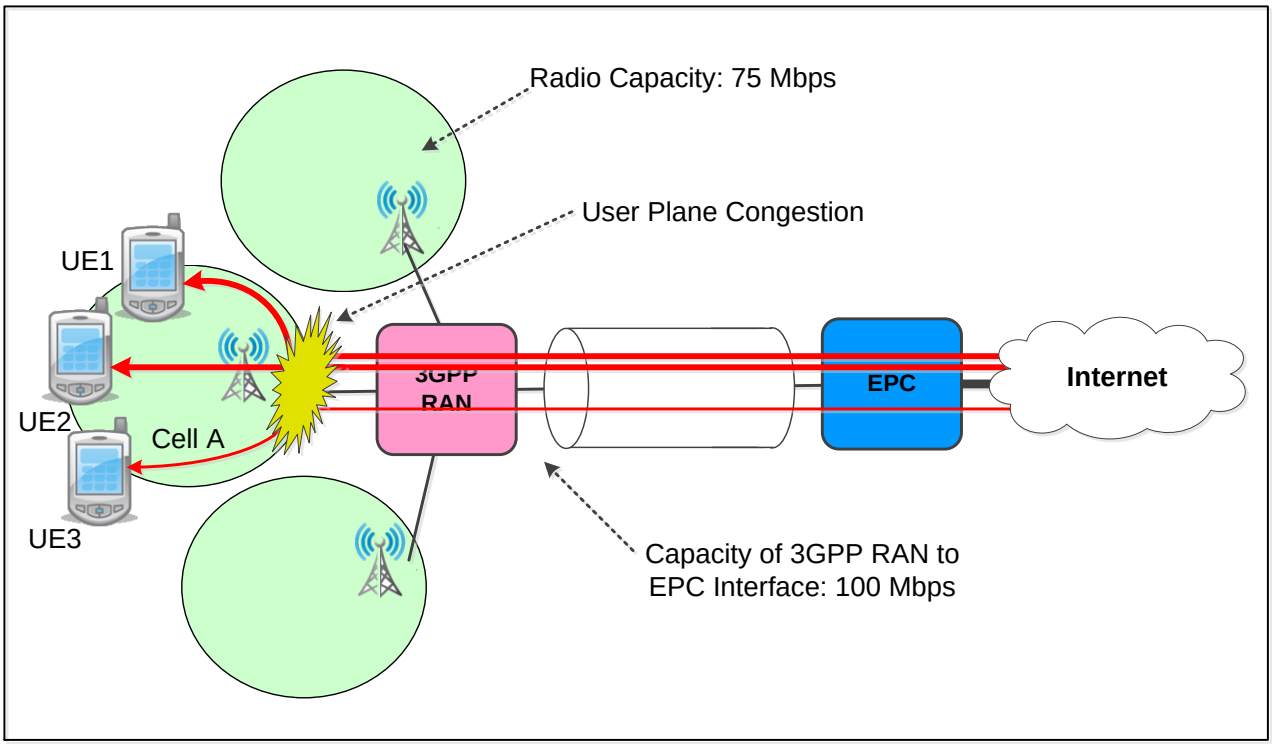

Figure 1 - User plane congestion due to full use of cell capacity (example capacities)

4.1.4 User plane congestion due to 3GPP RAN to EPC interface capacity limitation

When the user plane data volume of all the UEs being served by Cells A, B and C totals more than the actual capacity of the 3GPP RAN to EPC interface, there will be a potential impact on all the UEs involved. This may lead to excessive data rate reduction or service denial. Even though each cell may have the necessary capacity to support the UEs it is serving, the capacity of the 3GPP RAN to EPC interface has an impact on each UE and may in the worst case actually prevent UEs from being offered any capacity at all. This is illustrated in Figure 2.

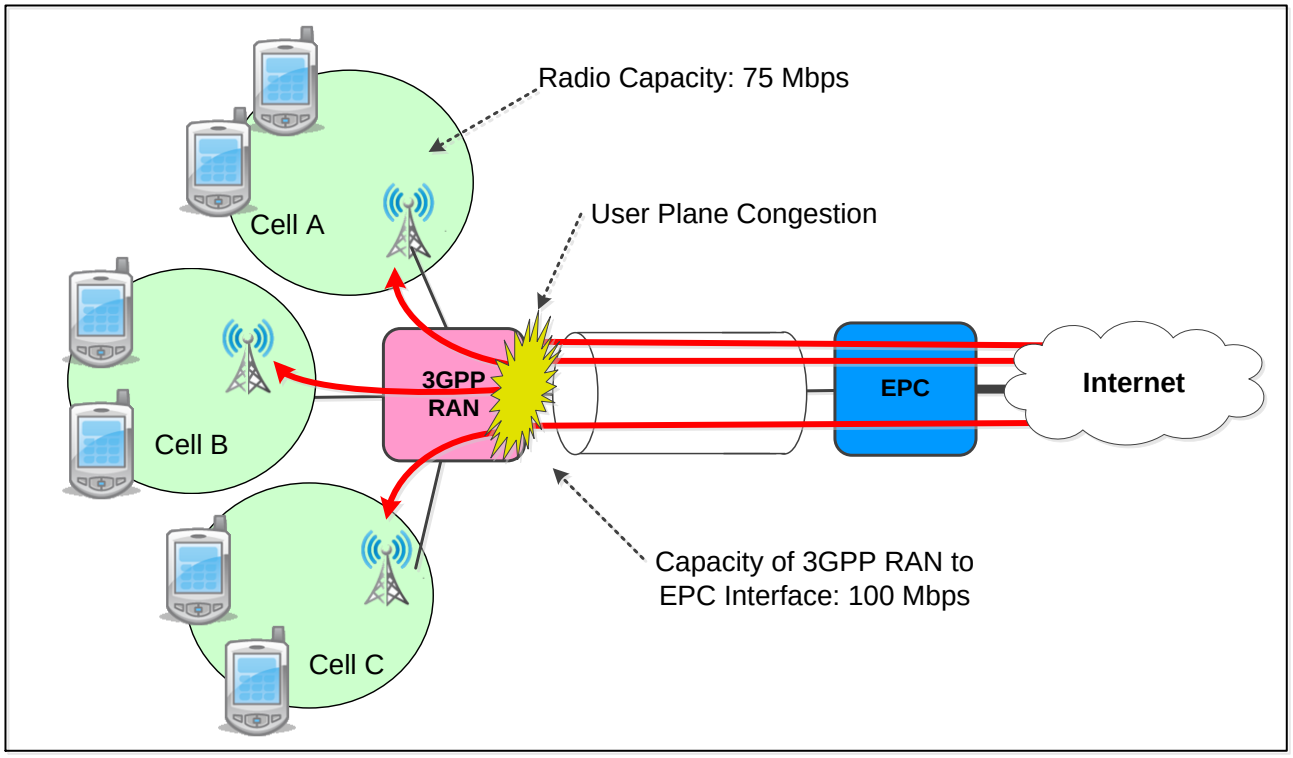

Figure 2 - User plane congestion due to 3GPP RAN to EPC interface capacity limitation (example capacities)

4.2 Use Case 1 – Managing congested RAN traffic using differentiated service subscription QoS attributes

4.2.1 Description

RAN design is resource constrained due to available radio spectrum limitations. This can lead to congestion in crowded cells, e.g. peak hour at train stations, peak business hours in business areas. There is a category of subscribers (e.g. business users) who would be willing to pay extra for a service plan that provides prioritized (higher QoS / MBR) access than other subscribers during congestion.

NOTE: This is very similar to "priority boarding queue" feature provided by airlines to their loyal / higher class of service travellers (first / business class) to avoid waiting and congestion at the boarding gate.

4.2.2 Pre-conditions

Alice and Bob have different subscription service profiles to address congestion situations.

Alice has an expensive platinum subscription plan that allows her "priority" access for mobile broadband Internet access.

Bob has a cheaper subscription plan that allows him "best effort" mobile broadband access.

4.2.3 Service flows

Both Alice and Bob are at the airport lounge waiting to board their flights to the SA1 meeting.

Due to peak time, the (E) UTRAN cell serving the airport / gate is over subscribed and congested.

Both Alice and Bob choose to download the SA1 contributions from ftp.3gpp.org.

Since Alice has a platinum priority service subscription, she gets prioritized treatment over Bob who has a "best effort" subscription.

