Advancing 3GPP Networks: Optimisation and Overload Management Techniques to Support Smart Phones

Version 1.0

June 2012

Security Classification – Non Confidential GSMA White Paper

Copyright Notice

Antitrust Notice

The information contain herein is in full compliance with the GSM Association's antitrust compliance policy

Table of Contents

1 Executive Summary

The focus of this white paper is to identify overload situations that arise in 3GPP networks due to heavy penetration of Smart Phones and tablet PCs due to the applications and services running on them. This paper classifies the various issues into a short list of different categories (for example, Small Data Packet Transmission and Video Services etc.). There is no unique or theoretically correct categorization, as some of the issues may overlap between categories. For the sake of effectiveness, a pragmatic approach is promoted over a theoretical and lengthy debate.

This paper documents various solutions for the above identified categories. These solutions range from existence of ‘ready to deploy’ standards based and proprietary vendor solutions to nearly available industry specifications awaiting implementation to new concepts and ideas barely written up on paper. Solutions have been identified based on 3GPP specifications, other Standards Development Organisations (SDO) or over-the-top solutions

This white paper can be used as a reference by operators in identifying solutions that would provide relief for their network conditions and to further advance implementation of solutions. It is proposed to forward this white paper to relevant Working Groups/SDOs where interested companies can further develop specifications using the normal SDO procedures.

2 Introduction

2.1 Purpose of the document

The GSMA has embarked on an activity called *Network Application Efficiency*. The GSMA activity includes multiple threads, including the development of application developer guidelines (how best to design network-friendly applications on different operating systems) and the work stream on Network Optimisation.

This document will serve as key input to the GSMA Network Optimisation work stream. The objectives of this document are as follows:

- to identify key issues related to network inefficiencies caused by Smart Phones and Smart Phone applications, and
- to identify possible solutions which require a degree of joint and aligned industry action.

2.2 Scope

This document covers various control plane and user plane overload and optimization techniques for 3GPP networks due to extensive penetration of Smart Phones. Impacts of Smart Phone usage and applications are documented. Techniques that can be used by operators based on their individual network situation are documented.

2.3 Definition of Terms

Term	Description
3GPP	3rd Generation Partnership Project (www.3gpp.org)
AAA	Authentication Authorization and Accounting
AKA	Authentication and Key Agreement
AN	Access Network
ANDSF	Access Network Discovery and Selection Function
ARP	Allocation and Retention Priority
APN-AMBR	Access Point Name-Aggregate Maximum Bit Rate
AS	Application Server
BSF	Bootstrapping Server Function
CBC	Cell Broadcast Centre
CBS	Cell Broadcast Solution
CCF	Charging Collection Function
CDF	Charging Data Function
CDR	Call Detail Record
CN	Core Network
CPC	Continuous Packet Connectivity
CSCF	Call and Session Control Function
DASH	Dynamic and Adaptive Streaming over HTTP
DCCA	Diameter Credit-Control Application
DHCP	Dynamic Host Control Protocol
DPI	Deep Packet Inspection
DRA	Dynamic Routing Agent
DRX	Discontinuous Reception
DTX	Discontinuous Transmission
EAB	Extended Access Barring
eNB	Evolved Node B
eMBMS	Evolved Multimedia Broadcast Multicast Service
ePDG	Evolved Packet Data Gateway
EPS	Evolved Packet System
E-UTRAN	Evolved – Universal Terrestrial Radio Access Network

FQDN	Fully Qualified Domain Name
FLUTE	File Delivery over Unidirectional Transport
GBA	Generic Bootstrapping Architecture
GTP	GPRS Tunnelling Protocol
GUMMEI	Globally Unique MME Identifier
GUTI	Globally Unique Temporary Identifier (consists of GUMMEI and M-TMSI)
HeNB	Home evolved NodeB
H-SLP	Home SUPL Location Platform
HLR	Home Location Register
HLS	HTTP Live Streaming
HS-DPCCH	High Speed-Dedicated Physical Control Channel
HSS	Home Subscriber Server
HTCP	Hypertext Cashing Protocol
IETF	Internet Engineering Task Force
IFOM	IP Flow Mobility and Seamless Offload
IP	Internet Protocol
I-CSCF	Interrogating Call and Session Control Function
I-SBC	IMS Session Border Controller
ITU	International Telecommunications Union
LSGW	LTE SMS GW
LTE	Long Term Evolution
MAPCON	Multi Access Packet Data Network Connectivity
MBMS	Multimedia Broadcast Multicast Service
MME	Mobility Management Entity
MMEGI	Mobility Management Entity Group Identity
MCC	Mobile Country Code
MNC	Mobile Network Code
M-TMSI	MME- Temporary Mobile Subscriber Identity
NAI	Network Access Identifier
NAS	Non-access Stratum
NAT	Network Address Translation
NMS	Network Management System
NNI-SBC	Network to Network Interface – Session Border Controller
NSRM	Network Service Request Manager
NT	Network Initiated Traffic
OA&M	Operations and Maintenance
OCS	Online Charging Server
PCC	Policy and Charging Control
P-CSCF	Proxy Call and Session Control Function
PLMN	Public Land Mobile Network
PCRF	Policy and Charging Rules Function
PDN	Packet Data Network
PDN GW / PGW	Packet Data Network Gateway (H=Home or V=Visited)
PLMN	Public Land Mobile Network
PS	Packet Switched
PSI	Public Service Identifiers
QCI	QoS Class Identifier
QoS	Quality of Service
RAN	Radio Access Network
RAT	Radio Access Technology
RNC	Radio Network Controller
RRC	Radio Resource Control (3GPP)
RTP	Real-time Transport Protocol
S-CSCF	Serving Call and Session Control Function
SLP	SUPL Location Platform
SNMP	Simple Network Management Protocol
SAE	System Architecture Evolution
SBC	Session Border Controller

SCG	Service Continuity Gateway
SD	Small Data Packet Transmission
SDO	Standards Development Organisation
SGW	Serving Gateway
SMS	Short Message Service
SR	Session Router
S-TMSI	S-Temporary Mobile Subscriber Identity (consists of MMEC and M-TMSI)
SIP	Session Initiation Protocol
TA	Tracking Area
TA-List	Tracking Area-List
TAI-List	Tracking Area Identity-List
TAU-List	Tracking Area Update-List
TCP	Transmission Control Protocol
TWAP	Trusted Wireless Access Proxy
TWAG	Trusted Wireless Access Gateway
UE	User Equipment (a.k.a. – mobile handset or access terminal)
UMTS	Universal Mobile Telecommunications System
VoIP	Voice over Internet Protocol
VS	Video Services

2.4 Document Cross-References

Ref	Document Number	Title
[1]	3GPP TS 22.001	“Principles of circuit telecommunication services supported by a Public Land Mobile Network (PLMN)” (http://www.3gpp.org/ftp/Specs/html-info/22001.htm)
[2]	3GPP TS 22.011	“Service accessibility” (http://www.3gpp.org/ftp/Specs/html-info/22011.htm)
[3]	3GPP TS 22.801	“Study on non-MTC Mobile Data Applications impacts (Release 12)” (http://www.3gpp.org/ftp/Specs/html-info/22801.htm)
[4]	3GPP TS 23.060	" General Packet Radio Service (GPRS)" (http://www.3gpp.org/ftp/Specs/html-info/23060.htm)
[5]	3GPP TS 23.203	“Policy and charging control architecture” (http://www.3gpp.org/ftp/Specs/html-info/23203.htm)
[6]	3GPP TS 23.261	“IP Flow Mobility and Seamless WLAN Offload” (http://www.3gpp.org/ftp/Specs/html-info/23261.htm)
[7]	3GPP TS 23.401	"GPRS enhancements for E-UTRAN access" (http://www.3gpp.org/ftp/Specs/html-info/23401.htm)
[8]	3GPP TS 23.402	“Architecture enhancements for non-3GPP Access” (http://www.3gpp.org/ftp/Specs/html-info/23402.htm)
[9]	3GPP TS 23.682	“Architecture enhancements to facilitate communications with packet data networks and applications” (http://www.3gpp.org/ftp/Specs/html-info/23682.htm)
[10]	3GPP TR 23.888	“System Improvements for Machine-Type Communications (Release 11)” (http://www.3gpp.org/ftp/Specs/html-info/23888.htm)
[11]	3GPP TS	“Access to the 3GPP Evolved Packet Core (EPC) via non-3GPP access networks” (http://www.3gpp.org/ftp/Specs/html-info/24302.htm)

	24.302	
[12]	3GPP TS 24.312	"Access Network Discovery and Selection Function (ANDSF) Management Object (MO)" (http://www.3gpp.org/ftp/Specs/html-info/24312.htm)
[13]	3GPP TS 26.247	"Transport and end-to-end Packet-switched streaming service (PSS); Progressive download and Dynamic adaptive stream over HTTP (3GP-DASH)" (http://www.3gpp.org/ftp/Specs/html-info/26247.htm)
[14]	3GPP TS 36.331	"Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification" (http://www.3gpp.org/ftp/Specs/html-info/36331.htm)
[15]	NSRM reference	http://www.qualcomm.com/media/documents/managing-background-data-traffic-mobile-devices
[16]	GSMA TS 20	Developer Guidelines for Network Friendly Apps http://www.gsma.com/documents/Smarter-Apps-for-Smarter-Phones!/22440/
[17]	GSMA TS 18	Fast Dormancy Best Practices http://www.gsma.com/newsroom/1-0-network-efficiency-task-force-fast-dormancy-best-practices/
[18]	OMA Push V2.3	http://www.openmobilealliance.org/Technical/release_program/push_v2_3.aspx
[19]	OMA DCD V1.0	http://www.openmobilealliance.org/Technical/Release_program/dcd_v1_0.aspx

3 Optimisation and Overload Topics

3.1 Small Data Packet Transmission (SD)

3.1.1 Problem Statement and Motivation

As Smart Phones proliferate, not only has the aggregate volume of user data exploded but also the amount of signalling traffic generated on the network has increased significantly. Some measurements of the amount of signalling traffic in mobile networks show an increase by 30-50% over the growth of user data traffic. Some applications require permanent connectivity even though they may transmit little packet data.

Many wireless data applications are characterized by transmissions of small data packets (frequent or infrequent). These transmissions can have adverse effects due to:

- Overhead associated with managing radio transmission resource allocation, resulting in high negative impact on system capacity.
- UE (User Equipment) battery consumption.

[3GPP TR 22.801] provides a broad statement of the above problem in a general sense in Section 4.1 of the document. Annex C of that document provides experimental evidence of the adverse impact stated above. This was obtained by tests from real deployed Release 6 UMTS (Universal Mobile Telecommunications System) networks providing confirmation that the issues related to support of Smart Phone applications are serious and need to be addressed in a timely manner. TR 22.801 investigates the service scenarios and use cases of different mobile data applications. Their impact to the current system is generalized in the document, and potential service and operational requirements, as well as possible system enhancements are identified.

Some specific Small Data Packet Transmission causes and artifacts are summarized and illustrated below.

Application examples include the following:

- Social networking applications generate numerous typically small-payload messages, such as status messages, location messages, presence updates, instant messages, keep-alives, etc.). Similarly, content feed of social networking applications, like Twitter, results in many often short messages which fan out to users subscribing to the feed.
- The emerging set of HTML5.0 applications such as WebRTC (Real-Time Communication) will likely enable further rise in devices and mobile-based applications that generate small-payload packets. This is due to an increasing popularity of development platforms such as AJAX and other web apps. AJAX enables exchange of data between a client and a server, and updating parts of a web page without reloading the whole page. This represents a part of the broad-based trend sometimes referred to as web-based or on-line computing.
- Advertising in free downloadable games generates much background traffic often containing small payloads.
- Application and transport keep alive messages. These messages are generated by either applications or the TCP (Transmission Control Protocol) /IP (Internet Protocol) stack in order to maintain end-to-end connections. These messages carry no information for the user yet repeatedly consume network resources.

Small data packets are exchanged frequently as many of these mobile data applications run concurrently on a UE, causing the UE to transition frequently from idle and active states resulting in increased control plane signalling in RAN (Radio Access Network) and Core Network (CN).