4.2.4 Post-conditions

Alice downloads 60 MB worth of SA1 documents successfully in 8 minutes, boards her flight, during which she will be able to review the contributions. She is happy that she subscribed for premium service that gives her priority access in congested cells.

Bob is only able to download 5 MB worth of SA1 documents i.e. only $1/8^{th}$ of the contributions. He is not happy and decides to upgrade his service plan to "platinum" after he comes back from the meeting.

4.2.5 Potential requirements

The network shall be able to detect user plane congestion.

When making QoS policy decisions, the network shall be able to take into consideration the RAN congestion level and the subscriber's profile when coping with traffic congestion.

The network shall be able to configure such RAN congestion-based policy rules.

4.3 Use Case 2 - User level traffic control

4.3.1 Description

Operators may provide some flat rate data plans to attract more users or encourage the users to use more network services, especially at the early phase of a new network. Normally these flat rate plans are well-designed to ensure both users' acceptance and operators' revenue. However, there are always a few heavy users as determined by operator policy, e.g. over a billing cycle or based on usage measurements, who consume much more network resources than the others. One method to control this is setting a volume threshold and reducing the data rate if the amount of data exceeds the threshold. However, this will unavoidably reduce these users' experience even though there are plentiful network resources. A compromise to improve the above situation is to reduce the data rate of heavy users only when RAN user plane congestion occurs.

Operators may also apply user level traffic control to roaming users according to roaming agreement or to guarantee local users' experience. In this case, to optimize the roaming users' experience, this user level traffic control, e.g. bandwidth control, is only applied when the roaming users are in congested cells.

4.3.2 Pre-conditions

Alice is a subscriber of operator A. Alice is a heavy user whose total used volume of this month has exceeded, e.g. 2GB. Alice's contract with operator A allows operator A to reduce Alice's maximum data rate to, e.g. 512 kbps (applicable to both uplink and downlink) when the RAN is congested.

Operator A and operator B have signed a roaming agreement. According to the roaming agreement, operator A is allowed to downgrade the maximum data rate for a UE roaming into its network to, e.g. 512 kbps, when the RAN is congested.

Bob is a subscriber of operator B. Bob is now roaming to operator A's network.

4.3.3 Service flows

Cell X of operator A's network is lightly loaded at 8 AM. The current data rate for each user ranges from, e.g. 1 Mbps to 10 Mbps. All the users enjoy good user experience at this time.

At 9 AM, cell X becomes user plane congested as there are more active users and some of the users are using downloading services. Now there are some local users, including Tom and John, whose data rate is lower than, e.g. 128kbps. The user experience under cell X now becomes poor.

The network detects this user plane congestion of cell X and that there are, e.g. 5 heavy users and 10 roaming users, among all active users under cell X. The network downgrades each heavy user's data rate to 512kbps. The network also downgrades each roaming user's maximum data rate to 512 kbps. As a result, the data rate of other local users, including Tom and John, improves to a certain degree, e.g. to more than 512 kbps.

Alice is one of the 5 heavy users in cell X. Bob is one of the roaming users in cell X. When Alice and Bob move to cell Y which is not user plane congested, the network detects this and deactivates their data rate limitation. Alice's data rate goes up to, e.g. 5 Mbps, and Bob's data rate goes up to, e.g. 10 Mbps.

At 11 AM, the network detects that cell X becomes lightly loaded. The network deactivates the data rate limitation of all the heavy users and the roaming users in cell X. The data rate of the heavy users and roaming users under cell X goes up to normal state.

4.3.4 Post-conditions

The efficiency of cell X is improved in the period between 9 AM and 11 AM. The overall user experience in cell X during this period is also improved.

4.3.5 Potential requirements

The network shall be able to detect user plane congested cells.

The network shall be able to identify active UEs accessing the network via the congested RAN.

According to the operator's policies, the network shall be able to select specific users (e.g. heavy users, roaming users, etc.) and adjust the QoS of existing connections or the application of relevant policies for new connections depending on the RAN load status.

4.4 Use Case 3 - Application data rate control

4.4.1 Description

The resources required to provide good user experience vary from application to application. For example, IM applications usually require frequent signalling but little user plane traffic, while P2P applications are extremely aggressive in user plane bandwidth occupation and may largely downgrade the user experience of other applications.

When the RAN is congested due to user plane traffic, operators may want to limit the data rate of some applications such as P2P applications and thereby release some resources for other applications or for more users. In this case, application level traffic control is needed.

In addition, user related information and content type may also need to be considered during above flow. It is possible for the operator to allocate higher data rate of the same applications for the "platinum" user than for the user with cheap tariff plan, or to allocate different data rates of the same applications with different content type, e.g. allocate higher date rates on image transfer during instant messaging and lower date rates for text transfer.

4.4.2 Pre-conditions

P2P downloading applications and streaming applications are defined by the operator as aggressive bandwidth consuming applications and, e.g. 300kbps, is the minimum data rate to guarantee the user experience of streaming applications.

4.4.3 Service Flows

When cell X is congested due to user plane traffic at, e.g. in the period 21:00 to 23:00, the user experience in cell X is poor. The network detects cell X's user plane congestion and enforces the following application level traffic control to optimize the RAN user plane resource usage in cell X:

- the data rate of the identified P2P downloading applications is limited to, e.g. 128kbps for a UE of a user with a low cost tariff plan, and 512kbps for a UE of a user with a "platinum" tariff plan;
- the data rate of each identified streaming application is limited to, e.g. 300kbps for the UE of a user with a low cost tariff plan, and no downgrade for a UE of a user with a "platinum" tariff plan;
- the radio resource allocated of each identified IM application is limited to, e.g. 64kbps for a UE with when using text transfer and 128kbps for a UE when using image transfer.

After an hour, the network detects that cell X becomes lightly loaded again at 23:00. The above application level control is deactivated for cell X. As a result, the P2P applications and the steaming applications' maximum data rate goes up to the users' subscribed maximum data rate.

4.4.4 Post-conditions

The efficiency of cell X is improved during the period 21:00 and 23:00. The overall user experience in cell X during this period is also improved.

4.4.5 Potential requirements

According to the operator's policies, the network shall be able to select specific applications and control the data rate of the identified applications based on RAN load status, at the same time taking into consideration the user related information (e.g. a "platinum" subscription user should have good experience even in case of congestion) and content type (e.g., text vs. image.)

4.5 Use Case 4 - Disaster service priority

4.5.1 Description

In this use case, a specific communication service is allocated resources preferentially while a cell is congested due to high data traffic volume during a disaster situation.

4.5.2 Pre-conditions

When disaster occurs, operator activates "disaster message board" (DMB) service. Operator may announce activation of such service, e.g. via disaster alerts broadcast via SMS in the affected area, or by other means. As one example of DMB (other ways are plausible), this service is structured to provide priority to any wireless traffic to and from a specific web site, which routes messages to and from the disaster area. To prevent abuse, messages may be limited to text or certain payload size. It should be noted that DMB is not a priority service such as eMPS, which requires subscription. In contrast to eMPS, DMB service applies to all UEs.

Bob and Alice are Nancy's parents. Their mobile addresses are known to each other (e.g. telephone number, mail address, etc.)

Nancy lives away from her parents while attending university.

A disaster occurs around the area where Nancy lives.

People who are living in the disaster area try to inform their relatives about their safety. There are also some people in the disaster area who are able to take videos of the disaster situation and are uploading these videos on their blogs. However, due to these actions, there is a considerable increase in data traffic causing the network to become severely congested, which prevents Nancy and many other people in the disaster area from contacting their relatives to inform them about their safety.

4.5.3 Service Flows

When the network becomes congested, activation of DMB effectively reduces the bandwidth available to users uploading video files to their blogs since some network resources in the affected area are used for prioritized disaster message board messaging, allowing users to reach their relatives without difficulty.

Therefore, when Nancy tries to inform her parents about her safety via the disaster message board service, she can be guaranteed sufficient resources to send the message.

4.5.4 Post-conditions

Bob and Alice can confirm Nancy's safety via the disaster message board.

4.5.5 Potential requirements

The following requirements are identified for this use case:

- The network shall be able to identify specific high priority communications (e.g. related to a disaster message board service.)
- If the RAN is congested, the network shall be able to (re-)allocate resources to such communications.
- During RAN congestion, the operator shall be able to select the communications which require preferential treatment and allocate sufficient resources for such communications in order to provide these services with appropriate service quality.
- During RAN congestion, the system may allow continuation of non high priority communications with possibly reduced resources.

4.6 Use case 5 – Managing congested traffic by using appropriate application attributes

4.6.1 Description

The RAN is resource constrained due to available radio spectrum limitations. This leads to frequent congestion in crowded cells, e.g. at peak hours at train stations, at peak business hours in business areas. The majority of mobile broadband traffic utilizes either a primary PDP context (for GPRS) or a default bearer (for EPC) using a background service class . Subscribers use applications such as social networking, Over-the-Top (OTT) audio/video, blogging, internet games, FTP, software patches and updates, etc.

When the RAN is congested due to user plane traffic, operators may want to limit the data rate of some applications such as software patches and updates, with the consideration of user level or content type, as specified in section 4.4, use case 3 Application data rate control.

4.6.2 Pre-conditions

Alice, Bob and Cindy are in the same congested cell/sector.

Alice is using her smart phone to check her social networking application status, browse friends' photos and to subsequently perform a social network application check-in at the local coffee shop.

Bob is using a mobile broadband USB stick with his laptop. His laptop starts an automatic download of a new version of a large software application or an upgrade patch, e.g. of 120 MB.

Cindy is using her smart phone to her social networking application to browse friend text blog.

4.6.3 Service Flows

Alice and Bob are accessing mobile broadband communications using different applications. Alice is using her smart phone to access a social networking application while Bob's laptop is automatically downloading a large software update.

During traffic peak time the (E)UTRAN cell serving their location is congested.

Since Alice is using an interactive social networking application her data packets are prioritized over Bob's large software application patch.

User level and content type may also need to be considered in the above case. If Bob is subscribed as the "platinum" user while Alice is subscribed with cheap plan, it is possible that Bob's large software application patch update is prioritized over Alice's social network application. In addition, although Cindy and Alice perform the same application, Cindy's service content requires less resource and therefore Alice's data packets are transmitted at higher overall data rate than Cindy's data packets.

4.6.4 Post-conditions

Alice, Bob and Cindy are all fully satisfied with their mobile broadband service. Although Bob's software update download is slower than it would have been absent congestion, since it's a background non-interactive application, Bob is not perceptibly impacted.

4.6.5 Potential requirements

The network shall be able to detect user plane congested cells.

The network shall be able to identify, differentiate and prioritize different applications' traffic such as social networking, OTT audio/video, blogging, internet games, FTP, software patches and software downloads, etc., based on the QoS attributes of the communications of these applications. User related information and content type also need to be considered together in such prioritization.

4.7 Use case 6 - Content delivery based on RAN congestion status

4.7.1 Description

Certain mobile content delivery based (also known as "Push") services are not time sensitive. Such a service may occur at a designated time or periodically. It is possible that localized congestion exists in the part of the RAN in which a UE is residing at the time when a service for its subscriber is planned to take place. The network would want to control when and when not to activate such services to the UE in order to avoid further congestion. When localized congestion in the RAN due to user plane traffic occurs, the operator may want to delay the service for those users in the congested area until such users move to an uncongested part of the RAN or the affected part of the RAN becomes uncongested. Moreover, data to be delivered by a service may have a period of validity. For example, a coupon delivered by a Push service is only valid during a designated period.

A third party application provider offering Push services is expected to include with the data to be pushed a validity period/"best before time" (i.e. a point in time after which delivery should no longer be attempted and the user plane Push content is discarded.) Variations may include no expiry time (information is valid even if there is a long delay in delivery.) In return, the operator may provide information to the third party service provider on when the Push data is delivered or if delivery fails. These variations may be considered "added value" and hence provide opportunity for network operators to gain a revenue stream in addition to the basic Push service.

4.7.2 Pre-conditions

Mike is subscribed to a third party Push advertisement service from a restaurant that provides coupons to him in the format of an image or video clip each day and these coupons are valid during the period from 18:00 to 20:00 the same day.

Alice and Bob are subscribed to operator X's Push service which pushes the top 5 popular video clips to them every day.

Cindy has subscribed to a Push newspaper service. This service provides the subscriber with the latest news by MMS twice each day. In order to enable subscribers to get the latest news, the network sends the newspaper to subscribers during a pre-specified time period. For example, the morning newspaper is usually sent between 07:00 and 09:00. The night newspaper is sent between 18:00 and 20:00.

4.7.3 Service flow

Mike's subscribed Push service usually pushes his coupons to him between 11:00 and 12:00. However, during this period, the part of the RAN where Mike's UE resides is congested, therefore the network instructs the third party Push service to pause Mike's service.

When the congestion abates, the network instructs the third party Push service to resume Mike's service. If the period of validity of the coupons has not expired, the Push service will deliver the coupons to Mike's UE. If the period of validity expired by the time the congestion abates, the push service may be cancelled rather than delivering the out-of-date

coupons. In this case, it should be noted that the cancellation only happens this day. In this service, some measures can be taken to efficiently manage the delivery of delay tolerant services that have a period of validity when the part of the RAN where Mike's UE resides is congested.

Alice powers on her UE in the morning in her apartment. The network detects that Alice's UE is in an uncongested part of the RAN. The Push server pushes today's popular video clips to Alice's UE.

Bob powers on his UE in the morning when he arrives at a subway station. The network detects that Bob's UE is in a congested part of the RAN, and pauses Bob's Push service. After Bob arrives at his office, the network detects that Bob's UE is in an uncongested part of the RAN and resumes Bob's Push service which delivers today's popular video clips to Bob's UE unless the service's expiry time has been reached.

During the pre-specified time period of sending the morning newspaper, the part of the RAN in which Cindy's UE resides is congested in the user plane. The network cancels delivering the morning newspaper to Cindy's UE.

4.7.4 Post-conditions

Mike receives his daily coupons via the push service before the period of validity expires even if congestion occurs but only if the congestion abates before the coupons expire. Alternatively, if the congestion does not abate before the coupons expire the push service to Mike's UE is cancelled to avoid unnecessary traffic in the RAN. Mike does not receive expired coupons.

The video clips are pushed to Alice and Bob's UEs when they are in uncongested cells unless the service's expiry time has been reached.

When the RAN congestion abates, if the time is still between 07:00 and 09:00, the network will allow Push of the morning newspaper to Cindy's UE. If the time is past 09:00, the network cancels the Push of the morning newspaper to Cindy's UE for today as the service's expiry time has been reached.

As traffic is more favourably distributed, or content is not delivered in some cases, congestion is eased to a certain degree.

4.7.5 Potential requirements

The network shall be able to detect cell congestion onset and abatement.

The network shall be able to identify whether the Push service target UE is in a user plane congested cell or not.

The network shall be able to provide a mechanism to defer its own Push services based on the cell congestion status and the operator's policy (e.g. the pre-specified time period of pushing.)

The network shall be able to inform third party provided Push services to defer until advised otherwise Push services based on the cell congestion status of the target UE's location and the operator's policy (e.g. the pre-specified time period of pushing).

4.8 Use Case 7 - Traffic compression and transcoding

4.8.1 Description

With the rapid growth in the number of smart phones, the data traffic generated by users accessing the Internet to support the applications on their UEs (e.g. smart phones) is increasing dramatically. Operators are facing major challenges in maintaining their networks' performance. In addition, overload from data services traffic creates risks to the ability of operators to maintain voice service quality. Therefore, in order to reduce the pressure on networks and relieve the problems of traffic congestion, operators may expect their networks to provide the capability to reduce load over the RAN, for example by compressing or transcoding traffic into a format that requires less bandwidth before sending it to UEs.

Web pages may consist of multiple types of content such as text, images, audio and video. Data traffic compression means content compression, such as text compression, image transcoding, etc.

Operators may expect the network to provide traffic compression using common compression/ transcoding mechanisms. Operators may also expect the network to configure flexible traffic compression based on the following attributes:

- User (category such as gold, silver, bronze; identifier)
- Application
- UE parameters (e.g. screen resolution)
- Time (e.g. time of day)
- Date (e.g. New Year's Eve)
- Location (e.g. stadium, shopping centre, etc.)
- Access network type

4.8.2 Pre-conditions

The RAN user plane is not congested. Pre-conditions for Alice are:

- Alice is a gold subscriber;
- Alice's UE has a high resolution screen;
- Alice's UE has an active data session with the mobile network; and
- Alice's UE receives content without any network treatment, e.g. network compression or adaptation.