The time interval between heartbeat messages is often several tens of seconds for some applications. IM and presence status update information does not have regular intervals, but may change frequently. The effect of presence status updates is compounded by generating large number of small data packets as an update message is fanned out (pushed) to all friends in the buddy list.

Analysis of packet traces of major internet POPs shows that a large fraction (about 40%) of the packets on the Internet today are less than 50 Bytes for IPv4 traffic. Similar observations have been seen in the traffic on wireless access technologies. These packets contain a variety of payloads such as TCP ACKs and application related payloads such as VoIP (Voice over Internet Protocol) silence suppression.

This trend of small packets is expected to be exacerbated as status messages, location messages, instant messages, keep alives, heartbeat messages, etc., as generated by the current generation of mobile data apps grow considerably over time. Furthermore, with IPv6, packets carrying small payload may increase in size.

Mobile data applications featuring short messages (IM, Social Networking, etc.) involve interactive communications between the client on the UE and application server in the internet cloud. The server and the application on the UE periodically exchange “heartbeat” messages (also known as keep-alives) to keep the application session alive and also to avoid the expiry of NAT (Network Address Translation) mapping which causes IP session disconnection.

In addition to periodic keep-alive messages, the applications also generate frequent status update messages to notify the users of status updates relating to the application. Some examples include presence information of buddies in an IM buddy list, update of user location upon user “check in”, update of “Facebook likes” to a user’s friends, etc.

Regarding frequency, keep-alive messages such as those in VoIP apps (e.g. Skype) generate keep-alive messages between every 30 seconds and every 8 minutes. Social networking apps such as FindMe, generate status update messages upon geographic position changes. The frequency of such messages ranges from sporadic in the course of a day (e.g. “home”, “work”, “gym”, then back to “home”) to as frequent as every 60 seconds. Social networking servers push content and presence update messages of the subscriber’s friends to the application on the UE (e.g. Facebook posts the activities when your friend “likes” a particular article or “becomes a fan” of a particular group). The frequency of such content and presence update messages is estimated to be in the order of once every few minutes (see Annex A of [3GPP TR 22.801]).

A few more aspects can aggravate the impact of status update and keep-alive messages:

- Messages can be mobile-originated (MO) or mobile-terminated (MT), e.g. periodic FindMe messages can come from change of location of your friends or can come from the updates of your own location.
- It is not uncommon for a UE to install multiple applications, where each application generates these update/keep-alive messages autonomously.

A large number of Smart Phones generating small data transmissions could lead to network overload (congestion) situations.

3.1.2 Solutions

In this section, a number of solutions already commercially available are listed; this includes network or UE implementation specific solutions, network parameter configuration and profiling, over the top solutions or solutions coming from features available in 3GPP releases up to and including Release 9.

3.1.2.1 Network or UE implementation specific solutions

- **SD-1**: Network Socket Request Manager – UE implementation specific
    - Current Smart Phones allow tasks to continue to run in the background while the phone is in standby. In addition the operating system tends to present the cellular interface to these applications as an always-on IP interface. As a result applications running in the background initiate traffic asynchronously which leads to frequent radio connections. For instance an Android phone with about 10 applications was found to generate 30-50 RRC (Radio Resource Control) connections per hour while in standby.

    http://www.qualcomm.com/media/documents/managing-background-data-traffic-mobile-devices

    - The NSRM (Network Socket Request Manager) intercepts application’s socket calls and holds them until a release time is determined. This typically allows grouping several applications’ data transfer over a single radio connection.
    - Timers should be randomised over time, and offset by a few seconds randomly each time to prevent alignment/traffic spikes.

- **SD-2**: Use of proxy in the network to manage data exchanged with a Smart Phone. This mechanism:
    - Minimizes data volume by:
        - Grouping, sorting, prioritizing, batching, compressing and grooming data on the server side, before communicating with the network
        - Optimization for mobile use cases
    - Minimizes signalling frequency by:
        - Sharing a single socket for multiple concurrent services
        - Pushing high priority content, holding low priority content in reserve for available windows

- Minimizes handset impact:
  - Low signalling frequency reduces radio activation and battery drain

Figure 1 depicts a possible architecture where a proxy client or a device OS could provide and additional layer of protection towards the network by reducing the number of "status checks" various applications perform. The proxy client would have a corresponding network proxy server, which would be polling the appropriate application servers on behalf of the UE and the network resources are utilized only when there is actual valid data to be sent to the device or the application.

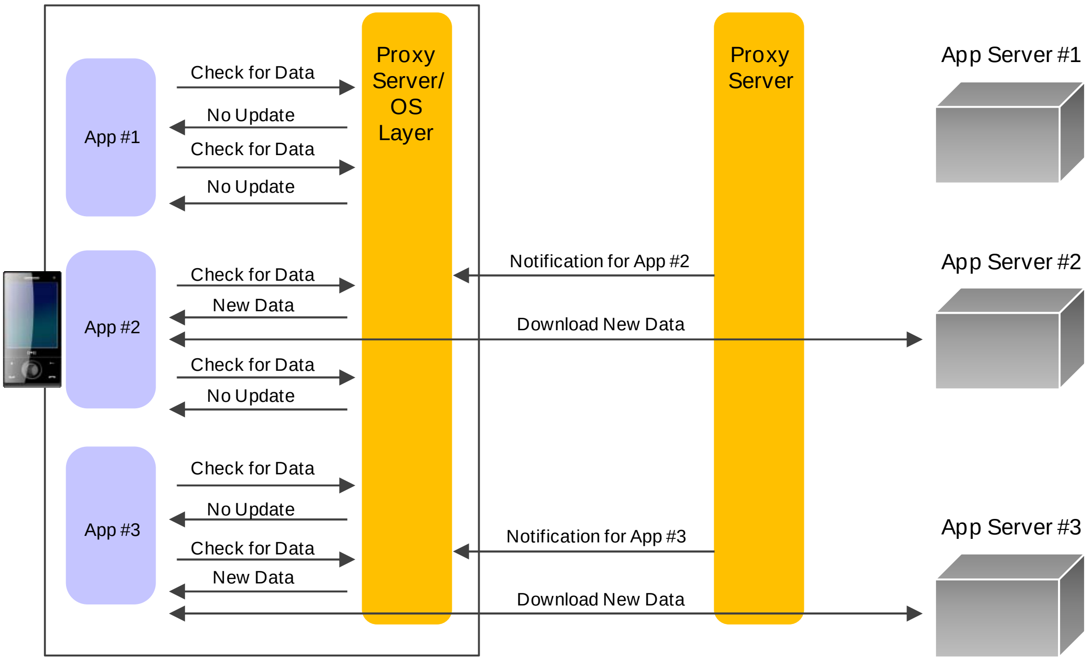

Figure 1: : Client Proxy-Network Proxy Solution

3.1.2.2 Network parameter configuration and profiling at signalling layer

- **SD-3**: Authentication Tuning

If the device is idle when it needs to send small data, it must start with a Service request to the network. Service requests are one of the main contributors to signalling load in the core network. For this reason, operators should tightly control the authentication procedure (AKA, Authentication and Key Agreement authentication) which can be triggered by the network during these service requests.

MME (Mobility Management Entity) re-authenticates the UEs based on various event triggers (e.g. – GUTI Attach, Tracking Area Update, Service Request, UE Initiated Detach and UE Initiated Extended Service Request) and even with selective authentication the MME would require new authentication vectors from the HSS (Home Subscriber Server) and could trigger additional load. Other applications i.e. AAA (Authentication Authorization and Accounting), CSCF (Call and Session Control Function) would also be requesting authentication vectors on various triggers.

- One solution is to download multiple authentication vectors to the network element (MME/AAA/CSCF) from the HSS and the network element would utilize the local vectors before reaching out to the HSS for new vectors. This solution may not be suitable for some access technologies.

- In EPS, there is an option to increase the lifetime of K_ASME in the MME as K_ASME is a high level key. This can automatically reduce the number of authentications hence reduces S6a signalling. As long as the K_ASME stays the same, there are no new security keys but the AS security keys can be renewed. K_ASME and NAS security keys can be renewed only through an EPS AKA authentication run.
- Network should permit the percentage of signalling events on which the UE is challenged to be configurable.
    - For example, the network should be able to challenge the UE only every 10th valid event trigger or every 10th Tracking Area Update i.e. 10% authentication scenario.
        - HSS should provide for ability to segment the vectors per application type (MME, AAA, CSCF etc.) to avoid overlap between clients and avoid re-sync issues. Refer to Annex C of 3GPP TS 33.102.

The figure below depicts an example of how each network element could request for more than one vector to reduce load on the HSS.

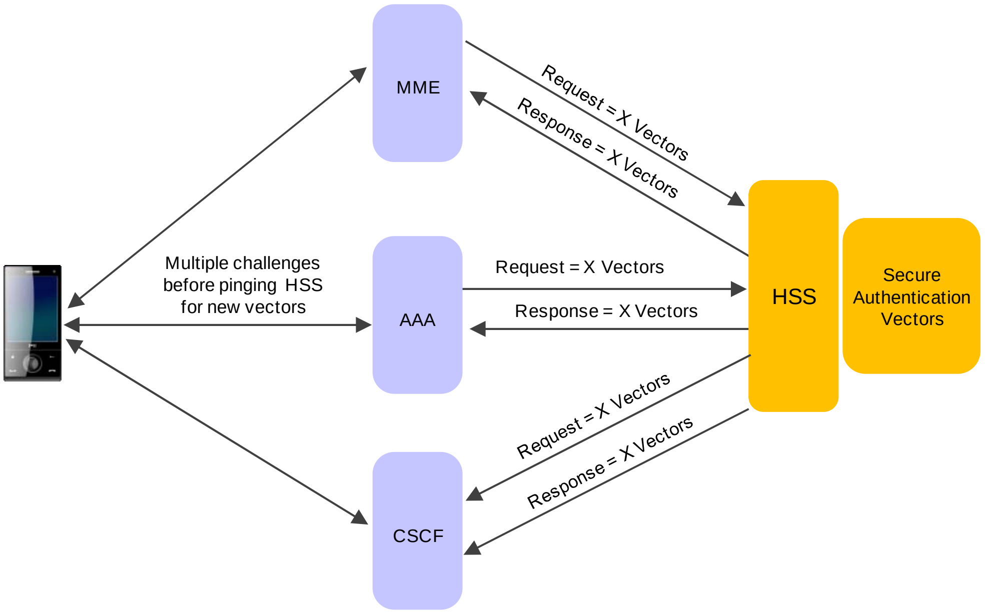

Figure 2: : Example Optimisation of Authentication

- **SD-4**: Overload Controls

Standard and/or proprietary congestion control or load balancing procedures, which are already available today, should be used by network equipment to prevent/minimize network congestion and failures. ACB is the basic feature for this, and the use of DSAC (3G), SSAC (LTE), PPAC and eMPS are also important to further enhance operator control (e.g., applying access control for UE in connected state) of its NW use. These overload mechanisms need to be tuned correctly to maintain a high performing network while providing a sufficient Smart Phone's user experience. Refer 3GPP TS 23.401 and TS 36.331.

3.1.2.3 Over the top mechanisms

A guideline for developing over the top mechanisms to address the issue of small data transmission is the subject of the GSMA document TS.20 (Developer Guidelines for Network Friendly Apps) [16].

- **SD-5**: Network Info API

Network Info API is a mechanism through which apps could learn when a network connection is present (and what type of network is being used). If such an API was exposed to applications, the applications make "opportunistic" decisions about what network transactions to defer until there is an existing connection (or when several requests have been issued, and justify the creation of a data connection). This approach puts the decision about how and when to use the network efficiently in the developer's hands.

3.1.2.4 3GPP features (both UE and Network are involved)

- **SD-6**: UMTS Release 99 feature: URA_PCH state and Cell_PCH state

This mechanism, present since Release 99 of UMTS is beneficial to keep UEs with very low data activity in a connected state. This allows to simultaneously save the UE battery (since the UE does not need to continuously monitor the downlink channels) and reduce the amount of signalling required to move the UE to a state where data can be exchanged (since the UE remains in connected state).