Pre-conditions for Bob are:

- Bob is a silver subscriber;
- Bob's UE has a high resolution screen;
- Bob's UE has an active data session with the mobile network; and
- Bob's UE receives content without any network treatment, e.g. network compression or adaptation.

Pre-conditions for Cindy are:

- Cindy is a silver subscriber;
- Cindy's UE has a low resolution screen;
- Cindy's UE has an active data session with the mobile network; and
- Cindy's UE receives content without any network treatment, e.g. network compression or adaptation.

4.8.3 Service Flows

The RAN becomes congested in the user plane.

The network identifies data traffic and enables the traffic load reduction capability, e.g. compression or adaptation to a codec/format with lower date rate to alleviate the congestion.

Alice's UE starts to access an image/video stream from a social networking website. Her UE receives the original high-quality image/video stream from the network.

Bob's UE accesses the same image/video stream. Based on Bob's silver subscription the network transcodes the image into a medium quality image e.g. lower colour depth but retaining the same resolution, same colour depth but lower resolution, or adapts the streaming codec to a lower data rate. His UE receives the transcoded medium-quality image/video stream.

Cindy's UE accesses the same image. As Cindy also has a silver subscription and additionally a low resolution screen UE, the network transcodes the image for her into a low quality image e.g. to a resolution appropriate for the screen of her UE, or adapts the steaming codec to a lower date rate. Her UE receives the transcoded low-quality image/video stream.

4.8.4 Post-conditions

The RAN user plane congestion abates. Traffic compression is discontinued.

Alice, Bob and Cindy receive uncompressed data.

4.8.5 Other Service Flows

4.8.5.1 Time-based compression

At 1400, the network enables traffic compression capability automatically.

Alice accesses an image on a social networking website using her UE. The network sends her UE a high-quality compressed image.

4.8.5.2 Location-based compression

At 18:00, Bob goes to the Olympic stadium to watch a football game. Traffic compression is enabled by default in the stadium area.

Bob accesses an image from a social networking website using his UE. The network sends his UE a medium-quality compressed image.

4.8.5.3 Application-based compression

Cindy stays at home to watch an on-line movie on a "TOP 20" web site. When Cindy's UE accesses the video, the network sends her UE a medium-quality compressed video.

4.8.6 Potential Requirements

The requirements derived from this use case are:

- The network shall be made aware of the RAN user plane congestion state;
- Based on RAN load status, per operator policies the network shall be able to reduce the traffic load (e.g. by compressing (like HTTP 1.1 web content into gzip format, transcoding a 16 bit TIFF image into an 8 bit TIFF image, etc., or, adapting the codec for video streaming(which may be lossy or lossless)), taking into account the UE related information (e.g. UE capabilities, subscription) to relieve RAN user plane congestion.

4.9 Use Case 8 - Servicing data connection requests/reactivations

4.9.1 Description

With the current situation where users are able to download and install multiple "data hungry" mobile applications ("apps") onto their UEs, and with the creators of such apps either prioritising their service/solution over the integrity of the mobile network or just generally being unaware of the perils of frequently transferring data, regardless of how small it is, mobile operator networks are facing an increasing data load and the corresponding challenges in meeting demand.

In order to reduce the pressure on networks and relieve the problem of traffic congestion in this scenario, it is proposed that the UE informs the network why a data connection is being requested (either a new connection or an existing one moving from idle to active) based on whether or not the user himself/herself actually initiated the request. With this information, the operator can then treat the data connection activation request differently e.g. deny the connection

request indefinitely, deny the connection request temporarily, accept the connection request anyway, etc. This creates the possibility to increase the likelihood of connections for data traffic of which the user is aware as he/she initiated it (known hereafter as Attended Data Traffic) being successfully established and having good through-put. This in turn has the benefit of increasing the user's perception of the state of the network, since the user is more likely to notice Attended Data Traffic not flowing than Unattended Data Traffic (i.e. data traffic that the user did not directly initiate) not flowing.

NOTE: An app knows when a request for data is user initiated or when the request is not intended for immediate rendering (on the screen), thus it could provide an indication when it requests data from the network. Additionally or alternatively, the UE could categorize connection requests into Attended or Unattended based on one or more of the following criteria:

- whether the screen/keyboard lock is activated
- how long since the user last pressed a key or touched the touch screen
- the app informing the lower layers of the UE (e.g. as part of the UE's API)
- known type of the app (for instance, an app monitoring a user's health – "mHealth" app – may need its data always treated as Attended Data Traffic)
- etc.

Operators may expect the network to differentiate data connection activations for Unattended Data Traffic and Attended Data Traffic for all users or a subset thereof based on the onset of congestion in the RAN and possibly one or a combination of the following criteria:

- User subscription (e.g. gold, silver, bronze)
- Time (e.g. time of day)
- Date (e.g. New Year's Eve)
- Location (e.g. stadium, shopping centre, etc.)
- RAT type

4.9.2 Pre-conditions

The RAN is not congested.

Pre-conditions for Alice are:

- Alice is a gold subscriber;
- Alice's UE has a social networking app installed that is set to check for friends' status updates every 10 minutes;
- Alice's UE is able to inform the network when a data connection is needed specifically due to an app's automatic update; and
- Alice's UE has the screen/keyboard lock on and is in her handbag.

Pre-conditions for Bob are:

- Bob is a silver subscriber;
- Bob's UE has a social networking app installed that is set to check for friends' status updates every 10 minutes;
- Bob's UE is able to inform the network when a data connection is needed specifically due to an app's automatic update; and
- Bob's UE has the screen/keyboard lock on and is in his pocket

Pre-conditions for Cindy are:

- Cindy is a silver subscriber;
- Cindy's UE has a social networking app installed that is set to check for friends' status updates every 10 minutes;

- Cindy's UE is able to inform the network when a data connection is needed specifically due to an app's automatic update; and
- Cindy is using her UE to browse the web

4.9.3 Service Flows

The RAN becomes congested.

Based on Alice's gold subscription, Alice's UE continues being able to establish a new data connection (or reactivate an idle data connection) to enable the social networking app to fetch friends' status updates every 10 minutes i.e. her service remains unaffected.

Based on Bob's silver subscription, Bob's UE discontinues being able to establish a new data connection (or reactivate an idle data connection) to enable the social networking app to fetch friends' status updates. If an active data connection exists already, then the social networking app should be able to fetch friends' status updates until such time as it is torn down or goes into idle.

Cindy continues to be able to browse the web and the social networking app on Cindy's UE continues being able to fetch friends' status updates every 10 minutes, until such time as she finishes her web browsing session (e.g. closes the browser app) and the data connection is torn down or goes into idle.

4.9.4 Post-conditions

The RAN congestion abates.

Alice, Bob and Cindy's status updates were perceptible to them did not suffer noticeably.

The volume of status update messages was lower than it would have been had the updates been conducted without regard to the attendance status of these users. Thus, the impact on network load is lessened.

4.9.5 Other Service Flows

4.9.5.1 Time- and date-based trigger

At 23:30 on 31st December 2012, the network stops serving requests for data connection establishments/reactivations for Unattended Data Traffic based on user category.

Alice's, Bob's and Cindy's UEs are subject to the same handling as in 4.9.3.

4.9.5.2 Location-based trigger

Bob goes to the Olympic stadium to watch the women's shot-put finals. Requests for data connection establishments/reactivations for Unattended Data Traffic are not serviced by default in the stadium area, based on user category.

Alice's, Bob's and Cindy's UEs are subject to the same handling as in 4.9.3.

4.9.5.3 Delay rather than prohibit data connection requests for Unattended Data Traffic

Alice's and Cindy's UEs are subject to the same handling as in 4.9.3.

Based on Bob's silver subscription, Bob's UE discontinues being able to establish a new data connection (or reactivate an idle data connection) every 10 minutes but is allowed every 20 minutes, thus enabling the social networking app to fetch friends' status updates less frequently. If an active data connection exists already, then the social networking app is able to fetch friends' status updates at the frequency desired i.e. every 10 minutes, until such time as it is torn down or goes into idle.