- **SD-7**: UMTS Release 7 feature: DTX (Discontinuous Transmission)/DRX (Discontinuous Reception) in active mode (CELL_DCH).

This mechanism (also known as Continuous Packet Connectivity - CPC), enables keeping UEs in active mode (aka CELL_DCH) when its traffic consists of frequent transmission of small amounts of data. When configured with CPC, UEs in CELL_DCH can apply DTX on UL physical channels providing gains in radio resource capacity, and also DRX on DL physical channels which saves UE battery.

- **SD-8**: UMTS Release 8 feature: NW controlled Fast Dormancy

This feature, also covered by the GSMA document [Fast Dormancy Best Practices TS.18], allows a UE to indicate to the radio network the end of a certain data transmission (known by the UE upper layers) ,which is controlled by the NW, and the radio network to decide a better radio state to transition the UE to (e.g. it could be active mode with DTX/DRX or a more dormant radio state, such as CELL_FACH or CELL/URA_PCH).

- **SD-9**: UMTS Releases 7 and 8 features to enhance operation in CELL_FACH

  A set of features has been defined to improve the performance of UEs in CELL_FACH, which becomes a very suitable RRC state for HSPA Smart Phones that generate small data. Enhancements include:

  - HSDPA in FACH (or HS-FACH) in Release 7 and EUL in FACH (or HS-RACH) in Release 8, which improve the user experience and the resource utilization for DL and UL data transmission in CELL_FACH (using HSPA for both signalling and bursty data), respectively.
  - CELL_FACH-DRX in Release 8, which improves UE battery consumption while UE stays in CELL_FACH.

  - The improved performance provided by the above CELL_FACH enhancements would allow Smart Phones to be kept longer in CELL_FACH state (compared to legacy CELL_FACH). This can minimize the overall number of UE transitions between RRC Connected and Idle states, reducing significantly the amount of radio signalling overhead necessary to handle frequent radio bearer releases and re-establishments, which have shown to impact both radio capacity (e.g. DL power) and UTRAN congestion (e.g. processing load).

- **SD-10**: LTE (Long Term Evolution) Release 8 features
  - LTE was designed to target short end to end delay, short connection setup delays and an always on experience. This resulted, for example in less nodes in the network, a fast connection setup between the UE and the eNB (Evolved Node B), and the inclusion of DRX from the first release (Release 8)

- **SD-11**: MBMS (Multimedia Broadcast Multicast Service) (Also please refer Section 3.3 of this document)

In the case of time synchronized small data, or data that can be intelligently scheduled whilst retaining context and usability, broadcast mechanisms can be used to reduce network load. Use cases such as scheduled App data population (e.g. newspapers, magazines, video clips), App updates, pre-caching of advertisements, and firmware updates may be suitable for broadcast. Small data could also be sent to a group of users (Refer to Section 3.5 of this document).

A number of solutions already standardized and included in 3GPP releases up to Release 11 are described below.

- **SD-12**: UMTS Release 11 feature on Further Enhancements to CELL_FACH

On the HSPA radio network side (UE and RAN), further optimization for small data transmission in CELL_FACH can be achieved by implementing other standard enhancements, like those defined under the Release 11 Work Item known as Further Enhancements to CELL_FACH [3GPP RP-111321]. Some suitable enhancements include: Standalone HS-DPCCH (High Speed-Dedicated Physical Control Channel), 2ms/10ms TTI handling, Signalling based interference control, UE battery life improvements and signalling reduction (aka 2nd DRX in CELL_FACH).

- **SD-13**: LTE Release 11 feature on Diverse Data Application is on-going and interested parties can contribute in 3GPP.

3.2 Network Initiated Traffic (NT)

3.2.1 Problem Statement and Motivation

Certain push mobile data services do not require real time delivery. Example applications are given below.

- Advertisements
- Service notifications
- Video clips
- Background traffic like advertising in free downloadable games

- Some multimedia messaging services.
- Software/firmware upgrades or patches

These services may be provided by the network operator, or from third-party providers. Usually these services are distributed to a large number of users in a short period, e.g. news reports in the morning rush hours and every evening. Customers subscribing to these services should receive the content periodically or in real time. However large bursts of data services can lead to network congestion and impact service experiences.

Dense simultaneous push services in a local area may cause the network congestion and service break. For push services that do not have strict real-time requirements, sending them to a large number of devices within a short duration is wasteful and unnecessary, especially when parts of the network are already congested. The delivery of such push services could be delayed until there are sufficient network resources available, which not only optimally utilize network resources, but also guarantee acceptable user experience.

However, for time synchronized push data (e.g. Software upgrades) that can be intelligently scheduled to a large number of Smart Phones, broadcast/multicast mechanisms (e.g. MBMS) can be used, which reduce network load optimizing resources utilization. See also 3.1.2.

In the case where the push services are provided by third-party providers, the lack of knowledge on the current congestion situations of different parts of the network results in the inability of third-party providers to schedule the push services more intelligently. In the case where the push services are provided by the operator, lack of mechanisms and interfaces to schedule the push services (which are not located in the EPC) based on the current network status at the core and Access Networks (AN) also means it is not possible to schedule the push services efficiently.

3.2.2 Solutions

In this section, a number of solutions already commercially available are listed; this includes network or UE implementation specific solutions, network parameter configuration and profiling, over the top solutions or solutions coming from features available in 3GPP releases up to and including Release 9.

3.2.2.1 Network or UE implementation specific solutions

- **NT-1**: Push server (see 3.1.2.1)

To reduce the frequency services are pushed to the device, application providers can be encouraged to use Push servers or push scheduling. Scheduling optimizations can be made to servers, especially for applications that do not need an immediate Push. Possibly those delay tolerant Pushes could be queued until the device becomes active, until there are several Push requests for the same device queued, or until the network is operating at off peak hours. A companion activity in the ANETF is focussed on Push server requirements that could address both SD-2 and NT-1.

3.2.2.2 Network parameter configuration and profiling

- **NT-2**: Management of Tracking Areas and small cell layers can reduce the signalling load put on the network when pushing services to a device.

- Operators could also consider a feature on the MME to perform heuristic paging where the MME keeps track of the eNodeB to which the UE most commonly attaches or was previously connected to and only paging those eNB before paging the TA or TA (Tracking Area)-list, thus reducing the signalling otherwise associated with continuous paging. Another possible solution is assigning the UE large TAI LIST to minimize the TAU traffic and then optimize paging via “guesses” on where the user potentially is based on estimations of movement direction, last known location and speed. The UE would get paged first based on these heuristics and only upon failure on the whole TAI List.
- When the device is idle it must be paged in order to receive the Push service. This can put a tremendous load on the network, especially if a service pushes too many devices at the same time. One way to reduce paging resources is to reduce the tracking area size. By limiting the area to be paged, the signalling load can be minimized if there is a high degree of confidence that the UE can be located on an initial attempt (e.g. because the UE is slow moving). Tracking area optimization is used in conjunction with paging methods to reduce signalling loads. Through the use of tracking area lists and TA-List management, Tracking Area Update (TAU) messaging can be reduced.
- When smaller paging areas are used, advanced tracking area capabilities are required to reduce the risk of “ping-pong” signalling events between adjacent tracking areas. The “ping-pong” effects at TA boundaries can be decreased if the TA-List is refreshed at each TA Update. Smaller TAs may induce more signalling due to more frequent tracking area updates (TAUs) and there is a greater risk of “ping-pong” signalling occurring between adjacent tracking areas as the UE moves in the boundaries of tracking areas. The end result of smaller TAs may mean no net benefit to either the subscriber (in improved battery life) or the operator (reduced signalling load) unless more sophisticated methods of tracking area management are used. The network operator can configure the system so that the UE is provided with a TA-List at either attachment or Tracking Area update. By providing a TA-List as opposed to a single TA, the UE needs only to send a TAU when it moves out of the TA-List it currently has or at the expiration of a periodic TAU timer. This reduces the network signalling load and saves UE battery life if the TA-Lists are defined intelligently.
- Management of TA-List can for instance be based on estimation of UE speed (the faster the UE the more frequently it will cross TA boundaries and the more it will generate TAU’s. Therefore the TAI (Tracking Area Identity)-List could be made larger as the UE is faster). More static UE’s could have a narrower list to minimize paging traffic. Also users that are relatively static should be moved to small cell layers if available and users that are moving faster to macro layers. The MME and the eNB are involved in these procedures.

- **NT-3**: Retain mobiles in connected mode

Since the control plane impact of Network Initiated Traffic is due to UE’s being in Idle mode that need to be brought back to active state, UE should be kept in active state as long as possible without compromising battery consumption too much, such that paging and the resulting idle to active signaling can be reduced. In UTRAN it is possible to keep UE’s in Cell-PCH or URA-PCH state to avoid the UE going Idle outright. This feature if more widely deployed could help for push service and also reduce the overall signalling load. In LTE keeping the UE connected longer is also possible via DRX cycles setting.

- **NT-4**: Retain mobiles in connected mode when IMS is used

In LTE with IMS, this solution becomes inherent as the SIP Push binding of OMA Push becomes directly usable over the SIP signalling channel. As with Push/SMS in 3G, SIP Push can be used to deliver "connectionless" events (not requiring a user plane bearer), via SIP MESSAGE.

http://www.openmobilealliance.org/Technical/release_program/push_v2_3.aspx

- **NT-5**: Authentication Tuning

If the device is idle when application data needs to be pushed to it, the device will respond to the page with a Service request which may require authentication. Therefore if the percentage of times this authentication is required is reduced, please refer to 3.1.2.2.

- **NT-6**: Overload Controls:

Even when a network is tuned optimally, extraordinary and multiple events occurring simultaneously can cause overload in the network.

  - A robust overload control strategy can mitigate overload caused by extraordinary events. There are many 3GPP based and proprietary congestion control procedures available for network equipment to prevent equipment failures (see 3GPP TS 22.001, 3GPP TS 23.401 e.g. Section 4.3.7.4 and TS 3GPP 23.060 Section 5.3.6). Some controls are executed within network elements such as the eNodeB, MME or data gateways. Other controls are used between entities to throttle the load one entity is putting on another such as the 3GPP overload control message that can be sent from MMEs to eNodeBs to have the eNodeB throttle traffic to the MME, or overload control messaging that can be used from the MME or the S4-SGSN to selectively limit the number of Downlink Data Notification requests the SGW (Serving Gateway) sends to the MME or S4-SGSN for downlink low priority traffic received for MSs in idle mode. The priority of traffic is based on the bearer ARP (Allocation and Retention Priority) and other policies. In this manner, low priority traffic related Downlink Data Notifications are not sent in favour of higher priority traffic related Downlink Data notifications.
  - There are also load balancing mechanisms, such as offloading part of the load on one MME to another MME in the MME pool. Proper values of timers used in the system are also important. These overload mechanisms need to be tuned correctly to allow the highest throughput while maintaining a high performing network.
  -

3.2.2.3 Over the top mechanisms

Operators can consider deployment of one or more solutions provided by the OS vendors. Operators may also consider the new Work Item in OMA. Please see Annex A in this document.

3.2.2.4 3GPP features

- **NT-7**: UMTS Release 7 and 8 features: Improvements to CELL_FACH, feature: Rel-8 NW controlled Fast Dormancy: See 3.1.2.4.

- **NT-8**: MBMS is useful in addressing the domain of push services that are of interest to a large number of users and these users can be delivered the same content based e.g. on subscription or because the operator by default delivers content to groups or all UE's (e.g. in given locations) or to the whole subscriber base (e.g. in a given location or in the whole network). Also refer to GF-2 and GF-7.

- **NT-9**: OMA has also defined an MBMS (as well as OMA BCAST and Cell Broadcast Service) binding for broadcast Push services.

- **NT-10**: UMTS Release 11 feature on Further Enhancements to CELL_FACH (will add a 2nd DRX timer which will make battery saving more like PCH): See 3.1.2.4.

- **NT-11**: LTE Release 11 feature on Diverse Data Application: See 3.1.2.4.

Please also see GF-8 for group paging and group triggering.

3.3 Video Services (VS)

3.3.1 Problem Statement and Motivation

Consumers, mobile device vendors, application developers, and content providers are collectively forcing mobile operators to reinvent themselves. Mobile networks must evolve to deliver the services, features, coverage, and high-quality experience that subscribers demand or subscribers will look for other providers.

Video delivery is both a challenge and an opportunity for mobile service providers - a challenge because the cost of transporting and delivering video, whether OTT and free, or premium and paid for, is much greater than the cost of traditional services such as voice or text Short Message Service (SMS), and an opportunity because there is high demand for a new type of video delivery service that meets subscribers' needs and wants. The challenge is to find a way to make the cost and revenues line up to produce return on investment.

In summary, how can a mobile service provider reduce the cost of mobile video, while at the same time enhancing the subscriber's watching experience?

3.3.1.1 Video services

The video services under considerations are:

- Live and on-demand video streaming
- Video clip download/upload/messaging
- Video monitoring
- Real-time communication

All these services can be provided as Over-the-Top applications or managed applications.

3.3.1.2 Overload situations

While video streaming and real-time communication services over dedicated radio bearers with guaranteed bit-rates are unlikely to suffer from congestion, services used over the top may not be mapped onto dedicated radio bearers but rather mapped to best effort radio bearers. In that case video clients compete for the same resources and also compete with other traffic (e.g. FTP, Web, p2p) and have no way to act in a fair manner. Hence congestion may occur which degrades QoS (Quality of Service).

Example scenarios of overload due to video services are:

- Live Content Streaming to many users at the same time and the same location over unicast bearers (e.g. breaking news TV coverage)

- High concentration of video streaming users in a stadium all watching the same content during a football game
- Simultaneous video service startup by many users in response to events
- Continuous streaming (security applications)
- Large data volume downloads related to video – Software and non-real time content affecting video users
- Over the top video services including HD quality (high resolutions and bitrates)
- 3D video services
- Client based adaptive bit rate applications requesting as much bandwidth as possible and competing with other clients sharing the same resource
- High concentration of video calls in small areas where visual events take place that users want to share

3.3.1.3 UE considerations

UE Considerations are:

- Battery consumption issues: One of the most important differentiator is the possibility to use video encoder and decoder hardware which is far less current consuming than SW based video. This consideration calls for the minimization of video codec fragmentation and the possibility offered to applications to use platform video hardware codecs.
- The proliferation of multiple codecs is problematic as the OTT content providers occasionally change codecs and require software plug-ins. Managed content providers tend to pick one codec and force all content to conform. This minimizes the number of representations required for each type of asset and assumes that the UE can decode the media. This is typically fMP4 (Fragmented MPEG4 as described in Dynamic and Adaptive Streaming, DASH-MPEG).
- Contention issues on the downlink and Uplink radio link. Depending upon the current network access issues, the UE may choose to access alternate network for offload.
- UE selection of best time. For example, podcast or advertising video content can be retrieved during off hours (e.g. at night).
- UE selection of best Access Network: for example video content can be streamed over LTE or offload to WLAN when available.
- UE selection of best delivery mode between Unicast and Broadcast.
- UE capabilities like display resolution. Content should be adapted to UE capabilities to maximize quality and minimize bandwidth and battery consumption.
- UE selection of video quality (frame rate and resolution)/bit rate based on user subscription (e.g. if volume is capped then maximising the instantaneous download rate maybe be counter-productive).

3.3.1.4 Network considerations

Network Considerations are:

- A Content Management System needs to provide content format and delivery options matching UE capabilities while the UE selects the client application, access network and delivery mode according to network environment.
- Networks supporting the 3GPP TS 23.203 - Policy and Charging Control architecture (PCC) are able to handle video services QoS and avoid congestion by establishing policies for these services whereby dedicated bearers with QoS are established.
- Network operator Service Agreements are generally in place with content providers in order to adapt content to network capabilities.

- Radio resource usage exhaustion.
- RAN backhaul capacity constraints.
- Ability to identify application QoS requirements – Uplink heavy like CCTV streaming vs. downlink traffic.
- Authentication of users to content. Need to ensure content authorization / key distribution is not overloaded and does not impact subscriber experience.
- Broadcast vs. Multicast vs. Unicast optimization.
- Broadcast and multicast would limit optimization based on user specific information but Unicast could burden resources.
    a) Identify and differentiate over the top video like applications vs. large data downloads within the operator domain like software downloads for devices

3.3.2 Solutions

Content, such as video, exists in two domains referred to as managed and un-managed relative to the network operator. Managed content affords the operator more opportunities to optimize content distribution through content preparation, content selection, content delivery methods, and content delivery points. Un-managed content is accommodated by adjusting network resources according to demand profiles. In some cases, un-managed content can be adapted or optimized on-the-fly, potentially subject to agreement with the content provider, where open protocols are used for the content retrieval and the context is not encrypted. The optimized content delivery will leverage both UE and network-based solutions services.

- Content compression: Content may be compressed to minimize the volume of data associated with an asset. The compression can be done as priority for video on-demand assets allowing the asset to have multiple representations in a catalogue. Alternatively, the content can be compressed on-the-fly for live content distribution and un-managed content requests without impacting end-user quality. The choice of codecs for content compression is evolving; however, the market has generally accepted the use of H.264 AVC for managed content. Unmanaged content encoding tends to evolve much quicker with the rapid introduction of new browser plug-in modules. HTML5 does not specify the use of a specific codec; therefore, the OTT content distributors are free to explore the use of new codecs independently from the network operator.
- Video Download Pacing: Most OTT videos are never watched all the way through. Over half are abandoned by 1 minute into the video, yet because the original server has no information about what the watcher will do, it will have downloaded the video at the fastest rate the network can support. A system that could anticipate which videos will be abandoned and could potentially save well over 50 percent of all of the bandwidth consumed by video downloads by not transmitting them over the most expensive and congested part of a mobile network-RAN. However this could be simply achieved by delivering the video content at just above the rate at which it is being viewed to make best use of radio resources.
- Network Caching of Content: Content that is popular may experience a high degree of demand. Caching the most popular content in the network allows bandwidth conservation in the backhaul (cache-fill) while reducing the latency incurred for stream initialization (content access). HTTP specifically provides a protocol description for caching content: RFC2756 – Hyper Text Caching Protocol. Enhanced cache control is provided by content distribution systems that leverage HTCP (Hypertext Cashing Protocol) methods for managed content. The industry has embellished the HTCP methods with vendor proprietary extensions (e.g. HTTP Live Streaming by Apple, HTTP Smooth Streaming by Microsoft, etc.); however, a new standard based on DASH-MPEG

(Dynamic Adaptive Streaming of HTTP) has been ratified for potential use across a diverse set of UE.

- Client Caching of Content: Content that a user has subscribed to may be queued for delivery when the client is unable or unwilling to retrieve the content. The content may be queued for packaging (encapsulation, encoding and encryption) allowing the network to prepare the content even when the client is off-line. In conjunction with the subscriber content packaging queue and content retrieval option, a content queuing function in the network could hold content for the subscriber until there is a sufficient network resources or when network resources are more economically viable for content delivery. Content may be packaged into discrete segments for incremental delivery or sent when the user is on broadband (Wi-Fi) or small cell coverage.
- Network-aware Content Adaptation: Content network in interaction with RAN decides whether to invoke content optimization functions such as trans-rating, pacing, or redirects to more compact asset formats. The network awareness could be provided as a historic network load and appropriate analytics applied to best optimize the content. The interaction may occur in several domains; therefore multiple standards bodies. RAN performance would be addressed by 3GPP while IP adaptation would be handled by IETF (Internet Engineering Task Force). The MPEG body would likely address the codecs.
- Client-aware Content Adaptation: Client applications will attempt to optimize the retrieval of content based on available network resources. Methods based on adaptive bit-rate streaming facilitate content delivery in the best possible manner given the available bandwidth. Mobile access incurs a wide variety of environments where bandwidth resources are variable; therefore, the client must adapt to the available resources. Clients should behave fairly but the network should protect itself from misbehaving clients by using Policy control functions. The network may offer different classes of service. The UE and its applications must be granted permission to use the different classes of service. Applications may register with a policy control point that allows that application access to the assigned class of service. The UE operating system and network must enforce access control for the assigned class of service.
- Provisioned Bandwidth: Applications may invoke the allocation of capacity or priority connections for content delivery enabling a more effective content delivery experience. Different classes of service may be provisioned for each connection type where content shedding can take place, preferably on lower class connections (e.g. using QCI (QoS Class Identifier) as an input to load shedding). The UE may leverage multiple access connections and profile application flows for distribution into the different connections based on the performance and characteristics of the application flow [3GPP TS 23.261, 24.303 and 23.402].
- Broadcast Bandwidth: MBMS broadcast and MBMS group based services enable popular downloads/streams that have a high probability of creating multiple concurrent streams per cell. The efficiency of MBMS versus unicast delivery has a dynamic characteristic where high concurrency may be short lived. Architecturally, the network may provide the content live or via carousel for a brief period while transitioning to unicast delivery as the popularity of the content decays over time. There is a strong UE dependency on the delivery architecture. The eMBMS (Evolved Multimedia Broadcast Multicast Service) protocol may be defined, but there are definitions of media encoding methods, integration of back-office environments, and client application behaviours that need to be defined.
- Non-Multimedia files like software will need the OS to work in co-ordination with the content delivery platform. Multimedia (videos) will be application based and can subscribe to open or closed broadcasts as needed. Efficiency is only possible if the UE and application are closely integrated. The UE OS will need to expose API's for the

application vendors to call on specific connections and define quality of service attributes available to the applications.

- Bandwidth Prioritization: Tiered user profiles in the form of QCI for dedicated bearers, or content profiles in the form of dedicated bearers for that specific traffic can be used to provide priority and prioritization or shedding in times of congestion. Furthermore, in 3GPP Release 10, eMPS (enhanced Multimedia Priority Services) can be used to increase the priority of a bearer on demand. This is typically used for emergency services.
- Content Context Routing: Applications can be configured to leverage dynamic site acceleration functions that are network-based such that diverse content types can be efficiently handled by different network-based functions. Traditionally, dynamic site acceleration functions reside in front of a data centre server bank. Invoking dynamic site acceleration at the edge of the network allows vectoring of different classes of content requests to different network-based content optimizations functions such as caches, proxy servers, and data centre resources. The introduction of a dynamic website accelerator northbound of the PGW may allow specific UE flows to be directed to a cache system while other flows are directed to an Origin Server.
- Bandwidth Augmentation and Off-load: The UE will have access to a variety of connection methods with different properties including coverage, bandwidth, cost, and mobility. Intelligent content applications will be able to assign transactions to the most appropriate connection method to optimize the perceived costs / benefits for content delivery. Initially, the optimization may leverage static policies established by either the operator or the subscriber. Eventually, the application will be able to dynamically associate transactions to connection options based on a dynamic set of criteria.
- Wi-Fi offload can be done either controlled by operator policy or by the end users own preferences. See below.
- Dynamic Adaptive Streaming over HTTP (DASH). And see Medium Term solution below.

  a) DASH may be a good method for 'reference media' distribution. The MPD can describe HLS (HTTP Live Streaming), MPEG2TS, HSS, and DASH from a single media representation. Eventually, DASH may replace HLS as a distribution method for unicast and broadcast.

  b) There are essentially two principal media container formats: MPEG2TS, ISO-BMFF. The former (MPEG2TS) is more easily represented in HLS and MPEG2TS streaming. The latter (ISO-BMFF) is more easily represented in HSS. MPEG-DASH supports both, MPEG2TS and ISO-BMFF based file formats for media segments. In both cases, the media is encoded in H.264 or MPEG4 with audio encoded in AAC. It is possible to create HLS/MPEG2TS from ISO-BMFF. Likewise, it is possible to remap MPEG2TS into HSS/ISO-BMFF. Fortunately, the first case (ISO-BMFF -> TS) is possible for HLS as there are many TS features that are not used by HLS.
- High Efficiency Video Encoding (HEVC). See below.