4.9.6 Potential Requirements

The network shall be able to provide mechanisms to detect RAN congestion onset and abatement.

The UE shall be able to inform the network whether a request for a data connection establishment/reactivation is for Unattended Data Traffic or Attended Data Traffic.

The network shall be able to identify whether a UE is in a congested cell and if so, based on operator policy (which may include subscriber consent), the system shall be able to prohibit or delay all or a particular selection of requests for data connection establishments/reactivations for Unattended Data Traffic from that UE.

According to the operator's policies, the system shall be able to identify specific applications requesting certain types of network resources (e.g., P2P, streaming data) and use their status to make decisions on controlling the data rate of the identified applications depending on RAN congestion status.

4.10 Use Case 9 - Voice and video media quality modification

4.10.1 Description

With the current operator push for increased bit rates to provide the likes of wideband voice (sometimes marketed as "HD Voice") and high definition (HD) video, the data traffic generated by such media is likely to increase as more and more UEs and operator networks become available to support it. As stated in clause 4.8.1, operators are facing major challenges in maintaining their networks' performance, therefore maintaining the ability to provide end user acceptable voice and video service quality is paramount. In order to reduce the pressure on networks and relieve the problem of traffic congestion, operators may expect their networks to provide the capability to renegotiate voice and video media into a format of a lower quality that requires less overhead e.g. lower bit rate. Such renegotiation can result in the existing codec continuing to be used (but at a lower bit rate) or a new one be used. In addition, the network should also include the flexibility to differentiate and independently reduce the bit-rate between the voice and video media in order to preserve the audio quality or connections for certain class of service plans, e.g. non-GBR.

Operators may expect such media quality renegotiation capability is not only used for voice /video call service, but also for packet-switched streaming service (PSS), e.g. video-on-demand (VOD.)

Operators may expect the network to renegotiate voice and video media based on one or more of the following attributes:

- User (category such as gold, silver, bronze)
- UE parameters (e.g. codecs supported)
- Time (e.g. time of day)
- Date (e.g. New Year's Eve)
- Location (e.g. stadium, shopping centre, etc.)
- RAT type

4.10.2 Pre-conditions

The RAN is not congested.

Pre-conditions for Alice are:

- Alice is a gold subscriber;
- Alice's UE supports one or more wideband voice/video codecs in addition to one or more SD voice/video codecs; and
- Alice's UE has an active wideband voice or HD video call with her friend Cooper.

Pre-conditions for Bob are:

- Bob is a silver subscriber;
- Bob's UE supports one or more wideband voice/video codecs in addition to one or more SD voice/video codecs; and
- Bob's UE has an active wideband voice or HD video call with his friend Dylan.

Pre-conditions for Cindi are:

- Cindi is a silver subscriber;
- Cindi's UE supports one or more SD voice/video codecs but no wideband/HD codecs; and;
- Cindi's UE has an active SD voice or SD video call with her friend Lauper.

Pre-conditions for Dave are:

- Dave is a bronze subscriber;
- Dave's UE supports one or more SD voice/video codecs but no wideband/HD codecs; and
- Dave's UE has an active SD or SD video call with his friend Brubeck.

Pre-conditions for Tom are:

- Tom has subscribed a VOD service;
- Tom's UE supports one or more wideband voice/video codecs in addition to one or more SD voice/video codecs; and
- Tom's UE is accessing a HD video of VOD service.

NOTE 1: It is not important in the above as to which party initiated the call, nor is it important as to whether or not Cooper, Dylan, Lauper and Dave are in the same cell as Alice, Bob, Cindi and Brubeck, respectively.

NOTE 2: It is accepted that the same codec may support both SD and wideband media flows.

4.10.3 Service Flows

The RAN becomes congested.

The network identifies the wideband voice/video call media and enables codec renegotiation for these calls [2].

Based on Alice's gold subscription, Alice is allowed to continue her call with Cooper using the originally negotiated wideband codec/media.

Based on Bob's silver subscription, Bob's call with Dylan is renegotiated to use an SD codec/media but his call still remains active.

Cindi also has a silver subscription but since her call is already using an SD codec/media, her call with Lauper continues as is i.e. using the originally negotiated SD codec/media.

Based on Dave's bronze subscription, Dave's call with Chris is (re-)negotiated so that the video service bit rate is reduced to zero but his voice call still remains active.

Tom's UE is allowed to continue accessing the VOD service but using an SD codec/media.

NOTE: Operators can apply policy to originating calls, terminating calls, or both. In the case of applying to both, the most oppressive policy will apply.

4.10.4 Post-conditions

The RAN congestion abates.

Alice, Bob and Cindi continue their respective calls.

Dave’s video service bit rate is reduced so that he can continue his voice call with Brubeck.

Tom’s UE is allowed to continue accessing the VOD service using an HD codec/media.

Due to the lower data rate for an SD codec compared to an HD codec, Bob’s call with Dylan and other similar calls are renegotiated to the SD rate help reduce RAN congestion.

Silver subscribers such as Bob and Dylan perceive lower voice quality on the SD call compared to an HD call. On the other hand, gold subscribers such as Alice continue to enjoy HD call quality.

4.10.5 Other Service Flows

4.10.5.1 Time- and date-based renegotiation of codec/media

At 23:30 on 31st December 2012, the network enables renegotiation of codec/media capability automatically.

Alice's, Bob's and Cindi's calls to Cooper, Dylan and Lauper are subject to the same handling as in 4.10.3.

4.10.5.2 Location-based renegotiation of codec/media

Bob goes to the Olympic stadium to watch the women's shot-put finals. Renegotiation of codec/media is enabled by default in the stadium area.

Bob's call to Dylan is subject to the same handling as in 4.10.3.

4.10.6 Potential Requirements

The network shall be able to provide mechanisms to detect the RAN congestion onset and abatement.

The network shall be able to identify whether a UE that has an on-going real-time communication, e.g. voice/video call, PSS, is in a congested cell and if so, based on operator policy (which may include subscriber consent), the network shall be able to instruct the UE to renegotiate the communication media parameters of such real-time communications so that they consume lower bandwidth, e.g., a lower bit rate or a different codec type. The system shall also be able to distinguish between the voice and video portion of the call; (re-)negotiate either one separately to a lower bit rate. The network shall allow the UEs to restore the original media parameter set, e.g. higher bit rate, if RAN congestion has abated.

The network shall also be able to distinguish between the voice and video portion of the call and (re-)negotiate the media parameters of either one separately to, e.g., lower bit rates or different codec types.

The network shall be able to notify the PSS Server of the changed media codec.

4.11 Use case 10 - Charging policy based on RAN congestion status

4.11.1 Description

Charging policy based on RAN congestion status refers to ability to monitor user plane traffic across all cells in the network, and when a cell is congested, the network can accordingly raise the service rate in congested cell. If a particular cell has spare capacity then a lower service rate is offered in that cell.

Charging policy based on RAN congestion status is a capacity of operators to manage the traffic in the network. To some extent, it can balance the traffic load between congested cells and lightly loaded cells. With rapidly changing usage patterns and traffic patterns, adding capacity to the network to prevent peak-usage congestion isn’t a cost-effective. This different charging policy influences subscriber demand and behaviour, making it possible to manage congestion without increasing supply.

4.11.2 Pre-conditions

The user plane status is divided into multiple levels. (E.g. three levels: light level, heavy level, and congested level)

The network shall be able to detect the congestion level of cells across the network.

The network shall be able to identify active UEs accessing the system via congested cells.

The network shall be able to identify active UEs accessing the system via lightly loaded cells.

4.11.3 Service Flows

Cell X is lightly loaded at 8 AM. As cell X becomes user plane congested at 10 AM, the user experience will become poor. The system informs users that data service rate will be higher and starts charging users at the higher service rate.

On another day at 11 AM, the system detects that cell X's user plane is lightly loaded. The system informs users that there is a low service rate and starts charging users at the lower rate.

4.11.4 Post-conditions

The users are charged based on RAN congestion status. Some users postpone the download of large files in congested cells. The risk of user plane congestion is potentially lower and average network utilization improves.

4.11.5 Potential Requirements

The network shall be able to detect the congestion level of cells across the network.

The network shall be able to identify active UEs accessing the system via congested cells.

The network shall be able to identify active UEs accessing the system via lightly loaded cells.