- **VS-1**: Video Compression
  - The choice of video codec affects quality and bandwidth. H.264 is now widely implemented and offers the best current video compression efficiency in deployed systems. Ensure that H.264 High Profile is supported and used by Smart Phones.
  - 3GPP have recently streamlined their video codec specifications so as to avoid codec proliferation. Note again that UE battery consumption is affected by the possibility to support video encoder/decoder hardware implementations.

  - A clearly defined list of supported codecs will avoid the proliferation of systems thereby allowing operators to implement the following solutions.

- **VS-2**: Adaptive Bit Rate Adoption:
- The short-term opportunity is to leverage the UE's ability to adapt to different content profiles for a given asset. This is applicable to both managed and un-managed content. While the network operator does not directly influence availability of un-managed adaptive bit rate content, the impetus to produce content appropriate for a wireless network is indirectly influenced by the network operator. The users will choose content that is effectively delivered over the wireless network resources.
  - The operator may offer optimized adaptive bit rate content that performs better on the wireless network resources and that shifts the demand for assets from un-managed content to managed content and use optimum codecs for the wireless infrastructure.

- **VS-3**: Client content caching and network off-load:
  - The availability of content at any time regardless of network connectivity is extremely important. Content can be divided into different content domains including live (potential for extremely high concurrency), popular video on-demand (high concurrency), long-tail video on-demand (low concurrency), and private video (no concurrency). The establishment of a content delivery profile in the application will allow the queuing and dissemination of content from optimal network locations. However, content that has high concurrency or high bit rate delivery requirements to the UE should be delivered using more economical connection methods or access types such as Wi-Fi or using Multicast.
  - Standardized profiles are needed that describe the characteristics of connection types such that applications can choose the appropriate connection based on per-subscriber policy. Connection characteristics such as relative cost, available bandwidth, latency, loss, and reliability are required for the application to optimize the available network resources.

- **VS-4**: Application based QoS control
  - P-GW/GGSN detects a particular traffic optionally with the support of a DPI (Deep Packet Inspection) function or a media proxy through PCC (e.g. online gaming) and initiates an upgrade or downgrade of QoS. QoS modification is applied end-to-end. This offers application prioritization based on application detection and enables “application detection and control over Gx” feature. Procedure for QoS modification is specified in 3GPP TS 23.401.
  - The operator may enter into arrangements with application providers to prioritize traffic that belongs to the application provider to improve quality of experience.

.

- **VS-5**: Network-based Caching:

Static content (even opaque content to the operator) may be cached in the network. The network cache will have little impact on the consumption of the air-link; however, it will off-load the backhaul capacity. The positioning of cache systems in closer proximity to the UE will also allow more efficient initialization of streaming sessions and content retrieval. The un-managed content must be dynamically detected and redirected for cache relevance while managed content can be specifically targeted for content distribution via caching systems. Caching of un-managed content is a challenge due to the perpetual change in content representation for over-the-top content. Caching of managed content is much more viable where the network operator specifies a deterministic set of content characteristics and representations.

- **VS-6**: Solution for broadcast scenarios

Within 3GPP, MBMS is defined from Release 6 onwards for UMTS, incorporating LTE from Release 9 onwards including Multicast/Broadcast over a single frequency network (MBSFN). MBMS is not currently widely deployed in any market, however the standards are stable and deployment interest is increasing in the context of LTE network optimization.

- eMBMS for eUTRAN and MBMS for UTRAN is defined by 3GPP. eMBMS handles the following services
    - Television and video streaming
    - File download, including static media
    - Carousel: Carousel is a service that combines aspects of both the Streaming and File download services described above. Similar to the streaming service this service includes time synchronisation. However, the target media of this service is only static media (e.g. text and/or still images). Time synchronization with other media is also required. For example, text objects are delivered and updated from time to time. Still images may also be collated to display low frame-rate video. In common with the download service this service also includes reliability (typically 100% reliability is not always necessary). The benefit of this service is that it is possible over a low bit-rate bearer.
- MBMS supports streaming services via RTP (Real-time Transport Protocol) and download services via FLUTE (File Delivery over Unidirectional Transport). From Release 9, MBMS supports carrying HTTP streaming content (called AHS in Release 9, DASH in Release 10) via the download delivery mechanism. DASH, including 3GP-DASH and MPEG-DASH is expected to become increasingly dominant over RTP for the delivery of video content. Utilizing DASH over MBMS is an efficient mechanism for delivering near real-time video in a scalable manner.
    - eMBMS can be used as an offload mechanism for popular content in areas with many users. It can also be used to push content towards UEs during off-peak hours (e.g. trailers and ads, software updates).

- **VS-7**: Offload Solutions
    - WLAN Offload for selective content is defined by 3GPP and can be supported with existing devices with minor enhancements. There are UE specific and vendor proprietary solutions currently available to facilitate Wi-Fi access selection. These are explored in Section 3.4 of this document.
    - Other relevant bodies are IEEE, WFA and WBA. Several initiatives are ongoing to align the interworking requirements, e.g. 3GPP-WFA alignment, NGMN and the GSMA Terminal Steering Group (TSG).

- **VS-8**: Femto/Small cells/H(e)NB

Solutions are defined by 3GPP and are transparent to the existing UEs in the network. These solutions may be used by operators to augment coverage and capacity as needed. H(e)NB deployment would require introduction of SeGW into the network architecture.

Operators may consider introducing HeNB (Home evolved NodeB) or HeNB-GW to limit the signalling load on the macro network MME.

- **VS-9**: Service based throttling

Locally define rules or rules enabled are provided through PCRF in P-GW/GGSN can be used to limit the bit rates for certain traffic profiles. Traffic Shaping and throttled bit rates can be used to define APN-AMBR (Access Point Name-Aggregate Maximum Bit Rate), which in turn can be used by the MME to set UE-AMBR that is applied by the eNB. Traffic Shaping / Throttling can also be used to ensure that each application does not exceed the bit rates reserved for this application thereby contributing to resource savings.

- **VS-10**: Network-assigned Priority - Allocating a special QCI / application specific QCI

P-GW/GGSN detects the applications using a certain predefined rules, properties and profiles with similar characteristics and are mapped to a predefined QCI for that bearer. P-GW / GGSN may request a dedicated bearer with appropriate QCI. Assigning a special QCI based on traffic characteristics can help optimize radio resource allocation. Tiered user profiles in the form of QCI for default bearers, or content profiles in the form of dedicated bearers for that specific traffic can be used provide priority and prioritization or shedding in times of congestion.

- **VS-11**: Enhanced RAN CORE Interaction for resource usage
    - Centralized node to maintain historic data of the radio resource usage and to interact with RAN in order to provide for optimized Radio parameters based on content being pushed.
    - Traffic usage, collected for example from network taps on the S1 interface, can provide historical usage which can be used to predict how much capacity is available and then be used to modify the manifest to suit the available bandwidth.

- **VS-12**: Cell on Wheels (COWs) and EPC in a Box

Pre-planned and expected load could potentially be handed over to localized networks for a short period of time.

- **VS-13**: UE function changes

UE Specifications may include a requirement to support a timing mechanism for unicast access to the network. However there needs to be a similar approach for coordinating receipt of broadcast content. See description in SD-2.

- **VS-14**: Dynamic an Adaptive Streaming over HTTP (DASH).
    - Live and on-demand video streaming standard where client controls downloading of a succession of segments of video files over HTTP. Reuses widely deployed standard HTTP servers/caches for scalable delivery.
    - Specified in both MPEG and 3GPP (TS 26.247), adopted by other forums, e.g. Open IPTV Forum.
    - Common industry profile being defined in the DASH promoter's group (http://dashpg.com/).

  - IMTC handling client-server interoperability testing.
  - Available as Unicast and also broadcast in eMBMS context.
  - DASH addresses several of the potential overload problems:
    - Content server offers content with various versions (bit rate/resolution) so as to match UE capabilities and access networks. Note, the server does not know the UE capabilities. It knows the set of possible capabilities and offers that set to the UE. The UE picks the capability that is most appropriate given the current environment.
    - Client driven adaptation can switch dynamically between those representations.
    - HTTP offers native support for content caching (both UE and network).
    - UE battery savings can be made by adapting segment size.
    - DASH264 profile should limit codec proliferation and increase content availability.
    - 3GPP DASH QoS support via PCC

For managed content all these optimizations are available to operators. Further OTT content may be optimized assuming the network operator has access to MPDs (manifest files) and video segments (not-encrypted). The MPD description can be re-written by the network operator to describe a different set of video codec options where the video segments have been transcoded to a more efficient format for delivery. Both the media and the MPD must not be encrypted; otherwise, the network operator must deliver the content as an opaque object with no optimization options.

- **VS-15**: Video Compression
  - The High Efficiency Video Coding (HEVC, H.265) standard is being defined in JCT-VC (Joint Collaborative Team on Video Coding of ITU-T and ISO/MPEG). It expected be published in January 2013.
  - Very significant efficiency gains are expected from this new codec. Video services standard and content providers should gradually adopt HEVC at a point where implementations are readily available.

- **VS-16**: Client Caching

Mechanisms are being discussed in fora like HQME/IEEE.

- **VS-17**: DASH improvements
  - 3GPP DASH Unicast/Broadcast (eMBMS) dynamic switching is being developed in 3GPP in Release 11.
  - DASH QoS handling via PCC is also being studied for improvements in 3GPP.
  - DASH content adaptation
    - Network based adaptation of content and content selection (MPD/manifest file). Should not impact UNI so as to improve DASH adoption and deployments. I.e. must be transparent to client-server interface. Over the top video services including too high resolutions and bitrates may be adapted via intermediate functions under certain conditions. No specific standardization work is required for these.
    - Client rate adaptation is currently not specified by 3GPP or MPEG. 3GPP is working on guidelines in Rel-11 so that clients are consistent in their

behaviour and do not request as much bandwidth as possible, waste resources and compete with other clients sharing the same resource.

- **VS-18**: Network-assigned Capacity:
    - The UE and/or its applications may be assigned connections that are appropriate for the type of content delivery. This model will use a fairly static policy where the content profile is well understood and the resources required to deliver the content are readily managed. However there needs to be a mechanism for the application to request access to specific connections and be granted access. There are proprietary solutions being developed to facilitate these procedures.
    - The availability of MBMS resources can be leveraged for specific content classes where there is a high-concurrency probability. Most likely, these content classes will be focused on specific venues where the UE is registered as a client of the MBMS connection method.

- **VS-19**: Optimization based on congestion awareness
    - EPC and RAN to interact to provide Video optimization based on network load and not just on UE capabilities. Video optimization would right size the video content for the available resources at the RAN.
    - Additional enhancements / optimization is possible by adding network and user awareness in the network, which could maintain historical network and user behaviour. Until 3GPP standardizes the procedure for collating network conditions (see VS-32), the solutions described here are vendor specific yet could provide for:
        - Optimized streaming content – identifying rates for the user / network and provide that into the content servers and as necessary intercept the offered media to limit it to lower bandwidth options. The intent is to limit the content choices based upon the radio bandwidth capacity availability.
        - UE and network together identify the best offload mechanism – either shedding low priority or low tier services vs. offloading.
    - Function would be responsible for identifying appropriate:
        - Access technology and mechanism – provide decisions into existing ANDSF, (Access Network Discovery and Selection Function).
        - Gate the timing of delivery – provide decisions into the application gateway or content provider (offline video download).
        - Prioritize operator approved content vs. over the top content – provide decisions into the PCRF or apply local policies.

- **VS-20**: Application API / OneAPI

Provide for a common 3GPP or GSMA defined API access to over the top applications that can invoke a specific PCC rule from the operator network. Charging could be user based or application / content based. This would require an application gateway API to interact with the network as well as any failures to apply the requested PCC. This concept can be progressed in GSMA NAPI and the interaction between the network and application providers needs to be investigated.

- **VS-21**: Content Authorization
  - Today the BSF (Bootstrapping Server Function) is used for authorizing the user for content and access to the user credentials is provided via the NAF access via Zh interface for MBMS and other applications.
  - The BSF is used in network application/service authorization. After running bootstrapping with UE and generating the bootstrapping key Ks, the BSF provides user specific keys (called Ks_NAF) to the NAF (e.g. MBMS application) so that the NAF can use these keys for its own application specific purposes, like authentication, encryption, authorization, etc. The BSF is not involved in the handling of the keys after it has given them to the NAF. Once the application/service is started, fine/medium-grained access content authorization is managed by the NAF, for example in (e)MBMS.
  - Note that GBA (Generic Bootstrapping Architecture) adds additional load on the HSS, which is already challenged, with network authentication load.
  - The BSF needs to contact the HSS via Zh only when GBA bootstrapping is done (and this is when the bootstrapping key Ks is generated). This may be quite seldom, since the BSF can derive user specific NAF keys (Ks_NAF) for many applications/services using the same bootstrapping key (Ks). New bootstrapping needs to be done only when the bootstrapping key expires (lifetime configured by the operator) or if an application requires a fresh bootstrapping key to be used. This was intentionally designed to minimize the load on HSS. Overall, any application/service that wants to leverage the HSS & SIM security association would impose an extra-load on the HSS, e.g. EAP-SIM/AKA or IMS. But GBA is the optimal approach, while others are sub-optimal.
  - Possible optimization would be to separate content authorization via an external entity via a download of user authorization credentials or priority based access to credentials to ensure HSS can focus on network authorization during an overload scenario. This would require NAFs to be defined with different priorities.
  - GBA is already doing that, i.e. "to separate content authorization via an external entity via download of user authorization credentials ". The external entity is the NAF. Note that GBA have possibility to categorize NAFs to NAF groups, e.g. based on priority.

- **VS-22**: UE function changes

Changes needed to the OS, applications or UE connection manager to synchronize when files, updates and content can be sent over multicast. GSMA's recommendations on application scheduling [16] define a timing mechanism for unicast access to the network. However there needs to be a similar approach for coordinating receipt of broadcast content. Hence changes to the OS, applications or UE connection manager to synchronize when files, updates and content can be sent over multicast.

- **VS-23**: Offload Solutions

See Section 3.4 of this document for further details.

- **VS-24**: Trusted 3rd Party access

- Utilize the existing Rx interface to provide access to operator PCC for operator trusted 3rd party access. Security concerns provide for certain obstacles on extending similar services to all over the top applications. The Rx interface to non 3GPP elements is a proprietary implementation that will require agreement between the PCRF and application vendors.
- Alternatively, the interface between the application and PCRF can be SOAP or Restful API's which are defined between the two vendors.

- **VS-25**: UE OS and browser for closer application integration

UE OS and browsers needs to provide for more APIs for application vendors to invoke specific connections and be able to define QoS attributes for the ability to handle differences between Non-Multimedia files like software vs. Multimedia download.

Providing APIs from device level to application shall allow for applications to choose the most appropriate connection method to optimize for content delivery. Connection API's will need to be exposed to the application in order to assign flows to the appropriate connection. There is a possibility for GSMA to provide input into W3C.

- **VS-26**: Negotiate end of transmission

Negotiate "size / duration of transfer" or "end marker" when bearer context is being established between the UE and the network. Upon detection of end of transmission, P-GW / GGSN can release the resources that are not required when there is no transmission. End of transmission can be indicated to the radio network to assist in deciding when the RRC connection can be released thus saving radio resources. Alternatively, P-GW/GGSN can initiate deactivation of bearers upon detection of end of transmission.

- **VS-27**: Network Augmentation:

The UE connection manager, the network and the application will have the intelligence to arbitrate between the most cost-effective connection methods. Specific transactions may leverage a macro wireless network coverage to insure continuity of the content availability, while other content delivery transactions may leverage the macro or micro wireless network to optimally deliver content based on content classes and priority. Client / Server solutions that perform network selection are available in the short/medium term, however solutions that are able to effectively arbitrate between the available networks will need extensions to the IEEE 802.11 specifications to not only be aware of the radio congestion but also the backhaul conditions.

- **VS-28**: Trusted 3rd Party access

The GSMA OneAPI architecture supports a set of network APIs, however these could be extended to provide network capacity and availability to allow 3rd Parties to access and contact the subscribers at the appropriate time and RAN connection to avoid congesting the network.

- **VS-29**: Content Context Routing

  UE applications will be able to distinguish between different classes of content for a given application and route the content retrieval requests to different network-based content delivery services. Some of the transactions may be optimized through dynamic site acceleration while the application may leverage local and remote cache infrastructures for static content. Caching and CDN solutions are currently available however there is little interaction with the UE. Enhancements on how the UE makes content requests will require standardization work.

- **VS-30**: Network-based Content Availability
  - The network infrastructure will automatically ascertain the demand for certain classes of content and dynamically instantiate more efficient content distribution methods through MBMS. UE devices will dynamically adapt to the more efficient distribution method while in proximity of the MBMS broadcast asset. As the concurrency for the content diminishes or the UE moves out of the coverage of the MBMS delivery method, the UE will revert to a unicast delivery method as the content has become less popular and has low concurrency.
  - The standards support the delivery method for eMBMS. It does not define the back-office systems required to enable the application environment. Access to content is controlled by the video application servers. The availability of content delivery options needs to be advertised to the client in a standardized manner.

- **VS-31**: OMA Dynamic Content Delivery (DCD)

The DCD vision was specified for managed content delivery to mobile devices, with intelligent pre-caching and a metadata-driven, syndicated/channelized content service design methodology.

http://www.openmobilealliance.org/Technical/Release_program/dcd_v1_0.aspx

- **VS-32**: RAN Congestion handling
  - Mechanisms to collect RAN feedback or via query/response to allow core network/application layer to react as needed and to provide application awareness to the RAN are desirable.

A simplified version of the network awareness platform is depicted below where information from the various network elements is collected and analysed to provide a near real time optimization of the network.

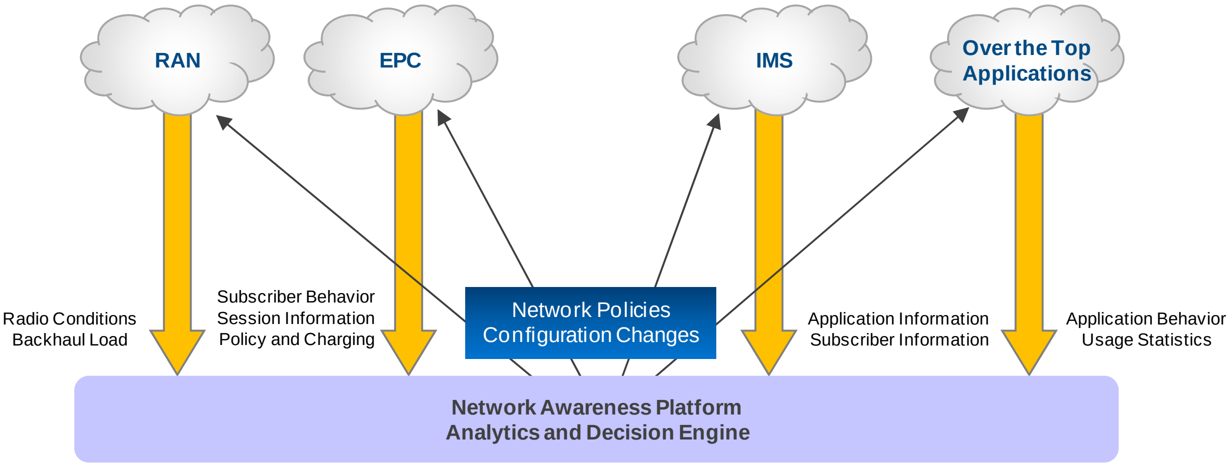

Figure 3: : Network Awareness and Feedback Function

3.4 Access Network Changes (AN)

3.4.1 Problem Statement and Motivation

Operators are looking to leverage other access technologies to provide additional coverage and capacity to their 3GPP coverage. WLAN technologies can do a lot to alleviate the load on the 3GPP coverage, however they also add to the signalling plane to the AAA/HSS and other elements that are needed when there is a RAT (Radio Access Technology) change. Another point to consider is that the availability of WLAN may not lead to a balanced usage between the different radio networks and the user experience could be impacted. WLAN networks are pervasive and available in the home as well as managed deployments in public spaces and offices. However access may be limited (e.g. UE does not successfully select and associate with an AP). This reduces the benefit and the solutions may not provide optimum user experience because mobility between 3GPP and WLAN can disrupt applications. WLAN can be used in a more seamless way if solutions are deployed to;

- Network Selection: Assist the UE automatically select suitable WLAN Access Points (using ANDSF rules and WLAN Standards features advertised as part of WFA Hot Spot 2.0 specifications).
- Automatic Authentication Procedures: Support for 3GPP EAP methods to authenticate on the WLAN network as part of HS2.0.
- Session Mobility: When needed, the UE and network should support seamless mobility between 3GPP and WLAN coverage that preserves the IP address / Prefix allocated to the UE and the application connection.

3.4.1.1 Examples scenarios:

- Change of access technologies between 3GPP and WLAN due to mobility
- Shifting applications from one access to another access due to user preference, operator policy, charging considerations, etc.

3.4.1.2 UE and Network considerations:

3.4.1.2.1 Network Considerations

- 3GPP to WLAN mobility is defined in 23.402 specifications There may be optimizations and changes needed to make the procedures more effective.
- There are many networks that the UE can connect to; including 3GPP networks, WLAN access in trusted and un-trusted mode. The mobility management and network selection function on the UE is controlled by network policies defined in 24.312 ANDSF specification. 3GPP ANDSF related specifications can be enhanced to address the capabilities of Hotspot 2.0.
- In case of WLAN access, the UE takes the final decision on where to camp (controlled by ANDSF policies). Nevertheless, leaving the UE to select the network may have undesired consequences (e.g. WLAN congestion), and it may be necessary to have more network control.
- Many devices have a local policy to restrict delivery of large files and content for when the UE is connected to WLAN. This policy is only effective if the UE does regularly connect to WLAN networks. If the user chooses not to connect to WLAN then methods are needed to modify such policies to ensure the content is delivered in a timely manner.
- In situations where WLAN and 3GPP networks are available there is a danger that the UE could frequently hand off between the networks and resulting re-authentication and mobility related signalling. There should be means to reduce control plane load and mobility related signalling. The UE should also prevent undue handoffs.

3.4.1.2.2 UE Considerations

- As described above, the UE needs to have controls to limit the number of handoffs to both reduce the (re)authentication procedures and reduce the impact of handoffs. The UE also needs to manage how long it maintains the authentication credentials to improve the handoff signalling.
- The UE has limited ability to measure the load on the network and make correct discussions about which access network offers the optimum experience. Proprietary UE-centric solutions with limited capabilities are available today, however network-based standard solutions are needed in the future.
- When selecting WLAN, some UEs have been observed to initiate the PDP context deactivation procedure to the 3G networks, and hence also when returning to 3GPP networks after leaving WLAN coverage, to (re-)initiate the PDP context activation procedure. This behaviour could lead to control plane loading especially where WLAN coverage exists sparsely, or when a group of UEs simultaneously enters WLAN coverage (e.g. train arrives at a station where WLAN coverage is provided). To mitigate control plane loading, it is preferable that such UE behaviour is improved.

3.4.2 Solutions

- **AN-1**: Context maintenance

The authentication entity can maintain the context or as described in Section 3.1.2.2 (SD-3) of this document**.**, maintain multiple vectors. This is available through vendor proprietary solutions.

- **AN-2**: WLAN Access Point Selection

Early proprietary solutions architectures allow local configuration or for operators to push preferred policies to devices.