Based on operator strategy, the network shall be able to provide mechanism to support different charging rates based on RAN load status.

4.12 Use Case 11 - Multiple applications traffic control over one bearer

4.12.1 Description

The majority of mobile broadband traffic utilizes primary PDP context (for GPRS) or default bearer (for EPC) for various applications data transmission, like social networking, video, blogging, internet games, FTP, software patches and updates, etc. All the packets flows for those applications share the same QoS treatment for the primary PDP context or the default bearer.

4.12.2 Pre-conditions

Alice is using her smart phone to check her social networking application status, browse friend photos and to subsequently perform a social network check-in application at the local coffee shop. At the same time, her smart phone starts an automatic download of a new version of a large software application or an upgrade patch of, e.g. 120 MB.

Alice's smart phone only establishes one bearer (e.g. the primary PDP context or the default bearer) to transmit all the packets for both social networking application and software upgrading application.

4.12.3 Service Flows

Due to peak time the (E)UTRAN cell serving her location is congested.

Due to congestion, the data rate on Alice's data sessions is reduced.

The network detects different applications are running and reduce the data rate for downloading services but maintain the data rate of social networking applications.

4.12.4 Post-conditions

Only software upgrading application is impacted due to the network congestion control, data transfer becomes slow, unacceptable delay for social network check-in.

Alice is happy as she can enjoy reasonable performance of her social networking app.

4.12.5 Potential requirements

The network shall be able to identify, differentiate and prioritize different applications with same QoS attributes such as social networking, video, blog, internet games, FTP, software patches and updates, etc.

4.13 Use Case 12 - Application Data Priority Control

4.13.1 Description

Operator may want to offer higher priority treatment in the network for the applications which are owned by the operator than those owned by service providers, even the type for the applications is the same, e.g. the IM application belonging to the operator might have higher priority than those belonging to the 3rd party. The user experience should be enhanced for those applications with higher priority.

4.13.2 Pre-conditions

Alice and Bob are at the same cell X.

Alice has installed the IM application A in her smart phone. The IM application A is owned by the operator (either developed by the operator or offered as a special agreement between the operator and the service provider.)

Bob has installed the IM application B in his smart phone. The IM application B is owned by the 3rd party. Normally, IM application B's data are transferred in OTT style with best effort.

Alice is using IM application A to communicate with her friend, Mike.

Bob is using IM application B to communicate with his friend, Rose.

4.13.3 Service Flows

When cell X is congested due to user plane traffic.

The network prioritises Alice's IM application and reduces the bandwidth for Bob's application.

4.13.4 Post-conditions

Alice enjoys communication with Mike without any adverse impact due to cell X congestion.

Bob has to experience message delays, and might decide to install the IM application A instead of IM application B for better user experience.

4.13.5 Potential requirements

According to the operator's policies, the network shall be able to set different priorities per application, e.g. on the basis of operator ownership of the application, and shall be able to control application data traffic according to the priority when RAN is congested.

4.14 Use Case 13 - Protocol Optimization

4.14.1 Description

In order to reduce the pressure on networks and relieve the problems of traffic congestion, the operator may, when the RAN congestion happens, provide the capability of optimizing application protocol to reduce application data before sending it to UEs.

The operator may enable the network to provide common protocol optimization capabilities, e.g. HTTP Multi-Part Response, HTTP Pipelining, Robust Header Compression (ROHC), etc.

The operators may enable the network to optimize different types of protocol, like HTTP, RTSP/RTCP, etc.

However, it may not be beneficial to turn on such protocol optimization all the time as there might be other trade-off considerations. For example, a protocol optimization may cause negative effects on the UE side, such as increased power consumption.

In order to keep the balance between UE power consumption and network load, the operator can decide when to enable protocol optimization capability.

4.14.2 Pre-conditions

The RAN user plane is not congested.

Pre-conditions for Alice:

- Alice's UE has an active data session with the mobile network;
- Alice's UE supports common protocol optimization mechanisms, e.g., HTTP Multi-Part Response capability, HTTP Pipelining;
- Alice is downloading web pages to her UE without network protocol optimization.

4.14.3 Service Flows

The RAN becomes congested in the user plane.

If the network enables the protocol optimization capability, when Alice's UE starts to access a social networking website, the application protocol is optimized between the network and her UE, e.g., Alice's UE fetches the full web page with the main text and all images within a single pipelined HTTP connection.

If the network doesn't enable the protocol optimization capability, when Alice's UE starts to access a social networking website, the application protocol is not optimized between the network and her UE.

If the network doesn't enable the protocol optimization capability, the web page download will result in unreduced data being transferred to the UE, hence exacerbating RAN congestion.

4.14.4 Post-conditions

Having enabled protocol optimization capability, the web page download results in less data being transmitted to the UE, thus lessening RAN congestion.

When congestion abates, protocol optimization is discontinued.

4.14.5 Potential Requirements

The requirements derived from this use case are:

- The network shall be made aware of the RAN user plane congestion state;

- During RAN user plane congestion, the network shall be able to enable the protocol optimization capability to reduce the number of interactions between the UE and the network so as to relieve RAN user plane congestion for users.

5 Considerations

5.1 Considerations on charging

No user plane congestion management considerations have been identified that relate to charging.

5.2 Considerations on security

No user plane congestion management considerations have been identified that relate to security.

5.3 Considerations on addressing

No user plane congestion management considerations have been identified that relate to addressing.

5.4 Considerations on aspects to avoid cell-level congestion

Annex A describes a situation in a healthy network where congestions are of short durations which can be seen to be around 90 % of the cases. There are also other types of situations were congestion can last for several minutes or more which might account for around 10 % of the cases.

In Annex A it is explained that when reactive solutions are used to mitigate congestion they may have a negative impact on system performance in case the feedback-loop, i.e. the delay from congestion detection until a reactive action can take effect, is too long in comparison to the load fluctuations of the cell. In addition, frequent provisioning of congestion information from the RAN to a higher level aggregating point may induce a significant increase of signalling into the system. This could potentially have a negative impact on overall system performance. Therefore suitable solutions targeting user plane traffic congestion in the RAN needs to fulfil the following suggested requirements:

- The system should react in a timely manner to manage a congestion situation, i.e., that the measures taken can take effect to help resolve the congestion.
- The signalling overhead in the system shall be minimized.

5.5 Considerations of MOCN networks in UPCON

Any use of application identification should consider the impact on Multi-Operator Core Network (MOCN) partner(s) gaining information on the use of the network by the other MOCN partner(s).

6 Potential Requirements

6.1 Introduction

The following requirements are consolidated from the requirements associated with the use cases included in this TR.

User plane congestion may last for a few seconds, a few minutes, or a few hours due to the radio environment changing, the mobile user moving and other reasons. A short-time burst of user plane traffic should not be identified as RAN congestion.

The solutions should be resilient to rapid changes in the level of congestion and be responsive to them.

6.2 Consolidated requirements

6.2.1 General

a) The network shall be able to provide mechanisms to detect RAN congestion onset and abatement.

b) The network shall be able to identify whether an active UE is in a user plane congested cell or not.

c) The network shall be able to configure such RAN congestion-based policy rules.

d) The system shall be able to provide mechanism to support different charging policy based on RAN congestion status.

e) The system should react in a timely manner to manage a congestion situation, i.e., that the measures taken can take effect to help resolve the congestion.

f) The signalling overhead in the system shall be minimized.

6.2.2 Prioritizing traffic

a) The network shall be able to identify, differentiate and prioritize different applications based on the QoS attributes of their communications.

NOTE: The applications may be social networking, OTT video, blogging, internet games, FTP, software patches and updates, non real time services, etc.

b) According to operator policies, during RAN congestion the operator shall be able to select the communications which require preferential treatment and allocate sufficient resources for such communications in order to provide these services with appropriate service quality.

c) According to operator policies, the network shall be able to select specific users (e.g. heavy users, roaming users, etc.) and adjust the QoS of existing connections or the application of relevant policies for new connections depending on the RAN congestion status and the subscriber's profile.

d) The network shall be able to identify, differentiate and prioritize different applications with same QoS attributes such as social networking, video, blog, internet games, FTP, software patches and updates, operator ownership, etc.