- Application based network selection whereby the device has a local policy to select WLAN for certain applications. These are typically user controlled.
- Location based network selection whereby the device is configured with information on appropriate WLAN network based on the current user location
- Periodic updates of WLAN Access network list pulled by the UE.
- Time based selection whereby the device is provided with information to choose a WLAN network.
- References of vendor solutions: http://www.birdstep.com

- **AN-3**: WLAN Quality Estimation

Using this mechanism, the operator controls the network selection performed by the terminal via proprietary solutions. When a WLAN network is detected, the UE assesses the suitability of the network by checking for connectivity to the Internet, bitrate available to the Internet and quality of the radio link. This information is used by the UE to select between the WLAN and the cellular network. Such framework ensures that users have a good WLAN experience, thereby reducing the number of users who disable WLAN all-together (references provided below).

http://www.qualcomm.com/about/research/projects/traffic-management/wqe

http://www.airsensewireless.com/

- **AN-4**: Battery Life Optimization

Use vendor provided policy to enable / disable the secondary radio in the device depending upon the location or network conditions

- **AN-5**: When a UE performs IEEE 802.1x authentication with an AP but moves to new AP, the authentication credentials can be re-used when IEEE 802.11r procedures are used. The first time a UE associates to an AP, it performs a full authentication (meaning back to H-AAA); every time a mobile associates to a new AP thereafter, the [fast] re-authentication is handled locally by the AAA Proxy.

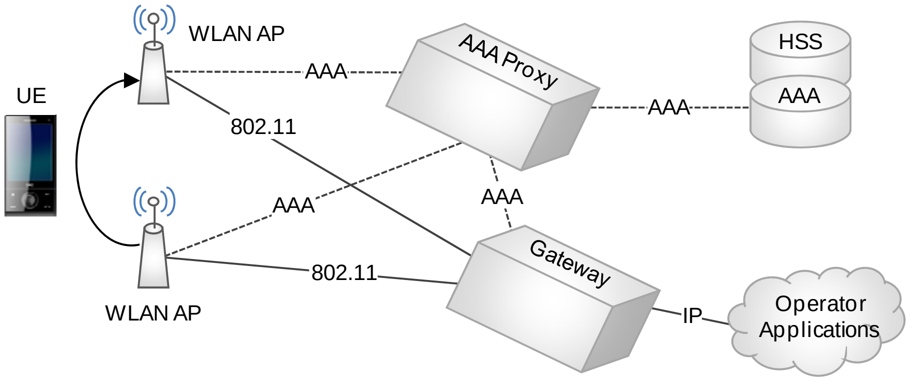

Figure 4: : Fast Transition with IEEE 802.11r procedures

- **AN-6**: Offload Mechanisms
    - ANDSF based architecture provides for pushing operator preferred policies
        - Application based network selection solution as relying on 3GPP DIDA, as specified in 3GPP TS 24.302 and 24.312.

- **AN-7**: Study additional enhancements / optimization to provide a feedback loop to avoid congesting non-3GPP networks. Historical and current WLAN conditions could be used to provide feedback into ANDSF to push timely updated policies and access lists to the appropriate UEs. This work could be combined with VS-30.
    - Define a common feedback mechanism from all network elements.
    - Load balance the different networks.

- **AN-8**: Study the Integration / interworking between the non-3GPP access mobility and security control with 3GPP session and mobility management.

As an example and interface between the Trusted Wireless Access Gateway and the MME, could be envisaged with similar functions as the S10 interface in LTE, to retrieve the context information and potentially avoid the need to query the AAA and also simplify the authentication procedures. It has nevertheless to be noted that:

- The same EAP-AKA’ security parameters cannot be used over 3GPP and Non 3GPP access as the Access Network Identifier (ANID) are different and the ANID is an input to the key derivation function per AKA’
- There is the need to contact the HSS at the mobility between 3GPP and non-3GPP, in order for the HSS to know towards which serving node it has

to send potential changes of subscription data / requests of immediate service termination

Figure 5 illustrates an idea to add an interface between the Trusted Wireless Access Gateway and the MME, shown in red below (identified by the letters SZZ as an example), with similar functions as the S10 interface in LTE, that can be used to retrieve the context information.

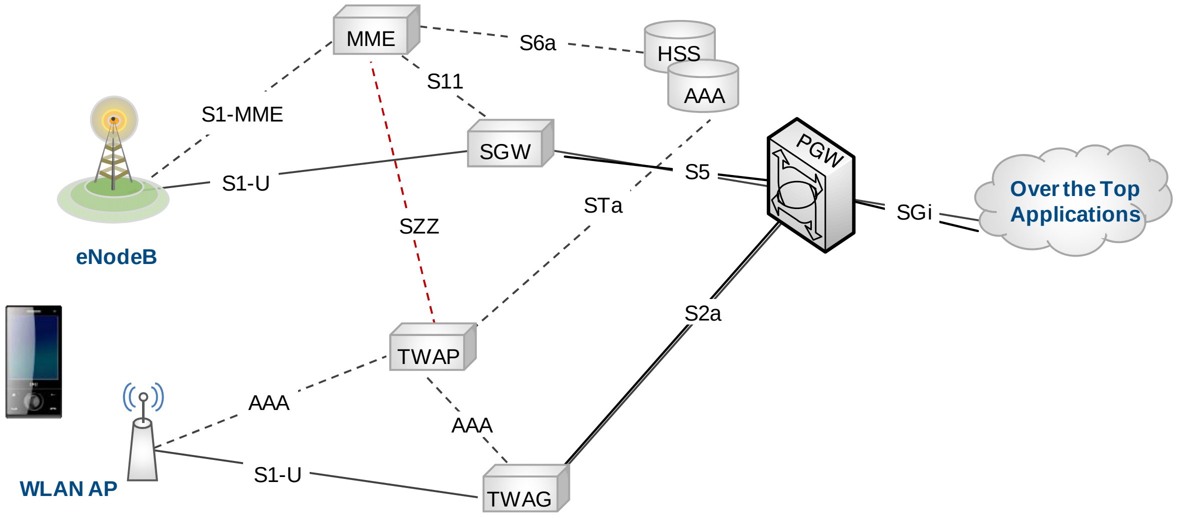

Figure 5: Possible 3GPP to WLAN Interface

- **AN-9**: ANDSF Enhancements
  - Improve ANDSF specifications to take into account network capabilities advertised per Hot Spot 2.0

3.5 Group Features (GF)

3.5.1 Problem Statement and Motivation

Groups are intended for use with devices and applications that can easily be grouped to enable optimization of network resource usage. A group of devices could be defined by the network operator or based on agreements between the service provider and the operator.

Certain applications (e.g. Facebook, Twitter, Skype) are generally used by a group of devices. In some cases, service providers subscribe for more than 1000 subscriptions and use a single application in these devices.

From both end-customer and mobile network operator perspective, there is benefit in optimized handling of groups of devices/subscriptions. This can be, e.g., the ability to trigger a group of devices with one trigger message, the ability to enforce a QoS policy for a group of devices, optimization of charging for groups of devices that need to be aggregated into one bill to a single customer.

Group based policing can be used to enforce a QoS policy for a group of devices/subscriptions. This allows greater flexibility to the application / application owner compared to individual policies for each of the devices/subscriptions, while at the same time ensuring the operator that the particular group of devices/subscriptions does not unduly load the network.

Group based triggering can be used by the application server to trigger a group of devices (e.g. notification server can trigger the devices for software upgrade notification). Devices that are members of the target group are configured to recognize the broadcast message as a trigger. The network broadcasts this message within a geographic area specified by the Service Capability Server.

Group based charging is to increase charging efficiency for group based applications. In many cases, the data volume of CDRs (Call Detail Record) generated by applications is greater than the volume of actual user data transmitted. In these cases it may be beneficial to create bulk CDRs to count chargeable events per group instead of CDR creation per individual subscription. Especially when UE initiated signalling needs to be charged (to reduce signalling overhead), group based charging can prevent a surge of CDRs that need to be handled, sent, and aggregated.

Examples:

- Subscription information by device type, by OS type
- QoS policy by group
- Group-based addressing (slow down/shut down devices with misbehaving apps, devices in particular location, etc.)
- Software upgrade notifications
- Non-subscription based group features
- Application based group features (Social networking)

3.5.1.1 Main Considerations

Benefit of applying group based features depends on the optimal grouping of devices. This is important because employing group based features (e.g. broadcasting, IP multicasting) could consume significant amount of core network and radio resources hence care needs to be taken in grouping devices appropriately. Grouping could be performed based on location, similar application running in the device, owner of the device (e.g. service provider subscribe for 1000’s of devices running the same application) or based on similar features subscribed and supported by the device.

3.5.2 Solutions

- **GF-1**: Cell broadcast to devices in a certain location as specified in 3GPP TS 23.041.

Note: Cell broadcast in EPS (Evolved Packet System) /LTE is currently limited to Warning message delivery (PWS) only.

Note: Cell Broadcast Solution (CBS) specified in 3GPP TS 23.041 cannot be deployed in a network sharing environment.

- The CBS service is analogous to the Teletext service offered on television, in that like Teletext, it permits a number of unacknowledged general CBS messages to be broadcast to all receivers within a particular region. CBS messages are broadcast to defined geographical areas known as cell broadcast areas. These areas may comprise of one or more cells, or may comprise the entire PLMN (Public Land Mobile Network). Individual CBS messages will be assigned their own geographical coverage areas by mutual agreement between the information provider and the PLMN operator. CBS messages may originate from a number of Cell Broadcast Entities (CBEs), which are connected to the CBC (Cell Broadcast Centre). CBS messages are then sent from the CBC to the cells, in accordance with the CBS's coverage requirements.
- CBS can be used for broadcast services in a proprietary manner (e.g. interfaces to CBE, addressing needed)

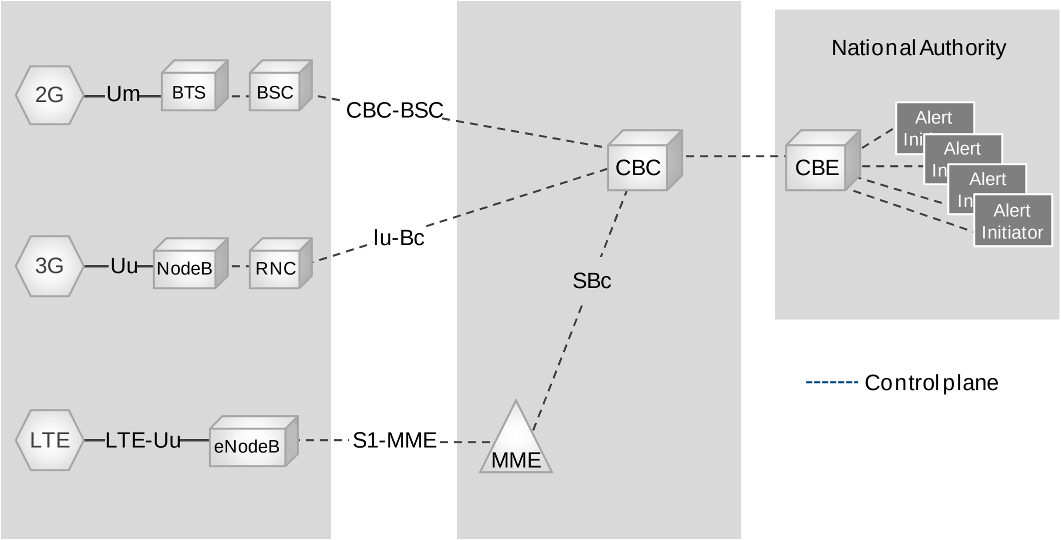

Figure 6: Cell broadcast architecture in GERAN, UTRAN, LTE

- **GF-2**: MBMS Broadcast to groups of devices in a particular location as specified in 3GPP TS 36.331:

The target design for Release 9 EPS functionality was limited to enable Mobile TV and scheduled file downloads. Therefore MBMS for EPS only supports MBMS Broadcast mode (as defined in 3GPP TS 23.246). The MBMS Broadcast Mode differs from Multicast mode in that there is no requirement for a user to join or activate the service. The MBMS Broadcast mode functionality in E-UTRAN differs from Release 8 and UTRAN in that E-UTRAN does not support counting, and data are broadcasted to predefined areas regardless whether there are any UEs in this area.

- **GF-3**: Enhance cell broadcast services (specified in 3GPP TS 23.041) for LTE:

The CBS service described in Section 3.5.2 of this document could be enhanced to support additional message identifiers, message content for broadcasting unacknowledged messages to devices within the cell. Enhancements to addressing are also needed.