6.2.3 Optimizing traffic

a) When RAN user plane congestion occurs, per operator policies the system shall be able to subject traffic to compression (e.g. compressing HTTP 1.1 web content into gzip format, transcoding a 16 bit TIFF image into an 8 bit TIFF image), taking into account UE capabilities in order to optimize traffic delivery to relieve RAN user plane congestion.

b) The network shall be able to identify whether a UE that has an on-going real-time communication, e.g. voice/video call, PSS, is in a congested cell and if so, based on operator policy (which may include subscriber consent), the network shall be able to instruct the UE to renegotiate the communication media parameters of such real-time communications so that they consume lower bandwidth, e.g., a lower bit rate or a different codec type. The system shall also be able to distinguish between the voice and video portion of the call; (re-)negotiate either one separately to a lower bit rate. The network shall allow the UEs to restore the original media parameter set, e.g. higher bit rate, if RAN congestion has abated.

c) The network shall also be able to distinguish between the voice and video portion of the call and (re-)negotiate the media parameters of either one separately to, e.g., lower bit rates or different codec types.

d) The network shall be able to notify the PSS Server of the changed media codec.

e) The network shall be able to enable the protocol optimization capability to reduce the number of interactions between the UE and the network.

6.2.4 Limiting traffic

a) The network shall be able to provide a mechanism to defer its own Push services based on the cell congestion status and the operator's policy (e.g. the pre-specified time period of pushing.)

b) The network shall be able to inform third party provided Push services to defer until advised otherwise Push services based on the cell congestion status of the target UE's location and the operator's policy (e.g. the pre-specified time period of pushing).

c) The UE shall be able to inform the network when a request for a data connection establishment/reactivation is for Unattended Data Traffic or Attended Data Traffic.

d) The system shall be able to prohibit or delay all or a particular selection of requests for data connection establishments/reactivations for Unattended Data Traffic from that UE

NOTE: An app knows when a request for data is user initiated or when the request is not intended for immediate rendering (on the screen), thus it could provide an indication when it requests data from the network. Additionally or alternatively, the UE could categorize connection requests into Attended or Unattended based on one or more of the following criteria:

- whether the screen/keyboard lock is activated
- how long since the user last pressed a key or touched the touch screen
- the app informing the lower layers of the UE (e.g. as part of the UE's API)
- known type of the app (for instance, an app monitoring a user's health – "mHealth" app – may need its data always treated as Attended Data Traffic)
- etc.

7 Conclusion and recommendations

The number of data intensive UEs and the applications installed on them along with other devices that require data traffic on mobile networks continues to grow. It is becoming more urgent that appropriate mechanisms be provided to manage user plane traffic when RAN congestion occurs.

A number of use cases have been identified where user plane congestion occurs and needs to be managed. The analysis has resulted in a consolidated set of unique requirements as captured in the previous section. Mechanisms to address these requirements need to be specified, taking into account other mechanisms available that may address some aspects of this problem space.

It is therefore recommended that the potential requirements identified in this TR are considered for the development of normative requirements.

Annex A (informative): Aspects on cell-level congestion in a healthy network

A.1 Forms of congestion

This annex describes the situation in a healthy network where congestions are of short durations which can be seen to be around 90 % of the cases. There are also other types of situations were congestion can last for several minutes or more which might account for around 10 % of the cases.

For this annex we define two forms of congestion:

- End-user congestion;

  From an end-user perspective, a congestion occurs when a service is not delivered to the expectation of the user, in the following denoted service congestion.

  The expectation for a service delivery is dependent on which service that is used (requirements on bandwidth, delay…) but differs also between subscriber groups (a premium subscriber have higher expectations than a subscriber with the cheapest subscription)

- Resource congestion;

  Resource congestion occurs when traffic offered cannot be carried due to resources being depleted e.g. a node receives more packets than it can transmit or buffer. Resource congestion may cause service congestion.

Taking preventive actions before or when resource congestion occurs could be means to prevent end-user congestion.

A.2 Example of Cell level congestion

Cell level congestion is the form of resource congestion that in many cases constitutes the weakest link in a 3GPP system. The graphs below illustrate the distribution of cell level congestion durations and the distribution of the time between congestion situations on cell level. These events were collected from two RNCs during a 5 day measurement campaign in a typical live WCDMA network in a major European city, providing an example on the dynamics of resource congestion on cell level.

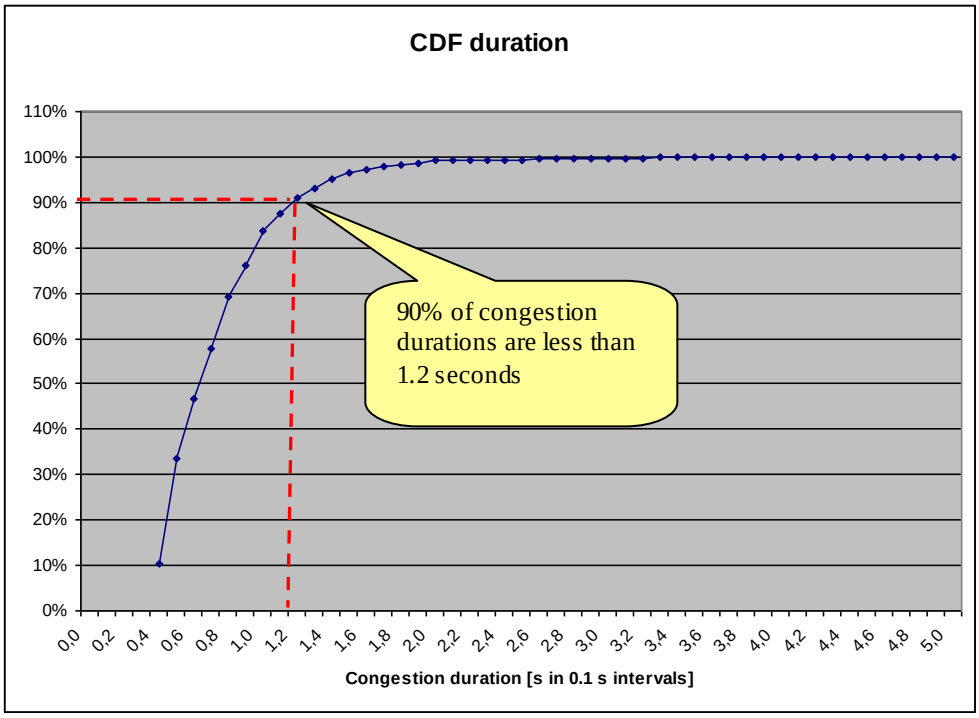

Figure A.1: Cumulative Distribution Function (CDF) of WCDMA cell-level

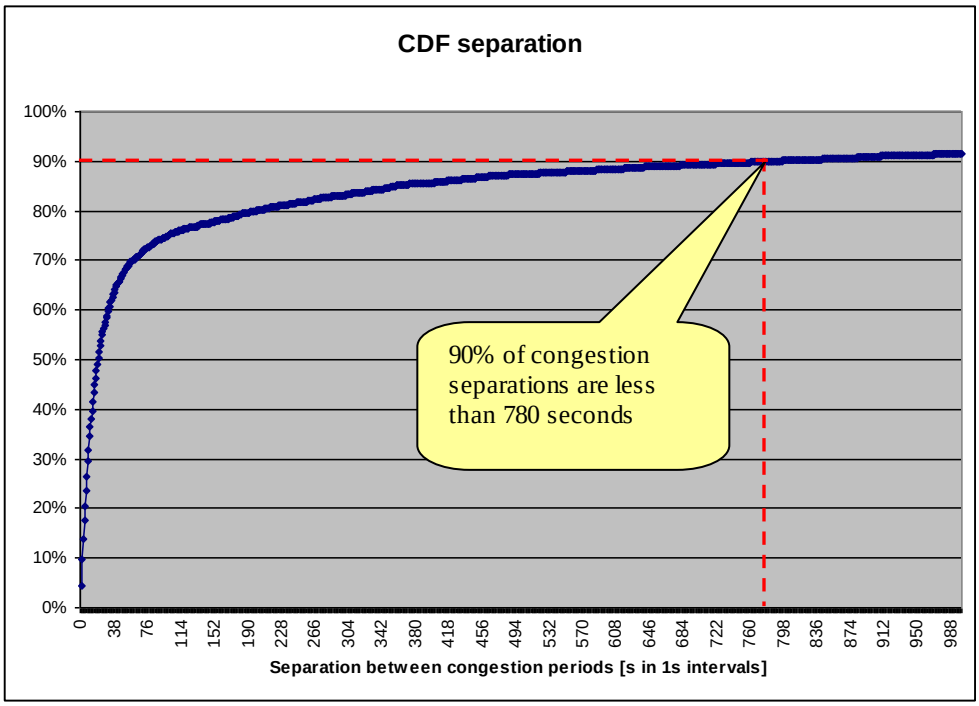

Figure A.2: Cumulative Distribution Function (CDF) of cell-level congestion separations (within individual cells).