- **GF-4**: Definition of globally unique group identifiers in the subscription data

A group of devices could be defined by the network operator or based on agreements between the service provider and the operator. Upon subscription, globally unique group identifier could be defined for devices within the group.

- **GF-5**: Introduction of group based charging using group identifier as a correlation identifier:

Group based charging is to increase charging efficiency for group based applications. In many cases, the data volume of CDRs generated by applications is greater than the volume of actual user data transmitted. In these cases it may be beneficial to create bulk CDRs to count chargeable events per group instead of CDR creation per individual subscription. Especially when UE initiated signalling needs to be charged (to reduce signalling overhead), group based charging can prevent a surge of CDRs that need to be handled, sent, and aggregated. Group ID can be used as a correlation identifier to enable group based charging.

- **GF-6**: Introduction of group based policing under the assumption that all devices are connected to the same P-GW/GGSN.

Group based policing can be used to enforce a QoS policy (e.g. group APN-AMBR) for a group of devices / subscriptions. This allows greater flexibility to the application / application owner compared to individual policies for each of the devices/subscriptions, while at the same time ensuring the operator that the particular group of devices/subscriptions does not unduly load the network.

- **GF-7**: IP Multicast for groups of devices in a particular location running the same application as specified in 3GPP TS 23.246
  - For GPRS in GGSN it is possible, by configuration, to enable distribution of MBMS payload by using IP Multicast in the backbone network. IP Multicast distribution is done from GGSN to RNC (Radio Network Controller) downstream nodes or from MBMS GW RNC downstream nodes.
  - For EPS in MBMS GW it is possible, by configuration, to enable distribution of MBMS payload by using IP Multicast in the backbone network. IP Multicast distribution is done from MBMS GW to eNodeB.

This can be beneficial when groups are devices are running the same application with similar characteristics and are being triggered at the same time. This can combined with group identifier and group subscriptions defined in the previous solutions.

- **GF-8**: Define group based features such as group triggering (specified in 3GPP TR 23.888):
  - When the network needs to trigger the devices in a group (to react on notification from the network e.g. for software upgrade notification), group triggering feature can be employed to save backhaul signalling and radio resources.
  - If the devices are idle when they are being triggered, then the network needs to page the devices. It is beneficial to page all the devices in a group with a single paging message in order to save bandwidth on the paging channel. When group paging is performed, it is important to ensure not all the devices try to access the network at the same time to avoid overload and congestion.

3.6 Application/Device Monitoring (AD)

3.6.1 Problem Statement and Motivation

Monitoring is an essential feature for the following reasons:

- To enable the network to react to misbehaving devices, misbehaving apps by device category, application category etc.
- Devices can be deployed in remote areas and in locations where they are not monitored actively by humans, probability of theft and vandalism is really high.

So it is expected that operator networks provide the mechanism to auto-detect suspicious activities e.g. change of association between UE and UICC, loss of connectivity, communication failure, change of location, denial of service attacks and in general, detect behaviour that is not aligned with subscribed features. These events are currently neither detected nor reported. Hence this feature requires the networks to enable detection of events and report these events as and when they occur so the service provider or the user can take appropriate action.

Example scenarios:

- Monitoring by device category, by application, etc. (to enable the network to react to misbehaving devices, misbehaving apps, etc.)
- Location changes, presence updates, etc.
- Using above network information to trigger services
- Detection of distributed denial of service attacks

3.6.2 Solutions

- **AD-1**: Usage Monitoring:

Monitoring excessive data usage can be detected by utilizing the usage monitoring feature specified in 3GPP TS 23.203.

  - Application layer registers with the PCRF for notification of excessive data usage. PCRF sets and sends the applicable thresholds to the PCEF or TDF for monitoring. When the usage exceeds a certain limit (e.g. threshold specified). PCEF or TDF notifies the PCRF or TDF which then notifies the AF via Rx interface.

Note: This support is available in the Rx interface only for sponsored data connectivity (e.g. ads sponsored by Google).

- **AD-2**: Enhanced charging records:

Include additional information in the CDRs. Charging records can be enhanced to include the following information. This can then be used by operators to either notify the user to take action or for statistical purposes:

  - Excessive data used by applications to monitor sponsored data connectivity (e.g. ads sponsored by Google).
  - UE behaviour with respect to subscribed features (e.g. UE trying to initiate CS service when subscribed only for PS services)

- **AD-3**: Usage monitoring feature for general data usage:

Usage monitoring feature and Rx interface specified in 3GPP TS 23.203 can be enhanced to support general data usage.

- **AD-4**: Enhancement in the packet core network to monitor additional events:

Application or a monitoring server registers with the HSS/PCRF which in turn registers with the serving node (or interworking function registers directly with the serving node) for notification of certain monitoring events. Packet core network enhanced to provide monitoring services for events such as the following:

    - Change in point of attachment (e.g. Location changes, presence updates)
    - Detection of distributed denial of service attacks
    - Association of the device and UICC (i.e. IMSI / IMEI mapping)
    - Alignment of subscribed features (e.g. e.g. UE trying to initiate CS service when subscribed only for PS services)

Detection of signalling loops (e.g. continuous PDP attempts, continuous attaches/RAU/TAUs, etc)

Annex A: Background on Industry Push Servers

The Open Mobile Alliance (OMA) document OMA-AD-Push-V2_2-20071002-C outlines the Push Architecture and related specifications, which together specify an enabler for a service to push content to mobile devices. A new work item in OMA WID 0263 - Always Online Infrastructure has been started.

UE OS Vendors have specific solutions available today and are summarised below.

Apple Push Notification Server (APNs):

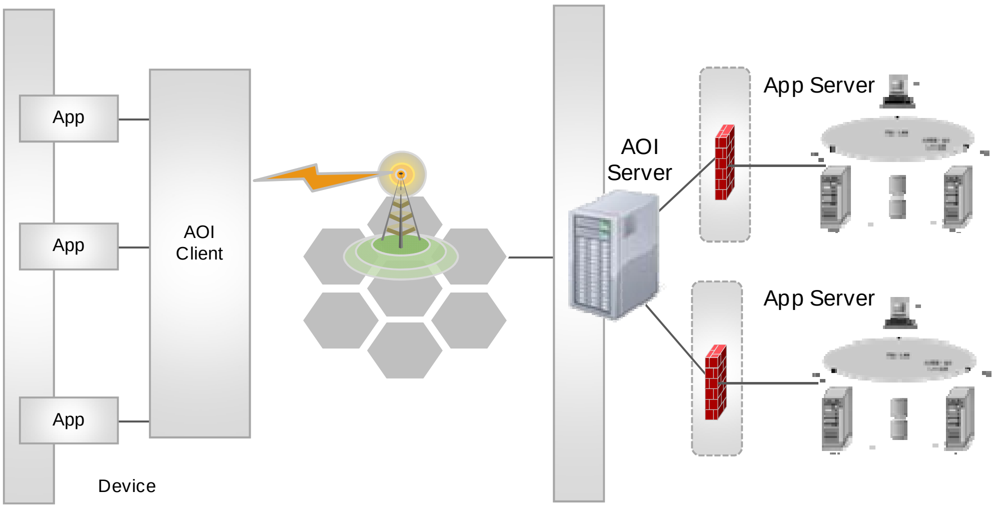

Figure 7: Apple Push Notification Server

Apple Push Notification service (APNs for short) is the centrepiece of the push notifications feature. It is a robust and highly efficient service for propagating information to devices such as iPhone, iPad, and iPod touch devices. Each device establishes an accredited and encrypted IP connection with the service and receives notifications over this persistent connection. If a notification for an application arrives when that application is not running, the device alerts the user that the application has data waiting for it.

Software developers (“providers”) originate the notifications in their server software. The provider connects with APNs through a persistent and secure channel while monitoring incoming data intended for their client applications. When new data for an application arrives, the provider prepares and sends a notification through the channel to APNs, which pushes the notification to the target device.

Apple Push Notification service transports and routes a notification from a given provider to a given device. A notification is a short message consisting of two major pieces of data: the device token and the payload. The device token is analogous to a phone number; it contains information that enables APNs to locate the device on which the client application is installed.

The flow of remote-notification data is one-way. The provider composes a notification package that includes the device token for a client application and the payload. The provider sends the notification to APNs which in turn pushes the notification to the device.

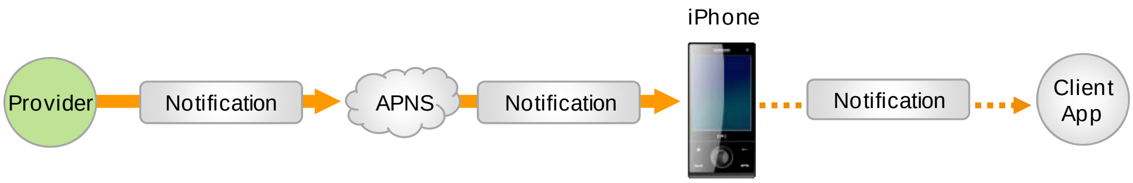

Figure 8: A push notification from a provider to a client application

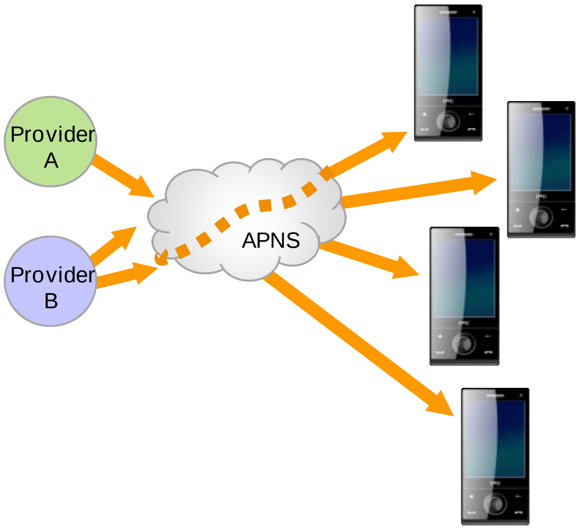

Figure 9: Push notifications from multiple providers to multiple devices

- Feedback Service: Sometimes APNs might attempt to deliver notifications for an application on a device, but the device may repeatedly refuse delivery because there is no target application. This often happens when the user has uninstalled the application. In these cases, APNs informs the provider through a feedback service that the provider connects with. The feedback service maintains a list of devices per application for which there were recent, repeated failed attempts to deliver notifications. The provider should obtain this list of devices and stop sending notifications to them. For more on this service, see “The Feedback Service.”
- Quality of Service: Apple Push Notification Service includes a default Quality of Service (QoS) component that performs a store-and-forward function. If an APN attempts to deliver a notification but the device is offline, the QoS stores the notification. It retains only one notification per application on a device: the last notification received from a provider for that application. When the offline device later reconnects, the QoS forwards the stored notification to the device. The QoS retains a notification for a limited period before deleting it.

GOOGLE C2DM (Cloud to device Messaging)

https://developers.google.com/android/c2dm/

- Android Cloud to Device Messaging (C2DM) is a service that helps developers send data from servers to their applications on Android devices. The service provides a mechanism that application servers can use to tell mobile applications to contact the server directly, to fetch updated application or user data. The C2DM service handles all aspects of queuing of messages and delivery to the target application running on the target device. It is therefore conceptually similar to APNs by apple.

- It is to be regarded as a source of push traffic rather than some tool to help in addressing issues related to Network initiated traffic.

Contributors

Name	Company
Kalyani Bogineni (Technical Lead)	Verizon
Guenter Klas	Vodafone
Chris Pudney	Vodafone
Deborah Barclay	Alcatel-Lucent
Alessio Casati	Alcatel-Lucent
Laurent Thiebaut	Alcatel-Lucent
Ravi Guntupalli	Cisco
Aeneas Dodd-Noble	Cisco
Vojislav Vucetic	Cisco
Scott Wainner	Cisco
Frederic Gabin	Ericsson
Magnus Olsson	Ericsson
Henrik Voigt	Ericsson
Devaki Chandramouli	Nokia Siemens Networks
Rainer Liebhart	Nokia Siemens Networks
Etienne Chaponniere	Qualcomm
Subramanian Ramachandran	Qualcomm