For this measurement congestion was defined as transmitted carrier power and uplink interference exceeding its respective threshold values (for some time). Notice that cell-congestion occurs as a result of quickly deteriorating radio conditions when the utilization of the resources in the cell is high.

Figure A.1 indicates that cell congestion durations are short, 90% below 1.2 seconds.

Figure A.2 indicates that cell congestion separations are relatively short for WCDMA, 90% below 780 seconds.

From this we can conclude that cell congestion is a relatively frequently occurring phenomenon in a healthy network. The reason to this is of course that operators would like to have a high utilization of the system, however when radio conditions temporarily gets deteriorated in a cell then it may be pushed into a state of congestion.

We can also conclude that the duration of cell-level congestion is very short. This is natural because RAN internal mechanisms, e.g. Radio Link re-configurations for existing RABs, admission control for new RABs etc., will kick in immediately to alleviate the situation.

A.3 The disadvantage with a regulating function with a too slow response time

In order to prevent end-user congestion it is possible to either take preventive actions before resource congestion occurs (i.e. proactive) or after it has occurred (i.e. reactive).

Reactive actions to congestion can either be deployed directly at the entity experiencing resource congestion or, alternatively, congestion information can be signalled to an external entity, e.g. client/server or an intermediate node, for further action. It is important to notice for all reactive solutions that, just as for any control system, the relation between the expected life time of the signalled information and the delay until a regulating action can take takes place is critical. In case those two parameters are ill proportioned there is a risk that the information will be stale by the time it reaches it destination with the risk for system instability as a consequence.

Applied to solutions intended to mitigate cell-congestion this means that the delay of any signalled congestion information must match the fluctuations of the resources that constitutes the cell load well, otherwise the negative impact on performance will be significant. This is visualized in the Figure A.3.

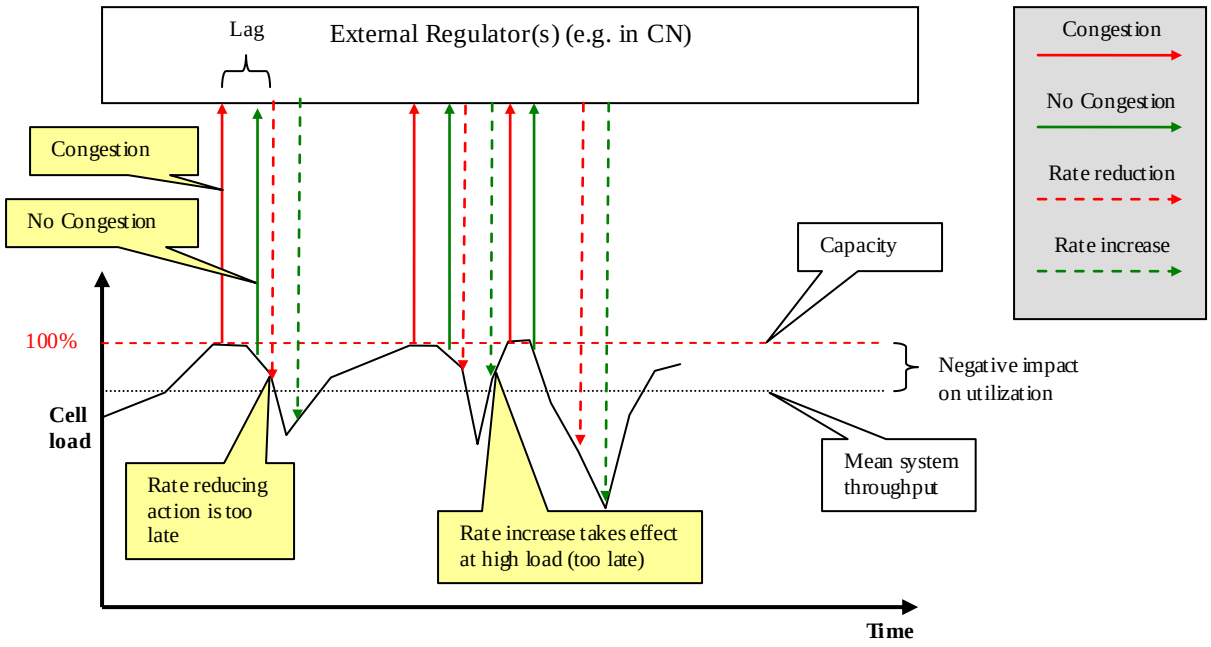

Figure A.3: Example of system performance impact from reactive actions to cell-congestion from an external entity (in this example a binary signal is assumed for reasons of simplicity).

Note that for a reactive solution where the rate reducing algorithm suffers from a lag (i.e. load/congestion information is signalled to an external regulator), there is a risk that the negative impacts on system performance resulting from the attempt to mitigate resource congestion may actually lead to unnecessary service congestion; If rate adapting actions are too late then end-user payload will get policed/shaped based on congestion policies although there may in fact be sufficient resources available in the system.

The other extreme for reactive solutions is to provide sufficient information to the entity controlling Radio Resource Management to enable that entity to adequately handle payload traffic even in congested situations.

Annex B (informative): Change history

Change history							
Date	TSG #	TSG Doc.	CR	Rev	Subject/Comment	Old	New
2012-02	SA1 #57	S1-120021			Initial skeleton presented to SA1#57	n/a	0.0.1
2012-02	SA1 #57	S1-120319			Initial version based on inputs to this SA1 meeting	0.0.1	0.1.0
2012 02	SA1 #57	S1-120320			Revisions per review of revised use cases in S1-120316, S1-120317 and S1-120318.	0.1.0	0.2.0
2012 05	SA1 #58	S1-121272			Incorporates agreements on: S1-121009 (with edits) S1-121011 (with edits) S1-121043 S1-121047 (with edits) S1-121048 (portion on Use Case 5) S1-121049 S1-121054 (with edits) S1-121071 (with edits) S1-121072 (with edits) S1-121074 S1-121125 S1-121144 (with edits) S1-121154 (parts) S1-121183 (parts) S1-121273 S1-121275 S1-121277 (with edits) S1-121279 (parts) S1-121281 S1-121282 (with edits) S1-121284 S1-121285 (editor's proposal) S1-121340 S1-121341 (with edits) S1-121343 (with edits) S1-121344 S1-121358	0.2.0	0.3.0
2012 05	SA1 #58	S1-121433			Editorial corrections	0.3.0	0.3.1
2012 06	SA#56	SP-120296			Raised to v.1.0.0 for presentation to SA#56. Same technical content as v.0.3.1.	0.3.1	1.0.0
2012 08	SA1 #59	S1-122242			Includes editorial corrections to insert missing spaces, delete extra spaces, and correct spelling to UK English. Incorporates agreements on: S1-121193 (with edits) S1-122076 S1-122113 S1-122116 S1-122245 S1-122246 S1-122247 S1-122249 S1-122250 S1-122251 S1-122252 S1-122253 S1-122254 S1-122255 S1-122256 (with edits) S1-122395 S1-122399	1.0.0	1.1.0

2012 08	SA1 #59	S1-122258			Correction of editorial errors in incorporating P-CRs.	1.1.0	1.1.1
2012 08	SA1 #59	S1-122259			Further editorial corrections in incorporating P-CRs. Delete unused reference numbers in §2. Adjust formatting of lists in §6 to match drafting rules in TR 21.801.	1.1.1	1.1.2

Change history											
TSG SA#	SA Doc.	SA1 Doc	Spec	CR	Rev	Rel	Cat	Subject/Comment	Old	New	WI
SP-57	SP-120537	S1-122259	22.805	-	-	-	-	Raised to v.2.0.0 by MCC for submission to approval by SA#57	1.1.2	2.0.0	FS_UPCON